# On the Analysis and Modeling of a Class of Computer Communication Systems

WESLEY W. CHU, MEMBER, IEEE, AND ALAN G. KONHEIM

*Abstract*—**Recent advances in computer communications are discussed including computer-traffic and channel error characteristics, optimal fixed message block size, statistical multiplexing, and loop systems. A unified model is developed and then used to analyze the queueing behavior of the star and loop systems. Numerical results for selected traffic intensities and message lengths, given in graphical form, provide insight into the performance of these systems.**

## Introduction

As COMPUTER communication systems such as time sharing and distributed computer systems grow in scale and complexity, the problem of being able to understand and to predict system behavior becomes increasingly important. It becomes clear that in order to synthesize a system that meets operating and performance criteria at minimum cost for computing, communication, and operation, two key problem areas need to be studied: first, the interfacing problem between computer and communication systems; second, the relationships among communication traffic sources, channels, and computer resource allocation mechanisms.

The difficulties in studying and understanding these problems are 1) computer designers' lack of knowledge in communication technology; 2) communication engineers' lack of knowledge in computer technology; and 3) the lack of tools and models with which to analyze the behavior of these complex systems. The first two difficulties may be resolved by exchange of information between computer designers and communication specialists. The third difficulty may be remedied by periodically summarizing important research related to computer communication systems that is scattered throughout various journals, conference proceedings, and technical reports. In this paper we aim at the last objective.

Our presentation will be divided into two distinct but related parts. In the first part, we present recent advances including computer traffic characteristics, channel error characteristics, selection of optimal fixed message block size, statistical multiplexing, and loop systems. In the second part, we develop a model for terminal-to-computer communication that results in a unified treatment of several types of computer communication systems. The results of this unified approach are summarized in Section II-A; the detailed analysis begins in Section II-B.

Manuscript received January 10, 1972; revised February 28, 1972. This work was supported in part by the U.S. Office of Naval Research under Contract N00014-69-A-0200-4027, NR 048-129, in part by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC 15-60-C-0285, and in part by the U.S. Air Force under Contract F44620-70-C-0063.

W. W. Chu is with the Department of Computer Science, University of California, Los Angeles, Calif. 90024.

A. G. Konheim is with the Mathematical Sciences Department IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y. 10598.

## I. Some Recent Advances in Computer Communications

### A. Computer Traffic Characteristics

It has become apparent that real progress in modeling and analysis depends upon more than elegant analytical results based upon convenient but unsupported assumptions. Measurement and observation are needed; a study of the computer traffic characteristics of in-house time-sharing systems has been undertaken by the Bell Telephone Laboratories to obtain estimates of system variables. Two types of systems under study are long holding time (connect to disconnect) and short holding time. Long holding time is characteristic of business and scientific applications that require extensive computation; a holding time typically of 15–30 min. Short holding time is characteristic of inquiry–response systems such as on-line banking, credit bureau, and production control; typically holding times of a few seconds to one or two minutes.

Jackson and Stubbs [1] and Fuchs and Jackson [2] have reported the results of long holding time. They show that the volume of computer-to-user traffic is an order of magnitude higher than that of user-to-computer traffic. The interarrival time between messages can be approximated by an exponential distribution; that is, the stream of messages can be assumed to constitute a Poisson process. Furthermore, the length of messages can be satisfactorily approximated by the geometrical distribution. During the call interval, the user is active only 5 percent of the time and the computer is active about 30 percent of the time. Thus, the channel is idle for a significant portion of the holding time. The traffic characteristics of short holding time are reported on by Dudick *et al.* [3]. The measured results from four such systems reveal that user send time (the total amount of time during which user characters are being transmitted) is less than 15 percent of the holding time. This parameter is important to the design of statistical multiplexors. The character interarrival times can be represented as a sum of two gamma distributions, the number of user segments

per call can be represented by a geometrical distribution, and the number of computer segments per call can be represented by a geometrical distribution. These measurements and estimated system variables not only provide us with insight into the behavior of the system and shed light on areas that need improvements, but are essential in the modeling and analysis of computer communication systems.

### *B. Channel Error Characteristics*

The communication channel provides the links between processors and terminals and plays an important role in computer communication systems. Thus, a characterization of channel performance is important for understanding the cause of errors, for suggesting the possible improvements in the design of transmission equipment and the design of efficient error-control systems, and for planning optimal computer communication systems.

To characterize telephone channel error performance, a survey measurement program of the telephone network to determine the error performance and the data speed capabilities is necessary. A series of such studies started by the Bell Systems in 1958 was directed towards this goal and several papers have been published on this subject [4]–[8]. Here we shall emphasize recent survey results (1969–1970) on high-speed voice-band data transmission performance on the switched telecommunication network [7] and low-speed data transmission performance on the switched telecommunication network [8]. In these two papers, the distributions of error per call are given on a bit, burst, and block basis. Information is also presented on the distribution of intervals between errors, the structure of burst errors, and the number of errors in blocks of various sizes. Such statistics provide information on channel reliability and are also useful in the design of efficient error-control procedures.

In the high-speed voice-band data transmission channel, toll traffic was used as a basis for the sampling plan, which resulted in the selection of approximately 600 dialed-toll connections between geographically dispersed local switching offices. Data rates of 1200, 2000, 3600, and 4800 bits/s are measured on the Bell System switch telecommunication network. The measured results show a substantial improvement of performance in comparison with the results of previous surveys. For example, the measured results for operation at 1200 and 2000 bits/s show that approximately 82 percent of the calls have an average error rate of 1 error in $10^5$ bits or better average over short, medium, and long haul calls, while the 1959 Alexander, Gryb, and Nast survey [4] shows only 63 percent of their test calls reached this ($10^{-5}$) performance level for operation at the same data rate. A general tendency for performance to degrade with transmission distance has been noted. These results also indicate that impulse noise accounted for a large percentage of the observed errors.

Low-speed data transmission corresponds to teletypewriters, computer ports, and other terminal devices that communicate by means of data organized in characters (comprised of several bits) using start–stop transmission at a rate of 300 bits/s. Character-error statistics, rather than bit-error statistics, are the parameters of interest in this type of transmission because the message consists of a display (in teletypewriters) or use of characters in most applications (in computers). Measurements were made on 534 connections with over 21 million characters (1 character = 10 bits) transmitted. Over 90 percent of the low-speed test calls contained about 36 000–54 000 characters. A character error rate of $10^{-4}$ or less is indicated for approximately 78 percent of all calls, while 95 percent of all calls have a lost character rate of $10^{-4}$ or less. Errors occurring in the messages are in bursts rather than at random. The number of character errors in a block increases with the block length. Since this is the first report on low-speed data, no comparison with any previous survey is possible. Further analysis of the statistics will give insight into the causes of errors, which in turn may suggest approaches to improve error performance.

### *C. Optimal Fixed Message Block Size*

The message outputs from a computer are usually in strings of characters or bursts. The variation of message length can best be described by a probability distribution. For ease in data handling and memory management, the random message length is usually partitioned into several fixed-size blocks. Due to the random length of the message, the last partitioned block usually is not filled by the message but is filled with dummy information.

For reasons of economics and reliability, error detection and retransmission are employed in many data communication systems [5], [9]. The optimal block size is an important parameter in the design of such systems. From the acknowledgment point of view, it is desirable to select the largest possible block size. Since each message block requires at least one acknowledgment signal, the fewer the number of blocks needed for a message, the less the channel capacity required for acknowledgments. On the other hand, since a larger message block has a higher channel wastage due to the last unfilled partitioned block and has also a higher probability of error, it is desirable to select the smallest possible block size. Thus there is a tradeoff in selecting the optimal block size.

Kucera [10], Balkovic and Muench [11], and Kirlin [12] have studied the optimal message block size for the error detection and retransmission system that maximizes transmission efficiency. Chu [13] considers an additional important parameter, the message (file) length, in determining the optimal message block size, which significantly affects the selection of the optimal message block size. His model considers average message (file) length, message length distribution, channel error characteristics (random error and burst error), overhead for addressing, error control, and acknowledgment delay. His criterion

for optimality is to minimize the time wasted in acknowledgments, retransmissions, and the waste in the last unfilled block.

### D. Statistical Multiplexing

Multiplexing is commonly used to share and efficiently utilize a communication channel. Currently data multiplexing has taken two forms: frequency division multiplexing (FDM) and synchronous time division multiplexing (STDM) commonly known as time division multiplexing. Frequency division multiplexing divides the channel bandwidth into several subchannels such that the bandwidth of each subchannel is greater than that required for a message channel. Because of the need to employ guard bands to prevent data signals from each of the data channels from feeding into adjacent channels and because of the relatively poor data transmission characteristics of the voiceband channel near the edges of its bandwidth, FDM does not make as efficient use of the voiceband as does STDM.

In STDM, each user (terminal) is assigned a fixed time duration or time slot on the communication channel for the transmission of messages from terminals to computer. The multiplexing apparatus scans the set of users in a round-robin fashion. After one user's time duration has elapsed, the channel is switched to another user. With appropriately designed synchronous operation, required buffering can be limited to one character per terminal. Addressing is usually not required since the user is identified by the time slot position. The STDM technique, however, also has certain disadvantages. It is inefficient in channel utilization to permanently assign a segment of bandwidth that is utilized only a portion of the time. Statistics collected from several typical operating time-sharing systems [1] showed that during a call (connect to disconnect), the user-to-computer traffic, in the long holding time case, is active only 5 percent of the time. Thus STDM would be very inefficient in channel utilization in such an environment since it allocates a time slot to each user independent of his activity. In order to increase channel utilization, statistical multiplexing or asynchronous time division multiplexing (ATDM) has been proposed [14], [15] for computer communications. The basic idea is to switch from one user to another user whenever the former is idle and the latter ready to transmit data. Thus the data is asynchronously or statistically multiplexed with respect to the users. With such an arrangement, each user would be granted access to the channel only when he has a message to transmit. The crucial attributes of such a multiplexing technique are 1) an address is required for each transmitted message and 2) buffering is required to handle statistical fluctuations in the input traffic.

The data structure for messages forming the input to the multiplexor buffer can be classified into four categories: constant-length messages; random-length message; mixed (constant and random length) message; random-length messages segmented into fixed-size blocks. The constant-length message input corresponds to teletype (TTY) input, each user types in one character at a time. The random-length message input corresponds to paper-tape input, cathode-ray tube (CRT) input or computer output. The mixed-message input corresponds to traffic from a mixture of types of input terminals such as CRT, TTY, etc. For ease in data handling and memory management, random-length messages are often segmented into fixed-size blocks that correspond to the last type of data structure. Since messages have random length, the last block of a message usually cannot be entirely filled. As a result, this type of data structure requires a larger buffer than that of random-length messages that are not segmented [14]. The buffer behavior in terms of buffer overflow probability and expected queueing delay due to buffering of these four types of data structures have been analyzed by finite waiting-room queueing models [14]–[17].

The buffer behavior of a statistical multiplexor for mixed-message inputs lies in between that of constant-length and random-length messages [17]. The output process of a statistical multiplexor has been studied by Pack [18].

The demultiplexor distributes messages to appropriate destinations according to their message addresses. Thus, the behavior of the demultiplexing buffer not only depends on traffic intensity but also on traffic scheduling to various destinations. In the case of a time-sharing system, message scheduling is determined by the job-scheduling algorithm of the computer's operating system. In the case of distributed computer systems, message scheduling is influenced by the message-routing algorithm. A simulation study of the demultiplexing buffer behavior revealed that, for a given input traffic volume, the best buffer behavior can be achieved by scheduling an equal amount of traffic to each destination [19]. Hence there is a close relationship among demultiplexing system performance and scheduling algorithms [20] in computer operating systems and/or message routing algorithms. Further research in this direction would be desirable. Results obtained in this area will be essential in the joint optimization of the overall performance of such computer communication systems.

Buffering is required in order to provide error control and message scheduling, which are two important functions in computer communication systems. Since statistical multiplexing requires buffering in order to handle statistical fluctuations, the multiplexing buffer can also be used for these functions.

From these studies we conclude that in an ATDM system, an acceptable buffer overflow probability can be achieved by a reasonable buffer size; the expected queueing delay is very small and acceptable for most applications. Hence, ATDM or statistical multiplexing is a feasible technique for data communications. Furthermore, ATDM greatly improves the transmission efficiency and system organization.

We have constructed a statistical multiplexor at the

University of California, Los Angeles (UCLA). Our preliminary experience has shown that the gain in communication cost, especially in long distance transmission, by employing ATDM in computer communication could far outweigh costs of overhead in addressing and storage for buffering. Statistical multiplexing should, therefore, have high potential for use in future computer communication systems.

### E. Loop Systems

A special distributed-computer-system architecture of considerable recent interest is the loop (ring) system. This type of system connects all terminals and/or computer by a common bus or loop. The major advantages are the simple routing algorithm and ease in control of information. Farmer and Newhall [21] proposed and constructed a loop system with bursty traffic, which connects various devices such as teletype, plotter, cathode-ray tube display, disk control unit, and computer together. Pierce [22] proposed a hierarchy of interconnected loop systems with random-length messages segmented into fixed-size blocks and provides a scheme for transferring information among the various levels of loops. Konheim and Meister [23] analyzed such a hierarchical system. Hayes and Sherman [24] studied the message delay due to buffering for a single-loop system. The data source is assumed to be of a bursty nature. The traffic generated by each user is assumed to be identically distributed with uniformly distributed destinations. Konheim and Meister [25] studied the loop system as a priority service system. Messages may enter the system at any input terminal located on the loop. Priority is assigned on the basis of position on the loop; the terminal closest to the computer has highest priority and priorities decrease with distance from the computer. The stationary queue lengths and average virtual waiting time are calculated. Spragins [26] has calculated the waiting time with Poisson arrival process. A variant of this priority scheme is the multidrop system in which the central processing unit (CPU) serves each terminal on the loop in sequence; when a terminal is served, it retains the use of the channel until its buffer is empty. The channel is then idle for a certain number of slots for framing and synchronization information before resuming the service of the next terminal. The stationary-queue-length distribution and delay have been reported in [27].

## II. Unified Model for a Class of Computer Communication Systems

In Section I, we discussed the recent advances in traffic characterization and analysis of several computer communication systems. This motivates us to develop a unified approach to the analysis of a variety of computer communication systems that is the theme of Section II. We begin with a summary of results deferring the analysis until Sections II-B to I.

### A. Summary of Results

We shall first consider the single terminal system as shown in Fig. 1; a single buffered terminal linked to a central station (e.g., a CPU) by a communication channel. This system transfers data from the terminal to the central station. We imagine that the time axis $0 \leq t < \infty$ has been divided into contiguous intervals $(j-1)\Delta \leq t < j\Delta$ $(1 \leq j < \infty)$; we call this the $j$th slot (Fig. 2). Each slot is capable of transferring a single *data unit*. The data unit may be viewed as a character, a byte, or a fixed-size block. In the terminology of queueing theory, the channel is a server and the data units play the role of customers. Since the data units are of fixed length, the service time of each customer is constant $\Delta$. Furthermore, the service operation may start only at times of the form $j\Delta$ $(0 \leq j < \infty)$.

The performances in such a system that are of interest are 1) the delay experienced by a group of $m$ data units and 2) the buffer size required to achieve a certain level of overflow probability.

For mathematical tractability, we have assumed in our analysis buffers of infinite capacity. If, in this case, the probability that a buffer contains more than $n$ data units is sufficiently small, then we can be reasonably confident that the system with a buffer of this capacity will exhibit approximately the same behavior.

The arrival of traffic at a buffer will be described by stationary random processes. Let $X$ be the random number of arrivals (in data units) during a slot interval and $\mu = E\{X\}$ and $\sigma^2 = \text{var}\{X\}$ be the expectation and variance of $X$. We show in Section III-D that the expected value and variance of the stationary buffer queue length are

$$E\{L^*\} = \frac{1}{2}\frac{\sigma^2}{1-\mu} + \tfrac{1}{2}\mu \tag{1}$$

$$\text{var}\{L^*\} = \frac{\mu_3}{3(1-\mu)} + \left(\frac{1}{2}\frac{\sigma^2}{1-\mu} + \tfrac{1}{2}(1-\mu)\right)^2 - \tfrac{1}{12}(2\mu-1)(2\mu-3), \tag{2}$$

where $\mu_3$ is the third central moment $E\{(X-\mu)^3\}$ of the input process. The traffic intensity (and channel utilization) $\rho$ is equal to $\mu$.

While the analysis in Section II will not assume a particular distribution for input traffic, numerical results will be presented for a compound Poisson process motivated in part by the results obtained in [2]. For this process, the times at which messages arrive are determined by a standard Poisson process (with rate $\lambda$ (messages/slot interval)) while the lengths of the messages are geometrically distributed with mean length $\bar{m} = 1/(1-q)$ with $0 \leq q < 1$. The probability of $j$ data units arriving at a terminal during a slot interval is

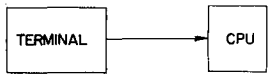


Fig. 1. The single terminal system.

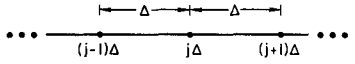


Fig. 2. Graphical representation of time slots.

$$\Pr\{X = j\} = \begin{cases} e^{-\lambda}, & j = 0 \\ e^{-\lambda} \sum_{k=1}^{j} \binom{j-1}{k-1} q^{j-k} \dfrac{(\lambda(1-q))^k}{k!}, & \end{cases}$$

$$1 \leq j < \infty. \qquad (3)$$

The mean and variance of $X$ are $\mu = \lambda/(1-q)$ and $\sigma^2 = \lambda(1+q)/(1-q)^2$. We show in Section II-D that for this process (3), $\Pr\{L^* \leq n\}$, the probability that the buffer contains no more than $n$ data units is

$$\Pr\{L^* \leq n\} = \begin{cases} \dfrac{1-\lambda-q}{1-q}, & n = 0 \\ \dfrac{1-\lambda-q}{1-q} \left[ e^{\lambda n} + \sum_{k=1}^{n-1} \sum_{j=1}^{n-k} \binom{n-k-1}{j-1} q^{n-k-j} \right. & \\ \left. \cdot e^{\lambda k} \dfrac{(-\lambda k(1-q))^j}{j!} \right], & 1 \leq n < \infty. \end{cases} \qquad (4)$$

Numerical results for selected message lengths are portrayed in Fig. 3. Comparison of (4) with the overflow probability in the finite waiting-room model [16] shows that the former is a good approximation for the low-overflow-probability region.

Next we consider the star system (Fig. 4), which consists of a central station and $N$ terminals linked by a common buffer.

When we assume that the traffic is independent from terminal to terminal, the number of data units that arrive at the common buffer during a slot interval is the sum of $N$ independent processes. The formulas for the expected value and variance of the stationary queue length in the buffer are (1) and (2) except that we replace $\mu$, $\sigma^2$, and $\mu_3$ by

$$\sum_{i=1}^{N} \mu^{(i)}, \qquad \sum_{i=1}^{N} \sigma^{(i)2}, \qquad \sum_{i=1}^{N} \mu_3{}^{(i)},$$

where $X^{(i)}$ denotes arrivals at the $i$th terminal and

$$\mu^{(i)} = E\{X^{(i)}\}, \ \sigma^{(i)2} = \operatorname{var}\{X^{(i)}\},$$

$$\mu_3{}^{(i)} = E\{(X^{(i)} - \mu^{(i)})^3\}.$$

If the $i$th arrival process is the compound Poisson process (3) with rate $\lambda^{(i)}$ and $q^{(i)} = q$ $(1 \leq i \leq N)$, then $\Pr\{L^* \leq n\}$ is given by (4) with $\lambda$ replaced by

$$\sum_{i=1}^{N} \lambda^{(i)}.$$

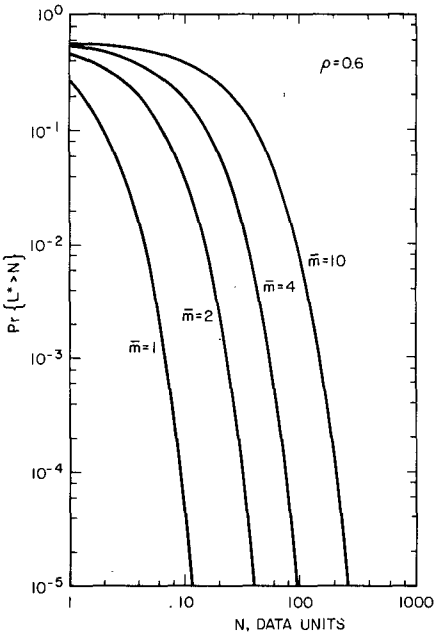


Fig. 3. The stationary-state probability, $\Pr\{L^* > N\}$, for the single terminal system with compound Poisson process input $\rho = 0.6$.

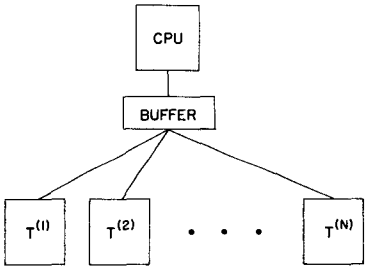


Fig. 4. The star system.

In an STDM loop system (Fig. 5), the rate of removing data units from the buffer at each terminal is generally much higher than the data arrival rate. As a result, the buffer's capacity may be limited to one data unit. If this is not the case, or if data tend to arrive in bursts, then buffering is required. With STDM each terminal is assigned a fixed proportion of the time slots, e.g., every $N$th slot. Service at the $i$th terminal $T^{(i)}$ is thus independent of the traffic at the remaining terminals and the analysis of the behavior of the buffer at $T^{(i)}$ is identical to that of the single terminal model (Fig. 1) after a reinterpretation of the input process. The expected value and variance of the stationary queue length in the buffer at $T^{(i)}$ are

$$E\{L^{(i)*}\} = \frac{1}{2} \frac{N\sigma^{(i)2}}{1 - N\mu^{(i)}} + \tfrac{1}{2} N\mu^{(i)} \qquad (5)$$

$$\operatorname{var}\{L^{(i)*}\} = \frac{N\mu_3{}^{(i)}}{3(1-N\mu^{(i)})} + \left( \frac{1}{2} \frac{N\sigma^{(i)2}}{1-N\mu^{(i)}} + \tfrac{1}{2}(1-N\mu^{(i)}) \right)^2 - \tfrac{1}{12}(2N\mu^{(i)} - 1)(2N\mu^{(i)} - 3), \qquad (6)$$

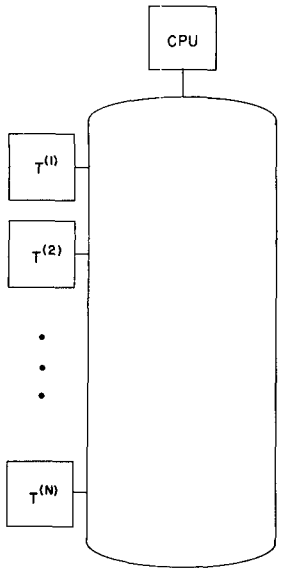


Fig. 5. The loop system.

where $N\mu^{(i)}$, $N\sigma^{(i)2}$, and $N\mu_3{}^{(i)}$ are the mean, variance, and third central moments of the arrival process at $T^{(i)}$ over $N$ contiguous slots; that is, between the start of consecutive service operations at $T^{(i)}$. When the input process is (3) then $\Pr\{L^{(i)*} \leq n\}$ is given by (4) when $\lambda$ is replaced by $N\lambda^{(i)}$ and $q$ by $q^{(i)}$.

To measure the delay in a queueing system with batch arrivals, it is natural to employ the notion of virtual customers; after the start of the $j$th slot we insert $m$ (virtual) data units. Their delay is the difference between queueing time (total time spent in system) and service time. The stationary expected delay for $m$ data units is

$$D_m{}^{(i)} = \frac{N+1}{2} + m(N-1) + N\left(E\{L^{(i)*}\} - \frac{N+1}{2}\mu^{(i)}\right). \quad (7)$$

In an ATDM loop system, the time slots no longer have a fixed relationship with the terminal. The number of service slots that a terminal receives in a fixed-time interval varies and the buffer behavior becomes a complex analytical problem. We first examine the queue discipline that favors terminals closer to the central station of the loop. This introduces a priority structure; a slot is available to the $i$th terminal when and only when the buffers at the first $i-1$ terminals are simultaneously empty. Note that the behavior of the buffer at $T^{(i)}$ is not influenced by the traffic at terminals $T^{(j)}$ with $j > i$. When all input processes are standard Poisson [(3) with $\lambda^{(i)} = \lambda$, $q^{(i)} = 0$ $(1 \leq i \leq N)$] the expected value of the stationary queue length at the $i$th terminal is

$$E\{L^{(i)*}\} = \tfrac{1}{2}\lambda \frac{(i-1)\lambda^2 + (1-(i-1)\lambda)^2}{(1-i\lambda)(1-(i-1)\lambda)^2} + \frac{1}{2}\frac{\lambda}{1-(i-1)\lambda} \quad (8)$$

corresponding to a system utilization $\rho = N\lambda$. Fig. 6 portrays the expected stationary queue length for selected utilizations with identical Poisson input processes. The stationary expected delay for $m$ data units (entering at the $i$th terminal) is

$$D_m{}^{(i)} = \frac{m + \frac{1}{2}\frac{i\lambda}{1-i\lambda} - \frac{1}{2}i\lambda + (i-1)\lambda}{1-(i-1)\lambda} - (m-1) \quad (9)$$

and may be attributed to two sources: 1) data units already buffered at the first $i$ terminals, and 2) data units that join the system at one of the first $i-1$ terminals before completion of the service of the $m$ data units. The second effect reflects the priority structure of the system. Fig. 7 displays (9) for a loop system with ten terminals.

Finally, we study in Section II-H the loop system under still another queue discipline—hub polling. This system, also commonly known as a multidrop network with polling, sequentially serves the $N$ terminals. Instead of allocating a fixed proportion of the channel to each terminal (as in STDM), the channel is assigned to a terminal until the buffer at the terminal is emptied. The channel then remains idle for $r$ slots (for framing and synchronization information) and the process continues with a poll of the next terminal. A single poll of each of the $N$ terminals is defined as a cycle of the system. We restrict our presentation to the case of identically distributed input processes and show that the stationary expected length of the polling cycle is

$$E\{\ell^*\} = \frac{Nr}{1-N\mu}, \quad (10)$$

which corresponds to a system utilization of $\rho = N\mu$. The expected value of the stationary queue length at a terminal is independent of the terminal position index $i$ and is given by

$$E\{L^{(i)*}\} = \frac{1}{2}\frac{\sigma^2}{1-N\mu} + \frac{1}{2}\frac{Nr\mu(1-\mu)}{1-N\mu} \quad (11)$$

while the stationary expected delay for $m$ data units is

$$D_m{}^{(i)} = \frac{1}{2}\frac{N\sigma^2}{1-N\mu} + \frac{1}{2}\frac{Nr(1-\mu)}{1-N\mu} + \tfrac{1}{2}(1-\mu). \quad (12)$$

When input traffic is given by (3) the average queue length as a function of utilization for a system with ten terminals and $r = 1$ slot is shown in Fig. 8.

### *B. The Arrival Process*

The next two sections consist of certain preliminary notions; the main development of the unified model starts in Section II-D.

We will denote random variables by capital letters $X, X_1, X_3{}^{(2)}, \cdots$. These variables will take values in the set of nonnegative integers. *The law of a random variable*

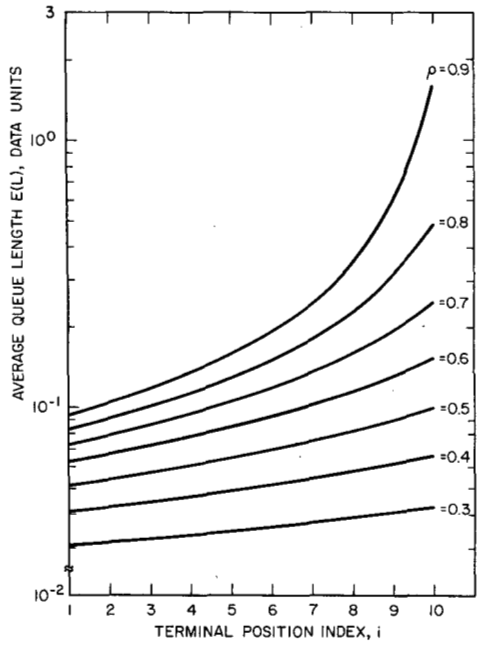


Fig. 6. Stationary expected queue length as a function of terminal position for an ATDM loop system with Poisson input process.

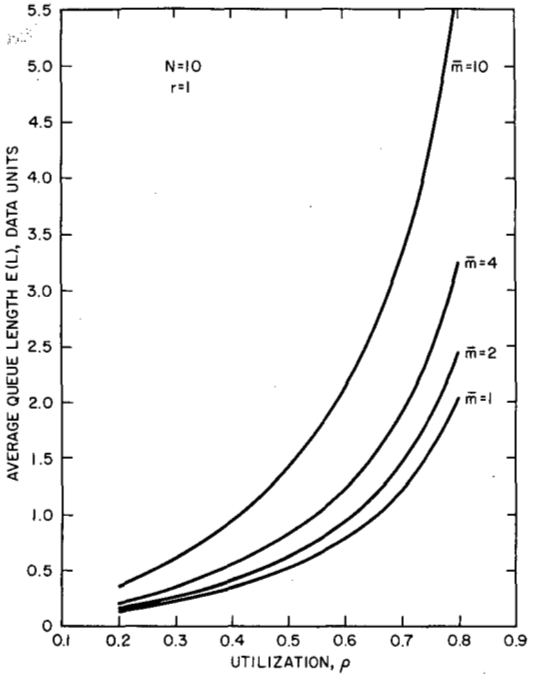


Fig. 8. Stationary expected queue length as a function of utilization for a loop system with hub polling, 10 terminals, $r = 1$, and compound Poisson process input.

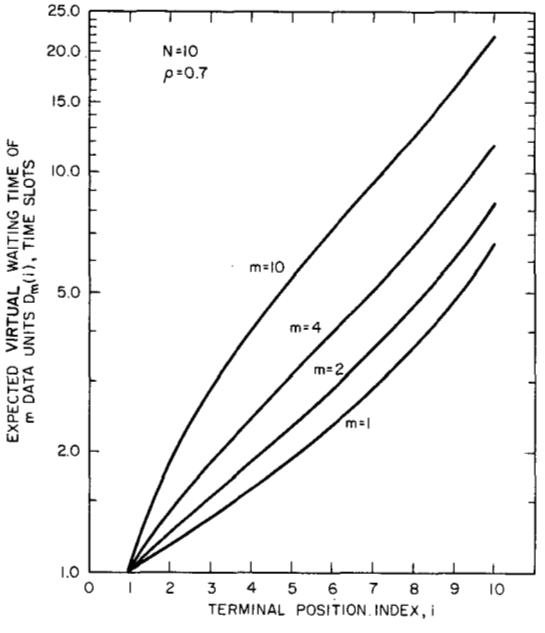


Fig. 7. Stationary expected delay for $m$ virtual data units as a function of terminal position for an ATDM loop system with Poisson input process $\rho = 0.7$.

$X$ is the (probability mass) function

$$p_X(k) = \Pr\{X = k\}, \qquad (0 \leq k < \infty)$$

to which we associate *the probability generating function*

$$P_X(z) = E\{z^X\} = \sum_{k=0}^{\infty} p_X(k) z^k. \tag{13}$$

Since $p_X(k) \geq 0$ $(0 \leq k < \infty)$ and

$$\sum_{k=0}^{\infty} p_X(k) = 1$$

the series (13) converges for all complex $z$ with $|z| < 1$ and thus $P_X(z)$ is an analytic function within the open unit disk $\{z : |z| < 1\}$. The correspondence between $p_X$ (the law) and $P_X$ (the generating function) is biunique; each determines the other. We use the notation $E\{\ \}$ throughout to denote the expectation of the random variable appearing within the brackets. The variance of $X$, denoted by var $\{X\}$, is defined by var $\{X\} = E\{X^2\} - E^2\{X\}$ whenever the indicated moments exist. Generating functions arise in a natural fashion in probability calculations. They play essentially the same manipulative role as do Laplace transforms in classical transient analysis. The moments of a random variable $X$ may be expressed in terms of the derivatives of its generating function $P_X$ at $z = 1$; in particular

$$E\{X\} = P'_X(1) \tag{14}$$

$$\text{var}\{X\} = P''_X(1) + P'_X(1) - (P'_X(1))^2. \tag{15}$$

With a slight abuse of notation we will write $E(P_X)$ and var $(P_X)$ for the right-hand sides of (14) and (15), respectively.

Three features are central in the description of a queueing problem.

1) The arrival process, which specifies how customers (in our case data units) arrive in the system.

2) The service process, which describes the nature and duration of the service operation.

3) The queue discipline, which prescribes the rules of interaction between service facility and customers.

In the service systems that we consider, the service operation is the transfer of data from a terminal to a central station and the server is the channel. The server will alternately be made available to the various sources of customers (terminals) according to the queue discipline and we will examine the effects of this service policy on system performance.

We assume that the time axis $0 \leq t < \infty$ is divided into contiguous intervals or *slots* of common duration as indicated in Fig. 2. The channel may transfer a single data unit within each slot and the service operation may take place only at times of the form $j\Delta$ with $0 \leq j < \infty$. We will specify an arrival process (at a terminal) by counting the number $X$ of data units that arrive in a time interval of length $\Delta$. The arrival process will have the form $\mathfrak{X} = \{X_j : 1 \leq j < \infty\}$, where $X_j$ is the number of data units that arrive during the $j$th slot. We shall assume throughout that the random variables are independent and identically distributed. We denote by $P(z)$ the common generating function $E\{z^{X_j}\}$ (independent of $j$). We will refer to $\mathfrak{X}$ as *a standard input process.*

The simplest example of a standard process is provided by coin tossing; here $X_j = 0$ or 1 (heads or tails) with probabilities $p(0)$ and $1 - p(0)$, respectively. The associated generating function is $P(z) = p(0) + (1 - p(0))z$.

The Poisson process provides us with a second example. A Poisson process $\mathfrak{N}$ counts the number of events $\mathfrak{N}_{[a,b)}$ that take place in the interval $a \leq t < b$. For our case, the event is the arrival of a data unit. Two properties are characteristic of a Poisson process.

1) $\mathfrak{N}_{[a_1,b_1)}$ and $\mathfrak{N}_{[a_2,b_2)}$ are independent provided $a_1 < b_1 \leq a_2 < b_2$.

2)

$$\Pr\{\mathfrak{N}_{[a,b)} = k\} = \frac{(\lambda(b-a))^k}{k!} \exp[-\lambda(b-a)].$$

The parameter $\lambda$ is called the rate of the process and is equal to the expected number of events in a unit interval. The random variables $\{X_j : 1 \leq j < \infty\}$ with $X_j = \mathfrak{N}_{[(j-1)\Delta, j\Delta)}$ constitute a standard process with $\mu = E\{X_j\} = \lambda\Delta$ and $\sigma^2 = \text{var } X_j = \lambda\Delta$. An important and simple generalization of the standard Poisson process is obtained by allowing for multiple arrivals at a Poisson epoch. Thus we assume that the arrival times of messages are determined by a standard Poisson process with rate $\lambda$ (in messages/unit time) but the number of data units that arrive at one of these Poisson epochs is governed by an auxiliary (and independent) law $\mathfrak{M}$. The resulting process is a compound Poisson process. If $M(z)$ is the generating function of this auxiliary law $\mathfrak{M}$ and $X_j = \mathfrak{N}_{[(j-1)\Delta, j\Delta)}$ then $\{X_j : 1 \leq j < \infty\}$ is a standard process and $P(z) = \exp \lambda(M(z) - 1)$. The choice $M(z) = z$ corresponds to the standard Poisson process.

We say that the message lengths are geometrically distributed if $M(z) = (1 - q)z/(1 - qz)$ with $0 \leq q < 1$. The average message length is $\bar{m} = 1/(1 - q)$ while the mean and variance of the resulting input process are

$$\mu = E\{X_j\} = \lambda\Delta/(1 - q)$$

and

$$\sigma^2 = \text{var}\{X_j\} = \lambda\Delta(1 + q)/[(1 - q)^2].$$

### C. The Problem of Gambler's Ruin

We continue the preliminaries with a short examination of the problem of gambler's ruin. With the proper interpretation, the equations describing the fortunes of our gambler will also describe fluctuations in the contents of a buffer. The connection will be supplied in Sections II-D to H.

A gambler begins with an initial capital of $X_0$ and plays a sequence of independent and identical games. The $i$th game produces a (nonnegative integral) gain of $X_i$ from which is deducted an entrance fee of one unit per game. The net worth of the gambler after the play of the $j$th game is thus

$$L_j = X_0 + X_1 + \cdots + X_j - j, \qquad (0 \leq j < \infty). \quad (16)$$

Play continues until ruin; that is, until $L_j \leq 0$. Gambler's ruin occurs after game $T$ where

$$T = \min\{k : L_k \leq 0\}. \quad (17)$$

The problem is to describe the law of $T$ in terms of the initial capital $X_0$ and the payoff on each game.

Introduce the generating functions $K(z) = E\{z^{X_0}\}$ and $P(z) = E\{z^{X_j}\}(1 \leq j < \infty)$. Note that the payoff process $\{X_j : 1 \leq j < \infty\}$ is a standard process in our sense.

In order to reflect real life we further assume that $E\{X_0\} < \infty$ and $E\{X_j\} < 1$, $(1 \leq j < \infty)$. That is, the gambler begins with a finite expected amount of capital and the game favors the house. As is well known (in mathematics as well as real life) the ruin of the gambler is certain; the event $\{T = \infty\}$ corresponding to an unlimited number of plays, is an event of probability zero.

We start with the recurrence formula

$$g_{n,k} = \sum_{j=1}^{k+1} g_{n-1,j} p(k - j + 1),$$

$$(1 \leq n < \infty; 0 \leq k < \infty) \quad (18)$$

with $p(k) = \Pr\{X_j = k\}$ and $g_{n,k}$ = the probability that ruin has not occurred prior to the $n$th game and $L_n = k$. Note that this even can occur in one of the following $k + 1$ mutually exclusive ways: 1) no ruin prior to game $n - 1$, $L_{n-1} = j$ followed by 2) a gain of $k - j + 1$ on the $n$th game with $1 \leq j \leq k + 1$. Equation (18) sums the probabilities of these $k + 1$ possible events. Define

$$G_n(z) = \sum_{k=0}^{\infty} g_{n,k} z^k \qquad G(z, w) = \sum_{n=0}^{\infty} G_n(z) w^n$$

the above series converging when $|z| < 1$, $|w| < 1$. Observe that $g_{n,0} = G_n(0) = \Pr\{T = n\}$ and hence $G(0, w) = E\{w^T\}$ is the probability generating function of the law of $T$. We start by translating (18) into a relationship in-

volving the generating functions. Multiply (18) by $z^k$ and sum $0 \leq k < \infty$ to obtain

$$G_n(z) = P(z)\frac{G_{n-1}(z) - G_{n-1}(0)}{z}, \qquad (1 \leq n < \infty)$$

and then by $w^n$ and sum $1 \leq n < \infty$ to obtain

$$G(z, w) = \frac{zK(z) - wP(z)G(0, w)}{z - wP(z)}. \tag{19}$$

The numerator in (19) contains an unknown boundary term $G(0, w)$. It is determined by an appeal to analyticity. According to Rouché's theorem [28] the denominator of (19) has a simple zero $z = \theta(w)$

$$\theta(w) - wP(\theta(w)) = 0 \tag{20}$$

of modulus less than 1 for each $w$, $|w| < 1$. By carefully examining (20) for $w$ real and $0 \leq w < 1$ we easily see that $\theta(w) \to 1$ as $w \to 1$. By differentiating (20) we find $(d^k)/(dw^k)\ \theta(w) \mid_{w=0} \geq 0$, $(1 \leq k < \infty)$ so that $\theta$ is itself the generating function of some probability law.

Since $G(z, w)$ is analytic in $|z| < 1$, $|w| < 1$, the vanishing of the denominator when $z = \theta(w)$ requires that the numerator must also vanish at this point. This yields the following proposition.

*Proposition 1:*

$$G(0, w) = E\{w^T\} = K(\theta(w))$$

providing us with the solution, the generating function of the law of $T$ in terms of the generating function of the initial capital $X_0$ and the payoff structure as reflected by the generating function $\theta$.

In general the explicit determination of $\theta$ (and hence $G(0, w)$) requires the solution of a (generally) transcendental equation (20) and may present some difficulties.[1] Even when the explicit form of $\theta$ is not calculated much can be learned from (20); if $\mu = E\{X_j\}$, $\sigma^2 = \text{var}\ \{X_j\}$ then by differentiating (20) at $w = 1$ we find

$$E(\theta) = \frac{1}{1 - \mu} \tag{21}$$

$$\text{var}\ (\theta) = \frac{\sigma^2}{(1 - \mu)^3}, \tag{22}$$

which implies

$$E\{T\} = E\{L_0\}\frac{1}{1 - \mu}$$

and

$$\text{var}\ \{T\} = \text{var}\ \{L_0\}\left(\frac{1}{1 - \mu}\right)^2 + E\{L_0\}\frac{\sigma^2}{(1 - \mu)^3}.$$

[1] A simple case of interest is $P(z) = \exp \lambda(z - 1)$. Here it is known [29] that $\theta(w)$ is given by

$$\theta(w) = \frac{1}{\lambda}\sum_{m=1}^{\infty}\frac{m^{m-1}}{m!}(\lambda we^{-\lambda})^m \quad (|\lambda we^{-\lambda}| < 1).$$

### D. The Single Terminal Model

We now return to the main development and show what the relationship is between gambler's ruin and computer communications systems. The system of Fig. 1 consists of a single buffered terminal linked to a CPU. Data from the terminal are generated by a standard process $\mathfrak{X} = \{X_j : 1 \leq j < \infty\}$ and the service operation is to transmit these data (one data unit per slot) from the buffer to the CPU. The state of the system can be described by the state variables

$L_j$ = the number of data units buffered at the terminal just prior to the start of the slot for $(j + 1)$th slot $(0 \leq j < \infty)$.

The system evolves according to

$$L_j = (L_{j-1} - 1)^+ + X_j, \quad 1 \leq j < \infty \tag{23}$$

where $a^+ = \max\ (a, 0)$. Indeed, a data unit is removed in the $j$th slot, provided that there is a data unit present (i.e., $L_{j-1} > 0$), leaving a queue of length $(L_{j-1} - 1)^+$. These data units are joined by $X_j$ new arrivals during the interval $(j - 1)\ \Delta \leq t < j\ \Delta$.

Before considering the analysis of (23) let us note some points of similarity and dissimilarity with (16), which describes the fortunes of the gambler. We may draw the following correspondences

removal of data units $\leftrightarrow$ entrance fee/play
arrival of data units $\leftrightarrow$ payoff
initial queue length $\leftrightarrow$ initial capital

between these two problems. When the gambler reaches ruin, the game is over, while ruin in this context corresponds to the emptying of the buffer. Play continues here and the buffer awaits the arrival of the next data unit(s). The essential point is that the emptying of the buffer marks a renewal point of the system; the system begins afresh, independent of the past, at such a renewal point. Technically, we are dealing with a Markov chain; in the problem of gambler's ruin the point 0 is an absorbing point while here it is a reflecting point. The generating function $\theta$ of Section II-C admits the following important interpretation; it is the generating function of *the renewal time* of the system of Fig. 1; the time between consecutive renewal points of the system.

Let

$$H_j(z) = E\{z^{L_j}\}$$

and

$$H(z, w) = \sum_{j=0}^{\infty} H_j(z)w^j$$

observing, as previously, that these generating functions are defined (and analytic) in $|z| < 1$, $|w| < 1$. The equation of evolution (23) translates into a relationship

involving the generating function $H(z, w)$ as in Section II-C

$$H(z, w) = \frac{zH_0(z) + w(z-1)P(z)H(0, w)}{z - wP(z)}, \quad (24)$$

where $H_0(z)$ describes the initial loading of the buffer. The boundary term $H(0, w)$ is determined as in Section II-C; the denominator of (24) vanishes when $z = \theta(w)$ and the analyticity of $H(z, w)$ requires that the numerator have this zero. Thus

$$H(z, w) = \frac{zH_0(z) - w\dfrac{1-z}{1-\theta(w)}P(z)H_0(\theta(w))}{z - wP(z)},$$

which provides the full transient solution for the system of Fig. 1, albeit not in a particularly transparent form. More manageable results can be obtained if we examine the limiting (in time) behavior of this system. It is known [30] that the random variables $\{L_j : 0 \leq j < \infty\}$ converge in law, $L^* = \lim L_j$ (in law), meaning that

$$H^*(z) = E\{z^{L^*}\} = \lim_{j\to\infty} H_j(z) \quad (25)$$

exists. $H^*$ is the generating function of the stationary distribution of the Markov chain $\{L_j : 0 \leq j < \infty\}$. The existence of the stationary distribution requires the hypothesis $E\{X_j\} < 1$ expressing the fact that data should not arrive at the terminal at a rate faster than it can be removed (from the buffer). To evaluate the limit of (25), we employ a Tauberian theorem [31]; accordingly

$$\lim_{w\uparrow 1}(1-w)H(z, w) = H^*(z) = \lim \frac{1}{n+1}\sum_{j=0}^{n} H_j(z)$$

in the sense that the existence of either of the limits implies the existence of the other and their equality. The right-hand limit is the so-called Ceasaro, 1 or $(C, 1)$ limit of the sequence $\{H_j(z) : 0 \leq j < \infty\}$, which admits a simple physical interpretation that we shall discuss later. We multiply (24) by $(1 - w)$ and let $w \uparrow 1$ obtaining

$$H^*(z) = (1-\mu)\frac{(z-1)P(z)}{z - P(z)} \quad (26)$$

with $\mu = E\{X_j\}$. If we write

$$H^*(z) = \sum_{n=0}^{\infty} \Pr\{L^* = n\}z^n$$

then the stationary-state probabilities $\Pr\{L^* = n\}$ can be found from (26) although a closed form determination may in general be tedious. In the special case $P(z) = \exp\lambda[(1-q)z/(1-qz) - 1]$, $\Pr\{L^* \leq n\}$ is given by (4).

Even when the stationary-state probabilities cannot easily be found, the moments of the stationary law of $L^*$ can be found from (26) by differentiation at $z = 1$ and we obtain $E\{L^*\}$ and var $\{L^*\}$ as given in (1) and (2).

The equality

$$H^*(z) = (C, 1)\lim H_j(z) = \lim \frac{1}{n+1}\sum_{j=0}^{n} H_j(z)$$

admits an interesting physical interpretation; the limit above is a time average and $\Pr\{L^* = n\}$ is the expected fraction of time (as measured in slots) that the contents of the buffer equals $n$. The choice $n = 0$ corresponds to the nonutilization of the server (channel) and hence the system utilization $\rho = 1 - \Pr\{L^* = 0\} = \mu$.

This waiting-line analysis provides also a measure of the delay. We will suppose that a group of $m$ data units enters the system at time $(j+1)\Delta + 0$; that is, just after the start of the $j$th slot.[2] This group will be delayed by the presence of $(L_j - 1)^+$ data units. Its delay is defined as the queueing time (total time spent in the system) less the service time. Thus the delay $1 + (L_j - 1)^+$ is independent of $m$. To find the stationary expected delay we take expectation and let $j \to \infty$ to obtain $D_m = E\{L^*\} + 1 - \mu$. We now will show how this model (with a suitable reinterpretation) presents a unified approach to the analysis of a class of computer communication systems.

### E. The Star System

The star system (Fig. 4) consists of a central station (CPU) and $N$ terminals linked by a common buffer. Each terminal is provided with a separate access to the buffer which collects data for the CPU. Data arrive at the $N$ terminals according to random processes $\mathfrak{X}^{(1)}, \mathfrak{X}^{(2)}, \cdots, \mathfrak{X}^{(N)}$

$$\mathfrak{X}^{(i)} = \{X_j^{(i)} : 1 \leq j < \infty\},$$

where the superscript refers to terminal number. Each process is a standard input process and we assume the $N$ processes independent. The state of the buffer is described by the state variable

$L_j$ = the total number of data units buffered just prior to the start of the $(j+1)$th slot

and the system evolves according to the equation

$$L_j = (L_{j-1} - 1)^+ + X_j^{(1)} + X_j^{(2)} + \cdots + X_j^{(N)}, \quad 1 \leq j < \infty. \quad (27)$$

This is just (23), where $X_j$ there is replaced by $X_j^{(1)} + X_j^{(2)} + \cdots + X_j^{(N)}$. The analysis of Section II-D is applicable; if $P^{(i)}(z) = E\{z^{X_j^{(i)}}\}$ is the generating function of the process $\mathfrak{X}^{(i)}$ then the generating function of the

[2] The choice of when the virtual customers enter the system is quite arbitrary; we could equally have chosen to have them enter at time $(j+1)\Delta + 0$. What we have in mind is a measure of delay for the next group of $m$ data units that enters after the start of the $j$th slot.

effective input process in the star system

$$\{X_i^{(1)} + X_i^{(2)} + \cdots + X_i^{(N)} : 1 \leq j < \infty\}$$

is

$$P(z) = P^{(1)}(z)P^{(2)}(z) \cdots P^{(N)}(z).$$

The laws of $\{L_j : 0 \leq j < \infty\}$ converge provided $\mu^{(1)} + \mu^{(2)} + \cdots + \mu^{(N)} < 1$ and the limiting law is given by

$$H^*(z) = E\{z^{L^*}\} = \lim H_j(z) = \left(1 - \sum_{k=1}^{N} \mu^{(k)}\right) \frac{(z-1)P^{(1)}(z)P^{(2)}(z) \cdots P^{(N)}(z)}{z - P^{(1)}(z)P^{(2)}(z) \cdots P^{(N)}(z)}.$$

The formulas for the expected value and variance of the stationary queue length $L^*$ are given in (1) and (2) when we reinterpret the input process; this requires that we replace $\mu$, $\sigma^2$, and $\mu_3$ in (1) and (2) by

$$\sum_{i=1}^{N} \mu^{(i)}, \qquad \sum_{i=1}^{N} \sigma^{(i)2}, \qquad \sum_{i=1}^{N} \mu_3^{(i)},$$

respectively. For the compound Poisson process of (3), $P^{(i)}(z) = \exp \lambda^{(i)}[(1-q)z/(1-qz) - 1]$ $(1 \leq i \leq N)$, a simplification is possible. If we let $\lambda = \lambda^{(1)} + \lambda^{(2)} + \cdots + \lambda^{(N)}$ then

$$E\{L^*\} = \frac{1}{2}\frac{(1+q)\lambda}{(1-q)(1-q-\lambda)} + \frac{1}{2}\frac{\lambda}{1-q}$$

$$\text{var}\{L^*\} = \frac{\lambda(1+4q+q^2)}{3(1-q)^2(1-q-\lambda)} + \left(\frac{1}{2}\frac{(1+q)\lambda}{(1-q)(1-q-\lambda)} + \frac{1}{2}\left(1 - \frac{\lambda}{1-q}\right)\right)^2 - \frac{1}{12}\left(\frac{2\lambda}{1-q} - 1\right)\left(\frac{2\lambda}{1-q} - 3\right)$$

and $\Pr\{L^* \leq n\}$ is given by (4) with $\lambda$ defined as previously.

### *F. The Loop System With Synchronous Time Division Multiplexing*

The loop system (Fig. 5) consists of a central station and $N$ buffered terminals. The terminals are linked by a single channel that is alternately made available to the terminals. Let us first consider the case in which data can be transmitted only from the terminals to the central station.

In STDM the set of slots

$$\mathcal{S} = \{s_i = [(j-1)\Delta, j\Delta) : 1 \leq j < \infty\}$$

is partitioned into $N$ subsets $\mathcal{S}^{(i)} = \{s_{kN+i} : 0 \leq k < \infty\}$ by assigning every $N$th slot to terminal $T^{(i)}$. Data from the $i$th terminal can be transmitted to the central station only in the slots $\mathcal{S}^{(i)}$ assigned to it. The input processes $\mathfrak{X}^{(1)}, \mathfrak{X}^{(2)}, \cdots, \mathfrak{X}^{(N)}$ are as in Section II-E and the state variables are

$L_i^{(i)}$ = the number of data units buffered at the $i$th terminal just prior to the arrival of the $(j+1)$th slot for $0 \leq j < \infty$, $1 \leq i \leq N$.

Note that for fixed $j$, the state variables $L_j^{(1)}, L_j^{(2)}, \cdots, L_j^{(N)}$ refer to the state of the system at different instances of time; the difference is related to the transit time of a slot from terminal to terminal. This difference will play no role in the analysis of this system or those to be discussed in Sections II-G and II-H.

The essential point here is that the contents of the buffers are independent random variables. The service capacity of the system has been divided in a fixed manner among the $N$ competing users independent of the actual traffic. The analysis of the queue length in the buffer at any terminal, say $T^{(i)}$, reduces to the analysis of the single terminal system (Fig. 1) with a reinterpretation of the input process. The state variable $L_j^{(i)}$ experiences changes due to two effects: 1) the removal of a data unit, and 2) the arrival of new data units. The first change can take place only at times of the form $j\Delta$ with $j - i + 1 = 0$ (modulo $N$).

Let $\mathfrak{L}_j^{(i)} = L_{(j+1)N+i-1}^{(i)}$ and

$$\mathfrak{Y}_j^{(i)} = \sum_{k=1}^{N} X_{jN+i+k-1}^{(i)}.$$

We have

$$\mathfrak{L}_j^{(i)} = (\mathfrak{L}_{j-1}^{(i)} - 1)^+ + \mathfrak{Y}_j^{(i)}, \quad 1 \leq j < \infty, 1 \leq i \leq N, \tag{28}$$

which is (23) with $X_i$ there replaced by $\mathfrak{Y}_j^{(i)}$. Note that $\mathfrak{Y}_j^{(i)}$ is the total number of data units that arrive during $N$ contiguous slots; that is, between the start of consecutive service operations at the $i$th terminal. The results of Section II-D show that the random variables $\{\mathfrak{L}_j^{(i)} : 0 \leq j < \infty\}$ converge in law provided $E\{\mathfrak{Y}_j^{(i)}\} = NE\{X_i^{(i)}\} = N\mu^{(i)} < 1$. The limiting law $L^{(i)*} = \lim \mathfrak{L}_j^{(i)}$ has generating function

$$H^{(i)*}(z) = E\{z^{L^{(i)*}}\} = (1 - N\mu^{(i)})\frac{(z-1)(P^{(i)}(z))^N}{z - (P^{(i)}(z))^N}$$

from which we obtain the moments $E\{L^{(i)*}\}$ and $\text{var}\{L^{(i)*}\}$ given in (5) and (6).

In measuring delay we encounter two phenomena; the delay due to customers already in the system and the delay due to the nature of the service operation. Suppose a group of $m$ data units enters the system at the $i$th terminal just after the start of slot $s_{jN+i+\alpha-1}$ with $0 \leq \alpha < N$. The contents of the buffer is

$$(L_{jN+i-1}^{(i)} - 1)^+ + \sum_{k=1}^{\alpha} X_{jN+k+i-1}^{(i)}.$$

To obtain the stationary expected delay (7), we average the above equation over possible entry times $(0 \leq \alpha < N)$, take expectation and let $j \to \infty$.

### *G. The Loop System With Asynchronous Time Multiplexing [25]*

An undesirable effect of the fixed-slot assignment of STDM is the creation of long queues and the attendant delays. In this section we study the same system under

a second service policy. The first terminal, in the sense of position on the loop, that requires a slot receives the slot. This service policy establishes a priority structure; $T^{(1)}$ has the highest priority and priority decreases with terminal position number. Priorities depend upon position of the terminal on the loop and we study the effect of priority on the grade of service at each terminal.

With the same state variables as in Section II-F, the system evolves according to the equations

$$L_j^{(i)} = (L_{j-1}^{(i)} - (1 - L_{j-1}^{(1)} - L_{j-1}^{(2)} - \cdots - L_{j-1}^{(i-1)})^+)^+ + X_j^{(i)}, \quad 1 \le j < \infty, 1 \le i \le N. \tag{29}$$

The $j$th slot is available to the $i$th terminal if and only if it is not taken by one of the first $i-1$ terminals; that is, if and only if $L_{j-1}^{(1)} + L_{j-1}^{(2)} + \cdots + L_{j-1}^{(i-1)} = 0$. Let $T_1^{(i)} < T_2^{(i)} < \cdots$ denote the times at which the event $L_j^{(1)} + L_j^{(2)} + \cdots + L_j^{(i-1)} = 0$ occurs. These are the epochs of service at the $i$th terminal. We may then replace (29) by

$$L_{T_j^{(i)}}^{(i)} = (L_{T_{j-1}^{(i)}}^{(i)} - 1)^+ + \sum_{k=1+T_{j-1}^{(i)}}^{T_j^{(i)}} X_k^{(i)}. \tag{30a}$$

For simplicity in notation, we let

$$\mathfrak{L}_j^{(i)} = L_{T_j^{(i)}}^{(i)} \qquad \mathcal{Y}_j^{(i)} = \sum_{k=1+T_{j-1}^{(i)}}^{T_j^{(i)}} X_k^{(i)}.$$

Then (30a) becomes

$$\mathfrak{L}_j^{(i)} = (\mathfrak{L}_{j-1}^{(i)} - 1)^+ + \mathcal{Y}_j^{(i)}. \tag{30b}$$

We recognize that (30b) takes the same form as (23) with $X_j$ there replaced by $\mathcal{Y}_j^{(i)}$. Note that $\mathcal{Y}_j^{(i)}$ is the sum of $T_j^{(i)} - T_{j-1}^{(i)}$ independent and identically distributed random variables. Suppose we prove the following.

*Proposition 2:* The renewal times $\{T_j^{(i)} - T_{j-1}^{(i)}: 1 < j < \infty\}$ are independent and identically distributed random variables with generating function

$$\theta^{(i)}(w) = E\{w^{(T_j^{(i)} - T_{j-1}^{(i)})}\},$$

where $\theta^{(i)}(w)$ is the unique solution of the equation

$$\theta^{(i)}(w) - w \prod_{k=1}^{i-1} P^{(k)}(\theta^{(i)}(w)) = 0, \tag{31}$$

which is of modulus less than one for each $w$, $|w| < 1$.

The random variables $\{\mathcal{Y}_j^{(i)}: 1 < j < \infty\}$ will then constitute a standard process in our sense and the results of Section II-D are applicable to the study of the system (30b). If $H_j^{(i)}(z) = E\{z^{\mathfrak{L}_j^{(i)}}\}$ then the $\{\mathfrak{L}_j^{(i)}: 0 \le j < \infty\}$ converges in law $E\{z^{\mathfrak{L}^{(i)*}}\} = H^{(i)*}(z) = \lim H_j^{(i)}(z)$ provided $E\{\mathcal{Y}_j^{(i)}\} < 1$ $(1 < j < \infty)$. Now $\{\mathcal{Y}_j^{(i)}: 1 < j < \infty\}$ is a compound process. Its generating function is $\theta^{(i)}(P^{(i)})$ and hence

$$H^{(i)*}(z) = (1 - E\{\mathcal{Y}_2^{(i)}\}) \frac{(z-1)\theta^{(i)}(P^{(i)}(z))}{z - \theta^{(i)}(P^{(i)}(z))}. \tag{32}$$

It remains to prove the proposition and calculate $E\{\mathcal{Y}_j^{(i)}\}$.

We start by observing that (29) implies

$$L_j^{(1)} + L_j^{(2)} + \cdots + L_j^{(i-1)} = (L_{j-1}^{(1)} + L_{j-1}^{(2)} + \cdots + L_{j-1}^{(i-1)} - 1)^+ + X_j^{(1)} + X_j^{(2)} + \cdots + X_j^{(i-1)}, \tag{33}$$

which again is a variant of (23) with $L_j$ replaced by $L_j^{(1)} + L_j^{(2)} + \cdots + L_j^{(i-1)}$ and $X_j$ by $X_j^{(1)} + X_j^{(2)} + \cdots + X_j^{(i-1)}$. In particular, the times at which the event $L_j^{(1)} + L_j^{(2)} + \cdots + L_j^{(i-1)} = 0$ occurs are the service epochs $\{T_j^{(i)}: 1 \le j < \infty\}$ at the $i$th terminal. As we have observed in Section II-D the increments $\{T_j^{(i)} - T_{j-1}^{(i)}: 1 < j < \infty\}$[3] are independent and identically distributed with generating function $\theta^{(i)}(w)$, the solution of (31). The mean and variance of the renewal time $T_j^{(i)} - T_{j-1}^{(i-1)}$ have already been calculated (21) and (22). We thus obtain the formulas

$$E\{T_j^{(i)} - T_{j-1}^{(i)}\} = \frac{1}{1 - \sum_{k=1}^{i-1} \mu^{(k)}} \tag{34}$$

$$\text{var}\{T_j^{(i)} - T_{j-1}^{(i)}\} = \frac{\sum_{k=1}^{i-1} \sigma^{(k)2}}{\left(1 - \sum_{k=1}^{i-1} \mu^{(k)}\right)^3}, \tag{35}$$

where $\mu^{(k)} = E\{X_j^{(k)}\}$ and $\sigma^{(k)2} = \text{var}\{X_j^{(k)}\}$. The compound process $\{\mathcal{Y}_j^{(i)}: 1 < j < \infty\}$ is the effective input process at the $i$th terminal and its mean and variance are

$$E\{\mathcal{Y}_2^{(i)}\} = \frac{\mu^{(i)}}{1 - \sum_{k=1}^{i-1} \mu^{(k)}} \tag{36}$$

$$\text{var}\{\mathcal{Y}_2^{(i)}\} = \frac{\mu^{(i)2} \sum_{k=1}^{i-1} \sigma^{(k)2} + \left(1 - \sum_{k=1}^{i-1} \mu^{(k)}\right)^2 \sigma^{(i)2}}{\left(1 - \sum_{k=1}^{i-1} \mu^{(k)}\right)^3} \tag{37}$$

so that the stability condition $E\{\mathcal{Y}_2^{(i)}\} < 1$ becomes $\mu^{(1)} + \mu^{(2)} + \cdots + \mu^{(i)} < 1$. In order that all terminals possess stationary distributions, we then require $\mu^{(1)} + \mu^{(2)} + \cdots + \mu^{(N)} < 1$. The stationary state probabilities may be found from (32), which we may now write, by virtue of (36), in the form

$$H^{(i)*}(z) = \frac{1 - \sum_{k=1}^{i} \mu^{(k)}}{1 - \sum_{k=1}^{i-1} \mu^{(k)}} \frac{(z-1)\theta^{(i)}(P^{(i)}(z))}{z - \theta^{(i)}(P^{(i)}(z))}. \tag{38}$$

The moments of $L^{(i)*}$ are given by (1) and (2) when we replace $\mu$, $\sigma^2$, and $\mu_3$ by the corresponding moments of the effective input process. The first two of these moments are given in (36) and (37).

A simplification in the formulas is achieved when we

[3] The first renewal epoch $T_1^{(i)}$ depends upon the initial loading of the buffers at the first $i-1$ terminals. If $L_0^{(j)} = 0$ $(1 \le j < i)$, then $E\{w^{T_1^{(i)}}\} = \theta^{(i)}(w)$ also. This will cause no difficulties since we are interested only in the limiting behavior.

assume that the terminal input traffic from all terminals are identical Poisson processes $P^{(i)}(z) = \exp \lambda(z - 1)$ $(1 \leq i \leq N)$. The expected value of the stationary queue length at the $i$th terminal is given in (8) and corresponds to a system utilization of $\rho = N\lambda$.

To calculate the delay for a group of $m$ data units that join at time $j\Delta + 0$ we observe that there are two sources of delay: 1) data units already buffered at terminals $T^{(1)}$, $T^{(2)}$, $\cdots$, $T^{(i)}$, and 2) data units that join subsequently at terminals $T^{(1)}$, $T^{(2)}$, $\cdots$, $T^{(i-1)}$ but before the completion of the service of these $m$ data units. The second source of the delay reflects the preferential service related to position on the loop. The problem of gambler's ruin provides us with the solution. Consider this problem with

$$L_0 = \text{initial capital} \sim m + \left(\sum_{k=1}^{i} L_i^{(k)} - 1\right)^+ + \sum_{k=1}^{i-1} X_i^{(k)}$$

$$X_k \sim X_{k+i}^{(1)} + \cdots + X_{k+i}^{(i-1)} (1 \leq k < \infty)$$

$$= \text{cumulative arrival process at the first } i - 1 \text{ terminals.}$$

The notation $\sim$ indicates that the two random variables are equal in distribution. The queueing time (total time spent in system) is equal to $1 + T$, where $T$ is the time to ruin. Using Proposition 1 and letting $j \to \infty$ we find

$$D_m^{(i)} = \frac{m + \frac{1}{2}\left(\sum_{k=1}^{i} \sigma^{(k)2}\right)\Big/\left(1 - \sum_{k=1}^{i} \mu^{(k)}\right) - \frac{1}{2}\sum_{k=1}^{i} \mu^{(k)} + \sum_{k=1}^{i-1} \mu^{(k)}}{1 - \sum_{k=1}^{i-1} \mu^{(k)}} - (m - 1). \tag{39}$$

When all input processes are identical standard Poisson processes $P^{(i)}(z) = \exp \lambda(z - 1)$, (39) reduces to (9).

*H. The Loop System With Hub Polling.*

We conclude with an examination of the loop system under another queue discipline—hub polling. The $N$ terminals are queried in sequence for service; if a terminal responds "yes," it retains control of the channel until its buffer is emptied. The channel is then idle for $r$ slots to provide for addressing information, end of message, check digits, and line control information. The same query is then addressed to the next terminal and the process repeats until all $N$ terminals have been polled. We call this *a cycle of the system*. The polling procedure is thereafter repeated.

The state of the system can be described by the vector-valued state variable $\boldsymbol{L}_i = (L_i^{(1)}, L_i^{(2)}, \cdots, L_i^{(N)})$ of Section II-F. The generating function of this vector variable is the analytic function of $N$ variables

$$F_i(z_1, z_2, \cdots, z_N) = E\{z_1^{L_i^{(1)}} z_2^{L_i^{(2)}} \cdots z_N^{L_i^{(N)}}\} = \sum_{\boldsymbol{k}=(k_1,k_2,\cdots,k_N)} \Pr\{\boldsymbol{L}_i = \boldsymbol{k}\} z_1^{k_1} z_2^{k_2} \cdots z_N^{k_N}$$

defined for $\boldsymbol{z}$ in the polydisk $\{(z_1, z_2, \cdots, z_N): |z_1| < 1, |z_2| < 1, \cdots, |z_N| < 1\}$. We limit our discussion to the special case of identically distributed input processes, i.e., $P(z) = P^{(i)}(z) (1 \leq i \leq N)$.

The idea of the analysis is to study the cyclic nature of the service operation. It will turn out that when we look at the state of the system at the start of a cycle, it will exhibit regularity. The process sampled at these points will constitute a Markov chain with stationary distribution. We will sketch the development; the details will be found in [27].

We begin by considering the operation "empty the buffer at $T^{(i)}$" and the effect this has on a generating function. If $F(z_1, z_2, \cdots, z_N)$ is the generating function of the state at the start of this operation, then

$$\prod_{k=1}^{N} P^r(z_k) F(z_1, z_2, \cdots, z_{i-1}, \theta\Big(\prod_{\substack{k=1 \\ k \neq i}}^{N} P(z_k)\Big), z_{i+1}, \cdots, z_N) \tag{40}$$

is the generating function at the start of the operation "empty the buffer at $T^{(i+1)}$." For proof of (40) note that $F(z_1, z_2, \cdots, z_N)$ provides a complete probabilistic description of the state of the system; in particular, the number of data units buffered at $T^{(i)}$. According to the gambler's ruin problem, this determines the length of time needed to empty this buffer. While this buffer is being emptied, data are simultaneously arriving at the remaining $N - 1$ terminals and (40) retains count of these arrivals. The initial term

$$\prod_{k=1}^{N} P^r(z_k)$$

represents a count of the arrivals during the $r$ idle slots. We will denote this operation with the notation $F \to \mathcal{E}_i F$ with $\mathcal{E}_i F$ given by (40). The effect of a single cycle in the polling process corresponds to the operator $\mathcal{C}F = \mathcal{E}_N \mathcal{E}_{N-1} \cdots \mathcal{E}_2 \mathcal{E}_1 F$. If $F$ describes the state of the system at the start of the $p$th cycle then $\mathcal{C}F$ describes the state at the start of the $(p+1)$th cycle. We seek a steady-state solution; that is, some distribution of the contents of the buffer at the start of a cycle $F$, which remains unchanged after a single cycle

$$\mathcal{C}F = F. \tag{41}$$

At this point, we may make use of the symmetry that results when all input processes have the same distribution; (41) may be replaced by

$$(\mathcal{E}_1 F)(z_1, z_2, \cdots, z_N) = F(z_2, z_3, \cdots, z_N, z_1). \tag{42}$$

Two important properties of the solution of (42) (or (41)) can be proved: 1) there exists a unique solution of (42), and 2) for any initial distribution $G$, $\mathcal{C}^p G \to F$ as $p \to \infty$. We may then regard $F$ as a stationary distribution for the contents of the buffers at the start of a cycle.

The actual calculation of the $F$ that satisfies (42) is quite straightforward. We introduce a family of auxiliary functions

$$H^{(k)}(z_1, z_2, \cdots, z_{N-1}) = \begin{cases} z_{1-k}, & -(N-2) \le k \le 0 \\ \theta\left(\prod_{i=1}^{N} P(z_i)\right), & k = 1 \\ H^{(1)}(H^{(k-1)}(z_1, \cdots, z_{N-1}), \cdots, \\ H^{(k-N+1)}(z_1, \cdots, z_{N-1})), \\ & 1 < k < \infty. \end{cases} \quad (43)$$

Each $H^{(k)}$ is a probability generating function. Then (40) and (42) yield the relationship

$$F(z_1, z_2, \cdots, z_N) = (P(z_N))^r \cdot \prod_{k=-(N-2)}^{0} (P(H^{(k)}(z_1, z_2, \cdots, z_{N-1}))^r F(H^{(1)}(z_1, \cdots, z_{N-1}), H^{(0)}(z_1, \cdots, z_{N-1}), \cdots, H^{-(N-2)}(z_1, \cdots, z_{N-1})). \quad (44)$$

We iterate (44) and find

$$F(z_1, z_2, \cdots, z_N) = \prod_{k=1}^{N} (P(z_k))^{(N-k+1)r} \cdot \prod_{k=1}^{\infty} P(H^{(k)}(z_1, z_2, \cdots, z_{N-1}))^{Nr}. \quad (45)$$

The validity of the argument depends upon proving the convergence of the infinite product (45). For this purpose it suffices to prove

$$\sum_{k=1}^{\infty} \frac{\partial}{\partial z_j} H^{(k)}(z_1, z_2, \cdots, z_{N-1}) < \infty$$

when $\mathbf{z} = (z_1, z_2, \cdots, z_N) = \mathbf{1} = (1, 1, \cdots, 1)$ and $1 \le j \le N-1$. Here we make use of the recursive definition (43); if

$$a_{k,j} = \frac{\partial}{\partial z_j} H^{(k)}(z_1, z_2, \cdots, z_{N-1})|_{z=1}$$

then (43) yields

$$a_{k,j} = \frac{\mu}{1-\mu} \sum_{t=1}^{N-1} a_{k-t,j}, \qquad 1 \le k < \infty. \quad (46)$$

We introduce the generating functions

$$A_j(z) = \sum_{k=1}^{\infty} a_{k,j} z^k, \qquad 1 \le j \le N-1; \ |z| < 1.$$

Then (46) yields

$$A_j(z) = \frac{\frac{\mu}{1-\mu} z \frac{1-z^{N-j}}{1-z}}{1 - \frac{\mu}{1-\mu} z \frac{1-z^{N-1}}{1-z}}$$

so that

$$A_j(1) = \sum_{k=1}^{\infty} \frac{\partial}{\partial z_j} H^{(k)}(z_1, \cdots, z_{N-1})|_{z=1} = \frac{\mu(N-j)}{1-N\mu}.$$

The infinite product converges whenever $N\mu < 1$, which again expresses the natural requirement that data should not enter the system at a rate faster than it can be removed. We have also proved

$$\frac{\partial}{\partial z_1} F(z_1, z_2, \cdots, z_N)|_{z=1} = \frac{Nr\mu(1-\mu)}{1-N\mu}. \quad (47)$$

Equation (47) gives the expected contents of the buffer at $T^{(1)}$ at the start of a cycle. A similar argument (involving generating functions) proves

$$\frac{\partial^2}{\partial z_1^2} F(z_1, z_2, \cdots, z_N)|_{z=1} = \sigma^2 Nr\left(\frac{1-(N+1)\mu+(2N-1)\mu^2}{(1-N\mu)^2}\right) - \frac{Nr\mu(1-\mu)}{1-N\mu} - \frac{N^2r^2\mu^2(1-\mu)^2}{(1-N\mu)^2}. \quad (48)$$

It remains for us to define what we mean by the limiting behavior of the system. Suppose $\{\tau_i : 1 \le i < \infty\}$ are the times at which the cycles start with $\tau_1 = 0$. The ratio

$$E\left\{\frac{1}{\tau_i} \sum_{i=0}^{\tau_i - 1} z^{L_j(1)}\right\} \quad (49)$$

is an average of the function $z^{L_j(1)}$ over the first $i-1$ cycles. It can be shown that this ratio converges as $i \to \infty$ with limit

$$H^{(1)*}(z) = \frac{1-N\mu}{Nr}\left\{\frac{1-F(z,1,\cdots,1)}{1-P(z)} + \frac{z(F(z,1,\cdots,1)-1)}{z-P(z)}\right\}, \quad (50)$$

which we can interpret as the time-averaged generating function for the contents of the buffer at terminal one. It averages the generating function $F_i(z, 1, \cdots, 1)$ over the first $i-1$ cycles with $i \to \infty$. The limit exists and is given by (50) for every initial distribution $F_0(z_1, z_2, \cdots, z_N)$ of the buffers. Since service at each of the terminals is the same, the superscript 1 appearing on the left-hand side of (50) can be replaced by any index $i$, $1 \le i \le N$. In principle, (50) provides us with the stationary-state probabilities $\Pr\{L^{(i)*} = n\}$, but their calculation is formidable. The stationary expected length of the queue at the $i$th terminal is found by differentiating (50) at $z = 1$. We require the intermediate results (47) and (48) to obtain the formula of equation (11).

It remains to describe the delay for a group of $m$ data units that enters as before just after the start of some slot. Suppose they enter at terminal $T^{(i)}$. There are two sources of delay for this group: 1) the group may enter in that part of a cycle during which the channel is available to $T^{(i)}$ or 2) in a part of the cycle during which the channel is either idle or available to one of the remaining terminals. In the former case, the group is delayed only by the data

units already buffered at $T^{(i)}$. In the latter case they are delayed by these units and by the time delay until the channel is next made available to $T^{(i)}$. Averaging over many cycles as in (49), we find the delay formula of (12).

### I. Models With Two-Way Traffic

We have restricted our attention thus far to the problem of one-way traffic, from a terminal to the central station. We comment here briefly on what modifications are possible when we admit the possibility of data flow in two directions. The channels are assumed to be half-duplex. Consider the loop system with asynchronous time division multiplexing and suppose that a slot may be used for either CPU-to-terminal or terminal-to-CPU transfer. The channel will be made available alternately to the CPU and terminals. The simplest model assumes that the $j$th slot contains data from the CPU with probability $q$ and is free (and hence available for terminal-to-CPU transfers) with probability $1 - q$. The CPU-to-terminal traffic is thus modeled by a 0th terminal with an input process $\mathfrak{X}^{(0)} = \{X_j^{(0)}: 1 \leq j < \infty\}$, which is a realization of coin tossing. The analysis of this model follows along the lines established in Sections II-D and II-G. The essential point is that a slot may be used for either terminal-to-CPU or CPU-to-terminal transfers but not both. We can generalize by allowing for periods during which the channel is not available (see [32]) and by more general CPU-to-terminal processes (see [33]).

A second model is from a view of store and forward networks. We first assign highest priority to the traffic from the terminals-to-CPU. Traffic from the CPU to a terminal is temporarily buffered at intermediate terminals when a conflict over a slot arises. That is, if a slot containing a data unit from the CPU for terminal $T^{(i)}$ reaches terminal $T^{(j)}$ (with $j < i$) and finds that the buffer there contains a data unit for the CPU, it is stored there, freeing the slot for this higher priority transmission. This buffered data unit remains at $T^{(j)}$ until a free slot appears and then resumes its journey. The analysis of this system has been carried out in [33] (see also [34]).

This completes the discussion of Section II. We note that the results in Section II are closely related to those in Section I. For example, the input traffic to the buffer is based on the measured computer-traffic characteristics; the fixed message block size used in the unified model can be determined from the fixed block-size model and channel error characteristics; the buffer behavior of the star system is related to that of the statistical multiplexor buffer; and the buffer behavior of the STDM loop system. The ATDM loop system and hub polling are special cases of the general loop system.

## III. Conclusions

Recent advances in computer communication systems have been summarized. Based on queueing theory, a unified model has been developed that can be used to analyze a class of computer communication systems including the star and loop systems. Such a unified approach provides us with more insight into the system behavior (delay, buffer overflow) and performance tradeoffs of these systems. These play an important role in the planning and optimization of computer communication systems.

## References


[1] P. E. Jackson and C. D. Stubbs, "A study of multi-access computer communications," in *1969 Spring Joint Computer Conf., AFIPS Conf. Proc.,* vol. 34. Washington, D. C.: Spartan, pp. 491–504.

[2] E. Fuchs and P. E. Jackson, "Estimates of distributions of random variables for certain computer communications traffic models," in *Proc. ACM Symp. on Problems in the Optimization of Data Communications Systems,* Pine Mountain, Ga., Oct. 1969, pp. 202–225.

[3] A. L. Dudick, E. Fuchs, and P. E. Jackson, "Data traffic measurements for inquiry–response computer communications systems," in *1971 Proc. Int. Fed. Information Processing Congr.,* Ljubljana, Yugoslavia, Aug. 1971.

[4] A. A. Alexander, R. M. Gryb, and D. W. Nast, "Capabilities of the telephone network for data transmission," *Bell Syst. Tech. J.,* vol. 39, pp. 431–476, May 1960.

[5] R. L. Townsend and R. N. Watts, "Effectiveness of error control in data communications over the switched telephone network," *Bell Syst. Tech. J.,* vol. 43, pp. 2611–2638, Nov. 1964.

[6] C. W. Farrow and L. N. Holzman, "Nationwide field trial performance of a multilevel vestigial sideband data terminal for switched network voice channels," in *Conf. Rec., 1968 IEEE Conf. Communication,* Philadelphia, Pa., June 1968, p. 782.

[7] M. D. Balkovic *et al.,* "1969–70 connection survey: High-speed voiceband data transmission performance on the switched telecommunications networks," *Bell Syst. Tech. J.,* vol. 50, pp. 1349–1384, Apr. 1971.

[8] H. C. Fleming and R. N. Hutchinson, Jr., "1969–70 connection survey: Low-speed data transmission performance on the switched telecommunications networks," *Bell Syst. Tech. J.,* vol. 50, pp. 1385–1405, Apr. 1971.

[9] S. Y. Tong, "A survey of error control techniques on telephone channels," in *Proc. 1970 Nat. Electron. Conf.,* Chicago, Ill., pp. 462–476.

[10] J. J. Kucera, "Transfer rate of information bits," *Comput. Des.,* pp. 56–59, June 1968.

[11] M. D. Balkovic and P. E. Muench, "Effects of propagation delay caused by satellite circuits on data communications systems that use block retransmission for error correction," *Conf. Rec., 1969 IEEE Conf. Communications,* Boulder, Colo., pp. 29/31–29/36, June 1969.

[12] R. L. Kirlin, "Variable block length and transmission efficiency," *IEEE Trans. Commun. Technol.,* vol. COM-17, pp. 350–355, June 1969.

[13] W. W. Chu, "Optimal fixed message block size for computer communications," in *1971 Proc. Int. Fed. Information Processing Congr.,* Ljubljana, Yugoslavia, Aug. 1971.

[14] W. W. Chu, "Design considerations of statistical multiplexors," in *Proc. ACM Symp. on Problems in the Optimization of Data Communications Systems,* Pine Mountain, Ga., 1969, pp. 39–60.

[15] W. W. Chu, "A study of asynchronous time division multiplexing for time-sharing computers," in *1969 Fall Joint Comput. Conf., AFIPS Conf. Proc.,* vol. 35. Montvale, N. J.: AFIPS Press, pp. 669–678.

[16] W. W. Chu, "Buffer behavior for batch Poisson arrivals and single constant output," *IEEE Trans. Commun. Technol.,* vol. COM-18, pp. 613–618, Oct. 1970.

[17] W. W. Chu and L. C. Liang, "Buffer behavior for mixed input traffic and single constant output rate," *IEEE Trans. Commun. Technol.,* vol. COM-20, pp. 230–235, Apr. 1972.

[18] C. D. Pack, "The effect of multiplexing on a computer communication system," submitted to *Commun. Ass. Comput. Mach.*

[19] W. W. Chu, "Demultiplexing considerations for statistical multiplexors," this issue, pp. 603–609.

[20] L. Kleinrock, "Scheduling, queueing and delays in time-shared systems and computer networks," in *Computer-Communication Networks,* N. Abramson and F. Kuo, Eds. Englewood Cliffs, N. J.: Prentice-Hall, 1972, ch. 4.

[21] W. D. Farmer and E. E. Newhall, "An experimental dis-

tributed switching system to handle bursty computer traffic," in *Proc. ACM Symp. Problems in the Optimization of Data Communications Systems,* Pine Mountain, Ga., Oct. 1969.
[22] J. R. Pierce, C. H. Cohen, and W. J. Kropfl, "Network for block switching of data," in *IEEE Conf. Rec.,* New York, N. Y., Mar. 1971.
[23] A. G. Konheim and B. Meister, "Waiting lines in multiple loop systems," *J. Math. Anal. Appl.,* to be published.
[24] F. J. Hayes and D. N. Sherman, "Traffic analysis of a ring switched data transmission system," *Bell Syst. Tech. J.,* vol. 50, Nov. 1971.
[25] A. G. Konheim and B. Meister, "Service in a loop system," *J. Ass. Comput. Mach.,* to be published.
[26] J. D. Spragins, "Loops used for data collection," presented at the 22nd Int. Symp. on Computer-Communication Networks and Teletraffic, Polytech. Inst. Brooklyn, Brooklyn, N. Y. Apr. 1972.
[27] A. G. Konheim and B. Meister, "Waiting lines and times in a system with polling," in preparation.
[28] L. V. Ahlfors, *Complex Analysis.* New York: McGraw-Hill, 1954, p. 124.
[29] E. M. Wright, "Solution of the equation $ze^z = a$," in *Proc. Roy. Soc. Edinburg,* vol. 65, pt. 2, pp. 193–203, 1959.
[30] D. V. Lindley, "The theory of queues with a single server," in *Proc. Cambridge Phil. Soc.,* vol. 48, 1952, pp. 277–289.
[31] E. W. Hobson, *The Theory of Functions of a Real Variable.* Cambridge, England: Cambridge Univ. Press, 1927.
[32] A. G. Konheim and B. Meister, "Two-way traffic in loop service systems," *Networks,* to be published.
[33] A. G. Konheim, "Service epochs in a loop system," presented at the 22nd Int. Symp. on Computer-Communication Networks and Teletraffic, Polytech. Inst. Brooklyn, Brooklyn, N. Y. Apr. 1972.
[34] A. G. Konheim and B. Meister, "Queues and waiting times in a loop with two-way traffic," submitted to *J. Comput. Syst. Sci.*

**Wesley W. Chu** (S'62–M'67), for photograph and biography please see page 609 of this issue.

**Alan G. Konheim** was born in Brooklyn, N. Y., on October 17, 1934. He received the B.E.E. and M.S. degrees from the Polytechnic Institute of Brooklyn, Brooklyn, N. Y., and the Ph.D. degree in mathematics from Cornell University, Ithaca, N. Y., in 1955, 1957, and 1960, respectively.

Since June 1960 he has been a member of the Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y. His principal research interests are in the field of probability theory. He was a Fulbright Scholar at the University of Heidelberg, Heidelberg, Germany, from 1966 to 1967.

Dr. Konheim is a member of the American Mathematical Society, the Mathematical Association of America, the Society for Industrial and Applied Mathematics, Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

# Analysis and Design of Reliable Computer Networks

ROBERT S. WILKOV, MEMBER, IEEE

***Abstract*—In the design of a computer network, one of the fundamental considerations is the reliability and availability of the communication paths between all pairs of centers in the network. These characteristics are strongly dependent on the topological layout of the communication links in addition to the reliability and availability of the individual computer systems and communication facilities. Based on graph theoretic models for computer and communication networks, many different reliability measures have been defined. Attempts have been made to characterize networks that are optimal with respect to these measures. In this paper, the most significant reliability criteria and their relevance to different applications will be discussed. Furthermore, we survey the status of current research on the different criteria. The difficulties and limitations on each reliability measure will be pointed out and what seem to be the most fruitful areas for further investigation will be indicated.**

## I. Introduction

IN THIS PAPER, a computer network is modeled by a linear graph in which the nodes or vertices correspond to computer centers in the network and the edges correspond to the communication links. In the design of a computer network, one of the fundamental considerations is the reliability and availability of the communication paths between all pairs of centers in the network. These characteristics strongly depend on the topological layout of the communication links in addition to the reliability and availability of the individual computer systems and communication facilities. Graph theoretic models for computer and communication networks have previously been used in the literature to characterize maximally reliable networks based on different reliability measures. The difficulties and limitations on each of these measures will be discussed in this paper, in addition to the problem areas that appear to be fruitful for further investigation.

In studies of communication and computer networks, reliability has been defined in a number of different ways. A network has been defined to be operational in the presence of failures provided communication paths exist between certain pairs of nodes. Alternatively, a network has been considered to be operational in the presence of failures if every node could communicate with a certain percentage of the other nodes. However, these definitions would be more meaningful if they quantitatively reflected the traffic-carrying capacity of the network in the presence of failures. For example, a line-switching network,

Manuscript received December 22, 1971; revised February 24, 1972.
The author is with the IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y.