ANALYSIS OF CONCENTRATED ALOHA SATELLITE LINKS*

Mart L. Molle and Leonard Kleinrock

Computer Science Department
University of California at Los Angeles
Los Angeles, California 90024

Abstract

A conventional ALOHA satellite link uses a transponder which blindly echoes all up-channel traffic on the down-channel. An ALOHA channel can never be fully utilized, so an intelligent satellite could statistically multiplex the successful packets from several slotted ALOHA up-channels onto a single down-channel to conserve bandwidth, and hence reduce cost. We refer to this as a concentrated ALOHA system. Throughput, delay and stability effects are considered, varying the number of up-channels per down-channel and the satellite buffer size. Up- and down-channel bandwidths are assigned independent linear costs, and all performance comparisons are between constant cost systems. It is shown that the marginal increase in system performance drops off so quickly that a small number of up-channels maximizes throughput if up-channel bandwidth has a non-zero cost. This small number is a function of the buffer size and the relative cost of up- to down-channel bandwidth. It is also shown that, even if satellite buffer space is free, a small buffer minimizes average delay for some previously studied protocols of this type. A new protocol which improves performance and allows a large buffer to be used effectively is introduced and analyzed. Solving for throughput and delay in concentrated ALOHA systems provides new analytic and numeric results for the G/D/1 queue with rest period equal to the service time.

1. Introduction

It is well known that message collisions limit the achievable throughput on a conventional ALOHA satellite link. Under the ALOHA protocol [ABRA 73], ground stations transmit without regard for other stations. Each new packet is sent as soon as it is generated; a packet that is not acknowledged within some timeout interval is assumed to have collided with another transmission. Lost packets are retransmitted after a further random delay (in order to reduce the risk of repeated collisions), and so on. One can easily show [ABRA 73, ROBE 72] that for an infinite population of sources, there is a fundamental limit to the channel utilization of $\frac{1}{2e}$ for unslotted ALOHA, and $1/e$ for slotted ALOHA. The problem is, of course, that distributed ground stations are unable to exchange control information that would allow them to coordinate their transmissions on the multi-access up-channel. However, the broadcast down-channel is used exclusively by the satellite. Collisions and unused slots need not be transponded by an intelligent satellite, so a down-channel of lower capacity may be used. This saving in down-channel bandwidth can be used to increase the bandwidth of the up-channel, which is the bottleneck in this system. For this reason, a number of authors [SPAN 78, DERO 78a,b] have suggested building intelligent satellites which can handle n up- and m down-channels on the same logical link, and which can provide a buffer for packets in the satellite.

DeRosa [DERO 78a] found the throughput and average delay in a system with n slotted ALOHA up-channels and a single down-channel, with the same capacity on all $n+1$ channels. The satellite processor could identify error-free packets in the up-channel traffic and would repeat one error-free packet on the broadcast down-channel whenever at least one such packet arrived in a time slot. No buffering at the satellite was considered; if more than one error-free packet arrived in a slot, all but one would be discarded. DeRosa et al [DERO 78b] later solved the state transition matrix numerically for the case of n up-channels, one down-channel and a small satellite buffer. They derived expressions equivalent to Eqs (2) (for throughput) and (7) (for delay) given below, independently of these authors, and showed that increasing either the number of up-channels or the size of the satellite buffer would increase the down-channel utilization. Since they assumed that up-channels are free, they concluded that a large number of up-channels should be used. Below, we generalize their model, improve the protocol, and provide analytic results.

Spaniol [SPAN 78] modelled a less intelligent satellite, which could recognize idle and busy channels, but could not distinguish between error-free packets and collisions. His model includes multiple up- and down-channels and a buffer in the satellite. However, he solved for throughput only for such simple cases as equal numbers of up- and down-channels (when no queue can form), and one less down- than up-channel (but only as a function of the buffer overflow probability, for which he gives no expression). He considered throughput and control of the system, but not delay. Throughput will always be lower than in DeRosa's model, since Spaniol's satellite may choose to accept a collision when an error-free packet is present.

2.The Model

Let us define an *n-concentrated ALOHA satellite link* to be a single logical communications link composed of a set of n identical multi-access up-channels connected to a single broadcast down-channel through a processing satellite. The satellite contains a store-and-forward buffer which may have infinite storage, or be limited to a maximum queue size of $B \geqslant 0$ packets, not including the packet currently being transponded. It is assumed that the satellite can distinguish error-free packet transmissions from both empty slots and collisions. Since the packets pass through the satellite buffer, the processor can perform error checking on each packet and immediately discard all packets that were not correctly received.

The satellite broadcast down-channel is the most power-limited resource in the system, so we shall define the utilization of a satellite link to be the steady-state probability that each satellite broadcast slot contains an error-free packet. Slots are synchronized so that one *period* contains exactly n complete up-channel slots (one per up-channel) and one complete down-channel slot. Either FDMA or TDMA may be used to split the bandwidth for the up-channels. If FDMA is used, ground terminals must randomly select one of the n channels in the chosen time slot [DERO 78]. A set of

* This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract MDA 903-77-C-0272, and by the Regents of the University of California through a Chancellor's Intern Fellowship.

FDMA channels is equivalent to a bulk-server system with service time equal to one period. A set of TDMA channels is equivalent to a single high-speed channel running n times faster than the period. TDMA makes buffer management particularly easy, because a complete packet arrives at the satellite on one channel (and can be checked for errors) before another packet arrives on the next channel. Ground stations maintain up-channel synchronization by using the satellite as a master clock at the period frequency. Hence, the satellite must broadcast "empty" packets whenever its buffer is exhausted.

We assume that there is an infinite population of terminals generating Poisson traffic at a combined rate G packets per period and transmitting under the slotted ALOHA access scheme without capture. Terminals do not distinguish between the up-channels,[1] so that the load on each up-channel will be G/n packets per slot. In this case, the probability of a success reaching the satellite in any slot will be the probability that exactly one terminal chooses to transmit in that slot, which is

$$s = \frac{G}{n} e^{-G/n} \qquad (1)$$

Since there are n up-channels, and each up-channel slot independently carries a successful packet with probability s, up to n packets may arrive in a period. We view these successful packets as a bulk arrival process at the satellite, with a binomial distribution of bulk size, whether the up-channels are implemented as parallel FDMA channels (when there really would be a bulk arrival) or as sequential TDMA channels.

It is worth mentioning at this point that our analysis depends on the assumption of using slotted-ALOHA random access for the up-channels only through Eq. (1). It is equally valid for any other infinite population random access scheme which cannot achieve full use of the up-channel. Of course, a processing satellite will provide less improvement if some other, more efficient access scheme is used.

Unlike previous work, we do *not* assume that the extra resources needed to support multiple up- and down-channels are free. In order to have a fair comparison with conventional slotted ALOHA links, we assign linear costs to bandwidth: we assume than an up-channel is X times as expensive as a down-channel of the same capacity, $0 \leqslant X < \infty$. Both concentrated and conventional ALOHA satellite links require one satellite broadcast channel. However, concentrated ALOHA uses n multi-access up-channels where conventional ALOHA uses only one. Performance comparisons are only made between constant cost systems. Hence, each channel in a concentrated ALOHA link will have a capacity that is only $\frac{X+1}{nX+1}$ times the channel capacity in a conventional ALOHA satellite link.

3. The G/D/1 Queue

We may view the satellite broadcast channel as a server and the store-and-forward buffer in the satellite as its queue. Throughput and delay on a concentrated ALOHA satellite link depend on the behavior of this queueing system: throughput depends on the probability of an empty buffer, and delay on the average queue size in the buffer. The arrival process in concentrated ALOHA is the sum of n independent identically distributed Bernoulli trials, with probability of success governed by Eq. (1), giving a binomial distribution of the number of arrivals per service time. Below we solve the more general problem where the number of arrivals per service time has an arbitrary distribution.

Consider the G/D/1 queue with constant rest period equal to the (constant) service time. The server is available only at slot boundaries; if no customer is waiting for service at such a boundary, the server 'takes a rest' until the next one. Erlich [ERLI 76] derived an expression for the z-transform for number in system for the G/D/1 queue with bulk service of up to m customers, constant rest period equal to the service time and infinite storage, but did not solve it to give the distribution, or any moments, of the number in system. In fact, her expression contained the unknown terms $p_0, \cdots, p_{m-1}$ explicitly, and hence cannot be solved analytically except for $m=1$. Below we present a much simpler derivation, equivalent to her result for $m=1$, and, in particular, extend it to the finite storage case. We also find explicit expressions for the average queue length with infinite storage, and for $\{p_k\}$ in some cases.

Let us define p_k, with z-transform $P(z)$, to be the equilibrium probability of k customers in the queue (*not* including any customer in service) just after the start of a service period[1], and analyze queue length as an imbedded Markov chain [KLEI 75a]. Note that ignoring customers in service simplifies the analysis by combining the two boundary equations for 0 or 1 in the system into a single, simpler boundary equation for 0 in the queue. Let us also assume an independent arrival process, where v_k is the probability of k arrivals in a service time, $V(z)$ is its z-transform, and $\bar{v}$ is its expected value. Since the server will be idle for a service period only if the queue was empty at the previous imbedded point and there were no new arrivals, the utilization must be

$$\rho = 1 - p_0 v_0 \qquad (2)$$

Let us begin with the *infinite storage* case. If we define the discrete convolution $f \circledast g$ of two non-negative probability density functions to be the sequence whose n^{th} term is

$$(f \circledast g)_n \triangleq \sum_{j=0}^{n} f_j g_{n-j}$$

then the balance equations for $\{p_k\}$ are simply

$$p_0 = (p \circledast v)_1 + (p \circledast v)_0 \qquad (3a)$$

$$p_k = (p \circledast v)_{k+1} \quad k \geqslant 1 \qquad (3b)$$

Using the convolutional property of z-transforms [KLEI 75a], we see that

$$\begin{aligned} P(z) &= \sum_{k=0}^{\infty} (p \circledast v)_{k+1} z^k + (p \circledast v)_0 \\ &= \frac{P(z)V(z) - (p \circledast v)_0}{z} + (p \circledast v)_0 \\ &= \frac{(p \circledast v)_0 (z-1)}{z - V(z)} \end{aligned}$$

With one application of l'Hôpital's rule, we find $P(1) = 1 = \frac{(p \circledast v)_0}{1-\bar{v}}$. Thus, in the infinite storage case, $\rho = \bar{v}$ from Eq. (2) (and hence $\bar{v}$ must not exceed one for stability), and in equilibrium

$$P(z) = \frac{(1-\bar{v})(z-1)}{z - V(z)} \qquad (4a)$$

[1] If the up-channels are TDMA and the satellite has a small buffer, a question of fairness arises. Since the buffer can never be full just after the satellite finishes transponding another packet, packets successfully arriving on the first TDMA channel can never be blocked, but there is a positive, increasing blocking probability on the following channels. Individual terminals will thus receive better service if they disregard the randomization. The question of channel-dependent traffic will not be considered here.

[1] Since we assume bulk arrivals at the end of each service period (i.e. FDMA channels), $\{p_k\}$ are valid for all time.

With *finite storage* limited to B in the queue, we must add an additional boundary equation to account for blocking as the queue overflows:

$$p_B = (p \circledast v)_{B+1} - p_{B+1}v_0 + \sum_{i=0}^{B} p_i \sum_{j=B+2-i}^{\infty} v_j \tag{3c}$$

Since $p_k \triangleq 0 \; \forall k > B$, we can include these redundant equations at will, as long as we are careful to subtract out the non-zero terms in the convolution (which were absorbed into the queue overflow probability):

$$p_k = (p \circledast v)_{k+1} - \sum_{i=0}^{B} p_i v_{k+1-i} = 0 \qquad k > B \tag{3d}$$

Eqs (3a) - (3d) define an infinite convolution with some boundary terms, so we can once again transform Eqs (3) to obtain

$$P(z) = \sum_{k=0}^{\infty} (p \circledast v)_{k+1} z^k + (p \circledast v)_0 + \sum_{i=0}^{B} p_i \sum_{j=B+2-i}^{\infty} v_j \left[z^B - z^{i+j-1} \right]$$
$$= \frac{(p \circledast v)_0 (z-1) + \sum_{i=0}^{B} p_i \sum_{j=B+2-i}^{\infty} v_j \left[z^{B+1} - z^{i+j} \right]}{z - V(z)} \tag{4b}$$

Eq. (4b) is not in a useful form, however, since $\{p_k\}$ are explicitly included. One could, of course, apply boundary conditions to find these unknowns, but that is tedious for large B. Since $P(1) \triangleq 1$, Eq. (4b) gives us the additional relation (after one application of l'Hôpital's rule)

$$1 - \bar{v} = (p \circledast v)_0 + \sum_{i=0}^{B} p_i \sum_{j=B+2-i}^{\infty} v_j [B+1-i-j]$$
$$= p_0 v_0 + \sum_{i=0}^{B} p_i \left[B+1-i-\bar{v} - \sum_{j=0}^{B-i} v_j [B+1-i-j] \right]$$
$$\therefore v_0(p_0 - 1) = \sum_{i=0}^{B-1} p_i \left[i + \sum_{j=0}^{B-i} v_j [B+1-i-j] - B - v_0 \right]$$

and we see that $p_0 = 1$ if $B=0$, and $p_0 = \frac{v_0}{1-v_1}$, $p_1 = 1 - \frac{v_0}{1-v_1}$ if $B=1$. We present the following alternate method for finding $\{p_k\}$ directly from Eqs (3) in the finite buffer case by "unravelling" the convolution. For convenience, let us define $\phi \triangleq \frac{v_0}{1-v_1}$ as the probability of no arrivals in a service time, conditioned on not having exactly 1 arrival.

Lemma 1:

Let $\quad F_k \triangleq \sum_{i=0}^{k} p_i \quad k \geqslant 0$

then $\quad F_k = \frac{F_0}{\phi^k} \left[1 - \sum_{j=2}^{k} \frac{v_j}{v_0} \phi^j \sum_{i=0}^{k-j} \frac{F_i}{F_0} \phi^i \right]$

proof: Substituting Eqs (3a) and (3b):

$$F_k = \sum_{i=1}^{k} \left[\frac{p_{i-1}}{\phi} - \sum_{j=2}^{i} \frac{v_j}{v_0} p_{i-j} \right] = \frac{F_{k-1}}{\phi} - \sum_{j=2}^{k} \frac{v_j}{v_0} F_{k-j}$$

By recursive substitution F_k may be expressed in terms of F_0 and other lower order sums:

$$F_k = \frac{F_0}{\phi^k} - \sum_{i=0}^{k-2} \phi^{-i} \sum_{j=2}^{k-i} \frac{v_j}{v_0} F_{k-i-j} = \frac{F_0}{\phi^k} \left[1 - \sum_{j=2}^{k} \frac{v_j}{v_0} \phi^j \sum_{i=0}^{k-j} \frac{F_i}{F_0} \phi^i \right]$$

□

Except for the rightmost summation, Lemma 1 expresses F_k in terms of F_0 and known quantities. However, this summation may be solved recursively:

Lemma 2:

Let $\quad \Lambda(k) \triangleq \sum_{i=0}^{k} \frac{F_i}{F_0} \phi^i$

then $\quad \Lambda(k) = k + 1 - \sum_{j=2}^{k} \frac{v_j}{v_0} \phi^j \sum_{i=0}^{k-j} \Lambda(i)$

proof: (By direct substitution of Lemma 1)

$$\Lambda(k) = \sum_{i=0}^{k} \frac{\phi^i}{F_0} \cdot \frac{F_0}{\phi^i} \left[1 - \sum_{j=2}^{i} \frac{v_j}{v_0} \phi^j \Lambda(i-j) \right] = k + 1 - \sum_{j=2}^{k} \frac{v_j}{v_0} \phi^j \sum_{i=0}^{k-j} \Lambda(i)$$

□

The following result follows immediately from Lemmas 1 and 2 and the observations that $F_B = 1$, and $p_k = F_k - F_{k-1}$:

Theorem 1:

$$p_0 = \frac{\phi^B}{1 - \sum_{j=2}^{B} \frac{v_j}{v_0} \phi^j \Lambda(B-j)}$$

$$p_k = \frac{p_0}{\phi^k} \left[1 - \phi - \sum_{j=2}^{k} \frac{v_j}{v_0} \phi^j \left[\Lambda(k-j) - \phi \Lambda(k-1-j) \right] \right] \qquad 1 \leqslant k \leqslant B$$

□

Theorem 1 provides an efficient numerical procedure for finding $\{p_k\}$, given $\{v_k\}$, n and B. Once $\Lambda(\cdot)$ has been calculated as a triangular set from Lemma 2, any of the p_k's can be computed independently using only known quantities.

4. Throughput in n-Concentrated ALOHA

Let S be the throughput in packets per period (i.e. the utilization of the satellite broadcast channel). In general, S < G, the offered load, because some packets are lost in ALOHA collisions on the up-channels (see Eq. (1)) or blocked at the satellite when its buffer overflows.

We may use the queueing results from the previous section to find p_0 and hence S. In this case, the number of arrivals per service time has a binomial distribution with n trials, each with probability of success s given by Eq. (1). Thence $V(z) = [1 + s(z-1)]^n$, $\bar{v} = ns$, and, for infinite storage, we have from Eq. (4a) that

$$P(z) = \frac{1 - ns}{1 - s \sum_{i=1}^{n} \binom{n}{i} [s(z-1)]^{i-1}} \tag{5}$$

and $\mathbf{S} = \min(1, Ge^{-G/n})$. When $n=2$, we have a birth-death process (for the number in queue but *not* for the number in system); Eq. (5) can be easily inverted, and we can immediately write the equilibrium solution for $n=2$ and any buffer size as

$$p_k = \frac{1 - \left(\frac{s}{1-s} \right)^2}{1 - \left(\frac{s}{1-s} \right)^{2(B+1)}} \left(\frac{s}{1-s} \right)^{2k}$$

$$\therefore \mathbf{S} = \frac{2s - \left(\frac{s}{1-s} \right)^{2(B+1)}}{1 - \left(\frac{s}{1-s} \right)^{2(B+1)}} = 1 - \frac{1 - 2s}{1 - \left(\frac{s}{1-s} \right)^{2(B+1)}}$$

In the limit as $B \to \infty$, we see that $\mathbf{S} \to 2s$ if $s < ½$ and $\mathbf{S} \to 1$ if $s \geqslant ½$. We also present closed form results for small systems in Table 1, using the results of Theorem 1.

Figures 1 and 2 show the throughput as a function of offered load for various buffer sizes. In Figure 1, it is assumed that up-channels are available at no cost. One can see that increasing either the number of up-channels or the satellite buffer size improves the system performance at all values of G. The speed with which

Table 1: Throughput with a Small Buffer

B	$S = 1 - p_0(1-s)^n$
0	$1 - (1-s)^n$
1	$1 - \phi\cdot(1-s)^n$
2	$1 - \dfrac{\phi^2\cdot(1-s)^n}{1-\binom{n}{2}\left(\frac{s\cdot\phi}{1-s}\right)^2}$
3	$1 - \dfrac{\phi^3\cdot(1-s)^n}{1-2\binom{n}{2}\left(\frac{s\cdot\phi}{1-s}\right)^2-\binom{n}{3}\left(\frac{s\cdot\phi}{1-s}\right)^3}$

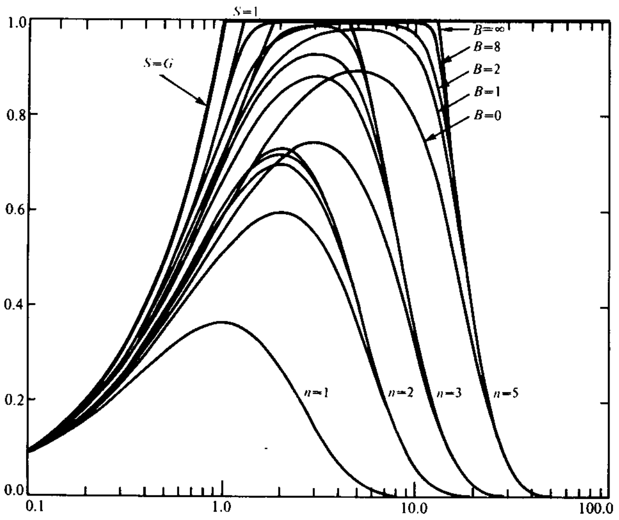

Figure 1: Mean Throughput, Free Uplinks

these families of curves converge to their limiting values is significant. With $B = 8$, throughput is already close to 1.0 with only three up-channels; with five up-channels, throughput remains very close to 1.0 over a broad load range (in fact, for $G < 1.0$, throughput is approaching the limiting case of $S=G$ with no ALOHA collisions and no blocking, i.e. $n=\infty$, $B=\infty$). Performance with three up-channels is most sensitive to buffer size, because three slotted-ALOHA up-channels have barely enough capacity to saturate the broadcast down-channel even with a large buffer to smooth the arrival rate. Two up-channels cannot supply packets fast enough to cause a serious backlog on the down-channel, while four or more up-channels have enough excess capacity to saturate the down-channel even without much buffer space.

When the relative cost of bandwidth X exceeds zero, the capacity of each channel must be reduced as we add more up-channels (recall that slot synchronization within a period implies that all channels are of equal capacity). Since we are using n up-channels and one down-channel for a concentrated ALOHA link, but the total cost must be the same as a conventional ALOHA link that used only one of each, the capacity of each channel must be reduced by a factor of $\frac{X+1}{nX+1}$. For Figure 2, we have assumed that $X = 1$ (i.e. up-channels are just as costly as down-channels). In this case, it is clear from the figures that only $n=2, 3$ can maximize throughput, whatever the buffer size. Although this up-channel cost may be unreasonably high given current technology, its message is clear: one should not commit too many resources to the same satellite link. At some point, the same resources may be reorganized into several simpler systems of greater total capacity. We now pursue this cost-performance tradeoff more fully using analytic methods.

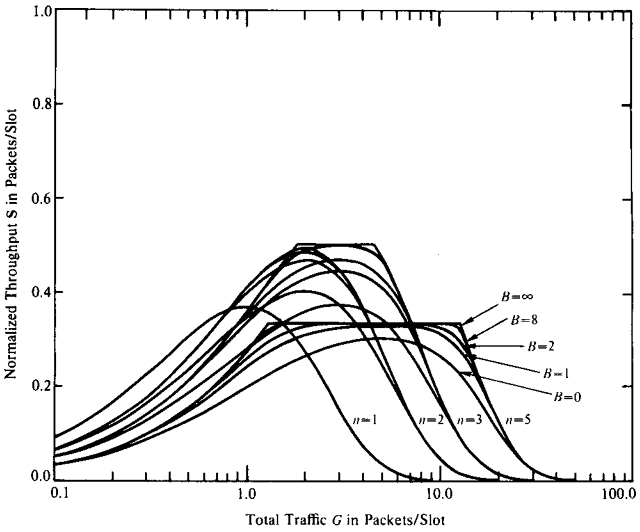

Figure 2: Mean Throughput, All Channels of Equal Cost

A sufficient condition to maximize the total throughput is to maximize s, the expected traffic arriving on each up-channel. From Eq. (1), one can easily see that s takes on its maximum value of e^{-1} when $G = n$. For fixed values of s and (finite) B, S is a monotonically increasing function of n, asymptotically approaching 1 as $n \to \infty$. However, the *normalized* throughput $\hat{S}$ of a constant cost system is approaching a *decreasing* limiting value, which is 0 as $n \to \infty$. Because the total budget is fixed, at some point the marginal gain in throughput from an extra up-channel no longer offsets the resulting loss of capacity per channel, resulting in a net loss of normalized throughput.

Let us fix s and B, and consider $S = S^{(n)}$ as a function of n. The relative up-channel cost $X = X^{(n)}$ where the $n+1^{st}$ channel gives a zero marginal gain in throughput is given by

$$S^{(n+1)}\cdot\frac{1+X}{1+(n+1)X} = S^{(n)}\cdot\frac{1+X}{1+nX}$$

$$\therefore X^{(n)} = \frac{S^{(n+1)} - S^{(n)}}{(n+1)\cdot S^{(n)} - n\cdot S^{(n+1)}} \qquad (6)$$

If the up-channel cost is greater than $X^{(n)}$, then $\hat{S}^{(n)} > \hat{S}^{(n+1)}$; if it is less than $X^{(n)}$, then $\hat{S}^{(n+1)}$ is greater. We lack general closed-form solutions for S, so $X^{(n)}$ must be found numerically. However, certain limiting cases can easily be obtained. Since we do have closed form solutions for S when $n=1$, 2, we can find $X^{(1)}$ explicitly for all buffer sizes:

$$X^{(1)} = \frac{(e-1)^{2(B+1)} - (e-1)}{e-2}$$

The dividing line starts at $e-1$ with no buffer, increasing (approximately exponentially for large B) to ∞ as $B \to \infty$. If $B=0$, $S = 1-(1-1/e)^n$, and we see that

$$X^{(n)}\big|_{B=0} = \frac{1}{e\left(\frac{e}{e-1}\right)^{n} - e - n}$$

and $X^{(n)} \rightarrow 0$ as $n \rightarrow \infty$, approximately exponentially for large n. In the limit as $B \rightarrow \infty$, $\mathbf{S} \rightarrow min(1, n/e)$, and we see that $X^{(1)} \rightarrow \infty$, $X^{(2)} \rightarrow \frac{e-2}{6-2e}$ and $X^{(n)} \rightarrow 0$ for all $n \geqslant 3$. So, for the infinite buffer case, only two or three up-channels can maximize throughput (the dividing line being $X \approx 1.275$). Figure 3 shows the X-B plane divided into regions according to the number of up-channels that maximizes normalized throughput. One can see that n=2, 3 dominate as B increases, with the boundaries spreading exponentially as the buffer size increases.

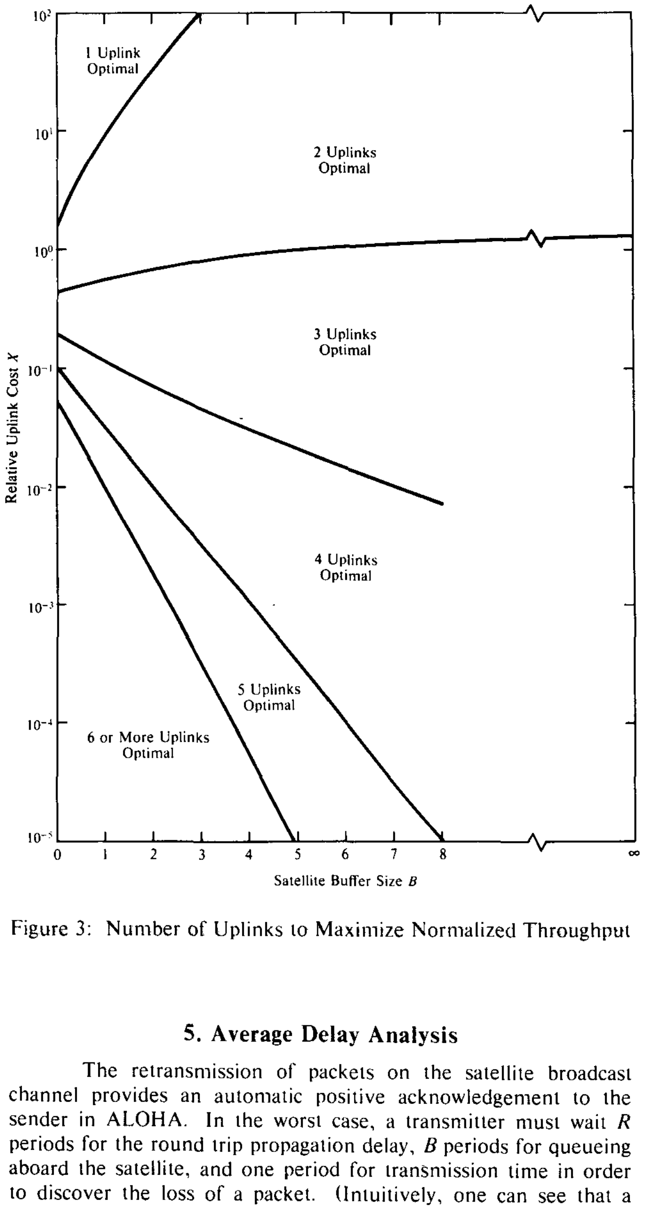

Figure 3: Number of Uplinks to Maximize Normalized Throughput

5. Average Delay Analysis

The retransmission of packets on the satellite broadcast channel provides an automatic positive acknowledgement to the sender in ALOHA. In the worst case, a transmitter must wait R periods for the round trip propagation delay, B periods for queueing aboard the satellite, and one period for transmission time in order to discover the loss of a packet. (Intuitively, one can see that a shorter timeout will actually increase the delay and decrease the throughput under heavy load. Whenever a station incorrectly assumes that its packet was lost when it did in fact successfully enter the satellite buffer, it may choose to transmit an (unnecessary) duplicate packet. If this duplicate is involved in a collision, another packet may needlessly be lost; if it successfully reaches the satellite, it will delay all other packets arriving in its delay busy-period [KLEI 76] (without adding to the throughput) and may cause a new packet to be lost unnecessarily if the satellite buffer overflows during that time.)

Let us begin by analyzing the naive protocol, also treated by DeRosa et al [DERO 78b], where the worst-case acknowledgement timeout of $R+B+1$ periods is always used. Let us assume that the probability of success for each attempt to send a packet is independent[1]. Each packet will be sent an average of $\frac{G}{\mathbf{S}}$ times, and the number of unsuccessful attempts per packet will be geometrically distributed, with mean $\bar{r} = \frac{G}{\mathbf{S}} - 1$. On each unsuccessful attempt, a packet will be delayed by the acknowledgement timeout of $R+B+1$ periods, plus an average of $\frac{K-1}{2n}$ periods for randomizing the time of the next retry over the next K slots. On the successful attempt, the packet will be delayed by $R+1$ slots for propagation and transmission, plus the average time W spent queued in the satellite buffer. Hence the average delay (measured in periods) from transmission of a packet to its successful reception is given by

$$D = R+W+1+\bar{r}\left(R+B+1+\frac{K-1}{2n}\right) \qquad (7)$$

where $W = \frac{\bar{N}_q}{\mathbf{S}}$ from Little's result. The expected queue length $\bar{N}_q$ can always be found from the set $\{p_k\}$ because these probabilities are valid for all time by our late bulk arrival assumption. In the infinite storage case, Eq. (4a) and the identity $\bar{N}_q = P'(z)|_{z=1}$ gives us a closed form expression for W in the G/D/1 case:

$$\bar{N}_q = \frac{\bar{v}}{2}\left[C_v^2 \frac{\bar{v}}{1-\bar{v}} - 1\right]$$

$$\therefore\ W = \frac{1}{2}\left[C_v^2 \frac{\bar{v}}{1-\bar{v}} - 1\right]$$

where $C_v = \frac{\sigma_v}{\bar{v}}$ is the coefficient of variation of the arrival bulk size. If we restrict ourselves to the binomial bulk size distribution, we find $C_v^2 = \frac{1-s}{ns}$, $\bar{v} = ns$, and

$$W = \frac{s(n-1)}{2(1-ns)} \qquad (8)$$

Figures 4 and 5 plot normalized delay as a function of normalized throughput for constant cost systems with two, three or five up-channels. Note that the transmission times and queueing delays (expressed in periods) must be scaled as n increases, but that the round-trip propagation time to the satellite, R, is a constant, independent of any access scheme. When up-channels are free, adding extra up-channels improves performance at all throughput levels. If we must pay for the extra up-channels from a fixed budget, adding extra up-channels is not necessarily a good idea. When all channel costs are equal, either $n=2$ or $n=3$ gives the best performance (although it is difficult to decide between those two).

Equation (7) leads to the surprising conclusion that, *even if satellite buffer space is free*, a very large buffer is undesirable. Whenever there is a positive offered load $G > 0$ and a finite number of up-channels $n < \infty$, there is a non-zero probability of losing packets in ALOHA collisions. Since the protocol always assumes the worst case queueing delay, no attempt to retransmit

[1] Kleinrock and Lam [KLEI 75b, LAM 75] distinguish between new and blocked packets in a more accurate model of the slotted ALOHA satellite channel. Their model explicitly introduces traffic *dependence* among the up-channel slots, and requires numerical solution. The *independence* assumption for up-channel traffic is basic to our analysis, so we shall use our simpler model.

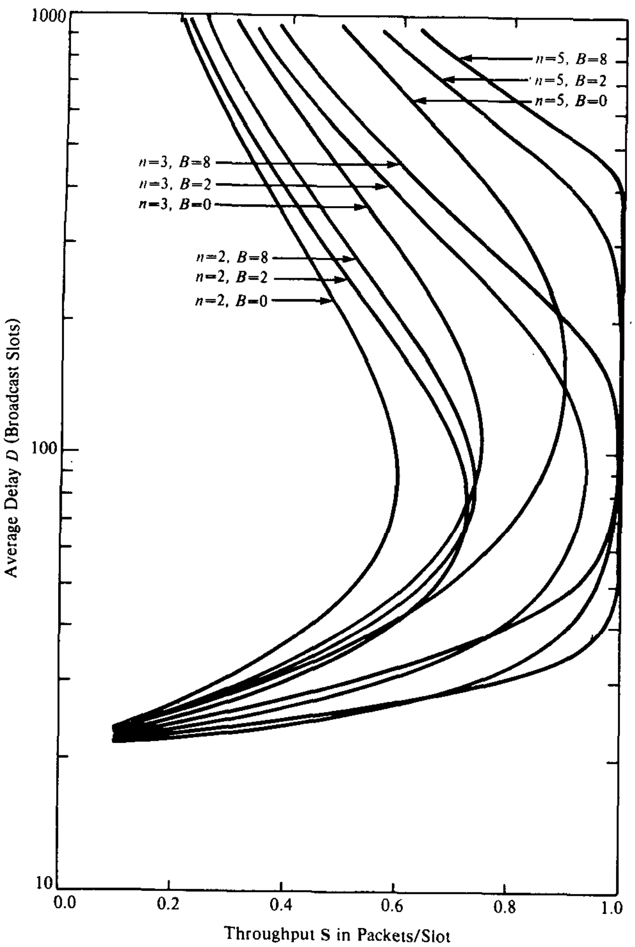

Figure 4: Average Packet Delay, Free Uplinks

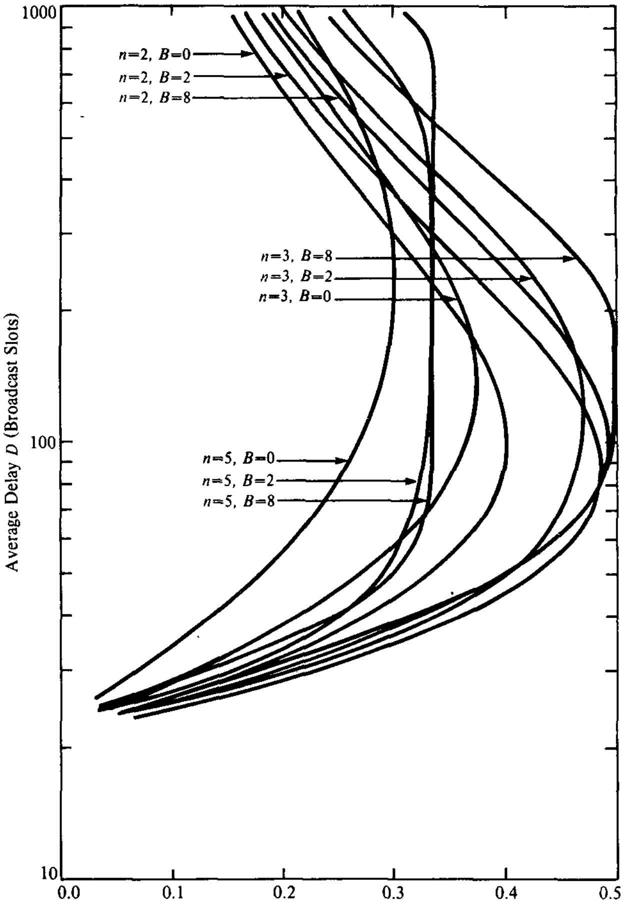

Figure 5: Average Packet Delay, All Channels of Equal Cost

these lost packets will occur within a finite time. Thus, the average delay D is unbounded for an infinite buffer system! This protocol also performs poorly with small buffer sizes. Returning to Figures 4 and 5, we see that $B=2$ gives lower average delays than $B=8$ until the systems approach their (lower) maximum throughput value. Figure 6 makes this even more clear by focusing on $n=3$, the system most sensitive to buffer size. One can see that the delay curves *cross* as throughput increases, with the optimal buffer size increasing with throughput.

For any given system, the optimum buffer size is a complicated function of S and n which we only solved numerically. We offer the following insight into the optimal buffer size as a function of S for fixed n. Clearly, in the limit as $\mathbf{S} \rightarrow 0$, $B=0$ is optimal since there will be no contention for the broadcast channel. As S approaches its maximum (usually 1), we find $W \rightarrow B$, and we may approximate Eq. (7) for high throughput by the following:

$$D = R+W+1+\left(\frac{G}{\mathbf{S}}-1\right)\left(R+B+1+\frac{K-1}{2n}\right)$$
$$\leqslant \frac{G}{\mathbf{S}}\left(R+B+1+\frac{K-1}{2n}\right) = D_u$$

Setting the derivative with respect to B equal to zero, we find an approximation to the optimum value of B:

$$\frac{\partial D_u}{\partial B} = 0 = \frac{G}{\mathbf{S}} + \frac{\partial}{\partial B}\left(\frac{G}{\mathbf{S}}\right)\cdot\left(R+B+1+\frac{K-1}{2n}\right)$$

Let us make the further approximation that G is not a function of B. Although G *is* a function of n, S and B, $W \approx B$ only when $\mathbf{S} \approx 1$; we see from Figure 1 that S is insensitive to small changes in G and B near its maximum value (such changes will only affect the rate at which the buffer overflows). In this case, we find:

$$\frac{G}{\mathbf{S}} \approx \frac{G}{\mathbf{S}^2}\cdot\left(R+B+1+\frac{K-1}{2n}\right)\frac{d\mathbf{S}}{dB}$$
$$\therefore \frac{d\mathbf{S}}{\mathbf{S}} = \frac{dB}{B+R+1+\frac{K-1}{2n}}$$
$$\therefore B = C\cdot\mathbf{S}-\left(R+1+\frac{K-1}{2n}\right)$$

where C is an arbitrary constant. Thus, it is clear that the optimal buffer size should grow no faster than linearly with the required throughput in a heavily utilized system.

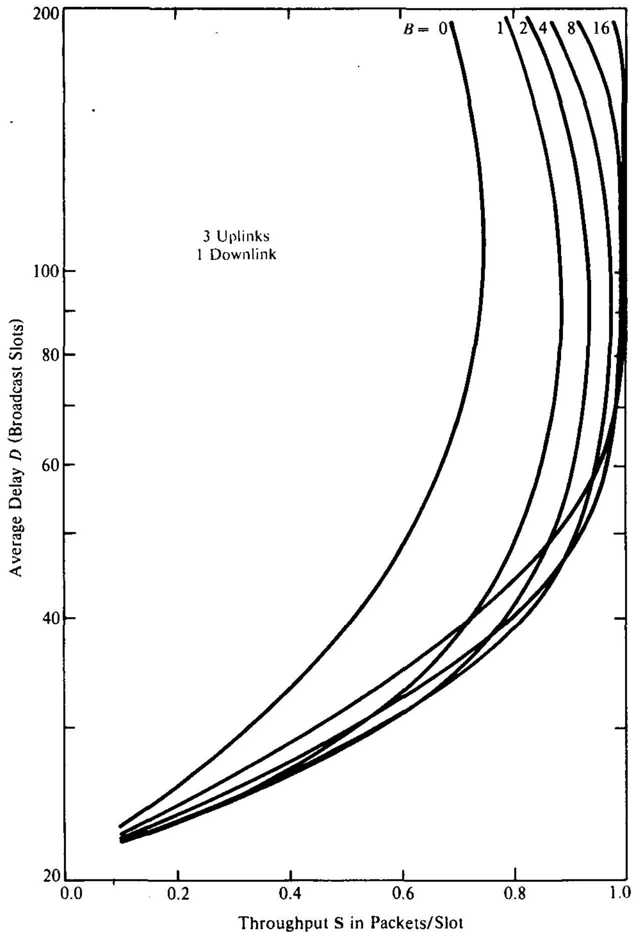

Figure 6: Optimizing Buffer Size for Minimum Delay

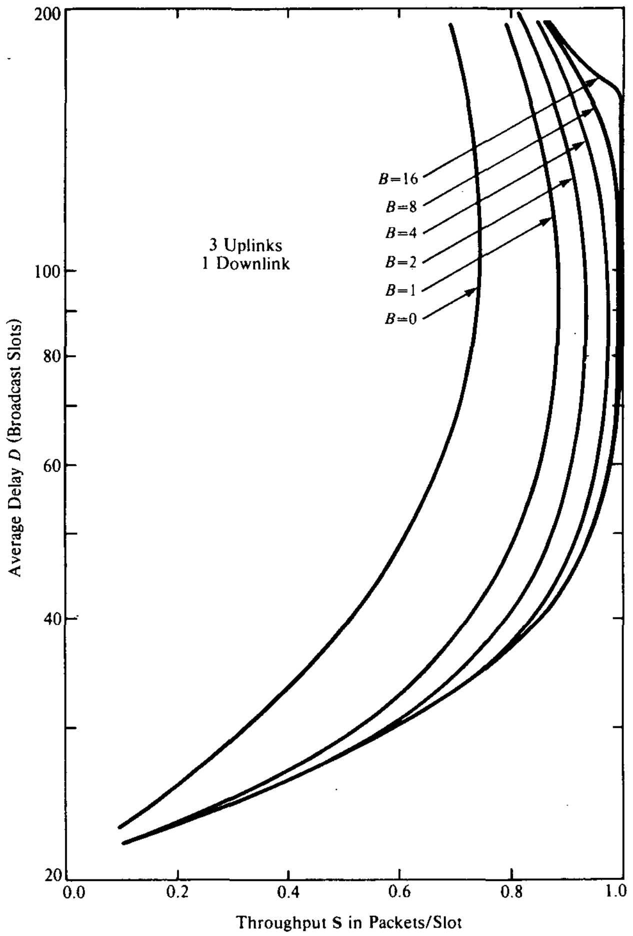

Figure 7: Mean Delay, Packets Carry Time Stamp

6. An Improved Protocol with Time Stamps on Packets

In the previous section, it was shown that the average delay is minimized when the satellite buffer changes size as a function of throughput. The buffer will usually be empty when traffic is light, so a larger buffer only increases the delay between retries. However, a larger buffer is less likely to overflow, reducing the number of retries in heavy traffic.

The following simple protocol change approximates this adaptive buffer size strategy and even outperforms it. Let each packet header carry its arrival time at the satellite, where the arrival time for an "empty packet" is defined to be its time of transmission on the broadcast channel. Although the satellite is the natural place to insert this time stamp, we can implement this protocol without forcing the satellite to do any additional processing. Each ground station must know its propagation time to the satellite to ensure that its transmissions reach the satellite within slot boundaries, so each ground station could fill in the time stamp in advance. The time stamp in "empty packets" is redundant, since the buffer must be empty by definition whenever an "empty packet" is transmitted.

Since each ground station can calculate the time at which its packet should have reached the satellite, it will know that its packet was lost as soon as *any* packet with a later time stamp is retransmitted, assuming FCFS at the satellite. Ground stations receive a positive or negative acknowledgement in the same average time, and the average (unnormalized) delay is reduced to

$$D = \frac{G}{\mathbf{S}} \cdot \left[R + W + 1 \right] + \bar{r} \cdot \left(\frac{K-1}{2n} \right) \tag{9}$$

Figure 7 shows the throughput and delay for the same set of systems as Figure 6, except for the protocol change. It is assumed that the overhead from adding an extra field to the packet header is insignificant. Comparing Eqs (7) and (9), we expect this protocol to reduce average delay by $\bar{r}(B-W)$ over the naive protocol. This is a significant improvement over a broad load range, since $\bar{r}$ is large under heavy load, and $B-W$ is large under light load. Figure 7 shows a real improvement over Fig. 6: the individual curves from the new protocol show considerably better performance than the individual curves from the original protocol, and even show some improvement over the optimum performance envelope obtained by varying the buffer size with the original protocol. It is also clear from Figure 7 that we now have a more "normal" system in the sense that performance always improves as the satellite buffer size increases. Using the results for maximizing throughput, we see that, if up-channels are no more expensive than down-channels, three up-channels with a large buffer will give us the most cost effective system. If up-channels are significantly more expensive than down-channels, use two up-channels instead of three.

7. Stability and Flow Control

Kleinrock and Lam [KLEI 75b, LAM 75] demonstrated the inherent instability of a regular ALOHA channel. Two equilibrium values exist for the average delay at any feasible throughput, with a region of instability beginning at the higher one. A slotted ALOHA channel will fail within a finite time because random load fluctuations will drive the system into the unstable region with probability one. The concept of "load lines" on the $\mathbf{S}-N_t$ plane is introduced to show how the average number of blocked ground stations affects the load on the system. Using a fluid approximation technique, they show that all infinite population ALOHA systems are unstable, but a finite population ALOHA system will be stable if the load line does not touch the unstable region.

We can apply this same analysis to the protocol introduced in the previous section. Unlike the previous work, where N_t was found from the steady state analysis of a Markov chain, we shall find the average number of blocked ground stations from Little's Result as

$$N_t = \mathbf{S}\cdot D = G\cdot(R+W+1)+\bar{\tau}\mathbf{S}\cdot\frac{K-1}{2n}$$

We shall use D instead of N_t in the analysis, since the shapes of the two curves plotted against $\mathbf{S}$ are about the same.

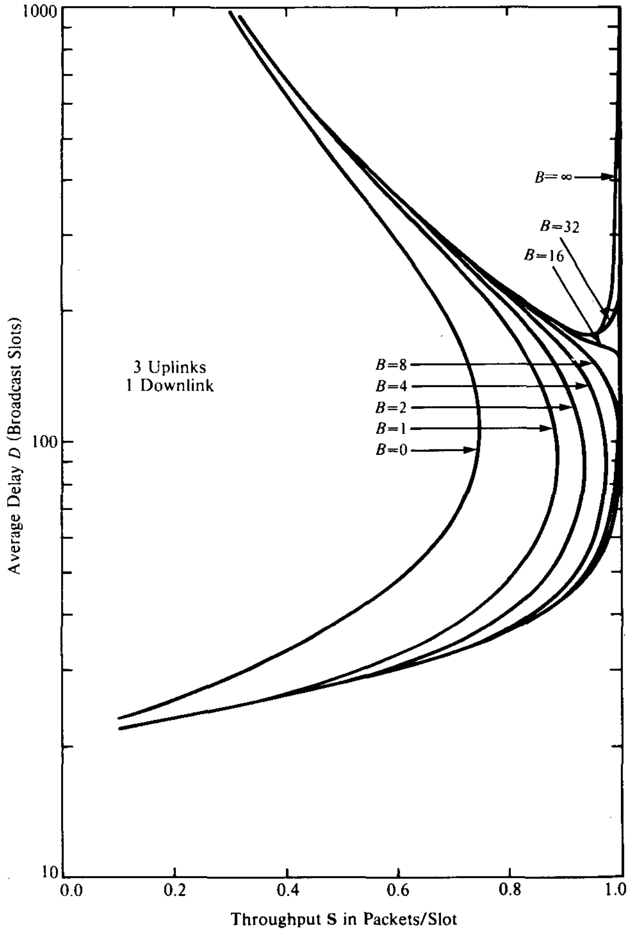

Figure 8: Anomalous Behavior with Very Large Buffers

Figure 7 showed little of the upper branches of the delay curves for this improved concentrated ALOHA protocol. However, it is clear that the delays for the *upper* branch of each curve are significantly reduced as well. Figure 8 increases the range of delay shown and also includes the $B=32, \infty$ curves. For very large buffer sizes, this protocol behaves in a most counter-intuitive way! As G approaches n from below ($G=n$ maximizes throughput), the system behaves much like an M/M/1 queue, with delay growing without bound as $\mathbf{S}\rightarrow 1$. In this case, the system is stable in the sense described by Kleinrock and Lam, because it automatically reduces the rate of blocked packet retries as load increases (reaching an infinite retransmission delay at $\mathbf{S}=1$), which is exactly how they propose to control an ALOHA channel. However, as G increases beyond n, the upper branch of the delay curve drops sharply, rapidly converging to the level of a system with little or no buffer space, and the system remains unstable until $\mathbf{S}\rightarrow 0$. We now see that this system has the peculiar property of being stable *only* for $\mathbf{S}=0$ and $\mathbf{S}=1$!

This behavior results from ground stations using W, rather than G, to control their retransmission rate. After the satellite processor has removed all collisions from the data stream, ground stations cannot tell whether W is less than B because the up-channels are idle (and W is a good measure of G), or full of collisions (when W is a *very bad* measure of G). Since the upper branch of the delay curve represents the start of an unstable region, we would like to push the upper branch of the delay curve to infinity; this system tries to resist this push, so other forms of direct dynamic flow control are needed.

8. Extensions

Other useful information could be added to the packet header. For example, the number of attempts before each packet is successfully transmitted would be available if each ground station inserted an "attempt number" field into the packet header. The short-term average number of attempts per packet is a fair measure of G. Ground stations could adaptively change the length of their randomized delay interval based on this information.

Another important issue worth considering is whether the data field of "empty packets" could be put to use. As LSI technology improves, one can imagine a satellite processor with the capacity to monitor the set of up-channels and distribute measurements and/or control information to the ground stations as the data in "empty packets". One important use could be to distribute an estimate of G based on the number of successes and collisions on the up-channels.

The retransmission delay may be reduced to its absolute minimum by adding a field to the down-channel packet header that shows whether each up-channel slot in the previous period was idle, carried a successful transmission, or carried a collision. This change requires additional processing by the satellite, since no ground station can predict this information. However, each ground station knows which channel it used, so the identity of the sender need not be included. Such a system would be even less stable in the sense of Kleinrock and Lam, since the delays on both branches of the delay curve would be further reduced.

Given the usage of each up-channel slot (i.e. idle, success or collision), ground stations have as much system status information as the satellite. Ground stations could estimate G, or perform any control function, as well as the satellite. There is no time advantage in doing these calculations either in the satellite or on the ground: information originating in the satellite is used to make a decision needed on the ground, and both the data and the decision will be delayed by the same propagation time. The only function which can be performed efficiently only at the satellite appears to be a channel reservation policy such as CPODA [JACO 78], where central coordination is desirable.

9. Conclusions

As expected, the extra resources used in a concentrated ALOHA system improve its performance over a conventional slotted ALOHA system. What *is* surprising, however, is how rapidly the marginal improvement drops off as more resources are added.

For example, when you must pay for each extra up-channel, even if the cost per up-channel is low compared to a down-channel, a small number of up-channels provides maximum throughput. The optimal number depends on both the relative cost of an up-channel and the size of the satellite buffer. If the satellite has an infinite buffer, three up-channels will be optimal as long as the relative cost of bandwidth X is less than about 1.275 (a reasonable assumption under current technology); if the relative cost of up-channel bandwidth $X > 1.275$, then two up-channels are optimal. If both B and the relative cost of an up-channel are small, more than three up-channels may be optimal. However, the minimum cost at which three up-channels is optimal approaches zero exponentially as B increases.

If the retransmission timeout includes the worst-case satellite queueing delay, a small buffer, whose size increases with S, minimizes the average delay for a fixed number of up-channels. This strategy suggests an improved protocol, based on adding the arrival time at the satellite to the packet header. A terminal knows its message was lost as soon as any newer packet is transponded, so the average queueing delay in the satellite controls the rate of blocked packet retries. This queueing delay provides a useful, positive flow control mechanism while the short-term offered traffic is less than the system's capacity, and a harmful negative flow control mechanism whenever the system becomes overloaded. For stability, some additional form of direct, dynamic flow control must be imposed. Unlike previously studied forms of concentrated ALOHA, this new protocol has the nice property that adding satellite buffer capacity always improves the system.

By forcing the satellite to do more processing (and adding more information to the packet header) we can further reduce the average delay by having the satellite announce whether each up-channel slot was idle, in use, or had a collision. This is all the information that is available to the satellite itself, so that any additional processing could be done on the ground.

An important unanswered question is the transient behavior of such systems under sudden load variations. From analytic results, one would expect them to be stable as long as G approaches n smoothly from below (in fact, the system actually *is* stable at $G=0$ and $G=n$), but unstable if G exceeds n. These improved protocols show promise for satellite communications and merit further research.

Bibliography

ABRA 73 Abramson, N., "Packet Switching with Satellites", *AFIPS Conference Proceedings*, vol. 42, NCC, New York, June 1973, pp.695-703.

DERO 78a DeRosa, J. K. and L. H. Ozarow, "Packet Switching in a Processing Satellite," *Proceedings of the IEEE*, vol. 66, No. 1, Jan. 1978, pp. 100-102.

DERO 78b DeRosa, J. K., L. H. Ozarow, and L. N. Weiner, "Efficient Packet Satellite Communications", Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Submitted for publication to *IEEE Transactions on Communications*.

ERLI 76 Erlich, Z., "On Centralized Bus Transportation Systems with Poisson Arrivals", Computer Systems Modeling and Analysis Group, School of Engineering and Applied Science, University of California, Los Angeles, UCLA-ENG-76124, December 1976.

JACO 78 Jacobs, I. M., Binder, R. and E. V. Hoversten, "General Purpose Packet Satellite Networks", *Proceedings of the IEEE*, vol. 66, no. 11, Nov. 1978, pp. 1448-1467.

KLEI 75a Kleinrock, L., *Queueing Systems, Vol I., Theory*, Wiley-Interscience, New York, 1975.

KLEI 75b Kleinrock, L. and S. Lam, "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation", *IEEE Transactions on Communications*, vol. COM-23, pp. 410-423, Apr. 1975.

KLEI 76 Kleinrock, L., *Queueing Systems, Vol II., Computer Applications*, Wiley-Interscience, New York, 1976.

LAM 75 Lam, S. and L. Kleinrock, "Packet Switching in a Multiaccess Broadcast Channel: Dynamic Control Procedures", *IEEE Transactions on Communications*, vol. COM-23, pp. 891-904, Sept. 1975.

LITT 61 Little, J., "A Proof of the Queueing Formula $L = \lambda W$", *Operations Research*, vol. 9, no. 2, March 1961, pp. 383-387.

MOLL 78 Molle, M. L., "Analysis of Concentrated ALOHA Packet Satellite Links", M.S. Thesis, Computer Science Department, University of California, Los Angeles, October 1978.

ROBE 72 Roberts, L., "ALOHA packet system with and without slots and capture," ARPA Network Information Center, Stanford Res. Inst., Menlo Park, Calif., ASS Note 8 (NIC 11290) June 1972.

SPAN 78 Spaniol, O, "Multifrequency ALOHA-Type Systems", Institut fur Informatik der Universital Bonn, Wegelerstrasse 6, D - 5300, Bonn, Submitted for publication to *Journal of the ACM*.