Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 8, October 1981
Printed in U.S.A.

Cutoff Calls and Telephone Equipment Reliability

By M. TORTORELLA

(Manuscript received March 31, 1981)

It is often difficult or expensive to measure cutoff calls, which are usually caused by failures and malfunctions in some component of the telephone network. Therefore, it is desirable to have an indirect method for estimating the number of cutoff calls caused by equipment failures in a switching system or facility. This paper discusses a mathematical model that can be used to determine the cutoff call rate in a network component as a function of the failure modes and failure rates in the component, and the call holding time distribution. It includes a discussion of a paradigm for developing reliability objectives that directly reflect service as it is seen by end users. The mathematical model, an M/M/c/c queuing system with server failures, is described. A strong law of large numbers and a central limit theorem for the number of cutoff calls—accumulated either according to the number of failures or over time—are developed. An example from a switching system is given to show how these results are applied in specific cases.

I. INTRODUCTION AND SUMMARY

The purpose of this paper is to describe a mathematical model for the rate of cutoff calls caused by failures and malfunctions in telephone equipment. The cutoff call behavior of almost any piece of telephone equipment that serves callers can be analyzed using this technique, but the primary applications we have in mind are large integrated systems containing many components, such as switching systems and transmission systems (trunk groups). The model relates the rate of cutoff calls produced by failures in the equipment and its subsystems to the failure modes in the equipment, their severity and frequency of occurrence, and the call-holding-time distribution.

The interaction of telephone call requests with service equipment has often been successfully described using queuing models. Therefore, it seems reasonable that a study of the effects of equipment failures on the calls in a telephone system should be feasible within the context of the classical queuing models of telephony. This is the approach adopted here, with the additional feature that the servers may be unreliable and subject to failures of a kind that cause the customer (if any), in service at a position whose server fails, to be dropped from the system at the time of the failure. Forys and Messerli have previously studied trunk groups containing unreliable servers.[1] Their interest was in characterizing the effect on arriving calls of one or more short-holding-time (hence, very likely to be malfunctioning) trunks in the group, whereas here the interest is primarily in the effects of unreliable servers that may fail singly or together in groups, on customers who are already in service.

The paper is divided into five sections. Section II contains a general discussion of reliability objectives as they apply to telephone equipment, and the paradigm for developing reliability objectives that directly reflect service as it is seen by the customer. We observe that the critical step that has been lacking is the ability to translate equipment reliability into rates of occurrence and duration of customer-perceivable problems, such as cutoff calls and network connection failures, that are produced by failures and outages.

In Section III, the structure of the mathematical models to be used is described. The basic structure is one of a queuing system with server failures, and, using this structure, the probability that a call in the system will be cut off is determined. The way one describes mathematically the system organization and failure modes is also covered in this section. The probability of cutoff can be computed under quite general conditions on the arrival process, the service times, and the queue discipline, because it depends only on what happens after the customer enters service.

Section IV describes a more specialized queuing model, in the context of which certain limit laws for the cumulative number of cutoff calls can be obtained. This is the M/M/c blocking system with server failures, and both a strong law and a central limit theorem are obtained. The eventual use of these limit laws, as the basis for constructing statistical tests for determining compliance with objectives, is also briefly discussed. Section V is devoted to the single-server case, and explicit calculation of all parameters of interest.

Finally, Section VI gives an example of the application of this theory to the estimation of cutoff call rates in a toll switching system. It is important to be able to do this kind of analysis because one may wish to predict cutoff call performance for a system that is still being designed. This technique is then an example of an indirect, albeit approximate, method of estimating a cutoff call rate for which no satisfactory direct method may be available.

Two appendices contain all proofs and other mathematical details that, otherwise placed, would interfere with the flow of the text.

II. RELIABILITY OBJECTIVES AND CUSTOMER SERVICE

2.1 General

It is currently recognized that the most desirable way to specify performance and service objectives for telephone network equipment is to use, in addition to economic information, considerations of how the operation of this equipment affects service as it is seen by the customer. In order to do this for reliability objectives, we need to realize that customers do not perceive outages, failures, and malfunctions as such. They are aware of them only insofar as they cause service problems detectable by users who generally are not aware of the internal operations of the telephone network. To achieve the goal of determining equipment reliability objectives based on customer needs and expectations, then, the following steps are required:

(*i*) Determine the customer-perceivable service effects of the reliability problems to be controlled.

(*ii*) Determine the quantitative relationships between the frequency and duration of reliability problems in the system or equipment and the rates of occurrence and duration of the service effects found in the first step.

(*iii*) Use these relationships to translate the customer service objectives for the system, which control the customer-perceivable effects stemming from reliability problems, into internal reliability objectives.

This paper focuses on the second step for a particular service effect: cutoff calls.

2.2 Service effects

From the customer's point of view, the primary service effects of failures and malfunctions are cutoff calls, ineffective attempts (network connection failures), isolation (line and toll), and transmission impairments (excessive loss, noise, etc.). Cutoff calls will be discussed at length below. Ineffective attempts, or network connection failures, can be caused by failures and malfunctions because the unavailability of a portion of the telephone network increases the network's blocking probability during the time this portion of the network is out of service. If the failed equipment is a customer's loop, or a part of the local central office that disables the customer's line functions, causing a customer to be unable to communicate with the local central office, the customer experiences *line isolation* for the duration of the failure.

If the failed equipment is a toll-connecting trunk group from a customer's local central office, the customer experiences *toll isolation,* meaning that toll calls to or from certain areas cannot be placed or received.

Transmission impairments can be caused by malfunctions such as equipment operating outside tolerances. These phenomena are well-understood and measurement plans are in place to return relevant information about transmission problems to maintenance forces so that abnormal conditions may be corrected. These will not be discussed further.

The rate of network connection failures and the duration of isolations are determined primarily by the duration of the outage. Thus, analysis of these service problems is helpful in determining reliability objectives and maintenance policies to limit outage duration. We will see below that the rate of cutoff calls is primarily driven by the rate of failures, so that analysis of cutoff calls is useful mainly in determining objectives for frequency of occurrence of outages. Of course, a comprehensive strategy for reliability management should deal with these complementary facets of equipment reliability in a unified way, and maintenance (service restoration and equipment repair) policies play an important part here. An objective for frequency of occurrence of failures, together with a maintenance policy, implies a certain total outage time for the equipment. Similarly, an objective for total outage time, together with a service restoration and equipment repair strategy, limits the number of times outages may occur. Although this paper deals only with cutoff calls and frequency of occurrence of outages, it should be borne in mind that a unified approach to reliability objectives, combining considerations not only of cutoff calls and outage frequency but also of network connection failures and outage duration, is most desirable.

2.3 Types of failures included

2.3.1 Causes of cutoff calls

A cutoff call is a connected (stable) call that has been terminated other than by an on-hook by either party. The event of termination is sometimes referred to as a *cutoff,* for short (as is a call that is so affected). The terminology is intended to connote an unintentional, unexpected interruption. International Telegraph and Telephone Consultative Committee (CCITT) terminology refers to a cutoff-causing failure in a switching system as a "premature release malfunction in an exchange."

Cutoff calls are caused by equipment failures (including recovery actions), and other external factors, such as radio fades and in-band talkoff (simulation of the 2600-Hz supervisory signal by a signal emitted by one of the parties). The termination takes place at the instant the failure or other event begins, so the rate of cutoffs is influenced primarily by the rate of failures (this is demonstrated in eq. (7)). Cutoffs are related to reliability, then, just as ineffective attempts or network connection failures are related to availability. To determine the rate of cutoff calls seen by a telephone user, the cutoff call performances of individual switching and transmission systems are combined in a network model. A suitable model is one for the reliability of a series system consisting of switching systems and trunk groups.

2.3.2 Scope of the model

The reliability problems covered by the model are those of failure and repair of entire systems and parts of systems, and those failures and malfunctions that may not completely disable a system or subsystem, but that cut off calls when they occur. In the first case, systems and subsystems will be considered to be either operating properly and fully available for use, or not operating at all and unavailable. Cutoff calls caused by improper operation, or operation outside tolerances, of a system or subsystem can also be treated. The key notion is that any event that causes cutoff calls when it occurs can be called a "failure" for purposes of this discussion. The model can accommodate many different "failure" modes, as long as the occurrence times and severities of these events can be characterized sufficiently well that failure processes and cutoff impacts (Section 3.2) can be assigned. In particular, the model could in principle include such events as radio fades and in-band talkoff as "failure modes." However, in studying cutoff calls as related to equipment reliability, this is not recommended, because these are external events, not caused by an equipment failure or malfunction which could be controlled by preventive or corrective action by the telephone company.

As for causes of failure, for the model there is no restriction on the cause of the failure or malfunction. All that is required is that one be able to list the kinds of events that cause cutoffs, and describe probabilistically the times between incidents for each kind of event. The scope of this work encompasses all failures which lead to cutoff calls, regardless of cause, including hardware (component failure), software and firmware faults, human intervention errors, office database errors, and so on.

2.4 Uses of the mathematical model

This model finds three primary applications in system analysis and design. First, it can be used to make the translation which allows system cutoff call objectives to determine reliability objectives and maintenance policies for switching and transmission systems. Reliability objectives should not be viewed as ends in themselves, but only as means by which objectives for those aspects of customer service that are affected by reliability problems can be met. Second, they have value as predictive tools. System designers can use the probability of cutoff as a figure of merit for hypothetical system designs, architectures, and reliability characteristics. Systems that have not yet been constructed can be compared for this aspect of service quality, and this comparison can be a factor in deciding among competing designs, for example. Its third major use is to provide a framework within which to perform statistical tests, based on observed cutoff call rates, to see whether objectives are being met. In systems where cutoff calls are not measured, the models enable inferences to be made about the cutoff call rate based on other kinds of data, such as reliability records of equipment failures and malfunctions. Since cutoff calls are often difficult or expensive to measure in a given system, these techniques provide another, perhaps more attractive, means of understanding this important service problem.

III. MODEL DESCRIPTION AND PROBABILITY OF CUTOFF

In this section, we discuss the structure of the mathematical model for cutoff calls and reliability of telephone equipment. It starts with an outline-like guide to the sequence of results which make up the mathematical model. As an aid to seeing where the details fit into the overall scheme, this guide can be referred to while reading the remainder of the paper. A queuing model with server failures is covered, as is the organization of the servers and failure modes. Physical interpretation is given, and some probabilistic insights are added to help clarify the ideas. Finally, the probability that a call that has been accepted by the system will be cut off is computed.

3.1 Outline of results

3.1.1 Relation of probability of cutoff to equipment reliability

The first important result obtained is in Section 3.6, where the probability that a call that has been accepted by the system will be cut off is computed. This probability can be thought of as a figure of merit for the system in question, and can be computed under weak assumptions about the arrival process, the holding times, and the interfailure times. However, the probability of cutoff, by itself, is not enough to give a good understanding of how a system will behave with respect to cutting off calls. In particular, there are two important questions on which knowing the probability of cutoff alone sheds no light. First, does the observed cutoff call rate have any relation to the probability of cutoff? Second, what is the structure of the stochastic process which counts the number of calls cut off in a time interval? How much variability can be expected in such a count, for example?

3.1.2 Measurements and consistent estimation of the probability of cutoff

Section IV is devoted to an exploration of these questions for a more specialized system, the M/M/c/c queue with server failures. In answer to the first question, Corollaries 5 and 6 show that the observed cutoff call rate converges to the probability of cutoff as given by eq. (8). This means that, in this case, measurements can be relied upon to consistently estimate the probability of cutoff, which may be controlled by an objective. Also, when a prediction about the cutoff probability in a new system is made, it can reasonably be expected that the cutoff call rate shown by the system in operation will approach the predicted value (subject, of course, to the quality of the inputs to the prediction).

3.1.3 Asymptotic distribution of the number of cutoff calls

In answer to the second question, Theorems 7 and 9 show that the number of cutoff calls is, when suitably normalized, asymptotically normally distributed. The asymptotic variance of the number of cutoff calls [Theorem 8(b)], together with the asymptotic normality, suggests the variability to be expected in the observed (normalized) number of cutoff calls: about 63 percent of observations fall within one standard deviation of the mean, etc. Finally, the asymptotic distribution of the number of cutoff calls could be used as the basis for a statistical test for determining whether the objective is being met, although this is not accomplished in this paper.

3.2 Mathematical description of cutoff call model

The equipment will be modeled as a c-server queuing system. Calls (requests for service) arrive at the system at times $\tau_1, \tau_2, \cdots$. Denote by τ_n' the time that the nth arrival enters service. If this is a blocking system and all servers are occupied at time τ_n, the nth arrival never enters service, and for later convenience, τ_n' will be taken to be $-\infty$ in this case. Throughout Section III the arrival process may be any arbitrary point process. Each call has associated with it a (nonnegative) holding time that it wishes to spend using the resources of the system. It is assumed that a single call occupies only a single server in the system during its entire holding time (this will be important later in discussion of the organization of the failure modes). The holding times are denoted by $Y_1, Y_2, \cdots$, and are taken to be mutually independent and identically distributed, and independent of the arrival process.

So far, we have just described an ordinary queuing model. The additional feature that distinguishes the models including equipment failure is that the servers may be unreliable. That is, at certain (random) times, all the servers, or certain groups of servers, may cease serving the customers at their positions, and the affected customers will be forced to depart prematurely from the system at these times. Adopting the natural physical terminology for the mathematical model, these customers will be said to have been "cut off." Suppose that there are m different failure modes in the system. That is, there are m different ways in which various groups of servers (and possibly all servers) can fail in such a way as to cause cutoffs at the instant the failure begins. Any particular server may be affected by many failure modes, and many different configurations of failed servers may be included in a single failure mode. For example, suppose a switching system having 1200 terminations (lines and trunks) is made up of ten identical units, each serving 120 terminations. Then this system has a failure mode at 120 servers (terminations)—this would not be counted as ten separate failure modes if all these units had the same failure characteristics. With each failure mode, associate a renewal process listing the times at which failures of this type occur. These m processes will be called "failure processes." Let F^i be the distribution of the interrenewal times for the ith process, and let λ_i be the reciprocal of the mean time between renewals, $\lambda_i^{-1} = \int_0^\infty x \, dF^i(x)$. Let the epochs in the ith failure process be denoted by S_1^i, S_2^i, $\cdots$. It is assumed that these failure processes are mutually independent and independent of the arrival- and holding-time processes. The latter independence assumption is reasonable when the arrivals have no prior knowledge about the state of the system at the time of arrival.

Also associated with the ith failure mode is a number p_i between zero and one. The quantity p_i represents the probability that a call in the system will be cut off when a failure of type i occurs, and is called the *cutoff impact* of failure mode i. The severity of a failure of type i is indicated by p_i. If $p_i = 1$ then the ith failure mode is an entire system failure, and, with probability one, all calls in service are cut off when such a failure occurs. If, on the other hand, p_i is close to zero, then this describes a minor failure, and fewer calls will be cut off when such a failure occurs. We will take $p_i \neq 0$ for every i since a failure mode with cutoff impact zero can be ignored.

3.3 Correspondence with physical situation

Imagine a call using the resources of some telephone system (for definiteness, say a switching system), in either the setup phase or the conversation (stable) phase. Many elements of the system are used to provide and maintain the conversation path that is the electrical connection from one side (incoming or originating) of the system to the other (outgoing or terminating). Failure of some of these elements may cause the call to be dropped from the system without an on-hook by either party. In the queuing model, it is not these elements that are thought of as the servers. Rather, a single call is thought of as occupying a single server, such as a pair of terminations or a path through a system, which may be subject to being disabled by the failure of some of these elements. From this point of view, any particular server may be affected by several failure modes.

3.4 Probabilistic interpretation

Before turning to the computation of the probability that a call that has been accepted by the system will be cut off, the following probabilistic heuristics are offered as an aid to clarifying the idea of the model.

The event that a call in the system is cut off can be conceptualized as a realization of a competition process. Suppose a call having holding time Y enters the system at time t. At the entrance time t, m clocks are set running, with the ith clock's running time having the distribution of the excess lifetime of the time between failures for the ith failure process at time t. If the holding time Y expires before any of the clocks run down, no failures occur and, hence, no cutoff can occur. If one of the clocks runs down first (say the jth one), a biased coin ($P\{\text{heads}\} = p_j$) is tossed. If the coin comes up heads, the call is cut off, and the experiment stops for this call. If the coin comes up tails, the call is not cut off, and the experiment continues, with the jth clock now running according to the distribution F^j. For this call, the experiment stops either when it has been cut off or when it departs normally from the system.

The computation, which is performed in the next section, follows this description by first determining the probability of no cutoff and then subtracting from one.

3.5 Probability of cutoff

With this section, we begin following the outline of Section 3.1. The sequence of results and their proofs is simply a mathematical translation of the description given in Section 3.4. Lemma 1, while of independent interest, is used here only in establishing the main result of this section, which is Theorem 2.

Lemma 1: *Let* $\{N(t)\colon t \geq 0\}$ *be a renewal counting process with interrenewal time distribution* F. *Then for* $t,\ y \geq 0$ *and* $k \geq 1$, *the probability that there are k renewals in the interval* $[t, t + y]$ *is given by*

$$\int_0^t g_k(t - s,\ y)dM_0(s), \tag{1}$$

where

$$g_k(u,\, y) = \int_u^{u+y} [F_{k-1}(u + y - x) - F_k(u + y - x)]dF(x), \tag{2}$$

with F_k the k-fold convolution of F with itself, F_0 equals V, the standard right-continuous unit step function with jump at the origin, and M_0 the augmented renewal function for the process. For $k = 0$, the probability that there are no renewals in this interval is given by $\int_0^t [1 - F(t + y - s)]dM_0(s)$.

Theorem 2: Let $\bar{M}_0^i$ be the augmented renewal function for the defective distribution $(1 - p_i)F^i$,

$$\bar{M}_0^i(x) = \sum_{k=0}^{\infty} (1 - p_i)^k F_k^i(x), \tag{3}$$

and let

$$\bar{g}_i(u,\, y) = 1 - F^i(u) - p_i \int_u^{u+y} \bar{M}_0^i(u + y - x)dF^i(x). \tag{4}$$

Then the probability that a call entering the system at time t is cut off is given by

$$1 - \int_0^{\infty} \left[\prod_{i=1}^{m} \int_0^t \bar{g}_i(t - s,\, y)dM_0^i(s) \right] dH(y). \tag{5}$$

In the limit as t approaches infinity, this becomes

$$1 - \int_0^{\infty} \prod_{i=1}^{m} \left[1 - \lambda_i p_i \int_0^{\infty} \bar{g}_i(u,\, y)du \right] dH(y). \tag{6}$$

If the arrival process is independent of the remaining queuing and failure processes, the probability that the nth call will be cut off, given that it enters the system, can be computed by integrating eq. (5) against the distribution of τ_n'. In case all the failure processes are stationary Poisson processes, the probability of cutoff is constant and does not depend on the entrance time of the call.

Corollary 3: Suppose $F^i(x) = 1 - e^{-\lambda_i x}$ for $i = 1, \cdots, m$. Then every call in the system has probability of cutoff given by

$$1 - \int_0^{\infty} \exp\left(- \sum_{i=1}^{m} \lambda_i p_i y \right) dH(y). \tag{7}$$

If, in addition, the call-holding-time distribution is exponential, $H(y) = 1 - e^{-\nu y}$, the probability of cutoff reduces to

$$\theta = \frac{\sum_{i=1}^{m} \lambda_i p_i}{\nu + \sum_{i=1}^{m} \lambda_i p_i}. \tag{8}$$

These are obtained by appropriate substitution in eq. (5).

3.6 Discussion

The probability that a call already in the system will be cut off has been computed for a queuing system with unreliable servers. The arrival process and queue discipline may be arbitrary; this is a reflection of the fact that the event of cutoff depends only on what happens after the call enters the system. The limiting argument used to establish eq. (6) can be carried out even if the arrival process depends on the service time process (as in systems with state-dependent arrival rates), although the probability that the nth call will be cut off is more difficult to compute in this case. We have assumed the service times are independent and identically distributed. This could be relaxed, but for most ordinary message telephone service applications it does not seem necessary to introduce this complication. As can be seen from eq. (7), great simplification results if it can be assumed that the failure processes are stationary Poisson processes. In practice, this assumption has often been used because, in studying large systems from a great distance, data that would enable one to characterize the failure processes in the system in more detail are often not available. When the conditions that obtain in the physical situation are difficult to identify exactly, it may not be possible to determine the information needed to make successful application of a more general model.

IV. A MARKOV MODEL AND SOME LIMIT LAWS

4.1 Introduction

In Section IV we deal, for a more specific queuing system, with the second two items in the outline in Section 3.1. There are many ways to particularize the general considerations discussed in Section III, depending on the underlying queuing model. For purposes of estimation of cutoff call rates in telephone systems, certainly it is desirable to allow the most general model possible. This might be a transient analysis of a queue in which, in addition to the exogenous arrivals, there may be feedback and retrials by rejected and cutoff customers, and general service and interfailure times. Unfortunately, analytic treatment of such a complicated model is not within reach. The asymptotic analysis of such general queues, even with perfectly reliable servers, is accomplished only approximately in many cases.

Here, instead, we will study about the simplest of stochastic models for this situation, the M/M/c/c queue with stationary Poisson failure processes. This decision results from informal consideration of the tradeoff between realism of description on one hand and possibility of successful execution of analysis on the other. Even in this simple case, there are many interesting difficulties. For example, solving numerically the Chapman-Kolmogorov equations (Appendix A) for the invariant distribution of the embedded chain (Section 4.3) is likely to be easier than obtaining qualitative insight through analytic solution of these equations. No representation is hereby made that the Markovian assumptions are particularly accurate in representing reality, or that the asymptotic results obtained well describe transient behavior. Nevertheless, the assumptions are not such gross distortions of the physical situation that they render such models useless, and the study of simpler models has several important virtues to recommend it. Solutions can be obtained, the general features of the underlying situation remain visible without the technical details that sometimes obscure the main ideas, directions for the generalizations that are likely to be successful on more complicated models are suggested, and, last but not least, results can be checked against data to determine if more general models are required. The Markovian model to be described has been successfully used in the switching systems area, and predictions made from it have shown reasonable agreement with data. This is not to say that further refinements of these models would not be valuable. Such refinements would be interesting and useful advances in the state of the art.

4.2 Specifications and notation

In the M/M/c blocking system, let α be the arrival rate, ν be the service rate, and let $\{A(t) : t \geq 0\}$ denote the arrival process. The m failure processes are all stationary Poisson processes with rates $\lambda_1, \cdots, \lambda_m$, all positive. (In the example in Section 3.2, the failure rate for the 120-termination *failure mode* would be ten times the failure rate of a single 120-termination *unit*.) The system will be assumed to recover instantaneously from failures, so that the only effect that a failure has is to cause some of the calls in the system to depart prematurely, before the completion of their intended holding times. Failures, therefore, have no effect on calls that are not already in the system. For example, they do not cause an increase in the blocking probability of the system. Clearly this is only an approximation to the true situation, but it seems to produce acceptable results, for several reasons. First, in practical cases, the ratio of average outage time to mean time between failures is usually small; here this small number has been replaced by zero. Secondly, in this approximation the total number of cutoff calls tends to be overestimated because more calls are accepted into the system than would be if the failure durations were positive. This means that more calls are exposed to the possibility of being cut off. Again, if the times between failures are long compared to the outage times, the cutoff call rate (number of cutoff calls divided by number of arrivals or number of accepted calls) will not be badly distorted by this approximation.

The failure processes interact with the queuing processes in the following way. Let $B(t)$ denote the number of busy servers at time t, $t \geq 0$, including the effects of failures (as below), and let $C(t, r)$ be the number of busy servers at time t in an ordinary (no server failures) M/M/c/c system when there are r in the system at time 0. Then, whenever a failure of type i occurs, the probability that a call in the system will be cut off is p_i, and the cutting-off events for each of the calls in the system at that time are assumed to be mutually independent, as are the cutting-off events corresponding to different failure times. (Simultaneous failures occur with probability zero since the distributions of the interfailure times are all continuous.) This models a situation in which the calls in service at any time are more or less regularly spread out over the servers in the system, and all parts of the system subject to a given failure mode are equally vulnerable. The independence, on the other hand, is invoked to reflect the fact that this regular distribution obtains only perhaps in a very broad, average way, and at any given failure epoch, the server occupancy might be quite irregular. At each epoch in each failure process, then, the number of calls cut off is a binomial random variable with parameters given by the number of busy servers at that epoch and the cutoff impact of that failure mode. That is, at time S_n^i, if $B(S_n^i) = k$, the number of calls cut off is binomially distributed with parameters k and p_i. Sometimes many of the calls in the system will be carried by the unit (group of servers) experiencing the failure; sometimes proportionately fewer calls will be carried on this unit. The binomial model provides an approximate description of this situation. This is a compromise between a very detailed model that keeps track of individual server busy and idle times and the individual identities and times of failure of server groups, and a deterministic model having the number of cutoffs at S_n^i equal to $p_i B(S_n^i)$, which is unrealistic for being too regular.

4.3 The embedded Markov chain

As defined, $B(t)$ is a pure jump process; even with cutoffs caused by failures accounted for, all sample paths can be assumed to be continuous from the right. Pool the failure processes and denote the resulting stationary Poisson process by $\{S_1, S_2, \cdots\}$. Define $B_n = B(S_n^-)$ $(n = 1, 2, \cdots)$; B_n is the number of busy servers just before the nth failure of any kind. The sequence $\{B_n: n = 1, 2, \cdots\}$ is a Markov chain, called the *embedded chain*, with state space equal to $\{0, 1, \cdots, c\}$. The survivors in the system at time S_n^+ have the same exponentially distributed service times as new arrivals do, and their number is determined only from B_n. The number of arrivals in $[S_n, S_{n+1}]$ is independent of the number of arrivals before S_n. Note that the strong Markov property is not required of the arrival process, for while the failure epochs are random times, they are not determined by the arrival process because of the assumed independence.

4.4 Properties of the embedded chain

Let W_n denote the number of calls cut off by the failure that occurs at time S_n. Then, for each n, the conditional distribution of W_n, given B_n, is a mixture of binomials:

$$P\{W_n = w | B_n = b\} = \frac{1}{\lambda} \sum_{i=1}^{m} \lambda_i \binom{b}{w} p_i^w (1 - p_i)^{b-w},$$

$$b = 0, \cdots, c; \; w = 0, \cdots, b. \quad (9)$$

Here λ_i/λ is the probability that the nth event in the pooled process comes from failure process i ($\lambda = \lambda_1 + \cdots + \lambda_m$). Denote the right-hand side of eq. (9) by q_{bw}.

Finally, note that the W_n's are conditionally independent, given the B_n's, because of the independence of the cutting-off events corresponding to different failure times. That is,

$$P\{W_{i_1} = w_1, \cdots, W_{i_n} = w_n | B_{i_1} = b_1, \cdots, B_{i_n} = b_n\}$$

$$= \prod_{k=1}^{n} P\{W_{i_k} = w_k | B_{i_k} = b_k\} \quad (10)$$

for all positive integers $n, i_1, \cdots, i_n$.

The properties of the B_n-process can be most readily obtained from the fundamental representation

$$B_{n+1} = C(S_{n+1} - S_n, B_n - W_n),$$

where the equality is equality in distribution. That is, the number of busy servers at (just before) S_{n+1} has the same distribution as the number of busy servers in an ordinary (no server failures) M/M/c/c system running for time $S_{n+1} - S_n$ with $B_n - W_n$ (the number of survivors in the system at time S_n^+) calls in the system at time zero. It has already been observed that $\{B_n: n = 1, 2, \cdots\}$ is a Markov chain; straightforward conditioning arguments and appeal to the independence of the failure and queuing processes establish that its transition probabilities are given by

$$P\{B_{n+1} = j \mid B_n = i\}$$

$$= \sum_{k=0}^{i} \sum_{r=1}^{m} \lambda_r \binom{i}{k} p_r^{i-k}(1 - p_r)^k \int_0^\infty P\{C(x, k) = j\} e^{-\lambda x} dx. \quad (11)$$

These are independent of n, so the chain has stationary transition probabilities. Denote them by p_{ij}. We remark that if $p_r = 1$ for every r, these reduce to

$$p_{ij} = \int_0^\infty P\{C(x, 0) = j\} \lambda e^{-\lambda x} dx,$$

so that $\{B_n\}$ are mutually independent in this case. Also, if the failure processes are not stationary Poisson, but are, say, renewal, then the p_{ij} are still well-defined, although they take a different form. In particular, they then depend on n, and while $\{B_n\}$ is still a Markov process, it does not have stationary transition probabilities. Some of the following results (particularly those about recurrence) continue to hold in this case, but limit laws are harder to obtain.

Riordan gives the distribution of $C(x, k)$:[2]

$$P\{C(x, k) = j\} = \frac{\rho^j}{j!} \left(\sum_{i=0}^{c} \frac{\rho^i}{i!} \right)^{-1}$$

$$+ \frac{c!}{j!} \rho^{c-k} \sum_{i=1}^{c} \frac{D_k(r_i) D_j(r_i)}{r_i D_c(r_i) D_c'(r_i + 1)} e^{r_i \nu x}, \quad (12)$$

where $\rho = \alpha/\nu$, the D_n are related to the Poisson-Charlier polynomials c_n (Ref. 3) by $D_n(s) = \rho^n c_n(-s)$, and $r_1, \cdots, r_c$ are the roots of $D_c(s + 1)$. These roots are all real and negative so that the $e^{r_i \nu x}$ all vanish as $x \to \infty$, and the $P\{C(x, k) = j\}$ approach the well-known Erlang equilibrium probabilities, independent of k. Equation (12) shows that $P\{C(x, k) = j\}$ is an analytic function of x that is not identically zero, so that its zeros, if any, are isolated. Thus, there is a set of positive measure in $[0, \infty[$ on which $P\{C(x, k) = j\} > 0$. This means that $\int_0^\infty P\{C(x, k) = j\} \exp(-\lambda x) dx$ is positive for every j and k, and so $p_{ij} > 0$ for every i and j. This positivity shows the $\{B_n\}$ chain to be irreducible and aperiodic. Since the chain is finite, all states are positive recurrent (Ref. 4, Section I.XV.6).

4.5 The induced Markov chain

The two-dimensional process $\{(B_n, W_n) : n = 1, 2, \cdots\}$ is again a Markov chain whose transition probabilities are given by $r_{s_i s_j} = q_{b_j w_j} p_{b_i b_j}$, where $s_i = (b_i, w_i)$. That is,

$$P\{(B_{n+1}, W_{n+1}) = s_j \mid (B_n, W_n) = s_i\} = r_{s_i s_j} = q_{b_j w_j} p_{b_i b_j}.$$

Use is made here of eq. (10). This chain will be called the *induced chain.*

It is desirable for the induced chain to inherit the properties of the embedded chain discussed in Section 4.4. To obtain this, it would be sufficient to have $q_{ij} > 0$ for all i and j. From eq. (9), this is satisfied, unless $p_i = 1$ for every i. The case $p_i = 1$ for every i is a trivial special case of what is to follow, because then $W_n = B_n$ with probability one, for every n. Also, for large systems with many failure modes, this case is of little interest. For these reasons, we will suppose that there is at least one i for which $p_i < 1$. Under this condition, $r_{s_i s_j} > 0$ for every i and j, and since the induced chain is also finite (its state space is $\{(b, w): b = 0, 1, \cdots, c, w = 0, 1, \cdots, b\}$), it is irreducible, aperiodic, and positive recurrent, just as the embedded chain was.

4.6 Stationarity

Since service objectives represent long term goals for system operation, it is appropriate to compare the equilibrium features of the model against the service objectives.

Since both the embedded and induced chains are positive recurrent, they are both ergodic. The embedded chain has an invariant distribution $\{u_k: k = 0, \cdots, c\}$ given by

$$u_k = \lim_{n \to \infty} p_{ik}^{(n)},$$

independent of i. As usual, the parenthesized superscript indicates the n-step transition probability. Furthermore, $u_k > 0$ for each k, $\sum_{k=0}^{c} u_k = 1$, and $u_k = \sum_{i=0}^{c} u_i p_{ik}$ (Ref. 4, Section I.XV.7). To say that the system has been in operation for a long time can be expressed by taking $\{u_0, \cdots, u_c\}$ to be the distribution of the number of busy servers at time zero. With this choice of initial distribution, $\{B_n\}$ becomes a strictly stationary process.

The induced chain also has an invariant distribution, denoted by $\{v_{(0,0)}, \cdots, v_{(c,c)}\}$. It is easy to see that $v_{(b,w)}$ is given by $v_{(b,w)} = q_{bw} u_b$, $b = 0, \cdots, c$; $w = 0, \cdots, b$. The induced chain can also be made strictly stationary by taking its initial distribution to be its invariant distribution.

4.7 A strong law of large numbers

The quantity of basic interest in this study is the cumulative number of cutoff calls, $\chi_n = W_1 + \cdots + W_n$. This section is devoted to describing a strong law of large numbers for χ_n and some of its ramifications. This addresses the second item in the outline of Section 3.1.

In general, $\{W_n: n = 1, 2, \cdots\}$ is not a Markov process. However, it can be written as a functional of the induced chain. The appropriate functional to choose is π_2, the projection onto the second coordinate: $W_n = \pi_2(B_n, W_n)$ for each n. π_2 is clearly a measurable function on the σ-field of the induced chain, and so the limit theorems of Sections V.5 and V.7 of Ref. 5 may be applied to $\{W_n\}$.

Let $Z(t)$ be the number of calls accepted by the system in $[0, t]$,

$$Z(t) = \sum_{n=0}^{\infty} V(t - \tau_n'),$$

and put $Z_n = Z(S_n^-)$.

Theorem 4: *χ_n/n converges with probability one to $\theta \lim_{n\to\infty} (EZ_n/n)$.*

Corollary 5: *χ_n/Z_n converges to θ in expectation and with probability one.*

The proofs of these results can be found in Appendix B.

Now this is not quite what is required for applications. Generally, one does not count either carried calls or cutoff calls indexed by the times of failure $S_1, S_2, \cdots$. Rather, what one does is keep a running count of these items indexed by a continuous time parameter. Accordingly, let $\chi(t)$ denote the total number of calls cut off in $[0, t]$; one has

$$\chi(t) = \sum_{n=1}^{\infty} W_n V(t - S_n) = \chi_{\max\{n: S_n \leq t\}}.$$

Corollary 6: *$\chi(t)/Z(t)$ converges to θ in expectation and with probability one.*

Applications of these results have been discussed in Section 2.1. In a stable Markovian environment, Corollary 6 says that the natural estimator of the probability of cutoff in the system, namely the cutoff call rate, is strongly consistent. The implication for measurement is that for systems in operation, measurements can be relied upon to estimate the underlying cutoff call rate that is characteristic of the system. The extension of these results to other than Markovian queues would provide even better approximations when the environment can be more precisely specified. The implication for system design is that once it is configured with certain failure modes, etc., its cutoff call rate, in the appropriate environment, will be as predicted, subject to sets of probability zero and the quality of the failure rate predictions.

Before turning to central limit theorems, a partial indication of the rate of approach to steady state will be given.[6] For this purpose, assume that $c = \infty$ (so that all arriving calls immediately enter service) and that $p_i = 1$ for all i (so that every time a failure occurs, all calls in the system are cut off). Then it can be shown that

$$E\left(\frac{\chi(t)}{Z(t)}\right) = E\left(\frac{\chi(t)}{A(t)}\right) = \frac{\lambda}{\lambda+\nu}(1-e^{-\lambda t})\left[1-\frac{1-e^{-(\lambda+\nu)t}}{(\lambda+\nu)t}\right]. \quad (13)$$

Eq. (13) can be used to estimate relative errors after different times. Let $R(t) = \left(\frac{\lambda}{\lambda+\nu}\right)^{-1}\left[\frac{\lambda}{\lambda+\nu} - E\left(\frac{\chi(t)}{A(t)}\right)\right]$; then $100R(t)$ is the percentage error in $E\left(\frac{\chi(t)}{A(t)}\right)$ as an estimate of θ after t time units. Using eq. (13),

$$R(t) = e^{-\lambda t} + \frac{1}{(\lambda+\nu)t}(1-e^{-\lambda t})(1-e^{-(\lambda+\nu)t}). \quad (14)$$

Measuring time in minutes, with $\lambda = 0.003$ (about three failures per day) and $\nu = 0.166$ (six-minute average call holding time), the percentage errors, from eq. (14), are 85 after one hour, 35 after six hours, 12 after 12 hours, and 2 after 24 hours.

4.8 A central limit theorem

The existence of an asymptotic normal approximation for the cumulative number of cutoff calls makes the construction of statistical tests easier. In this section, we discuss these approximations in discrete and continuous time. This addresses the third item in the outline of Section 3.1.

The central limit theorem for χ_n follows directly from the central limit theorem for functionals defined on a Markov chain, for example, see Theorem V.7.5 in Ref. 5.

Theorem 7: *There are positive numbers μ and σ for which*

$$\lim_{n\to\infty} P\left\{\frac{\chi_n - \mu n}{\sigma\sqrt{n}} \le x\right\} = \Phi(x),$$

where $\Phi(x)$ is the standard normal integral.

This requires little discussion: the condition (D_0) and the moment condition of theorem V.7.5 of Ref. 5 are satisfied because the induced chain is finite and positive recurrent. The interesting results are the values of the centering and scale parameters. It is easy to see that

$$\mu = EW_1 = \sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_iEB_1 = \sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\sum_{b=0}^{c}bu_b = \sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\sum_{b=0}^{c}\frac{b}{m_{bb}}, \quad (15)$$

where m_{ab} is the mean first passage time from state a to state b in the embedded chain. (If $a = b$, this is a mean recurrence time).

Theorem 8(*a*): *The asymptotic variance of the partial sums of the B_k's is*

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left(\sum_{k=1}^{n} B_k\right) = \sum_{a=0}^{c} \sum_{b=0}^{c} \frac{ab}{m_{aa}m_{bb}} \left(\frac{m_{bb}^{(2)}}{m_{bb}} - 2m_{ab}\right) + \sum_{b=0}^{c} \frac{b^2}{m_{bb}}, \qquad (16)$$

where $m_{bb}^{(2)}$ is the second moment of the recurrence time for state b in the embedded chain.

Theorem 8(*b*): *The scale constant in the central limit theorem is*

$$\sigma^2 = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} p_i(1 - p_i) \sum_{b=0}^{c} \frac{b}{m_{bb}} + \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} p_i^2 \sum_{b=0}^{c} \frac{b^2}{m_{bb}}$$

$$+ \left(\sum_{i=1}^{m} \frac{\lambda_i}{\lambda} p_i\right)^2 \sum_{a=0}^{c} \sum_{b=0}^{c} \frac{ab}{m_{aa}m_{bb}} \left(\frac{m_{bb}^{(2)}}{m_{bb}} - 2m_{ab}\right). \qquad (17)$$

We have written the centering and scale parameters in terms of the moments of the recurrence times for the embedded chain. These can be found by solving for the invariant distribution of the embedded chain (Appendix A). The mean recurrence times are then just the reciprocals of the elements of the invariant distribution, and the second moments can be obtained from the first moments by using Theorem I.11.7 of Ref. 7. The mean first passage times m_{ij} can be found by solving another system of linear equations, for example see Theorem 6–7A of Ref. 8. For even moderate values of c, it appears that the wisest thing to do in applications is to solve the system of eqs. (25) numerically. The single-server case is treated explicitly in the next section, and it can be seen that even in this case, the computations are extensive.

In continuous time, the central limit theorem looks slightly different. This is because counting the number of cutoffs according to the number of failures, rather than over time, introduces a random time transformation with scale λ.

Theorem 9: *The distribution of the normalized cumulative number of cutoff calls over time,*

$$\frac{\chi(t) - \lambda\mu t}{\sigma\sqrt{\lambda t}}, \qquad (18)$$

converges weakly to the cumulative distribution function (*cdf*) *of a normal random variable having mean zero and variance* $1 + \mu^2/\sigma^2$.

V. THE SINGLE-SERVER SYSTEM

In this section we discuss in detail the results of the previous sections as they apply to the single-server system. We will explicitly solve for the invariant distributions and, thereby, be able to represent the parameters of the limit laws in terms of the arrival, service, and failure rates.

If there is only a single server, we will suppose that there is only one failure mode, of rate λ and cutoff impact p. Certainly if all failures are complete failures, $p = 1$. We can allow $p < 1$ to account for malfunctions which may only sometimes cut off calls. Other failure modes with other severities could be allowed. Solving the Chapman-Kolmogorov system of eqs. (25) and making use of Theorem 10 we obtain

$$u_0 = r_0 = \frac{p\lambda + \nu}{p\lambda + \nu + \alpha}, \quad u_1 = r_1 = \frac{\alpha}{p\lambda + \nu + \alpha}. \tag{19}$$

From eqs. (11) and (12) we obtain the transition probabilities

$$\begin{aligned} p_{00} &= \frac{\nu + \lambda}{\lambda + \nu + \alpha}, \\ p_{01} &= \frac{\alpha}{\lambda + \nu + \alpha}, \\ p_{10} &= \frac{\nu + p\lambda}{\lambda + \nu + \alpha}, \\ p_{11} &= \frac{(1 - p)\lambda + \alpha}{\lambda + \nu + \alpha}. \end{aligned} \tag{20}$$

The mean first passage and recurrence times can then be obtained as indicated in Section 4.8:

$$\begin{aligned} m_{00} &= \frac{p\lambda + \nu + \alpha}{p\lambda + \nu}, \\ m_{01} &= \frac{\lambda + \nu + \alpha}{\alpha}, \\ m_{10} &= \frac{\lambda + \nu + \alpha}{p\lambda + \nu}, \\ m_{11} &= \frac{p\lambda + \nu + \alpha}{\alpha}. \end{aligned} \tag{21}$$

Let σ_{11}^2 be the variance of the recurrence time for state 1. By using Theorem I.11.7 of Ref. 7, we obtain

$$\sigma_{11}^2 = \frac{(p\lambda + \nu)((2 - p)\lambda + \nu + \alpha)}{\alpha^2}. \tag{22}$$

Since σ^2 from Theorem 8(b) reduces to $p\sigma_{11}^2/m_{11}^3$ in case $c = 1$, we obtain

$$\sigma^2 = \frac{p\alpha(p\lambda + \nu)((2 - p)\lambda + \nu + \alpha)}{(p\lambda + \nu + \alpha)^3}. \tag{23}$$

This is the scale constant for Theorem 7. The centering constant is

$$\mu = \frac{p\alpha}{p\lambda + \nu + \alpha}.$$

VI. APPLICATIONS

These results have been applied at Bell Laboratories to predict the cutoff call performance of certain toll switching systems, and to evaluate reliability objectives for these systems on the basis of a determination of whether the cutoff call objective for the system can be met if these reliability objectives are followed.

In one such example, a system terminating 22,000 trunks was considered. Thirteen failure modes that were significant for cutoff calls in the system were identified. Table I lists, for each failure mode, the size of the unit failing or the number of terminations affected by the failure, the failure rate expressed as a mean number of failures per year, and the cutoff impact. For most of the failure modes, there was more than one type of unit or subsystem of the given size. The failure rates of all the units or subsystems of a single size were added together to obtain the failure rate for that failure mode. This is done because we are going to assume uniform distribution of calls over terminations, as discussed in Section 4.2. If more precise information on the distribution of calls over terminations or location of failed units is available, it may be more reasonable not to pool, but to carry individual information, as appropriate.

Every stable call in the system must occupy two terminations, one incoming and one outgoing. For a particular call, the failed unit or subsystem may be on the incoming side of the switch, the outgoing side, or both, or neither. Then the estimation of the cutoff impact of a failure mode is like a problem in sampling without replacement in which one counts the number of paths through the switch that contain the failed unit or subsystem. If the total number of terminations on the switch is N and the number of terminations affected by a failure of type i is n_i, then the cutoff impact for failure mode i is

Table I—Failure modes, frequencies, and cutoff impacts for example in Section VI

Failure Mode	Terminations Affected	Failures per Year	Cutoff Impact
1	22,000	0.248	1.0
2	5,500	0.195	0.438
3	4,080	0.077	0.337
4	2,040	0.0004	0.177
5	1,920	0.355	0.167
6	840	0.482	0.075
7	512	10.819	0.046
8	128	66.667	0.012
9	120	0.263	0.011
10	32	22.727	0.003
11	16	20.0	0.0015
12	8	217.391	0.0007
13	1	1030.0	0.0001

$$p_i = \frac{n_i(2N - n_i - 1)}{N(N-1)}. \tag{24}$$

Using eq. (8) we find that in a Markovian environment, the probability that a call entering the system will be cut off because of one of these failures is 0.24×10^{-4}, when the mean call-holding time is six minutes. Based on this, it was concluded that a sufficient margin of safety existed to ensure that the system's cutoff call objective would be met, even after allowing for possible errors in the specification of failure modes and rates, and other possibilities that could not be accounted for in the analysis.

VII. ACKNOWLEDGMENT

I would like to thank J. J. Kulzer and R. F. Serfozo for helpful discussions. I would also like to thank N. A. Marlow, who showed me how to obtain Theorem 8(a), and the referee, who suggested some simplifications in Theorem 4 and its corollaries.

APPENDIX A

The Invariant Distributions in Discrete and Continuous Time

As pointed out in Section 4.8, the centering and scale constants for the strong law and the central limit theorem are all written in terms of the mean first passage and recurrence times for the $\{B_n\}$ process. It appears from eqs. (11) and (12) that use of theorems I.7.1 and I.6.1 of Ref. 7 to find the invariant distribution of $\{B_n\}$ will require significant effort. In this appendix, we will derive the Chapman-Kolmogorov equations for the $\{B(t)\}$ process. Finding the invariant distribution of the $\{B(t)\}$ process by solving these equations is easier than solving for the invariant distribution of the discrete-time process using the transition probabilities in eq. (11). It is also a more attractive procedure numerically, because the matrix of coefficients is upper triangular with only a single nonzero subdiagonal, consisting of all α's. Finally, these results are tied together by Theorem 10, which indicates that these two invariant distributions are identical.

Let $r_n(t) = P\{B(t) = n\}$. Then for $h \geq 0$, we can write $r_n(t + h) = \sum_{k=0}^{c} P\{B(t + h) = n \,|\, B(t) = k,\, S_j \notin [t, t + h],\, \forall\, j\} P\{B(t) = k,\, S_j \notin [t, t + h],\, \forall\, j\} + \sum_{k=0}^{c} P\{B(t + h) = n \,|\, B(t) = k,\, S_j \in [t, t + h],\, \exists\, j\}$ $P\{B(t) = k,\, S_j \in [t, t + h],\, \exists\, j\}$.

To simplify the following display, in the first sum, all terms involving both an arrival and a departure in $[t, t + h]$ have an h^2 in them, and so can be left off. Similarly, in the second sum, because of the λh that will appear in front, all terms involving either an arrival or a departure can be left off. We obtain, omitting terms $o(h)$ or higher,

$$r_0(t+h) = (1-\lambda h)(1-\alpha h)[r_0(t) + \nu h r_1(t)] + \lambda h \sum_{k=0}^{c} q_{kk} r_k(t),$$

$$\begin{aligned} r_n(t+h) &= (1-\lambda h)(1-\alpha h)[(1-n\nu h)r_n(t) + (n+1)\nu h r_{n+1}(t)] \\ &\quad + (1-\lambda h)\alpha h r_{n-1}(t) + \lambda h \sum_{k=n}^{c} q_{k,k-n} r_k(t),\ 1 \le n \le c-1, \end{aligned}$$

$$r_c(t+h) = (1-\lambda h)[(1-c\nu h)r_c(t) + \alpha h r_{c-1}(t)] + \lambda h(1-c\nu h)q_{c0} r_c(t).$$

Collecting terms, simplifying, dividing by h, and letting $h \to 0$, we obtain

$$r_0'(t) = -\alpha r_0(t) + \nu r_1(t) - \lambda[r_0(t) - \sum_{k=0}^{c} q_{kk} r_k(t)],$$

$$\begin{aligned} r_n'(t) &= -(\alpha + n\nu) r_n(t) + \alpha r_{n-1}(t) + (n+1)\nu r_{n+1}(t) \\ &\quad - \lambda[r_n(t) - \sum_{k=n}^{c} q_{k,k-n} r_k(t)],\ 1 \le n \le c-1, \end{aligned}$$

$$r_c'(t) = -c\nu r_c(t) + \alpha r_{c-1}(t) - \lambda[r_c(t) - q_{c0}\, r_c(t)].$$

In equilibrium, we look for solutions with $r_j = \lim_{t\to\infty} P\{B(t) = j\}$ and $\lim_{t\to\infty} r_j'(t) = 0$. Then these equations become

$$\begin{aligned} 0 &= -(\alpha+\lambda) r_0 + \nu r_1 + \lambda \sum_{k=0}^{c} q_{kk} r_k \\ 0 &= \alpha r_{n-1} - (\alpha + n\nu + \lambda) r_n + (n+1)\nu r_{n+1} \\ &\qquad + \lambda \sum_{k=n}^{c} q_{k,k-n} r_k,\ 1 \le n \le c-1 \qquad (25) \\ 0 &= \alpha r_{c-1} - (c\nu + \lambda - \lambda q_{c0}) r_c \\ 1 &= \sum_{k=0}^{c} r_k, \end{aligned}$$

where the condition that $\{r_0, \cdots, r_c\}$ be a probability distribution has been added. These are the equations used to solve for the invariant distribution of the continuous-time process. Writing $\rho = \alpha/\nu$, $\lambda' = \lambda/\nu$, and $\mathbf{r} = (r_0, \cdots, r_c)^T$, the first $c+1$ equations can be written in matrix form as

$$[M(\rho) + \lambda' Q]\mathbf{r} = 0,$$

where $M(\rho)$ is the standard matrix for the M/M/c/c birth-death process (Ref. 9, Section 2.1), and

$$Q = \begin{pmatrix} 0 & q_{11} & q_{22} & \cdot & \cdot & q_{cc} \\ 0 & q_{10} - 1 & q_{21} & \cdot & \cdot & q_{c,c-1} \\ 0 & 0 & q_{20} - 1 & \cdot & \cdot & q_{c,c-2} \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & \cdot & q_{c1} \\ 0 & 0 & 0 & 0 & 0 & q_{c0} - 1 \end{pmatrix}.$$

The equations in this form show clearly that when $\lambda = 0$, we recover the ordinary M/M/c/c system, as expected. The $M(\rho)$ matrix is tridiagonal and Q is upper triangular, leading to the attractive form for numerical work mentioned above.

It remains to show that the two invariant distributions, for continuous time and for discrete time, are identical.

Theorem 10: $r_j = u_j, j = 0, \cdots, c$.

Proof. Define $B^*(t) = \sum_{n=1}^{\infty} B_n I(S_n \le t < S_{n+1})$, where I denotes the indicator function. Since $\{S_n\}$ is a Poisson process, $B^*(t)$ is a Markov process which will be thought of as a semi-Markov process embedded in the continuous-time busy server process. The distribution of the time between transitions in this process is exponential(λ), regardless of the starting state, and so the expected time to the next transition, starting from state i, is $1/\lambda$ for every i. From Section 6.3(ii) of Ref. 10 we obtain that $\lim_{t\to\infty} P\{B^*(t) = j \,|\, B^*(0) = i\} = u_j$ for $j = 0, \cdots, c$. Next, the distribution of the time from an arbitrary epoch back to the most recent failure is also exponential(λ), so that using Section 6.3(iv) of Ref. 10, we obtain

$$r_j = \sum_{i=0}^{c} u_i \int_0^{\infty} P\{B(S_n + t) = j \,|\, B_n = i\} \lambda e^{-\lambda t} dt,$$

regardless of the value of n because of the stationarity of $\{B_n\}$. For t with $S_n + t < S_{n+1}$, one gets $B(S_n + t) = j$ by having k survivors in the system at time S_n^+ and letting the M/M/c/c system evolve from there ($k = 0, 1, \cdots, i$). This has probability $\sum_{k=0}^{i} q_{i,i-k} P\{C(t, k) = j\}$, so

$$r_j = \sum_{i=0}^{c} u_i \sum_{k=0}^{i} q_{i,i-k} \int_0^{\infty} P\{C(t, k) = j\} \lambda e^{-\lambda t} dt = \sum_{i=0}^{c} u_i p_{ij} = u_j. \quad ■$$

APPENDIX B

Proofs

In this appendix, we provide proofs for Lemma 1, Theorems 2 and 4, Corollaries 5 and 6, and Theorems 8 and 9. The blot symbol ■ signifies the end of a proof.

Proof of Lemma 1: For $t, y > 0$ and $k \ge 2$, begin by writing

$$P\{N(t+y)-N(t)=k\}=\sum_{j=0}^{\infty}P\{N(t+y)=k+j,\,N(t)=j\}$$

$$=\sum_{j=0}^{\infty}P\{S_{k+j}\le t+y<S_{k+j+1},\,S_j\le t<S_{j+1}\},$$

where the interrenewal times are $X_1, X_2, \cdots, S_n = X_1 + \cdots + X_n$, and $S_0 = 0$. Now condition on $S_j = s$, $X_{j+1} = x$, and $X_{j+2} + \cdots + X_{j+k} = u$. Using the independence and identical distribution of the interrenewal times, together with some algebraic simplification, leads to eqs. (1) and (2). The sum and integral can be interchanged because of the uniform convergence of the renewal function on compact intervals. The proof is similar for the cases $k = 0$ and 1. ■

Proof of Theorem 2: Let $N_i(t)$ be the renewal counting process for failure mode i, and let T_n stand for the event that the nth arriving call is accepted into the system and survives to the end of its intended holding time without being cut off. Then there is a version of $P\{T_n | Y_n = y, \tau'_n = t\}$ that is given by

$$\sum_{k_1=0}^{\infty}\cdots\sum_{k_m=0}^{\infty}P\{T_n|N_i(t+y)-N_i(t)=k_i,\,i=1,\cdots,m\}$$

$$\cdot P\{N_i(t+y)-N_i(t)=k_i,\,i=1,\cdots,m\}$$

$$=\sum_{k_1=0}^{\infty}\cdots\sum_{k_m=0}^{\infty}\prod_{i=1}^{m}(1-p_i)^{k_i}P\{N_i(t+y)-N_i(t)=k_i\}$$

$$=\sum_{k_1=0}^{\infty}\cdots\sum_{k_m=0}^{\infty}\prod_{i=1}^{m}(1-p_i)^{k_i}\int_0^t g^i_{k_i}(t-s,y)dM^i_0(s)$$

$$=\prod_{i=1}^{m}\sum_{k=0}^{\infty}(1-p_i)^{k}\int_0^t g^i_{k}(t-s,y)dM^i_0(s),$$

where the superscript i on the g indicates the function from eq. (2) which belongs to failure mode i. Now insert the expression for g^i_k from Lemma 1, simplify, and use eq. (4). This leads to the desired conditional probability's being given by

$$\prod_{i=1}^{m}\int_0^t \bar{g}_i(t-s,y)dM^i_0(s).$$

Equation (5) is then obtained by unconditioning on the holding-time distribution and subtracting from one to obtain the probability of cutoff. To obtain eq. (6), first observe that since $\bar{g}_i$ is directly Riemann integrable, the basic renewal theorem (Ref. 4, Section II.XI.1) applies, yielding

$$\lim_{t\to\infty} \int_0^t \bar{g}_i(t-s, y)dM_0^i(s) = \lambda_i \int_0^\infty \bar{g}_i(u, y)du$$

$$= 1 - \lambda_i p_i \int_0^\infty \int_u^{u+y} \bar{M}_0^i(u+y-x)dF^i(x)du.$$

The Lebesgue bounded convergence theorem then allows the interchange of limit with the integrals in eq. (5), yielding eq. (6). ■

Proof of Theorem 4: The existence of the a.s. limit as $n \to \infty$ of χ_n/n follows from standard results about regenerative processes. These results, in a Markov chain setting (e.g., Theorem I.15.2 of Ref. 7), show that χ_n/n converges w.p. 1 to

$$\sum_{b=0}^{c} \sum_{w=0}^{b} wv_{(b,w)},$$

which, upon reversing order of summation, is seen to be equal to EW_1 since $\{W_n\}$ is a stationary process. Note that one also has $EW_1 = \lim_{n\to\infty} E\chi_n/n$. To complete the proof, straightforward calculations show that $E\chi_n = \theta EZ_n$, and that $\lim_{n\to\infty} EZ_n/n$ exists. ■

Proof of Corollary 5: Write $\chi_n/Z_n = (\chi_n/n)(n/Z_n)$ to obtain the result. ■

Proof of Corollary 6: Use theorem 8.1 of Ref. 12. ■

Proof of Theorem 8(*a*): Let $V_j(n)$ denote the number of visits to state j in the first n transitions of the embedded chain. Then

$$\sum_{k=1}^{n} B_k = \sum_{j=0}^{c} jV_j(n),$$

and it follows that

$$B_1 = \sum_{j=0}^{c} jV_j(1) \quad \text{and} \quad B_k = \sum_{j=0}^{c} j[V_j(k) - V_j(k-1)], \quad k \geq 2. \tag{26}$$

Using Lemma 7.3 of Ref. 5 and the stationarity of the embedded chain, our first step is

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left(\sum_{k=1}^{n} B_k\right) = \operatorname{Var} B_1 + 2 \sum_{k=2}^{\infty} [EB_1B_k - (EB_1)^2]. \tag{27}$$

The variance of B_1 is easily seen to be

$$\operatorname{Var} B_1 = \sum_{b=0}^{c} \frac{b^2}{m_{bb}} - \sum_{a=0}^{c} \sum_{b=0}^{c} \frac{ab}{m_{aa}m_{bb}}. \tag{28}$$

For the second term, use the representation in eq. (26), exchange order of summation, and sum by parts to obtain

$$\sum_{k=2}^{\infty} [EB_1B_k - (EB_1)^2]$$

$$= \sum_{i=0}^{c} \sum_{j=0}^{c} ij \lim_{R\to\infty} \left(E[V_i(1)V_j(R) - V_i(1)V_j(1)] - \frac{R-1}{m_{ii}m_{jj}} \right). \quad (29)$$

To simplify this, observe that $EV_i(1)V_j(1) = EV_i(1)^2$ when $j = i$ and is zero otherwise. Also, $P\{V_i(1) = x\} = 1 - u_i$ for $x = 0$, it equals u_i for $x = 1$, and is zero for $x \geq 2$, so that $EV_i(1)^2 = u_i = 1/m_{ii}$. Equation (29) becomes

$$\sum_{i=0}^{c} \sum_{j=0}^{c} ij \lim_{R\to\infty} \left[EV_i(1)V_j(R) - \frac{R-1}{m_{ii}m_{jj}} \right] - \sum_{i=0}^{c} \frac{i^2}{m_{ii}}. \quad (30)$$

Further simplifying, observe that

$$\begin{aligned} EV_i(1)V_j(R) &= E(V_j(R) \mid V_i(1) = 1)P\{V_i(1) = 1\} \\ &= E(V_j(R) \mid B_1 = i)u_i, \end{aligned}$$

so that the limit to be evaluated in eq. (30) becomes, after factoring out the common term $1/m_{ii}$,

$$\lim_{R\to\infty} \left[E(V_j(R) \mid B_1 = i) - \frac{R-1}{m_{jj}} \right]. \quad (31)$$

Now, letting I stand for the indicator function, $V_j(R) = \sum_{n=1}^{R} I(B_n = j)$, so that $E(V_j(R) \mid B_1 = i) = \sum_{n=0}^{R-1} p_{ij}^{(n)}$. When $j = i$ we obtain immediately, using Theorem I.6.5 of Ref. 7, that the limit in eq. (31) is given by

$$\frac{m_{ii}^{(2)} + m_{ii}}{2m_{ii}^2}.$$

When $j \neq i$, add and subtract $\sum_{n=0}^{R-1} p_{jj}^{(n)}$ in eq. (31), and use Theorem I.11.4 together with Theorem I.6.5 of Ref. 7 to obtain that the limit in eq. (31) is given, in this case, by

$$\frac{m_{jj}^{(2)} + m_{jj}}{2m_{jj}^2} - \frac{m_{ij}}{m_{jj}}.$$

We obtain, finally,

$$2 \sum_{k=2}^{\infty} [EB_1B_k - (EB_1)^2] = \sum_{i=0}^{c} \sum_{j=0}^{c} \frac{ij}{m_{ii}} \left(\frac{m_{jj}^{(2)} + m_{jj}}{m_{jj}^2} - \frac{2m_{ij}}{m_{jj}} \right). \quad (32)$$

Combining eqs. (27), (28), and (32) yields eq. (16), as was to be proved. ■

Proof of Theorem 8(*b*): From Theorem 7.5 of Ref. 5, the scale constant for the central limit theorem for the induced chain is the asymptotic variance of χ_n. We have

$$\operatorname{Var} \chi_n = E \sum_{k=1}^{n} W_k^2 + 2E \sum_{k=1}^{n} \sum_{j<k} W_jW_k - \left(E \sum_{k=1}^{n} W_k \right)^2.$$

The last term on the right is equal to

$$-\left(\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\right)^2\left(E\sum_{k=1}^{n}B_k\right)^2.$$

Using the conditional independence (eq. (10)), one shows that

$$EW_jW_k=\left(\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\right)^2 EB_jB_k \quad \text{for} \quad j\neq k,$$

and using eq. (9), one shows that

$$EW_k^2=\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i(1-p_i)EB_k+\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i^2EB_k^2.$$

Combining these and simplifying leads to

$$\frac{1}{n}\operatorname{Var}\chi_n=\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i(1-p_i)\sum_{b=0}^{c}\frac{b}{m_{bb}}$$
$$+\left[\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i^2-\left(\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\right)^2\right]\sum_{b=0}^{c}\frac{b^2}{m_{bb}}+\frac{1}{n}\left(\sum_{i=1}^{m}\frac{\lambda_i}{\lambda}p_i\right)^2\operatorname{Var}\left(\sum_{k=1}^{n}B_k\right).$$

The remainder of the proof consists in using Theorem 8(a) followed by algebraic manipulation. ■

Proof of Theorem 9: Begin by writing

$$\frac{\chi_n-\mu\lambda S_n}{\sigma\sqrt{n}}=\frac{\chi_n-\mu n}{\sigma\sqrt{n}}-\frac{\mu}{\sigma}\frac{S_n-n/\lambda}{\sqrt{n}/\lambda}. \tag{33}$$

By Theorem 7, the distribution of the first term on the right converges to the standard normal cdf. The distribution of the second term converges to the cdf of a normal random variable having mean zero and variance μ^2/σ^2. We will show that for each n, these two terms are independent.

The stochastic process $B^*(t)$ defined in the proof of Theorem 10 is a Markov pure jump process, and $X_1 = S_1$ and B_1 are independent because of the independence of the failure process and the arrival and service time processes (or use Theorem 15.28 of Ref. 11). Since the σ-field of the W's is contained in that of the B's, $S_1 = X_1$ and $\chi_1 = W_1$ are independent too. By Proposition 15.27 of Ref. 11, S_1 is a Markov time for the process, so that the process $B_1^*(t) = B^*(t + S_1)$ for $t \geq 0$ is a Markov process whose initial distribution is $P\{B_1 = b\}$, $b = 0, \cdots, c$. But because of the stationarity of $\{B_n\}$, $P\{B_1 = b\} = u_b$, $b = 0, \cdots, c$, so that $B_1^*(t)$ and $B^*(t)$ are equivalent processes. Hence, X_2 and B_2 are independent, and so are X_2 and W_2, from which it follows that S_2 and χ_2 are independent. The result for S_n and χ_n follows by induction.

It follows that the limit of the distribution of the quantity in eq. (33)

is the cdf of a normal random variable having mean zero and variance $1 + \mu^2/\sigma^2$. Now apply Theorem 8.1 of Ref. 12 to obtain the final result. The sufficient condition of that theorem is satisfied, because, using the notation of Ref. 12, $M^*(n) \le W_{n+1} + \mu\lambda X_{n+1}$ with probability one. ■

REFERENCES

1. L. J. Forys and E. Messerli, "Analysis of Trunk Groups Containing Short-Holding-Time Trunks," B.S.T.J., *54,* No. 6 (July–August 1975), pp. 1127–53.
2. J. Riordan, *Stochastic Service Systems*, New York: John Wiley, 1962.
3. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53,* No. 3 (March 1974), pp. 525–51.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vols. I and II, New York: John Wiley, 1950.
5. J. L. Doob, *Stochastic Processes*, New York: John Wiley, 1953.
6. N. A. Marlow, private communication.
7. K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, New York: Springer Verlag, 1967.
8. E. Parzen, *Stochastic Processes*, San Francisco: Holden-Day, 1962.
9. L. Kosten, *Stochastic Theory of Service Systems*, Oxford: Pergamon Press, 1973.
10. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, New York: John Wiley, 1974.
11. L. Breiman, *Probability*, Reading, Mass.: Addison-Wesley, 1968.
12. R. F. Serfozo, "Functional Limit Theorems for Stochastic Processes Based on Embedded Processes," Adv. Appl. Prob., *7* (1975), pp. 123–9.