THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 59, No. 5, May-June 1980
Printed in U.S.A.

Measuring the Utilization of a Synchronous Data Link: An Application of Busy-Period Analysis

By T. YUM

(Manuscript received August 29, 1979)

The utilization of links in a communication network is an important and easily measurable quantity; it is the fraction of time that a link is busy. We focus on the estimation of link utilization through measurements and consider, in particular, the duration of measurement required for a prescribed accuracy of the estimate. Two conflicting factors involved in the choice of measurement time are the desirability of a measurement of short, and thus recent, duration and the desirability of an interval of sufficient duration to assure statistical accuracy. We assume that the composite input of messages to an outgoing buffer of a synchronous data link constitutes a Poisson process. The message length is arbitrarily distributed and is in units of "packets." This allows us to model the transmission facilities as a continuous-time-input, discrete-time-output $M|G|1$ *queue. We analyze the output process of such a system and determine the time needed to measure link utilization with a prescribed accuracy. We also present an application to the Common Channel Interoffice Signaling network as an illustration of the analysis.*

I. INTRODUCTION

Measurements made on a computer-communication network can provide information regarding the throughput, delay, congestion, and deadlock in the network. These kinds of information are used not only to validate and improve network designs but also to provide for real-time control of the traffic flow.

The utilization of links in a communication network is an important and easily measurable quantity. It is merely the fraction of time that a link is busy. This paper focuses on the estimation of link utilization through measurements and considers, in particular, the duration of measurement required for a prescribed accuracy of the estimate. There are two conflicting factors involved in the choice of measurement time: on the one hand, we know that the most recent information should be used in any on-line control mechanism due to the nonstationarity of the "utilization" process. Thus a measurement of short duration is desirable. On the other hand, we want to measure an interval of sufficient duration to assure statistical accuracy. Thus we are interested in finding the shortest measurement time T such that the error is less than ϵ with confidence level $1 - \alpha$. We shall be considering both relative and absolute error criteria.

We start with the modeling and analysis of the data link in Sections II and III. The results are used to derive the measurement time T in Section IV and are applied to the CCIS network in Section V.

II. THE QUEUING MODEL FOR THE DATA LINK

It is customary to model a computer-communication network as a network of interconnected single server queues. This kind of modeling allows us to study the delay-blocking performance of the network as well as the utilization of each link. For the purpose of the present work, we focus on a synchronous line (defined at the end of this section) in such a network and model it as an $M|G|1$ queue. The Poisson arrival assumption is widely used in the literature.[1,2] To justify this assumption intuitively, let us study in a little more detail the arrival process to a particular link (or queue) in a network node. Basically, there are two kinds of arrivals, the external ones and the internal ones. The external arrivals of messages to a particular link are usually generated by a large number of independent "users."* It is therefore a superposition of a large number of independent random processes and can be considered to be approximately random. The internal arrivals to a link, on the other hand, are the messages being relayed through that link toward their destinations. They are the portions of the departure processes from neighboring nodes that have to be sent out again to other neighboring nodes through that link. This complicated internal arrival process is therefore another superposition of primarily unrelated processes from neighboring nodes. Combining with the external arrival process, we have a composite process which indeed is quite random. This points immediately to a Poisson process which is memoryless and has uncorrelated arrivals.

A common practice in modern data network is to segment messages into "packets" of some fixed size. This is done to improve both network

* For the CCIS network application in Section V, the external messages are initiated by telephone calls.

utilization and transmission reliability. Error checking and retransmission can be done on each packet, and packets can be interleaved and routed by the most cost-effective available route. The segmentation of messages into packets is modeled here; therefore, message lengths are in units of packets. For convenience, we define the time to transmit a packet on the synchronous line as a unit service interval.

The distribution of message length X differs widely for different networks and may not even follow any well-known distribution. But since these distributions can easily be measured, we shall assume an arbitrary (or general) distribution with $B(n)$, $b(n)$, and $B^*(z)$ being the distribution function, the probability mass function, and the generating function of X, respectively.

As a final remark on modeling, what we meant by synchronization is that a system clock is maintained and the output of packets must be at the beginning of one of the equally spaced time slots. This renders the output a discrete time process. The queuing model therefore is neither purely continuous nor purely discrete. Fortunately, this seemingly complex phenomenon adds little complexity to the analysis, as we see in the next section.

III. THE BUSY-IDLE PERIOD ANALYSIS

A lucid discussion on the busy period analysis of the $M|G|1$ queue can be found in Ref. 3. We follow a similar approach here, but we must generalize previous results to take into account the synchronization effect. The reader can easily find the differences in the results. A special result of this section, namely, the means and variances of the busy and idle periods, is to be used in finding the optimal measurement time T in the next section. We derive the idle period statistics here and refer the reader to the appendix for the lengthy derivation of the busy period results.

Focusing on the departure process of the queuing system, we observe that the system passes through alternating cycles of busy and idle periods. Consider first the idle period. Since it terminates immediately upon the arrival of a message and the time until the next message arrival has exponential distribution, the length of idle period is also exponentially distributed. This, however, is not true when the output stream of messages are synchronized. Let there be an arrival to an empty queue. Due to synchronization, the message is not served until the beginning of the next time slot. Thus the idle period is in effect exponentially distributed, but rounded off (or discretized) to the next time unit.

Consider the exponential density function in Fig. 1. After discretization, all the "mass" from $t = n - 1$ to $t = n$ is concentrated at $t = n$.

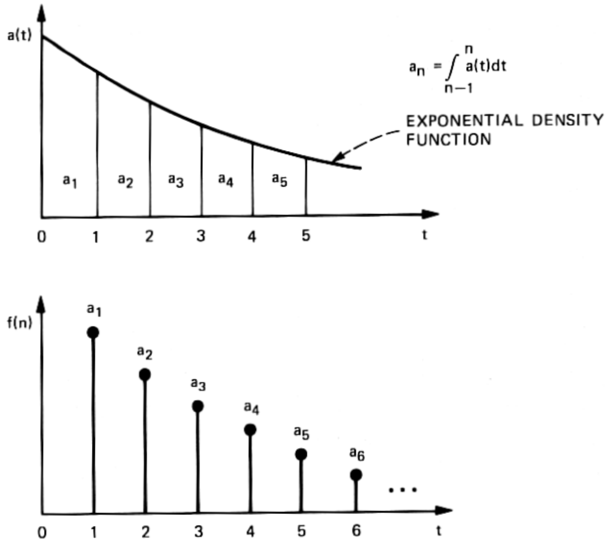


Fig. 1—Discretization of a density function.

Therefore, the probability mass function of the length of discretized idle period I is

$$f(n) = \int_{n-1}^{n} \lambda e^{-\lambda t}\, dt = e^{-\lambda n}[e^{\lambda} - 1] \qquad n = 1, 2, 3, \cdots, \tag{1}$$

where λ is the arrival rate of the Poisson process. The mean and variance of I are

$$\bar{I} = \frac{1}{1 - e^{-\lambda}} \tag{2}$$

$$\sigma_I^2 = \frac{e^{-\lambda}}{[1 - e^{-\lambda}]^2}. \tag{3}$$

The mean and variance of the idle period for the nonsynchronous system are $1/\lambda$ and $1/\lambda^2$, respectively. Comparing with (2) and (3), we see that synchronization increases the mean and variance of the idle period. This can be explained intuitively by the fact that synchronization delays the termination of the idle period until the next time slot.

We now turn to the derivation of the busy period statistics. Let there be an arrival to an empty queue at t_1. The busy period initiated by that arrival does not start until the beginning of the earliest time slot after t_1. We denote the length of this interval as D and show schematically the situation in Fig. 2. We continue the derivation in the appendix and eventually come up with the mean and variance of the busy period U as:

$$\bar{U} = \frac{\lambda}{1-e^{-\lambda}} \frac{\bar{X}}{1-\rho} \tag{4}$$

$$\sigma_U^2 = \frac{\lambda}{1-e^{-\lambda}} \frac{\sigma_X^2 + \rho\bar{X}^2}{(1-\rho)^3} + \frac{\lambda}{(1-e^{-\lambda})^2}(1-e^{-\lambda}-\lambda e^{-\lambda}) \frac{\bar{X}^2}{(1-\rho)^2}, \tag{5}$$

where $\bar{X}$ is the average message length and $\rho \triangleq \lambda\bar{X}$ is the utilization of queuing system. Note that the busy period for the nonsynchronous system is merely Y (defined in the appendix). Its mean and variance are given by (13) and (14), respectively, in the appendix. Comparing (4) and (5) with (13) and (14), we see that synchronization also increases the mean and variance of the busy period.* This presumably is due to the additional arrivals in the time interval D.

This concludes the busy-idle period analysis of a synchronous data link. In the next section, we use the above results to derive the optimal measurement time T.

IV. DERIVATION OF MEASUREMENT TIME

In the operation of a data network, people usually want to have an estimate of link utilization from time to time as network conditions change. Of interest are the frequency and duration of measurements required to assure a particular quality for the utilization estimate. We derive in the following the measurement duration T needed for any given desired accuracy. The results depend on the accuracy (or error) criteria used. We consider two of them: the absolute error δ and the relative error ϵ.

Consider a time interval of T slots at the output end of the $M|G|1$ queue of the last section. Let $N(T)$ be the number of busy periods in T and U_i, $i = 1, 2, \cdots, N(T)$ be the length of the busy periods in T. If there are R message-carrying slots (therefore, T-R idle slots), an estimate of the utilization ρ over the interval $[0, T]$ is

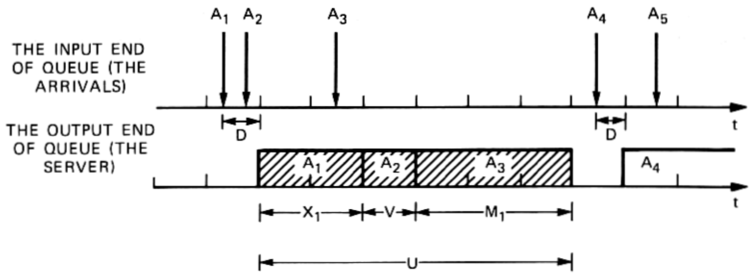
$$\hat{\rho}(T) = \frac{R}{T}. \tag{6}$$

We note that R is composed of the sum of busy periods:

$$R = U_1 + U_2 + \cdots + U_{N(T)}.$$

Now the central limit theorem for $N(T)$ says that, for large T, $N(T)$

* Note that $\lambda/(1-e^{-\lambda}) > 1$ for $\lambda > 0$ in (4) and (5).



A_I = THE I–TH ARRIVAL TO THE QUEUE

X_1 = SERVICE TIME OF THE FIRST MESSAGE

V = SUB–BUSY–PERIOD GENERATED BY THE ARRIVALS IN D

M_1 = SUB–BUSY–PERIOD GENERATED BY THE ARRIVALS IN THE SERVICE TIME OF THE FIRST MESSAGE

U = THE BUSY PERIOD

D = THE INTERVAL FROM THE FIRST ARRIVAL TO AN EMPTY SYSTEM UNTIL THE BEGINNING OF THE EARLIEST TIME SLOT

Fig. 2—The sub-busy-periods in a busy period.

is approximately normal with mean and variance[4]

$$\bar{N}(T) = \frac{T}{\bar{U} + \bar{I}}$$

$$\sigma^2_{N(T)} = \frac{T(\sigma^2_U + \sigma^2_I)}{(\bar{U} + \bar{I})^3}.$$

R is a sum of $N(T)$ i.i.d. random variables U. Hence, for large T, R is also approximately normal. Therefore, from (6), $\hat{\rho}(T)$ is also approximately normal. Taking expectation of $\hat{\rho}(T)$, we have

$$E[\hat{\rho}(T)] = \frac{1}{T} E[R]$$

$$= \frac{1}{T} E[U] \cdot E[N(T)]$$

$$= \frac{1}{T} \cdot \bar{U} \cdot \frac{T}{\bar{U} + \bar{I}}$$

$$= \frac{\bar{U}}{\bar{U} + \bar{I}} = \rho.$$

Thus the $\hat{\rho}$ defined in (6) is an unbiased estimator of ρ. From (6), the variance of $\hat{\rho}(T)$ is given by

$$\sigma^2_{\hat{\rho}(T)} = \frac{1}{T^2}\,\sigma_R^2$$

$$= \frac{1}{T^2}\,[\bar{N}(T)\sigma_U^2 + \bar{U}^2\sigma^2_{N(T)}]$$

$$= \frac{1}{T}\left[\frac{2\bar{U}\sigma_U^2(\bar{U}+\bar{I}) + \bar{I}^2\sigma_U^2 + \bar{U}^2\sigma_I^2}{(\bar{U}+\bar{I})^3}\right]$$

$$= \frac{1}{T}\left[\frac{\rho(1+\rho^2)(\sigma_X^2+\bar{X}^2)}{\bar{X}(1-\rho)^2} - \frac{2\rho^3}{(1-\rho)(e^{\rho/\bar{X}}-1)}\right], \qquad (7)$$

upon substituting the means and variances of the busy and idle periods from the previous section.

As we have mentioned before, our objective is to find the minimum value of T such that the following inequality is satisfied:

$$P[|\hat{\rho}(T) - \rho| \le \epsilon\rho] \ge 1 - \alpha. \qquad (8)$$

Let

$$W = \frac{\hat{\rho}(T) - \rho}{\sigma_{\hat{\rho}(T)}}.$$

Then, equivalently, we can express (8) as

$$P\left[|W| \le \frac{\epsilon\rho}{\sigma_{\hat{\rho}(T)}}\right] \ge 1 - \alpha. \qquad (9)$$

Now, from the table of $N(0, 1)$*, we can find a quantile w that gives a cumulative probability of $1 - (\alpha/2)$. That is, w is given as

$$\int_{-\infty}^{w} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,dy = 1 - \frac{\alpha}{2}.$$

We interpret w as the number of standardized $N(0, 1)$ deviations such that $P[|W| \le w] \ge 1 - \alpha$. Comparing this with (9), we have

$$\frac{\epsilon\rho}{\sigma_{\hat{\rho}(T)}} = w$$

or

$$\sigma^2_{\hat{\rho}(T)} = \frac{\epsilon^2\rho^2}{w^2}.$$

* This is the normal density function with zero mean and unit variance.

From (7), we have finally

$$T = \left(\frac{w}{\epsilon\rho}\right)^2 \left[\frac{\rho(1+\rho^2)(\sigma_X^2 + \bar{X}^2)}{\bar{X}(1-\rho)^2} - \frac{2\rho^3}{(1-\rho)(e^{\rho/\bar{X}} - 1)}\right], \tag{10}$$

where w is a constant for a specific confidence level $1 - \alpha$. One can specify ϵ as an acceptable upper limit for the relative error or $\epsilon\rho$ as the acceptable absolute error in $\hat{\rho}$.

In the following discussion, we investigate the dependence of T on ρ for the absolute and the relative error criterion.

(*i*) Consider a relative error objective ϵ. From (10), we see that there are two poles, at $\rho = 0$ and $\rho = 1$. When the traffic is light (i.e., ρ small), the error bound $\epsilon\rho$ specified by the relative error criterion is small too. This means that the measure duration T must be long to accumulate enough "samples" (message-carrying slots) for an accurate estimate, and therefore account for the pole at the origin. When the traffic is heavy (i.e., ρ close to 1), the term $(1 - \rho)^2$ in the denominator dominates. This term comes from the variance of the busy period and accounts for the pole at $\rho = 1$.

(*ii*) If we specify an absolute error objective, the denominator $(\epsilon\rho)^2$ should be replaced by δ^2 where δ is the specified objective. The dependence of T on ρ will be via the terms in the square bracket of (10). A zero is at $\rho = 0$ and a pole at $\rho = 1$. For this case, T increases in ρ.

Figures 3 and 4 show the curves for both error criteria for a specific example. These figures also show how seemingly similar specifications can result in entirely different conclusions. Both error criteria are used extensively; the choice depends entirely on the designer. With a thorough understanding of both, however, it is trivial to transform one to the other.

V. EXAMPLE

We use the Common Channel Interoffice Signaling (CCIS) network to illustrate the analysis. Presently, a CCIS link is engineered to accommodate the signaling load for around 1500 trunks under normal condition. With a busy-hour average of five attempts in each direction per trunk, a link utilization of $\rho = 0.28$ is obtained. The message length, in "signal units," has the following distribution:

$$b(n) = \begin{cases} 0.603 & n = 1 \\ 0.261 & n = 3 \\ 0.130 & n = 4 \\ 0.006 & n = 5. \end{cases}$$

Each signal unit is 28 bits long. The mean and variance of the message

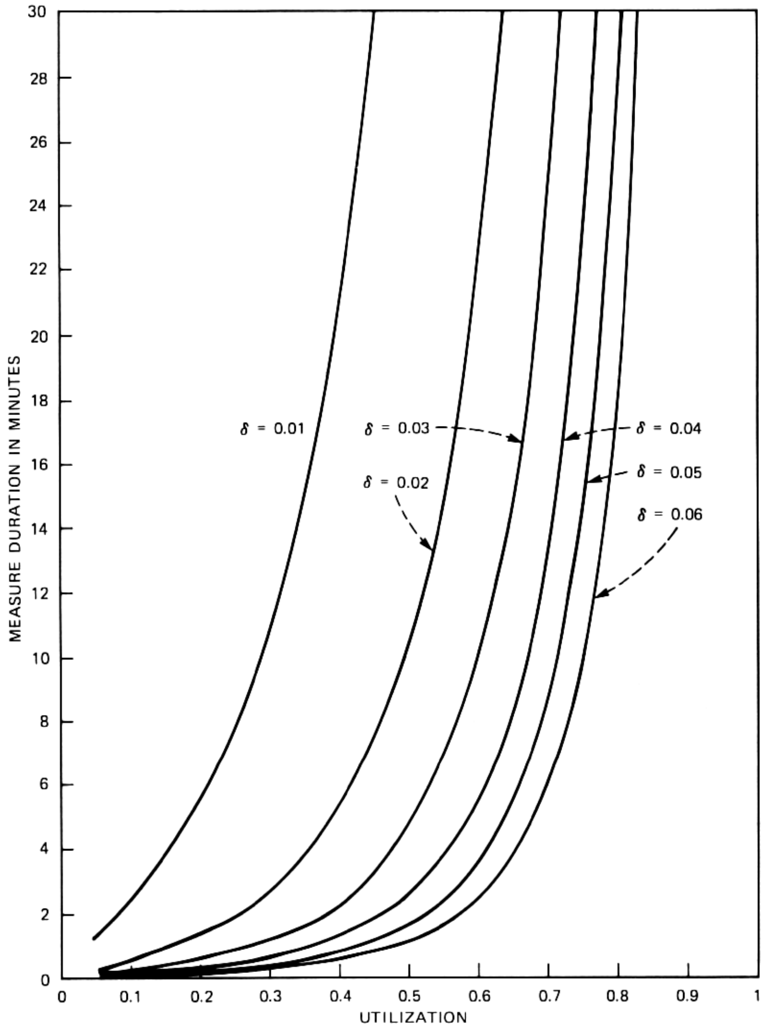


Fig. 3—Measure duration: Absolute error criterion.

length in signal units are 1.936 and 1.434. The CCIS link currently operates at a rate of 2400 bits/s or, equivalently, 85.7 signal units/s. One out of every 12 signal units is used for acknowledgment. The equivalent "message" capacity therefore is $85.7 \times 11/12 = 78.57$ signal units/s. With this capacity and with the nonrandom acknowledgment traffic deleted, we have an effective utilization of

$$\rho_e = \frac{0.28 - 0.0825}{1 - 0.0825} = 0.2153.$$

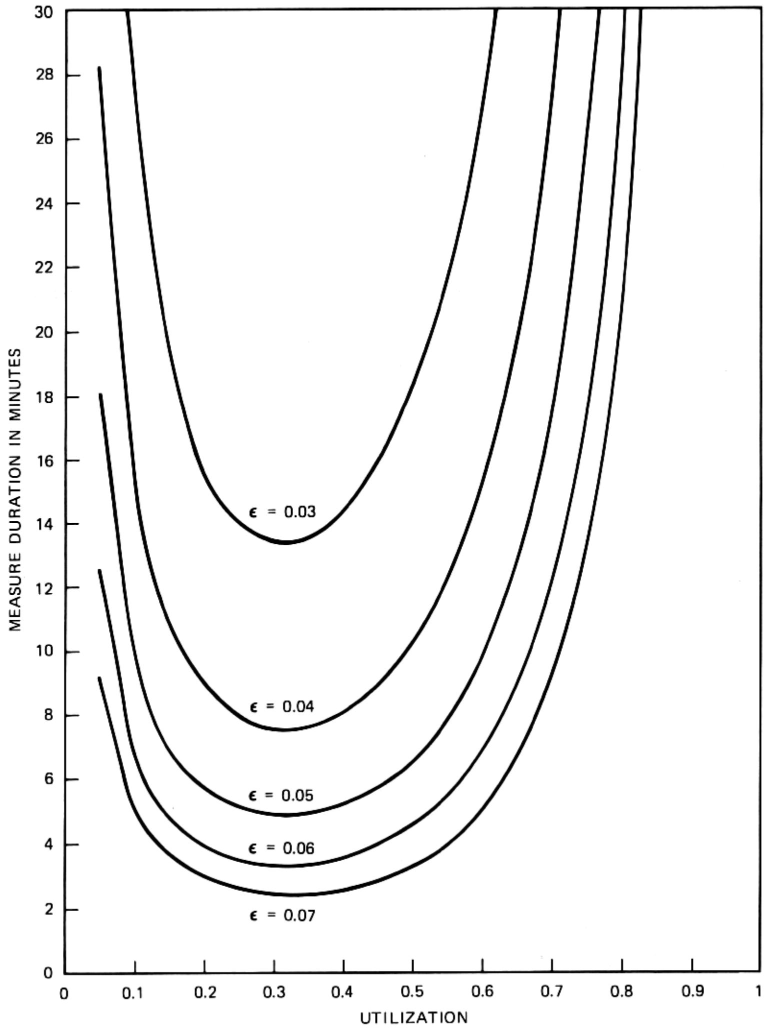


Fig. 4—Measure duration: Relative error criterion.

(*i*) The CCIS links operate in pairs so that, when one link fails, the other can take up the combined load. Consider first the operation of the network under normal conditions, i.e., $\rho_e = 0.2153$. Let $\alpha = 0.05$, $\epsilon = 0.05$; then $w = 1.96$. From (10), we have $T = 25297$ units of time = 5 min, 22 s. Note that $\bar{N}(T) = T/(\bar{U} + \bar{I}) = 768$. This justifies the normality of R, and hence $\hat{\rho}(T)$ in (6). Now let the maximum absolute error be $\delta = 0.03$. The corresponding optimal measure time is $T = 41$ s. ($\bar{N}(T) = 98$).

(*ii*) Next, let us consider the case when one of the link pair fails. Its mate link then must handle a total of $2\rho_e = 0.4306$. Proceed with the calculation as before. For $\epsilon = 0.05$, we have $T = 5$ min, 30 s. For $\delta = 0.03$, we have $T = 2$ min, 48 s. Figures 3 and 4 show the measurement time T versus the utilization ρ_e for the absolute error and the relative error, respectively.

If we need more confidence in the result, say, 99-percent confidence ($\alpha = 0.01$), we can use the following elementary "ratio and proportion" technique. From the table of normal distribution, the w that corresponds to $\alpha = 0.01$ is found to be $w_{0.01} = 2.576$. From (10), we know that T is proportional to w^2. Therefore, the conversion factor for changing from $\alpha = 0.05$ to $\alpha = 0.01$ is

$$\left(\frac{w_{0.01}}{w_{0.05}}\right)^2 = 1.727.$$

Thus in (*ii*) above with $\epsilon = 0.05$, $\rho = 0.4306$ and $\alpha = 0.01$, we require $T = 5.5 \text{ min} \times 1.727 = 9.5$ min.

VI. SUMMARY AND ACKNOWLEDGMENT

We have focused our attention on the output of a synchronous data link. We analyzed the busy-idle period and specialized our results to traffic intensity measurements. As a specific example, we studied the CCIS link and calculated the required measure duration for different utilization levels ρ, error criteria, confidence levels $(1 - \alpha)$, and accuracy.

I would like to thank Tony T. Lee, Jerry Gechter, and Charles D. Pack for their helpful suggestions in the course of this work.

APPENDIX

The Busy Period Statistics

The length of the busy period U is composed of two sub-busy-periods (SBP's) V and Y. The V-SBP is generated by the arrivals in D. The Y-SBP constitutes the service time of the first message and those SBP's generated by the arrivals during the service time of that first message. In other words,

$$Y = X_1 + M_1 + M_2 + \cdots M_K, \tag{11}$$

with X_1 the length of the first message, M_i the length of the SBP generated by the ith arrival during the service time of the first message and K the total number of arrivals during that first service time. Observe that the M and the Y SBP's are both initiated by single messages with lengths drawn independently from the same distribution. Moreover, each SBP continues until the system catches up to the work load (i.e., the work load of the system drops back to the level just before the start of the SBP). Thus the M_i's are independent, identically distributed, and have the same distribution as Y.

Let $y(n) = \text{Prob}[Y = n]$ and $Y^*(z)$ be its generating function. We condition Y on two events: the length of the first message and the number of arrivals during the service of the first message. From (11), we can write

$$Y^*(z)\,|_{X_1=n, K=k} = E[z^{n+M_1+\cdots M_k}].$$

Since all the M-SBP's are independent and identically distributed as the Y-SBP and n is nonrandom, we may write

$$Y^*(z)\,|_{n,k} = z^n[Y^*(z)]^k.$$

Moreover, conditioned on $X_1 = n$, K has a Poisson distribution with mean λn. We may therefore remove the condition on K:

$$Y^*(z)\,|_n = \sum_{k=0}^{\infty}\left[\frac{(\lambda n)^k e^{-\lambda n}}{k!}\right] z^n[Y^*(z)]^k = z^n e^{-\lambda n[1-Y^*(z)]}.$$

Similarly, we remove the condition on X_1:

$$Y^*(z) = \sum_{n=1}^{\infty} b(n)[ze^{-\lambda[1-Y^*(z)]}] = B^*[ze^{-\lambda[1-Y^*(z)]}]. \tag{12}$$

Thus we arrive at a functional equation in $Y^*(z)$, which is usually impossible to solve and invert. However, we can solve for the ith moment h_i by differentiating (12) i times. Carrying this out, we have the mean and variance of Y given as

$$\bar{Y} = h_1 = \frac{\bar{X}}{1-\rho} \tag{13}$$

$$\sigma_Y^2 = h_2 - h_1^2 = \frac{\sigma_X^2 + \rho\bar{X}^2}{(1-\rho)^3}, \tag{14}$$

where $\rho \triangleq \lambda\bar{X}$ is the utilization of the system.

We now turn to the V-SBP, which is initiated by L random arrivals in D. Each of these arrivals initiates an SBP which behaves statistically the same as the M-SBP's that we have just described. Hence, we can write

$$V = M_1 + M_2 + \cdots M_L. \tag{15}$$

Proceeding as before, we condition V on $L=l$ and $D=s$. Taking the z-transform on (15), we have

$$V^*(z)\,|_{L=l, D=s} = E[z^{M_1+M_2+\cdots M_l}] = [Y^*(z)]^l.$$

Removing the conditioning on L, we have

$$V^*(z)\,|_{D=s} = \sum_{l=0}^{\infty} \frac{(\lambda s)^l}{l!} e^{-\lambda s}[Y^*(z)]^l = e^{-\lambda s[1-Y^*(z)]}. \qquad (16)$$

We now digress a moment to find the density of D. Figure 5 shows an arrival A_1 in the interval (0, 1) (note again that the unit time is the time to transmit one packet). Let $C \triangleq 1 - D$. Then C is the arrival time of the first message in (0, 1). We know there is at least one arrival in (0, 1); A_1 is the first one, and there may be more in the same slot. Therefore, the random variable C is actually the arrival time of the first message, given that there is at least one arrival in (0, 1). We may therefore write

$$\text{Prob}[C \le x] = \frac{\text{Prob}[\text{arrival time} \le x]}{\text{Prob}[\text{at least one arrival in } (0, 1)]}$$

$$= \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda}} \qquad 0 \le x \le 1. \qquad (17)$$

Now since $D = 1 - C$, the density function of D is

$$f_D(y) = f_C(1 - y) = \frac{\lambda e^{-\lambda}(1 - y)}{1 - e^{-\lambda}} \qquad 0 \le y \le 1. \qquad (18)$$

Figure 5 also shows the two density functions.

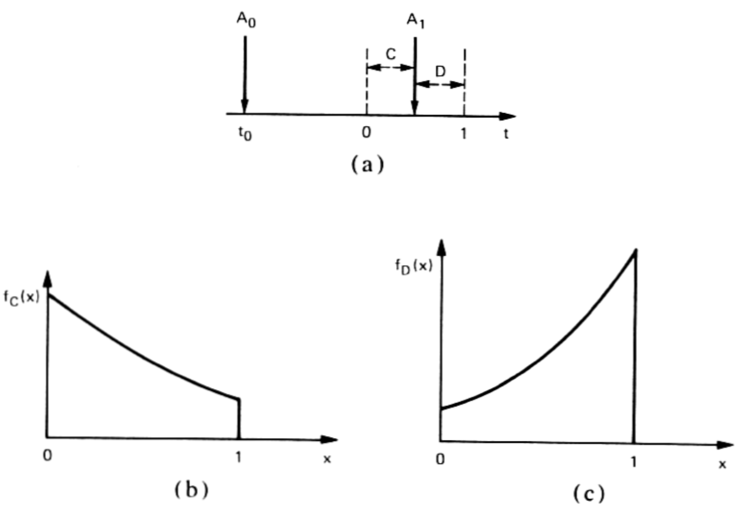


Fig. 5—The densities of C and D. (a) An arrival in the time slot (0,1). (b) Density of C. (c) Density of D.

We are now ready to remove the conditioning on D:

$$V^*(z) = \int_0^1 e^{-\lambda s[1-Y^*(z)]}\left[\frac{\lambda e^{-\lambda(1-s)}}{1-e^{-\lambda}}\right] ds = \frac{e^{\lambda Y^*(z)}-1}{Y^*(z)(e^{\lambda}-1)}. \tag{19}$$

Now V and Y are each generated by Poisson arrivals in disjoint time intervals, hence they are independent. Equation (19) says the distribution of V can be expressed as a function of the distribution of Y, not that V and Y are statistically dependent. Hence from $U = V + Y$, we have

$$U^*(z) = V^*(z)\,Y^*(z) = \frac{e^{\lambda Y^*(z)}-1}{e^{\lambda}-1}. \tag{20}$$

Extracting the moments, we obtain the mean and variance of U as given in (4) and (5).

REFERENCES

1. T. Stern, "A Class of Decentralized Routing Algorithms Using Relaxation," IEEE Trans. Commun., *COM-25* (Oct. 1977), pp. 1092–1102.
2. C. Brown and M. Schwartz, "Adaptive Routing in Centralized Computer Communication Networks with Application to Processor Allocation," Proc. ICC, June 1975.
3. L. Kleinrock, *Queuing Systems*, Vol. I, New York: Wiley-Interscience, 1975, pp. 206–216.
4. W. Feller, *An Introduction to Probability Theory and Its Application*, 2nd ed., Vol. II, New York: John Wiley, 1971, p. 372.