PAPER Special Issue on Satellite Communications Networking and Applications

In Search of the Minimum Delay Protocol for Packet Satellite Communications

Eric W. M. WONG† *and* Tak-Shing Peter YUM†, *Nonmembers*

SUMMARY 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 (i.e. Aloha slots) should be filled with a packet rate of one per slot whenever possible. In other words, the utilization of Aloha slots should be maximized.
3) an optimum balance between transmitting packets and making reservations before transmission should be maintained.

key words: *satellite communications, protocols, minimum delay*

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.(1)-(5)

In this paper, 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.(6)
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.(7) 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 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 Ref. (3), 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

Manuscript received August, 31, 1992.
Manuscript revised January, 12, 1993.

† The authors are with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

iould depend on the channel state and the channel iading condition.

In the following, we shall first describe the packet itellite system. We then design the protocol to be ptimized and derive its throughput and delay characristics. Finally, we minimize the delay analytically ith respect to all tunable parameters to obtain the inimum delay protocol in ξ as well as the set of nditions for maintaining minimum delay.

The Packet Satellite System

Consider a packet satellite system. Besides the plink data channel used for transmitting packets, let iere also be an uplink narrow-band control channel ır making reservation and a downlink announcement iannel for broadcasting successful reservation. In ractice, the control channel and the announcement iannel can be subchannels on the up- and the downnk data channels respectively. The data channel is otted with slot size equal to one packet transmission me. The control channel is divided into minislots ith M (need not be an integer) minislots per slot Fig. 1). In contrast to other protocols such as that in efs. (2)-(4), framing of slots is not needed. There are ro types of slots. The Aloha slots are for transmitting ackets without prior reservations whereas the eserved slots are for transmitting packets with sucssful reservations. The control channel serves two urposes:

to make reservations for transmissions on the data iannel and

to make spare reservations for retransmissions in ise the transmissions in Aloha slots fail. The inouncement channel is used to broadcast the locaons of the Reserved slots to all stations. All noneserved slots are treated as Aloha slots.

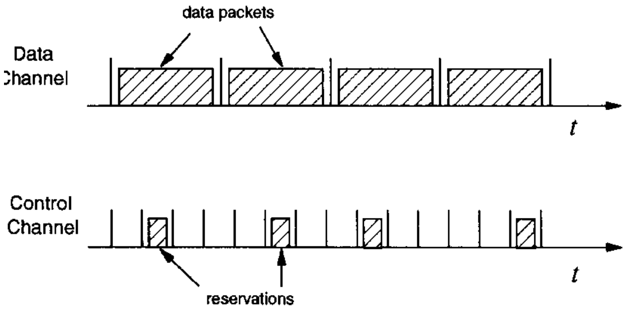
The Transmission Protocol

Consider the arrival of a packet. If it hits an loha slot, it will either, with the probability f_1, make reservation on the control channel and await its signed Reserved slot, or with the remaining probabil y $1-f_1$, be transmitted in the current Aloha slot. In the latter case, the packet can, with probability a, 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 a. 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. 2.

The protocol being designed takes all optimizable options, namely, the parameters a, f_1 and f_2, into consideration. The optimal setting of these parameters in different traffic conditions (learnt through measurements on the channel) guarantees minimum average packet delay. Optimal control parameters are computed offline and stored. Whenever there is a significant change of traffic rate, new optimal control values are looked up and used. As the protocol needs to maintain a reservation queue some very simple on-board processing is required.

The stability of the protocol can be appreciated in a very intuitive manner. When the traffic is very heavy, both f_1 and f_2 are set to 1. This means that all packets will have to make reservations before transmission. It therefore behaves purely as a reservation protocol and



ig. 1 Slots for data packets and minislot for reservations.

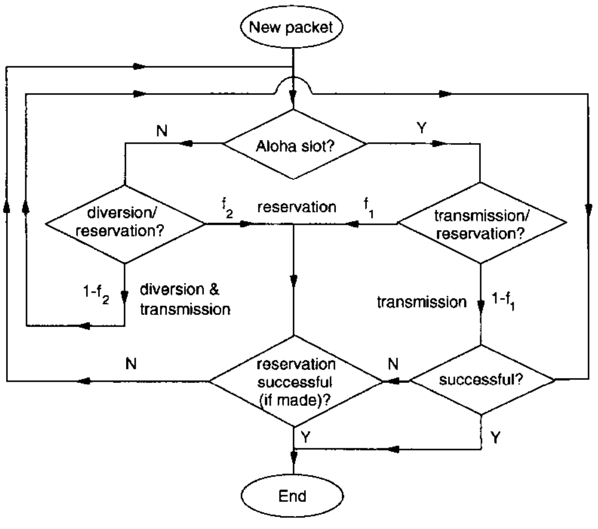


Fig. 2 Flow chart of the minimum delay protocol.

is always stable. On the other hand when traffic is very light both f_1 and f_2 will be zero, meaning that all packets are transmitted without reservation. This therefore is just slotted Aloha.

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

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

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,

$$\begin{aligned}
x&=[\text{av. no. of successful reservation per slot}]\\
&=\left[\begin{array}{c}\text{av. no. of uncollided}\\ \text{reservations}\\ \text{in } M \text{ minislots}\end{array}\right]\\
&\quad-\Pr\left[\begin{array}{c}\text{a slot is}\\ \text{of the}\\ \text{Aloha type}\end{array}\right]\left[\begin{array}{c}\text{av. no. of spare res'ns}\\ \text{to be ignored}\\ \text{in an Aloha slot}\end{array}\right]\\
&=M\lambda_m e^{-\lambda_m}-(1-x)\Pr\left[\begin{array}{c}\text{a packet is}\\ \text{succ. tx'ed in}\\ \text{an Aloha slot}\end{array}\right]\\
&\quad\cdot\Pr\left[\begin{array}{c}\text{a spare}\\ \text{res'n}\\ \text{is made}\end{array}\right]\Pr\left[\begin{array}{c}\text{this spare}\\ \text{res'n is not}\\ \text{collided}\end{array}\right]\\
&=M\lambda_m e^{-\lambda_m}-(1-x)\,(\lambda_a e^{-\lambda_a})\,\alpha e^{-\lambda_m}.
\end{aligned} \tag{2}$$

Next, λ_r is related to λ_a by

$$\begin{aligned}
\lambda_r&=\left[\begin{array}{c}\text{Av. no. of}\\ \text{packets arrived}\\ \text{to a slot}\end{array}\right]\left\{f_1\Pr\left[\begin{array}{c}\text{a slot is}\\ \text{of the}\\ \text{Aloha type}\end{array}\right]\right.\\
&\quad\left.+f_2\Pr\left[\begin{array}{c}\text{a slot is}\\ \text{of the}\\ \text{reserved type}\end{array}\right]\right\}\\
&=[\lambda_r+\lambda_a(1-x)][f_1(1-x)+f_2x].
\end{aligned} \tag{3}$$

Finally, the throughput S is given by

$$\begin{aligned}
S&=x\Pr\left[\begin{array}{c}\text{a Res. slot}\\ \text{contains a}\\ \text{succ. tx'n}\end{array}\right]+(1-x)\Pr\left[\begin{array}{c}\text{an Aloha slot}\\ \text{contains a}\\ \text{succ. tx'n}\end{array}\right]\\
&=x+(1-x)\lambda_a e^{-\lambda_a}
\end{aligned} \tag{4}$$

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

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

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,[8] the mean waiting time D_3 in this queueing system is obtained as

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

Note that D_3 with $M\rightarrow\infty$ was derived in Ref. (3) 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

$$\begin{aligned}
D_5&=\Pr\left[\begin{array}{c}\text{a slot is}\\ \text{of the}\\ \text{reserved type}\end{array}\right]\left[\begin{array}{c}\text{the fraction of}\\ \text{traffic diverted}\\ \text{from a Reserved slot}\end{array}\right]\\
&\quad\cdot\left[\begin{array}{c}\text{av. duration}\\ \text{between two}\\ \text{Aloha slots}\end{array}\right]\frac{I-1}{2}\\
&=x(1-f_2)\frac{I-1}{2(1-x)}.
\end{aligned}$$

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

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

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

D_7 is the average delay due to retransmissions and is iven 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

$$D(\alpha,f_1,f_2)=1.5+\frac{x(1-M^{-1})}{2(1-x)}+\frac{x+S}{S}R+D_5+D_6+\left(R+\frac{J-1}{2}+D_5+D_6\right)\cdot\frac{\lambda_r+\lambda_a(1-x)-S}{S} \qquad (5)$$

For a given S and M and under constraints Eqs. 1) to (4), we can numerically minimize $D(\cdot)$ in Eq. 5) with respect to α, f_1 and f_2 to obtain the minimum elay protocol in ξ. But in order to find the conditions o maintain minimum delay and to understand the perational mechanism of the protocol for all values of and M, we have to resort to analytical method. We rst break Eq. (5) into two parts:

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

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

$$D_I=1.5+\frac{x(1-M^{-1})}{2(1-x)}+[x+\lambda_r+\lambda_a(1-x)]\frac{R}{S} \qquad (6a)$$

$$D_{II}=D_5+D_6+\left(\frac{J-1}{2}+D_5+D_6\right)\cdot\frac{\lambda_r+\lambda_a(1-x)-S}{S} \qquad (6b)$$

The analytical optimization process involves the llowing two steps:

Since D_I is the dominating term, we shall minimize D_I first with respect to α and λ_a under constraints Eqs. 1), (2) and (4).

By using the optimized α and λ_a from step 1, D_{II} is inimized with respect to f_1 and f_2 under constraint q. (3).

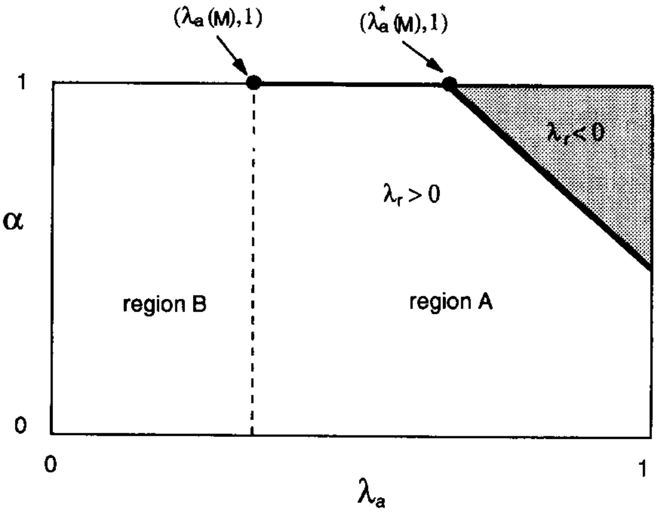
This two step process gives only a sub-optimal solution. It is chosen because simultaneous minimization of $D(\cdot)$ with respect to α, f_1 and f_2 is analytically too difficult. The optimized D_I, denoted as D_I^*, is a natural lower bound of $D(\cdot)$. In Sect. 8, we will show numerically that the 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.

6. Minimization of D_I

Figure 3 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, $d\lambda_r/d\alpha \geqq 0$ at $\alpha=1$. These conditions determine the value of the boundary point $\hat{\lambda}_a(M)$ such that in region A, $\hat{\lambda}_a(M)<\lambda_a\leqq 1$ and in region B, $0\leqq\lambda_a\leqq\hat{\lambda}_a(M)$. 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 $\lambda_r \geqq 0$ and $\alpha=1$, and in region B it is at $(\lambda_a=\hat{\lambda}_a(M), \alpha=1)$. We then show that, for $M\geqq 3$, 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 $M\geqq 3$. For $M<3$, we will show via an example in Sect. 9 that the $\alpha=1$ solution is optimal



—— : the locus of the optimal α in region A.

Fig. 3 The (λ_a, α) space for delay minimization.

only in a restricted range of throughput. Outside that range, delay minimization has to be entirely numerical. The $M \geqq 3$ is the more interesting case because $S_{\max} < 1$ for $M \leqq 2$ (shown in Ref. (6)) while $S_{\max} = 1$ for $M \geqq 3$ (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.

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 Eq. (1) with respect to α and using Eq. (A·4) and Eq. (4), we get

$$\begin{aligned} \frac{d\lambda_r}{d\alpha} &= M\ \frac{d\lambda_m}{d\alpha} - \lambda_a(1-x) \\ &= \frac{M(S-x) - \lambda_a(1-x)(M - xe^{\lambda_m})}{M - xe^{\lambda_m}} \\ &= \frac{\lambda_a(1-x)[xe^{\lambda_m} + Me^{-\lambda_a} - M]}{M - xe^{\lambda_m}} \end{aligned} \tag{7}$$

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

$$xe^{\lambda_m} + Me^{-\lambda_a} - M = 0. \tag{8}$$

Substitute λ_m from Eq. (8) into Eq. (2) and set $\alpha=1$, we obtain

$$\ln\left[\frac{(1-e^{-\lambda_a})M}{x}\right] = 1 - e^{-\lambda_a} + (1-x)\lambda_a e^{-\lambda_a} \tag{9}$$

At a given value of S and M, Eqs. (4) and (9) can be solved simultaneously for λ_a and x which are the required $\hat{\lambda}_a(M)$ and $\hat{x}(M)$.

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 $\lambda_a > \hat{\lambda}_a(M)$, $[\cdot]$ in Eq. (7) is negative at $\alpha=1$. Lemma 4 states that λ_m decreases with α. Hence $[\cdot]$ in Eq. (7) is also negative for $\alpha < 1$. 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 Eq. (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 $\alpha=1$, $\lambda_a \leqq \lambda_a^*(M)$ implies $\lambda_r \geqq 0$ from Lemma 7. Therefore, for a given λ_a, D_I is minimized at $\alpha=1$ by Theorem 1. Using Eq. (1) and setting $\alpha=1$, we obtain D_I as

$$D_I(\lambda_a) = 1.5 + \frac{x(1-M^{-1})}{2(1-x)} + \frac{x + M\lambda_m}{S}\,R \tag{10}$$

To minimize $D_I(\lambda_a)$ with respect to λ_a, Eq. (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 $\lambda_r \geqq 0$. 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 $\lambda_r \geqq 0$ 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 Eq. (4), substituting into (6a) with $\lambda_r = 0$, and differentiating with respect to λ_a, we have

$$\frac{dD_I}{d\lambda_a} = \frac{R(1-S)[1 - e^{-\lambda_a} + \lambda_a e^{-\lambda_a}(1-\lambda_a)]}{S(1-\lambda_a e^{-\lambda_a})^2} - \frac{(1-M^{-1})}{1-S}e^{-\lambda_a}(1-\lambda_a) \quad S < S_c(M). \tag{11}$$

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

$$\begin{aligned} \frac{dD_I}{d\lambda_a} &\geqq \frac{R(1-S)[1 - e^{-S} + Se^{-S}(1-S)]}{S(1-Se^{-S})^2} - e^{-S} \\ &\equiv \phi(S) \quad S < S_c(M). \end{aligned}$$

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

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

For $R \geqq 1$, $\phi(S_c(\infty)) > 0$. 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 $S \geqq S_c(M)$, we set $\lambda_a^*(M) = 1$ and solve for x, λ_m and λ_r simultaneously from Eqs. (1), (4) and (A·5). By substituting them into Eq. (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 Eqs. (A·5) and (A·6) and substitute them into Eq. (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.

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 $M \geqq 3$. Therefore, finding the exact minimum delay in region B is not important because the global minimum delay point for $M \geqq 3$ 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

$$\Lambda_m=\frac{\Lambda_r+\alpha\Lambda_a(1-e^{-\Lambda_a})(1-x)}{M} \qquad (12)$$

The average number of successful reservations per slot x is

$$x=M[\text{av. no. of successful reservation in a minislot}]$$
$$=M\Lambda_m e^{-\Lambda_m} \qquad (13)$$

Substituting Eq. (13) into Eq. (4), we have

$$S=M\Lambda_m e^{-\Lambda_m}+(1-M\Lambda_m e^{-\Lambda_m})\Lambda_a e^{-\Lambda_a} \qquad (14)$$

From Eq. (6a), we obtain D_I as

$$D_I(\Lambda_a)=1.5+\frac{x(1-M^{-1})}{2(1-x)}+\frac{x+\Lambda_r+\Lambda_a(1-x)}{S}R. \qquad (15)$$

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 Eqs. (4) and (12) and setting the optimal value of $\alpha=1$, Eq. (15) becomes

$$D_I(\Lambda_a)=1.5+\frac{x(1-M^{-1})}{2(1-x)}+\frac{M\Lambda_m+S}{S}R. \qquad (16)$$

To minimize $D_I(\Lambda_a)$, Eq. (16) stipulates that x and Λ_m should both be as small as possible. Lemmas 1 and 15 state that Λ_a should be as large as possible. Therefore, the delay is minimized at $\Lambda_a=\hat{\lambda}_a(M)$ and $\alpha=1$. Q. E. D.

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 $M \geqq 3$.

Proof:

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

First, we consider region A. From Eq. (10) we obtain the minimum delay in this region as

$$D_I(\lambda_a=\lambda_a^*(M))=1.5+\frac{x^*(M)(1-M^{-1})}{2(1-x^*(M))}+\frac{x^*(M)+M\lambda_m}{\mathrm{S}}R \qquad (17)$$

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

Next, we consider region B. Since $\hat{x}(M)>M\Lambda_m$ from Eq. (13), we obtain the minimum delay in this region from Eq. (16) as

$$D_I(\Lambda_a=\hat{\lambda}_a(M))>1.5+\frac{\hat{x}(M)(1-M^{-1})}{2(1-\hat{x}(M))}+\frac{\hat{x}(M)+S}{S}R \qquad (18)$$

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,

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

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

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

(ii) The proof for $S\geqq S_c(\infty)$ is similar. Q. E. D.

7. Minimization of D_{II}

From Eq. (6b), we can see that minimizing D_{II} is equivalent to minimizing D_5+D_6 where

$$D_5+D_6=x(1-f_2)\frac{I-1}{2(1-x)}+\left[(1-x)\alpha(1-f_1)+\frac{\lambda_r}{\lambda_r+\lambda_a(1-x)}\right]\frac{K-1}{2M} \qquad (19)$$

Substituting f_2 from Eq. (3) into Eq. (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 $f_2 \leqq 1$ as governed by Eq. (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 $I=J=K=10$. Figures 4 and 5 show the average delay of the UCA protocol,(3) the Controlled Multiaccess protocol,(5) 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^*.

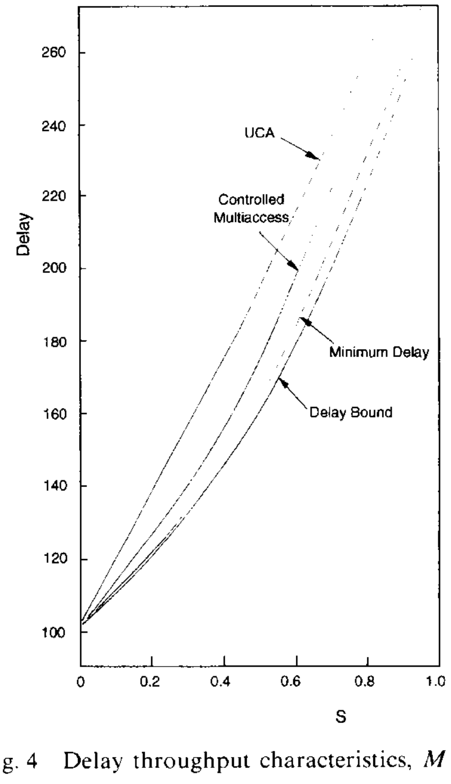


Fig. 4 Delay throughput characteristics, $M=3$.

Figure 6 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 $S \leqq 0.95$, ten minislots per slot is sufficient to give a near optimal performance.

9. Conclusions

The minimum delay protocol designed in this paper is under the assumptions of Poisson arrivals and single copy transmission. Steady state analysis is used to obtain the optimal protocol parameters. For cor-

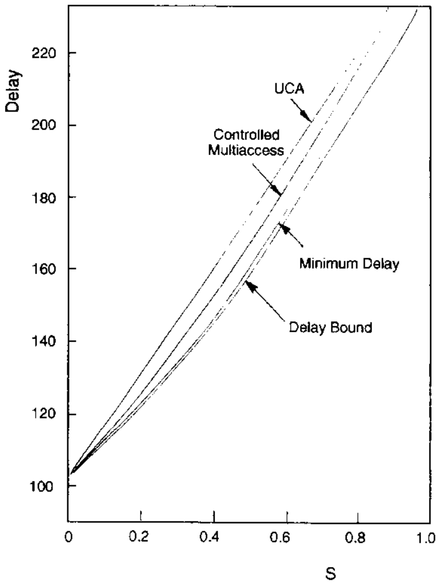


Fig. 5 Delay throughput characteristics, $M=6$.

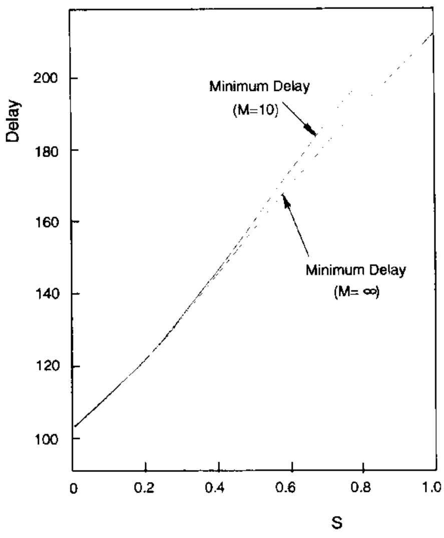


Fig. 6 $M=10$ is quite sufficient for near-optimal delay performance.

related 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 this paper. 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.

References

(1) Bose, S. and Rappaport, S. S., "High Capacity: Low Delay Packet Broadcast Multiaccess," *IEEE Trans. Aerosp. & Electron. Syst.* vol. AES-16, no. 6, pp. 830-838, Nov. 1980.

(2) Chang, J. F. and Lu, L. Y., "Distributive Demand-Assigned Packet Switching with Trailer Transmissions," *IEEE Trans. Aerosp. & Electron. Syst.* vol. AES-20, no. 6, pp. 775-787, Nov. 1984.

(3) Lee, H. W. and Mark, J. W., "Combined Random/Reservation Access for Packet Switched Transmission over a Satellite with On-Board Processing: Part I-Global Beam Satellite," *IEEE Trans. Commun.* vol. COM-31, no. 10, pp. 1161-1171, Oct. 1983.

(4) Yum, T. S. P. and Wong, E. W. M., "The Scheduled-Retransmission Multiaccess Protocol for Packet Satellite Communication," *IEEE Trans. Inf. Theory*, vol. IT-35, no. 6, pp. 1319-1324, Nov. 1989.

(5) Wong, E. W.M. and Yum, T. S.P., "The Controlled Multiaccess Protocol for Packet Satellite Communication," *IEEE Trans. Commun.*, vol. COM-39, no. 7, Jul 1991.

(6) Wong, E. W. M. and Yum, T. S. P., "Delay Bounds for Packet Satellite Protocols," *IEEE GLOBECOM '89*, Nov. 1989.

(7) Kleinrock, L. and Lam, S. S., "Packet-Switching in a Multi-Access Broadcast Channel: Performance Evaluation," *IEEE Trans. Commun.* vol. COM-23, pp. 410-423, Apr. 1975.

(8) Kleinrock, L., *Queueing Systems, Vol. 1: Theory*, John Wiley, pp. 191, 1975.

Appendix

Lemma 1: x and λ_a are inversely related.
Proof: This follows from differentiating Eq. (4).
Q. E. D.

Lemma 2: For fixed α, λ_a and x,

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

(iii) $\frac{d\lambda_r}{dM}<0$.

Proof:
(i) Solving for $(1-x)\lambda_a e^{-\lambda_a}$ in Eq. (4) and substituting into Eq. (2), we have

$$xe^{\lambda_m}=M\lambda_m-\alpha(S-x). \qquad (\text{A}\cdot 1)$$

For fixed α, λ_a and x, differentiating $M\lambda_m$ in (A・1) with respect to M, we have

$$\frac{d(M\lambda_m)}{dM}=\frac{-\lambda_m xe^{\lambda_m}}{M-xe^{\lambda_m}} \qquad (\text{A}\cdot 2)$$

Since $S>x$ (from Eq. (4)), we have from Eq. (A・1).

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

Substituting into the denominator of Eq. (A・2), we obtain

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

(ii) Differentiating λ_m in Eq. (A・1) with respect to M, we obtain

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

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

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

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 Eq. (4)) and λ_m (from Eq. (1)) into Eq. (A・1), set $\alpha=1$ and $\lambda_r=0$, and then differentiate with respect to M, we have $d\lambda_a/dM<0$. Differentiating Eq. (4), we have

$$\frac{dx}{dM}=\frac{-(1-x)(1-\lambda_a)e^{-\lambda_a}}{1-\lambda_a e^{-\lambda_a}}\frac{d\lambda_a}{dM} \qquad (\text{A}\cdot 3)$$

Differentiating $(x+M\lambda_m)$ using Eq. (1) and substituting by Eq. (A・3), we have

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

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

Lemma 4: For fixed λ_a, x and M, $d\lambda_m/d\alpha>0$.
Proof: Differentiating Eq. (A・1) with respect to α, we have

$$\frac{d\lambda_m}{d\alpha}=\frac{S-x}{M-xe^{\lambda_m}}>0. \qquad (\text{A}\cdot 4)$$

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 $\lambda_a=\lambda_m=1$.
(iii) the minimum M (denoted as M^*) for maximum throughput is $M^*=e$.
Proof: (i) and (ii): Setting $\alpha=1$ in Eq. (2) and solve for x, we have

$$x = \frac{M\lambda_m - \lambda_a e^{-\lambda_a}}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \qquad (\mathrm{A}\cdot 5)$$

Substituting into Eq. (4), we obtain

$$S = \frac{M\lambda_m + (e^{\lambda_m} - M\lambda_m - 1)\lambda_a e^{-\lambda_a}}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \qquad (\mathrm{A}\cdot 6)$$

where $\lambda_a \leqq 1$ and $\lambda_m \leqq 1$. By differentiating Eq. (A·6) 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 Eq. (A·6) 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 Eq. (A·5) into Eq. (1) and solving for λ_r, we have

$$\lambda_r = \frac{M\lambda_m e^{\lambda_m} - M\lambda_m \lambda_a e^{-\lambda_a} - \lambda_a e^{\lambda_m} + M\lambda_a \lambda_m}{e^{\lambda_m} - \lambda_a e^{-\lambda_a}}. \qquad (\mathrm{A}\cdot 7)$$

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 Eq. (A·7) 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, $[\cdot]$ 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: $\lambda_a \geqq S$ at $\lambda_r = 0$.

Proof: Substituting Eq. (1) into Eq. (2) and then into Eq. (4) and setting $\lambda_r = 0$, we have

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

Since $(1-x) \leqq 1$ and $[\cdot] \leqq 1$, we have $\lambda_a \geqq S$. Q. E. D.

Lemma 12: $\hat{x}(M)$ is a monotonically increasing function of M.

Proof: From Eq. (8), we have

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

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

Lemma 13: For $M \geqq e$ 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, $[\cdot]$ in Eq. (7) is negative at $\lambda_a = \lambda_a^*(e)$. Substituting Eqs. (1) and (2) into $[\cdot]$ in Eq. (7), we have

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

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

Lemma 14: For a given Λ_a, Λ_r is minimized at $\alpha = 1$.

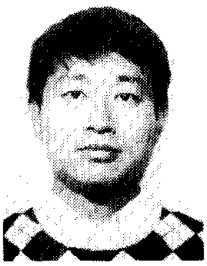
Proof: Substituting Eq. (12) into Eq. (13) and differentiating with respect to α, we have

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

Lemma 15: Λ_m and Λ_a are inversely related.

Proof: It follows from differentiating Eq. (14) with respect to Λ_m and Λ_a. Q. E. D.

Eric W. M. Wong was born in Hong Kong on May 28, 1963. He received the B.Sc. and M.Phil. degrees from The Chinese University of Hong Kong, Shatin, in 1988 and 1990, respectively. He is currently pursuing the Ph.D. degree at the University of Massachusetts, Amherst. His research interest is in satellite communications, telecommunications networks and dynamic routing. Mr. Wong received First Prize in the 1988 IEEE Hong Kong Section Student Paper Contest.



Tak-Shing Peter Yum received his BSc, MSc, MPh and PhD from Columbia University in 1974, 1975, 1977 and 1978 respectively. He worked in Bell Telephone Laboratories, USA for two and a half years and taught in National Chiao Tung University, Taiwan, for 2 years before joining The Chinese University of Hong Kong in 1982. He has published original research on packet switched networks with contributions in routing algorithms, buffer management, deadlock detection algorithms, message resequencing analysis and multiaccess protocols. In recent years, he branched out to work on the design and analysis of cellular network, lightwave networks and video distribution networks. He believes that the next challenge is designing an intelligent network that can accommodate the needs of individual customers.