DELAY MINIMIZATION FOR PACKET SATELLITE COMMUNICATION SYSTEMS

WONG, Wing-ming Eric

(黄永明)

A DISSERTATION SUBMITTED IN PARTIAL
FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF PHILOSOPHY

GRADUATE SCHOOL
THE CHINESE UNIVERSITY OF HONG KONG

HONG KONG

1990

ACKNOWLEDGMENTS

I would like to express my deepest appreciation and gratitude to my supervisor, Dr. T.S. Yum, for his continuous guidance, support and encouragement throughout the years of master studies. His valuable comments and suggestions made the completion of this thesis easier than it should be.

ABSTRACT

In this thesis, two very tight delay lower bounds are derived for packet satellite protocols with memoryless packet arrival process and single copy transmission. One bound is for protocols with contention-free reservation and the other is for protocols with contention-based reservation. The derivation indicates that for minimum delay, a protocol should strive to maintain a balance between transmitting packets immediately and making reservations before transmissions. Moreover, under the conditions of Poisson arrivals and single copy transmission, we designed a minimum delay protocol for packet satellite communications. The approach is to assume a hybrid random-access/reservation protocol, derive its average delay and minimize the delay with respect to all tunable system parameters. We found that for minimum average delay (1) a spare reservation should *normally but not always* be made for each packet transmission, (2) all unreserved slots should be filled with a packet rate of one per slot whenever possible, and (3) an optimum balance between transmitting packets and making reservations before transmission should be maintained.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Nowadays, communications satellites have carried the dominant portion of long distance communications [PRAT 86]. They handle most international telephone traffic, all international and almost all domestic long-distance television programs. The proportion of new domestic voice and data channels is also rapidly growing. Direct satellite broadcasting is coming soon, and electronic mail and personal two-way satellite radios have also been proposed.

At an altitude of about 36,000 km, the satellites, which act as a replay, can receive, amplify, and retransmit radio signals for most of a hemisphere. An earth station, through a satellite, can communicate with others distributed on nearly half of the world. With three satellites, one user can communicate with the other anywhere. Together with its broadcasting nature, the satellite is more suitable for long-distance television communication than other communication media.

1.1 Advantages and Disadvantages

There are a number of advantages in satellite communications:

1) no acknowledgement is needed for the protocols because of its broadcasting nature.
2) no routing problem
3) the size of the network can be increased by easily assigning more bandwidth rather than performing a complicated heuristic topology optimization.
4) mobile users can easily be accommodated.

It also has a number of disadvantages:

1) launches, satellites and antennas are expensive.

2) the performance is affected by the weather especially raining.

3) there is no privacy for each user since a satellite is a broadcast medium.

4) the technology is very difficult to be upgraded since the satellite is in the sky.

In particular, the main problem faced by packet satellite networks is the long round trip propagation delay. A lot of conventional channel allocation methods in local area network are no longer applicable (e.g. carrier sensing and polling).

1.2 Satellite System Engineering

From the view point of satellite system engineering, the design of a satellite communications system is a complicated process since it involves a lot of considerations.

The first consideration is the satellite itself. Since it is extremely expensive to put a kilogram into synchronous orbit, the satellite has to be made as small and lightweight as possible and consume a minimum of energy. Since launches and satellites are expensive and the maintenance is very difficult in the sky, it must be guaranteed that a satellite in the orbit will function without maintenance for many years and can stand for severe thermal cycling and constant bombardment of radiation and particles. Moreover, the development in communication technology is quick and unexpected but the components in the launched satellite is basically unexchangable. Therefore, the satellite should be designed to be as flexible as possible.

The other consideration is loss. Due to the distance of about 36,000 km between the satellite and earth stations, the inverse square losses are enormous and the rain losses are added at 10 GHz. On the uplink, large antenna with powerful transmitter can be used, although these are expensive and inconvenient. However, on the down link, the antenna size and transmitter power are extremely limited in the satellite. Therefore, much attention must be paid to antenna gain, transmitter efficiency, receiver noise figure, and the like. Due to the limited hardware on the satellite, a lot of effort is required in the software aspects for compensation. Specifically, much work goes into improving modulation and coding skills for detecting and correcting the transmission errors introduced by noise.

Multiple access is also an important problem in satellite communications. In the satellite system, a great population of users are scattered over a whole country or even an entire hemisphere and hence they are uncoordinated in topological nature. The traffic load in satellite systems is normally varied with the time. Therefore, some flexible and efficient channel allocation schemes are very desired for a large but changing number of independent users with varying traffic load.

1.3 Channel Allocation Methods

The main problem faced by packet satellite networks is the long round trip propagation delay. A lot of conventional channel allocation methods in local area network are no longer applicable such as carrier sensing and polling. A number of options have been proposed for packet satellite communications:

(i) Fixed Channel Assignment

A channel is divided into N equal portions where N is the number of users. If the partition is in time domain, it is called time-division multiplexing (TDM) or if in frequency domain, called frequency-division multiplexing (FDM). Since each user has its own transmission period, there is no interference between users. Therefore, this assignment is suitable for the users with regular traffic. However, in most computer system, data traffic is extremely bursty and hence most channels are idle for most of the time.

(ii) Random Access Assignment

In this assignment, a population of users will simultaneously content a channel. The most famous and simple one is the Aloha scheme in which the users just transmit whenever they have data to be sent. For the bursty traffic condition, random access assignment is more flexible and efficient than fixed channel assignment. However, the delay in this scheme is unbounded. Moreover, the system will rapidly downgrade when the traffic become heavy and a lot of packets get collided with each other.

(iii) Reservation Channel Assignment

When a user has data to transmit, he reserves in advance. If his reservation is accepted, then he can transmit at the prescribed time. This demanded-type scheme can give the maximum channel throughput close to one and has better system stability than the random access scheme. However, at least one round trip propagation delay (270 msec) is needed for each user to exchange the reservation information with the satellite. This is a great delay overhead, especially for real time applications.

(iv) Hybrid Random Access/Reservation Channel Assignment

This hybrid scheme works alike the random access scheme under light traffic conditions while alike the reservation scheme under heavy traffic condition. It combines the advantages of the random access channel assignment with low delay under light traffic condition and the reservation channel assignment with high maximum throughput. However, this scheme is relatively more complicated.

1.4 Outline of this Thesis

In chapter 2, two very tight delay lower bounds are derived for packet satellite protocols with memoryless packet arrival process and single copy transmission. One bound is for protocols with contention-free reservation and the other is for protocols with contention-based reservation. The derivation indicates that for minimum delay, a protocol should strive to maintain a balance between transmitting packets immediately and making reservations before transmissions.

In chapter 3, under the conditions of Poisson arrivals and single copy transmission, we designed a minimum delay protocol for packet satellite communications. The approach is to assume a hybrid random-access/reservation protocol, derive its average delay and minimize the delay with respect to all tunable system parameters. We found that for minimum average delay,

1) a spare reservation should *normally but not always* be made for each packet transmission,
2) all unreserved slots should be filled with a packet rate of one per slot whenever possible,
3) an optimum balance between transmitting packets and making reservations before transmission should be maintained.

CHAPTER 2

DELAY BOUNDS

2.1 Introduction

In multiaccess communication systems, the average packet delay is bounded below by the G/G/1 queuing delay with the same interarrival and service time distributions. This delay bound is very loose for packet satellite systems where the round trip propagation delay is long and carrier sensing is not possible. A tighter delay bound is desirable for assessing the possible delay improvement on existing protocols and for deciding whether a particular delay requirement can ever be satisfied.

In this chapter, two new delay lower bounds are derived for packet satellite systems with contention-free and contention-based reservations respectively. The class of protocols whose delays we are trying to bound is of the hybrid random-access/reservation type. This class of protocols includes random access protocols and reservation protocols as special cases and is sufficiently general to be of interest. The environment in which the protocols are to operate is defined by a set of conditions. We shall call this environment ξ and the delay bounds are for the protocols operating in ξ. The conditions defining ξ are:

1) *The packet arrival process is of the memoryless type.* For a finite population model this refers to the Bernoulli process and for infinite population model, Poisson.

2) *Transmitting multiple copies of the same packet and making multiple reservations for the same packet are not allowed.* Transmitting multiple copies and making multiple reservation might give slightly smaller delay when the traffic is light. Since we have not done any investigation on this, we shall not consider this option.

3) *A single uplink channel is considered.* This condition is not really restrictive because multiple channel systems involve three kinds of inefficiencies:

   a) the overhead in partitioning a channel into several TDM or FDM subchannels,

   b) longer transmission time on lower bit rate subchannels,

   c) multiple reservation queues on the satellite give a longer average delay than a single reservation queue.

4) *Only the slotted channel is considered.* The unslotted channel gives slightly better delay performance only at very very low traffic conditions.

In the following, we will describe the packet satellite system and design an idealized protocol for deriving the delay lower bounds.

2.2 The Packet Satellite System

Consider a packet satellite system serving a population of users. Besides the uplink data channel, let there also be an uplink narrow-band control channel for making reservations and a downlink announcement channel for broadcasting successful reservations. In practice, the control channel and the announcement channel can be piggybacked on the up- and down-link data channels respectively. The data channel is slotted with slot width equal to one packet transmission time. There are two types of slots. Aloha slots are for transmitting packets immediately whereas Reserved slots are for packets with successful reservations. The announcement channel broadcasts the locations of the Reserved slots so that other stations will refrain from transmitting on these slots. All non-Reserved slots are treated as Aloha slots.

2.3 The Idealized Protocol with Contention-Free Reservation

Many protocols were proposed for the above system and an extensive survey can be found in [CHIT 88]. To obtain a delay lower bound for all possible protocols in ξ, we hypothesize an idealized protocol by assuming

1) contention-free reservation,
2) no reservation overflow in the reservation queue,
3) an optimal balance of the packet traffic rate and the reservation traffic rate in the system,
4) the traffic statistics after the balancing process is memoryless.

These idealized assumptions guarantee that no practical protocols of the hybrid random-access/reservation type will have a smaller delay than the idealized protocol. The delay of this idealized protocol is therefore a delay lower bound for all practical protocols of the hybrid random-access/reservation type in ξ.

Consider the arrival of a new packet. If it hits an Aloha slot, it will either make a reservation on the control channel for future transmission or be transmitted in the current Aloha slot with a spare reservation made on the control channel. This spare reservation assures that, in case of a collision in the Aloha slot, the retransmission is always successful. If the transmission is successful, the spare reservation is discarded. On the other hand, if the arriving packet hits a Reserved slot, it will either make a reservation right away or be transmitted in one of the future Aloha slots.

In a practical protocol, some form of strategy is needed to optimally balance the random-access traffic and the reservation traffic. Since the idealized protocol is used for deriving a delay lower bound, it need not be realizable. An optimal traffic balancing strategy can therefore be assumed as built-in.

All reservations are processed by the satellite and for each successful reservation a Reserved slot is assigned on the uplink data channel. Since all reservations are assumed to be successful, a packet will encounter at most one collision before successful transmission. A flow chart summarizing this protocol is shown in Fig. 2.1.

In the next section, we shall derive the delay of the idealized protocol assuming a finite population model. A similar bound for infinite population model can be obtained either by letting the population size N go to infinity or by starting from the Poisson arrival model. These bounds turn out to be expressible in closed forms. To tighten these bounds, we relax the assumption of contention-free reservation. The resulting delay lower bound for the protocols with contention-based reservation is derived in section 2.4.

2.4 Delay Lower Bound for Protocols with Contention-Free Reservation

Let there be N users in the system. Let λ_a be the average number of transmissions in an Aloha slot and λ_r be the average number of transmission reservations (i.e. excluding the spare reservations) per slot on the control channel. Let the average number of successful reservations per slot be x. Since each successful reservation is assigned a Reserved slot, x is the average number of packets transmitted through reservation per slot. This also means that x is equal to the probability that a slot is of the reserved type. With the assumption that all reservations are successful, x is derived as:

$$x \equiv [\text{av. no. of successful reservations per slot}]$$

$$= \begin{bmatrix} \text{av. no. of tx'n} \\ \text{reservations} \\ \text{per slot} \end{bmatrix} + \begin{bmatrix} \text{av. no. of remaining} \\ \text{spare reservations} \\ \text{from an Aloha slot} \end{bmatrix} \Pr \begin{bmatrix} \text{a slot is} \\ \text{of the} \\ \text{Aloha type} \end{bmatrix}$$

$$= \lambda_r + \left[\lambda_a + \lambda_a \left(1 - \frac{\lambda_a}{N} \right)^{N-1} \right] (1 - x) \tag{2.1}$$

where $\lambda_a \left(1 - \frac{\lambda_a}{N}\right)^{N-1}$ is the average number of successful transmissions in an Aloha slot.

The throughput S of the idealized protocol is given by

$$S = \Pr \begin{bmatrix} \text{a slot is} \\ \text{of the} \\ \text{res. type} \end{bmatrix} \Pr \begin{bmatrix} \text{a res. slot} \\ \text{contains a} \\ \text{succ. tx,n} \end{bmatrix} + \Pr \begin{bmatrix} \text{a slot is} \\ \text{of the} \\ \text{Aloha type} \end{bmatrix} \Pr \begin{bmatrix} \text{an Aloha slot} \\ \text{contains a} \\ \text{succ. tx'n} \end{bmatrix}$$

$$= x + (1 - x) \lambda_a \left(1 - \frac{\lambda_a}{N} \right)^{N-1}. \tag{2.2}$$

Solving x from (2.1) and substituting into (2.2), we have

$$S = \frac{\lambda_r \left[1 - \lambda_a \left(1 - \frac{\lambda_a}{N} \right)^{N-1} \right] + \lambda_a}{1 + \lambda_a - \lambda_a \left(1 - \frac{\lambda_a}{N} \right)^{N-1}}. \tag{2.3}$$

By differentiating (2.3), we observed two properties:

Property 1: S is a monotonically increasing function of λ_a and λ_r.

Property 2: For a given S, λ_a and λ_r are inversely related functions.

The average delay D of the idealized protocol consists of five terms denoted as D_1 to D_5. The average synchronization delay D_1 is equal to 0.5 slot. The expected reservation delay

D_2 is equal to the round trip propagation delay R (in unit of slots) multiplied by the probability of transmission through reservation or $D_2=(x/S)R$. The average waiting time in the reservation queue formed by the reservation traffic, denoted by D_3, is given by the average delay of a discrete-time $M/D/1$ queue with a composite Bernoulli arrival process of rate x. From the Pollaczek-Khinchin mean value formula [KLEI 75a], we have

$$D_3 = \frac{x(1-N^{-1})}{2(1-x).}$$

The combined packet transmission and propagation time D_4 is equal to $(1+R)$. The average delay of traffic diversion from the Reserved slots to the Aloha slots is denoted as D_5. Adding up the five terms, we have

$$D = \frac{x(1-N^{-1})}{2(1-x)} + R\left(1+\frac{x}{S}\right) + 1.5 + D_5. \tag{2.4}$$

For the idealized protocol, parameter x in (2.4) should be chosen such that D is minimum. However, as D_5 involves the specification of the traffic diversion process and is in general much smaller than the round trip propagation delay R, we shall neglect D_5 in the optimization process. In doing so, the delay obtained is only a lower bound for the idealized protocol. This bound is obviously also a lower bound for all protocols in ξ. Let

$$D_L = \frac{x(1-N^{-1})}{2(1-x)} + R\left(1+\frac{x}{S}\right) + 1.5. \tag{2.5}$$

To minimize D_L for a given value of S, (2.5) stipulates that x should be as small as possible. From (2.2), x can be expressed as

$$x = 1 - \frac{1-S}{1-\lambda_a\left(1-\frac{\lambda_a}{N}\right)^{N-1}}.$$

Differentiating x with respect to λ_a, x is found to have a single minimum at λ_a=1. But λ_a and λ_r must also satisfy (2.3). Therefore, substituting λ_a=1 into (2.3) and solving for λ_r, we obtain

$$\lambda_r = \frac{S[2-(1-N^{-1})^{N-1}]-1}{1-(1-N^{-1})^{N-1}}. \tag{2.6}$$

Since λ_r must be non-negative, this means that for the above "λ_a=1" solution to be valid,

$$S \geq \frac{1}{2-(1-N^{-1})^{N-1}} \equiv S_c. \tag{2.7}$$

At the boundary point $S=S_c$, we have λ_a=1 and λ_r=0. For $S<S_c$, the $\lambda_r \geq 0$ constraint is binding. Therefore, we set λ_r=0 in (2.3) to obtain

$$S = \frac{\lambda_a}{1+\lambda_a-\lambda_a\left(1-\frac{\lambda_a}{N}\right)^{N-1}} \tag{2.8}$$

and from which the constrained optimum value of λ_a, denoted as λ_a^* can be solved numerically. Using Property 1 and in comparison with the $S=S_c$ case, λ_a^* can be shown to be always less than one. The above "λ_r=0" solution is indeed optimum since Property 2 states that if λ_r is increased, λ_a will be decreased resulting in the increase of x. Substituting the optimum λ_a and λ_r into (2.4), we obtain the delay lower bound $D_L(S, R, N)$ of the idealized protocol as

$$D_L(S,R,N) = \begin{cases} \left[S-\left(1-\frac{\lambda_a^*}{N}\right)^{N-1}\right]\frac{(1-N^{-1})}{2(1-S)} + R\frac{2S-(1+S)\lambda_a^*\left(1-\frac{\lambda_a^*}{N}\right)^{N-1}}{S\left[1-\lambda_a^*\left(1-\frac{\lambda_a^*}{N}\right)^{N-1}\right]} + 1.5 & S < S_c \\ \frac{S(1-N^{-1})-(1-N^{-1})^N}{2(1-S)} + R\frac{2S-(1+S)(1-N^{-1})^{N-1}}{S[1-(1-N^{-1})^{N-1}]} + 1.5 & S \geq S_c. \end{cases} \tag{2.9}$$

It can be shown that

$$D_L(S,R,N) < D_L(S,R,N+1) \qquad N = 1,2,\ldots$$

In the limit $N \to \infty$, (2.9) becomes

$$D_L(S,R,\infty) = \begin{cases} \dfrac{S - e^{-\lambda_a^*}}{2(1-S)} + R\dfrac{2S - (1+S)\lambda_a^* e^{-\lambda_a^*}}{S\left[1 - \lambda_a^* e^{-\lambda_a^*}\right]} + 1.5 & S < \dfrac{e}{2e-1} \\ \dfrac{S - e^{-1}}{2(1-S)} + R\dfrac{2S - (1+S)e^{-1}}{S(1-e^{-1})} + 1.5 & S \geq \dfrac{e}{2e-1}. \end{cases} \qquad (2.10)$$

which can be independently derived by assuming a Poisson arrival process.

The control channel may be regarded as a pure overhead because it is not used for transmitting data packets. For protocols with control channels consuming a fixed ratio w of the total bandwidth, the effective throughput S becomes

$$S\,|_{\text{with overhead}} = (1-w)S\,|_{\text{without overhead}}\,.$$

Fig. 2.2 shows the average delay of the UCA protocol [LEE 83] with contention-free reservation, the average delay of the C-MA (Controlled Multiaccess) protocol [WONG 88] (20 minislots per slot and a maximum of 10 reservations in the reservation queue) and the delay lower bound. Poisson arrival process and zero control channel overhead are assumed in all three cases. We see that both UCA and C-MA protocols have very good delay performance because at most 5% delay reduction can be hoped for. As both UCA and C-MA are not the minimum delay protocol, the difference between the lower bound and the delay of the unknown minimum delay protocol is less than 5% for R=100. Fig. 2.2 also shows that for R large, the $M/D/1$ bound is too loose to be of any use.

2.5 Delay Lower Bound for Protocols with Contention-Based Reservation

In section 2.4, we derived the delay lower bound assuming a contention-free control channel. Here, we relax this assumption by choosing the control channel to be of the slotted Aloha type. Let the control channel be divided into minislots and let there be M minislots to a slot. Let the arrival of input packets be a Poisson process. As before, we first design an idealized protocol and derive its average delay. This delay is therefore a lower bound for all hybrid protocols with contention-based reservations in ξ.

For the idealized protocol under contention-based reservation, we made three more assumptions in addition to assumptions 2, 3 and 4 in section 2.3. First, we assume that all packets which are successfully transmitted in Aloha slots did not make any spare reservations. This "noncausal" assumption guarantees that there is no spare reservation from successful packets to interfere with the other reservations and hence a smaller delay will result. Second, we assume that all collided packets have made spare reservations because doing so will provide an extra chance of obtaining a Reserved slot for retransmission. When a reservation collides with the other reservations, the stations concerned will reattempt the channel after a random delay. Third, we assume that the combined arrival of normal and spare reservations to the control channel is given by a Poisson process. This is an idealized assumption because packets collided on the Aloha slots will tend to have their spare reservations aggregated together on the control channel. Assuming these reservations to be uncorrelated and modeling them as Poisson arrivals will give an underestimated delay. But for obtaining a delay lower bound this is acceptable.

Let the combined arrival rate of normal and spare reservations to the control channel be λ_m per *minislot* where

$$\lambda_m = \frac{\lambda_r + \left(\lambda_a - \lambda_a e^{-\lambda_a}\right)(1-x)}{M} \tag{2.11}$$

Here as before x is the average number of successful reservations per slot and is given by

$$\begin{aligned} x &= M[\text{av. no. of successful reservations in a minislot}] \\ &= M\lambda_m e^{-\lambda_m}. \end{aligned} \tag{2.12}$$

As a check, by setting $M \to \infty$ (2.12) degenerates to the case of no reservation contention. Substituting x from (2.12) into (2.2) and letting $N \to \infty$, we obtain the throughput S of the idealized protocol as

$$S = M\lambda_m e^{-\lambda_m} + \left(1 - M\lambda_m e^{-\lambda_m}\right)\lambda_a e^{-\lambda_a}. \tag{2.13}$$

By differentiating (2.13) with respect to λ_m and λ_a, and noting that $x<1$, we obtain:

Property 3: For a given value of S, λ_m and λ_a are inversely related.

Substituting (2.12) into (2.11) and solving for λ_r, we have

$$\lambda_r = M\lambda_m - \left(1 - M\lambda_m e^{-\lambda_m}\right)\left(\lambda_a - \lambda_a e^{-\lambda_a}\right). \tag{2.14}$$

By differentiating λ_r with respect to λ_m and λ_a, we obtain

Property 4: λ_r is a monotonically increasing function of λ_m for a fixed λ_a.

Property 5: λ_r is a monotonically decreasing function of λ_a for a fixed λ_m.

Property 3 states that for a fixed S, λ_a will decrease when we increase λ_m. But the decrease of λ_a causes an increase of λ_r according to Property 5. Also, Property 4 states that increasing λ_m causes a corresponding increase of λ_r. Therefore, we conclude:

Property 6: λ_r is a monotonically increasing function of λ_m for a given S.

The average packet delay consists of seven terms denoted as D_1 to D_7. D_1, D_2, D_4 and D_5 are the same as that in section 2.3. D_3 is the mean waiting time in the satellite reservation queue and is given by the waiting time on a discrete-time $M/D/1$ queue with the distribution of the number of arrivals per slot U given by

$$\Pr[U=k] = \binom{M}{k}\left(\frac{x}{M}\right)^k\left(\frac{M-x}{M}\right)^{M-k}$$

This queueing system is exactly the same as that analyzed in section 2.3. Therefore, we have

$$D_3 = \frac{x(1-M^{-1})}{2(1-x)}$$

The additional propagation delay due to retransmissions D_6 is given by

$$D_6 = R\left[\frac{\lambda_r + \lambda_a(1-x)}{S} - 1\right]. \tag{2.15}$$

where $[\bullet]$ is the expected number of retransmissions. Eliminating λ_r and λ_a with the use of (2.2) and (2.11), we have

$$D_6 = RS^{-1}(M\lambda_m - x)$$

Neglecting D_5 and the randomization delay for retransmission D_7 for the delay lower bound of the idealized protocol, we obtain

$$D_L = 0.5 + \frac{x(1-M^{-1})}{2(1-x)} + \frac{xR}{S} + 1 + R + \frac{M\lambda_m - x}{S}R$$

$$= \frac{M-1}{2\lambda_m^{-1}e^{\lambda_m} - 2M} + R\left(\frac{M\lambda_m}{S} + 1\right) + 1.5. \tag{2.16}$$

where λ_m is to be chosen for minimum D_L. As a check, by setting $M \rightarrow \infty$ (2.16) degenerates to the contention-free case. To minimize (2.16) for a given S, λ_m should be as small as possible. To minimize λ_m, Property 3 states that λ_a should be as close to one as possible. Following the approach in section 2.3 and using (2.13) and (2.14), we obtain

$$S_c \equiv S \mid_{\lambda_a = 1, \lambda_r = 0}.$$

For $S \geq S_c$, we choose λ_a=1 and solve for λ_m from (2.13). For $S<S_c$, we choose λ_r=0 and solve for λ_a and λ_m simultaneously from (2.13) and (2.14). This choice of λ_r=0 results in minimum delay because any other choice of λ_r will cause an increase of λ_m by Property 6. Substituting the computed λ_m into (2.16), a delay lower bound for the idealized protocol can be explicitly evaluated. This lower bound is also a bound for all protocols with contention-based reservation operating in environment ξ as defined in section 2.1.

Fig. 2.3 shows the average delays of the UCA and C-MA protocols. Here, with only 3 minislots per slot, the contention-based reservation bound is much tighter than that for the contention-free reservation.

Fig. 2.4 shows the delay lower bounds for M=3, 5, 10 and ∞. We see that the bounds are very close for $S \leq 0.5$. We also notice that for $M \geq 5$, the bound for contention-free reservation (i.e. $M=\infty$) is a good approximation to that for contention-based reservation.

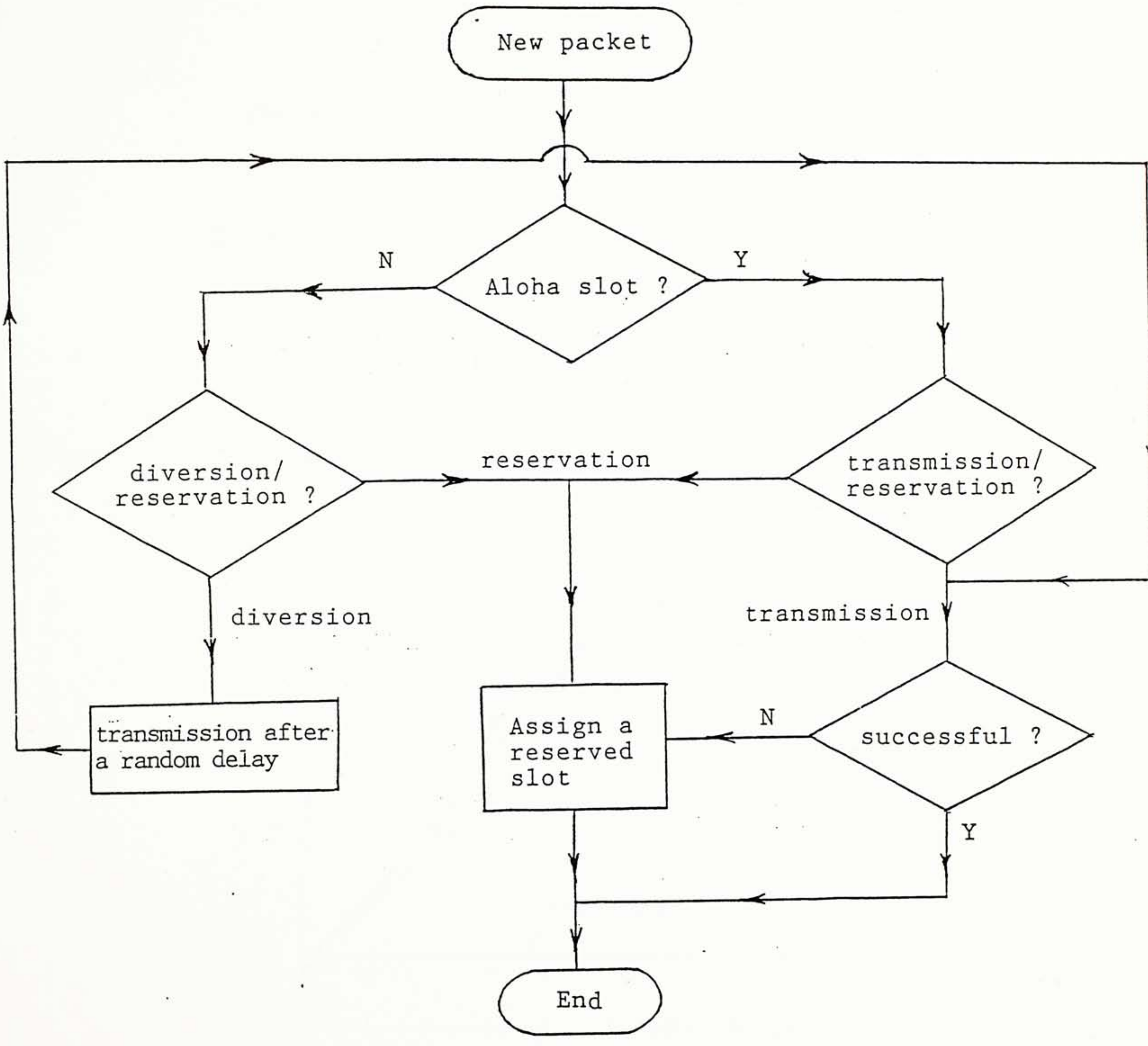


Fig.2.1. Flow chart of the Idealized protocol with contention-free reservation.

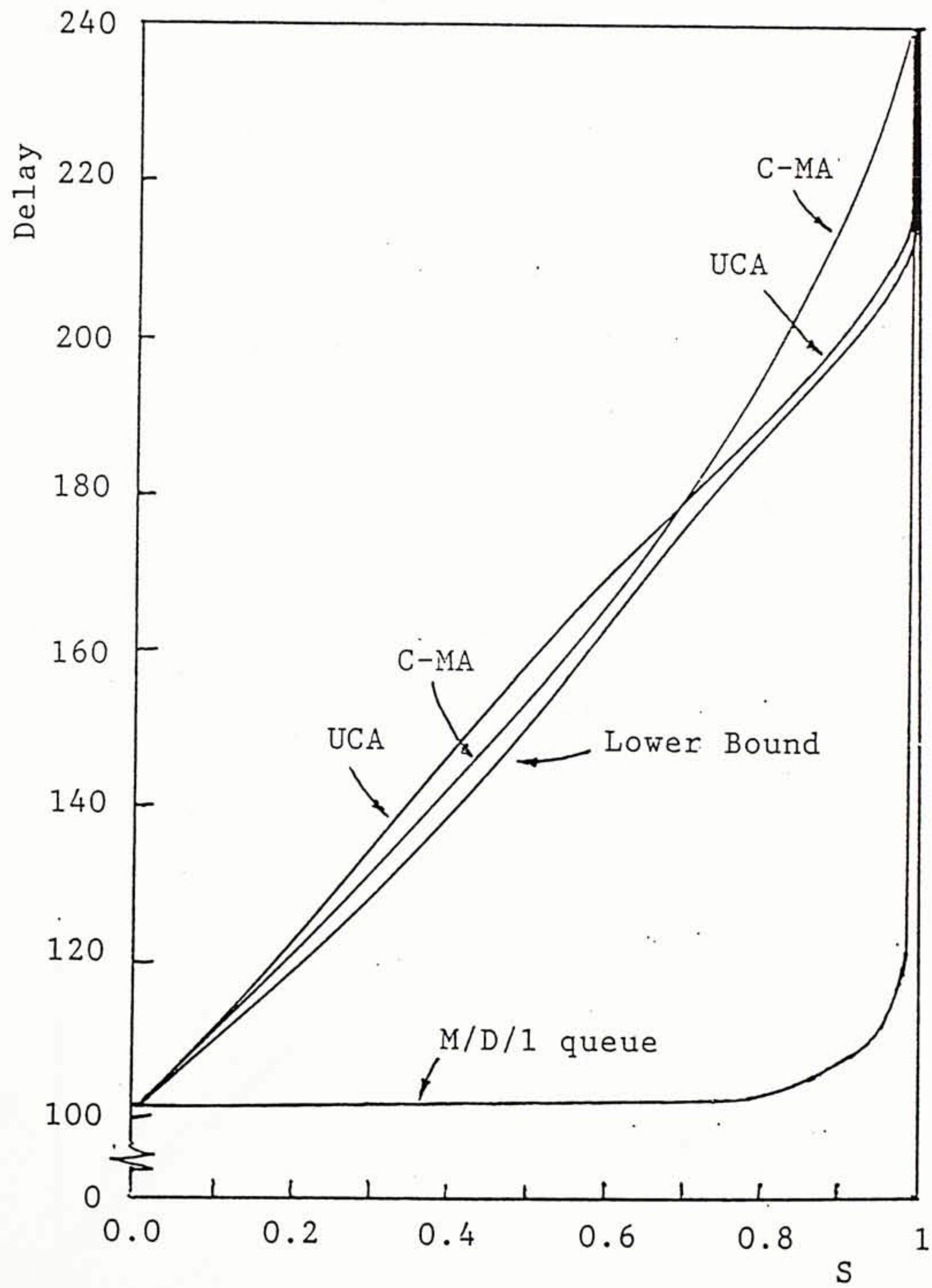


Fig.2.2.Delay and Delay Bounds for R=100.

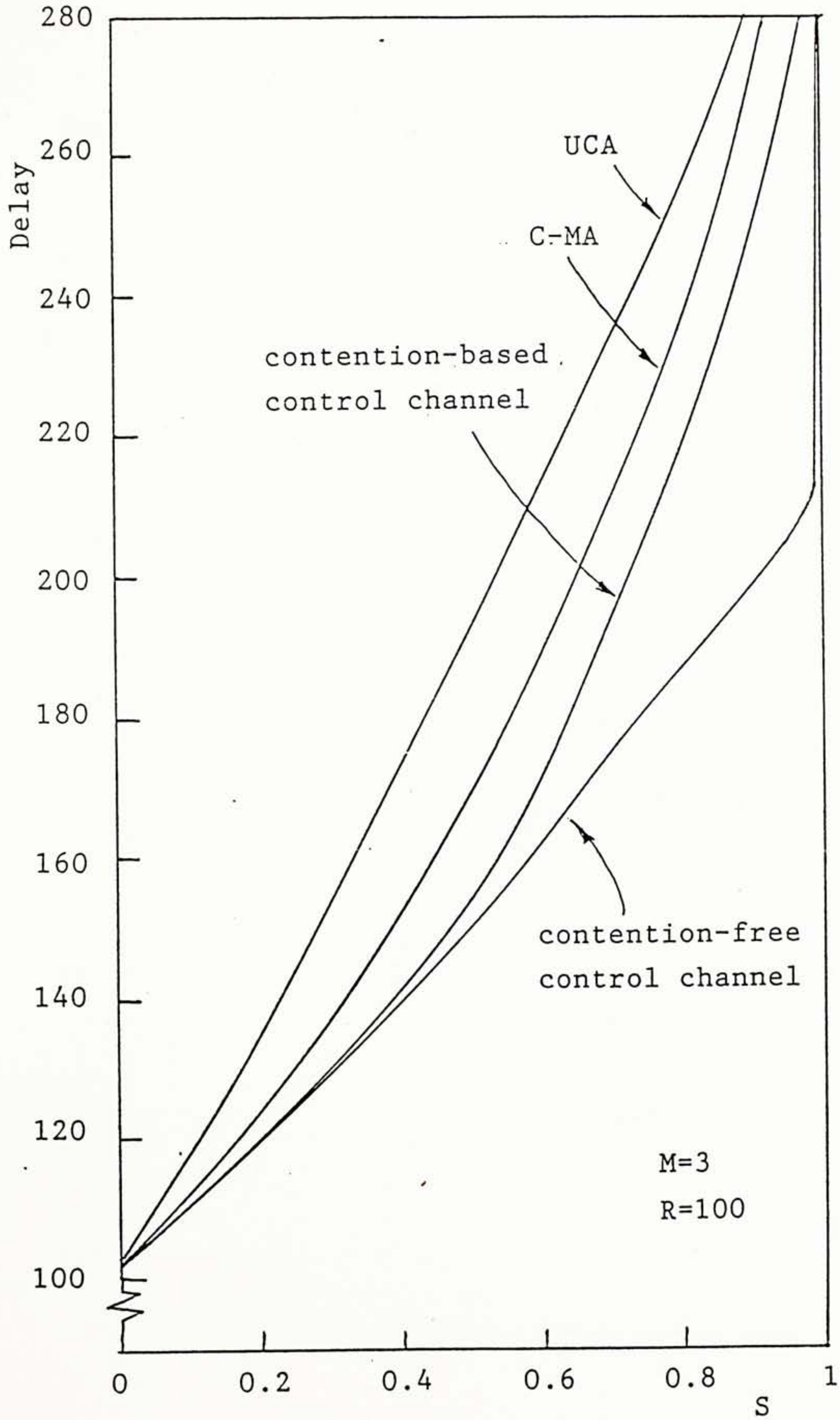


Fig.2.3.Comparison of delay bounds for contention-based and contention-free control channels.

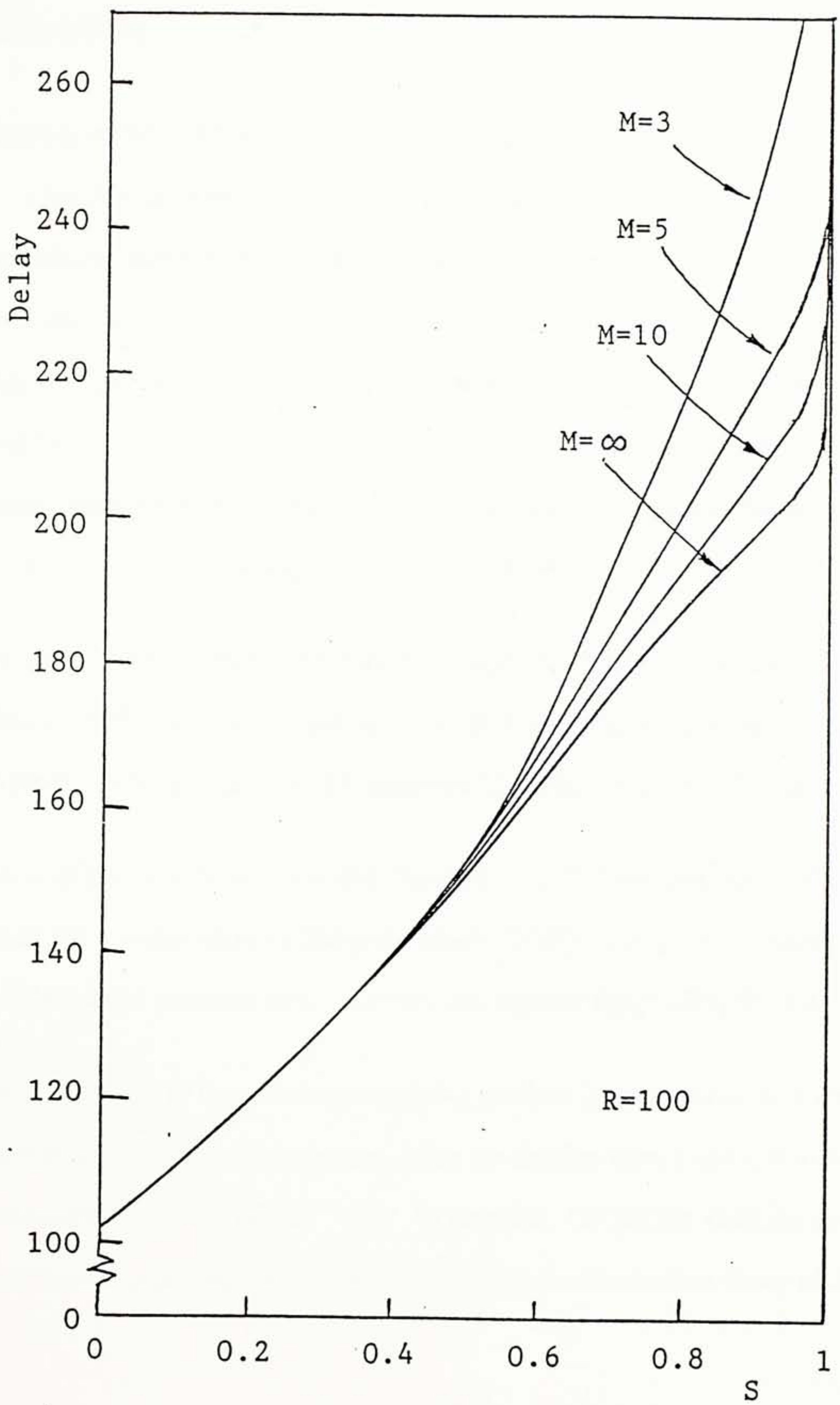


Fig.2.4 Delay bounds for various M values.

CHAPTER 3

IN SEARCH OF THE MINIMUM DELAY PROTOCOL

3.1 Introduction

Multiaccess protocols for packet satellite systems usually take on one of the following three types: 1) random-access, 2) reservation and 3) hybrid random-access/reservation. Types 2 and 3 protocols are inherently more complicated than type 1 because extra processing, either on-board or at each earth station, is required. Type 3 is a synthesis of type 1 and type 2, taking the advantages of the low delay property of type 1 and the high throughput property of type 2. Because of that, type 1 and type 2 can also be considered as special cases of type 3 protocols. In recent years, there have been constant efforts to design better and better type 3 protocols [BOSE 80], [CHAN 84], [LEE 83], [YUM 87], and [WONG 88].

In this chapter, we attempt to find the minimum delay protocol under a set of conditions. These conditions define the environment of the protocol and the protocol is optimal only in this environment. We shall call this environment ξ. The conditions defining ξ are:

1) *The arrival of packets to the satellite channel is a Poisson process.* We would like to caution that for a population sufficiently small, TDMA can give a smaller delay than the best possible hybrid protocol over a certain throughput range [WONG 89].

2) *The combined arrival of new and reattempting packets is assumed to be a Poisson process.* For mean retransmission randomization delay no smaller than 5 slots, it was found that the above assumption is valid [KLEI 75b]. In practice, for packet satellite systems inherent with long round trip propagation delay, an average randomization delay of 5 slots or more

is also desirable to uncorrelate the retransmission of collided packets. This uncorrelation process is vital since one more collision means a penalty of one more round trip propagation delay.

3) *Transmitting multiple copies of the same packet and making multiple reservations for the same packet are not allowed.* We suspect that transmitting multiple copies and making multiple reservations might lead to a slight reduction of the overall delay under certain throughput range. But since we have not done any investigation on this, we shall not consider this option.

4) *Only a single uplink channel is considered.* This condition is really not restrictive because multiple channel systems involve three kinds of inefficiencies:

   i) additional overhead in partitioning a channel into several TDM or FDM subchannels,
   ii) longer transmission time on lower bit rate subchannels,
   iii) longer average delay on multiple reservation queues on the satellite.

5) *Only the slotted channel is considered.* The unslotted channel gives slightly better delay performance only at very very low traffic conditions.

6) A control channel is used for transmitting reservation information. We assume *the bandwidth occupied by the control channel is a fixed percentage of the total bandwidth.* In [LEE 83], a scheme was proposed that allows the dynamic sharing of control and data channel bandwidths. Such a scheme, although elegant, was also reported to be more complicated with only a slight improvement of delay performance when the number of minislots per slot is more than 4.

Under the above conditions, there are still a number of options in the design of protocols. We attempt to isolate all the available options and minimize the average packet delay with respect to these options. The resulting protocol is then the minimum delay protocol in ξ. What are the remaining options under the above conditions? Obviously, a station with a packet can choose to transmit immediately, to make a reservation immediately, to make a spare reservation immediately with packet transmission, or to defer transmission until a later time. The optimal choice should depend on the channel state and the channel loading condition.

In the following, we shall first describe the packet satellite system. We then design the protocol to be optimized and derive its throughput and delay characteristics. Finally, we minimize the delay analytically with respect to all tunable parameters to obtain the minimum delay protocol in ξ as well as the set of conditions for maintaining minimum delay.

3.2 The Packet Satellite System

Consider a packet satellite system. Besides the uplink data channel used for transmitting packets, let there also be an uplink narrow-band control channel for making reservation and a downlink announcement channel for broadcasting successful reservation. In practice, the control channel and the announcement channel can be subchannels on the up- and the down-link data channels respectively. The data channel is slotted with slot size equal to one packet transmission time. The control channel is divided into minislots with M (need not be an integer) minislots per slot. Let there be two types of slots. The *Aloha slots* are for transmitting packets without prior reservations whereas the *Reserved slots* are for transmitting packets with successful reservations. The control channel serves two purposes:

1) to make reservations for *transmissions* on the data channel *and*

2) to make spare reservations for *retransmissions* in case the transmissions in Aloha slots fail. The announcement channel is used to broadcast the locations of the Reserved slots to all stations. All non-Reserved slots are treated as Aloha slots.

3.3 The Transmission Protocol

Consider the arrival of a packet. If it hits an Aloha slot, it will either, with the probability f_1, make a *reservation* on the control channel and await its assigned Reserved slot, or with the remaining probability $1-f_1$, be transmitted in the current Aloha slot. In the latter case, the packet can, with probability α, make a *spare reservation* on the control channel. In case of a collision in the Aloha slot, this spare reservation, if successful, allows the packet to be transmitted in a Reserved slot after a round trip propagation delay (RTPD). If the transmission on the Aloha slot is successful, its spare reservation, if made, is ignored by the satellite. When a station wants to make a reservation or a spare reservation, it does so by marking its identity randomly on one of the K subsequent minislots.

If the arrival packet hits a Reserved slot, it will either, with probability f_2, make a reservation immediately or, with the remaining probability $1-f_2$, be transmitted randomly on one of the I up-coming Aloha slots. In the latter case a spare reservation will also be made with probability α. For each successful reservation, a Reserved slot on the uplink data channel is assigned. Packets with unsuccessful transmission or unsuccessful reservation (including spare reservation) will reattempt the system on one of the J subsequent slots.

A flow chart summarizing this protocol is shown in Fig. 3.1.

3.4 Throughput Analysis

Let λ_a be the average number of transmissions in an Aloha slot and λ_r be the average number of *ordinary reservations* per slot on the control channel. Due to random bifurcation and merging of Poisson processes, the combined arrivals of ordinary and spare reservations to the control channel is also a Poisson process with *per minislot* rate of

$$\lambda_m = \frac{\lambda_r + \alpha\lambda_a(1-x)}{M} \tag{3.1}$$

where x be the probability that a slot is of the reserved type. To find x, note that all successful reservations (to be quantified) are assigned a Reserved slot each. Hence, the average number of successful reservations per slot is equal to the average number of packets transmitted through reservation per slot, which in turn is equal to x. Mathematically,

$$\begin{aligned}
x &= [\text{ av. no. of successful reservation per slot}] \\
&= \begin{bmatrix}\text{av. no. of uncollided} \\ \text{reservations} \\ \text{in } M \text{ minislots}\end{bmatrix} - \Pr\begin{bmatrix}\text{a slot is} \\ \text{of the} \\ \text{Aloha type}\end{bmatrix}\begin{bmatrix}\text{av. no. of spare res'ns} \\ \text{to be ignored} \\ \text{in an Aloha slot}\end{bmatrix} \\
&= M\lambda_m e^{-\lambda_m} - (1-x)\Pr\begin{bmatrix}\text{a packet is} \\ \text{succ. tx'ed in} \\ \text{an Aloha slot}\end{bmatrix}\Pr\begin{bmatrix}\text{a spare} \\ \text{res'n} \\ \text{is made}\end{bmatrix}\Pr\begin{bmatrix}\text{this spare} \\ \text{res'n is not} \\ \text{collided}\end{bmatrix} \\
&= M\lambda_m e^{-\lambda_m} - (1-x)\left(\lambda_a e^{-\lambda_a}\right)\alpha e^{-\lambda_m}.
\end{aligned} \tag{3.2}$$

Next, λ_r is related to λ_a by

$$\begin{aligned}
\lambda_r &= \begin{bmatrix}\text{Av. no. of} \\ \text{packets arrived} \\ \text{to a slot}\end{bmatrix}\left\{f_1 \Pr\begin{bmatrix}\text{a slot is} \\ \text{of the} \\ \text{Aloha type}\end{bmatrix} + f_2 \Pr\begin{bmatrix}\text{a slot is} \\ \text{of the} \\ \text{reserved type}\end{bmatrix}\right\} \\
&= [\lambda_r + \lambda_a(1-x)]\,[f_1(1-x) + f_2 x].
\end{aligned} \tag{3.3}$$

Finally, the throughput S is given by

$$S = x\Pr\begin{bmatrix}\text{a Res. slot}\\ \text{contains a}\\ \text{succ. tx'n}\end{bmatrix} + (1-x)\Pr\begin{bmatrix}\text{an Aloha slot}\\ \text{contains a}\\ \text{succ. tx'n}\end{bmatrix}$$

$$= x + (1-x)\lambda_a e^{-\lambda_a} \qquad (3.4)$$

The control channel may be regarded as a pure overhead because it is not used for transmitting data packets. Let w be the ratio of the control channel bandwidth to the total channel bandwidth, then

$$S\,|_{\text{with overhead}} = (1-w)S\,|_{\text{without overhead}}\,.$$

3.5 Delay Analysis

The average packet delay $D(\alpha, f_1, f_2)$ consists of seven terms denoted as D_1 to D_7. D_1=0.5 is the average synchronization delay in slots. D_2 is the expected reservation delay and is equal to the round trip propagation delay R (in unit of slots) multiplied by the probability of transmission through reservation or $D_2=(x/S)R$. D_3 is the average waiting time in the satellite reservation queue. For integral values of M, D_3 is given by the waiting time on a discrete-time $M/D/1$ queue with the distribution of the number of arrivals per slot U given by

$$\Pr[U=k] = \binom{M}{k}\left(\frac{x}{M}\right)^k\left(\frac{M-x}{M}\right)^{M-k}.$$

From the Pollaczek-Khinchin mean value formula [KLEI 75a], the mean waiting time D_3 in this queuing system is obtained as

$$D_3 = \frac{x(1-M^{-1})}{2(1-x)}.$$

Note that D_3 with $M \to \infty$ was derived in [LEE 83] as the waiting time in the reservation queue with reservations always successful. D_4=(1+R) is the packet transmission and propagation time. D_5 is the average delay of traffic diversion from the Reserved slots and is given by

$$D_5 = \Pr\begin{bmatrix}\text{a slot is}\\ \text{of the}\\ \text{reserved type}\end{bmatrix}\begin{bmatrix}\text{the fraction of}\\ \text{traffic diverted}\\ \text{from a Reserved slot}\end{bmatrix}\begin{bmatrix}\text{av. duration}\\ \text{between two}\\ \text{Aloha slots}\end{bmatrix}\frac{I-1}{2}$$

$$= x(1-f_2)\frac{I-1}{2(1-x)}.$$

D_6 is the randomization delay for the reservations and is given by

$$D_6 = \left\{\Pr\begin{bmatrix}\text{a slot is}\\ \text{of the}\\ \text{Aloha type}\end{bmatrix}\begin{bmatrix}\text{the fraction of}\\ \text{"Aloha" traffic}\\ \text{with spare res'ns}\end{bmatrix} + \begin{bmatrix}\text{the fraction}\\ \text{that makes}\\ \text{ordinary res'ns}\end{bmatrix}\right\}\frac{K-1}{2M}$$

$$= \left[(1-x)\alpha(1-f_1) + \frac{\lambda_r}{\lambda_r + \lambda_a(1-x)}\right]\frac{K-1}{2M}.$$

D_7 is the average delay due to retransmissions and is given as

$$D_7 = [\text{av. delay per retx'n}]\ [\text{av. no. of retx'n}]$$

$$= \left[R + \frac{J-1}{2} + D_5 + D_6\right]\left[\frac{\lambda_r + \lambda_a(1-x)}{S} - 1\right].$$

Adding up the seven terms, we have

$$D(\alpha, f_1, f_2) = 1.5 + \frac{x(1-M^{-1})}{2(1-x)} + \frac{x+S}{S}R + D_5 + D_6 + \left(R + \frac{J-1}{2} + D_5 + D_6\right)\frac{\lambda_r + \lambda_a(1-x) - S}{S} \tag{3.5}$$

For a given S and M and under constraints (3.1) to (3.4), we can numerically minimize $D(\bullet)$ in (3.5) with respect to α, f_1 and f_2 to obtain the minimum delay protocol in ξ. But in order to find the conditions to maintain minimum delay and to understand the operational mechanism of the protocol for all values of S and M, we have to resort to analytical method. We first break (3.5) into two parts:

$$D(\bullet) = D_I + D_{II}$$

where D_I includes the waiting time for reservation and the propagation delay and D_{II} includes all the randomization delays. Specifically,

$$D_I = 1.5 + \frac{x(1-M^{-1})}{2(1-x)} + [x + \lambda_r + \lambda_a(1-x)]\frac{R}{S} \tag{3.6a}$$

$$D_{II} = D_5 + D_6 + \left(\frac{J-1}{2} + D_5 + D_6\right)\frac{\lambda_r + \lambda_a(1-x) - S}{S} \tag{3.6b}$$

The analytical optimization process involves the following two steps:

1. Since D_I is the dominating term, we shall minimize D_I first with respect to α and λ_a under constraints (3.1), (3.2) and (3.4).
2. By using the optimized α and λ_a from step 1, D_{II} is minimized with respect to f_1 and f_2 under constraint (3.3).

This two step process gives only a sub-optimal solution. It is chosen because simultaneous minimization of $D(\bullet)$ with respect to α, f_1 and f_2 is analytically too difficult. The optimized D_I, denoted as D_I^*, is a natural lower bound of $D(\bullet)$. In section 3.8, we will show numerically that the difference between the sub-optimal solution and D_I^* is insignificant. The closeness of the sub-optimal delay to the delay lower bound implies:

1. D_I indeed dominates over D_{II}.

2. The α, f_1 and f_2 parameters found by the above process are very close to the optimal ones.

3. Condition 2 in ξ is not really restrictive since choosing any smaller randomization parameters can at most reduce the overall delay to D_I^*.

To analytically minimize $D(\cdot)$, we need some lemmas. As these lemmas are self-contained, we place them in the appendix.

3.6 Minimization of D_I

Fig. 3.2 shows that the (λ_a, α) space is divided into two rectangular regions A and B such that in region A, $d\lambda_r/d\alpha < 0$ at $\alpha = 1$ and in region B, $d\lambda_r/d\alpha \geq 0$ at $\alpha = 1$. These conditions determine the value of the boundary point $\hat{\lambda}_a(M)$ such that in region A, $\hat{\lambda}_a(M) < \lambda_a \leq 1$ and in region B, $0 \leq \lambda_a \leq \hat{\lambda}_a(M)$. We make this particular partitioning because, as we shall show later, the locus of the optimal α lies on the boundary of region A. We shall further show that in region A, the minimum delay point is at $(\lambda_a = \lambda_a^*(M), \alpha = 1)$ where $\lambda_a^*(M)$ is the maximum value of λ_a for $\lambda_r \geq 0$ and $\alpha = 1$, and in region B it is at $(\lambda_a = \hat{\lambda}_a(M), \alpha = 1)$. We then show that, for $M \geq 3$, the minimum delay in region A is always smaller than the minimum delay in region B and hence the optimal (λ_a, α) is at $(\lambda_a^*(M), 1)$ for $M \geq 3$. For $M < 3$, we will show via an example in section 3.9 that the $\alpha = 1$ solution is optimal only in a restricted range of throughput. Outside that range, delay minimization has to be entirely numerical. The $M \geq 3$ is the more interesting case because $S_{\max} < 1$ for $M \leq 2$ (shown in [WONG 89]) while $S_{\max} = 1$ for $M \geq 3$ (from Lemma 5 in the appendix).

We now proceed to the details of the derivation. For each region, we first find the optimal α's for specific λ_a's. Then, using these α's, we minimize D_I with respect to λ_a.

3.6.1 Determination of $\hat{\lambda}_a(M)$ and $\hat{x}(M)$

$\hat{\lambda}_a(M)$ and $\hat{x}(M)$ are defined as the values of λ_a and x at $\alpha = 1$ and $d\lambda_r/d\alpha = 0$. Differentiating (3.1) with respect to α and using (A4) and (3.4), we get

$$\frac{d\lambda_r}{d\alpha} = M\frac{d\lambda_m}{d\alpha} - \lambda_a(1-x)$$

$$= \frac{M(S-x) - \lambda_a(1-x)(M - xe^{\lambda_m})}{M - xe^{\lambda_m}}$$

$$= \frac{\lambda_a(1-x)\,[xe^{\lambda_m} + Me^{-\lambda_a} - M]}{M - xe^{\lambda_m}} \qquad (3.7)$$

Since $x<1$, $d\lambda_r/d\alpha = 0$ if and only if

$$xe^{\lambda_m} + Me^{-\lambda_a} - M = 0. \qquad (3.8)$$

Substitute λ_m from (3.8) into (3.2) and set $\alpha = 1$, we obtain

$$\ln\left[\frac{(1-e^{-\lambda_a})M}{x}\right] = 1 - e^{-\lambda_a} + (1-x)\lambda_a e^{-\lambda_a} \qquad (3.9)$$

At a given value of S and M, (3.4) and (3.9) can be solved simultaneously for λ_a and x which are the required $\hat{\lambda}_a(M)$ and $\hat{x}(M)$.

3.6.2 The minimum delay point in region A

Theorem 1:

In region A, D_I is minimized by maximizing α without rendering λ_r negative.

Proof:

Lemma 9 states that for $\lambda_a > \hat{\lambda}_a(M)$, [•] in (3.7) is negative at α=1. Lemma 4 states that λ_m decreases with α. Hence [•] in (3.7) is also negative for α<1. Therefore $d\lambda_r/d\alpha < 0$ for all α. It means that maximizing α will minimize λ_r. For a given λ_a (x is fixed by (3.4)), D_I is minimized by minimizing λ_r or maximizing α.

Q.E.D.

Theorem 2:

The minimum delay point in region A occurs at $\alpha = 1$ and $\lambda_a = \lambda_a^*(M)$.

Proof:

(i) $\lambda_a \in (\hat{\lambda}_a(M), \lambda_a^*(M)]$:

At $\alpha = 1$, $\lambda_a \leq \lambda_a^*(M)$ implies $\lambda_r \geq 0$ from Lemma 7. Therefore, for a given λ_a, D_I is minimized at $\alpha = 1$ by Theorem 1. Using (3.1) and setting $\alpha = 1$, we obtain D_I as

$$D_I(\lambda_a) \quad = \quad 1.5 \quad + \quad \frac{x(1-M^{-1})}{2(1-x)} \quad + \quad \frac{x+M\lambda_m}{S}R \qquad (3.10)$$

To minimize $D_I(\lambda_a)$ with respect to λ_a, (3.10) stipulates that x and λ_m should both be as small as possible. To minimize x and λ_m, Lemmas 1 and 5(i) state that λ_a should be as close to one as possible, while maintaining $\lambda_r \geq 0$. Therefore, D_I is minimized at $\lambda_a = \lambda_a^*(M)$ and $\alpha = 1$.

(ii) $\lambda_a \in (\lambda_a^*(M), 1]$

This case exists only when $\lambda_a^*(M) < 1$. From the definition of $\lambda_a^*(M)$, the constraint $\lambda_r \geq 0$ is binding for $\lambda_a^*(M) < 1$. Therefore, $\lambda_r = 0$ at $\lambda_a = \lambda_a^*(M)$ and $\alpha = 1$. From Lammas 5(ii) and

10, we have $S < S_c(M)$. Also, by Lemma 7, $\lambda_r < 0$ for a given $\lambda_a > \lambda_a^*(M)$ at $\alpha = 1$. Therefore, from Theorem 1 for a given $\lambda_a > \lambda_a^*(M)$ the minimum delay occurs at $\lambda_r = 0$. Next, we minimize D_I with respect to λ_a by setting $\lambda_r = 0$. Solving x from (3.4), substituting into (3.6a) with $\lambda_r = 0$, and differentiating with respect to λ_a, we have

$$\frac{dD_I}{d\lambda_a} = \frac{R(1-S)\left[1 - e^{-\lambda_a} + \lambda_a e^{-\lambda_a}(1-\lambda_a)\right]}{S\left(1-\lambda_a e^{-\lambda_a}\right)^2} - \frac{(1-M^{-1})}{1-S} e^{-\lambda_a}(1-\lambda_a) \qquad S < S_c(M). \quad (3.11)$$

This derivative can be shown to be an increasing function of λ_a. Since Lemma 11 stipulates that $\lambda_a \geq S$, $dD_I/d\lambda_a$ is minimized at $\lambda_a = S$. Setting $\lambda_a = S$, (3.11) becomes

$$\frac{dD_I}{d\lambda_a} \geq \frac{R(1-S)[1 - e^{-S} + Se^{-S}(1-S)]}{S(1-Se^{-S})^2} - e^{-S} \equiv \phi(S) \qquad S < S_c(M).$$

Noting that $\frac{d\phi(S)}{dS} < 0$ and $S_c(\infty) > S_c(M)$, we have,

$$\frac{dD_I}{d\lambda_a} \; > \; \phi(S_c(M)) \; > \; \phi(S_c(\infty)).$$

For $R \geq 1$, $\phi(S_c(\infty)) > 0$. Therefore, $dD_I/d\lambda_a > 0$ and the delay is minimized at the minimum possible value of λ_a, i.e. at $\lambda_a = \lambda_a^*(M)$ with $\alpha = 1$.

Q.E.D.

To summarize, after setting $\alpha = 1$, if $S \geq S_c(M)$, we set $\lambda_a^*(M) = 1$ and solve for x, λ_m and λ_r simultaneously from (3.1), (3.4) and (A5). By substituting them into (3.10), D_I^* can be found. If $S<S_c(M)$, the choice $\lambda_a(M) = 1$ will render λ_r negative. Therefore, we choose $\lambda_r = 0$ and solve for $\lambda_a^*(M)$ and λ_m simultaneously from (A5) and (A6) and substitute them into (3.10)

to find D_I^*. The choice of $\lambda_r = 0$ results in minimum delay because from Lemma 7, an increase of λ_r will cause a decrease of λ_a and hence an increase of D_I. As λ_a is the traffic rate to the Aloha slots. The above says that for minimum delay the Aloha slots should be filled with a packet rate of one per slot whenever possible.

3.6.3 The minimum delay point in region B

In region B, the locus of the optimal α as λ_a varies is generally not on the boundary of the region. Locating the minimum delay point in this region appears to be analytically very difficult. What we shall do instead, is to find a lower bound of this minimum delay and to prove that this lower bound is always larger than the minimum delay in region A for $M \geq 3$. Therefore, finding the *exact* minimum delay in region B is not important because the global minimum delay point for $M \geq 3$ is in region A. The delay lower bound is obtained by making a *noncausal* assumption. Let us assume that all packets which are successfully transmitted in the Aloha slots did not make any spare reservations on the control channel. This noncausal assumption guarantees that there is no spare reservation from successful packets to interfere with the other reservations and hence will result in a smaller average delay.

Under the noncausal assumption, let Λ_a be the average number of transmissions in an Aloha slot, Λ_r be the average number of ordinary reservations per slot on the control channel. Then, the combined rate of ordinary and spare reservations per minislot to the control channel, denoted as Λ_m, is

$$\Lambda_m = \frac{\Lambda_r + \alpha\Lambda_a(1 - e^{-\Lambda_a})(1 - x)}{M} \tag{3.12}$$

The average number of successful reservations per slot x is

$$x = M[\text{av. no. of successful reservation in a minislot}]$$
$$= M\Lambda_m e^{-\Lambda_m} \quad (3.13)$$

Substituting (3.13) into (3.4), we have

$$S = M\Lambda_m e^{-\Lambda_m} + \left(1 - M\Lambda_m e^{-\Lambda_m}\right)\Lambda_a e^{-\Lambda_a} \quad (3.14)$$

From (3.6a), we obtain D_I as

$$D_I(\Lambda_a) \;=\; 1.5 \;+\; \frac{x(1-M^{-1})}{2(1-x)} \;+\; \frac{x+\Lambda_r+\Lambda_a(1-x)}{S}R. \quad (3.15)$$

Lemma 14 states that for a given Λ_a, $D_I(\Lambda_a)$ is minimized at $\alpha = 1$.

Theorem 3:

Under the noncausal assumption, the minimum delay point in region B is at $\Lambda_a = \hat{\lambda}_a(M)$ and $\alpha = 1$.

Proof:

From (3.4) and (3.12) and setting the optimal value of $\alpha = 1$, (3.15) becomes

$$D_I(\Lambda_a) \;=\; 1.5 \;+\; \frac{x(1-M^{-1})}{2(1-x)} \;+\; \frac{M\Lambda_m+S}{S}R. \quad (3.16)$$

To minimize $D_I(\Lambda_a)$, (3.16) stipulates that x and Λ_m should both be as small as possible. Lemmas 1 and 15 state that Λ_a should as large as possible. Therefore, the delay is minimized at $\Lambda_a = \hat{\lambda}_a(M)$ and $\alpha = 1$.

Q.E.D.

3.6.4 Delay comparison in the two regions

Theorem 4:

The minimum delay in region A is always smaller than the minimum delay in region B for $M \geq 3$.

Proof:

(i) $S < S_c(\infty)$:

First, we consider region A. From (3.10) we obtain the minimum delay in this region as

$$D_I(\lambda_a = \lambda_a^*(M)) \;=\; 1.5 \;+\; \frac{x^*(M)\,(1-M^{-1})}{2(1-x^*(M))} \;+\; \frac{x^*(M)+M\lambda_m}{S}R \qquad (3.17)$$

where $x^*(M)$ denotes the optimized x found before.

Next, we consider region B. Since $\hat{x}(M)<M\Lambda_m$ from (3.13), we obtain the minimum delay in this region from (3.16) as

$$D_I(\Lambda_a = \hat{\lambda}_a(M)) \;>\; 1.5 \;+\; \frac{\hat{x}(M)\,(1-M^{-1})}{2(1-\hat{x}(M))} \;+\; \frac{\hat{x}(M)+S}{S}R \qquad (3.18)$$

For M=3, numerical results shows that $\hat{x}(M)+S > x^*(M)+M\lambda_m$ for $S < S_c(\infty)$. $[\hat{x}(M)+S]$ increases with M by Lemma 12. Under both "$\lambda_a^*(M)=1$" and "$\lambda_r=0$" conditions, $[x^*(M)+M\lambda_m]$ decreases with increasing M from Lemmas 2(i) and 3. Therefore,

$$\hat{x}(M)+S > x^*(M)+M\lambda_m \qquad \text{for} \quad M \geq 3.$$

Together with $\hat{x}(M) > x^*(M)$ (from Lammas 1 and 13), we have

$$D_I(\lambda_a = \lambda_a^*(M)) < D_I(\Lambda_a = \hat{\lambda}_a(M)) \quad \text{for} \quad M \geq 3.$$

(ii) The proof for $S \geq S_c(\infty)$ is similar.

Q.E.D.

3.7 Minimization of D_{II}

From (3.6b), we can see that minimizing D_{II} is equivalent to minimizing $D_5 + D_6$ where

$$D_5 + D_6 \;=\; x(1-f_2)\frac{I-1}{2(1-x)} \;+\; \left[(1-x)\alpha(1-f_1) + \frac{\lambda_r}{\lambda_r + \lambda_a(1-x)}\right]\frac{K-1}{2M} \tag{3.19}$$

Substituting f_2 from (3.3) into (3.19), we have

$$D_5 + D_6 \;=\; \left[x - \frac{\lambda_r}{\lambda_r + \lambda_a(1-x)}\right]\frac{I-1}{2(1-x)} \;+\; \left[\alpha(1-x) + \frac{\lambda_r}{\lambda_r + \lambda_a(1-x)}\right]\frac{K-1}{2M} \;+\; f_1\left[\frac{I-1}{2} - \frac{\alpha(1-x)(K-1)}{2M}\right]$$

We choose $I = K$ to make $[\cdot]$ of the last term positive. Therefore, to minimize $D_5 + D_6$ (or D_{II}), f_1 should be chosen as small as possible while maintaining $f_2 \leq 1$ as governed by (3.3).

3.8 Numerical Examples

Numerical results show that for M=2 and $0.77<S<0.83$, the minimized D_I occurs at $\alpha < 1$. This means that making spare reservation for all packets transmitted in the Aloha slot is not always the best for small values of M. This is also to be expected since spare reservations have a high chance to collide with ordinary reservations when M is small. In practice, M

rarely needs to be set as low as 2 and so for all practical purpose, always making a spare reservation with each transmission in the Aloha slot (i.e. setting $\alpha = 1$) is the optimal operating condition.

Let R=100, w=0 and I=J=K=10. Figs. 3.3 and 3.4 show the average delay of the UCA protocol [LEE 83], the Controlled Multiaccess protocol [WONG 88], and the Minimum Delay protocol for M=3 and M=6 respectively. We choose UCA and Controlled Multiaccess for comparison because they have the best delay performance found in literature. They are, however, also more complicated. As expected, the Minimum Delay protocol has an average delay smaller than the other two protocols. Moreover, this delay is less than 2.5% higher than its lower bound D_I^*.

Fig. 3.5 compares the average delay of the Minimum Delay protocol for M=10 and $M = \infty$. As there is less than 5% difference in the two delays for $S \leq 0.95$, ten minislots per slot is sufficient to give a near optimal performance.

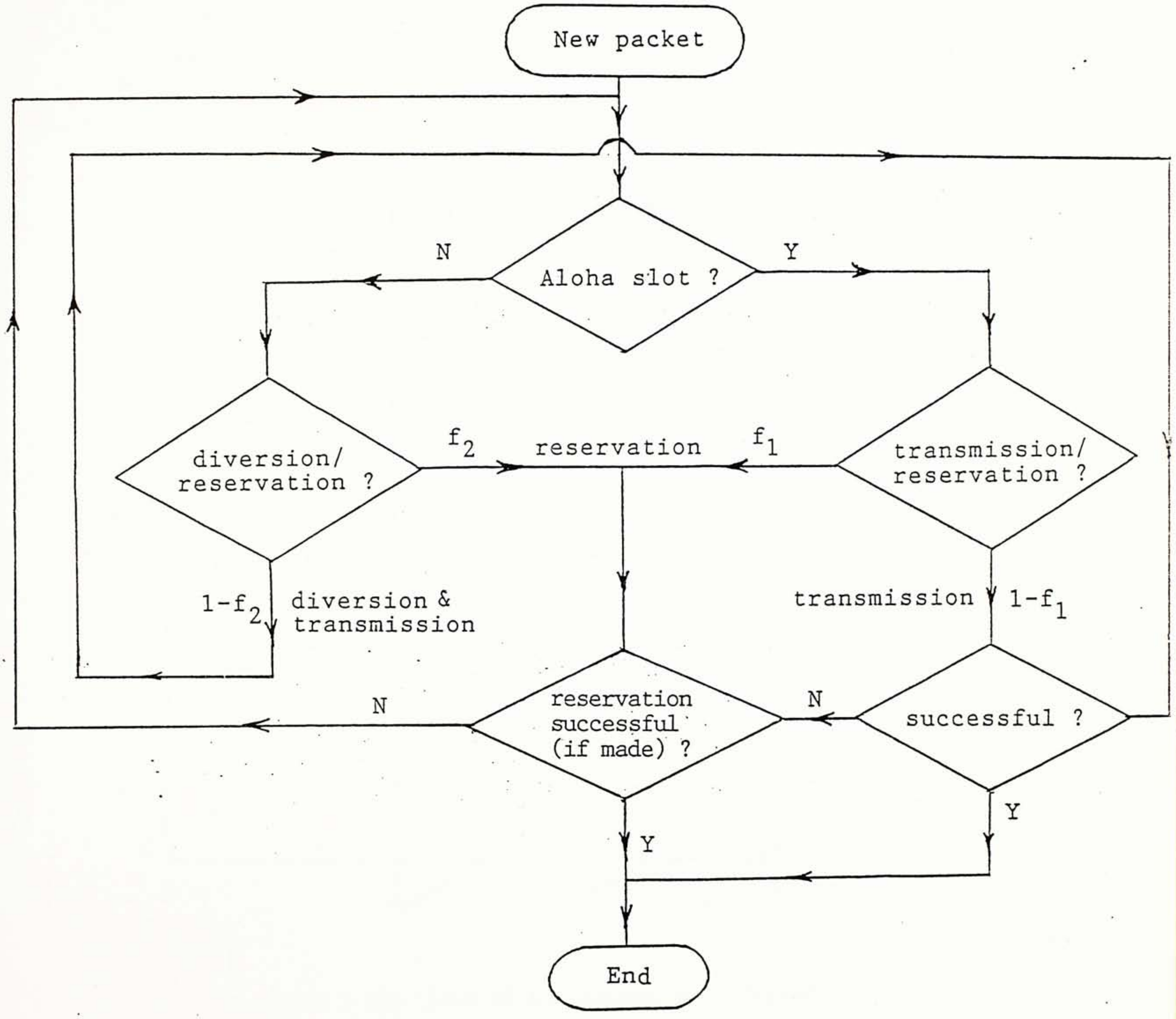
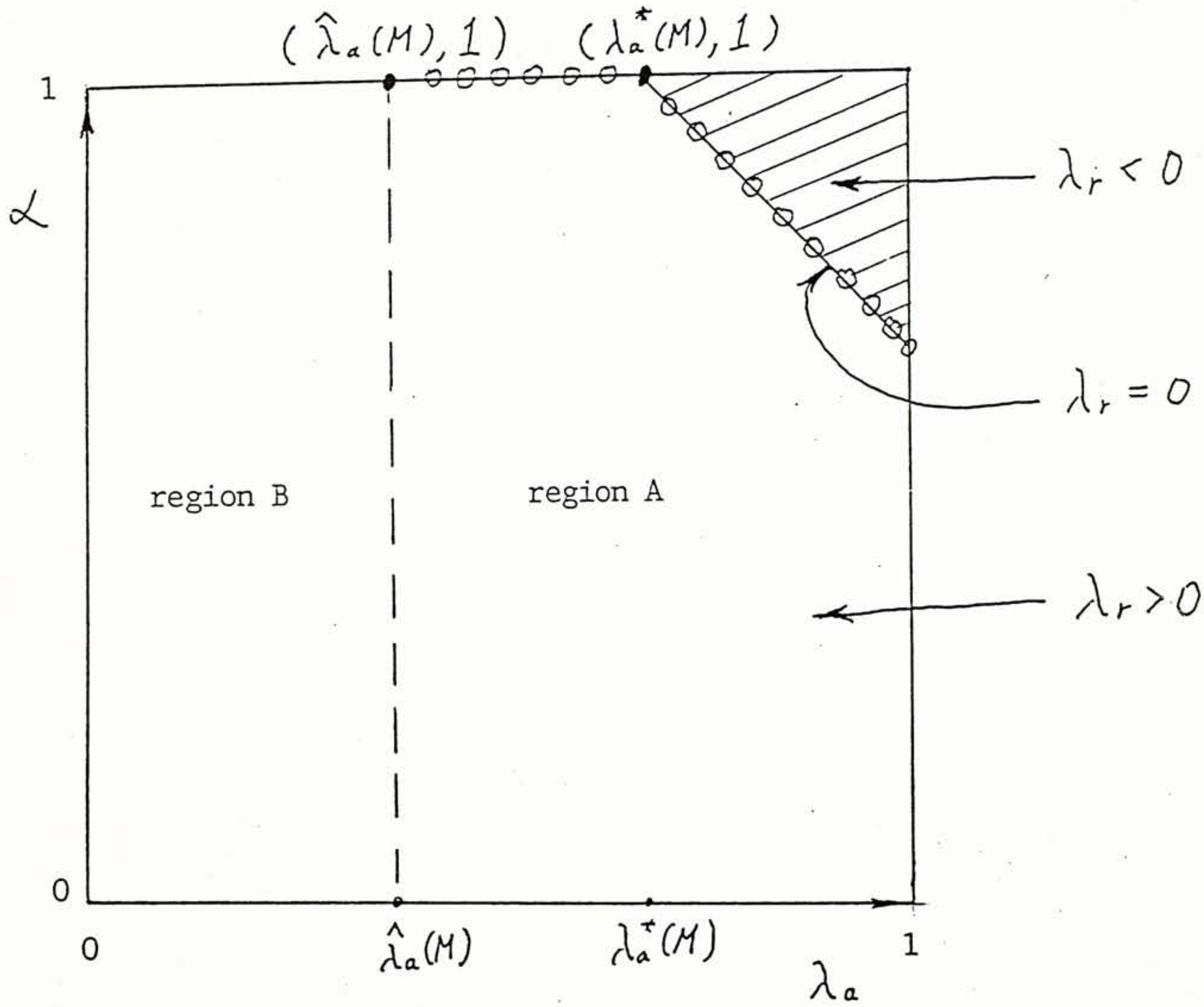


Fig.3.1.Flow chart of the minimum delay protocol.



o-o-o-o : the locus of the optimal α in region A

Fig.3.2. The (λ_a, α) space for delay minimization.

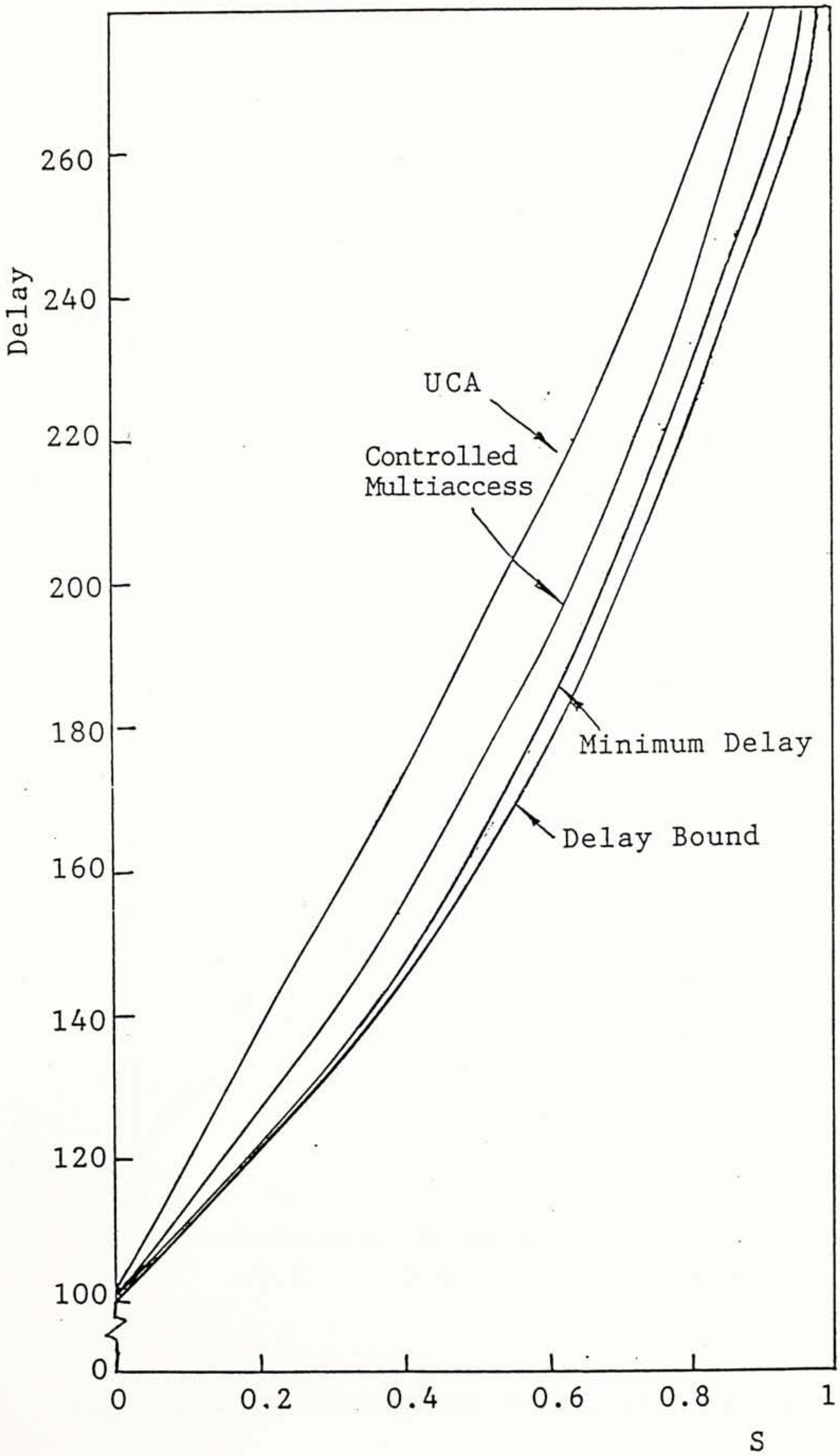


Fig.3.3.Delay Throughput characteristics, M=3.

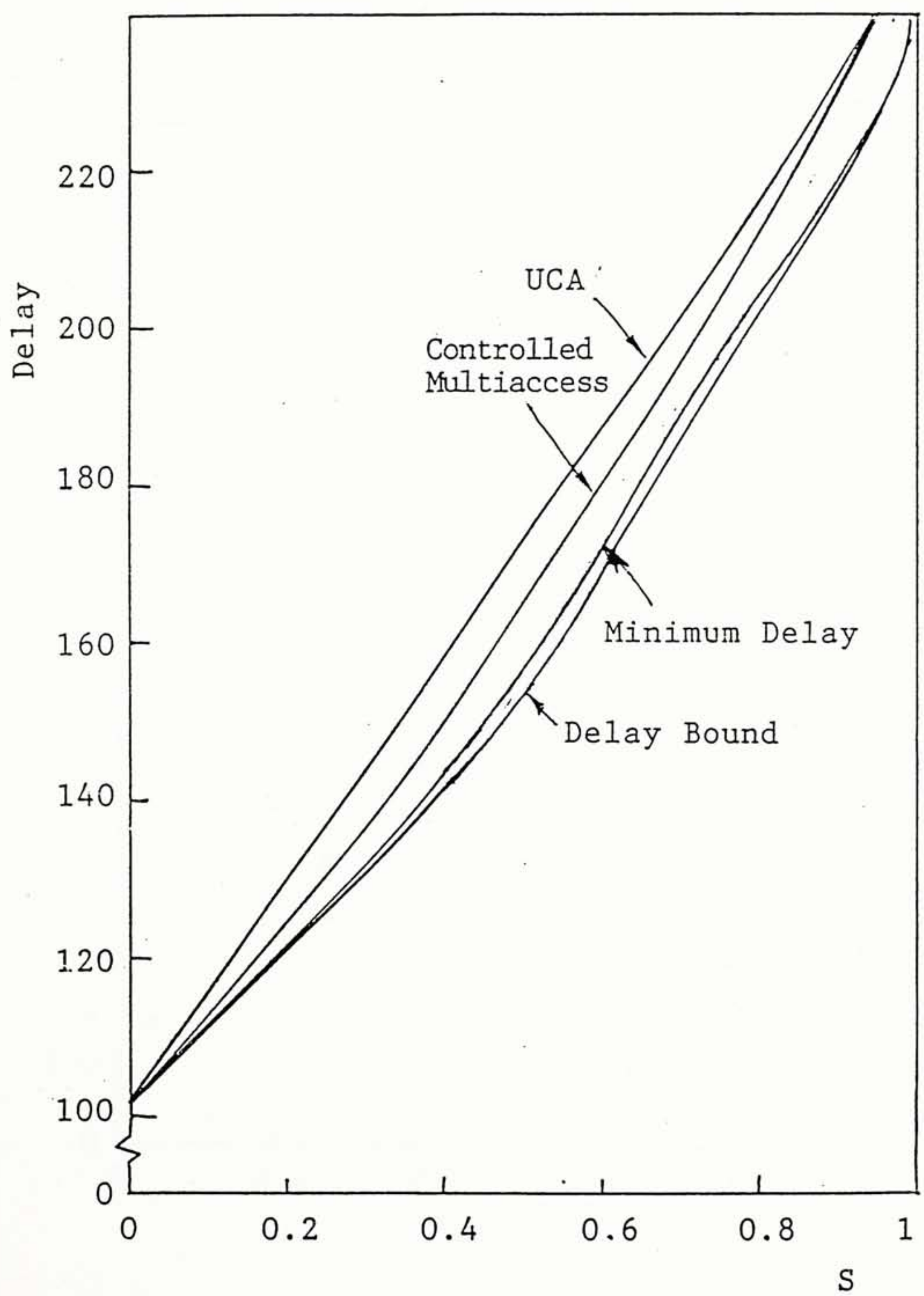


Fig.3.4.Delay Throughput characteristics, M=6.

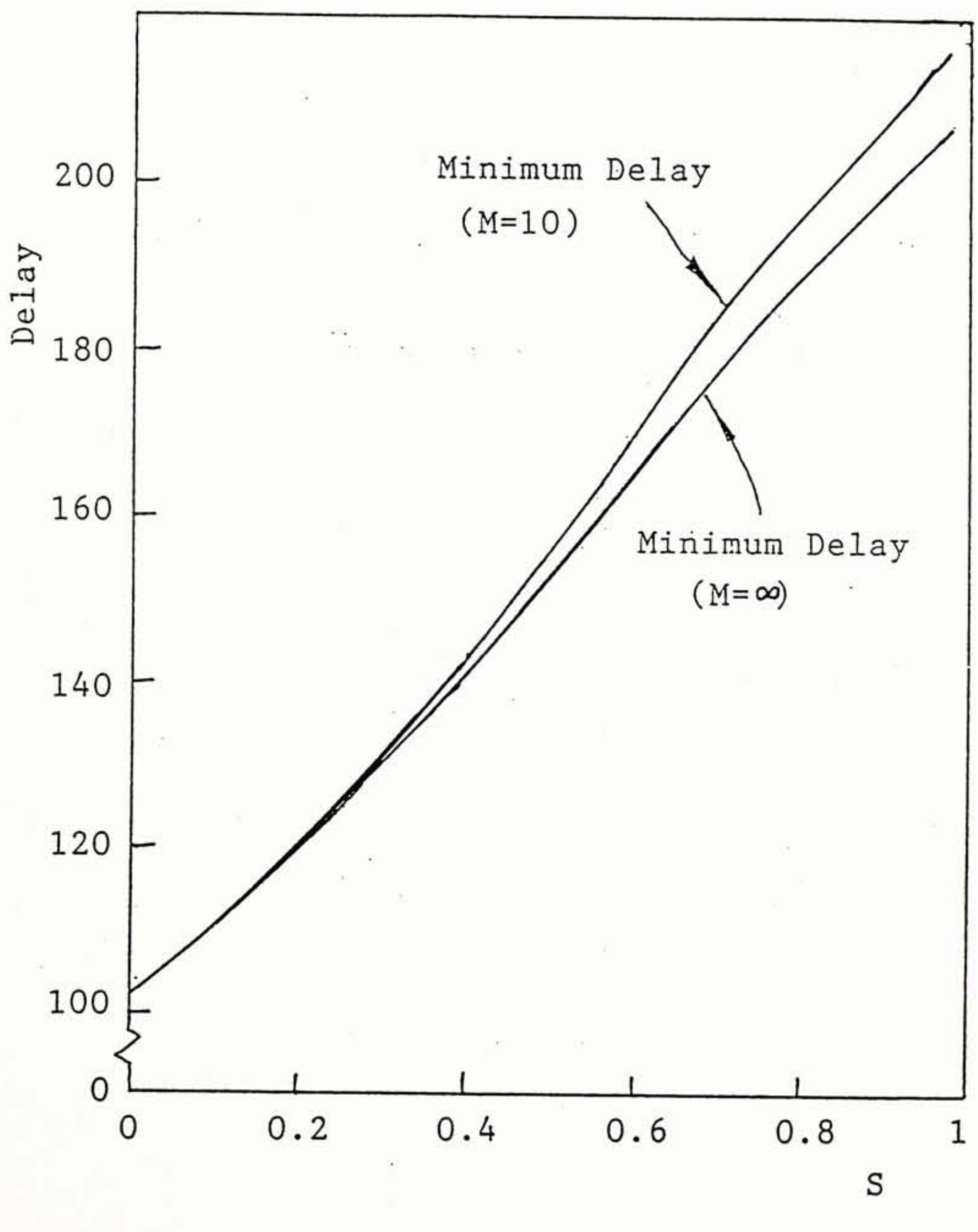


Fig.3.5. M=10 is quite sufficient for near-optimum delay performance.

CHAPTER 4

CONCLUSIONS

In chapter 2, two delay lower bounds are derived for packet satellite protocols under a set of operating conditions. They are shown to be very simple and very tight. They can be used for assessing the possible delay improvements of existing protocols and for deciding whether a particular delay requirement can ever be satisfied.

In chapter 3, the minimum delay protocol is under the assumptions of Poisson arrivals and single copy transmission. Steady state analysis is used to obtain the optimal protocol parameters. For correlated and non-stationary input processes, some form of adaptive control is needed for satisfactory performance. The design and optimization of these "adaptive" protocols appears to be a real challenge.

Only the overall average delay is minimized in chapter 3. In practice, for systems with different classes of traffic where each class has a different delay requirement, the protocol design appears to be very complicated. This is particularly true when the options of multiple transmission copies per packet and multiple reservations per packet are allowed. Multiaccess communication is indeed a fascinating field of research.

REFERENCES

[BOSE 80] S. Bose and S.S. Rappaport, "High Capacity: Low Delay Packet Broadcast Multiaccess," IEEE Trans. on AES, AES-16, No. 6, pp.830-838, Nov. 1980.

[CHAN 84] J.F. Chang and L.Y. Lu, "Distributive Demand-Assigned Packet Switching with Trailer Transmissions," IEEE Trans. on AES, AES-20, No.6, pp.775-787, Nov. 1984.

[CHIT 88] D.M. Chite AND J.S. McCoskey, "VSAT Networks: Architectures, Protocols, and Management," IEEE Communications Magazine, July 1988, VOl.26, No.7.

[KLEI 75a] L. Kleinrock, *Queueing Systems*, Vol.1: Theory, John Wiley, 1975, pp.191.

[KLEI 75b] L. Kleinrock and S.S. Lam, "Packet-Switching in a Multi-Access Broadcast Channel: Performance Evaluation," IEEE Trans. on Commun., COM-23, April 1975, pp.410-423.

[LEE 83] H.W. Lee and J.W. Mark, "Combined Random/Reservation Access for Packet Switched Transmission over a Satellite with On-Board Processing: Part I-Global Beam Satellite," IEEE Trans. on Commun., COM-31, No. 10, pp.1161-1171, Oct. 1983.

[PRAT 86] T. Pratt and C.W. Bostian, *Satellite Communications*, John, Wiley & Sons, Inc., 1986.

[WONG 88] Eric W.M. Wong and Tak-Shing Yum, "The Controlled Multiaccess Protocol for Packet Satellite Communication," IEEE Proc. of ICC'88, June 1988. Also to appear in IEEE Trans. on Communications.

[WONG 89] Eric W.M. Wong and Tak-Shing Yum, "Delay Bounds for Packet Satellite Protocols," IEEE GLOBECOM '89, Nov. 1989.

[YUM 87] Tak-Shing Yum and Eric W.M. Wong, "The Design and Analysis of the Scheduled-Retransmission Multiaccess Protocol for Packet Satellite

Communications," IEEE, International Conference on Communications, Seattle, Washington, USA, June 1987. Also, to appear in IEEE Trans. on Information Theory.

APPENDIX

Lemma 1:

x and λ_a are inversely related.

Proof:

This follows from differentiating (3.4).

Q.E.D.

Lemma 2:

For fixed α, λ_a and x,

$$(i)\frac{d(M\lambda_m)}{dM} < 0, \quad (ii)\frac{d\lambda_m}{dM} < 0, \quad \text{and} \quad (i)\frac{d\lambda_r}{dM} < 0.$$

Proof:

(i) Solving for $(1-x)\lambda_a e^{-\lambda_a}$ in (3.4) and substituting into (3.2), we have

$$xe^{\lambda_m} = M\lambda_m - \alpha(S-x). \tag{A1}$$

For fixed α, λ_a and x, differentiating $M\lambda_m$ in (A1) with respect to M, we have

$$\frac{d(M\lambda_m)}{dM} = \frac{-\lambda_m x e^{\lambda_m}}{M - xe^{\lambda_m}} \tag{A2}$$

Since $S>x$ (from (3.4)), we have from (A1)

$$xe^{\lambda_m} < M\lambda_m < M.$$

Substituting into the denominator of (A2), we obtain

$$\frac{d(M\lambda_m)}{dM} < 0.$$

(ii) Differentiating λ_m in (A1) with respect to M, we obtain

$$\frac{d\lambda_m}{dM} = \frac{-\lambda_m}{M - xe^{\lambda_m}} < 0.$$

(iii) From (3.1) we have $M\lambda_m = \lambda_r + \alpha\lambda_a(1-x)$. Differentiating, we have

$$\frac{d\lambda_r}{dM} = \frac{d(M\lambda_m)}{dM} < 0. \qquad \text{Q.E.D.}$$

Lemma 3:

$$\frac{d(x + M\lambda_m)}{dM} < 0,$$

for $\alpha = 1$, $\lambda_r = 0$ and S fixed.

Proof:

Substitute x (from (3.4)) and λ_m (from (3.1)) into (A1), set $\alpha = 1$ and $\lambda_r = 0$, and then differentiate with respect to M, we have $d\lambda_a/dM < 0$. Differentiating (3.4), we have

$$\frac{dx}{dM} = \frac{-(1-x)(1-\lambda_a)e^{-\lambda_a}}{1-\lambda_a e^{-\lambda_a}}\frac{d\lambda_a}{dM} \qquad \text{(A3)}$$

Differentiating $(x + M\lambda_m)$ using (3.1) and substituting by (A3), we have

$$\frac{d(x+M\lambda_m)}{dM} = (1-x)\left[1 - \frac{(1-\lambda_a)^2 e^{-\lambda_a}}{1-\lambda_a e^{-\lambda_a}}\right]\frac{d\lambda_a}{dM} < 0$$

since$[\cdot] > 0$. Q.E.D.

Lemma 4:

For fixed λ_a, x and M, $d\lambda_m/d\alpha > 0$.

Proof:

Differentiating (A1) with respect to α, we have

$$\frac{d\lambda_m}{d\alpha} = \frac{S - x}{M - xe^{\lambda_m}} > 0. \tag{A4}$$

Q.E.D.

Lemma 5:

At $\alpha = 1$,

(i) λ_m and λ_a are inversely related.

(ii) S is a monotonically increasing function of λ_a and λ_m; hence it is maximized at λ_a=λ_m=1.

(iii) the minimum M (denoted as M^*) for maximum throughput is M^*=e.

Proof:

(i) and (ii): Setting $\alpha = 1$ in (3.2) and solve for x, we have

$$x = \frac{M\lambda_m - \lambda_a e^{-\lambda_a}}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \tag{A5}$$

Substituting into (3.4), we obtain

$$S = \frac{M\lambda_m + \left(e^{\lambda_m} - M\lambda_m - 1\right)\lambda_a e^{-\lambda_a}}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \tag{A6}$$

where $\lambda_a \leq 1$ and $\lambda_m \leq 1$. By differentiating (A6) with respect to λ_m and λ_a, we obtain (i) and (ii) of Lemma 5.

(iii) Setting $S=\lambda_a=\lambda_m=1$ in (A6) and solving for M, we obtain $M^*=e$.

Q.E.D.

Lemma 6:

At $\alpha=1$ and for fixed λ_a, λ_r is a monotonically increasing function of λ_m.

Proof:

Substituting (A5) into (3.1) and solving for λ_r, we have

$$\lambda_r = \frac{M\lambda_m e^{\lambda_m} - M\lambda_m\lambda_a e^{-\lambda_a} - \lambda_a e^{\lambda_m} + M\lambda_a\lambda_m}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \quad \text{(A7)}$$

By differentiating λ_r with respect to λ_m, we obtain Lemma 6.

Q.E.D.

Lemma 7:

At $\alpha=1$, λ_r and λ_a are inversely related.

Proof:

Lemma 5(i) stipulates that λ_a decreases with increasing λ_m for a fixed S. However, from differentiating (A7) we know that the decrease of λ_a causes an increase of λ_r for a fixed λ_m. Also, Lemma 6 states that increasing λ_m causes a corresponding increase of λ_r for a fixed λ_a. Therefore λ_r is a monotonically decreasing function of λ_a for a fixed S.

Q.E.D.

Lemma 8:

λ_r is a monotonically increasing function of S for a fixed λ_a.

Proof:

This follows from Lemmas 5(ii) and 6.

Q.E.D.

Lemma 9:

$[xe^{\lambda_m} + Me^{-\lambda_a} - M]$ is a monotonically decreasing function of λ_a at $\alpha = 1$.

Proof:

As λ_a increases at $\alpha = 1$, x and λ_m will decrease according to Lemmas 1 and 5(i) respectively. Therefore, $[\bullet]$ decreases with increasing λ_a.

Q.E.D.

Lemma 10:

$S_c(M) \equiv S\,|_{\lambda_a = 1, \lambda_r = 0, \alpha = 1}$ increase with M.

Proof:

For fixed α, λ_a and x, as M increases, λ_r will decrease according to Lemma 2(iii). On the other hand, Lemma 8 states that λ_r increases with S for fixed M and λ_a. Therefore, $S_c(M)$ increases with M.

Q.E.D.

Lemma 11:

$\lambda_a \geq S$ at $\lambda_r = 0$.

Proof:

Substituting (3.1) into (3.2) and then into (3.4) and setting λ_r=0, we have

$$S = \lambda_a(1 - x)\left[\alpha\left(1 - e^{-\lambda_a}\right)e^{-\lambda_m} + e^{-\lambda_a}\right].$$

Since $(1 - x) \leq 1$ and $[\bullet] \leq 1$, we have $\lambda_a \geq S$.

Q.E.D.

Lemma 12:

$\hat{x}(M)$ is a monotonically increasing function of M.

Proof:

From (3.8), we have

$$\frac{\hat{x}(M)}{1-e^{-\hat{\lambda}_a(M)}} = M e^{-\lambda_m}.$$

As M is increased, λ_m decreases according to Lemma 2(ii). Therefore, $\frac{\hat{x}(M)}{1-e^{-\hat{\lambda}_a(M)}}$ increases with M. But as $\hat{x}(M)$ is increased, Lemma 1 states that $1-e^{-\hat{\lambda}_a(M)}$ is decreased. Therefore $\frac{\hat{x}(M)}{1-e^{-\hat{\lambda}_a(M)}}$ is increased if and only if $\hat{x}(M)$ is increased.

Q.E.D.

Lemma 13:

For $M \geq e$ and $\alpha = 1$, $\lambda_a^*(M) > \hat{\lambda}_a(M)$

Proof:

Numerical results show that $\lambda_a^*(e) > \hat{\lambda}_a(e)$. Therefore by Lemma 9, [•] in (3.7) is negative at λ_a=$\lambda_a^*(e)$. Substituting (3.1) and (3.2) into [•] in (3.7), we have

$$[\bullet] = \frac{\lambda_r + (1-e^{-\lambda_a})[-M + \lambda_a(1-x)]}{M}.$$

If $\lambda_a^*(M) = 1$, we have λ_r decreasing with increasing M by Lemma 2(iii) and hence [•] remains negative. On the other hand, if $\lambda_a^*(M) < 1$, the constraint $\lambda_r \geq 0$ is binding, i.e. $\lambda_r = 0$ and [•] remains negative for M > e since $[-M + \lambda_a(1-x)]$ in [•] is always negative. Therefore by Lemma 9, $\lambda_a^*(M) > \hat{\lambda}_a(M)$ for $M \geq e$.

Q.E.D.

Lemma 14:

For a given Λ_a, Λ_r is minimized at $\alpha = 1$.

Proof:

Substituting (3.12) into (3.13) and differentiating with respect to α, we have

$$\frac{d\Lambda_r}{d\alpha} = -\Lambda_a\left(1 - e^{-\Lambda_a}\right)(1 - x) < 0.$$ Q.E.D.

Lemma 15:

Λ_m and Λ_a are inversely related.

Proof:

It follows from differentiating (3.14) with respect to Λ_m and Λ_a

Q.E.D.