DELAY ANALYSIS OF A PACKET-SWITCHED TDMA SYSTEM

Simon S. Lam

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

A satellite Time Division Multiple Access (TDMA) channel for transmitting packetized data messages is considered. The queue length distribution is obtained using generating functions for an earth station which uses a TDMA channel for packet broadcast. A formula for the expected message delay is given. The analysis assumes Poisson message arrivals. The number of packets in a message has a general distribution. This work is motivated by the large frame time in satellite TDMA systems. (The analysis, however, is applicable to a general TDMA system.) It is shown that the expected message delay is equal to the message delay (of a FDMA channel at the same data rate) given by the Pollaczek-Khinchin formula plus a term which vanishes to zero in the limit of zero frame time. The delay analysis is also generalized to allow a nonpreemptive priority queue discipline; expected message delay formulas are given.

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 study the delay performance of a TDMA channel used for transmitting packetized data messages. In a TDMA system, the transmission capacity is shared using Synchronous Time Division Multiplexing (STDM). STDM is widely used in teleprocessing computer systems for sharing point-to-point terrestrial lines [2,3]. Each terminal in such a system is assigned a fixed time slot within a time frame, buffering at the multiplexor is limited to one unit of data per terminal and addressing is not required. (See Fig. 1.) However, since the data stream generated by a single terminal is typically very "bursty" with long periods of inactivity, it is often inefficient to permanently assign channels to individual terminals. ATDM is proposed to remedy this inefficiency [2,3] the main attributes of ATDM are: (i) data units from all terminals are buffered on a common queue, and (ii) source address is required on every transmitted unit of data. (See Fig. 2.) The increase in channel utilization is obtained as a result of statistically averaging the random demands of all terminals attached to the multiplexor. Now consider Fig. 3 in which the point-to-point channel in Fig. 2 is replaced by a satellite channel. By attaching also the destination address(es) on each unit of data, it can be broadcasted to one or more geographically distributed destinations. This "downlink demultiplexing" capability provides further gains in channel utilization by statistically averaging in the same channel traffic loads to multiple destinations.

This paper is concerned with the delay analysis of a TDMA channel fixed assigned to an earth station such as shown in Fig. 3. (This work is motivated by the large frame time in satellite TDMA systems. The model and analysis, however, are applicable to a general TDMA system.) The aggregate message arrivals to the earth station is assumed to be an independent Poisson process. The number of packets comprising a message has a general distribution. We solve for the steady-state probability generating function of the earth station queue size. From this, an explicit formula for the expected message delay (total time in queue and transmission) is obtained. Next we consider a nonpreemptive priority discipline for the earth station queue and obtain expected message delay formulas for the priority classes.

An STDM model was previously considered by Chu and Konheim [2] as a special case of a unified model for a class of computer communication systems. 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.

THE ANALYSIS

Consider a TDMA system in which time is slotted and the time slots are organized into frames of M slots indexed from 1 to M. Time slots with the same index in consecutive frames form a TDMA channel. (See Fig. 4.) The amount of data that can be transmitted into a time slot may, for examples, be a packet, a byte, a bit, etc. depending on the specific system. For purposes of this paper, it will be referred to as a packet. Let the duration of a frame be T seconds. An earth 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. This is equivalent to a single-server queuing system in which the service time of a packet is exactly 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. This last statement is illustrated in the upper part of Fig. 5 in which S_i represents the service time (as defined above) and W_i the waiting time of the i^{th} message. Note that the idle period I is exponentially distributed under the assumption of Poisson arrivals. Y is equal to X minus T/M. The functional relationship between Y and I is shown in the bottom part of Fig. 5.

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 distribution $B(x)$ 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 [4]). If $\lambda b_1 < 1$, then

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

where

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

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 an earth 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 $\{g_\ell\}_{\ell=1}^{\infty}$ with first and second moments L_1 and L_2 where

$$g_\ell = \text{Prob}[L = \ell].$$

The message service time distributions are given by the following transforms

$$B^*(s) = \sum_{\ell=1}^{\infty} g_\ell\, [\beta^*(s)]^\ell$$

$$\hat{B}^*(s) = \hat{\beta}^*(s) \sum_{\ell=1}^{\infty} g_\ell\, [\beta^*(s)]^{\ell-1}$$

where

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

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

$$F(y) = \text{Prob}[Y \le y]$$

The lower part of Fig. 5 illustrates Y as a function of the idle period I which is exponentially distributed. F(y) is obtained to be

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

where

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

The first and second moments of Y are given by

$$\bar{Y} = (T - \frac{T}{M} - \frac{1}{\lambda}) + \frac{T\, e^{-\lambda(T - \frac{T}{M})}}{1 - e^{-\lambda T}} \qquad (3)$$

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

The first and second moments of X are thus given by

$$\bar{X} = \bar{Y} + \frac{T}{M} \qquad (5)$$

$$\overline{X^2} = \overline{Y^2} + 2\frac{T}{M}\bar{Y} + (\frac{T}{M})^2 \qquad (6)$$

The first and second moments of message service times are

$$b_1 = L_1 T \qquad (7)$$

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

$$b_2 = L_2 T^2 \qquad (9)$$

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

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

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

$$= \frac{(1 - \lambda L_1 T)\,(1 - e^{-\lambda T})}{\lambda T e^{-\lambda(T - T/M)}} \qquad (11)$$

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

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

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

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

The expected service time of a message is

$$\bar{S} = (1 - P_0)b_1 + P_0 \hat{b}_1$$

$$= b_1 - P_0(b_1 - \hat{b}_1)$$

$$= \frac{\hat{b}_1}{1 - \lambda(b_1 - \hat{b}_1)} \qquad (14)$$

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

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

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

$$\bar{N} = \lambda L_1 T - \frac{\lambda T}{2} + \frac{\lambda T}{M} + \frac{\lambda^2 L_2 T^2}{2(1 - \lambda L_1 T)} \qquad (16)$$

$$\bar{D} = L_1 T - \frac{T}{2} + \frac{T}{M} + \frac{\lambda L_2 T^2}{2(1 - \lambda L_1 T)} \qquad (17)$$

$$\bar{S} = \frac{1}{\lambda} + \frac{(L_1 T - \frac{1}{\lambda})(1 - e^{-\lambda T})}{\lambda T e^{-\lambda(T - T/M)}} \qquad (18)$$

$$\bar{W} = L_1 T - \frac{T}{2} + \frac{T}{M} + \frac{\lambda L_2 T^2}{2(1 - \lambda L_1 T)} - \frac{1}{\lambda} - \frac{(L_1 T - \frac{1}{\lambda})(1 - e^{-\lambda T})}{\lambda T e^{-\lambda(T - T/M)}} \qquad (19)$$

The expected message delay given in Eq. (17) is a very interesting result. Suppose the satellite transmission capacity is C bps. The transmission rate of a TDMA channel is thus $\frac{C}{M}$ bps. Now consider a FDMA channel which transmits continuously at $\frac{C}{M}$ bps. 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.

$$\bar{D} = \bar{D}_{PK} - \frac{T}{2} + \frac{T}{M} \qquad (20)$$

This last is an interesting result because of its simplicity. It is not obvious, however, because one might expect from Welch's model [4] 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. Eq. (20) is less surprising if we compare it to Skinner's result [6] and consider the possibility of a negative server-walking time.

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

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

NON-PREEMPTIVE 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 non-preemptive 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 is 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 $\{g_\ell^{(k)}\}_{\ell=1}^{\infty}$ 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 priority 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 $\sum_{i=1}^{K} \rho_i < 1$, then the expected waiting time of a class k message is

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

where

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

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

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

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.

Observation. 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

$$\bar{W}_\ell = \frac{V}{(1 - \lambda \sum_{i=1}^{\ell-1} i\, g_i\, T)(1 - \lambda \sum_{i=1}^{\ell} i\, g_i\, T)} \tag{26}$$

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

$$\bar{S}_\ell = P_0[(\ell-1)T + \bar{X}] + (1-P_0)\ell T$$

$$= \ell T - P_0(T - \bar{X}) \tag{28}$$

NUMERICAL EXAMPLES

In this section we examine the delay performance tradeoffs through some numerical examples. Let the satellite transmission rate be C bps. In Fig. 6, we consider a system with C = 1.5 Mbps, M = 450, T = 0.3 second, and 1000 bits/packet. Messages which arrive to the earth station under consideration consists of either single-packets or eight-packets with $g_1 = \alpha$ and $g_8 = 1 - \alpha$. Three cases are shown for α = 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. 6 shows typical delay versus traffic intensity curves with expected delay going to infinity as ρ approaches unity.* Fig. 6 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 $\frac{T}{2} - \frac{T}{M}$, the normalized difference decreases as ρ increases or as the average message length increases (α decreases).

In Fig. 7, 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 Mbps, 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. 7. As shown by Eq. (20), the slope of each curve is $-(\frac{1}{2} - \frac{1}{M})$. The average message delay increases to D_{PK} in the $T \to 0$ limit as predicted. The reader is cautioned not to interpret Fig. 7 as evidence that a large packet size is necessarily desirable for TDMA. The example in Fig. 7 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. 7 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. 8, 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 Mbps, M = 450, T=0.3 second, 1000 bits per packet and the message length distribution, $g_1 = 0.3$ and $g_\ell = 0.1$ for $\ell = 2,3,\ldots,8$. Note in Fig. 8 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 $\ell \leq 7$; for $\ell = 8$, the difference actually increases as ρ increases to 1 (conservation law! [5, Vol. II]). 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).

CONCLUSIONS

We have considered the delay performance of a satellite TDMA channel for transmitting packetized data messages. The steady-state probability generating function of the earth 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 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 is desirable since data traffic typically consists of different classes of messages with disparate delay constraints. An example of priority based upon message length is considered. We note that if the message length is limited to a single packet, our TDMA priority model becomes the same as some models for loop systems [2] and our results can be applied there.

We have considered in this paper systems in which the earth station under consideration (see Fig. 3) is of moderate size such that its combined data traffic output is sufficiently "smooth" to warrant efficient use of a dedicated TDMA channel. This is probably a realistic situation for current domestic satellite earth stations. If, in future, the trend is toward small earth stations characterized by a bursty data

* The satellite propogation delay of roughly 0.27 second is not included in our delay values.

traffic output, other satellite packet switching techniques such as slotted ALOHA [9-11], packet reservation [12] should be considered.

ACKNOWLEDGEMENT

The author would like to thank W-M. Chow, S. Lavenberg and L.S. Woo of the 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 Comm., Vol. COM-20, June 1972, pp. 645-660.

[3] Martin, J., Systems Analysis for Data Transmission, Prentice-Hall, Englewood Cliffs, N.J., 1972.

[4] 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.

[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] Skinner, C. E., "A Priority Queueing System with Server-Walking Time," Operations Research, Vol. 15, 1967, pp. 278-285.

[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] Abramson, N., "Packet Switching with Satellites," National Computer Conference, AFIPS Conference Proceedings, Vol. 42, 1973, pp. 695-702.

[10] Kleinrock, L. and S.S. Lam, "Packet-Switching in a Slotted Satellite Channel," National Computer Conference, AFIPS Conference Proceedings, Vol. 42, 1973, pp. 703-710.

[11] Kleinrock, L. and S.S. Lam, "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation," IEEE Trans. on Comm., Vol. COM-23, April 1975, pp. 410-423.

[12] Roberts, L.G., "Dynamic Allocation of Satellite Capacity through Packet Reservation," National Computer Conference, AFIPS Conference Proceedings, Vol. 42, 1973, pp. 711-716.

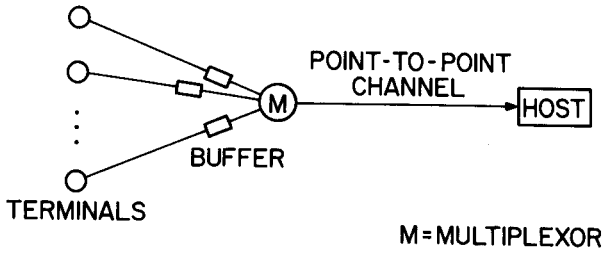

Figure 1 STDM System Configuration

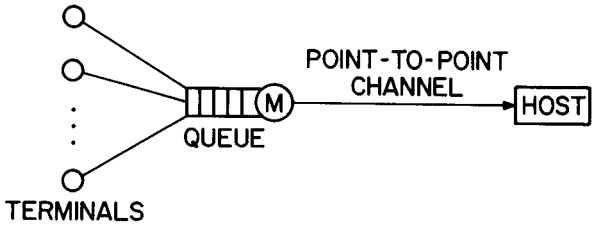

Figure 2 ATDM System Configuration

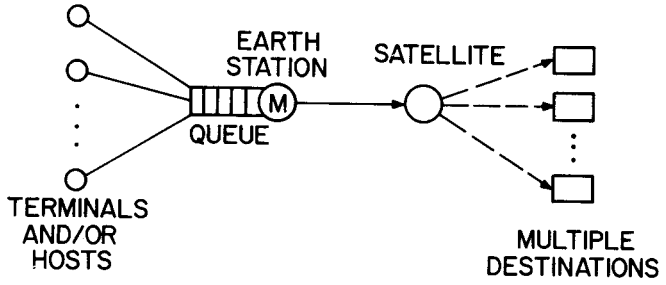

Figure 3 Satellite Packet Switching

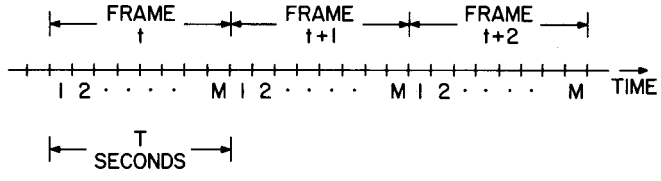

Figure 4 TDMA Channels

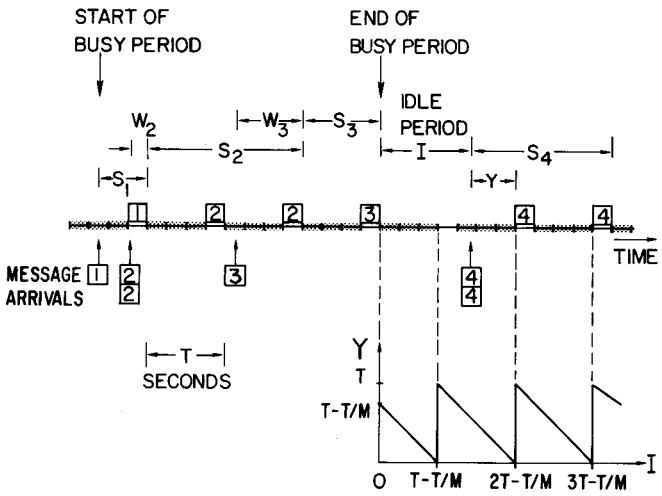

Figure 5 Illustration of Busy and Idle Periods, Waiting and Service Times

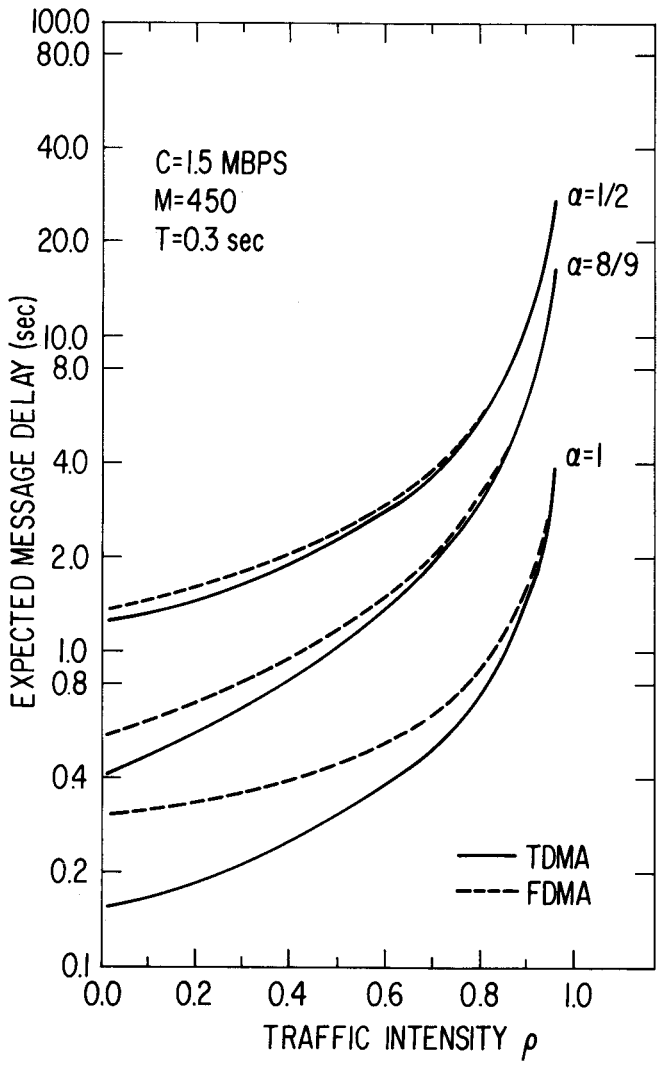

Figure 6 Expected Message Delay Versus Traffic Intensity

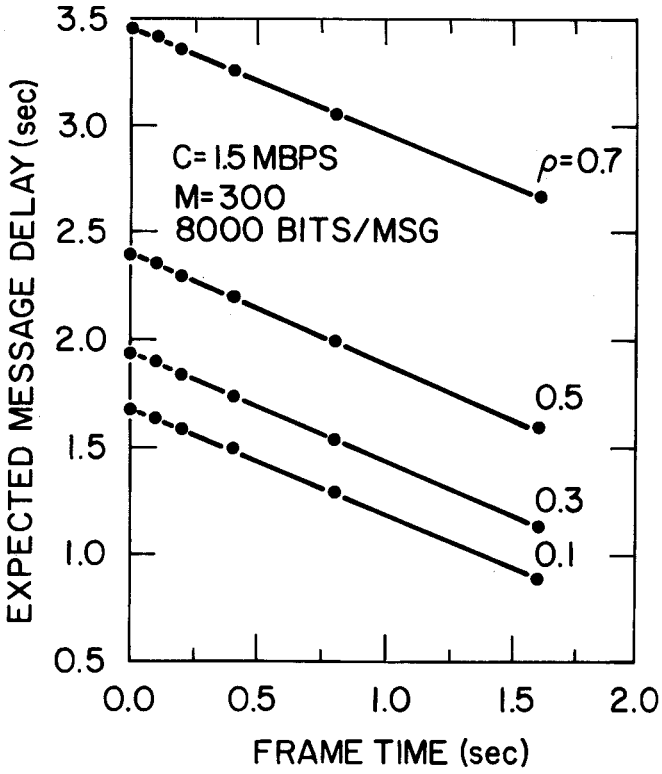

Figure 7 Expected Message Delay Versus Frame Time

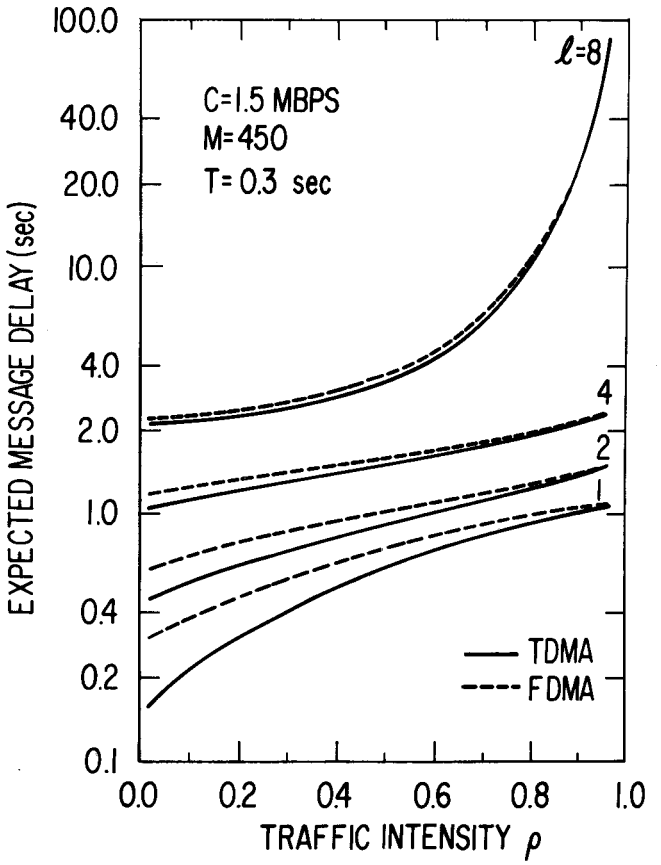

Figure 8 Expected Message Delay Versus Traffic Intensity for SMF Queue Discipline