Delay Analysis of a Time Division Multiple Access (TDMA) Channel

SIMON S. LAM, MEMBER, IEEE

Abstract—The delay performance of a Time Division Multiple Access (TDMA) channel for transmitting data messages is considered. The channel is assumed to be fixed assigned to a station with unlimited buffer capacity and Poisson message arrivals. Each message gives rise to one or more packets for transmission into fixed-length time slots. The steady-state probability generating function of the queue size is derived. A formula for the expected message delay is given. The analysis is then generalized to a nonpreemptive priority queue discipline; expected message delay formulas are given for the priority classes.

Paper approved by the Editor for Computer Communication of the IEEE Communications Society for publication after presentation at the National Telecommunications Conference, Dallas, TX, November 29-December 1, 1976. Manuscript received September 15, 1976; revised July 7, 1977.

The author was with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598. He is now with the Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712.

INTRODUCTION

As satellite communications mature, the trend is toward digital transmission and Time Division Multiple Access (TDMA) in preference to the use of Frequency Modulation and Frequency Division Multiple Access (FDMA) which predominate today [1]. The purpose of this paper is to analyze the delay performance of TDMA channels for transmitting data messages. Although this work was originally motivated by satellite TDMA systems, the analysis to be presented is not predicated upon satellite channels.

A station with a fixed assigned TDMA channel and unlimited buffer capacity for queueing is considered. It is assumed that data messages arrive at the station according to an independent Poisson process. From each message, one or more fixed-length packets are formed. (A packet corresponds to the amount of data that can be transmitted into a TDMA time slot.) The number of packets comprising a message is given by a general probability distribution. We solve for the steady-state probability generating function of the station queue size. From this, an explicit formula for the expected message delay (total time in queue and transmission) is obtained. Next, the analysis is generalized to a nonpreemptive priority queue discipline. Expected message delay formulas are obtained for the priority classes.

Consider a TDMA system in which time is slotted and time slots are organized into frames of, say, M slots indexed from 1 to M as shown in Fig. 1. Time slots with the same index in consecutive frames form a TDMA channel. (This is also refered to as *synchronous time division multiplexing.*) Let the duration of a frame be T seconds. A station using a TDMA channel transmits one packet of data into a time slot of T/M seconds (provided the queue is nonempty); it then becomes idle for $(M-1)T/M$ seconds before it can transmit another packet. For the purposes of this paper, it is immaterial to the station how the rest of the frame is shared among other users.

Synchronous time division multiplexing was previously studied by Chu and Konheim [2]. Their model permits a general distribution for the number of packet arrivals within a time slot and they solved for the probability generating function of the queue size (in number of packets) at time instants just prior to the beginning of a time slot. They also obtained the expected delay experienced by a "virtual" message arrival. By assuming Poisson message arrivals and employing a different analytic approach, we obtained different results for the steady-state probability generating function of the queue size (in number of messages) as seen by a random observer as well as the expected delay actually experienced by messages.

It later came to our attention that Hayes [3], using yet another analytic approach, obtained the probability generating function of message delay for the nonpriority case considered below. His expected message delay formula is identical to ours. The nonpreemptive priority queue discipline was also studied under more restrictive assumptions within the context of loop systems by Spragins [4]. In particular, our model reduces to his if we have exactly one slot per frame and one packet per message.

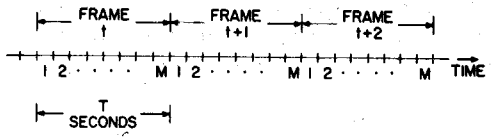

Figure 1. TDMA channels.

THE ANALYSIS

We shall consider a station utilizing a TDMA channel as a single-server queue in the usual sense [5] by adopting the following definition of service time. The service time of a packet includes, by definition, the time during which the packet is at the head of the queue with no transmission in progress and its transmission time of T/M seconds. The service time of a packet is thus exactly equal to T seconds except for the first packet of a message which arrives to find an empty system; the service time of such packets is a random variable (defined to be X) distributed between T/M and $T + T/M$. The service time of a message is the aggregate service time of its constituent packets. The above definition is illustrated in the upper part of Fig. 2 in which we show four consecutive message arrivals and their subsequent departures. In this figure, S_i represents the service time and W_i the waiting time of the ith message. Examples of busy and idle periods are also shown. Note that the idle period I is exponentially distributed under the assumption of Poisson arrivals. The random variable Y, defined to be $X - T/M$, is a function of I. The functional relationship is illustrated in the bottom part of Fig. 2.

Let N_t be the number of messages in the system (both queue and service) at time t. We now proceed to find the steady-state probability generating function of N_t by studying the following queuing system. Consider a single-server queue with Poisson arrivals at λ messages per second. The service time distribution of a message which initiates a busy period is $\hat{B}(x)$ with first and second moments $\hat{b}_1$ and $\hat{b}_2$. All subsequent messages in the same busy period have service times drawn independently from the distfibution $B(x)$ with first and second moments b_1 and b_2. Let

$$P_n = \lim_{t\to\infty} \text{Prob}\,[N_t = n]$$

and define the transforms

$$P(z) = \sum_{n=0}^{\infty} z^n P_n$$

$$B^*(s) = \int_0^{\infty} e^{-sx}\, dB(x)$$

$$\hat{B}^*(s) = \int_0^{\infty} e^{-sx}\, d\hat{B}(x).$$

Theorem 1 (Welch [6]): If $\lambda b_1 < 1$, then

$$P(z) = \frac{P_0[z\hat{B}^*(\lambda - \lambda z) - B^*(\lambda - \lambda z)]}{z - B^*(\lambda - \lambda z)} \tag{1}$$

where

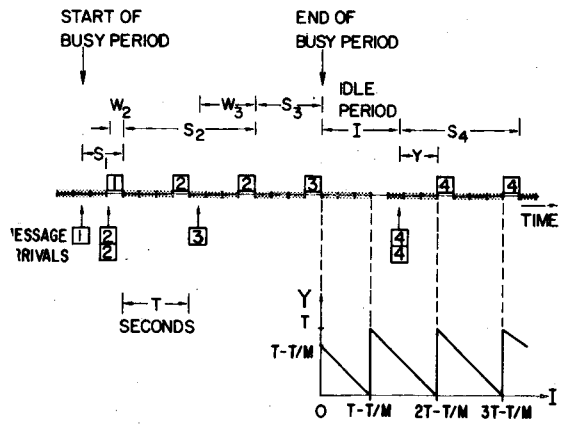

Figure 2. Illustration of busy and idle periods, waiting and service times.

$$P_0 = \frac{1 - \lambda b_1}{1 - \lambda(b_1 - \hat{b}_1)}. \tag{2}$$

Proof: The proof is similar to that for the $M/G/1$ queue [5, Vol. I]. Eq. (1) can be easily obtained by considering the imbedded Markov chain of queue size at service completion times. Also, since it is assumed that the arrival process is Poisson and that messages arrive and depart individually, the steady-state probability generating function of the queue size at service completion times is equal to the steady-state probability generating function of N_t sampled at random. Eq. (2) is obtained by evaluating Eq. (1) at $z = 1$. Q.E.D.

The above theorem is now used to obtain formulas for the expected message delay and expected queue size of a station using a TDMA channel. Higher moments of message delay and queue size can be obtained in a similar manner. However, only averages will be considered in what follows.

Let the number of packets in a message be L which is a random variable given by the probability density $\{g_l\}_{l=1}^{\infty}$ with first and second moments L_1 and L_2 where

$$g_l = \text{Prob}\,[L = l].$$

The message service time distributions are given by the following transforms

$$B^*(s) = \sum_{l=1}^{\infty} g_l[\beta^*(s)]^l$$

$$\hat{B}^*(s) = \hat{\beta}^*(s) \sum_{l=1}^{\infty} g_l[\beta^*(s)]^{l-1}$$

where

$$\beta^*(s) = e^{-sT}$$

$$\hat{\beta}^*(s) = e^{-st/M} \int_0^T e^{-sy}\, dF(y)$$

$$F(y) = \text{Prob}\,[Y \leqslant y].$$

The lower part of Fig. 2 illustrates Y as a function of the idle

period I which is exponentially distributed. (We note that this dependence does not affect the proof of Theorem 1.) $F(y)$ is obtained to be

$$F(y)=\begin{cases}\dfrac{e^{-\lambda(T-\frac{T}{M})}[e^{\lambda y}-1]}{1-e^{-\lambda T}}+U\left(y-T+\dfrac{T}{M}\right)\\ \quad\cdot[1-e^{-\lambda(T-\frac{T}{M}-y)}] & 0\leqslant y\leqslant T\\ 1 & y>T\end{cases}$$

where

$$U(x)\triangleq\begin{cases}1 & x\geqslant 0\\ 0 & x<0.\end{cases}$$

The first and second moments of Y are given by

$$\bar{Y}=\left(T-\frac{T}{M}-\frac{1}{\lambda}\right)+\frac{Te^{-\lambda(T-\frac{T}{M})}}{1-e^{-\lambda T}} \tag{3}$$

$$\overline{Y^2}=\left(T-\frac{T}{M}-\frac{1}{\lambda}\right)^2+\left(\frac{1}{\lambda}\right)^2+\left(T^2-2\frac{T}{\lambda}\right)\frac{e^{-\lambda(T-\frac{T}{M})}}{1-e^{-\lambda T}} \tag{4}$$

The first and second moments of X are thus given by

$$\bar{X}=\bar{Y}+\frac{T}{M} \tag{5}$$

$$\overline{X^2}=\overline{Y^2}+2\frac{T}{M}\bar{Y}+\left(\frac{T}{M}\right)^2. \tag{6}$$

The first and second moments of message service times are

$$b_1=L_1T \tag{7}$$

$$\hat{b}_1=(L_1-1)T+\bar{X} \tag{8}$$

$$b_2=L_2T^2 \tag{9}$$

$$\hat{b}_2=\overline{X^2}+2\bar{X}(L_1-1)T+(L_2-2L_1+1)T^2. \tag{10}$$

The probability of an empty system is obtained from Eqs. (2), (3), (5), (7) and (8).

$$P_0=\frac{1-\lambda b_1}{1-\lambda(b_1-\hat{b}_1)}=\frac{(1-\lambda L_1T)(1-e^{-\lambda T})}{\lambda Te^{-\lambda(T-T/M)}}. \tag{11}$$

The expected number of messages in the system $\bar{N}$ is obtained by evaluating the first derivative of Eq. (1) at $z=1$.

$$\bar{N}=\frac{\lambda\hat{b}_1}{1-\lambda(b_1-\hat{b}_1)}+\frac{\lambda^2(\hat{b}_2-b_2)}{2[1-\lambda(b_1-\hat{b}_1)]}+\frac{\lambda^2b_2}{2(1-\lambda b_1)}. \tag{12}$$

An application of Little's formula [5, Vol. II, p. 17] gives the expected message delay

$$\bar{D}=\frac{\hat{b}_1}{1-\lambda(b_1-\hat{b}_1)}+\frac{\lambda(\hat{b}_2-b_2)}{2[1-\lambda(b_1-\hat{b}_1)]}+\frac{\lambda b_2}{2(1-\lambda b_1)}. \tag{13}$$

The expected service time of a message is

$$\begin{aligned}\bar{S}&=(1-P_0)b_1+P_0\hat{b}_1\\&=b_1-P_0(b_1-\hat{b}_1)\\&=\frac{\hat{b}_1}{1-\lambda(b_1-\hat{b}_1)}\end{aligned} \tag{14}$$

which is just the first term in the expected message delay formula. Hence, the expected waiting time is

$$\bar{W}=\frac{\lambda(\hat{b}_2-b_2)}{2[1-\lambda(b_1-\hat{b}_1)]}+\frac{\lambda b_2}{2(1-\lambda b_1)}. \tag{15}$$

Now if we substitute Eqs. (3)–(10) into Eqs. (12)–(15), we obtain the following results:

$$\bar{N}=\lambda L_1T-\frac{\lambda T}{2}+\frac{\lambda T}{M}+\frac{\lambda^2L_2T^2}{2(1-\lambda L_1T)} \tag{16}$$

$$\bar{D}=L_1T-\frac{T}{2}+\frac{T}{M}+\frac{\lambda L_2T^2}{2(1-\lambda L_1T)} \tag{17}$$

$$\bar{S}=\frac{1}{\lambda}+\frac{\left(L_1T-\frac{1}{\lambda}\right)(1-e^{-\lambda T})}{\lambda Te^{-\lambda(T-T/M)}} \tag{18}$$

$$\bar{W}=L_1T-\frac{T}{2}+\frac{T}{M}+\frac{\lambda L_2T^2}{2(1-\lambda L_1T)}-\frac{1}{\lambda}-\frac{\left(L_1T-\frac{1}{\lambda}\right)(1-e^{-\lambda T})}{\lambda Te^{-\lambda(T-T/M)}}. \tag{19}$$

The *expected message delay* given in Eq. (17) is a very interesting result. Suppose the burst transmission rate is C bits/s. The data rate of a TDMA channel is thus C/M bits/s. Now consider a FDMA channel which transmits continuously at C/M bits/s. The expected message delay for such a channel is given by the Pollaczek-Khinchin (P-K) formula [5, Vol. I. Eq. (5.6)] to be

$$\bar{D}_{PK} = L_1 T + \frac{\lambda L_2 T^2}{2(1-\lambda L_1 T)}.$$

Thus the expected message delay for a TDMA channel can be expressed in terms of the expected message delay for a FDMA channel at the same data rate as follows.

$$\bar{D} = \bar{D}_{PK} - \frac{T}{2} + \frac{T}{M}. \tag{20}$$

This last is an interesting result because of its simplicity. It is not obvious, however, because one might expect from Welch's model [6] that as the traffic intensity

$$\rho \triangleq \lambda L_1 T$$

approaches 1, the difference $\bar{D}_{PK} - \bar{D}$ would vanish to zero because the expected duration of a busy period becomes infinite.

Finally, we note that if the duration T of a time frame is shrunk to zero (by, for example, decreasing the packet length) while the distribution of message lengths remains fixed, we have

$$\lim_{T \to 0} \bar{D} = \bar{D}_{PK}.$$

NONPREEMPTIVE PRIORITY QUEUE DISCIPLINE

We next consider the same system as in the previous section except that now messages belong to a number of priority classes. The priority level of a message may depend, for examples, upon its length, source, destination(s) and/or its type (data or control). The discipline being considered is the nonpreemptive head-of-the-line discipline of Cobham [7].

Let there be K message priority classes indexed from 1 to K where 1 denotes the highest priority level, K the lowest. Messages in the k^{th} priority class are assumed to arrive according to an independent Poisson process at λ_k messages per second. Define

$$\lambda = \sum_{k=1}^{K} \lambda_k.$$

The number of packets in a class k message is given by the probability density $\{g_l^{(k)}\}_{l=1}^{\infty}$ with first and second moments $L_1^{(k)}$ and $L_2^{(k)}$ respectively. At the service completion of a message, the server will serve next a message with the highest priority level. The first-in-first-out (FIFO) rule is assumed for messages which belong to the same prioirity class. Define

$$\rho_k = \lambda_k L_1^{(k)} T.$$

The first two moments $b_1^{(k)}$, $\hat{b}_1^{(k)}$, $b_2^{(k)}$ and $\hat{b}_2^{(k)}$ of message service times for the K priority classes are defined as in the previous section. Let

$$b_1 = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} b_1^{(k)}$$

$$\hat{b}_1 = \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} \hat{b}_1^{(k)}.$$

Theorem 2: If $\Sigma_{i=1}^{K} \rho_i < 1$, then the expected waiting time of a class k message is

$$\bar{W}_k = \frac{V}{\left(1 - \sum_{i=1}^{k-1} \rho_i\right)\left(1 - \sum_{i=1}^{k} \rho_i\right)} \qquad k = 1, 2, \cdots, K \tag{21}$$

where

$$V = \frac{P_0}{2} \sum_{k=1}^{K} \lambda_k \hat{b}_2^{(k)} + \frac{1 - P_0}{2} \sum_{k=1}^{K} \lambda_k b_2^{(k)} \tag{22}$$

where P_0 is given by Eq. (2). The expected service time of a class k message is

$$\bar{S}_k = P_0 \hat{b}_1^{(k)} + (1 - P_0) b_1^{(k)}. \tag{23}$$

The expected message delay of a class k message is

$$\bar{D}_k = \bar{S}_k + \bar{W}_k. \tag{24}$$

Proof: Eq. (21) is obtained by noting that Cobham's result [7] can be applied directly here except for V, which is the expected remaining service time of the message in service found by an arbitrary arrival. From Wolff [8], we see that V is given by

$$V = \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{i=1}^{n(\tau)} \frac{S_i^2}{2} \tag{25}$$

where $n(\tau)$ is a random variable denoting the number of message arrivals in the time interval $[0, \tau]$, and S_i is the service time of the i^{th} message arrival. By considering arrivals from each priority class separately and noting that P_0 is equal to the fraction of (Poisson) arrivals which find the system empty, Eq. (22) is obtained from Eq. (25) in the limit of $\tau \to \infty$. Derivations of Eqs. (23) and (24) are obvious. Q.E.D.

Observations: In the special case of $K = 1$ (no priority classes), Eq. (24) for the expected message delay reduces to Eq. (13) in the previous section.

As an example, we consider the shortest-message-first (SMF) priority discipline with FIFO between messages of equal length. The expected waiting time of a message of l packets is

$$\bar{W}_l = \frac{V}{\left(1 - \lambda \sum_{i=1}^{l-1} i g_i T\right)\left(1 - \lambda \sum_{i=1}^{l} i g_i T\right)} \tag{26}$$

where

$$V = \frac{\lambda P_0}{2} \hat{b}_2 + \frac{\lambda(1 - P_0)}{2} b_2. \tag{27}$$

The expected service time of a message of l packets is

$$\bar{S}_l = P_0[(l-1)T + \bar{X}] + (1 - P_0) lT$$
$$= lT - P_0(T - \bar{X}). \tag{28}$$

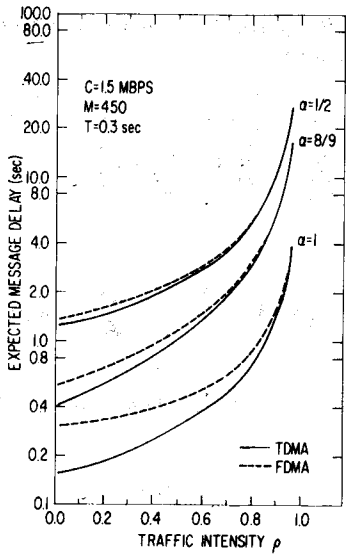

Figure 3. Expected message delay versus traffic intensity.

NUMERICAL EXAMPLES

In this section we examine the delay performance tradeoffs through some numerical examples. Let the burst transmission rate be C bits/s. In Fig. 3, we consider a system with $C = 1.5$ Mbits/s, $M = 450$, $T = 0.3$ second, and 1000 bits/packet. Messages which arrive at the earth station under consideration consist of either single-packets or eight-packets with $g_1 = \alpha$ and $g_8 = 1 - \alpha$. Three cases are shown for $\alpha = 1$, 8/9 and 1/2 corresponding to (1) single-packet messages only, (2) single-packet and eight-packet messages at equal packet rate, and (3) single-packet and eight-packet messages at equal message rate. Fig. 3 shows typical delay versus traffic intensity curves with expected delay going to infinity as ρ approaches unity. Fig. 3 also shows the expected message delay given by the P-K formula of a FDMA channel at the same data rate. Note that although the absolute difference in delay between TDMA and the P-K formula is equal to $(T/2) - (T/M)$, the normalized difference decreases as ρ increases or as the average message length increases (α decreases).

In Fig. 4, we consider the effect of the frame time T on the expected message delay. We assume that each message is 8000 bits long. At $C = 1.5$ Mbits/s, $M = 300$ and $T = 1.6$ seconds, each message can be transmitted as a single packet. Now suppose the frame time T is halved by splitting each packet into two. This process is repeated until T approaches 0. (The $T = 0$ limiting case corresponds to a FDMA channel at the same data rate.) The expected message delay is plotted versus T in Fig. 4. As shown by Eq. (20), the slope of each curve is $-(\frac{1}{2} - 1/M)$. The average message delay increases to D_{PK} in the $T \to 0$ limit as predicted. The reader is cautioned not to interpret Fig. 4 as evidence that a large packet size is necessarily desirable for TDMA. The example in Fig. 4 assumes a fixed message length. When the message length is random, a large packet size may be inefficient due to the large unfilled portion in the last packet of each message. Fig. 4 does tell us that if there is a natural minimum data block size in the distribution of message lengths, it should not be split up further into smaller units.

In Fig. 5, we show the delay versus traffic intensity curves for a TDMA channel with the shortest-message-first queue discipline. We assume in this example that $C = 1.5$ Mbits/s, $M = 450$, $T = 0.3$ second, 1000 bits per packet and the message length distribution, $g_1 = 0.3$ and $g_l = 0.1$ for $l = 2, 3, \cdots, 8$. Note in Fig. 5 that with a priority discipline, not only do high priority messages (short messages in this case) have a smaller delay than low priority messages (long messages in this case), they also have a finite delay even when the traffic intensity is equal to 1. It is interesting to note that the difference in delay between TDMA and FDMA vanishes to zero as ρ increases to 1 for $l \leqslant 7$; for $l = 8$, the difference actually increases as ρ increases to 1 (conservation law! [5, Vol. II]) although this is not observable in Fig. 5 due to the logarithmic scale.

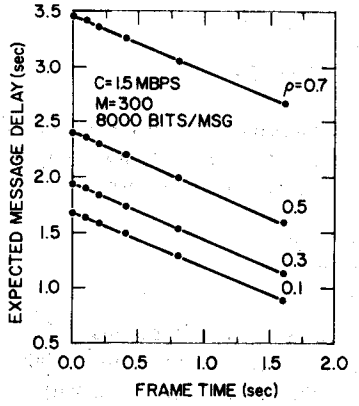

Figure 4. Expected message delay versus frame time.

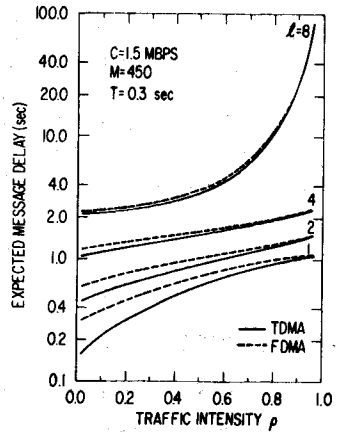

Figure 5. Expected message delay versus traffic intensity for SMF queue discipline.

Finally in Fig. 6 we show for the same example expected message delay versus message length for the SMF queue discipline. Note that the expected message delay increases with the message length. This is often a desirable feature since most applications typically require a much smaller delay constraint for short messages (interactive data traffic) than long messages (batch data traffic).

CONCLUSION

We have analyzed the delay performance of a TDMA channel for transmitting data messages. The steady-state probability generating function of the station queue size is obtained. Explicit formulas for the expected message delay and queue size are given. (Higher moments of message delay and queue size

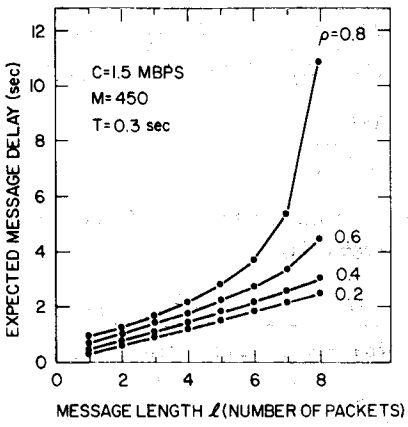

Figure 6. Expected message delay versus message length for SMF queue discipline.

can be readily obtained from the probability generating functions). The delay analysis is also generalized to obtain delay formulas for a system using a nonpreemptive priority queue discipline. A priority queue discipline can be used to take advantage of data traffic consisting of different classes of messages with disparate delay constraints. An example of priority based upon message length is illustrated.

This paper was motivated by satellite systems for data traffic. We have made the assumption that the station under consideration has sufficient data traffic to warrant efficient use of a fixed assigned TDMA channel. This is probably a realistic situation for current domestic satellite systems. If, in the future, the trend is toward small earth stations characterized by bursty data traffic, other satellite packet switching techniques such as slotted ALOHA and packet reservation should be considered [9].

ACKNOWLEDGMENT

The author would like to thank W-M. Chow, S. Lavenberg and L.S. Woo of the IBM Thomas J. Watson Research Center for helpful discussions on the analysis.

REFERENCES

[1] Schmidt, W. G., "Satellite Time-Division Multiple Access Systems: Past, Present and Future," *Telecommunications*, Vol. 7, August 1973, pp. 21-24.

[2] Chu, W. W. and A. G. Konheim, "On the Analysis and Modeling of a Class of Computer Communication Systems," *IEEE Trans. on Commun.*, Vol. COM-20, June 1972, pp. 645-660.

[3] Hayes, J. F., "Performance Models of an Experimental Computer Communication Network," *Bell System Technical Journal*, Vol. 53, February 1974, pp. 225-259.

[4] Spragins, J., "Simple Derivation of Queueing Formulas for Loop Systems," *IEEE Trans. on Commun.*, Vol. COM-25, April 1977, pp. 446-448.

[5] Kleinrock, L., *Queueing Systems Vol. I: Theory*, Wiley-Interscience, New York, 1975; *Queueing Systems Vol. II: Computer Applications*, Wiley-Interscience, New York, 1976.

[6] Welch, P. D., "On a Generalized M/G/1 Queueing Process in which the First Customer of Each Busy Period Receives Exceptional Service," *Operations Research*, Vol. 12, 1964, pp. 736-752.

[7] Cobham, A., "Priority Assignment in Waiting Line Problems," *Operations Research*, Vol. 2, 1954, pp. 70-76; "A Correction," *ibid.*, Vol. 3, 1955, p. 547.

[8] Wolff, R. W., "Work-Conserving Priorities," *Journal of Applied Probability*, Vol. 7, 1970, pp. 327-337.

[9] Lam, S. S., "Satellite Multiaccess Schemes for Data Traffic," *Conf. Rec. International Conf. on Communications*, Chicago, Il, June 1977, pp. 37.1-19 to 37.1-24.