IEEE Catalog No. 81CH1699-8
Library of Congress No. 81-82786
IEEE Computer Society Order No. 380

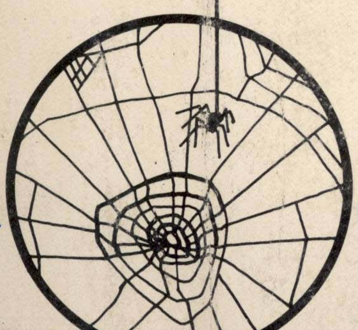
A symposium sponsored by

IEEE Computer Society
Technical Committee on
Computer Communications

U.S. Department of Commerce
National Bureau of Standards
Institute for Computer Sciences and Technology

December 8, 1981
National Bureau of Standards
Gaithersburg, Maryland

PROCEEDINGS



computer networking symposium

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE COMPUTER SOCIETY PRESS

IEEE Catalog No. 81CH1699-8
Library of Congress No. 81-82786
IEEE Computer Society Order No. 380

A symposium sponsored by

IEEE Computer Society
Technical Committee on
Computer Communications

U.S. Department of Commerce
National Bureau of Standards
Institute for Computer Sciences and Technology

December 8, 1981
National Bureau of Standards
Gaithersburg, Maryland

PROCEEDINGS

computer networking symposium

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE COMPUTER SOCIETY PRESS

IEEE Catalog No. 81CH1699-8
Library of Congress No. 81-82786
IEEE Computer Society Order No. 380

Order from: IEEE Computer Society
Post Office Box 80452
Worldway Postal Center
Los Angeles, CA 90080

IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08854

The Institute of Electrical and Electronics Engineers, Inc.

SYMPOSIUM COMMITTEE

Conference Chairperson
Robert Rosenthal
National Bureau of Standards

Program Chairperson
Thomas R. Stack
Network Analysis Corporation

Program Committee
Maurice France
TRW

Van Foster
U.S. State Department

James G. Hiemenz
U.S. House Information
Systems

NBS Liaison
Robert Toense
National Bureau of Standards

Treasurer and Registration
Gerrie P. Katz
IEEE Computer Society

Publications
C.G. Stockton
IEEE Computer Society

Local Arrangements Chairperson
Robert Toense
National Bureau of Standards

National Bureau of Standards
Gaithersburg, Maryland
December 8, 1981

Table of Contents

Session IA: Local Networks

Thomas R. Stack
Network Analysis Corporation

INTEGRATION OF VOICE, DATA, AND IMAGE TRAFFIC

ON A WIDEBAND LOCAL NETWORK*

Kathryn Hanson   Wushow Chou   Arne Nilsson

North Carolina Department of Administration   North Carolina State University

Abstract

A hybrid access method (HAM) is proposed for data, voice, and image traffic on a wideband, local communication network. The mixture of bursty and stream traffic from a large population of users requires: uninterrupted bandwidth for stream traffic; fast response time for bursty traffic; access by simple interface devices; method suitability for wide channel bandwidth, in particular, for optical fiber; and no bandwidth monopolization by either type of traffic.

HAM meets these constraints by dividing the channel into time frames, each with a fixed number of minislots for reservations, a specified maximum number of slots reserved by stream traffic, and a specified minimum number of slots for bursty traffic. The boundary between the bursty and stream portions of the frame shifts in response to stream traffic demand.

HAM is analyzed by modified reservation-Aloha and slotted-Aloha, and by diffusion approximation or analysis of models based on M/M/N (with varying service rate) or M/G/1 queues, depending on traffic load. Results show HAM appropriate for a range of mixes of traffic types and network operating characteristics.

Introduction

The networks to support data, voice, facsimile, video, and conference communications are expanding to the point that the duplication of facilities is becoming economically noticable. The problem is relatively new, but it is one of increasing interest as technology improves for integrated networks[1]. Demand for networks that handle more than one type of traffic is growing with the technology.

The communications can be considered generically as a mixture of bursty and stream traffic. Bursty traffic is predominantly represented by interactive data terminals and, in some cases, telephone voice conversations. Digitized voice conversation is a stream transmission at channel bandwidths in the thousands of bits per second range. In the hundred million bits per second range, individual packets appear more like bursty than stream traffic even when voice is digitized at high rates. Stream traffic can include video, facsimile, and bulk data.

A large population of users is expected to share the network. Unless the channel is effectively split, a stream transmission can block or cause intolerable delays for bursty traffic. If stream channel use is constant, one solution is separate subchannels. However, if the number of stream transmissions varies, as when file transfers are made primarily at night and interactive traffic activity is high during the day, then use of separate subchannels requires extra capacity and the system lacks flexibility to respond readily to dynamically changing loads of each traffic type.

The value of a movable boundary in reducing bursty traffic delay is shown in Figure 1. The delays for stream traffic at corresponding fixed slot availabilities are also shown. Loads of bursty and stream traffic are held constant as the boundary moves. Setting a minimum of seven slots per frame for bursty traffic results in an average delay of less than one msec. Allowing bursty traffic in unused stream slots can reduce delay almost an order of magnitude. The stream traffic leaves an average of 26 slots for bursty traffic. Given a movable boundary, the stream traffic delay is 0.5 msec excluding transmission time. An alternative perspective is that more bursty traffic can be supported than when the boundary is fixed, given the same delay constraint for bursty traffic.

Parameters for the figure are 150 Mbps operating bandwidth; 1 km network distance; 100 users each generating an average of one bursty data message of approximately 1000 bits every ten seconds; 500 users each generating an average of one 5 minute-duration conversation, digitized at 56 Kbps, every 33 minutes; and 50 users each generating an average of one 5 minute-duration steam transmission every 2.75 hours. Half of the streams are 1.544 Mbps and half are 6.312 Mbps. Delays are also shown where the number of users is increased by 15 percent.

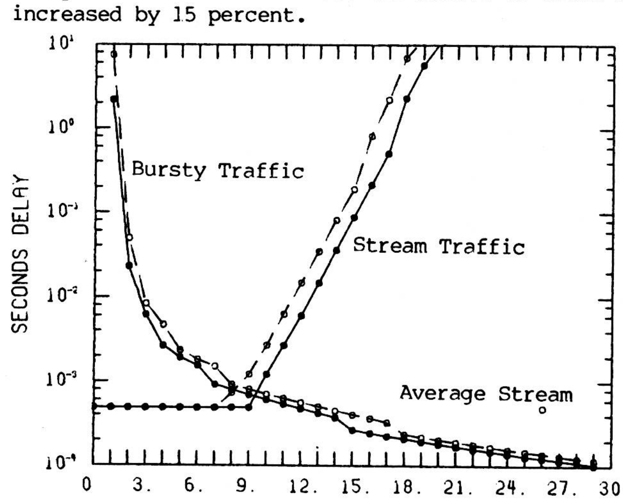


FIGURE 1: SLOTS AVAILABLE TO BURSTY TRAFFIC

CH1699-8/81/0000/0003$00.75 © 1981 IEEE

Effectiveness of the Solution

The hybrid access method (HAM) allows relatively simple and uniform stations to handle all traffic types. Previously developed protocols are suitable only for traffic with a uniform set of message characteristics and delay requirements, or they unnecessarily waste available bandwidth on a channel operated at a very high bandwidth.

For the mixed traffic and wideband medium environment, the following constraints apply:

1. Stream transmissions require assurance of bandwidth without interruptions.
2. Bursty traffic must have fast response time.
3. The majority of devices on the network are expected to be dumb terminals with simple network interfaces. They must have access to the network without the burden of a heavy processing load.
4. A dynamically varying mixture of traffic, including wideband streams, must be supported.
5. Fiber optics is likely to be used to provide needed bandwidth.
6. Stream traffic must be restricted to insure that it does not use all the bandwidth.

These constraints have major implications for design of the access method. The first implies that reservations are necessary for stream traffic. Given the large number of users, the second constraint implies that control must be decentralized. A contention scheme that does not require extensive tracking for its use is needed for bursty traffic if constraint three is to be met. Fixing the number of minislots per frame also contributes to simplicity of access devices. Because of contraint four, the channel must have wide bandwidth. On a local network, then, propagation time will be short but the ratio of propagation time to typical message transmission time in a slot will be large compared with currently prevalent ratios on local networks.

Constraints four and five each rule out carrier sense multiple access (CSMA) as the appropriate protocol for contention. CSMA does not give the best throughput when the ratio of propagation to transmission time is large; and a possible reflection problem exists if fiber optics is used. Finally, constraint six implies the requirement of a guaranteed minimum number of slots for bursty traffic which, because of constraint three, should be set to be available on a regular time schedule. Together these imply that a reasonable access method will combine a reservation scheme and a contention scheme without dependence on carrier sensing before transmission. HAM is designed to meet these constraints.

Multiple Access Protocols

In local networks, multiple access is a means for facility sharing. Allocation of the channel can be fixed or dynamic; control of dynamic allocations can be centralized or distributed; and distributed control can be by pure contention or reservation schemes.

An analogy can be made between switching techniques and multiple access protocols. Reservation schemes have characteristics related to circuit-switched transmission. Contention schemes are comparable to packet switching. Some access protocols combining bursty and stream traffic function much like hybrid switching methods. The slotted envelope network (SENET) approach[2], which reserves some portions of all time frames for circuit switching calls and some for packet switching data, includes a movable boundary feature to allow use by data of residual circuit switching capacity.

Contention Schemes. The primary contention schemes are Aloha[3,4], S-Aloha[5,6], CSMA[7,8], CSMA/CD (CSMA with collision detection)[9-11]. Several other variations of CSMA have been proposed[12-15]. These protocol presentations and reviews[16,17] emphasize bursty traffic in environments with either lower operating bandwidths or higher propagation delays than those for which HAM is designed. Contention schemes cannot insure the small variance in packet delivery delay necessary for stream transmissions.

Reservation Schemes. The purpose of reservations is to eliminate repeated contention for a part of the channel when a user transmits packets in close succession; but delay added by reservations would be significant for bursty traffic. Reservation-Aloha (R-Aloha)[18,19] is an implicit scheme in that each successful use of a slot implicitly reserves the corresponding slot in the next frame. The other major reservation scheme type is explicit, that is, a request is made explicitly for future transmission time[20-22].

Schemes have been designed for traffic with differing characteristics and requirements[23-25]. However, none of these schemes meets the needs of mixed data, voice, and image traffic on a local network supporting many users.

Hybrid Access Method Description

HAM allows simple user-network interface devices. Stations serving bursty traffic only need to keep synchronization and timing information and be able to retransmit when a message is not acknowledged. These stations transmit in any of a fixed set of slots each frame. Stations with greater capabilities are required for stream traffic and can be advantageous for bursty traffic, but again the protocol is kept as simple as possible. These user stations keep track of: the minislots in use, so they will not transmit over an ongoing reservation; the requests not yet satisfied; and the stream slots available when they make a request, so they can recognize the slots assigned to them for each frame. The slots not reserved for stream traffic are also recognized and can be used for bursty traffic.

Reservations for Stream Traffic

The number of minislots per frame is fixed primarily for simplicity in implementation. A user station contends for a minislot to transmit a reservation specifying the requested number of slots per frame. Once a minislot is acquired, the user continues to transmit the request in it to hold the minislot until the stream transmission is completed.

The minislots are a queue for the slots available to satisfy stream requests. Requests are met on a first-come-first-served basis[28] (or on a prioritized basis if the reservation also specifies priority). Minislot reservations are ranked in the order they will be served. Slots that will be used for stream traffic in the next frame are counted from the top of the ranking. The number of occupied slots can range from zero to the maximum number of slots available for stream traffic. By monitoring reservations, a user station knows which slots are available for bursty access in the next frame. Selection of the total number of minislots is made to allow a large percentage of the maximum number of outstanding stream requests to be queued. In this way the minislot reservations provide information to all users for flow control.

The hybrid scheme provides for zero delivery time variance for stream traffic unless preemptions are allowed by higher priority traffic. When some stream traffic does not have stringent delay requirements, as would be the case for bulk rather than real time traffic, tradeoffs are possible between the number of slots assigned to each stream transmission and the number of frames duration. For real-time traffic, allocated bandwidth must meet the message bandwidth requirement. (Or the users must have reduced their digitization rates to be able to fit within the allocated bandwidth. Inclusion of priority status in the reservations could be used to force some users to reduce their rates more than than others, or to drop low priority messages.)

The frame must be at least twice as long as the maximum propagation time, plus the time for all reservation minislots if a user is to recognize a successful reservation (or conflict) and be able to begin transmitting a stream message (or retransmit a request) in the next frame. This condition is obviously met on the local network because minislot and slot sizes must be as long as the propagation time plus request or message transmission time. Otherwise, transmissions in adjacent slots, and not just the same slot can collide.

Slot size selection and frame duration are interrelated with each other and with traffic characteristics and the relative mix of traffic types and of messages and acknowledgements. Longer frames result in greater average delay for stream transmissions. Too large a slot size reduces the maximum achievable utilization; stream traffic might not be generated fast enough to fill one slot each frame and bandwidth would be wasted.

When voice is handled as bursty traffic, slot utilization might be very low or the wait for enough generated bits to fill more of a slot might result in a delivery time variance that can not be acceptably handled by the buffers at the receiver. On the other hand, too small a slot size reduces effective utilization because a constant overhead for synchronization and addressing is required per slot. The impact of slot size selection on maximum utilization of a slot is illustrated at the end of the next section. Minislot size is based on the number of users to be served and the importance of each reservation request being correctly received.

Random Access for Bursty Traffic

In HAM, bursty traffic is assured a minimum number of slots per frame and has access to slots unused by stream traffic. The minimum number cannot be as low as the average number of slots required for bursty traffic without risking high bursty traffic delays. These would occur over periods when stream traffic uses most of its available slots and bursty traffic is above average[27,28].

Voice requirements are met in the bursty environment. Five msec or more buffering time typically will not fill one slot per frame; redundancies can be built in to minimize the effect of lost packets and the need for retransmissions. Up to 200 msec average variable talkspurt delay does not unduly affect speech activity[29].

S-Aloha rather than CSMA is the contention method used for bursty traffic. The ratio "a" of propagation time to transmission time is a measure of the opportunity for two transmissions to collide under CSMA. For large values of "a" and reasonable traffic loads of new arrivals plus retransmissions, carrier sensing does not give maximum throughput. The incorporation of collision detection (CSMA/CD) is particularly unnecessary because of the likelihood that a user will not detect a collision when one occurs.

Propagation time imposes an upper bound on the maximum channel bandwidth that can be used by any access method. This bound decreases with increasing network distance or bandwidth or with decreasing message length as shown in Table 1.

Table 1: Upper Bounds on Utilization

	Decreasing message size)			
	Maximum Message Size 2034 Bits		Maximum Message Size 752 Bits	
	0.5 km	1.0 km	0.5 km	1.0 km
	(Increasing distance)			
At 100 Mbps				
Required Slot Size (Bits)	1254	1484	742	972
Maximum Slot Utilization	.82	.69	.69	.53
Ratio Propagation/Max. Message Transmission Time	.225	.449	.449	.898
(Increasing bandwidth)				
At 150 Mbps				
Required Slot Size (Bits)	1369	1714	857	1202
Maximum Slot Utilization	.75	.60	.60	.43
Ratio Propagation/Max. Message Transmission Time	.337	.674	.674	1.348

Analytic Model

The analysis of HAM is an integration of three processes: (I) the acquisition of minislots for stream traffic reservations, (II) the acquisition of slots for stream traffic, and (III) the acquisition of slots for bursty traffic. The interrelationships of these processes are shown in Figure 2.

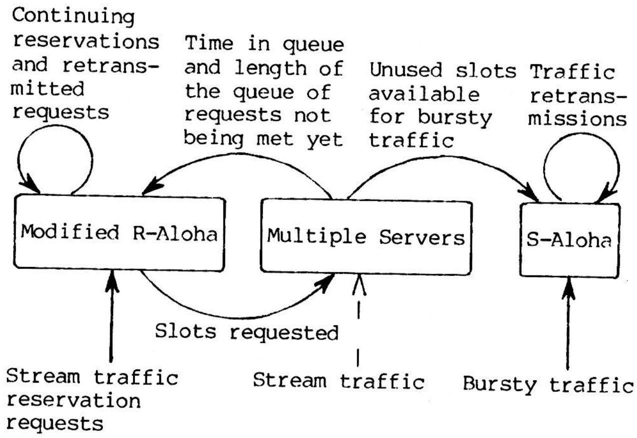


Figure 2: Analytic Model

The impact of (I) on stream traffic throughput and delay is significant when the number of minislots for making reservations is small enough that there is frequent contention. Additional delay beyond that for R-Aloha is incurred after a minislot is reserved and before transmission can begin because the user must wait for slot(s) to become available.

The multiple server system for (II) does not match any of the standard batch or bulk arrival or service queuing systems -- all of a specified number of slots requested for a transmission must be available at the same time, they free up at the same time, and stream slots can be left idle while requests are waiting in the queue of minislots.

Stream Traffic Service

An analytic formulation for (II) is needed to determine the expected number of slots available for bursty traffic and the expected waiting time in the queue of reservation minislots. Arrivals to the users are Poisson. Thus, arrivals of requests at the minislots and for service can be assumed to be Poisson also, provided sufficient minislots are available to hold all arrivals and provided there is little contention. Queuing for service occurs when an insufficient number of slots are available to satisfy all stream traffic requests in the minislots.

There can be queuing (backlogging) as well at each user station when the number of minislots is insufficient to hold all arrivals. If stream slots are available but the minislots are full, straightforward analysis at high utilizations of stream slots would overestimate service and thus underestimate delay.

Analytic determination of the steady state probabilities in a multivariate queuing model, in the case where the number of minislots is not a factor, has been found to be computationally intractable[30]. However, there are several other approaches for approximately modeling the multiserver portion of HAM analytically.

The system is certainly of the form G/G/NSS, a generally intractable model. NSS is the number of slots available for stream traffic. If the system is heavily loaded, that is, effective utilization of slots available for stream traffic is 65 percent or more, a good approximate solution can be found by diffusion approximation, which is based on an assumption that the number of slots in use is close to the number of slots available. This will not be the case when utilization is moderate or low.

If the system is very lightly loaded with utilization of slots no more than 15 percent, no queuing is expected. (There will be no queuing in the minislots either if the maximum possible number of slots requested is no more than the number of slots for stream traffic.)

When the system is more moderately loaded, another model formulation is necessary. Of course, if stream bandwidth is a constant K slots for all requests, an M/M/min([NSS/K],NZF) model is exact and results are readily calculable. NZF is the number of minislots per frame. If requests are for different numbers of slots and utilization of stream slots is no more than 25 percent, the system model is based on an M/G/1 queue with service completions independent for each slot. Up to this level of utilization, the effect of assuming release of slots on an individual basis instead of a block basis is not marked.

For utilization of stream slots between 25 and 35 percent, a linear combination of the M/G/1 based model and an M/M/NSS based model is used. The M/M/NSS queue has service rate dependent on the number in the system. The system model is based on this M/M/NSS queue when utilization of stream slots is between 35 and 65 percent.

Heavy Traffic Load. The number of servers, the variances, and the means of the arrival and service processes are needed to solve the forward Kolmogorov, or Fokker-Planck, equation for diffusion approximation. Approximate solution[31] of the equation requires specification of the mean and coefficient of variation (squared) of the interarrival and interdeparture processes. The expected number of arrivals to the multiple servers is AE(y), where E(y) is the average number of slots per frame bandwidth of each stream transmission and A is the arrival rate per frame of all messages. $AE(y^2)$ is the incremental variance. Well known results from renewal theory yield the interarrival process mean, 1/[AE(y)], and the coefficient of variation (squared), $E(y^2)/E(y)$. Results are analogous for the interdeparture process.

The mean of the interarrival process can be written 1/[(M)(ASU)] where ASU is the average number of slots used for stream transmissions.

ASU = min[AE(y)/M, NSS, NZF E(y)]

That is, the average number of slots used cannot exceed the steady state utilization (RHO = A/M, where M is the service rate) times the average number of slots per request [E(y)], or the maximum number of slots available (NSS), or the number of minislots (NZF) times the average number of slots per request.

Light Traffic Load. An approximation of the service process can be made by assuming that service slots free up one at a time instead of in blocks. Then stream duration is distributed as the maximum of a specific number of exponentially distributed random variables; that number is the number of slots per frame required for the stream transmission. The formulation is tractable and it provides sufficiently accurate results under light traffic loads, where queuing time in the minislots is not the predominant part of delay.

The Laplace-Stieltjes transform[32] for the equilibrium waiting time in the queue given this service process is

$$W^*(s) = (1-pq)(1-pd) + (1-pq)\ pd\ D^*(s) + \frac{pq\ [1-D^*(s)]\,[1-A\ E(B)]}{[s-A+A\ B^*(s)]\ E(d)}\ B^*(s)$$

pq is the steady state probability that the queue is not empty. pd is the steady state probability that a request made when the queue is empty will be delayed because insufficient service slots are available. D is the initial delay random variable; it is the delay of a request, made when the queue is empty, waiting for sufficient slots to be available to satisfy it. "A" is the request arrival rate. The arrival process is Poisson. B is the interservice random variable of requests in the queuing period; it is the time between one request in the queue beginning to be met and the next request in the queue beginning to be met. The expected number in the queue is Lq=AE(W).

A new queuing system is defined where service time is the time spent at the head of the queue. Such a system has <u>one</u> server and Poisson arrivals, that is, it is an $\overline{M/G/1}$ queue with busy period equivalent to the queuing period of the original system. "Service times" are distributed as the interservice random variable B except for the requests that initiate a queuing period. Those have "service time" distributed as D.

Calculations require specification of the matrix of the expected number of visits to transient state "j slots in use" starting in transient state "i slots in use" before absorption when there is a request requiring more slots than available. The transition probabilities are modified by F(i), the probability that at the time of an arrival the number of requests being served is <u>less than</u> the number of minislots, NZF, given that the number of slots in use is "i." Inclusion of the factor F can be critical in the case of an access method like HAM where reservations must be made for service and the maximum number of reservations is restricted.

Estimation of F is based on recognition of breakpoints[33] at the number of slots in use if all but one of the minislots are full and each request is for the minimum number of slots that can be requested for a stream transmission; and at the number of slots in use if all of the minislots are in use and each request if for the maximum number of slots that can be requested. The lower the utilization of minislots, the larger the value of "i" for which all requests are likely to be held by the minislots.

For utilization of stream slots between 25 and 35 percent, the length of the queue is calculated to be (g)(Lq)+(h)(L-Z) where g = (0.35-RHOH)(10), h = 1-g, and (L-Z) is the length of the queue calculated assuming moderate load, which is described in the next section.

<u>Moderate Traffic Load</u>. An approximate model based on the service rate varying with the number in the system was proposed[30] for a queuing system where a user can request multiple servers. Although the model assumed what would be an infinite number of minislots for reservations under the HAM scheme, the structure proved to be a reasonable starting point for a piecewise linear approximation[33] of the service rate given the number of reservation requests in the minislots. Breakpoints occur approximately at the largest number of requests such that all can be served simultaneously; at the largest number of requests such that all can be served simultaneously assuming each request is for E(y) slots; and at the maximum number of requests that can be in service.

With the values for the service rate MU(n), the steady state probabilities Q(n) are calculated as for an M/M/N queuing system. The average number of requests in the minislots is as follows.

If NZF < N3:

$$L = \sum_{n=1}^{NZF-1} nQ(n) + NZF\ \Big[\sum_{n=NZF}^{N3-1} Q(n) + \frac{Q(N3)}{1-\theta}\Big]$$

If NZF = N3:

$$L = \sum_{n=1}^{N3-1} nQ(n) + NZF\ \frac{Q(N3)}{1-\theta}$$

If NZF > N3:

$$L = \sum_{n=1}^{N3-1} nQ(n) + Q(N3)\Big[\frac{N3+\theta}{1-\theta} + \Big(\frac{\theta}{1-\theta}\Big)^2 (1-\theta^{NZF-1-N3})\Big]$$

The mean number of customers in service given n requests in the minislots is s(n)=MU(n)TFS. The expected number in service can be written as

$$Z = \sum_{n=1}^{\infty} s(n)\ Q(n) = \sum_{n=1}^{N3-1} s(n)\ Q(n) + \sum_{n=N3}^{\infty} s(n)\ Q(n)$$

where the last term is

$$\sum_{n=N3}^{\infty} \frac{CMS}{\theta}\ TFS\ Q(N3)\ \theta^{n-N3} = \frac{s(N3)Q(N3)}{1-\theta}$$

The expected number of requests waiting for service in the queue of minislots is L-Z.

<u>Bursty Traffic Service</u>. Given the average number of slots available for bursty traffic, which is determined in the analysis of process (II), an S-Aloha analytic formulation can be applied directly in the case where a user stops generating bursty traffic each time he has an outstanding bursty message. For true Poisson arrivals, the population is effectively infinite. Where the probability of a message being sent in a slot is constant and finite, the model remains applicable.

Propagation delay is assumed to be zero in the analysis. However, its effect is accounted for in calculations either of slot size or of the number of slots over time required to have a specified number of slots available for one type of traffic. The average delay waiting for retransmission after a collision is ARD=NSP+(K+1)/2, where NSP is the number of slots for propagation delay; NSP=0. The Markovian model assumes a memoryless geometric distribution for retransmission delay; it requires a value for the probability of each message being independently retransmitted in any slot. Probability P=1/ARD has the same mean as the uniform retransmission over the next K available slots.

The rate at which new messages are transmitted in a slot is

$$S = \frac{CMB + CMV}{ASA}, \text{ where}$$

CMB and CMV are the rate of arrivals per frame of bursty data messages and voice packets, respectively. Therefore, the average message delay, in slots, is

$$Db = Bb + NSP + 1 + \frac{ASA + (NSF - ASA)2}{2\ NSF}, \text{ where}$$

Bb, the average backlog time per bursty message in slots, is a function of ARD and the arrival rate of new messages, S. By Little's result, it is equal to the ratio of the expected backlog to the steady-state throughput rate. The ratio is increased by a factor of (NSF/ASA)(1-T)2 to account for the effect of framing and the corresponding gap between slots for bursty traffic in successive frames. NSF/ASA is the ratio of the number of slots per frame to the number of slots available for bursty traffic. The effect of this expansion factor varies with the steady state traffic throughput rate, T.

The remaining terms contribute to average message delay, Db. NSP is the number of slots, if any, required for propagation delay. The single slot is the transmission time, where it is assumed a bursty traffic message fills a slot. A half-slot delay is incurred on the average waiting for the beginning of a slot when in the appropriate portion of the frame, which is ASA/NSF of the time. The last term is the average time, in slots, spent waiting for the portion of the frame in which bursty traffic can be transmitted. (NSF-ASA)/NSF is the probability of a message arriving in the wrong portion of the frame. (NSF-ASA)/2 is the average time such a message would have to wait.

The number of slots per frame used successfully for bursty traffic must equal the steady state arrival rate per frame when steady state is achieved. Thus, the channel throughput rate (excluding the effect of controls) in bursty messages per second is

$$Tb = \frac{CMB + CMV}{TFO}, \text{ where}$$

TFO is the number of seconds per frame. The more slots available per frame for a given number of message transmissions per second, the lower the utilization, the fewer the collisions, and the shorter the delays. Delay for voice packets typically is dominated by the average buffering time for packetization.

Stream Traffic Reservations. Reservations have a significant impact on stream traffic throughput and delay when the number of minislots for making reservations is small enough that contention occurs frequently. Then additional control is required, as it is for S-Aloha, to maintain network stability. S-Aloha, as described in the preceding section, provides a foundation for the R-Aloha model[19]. However, the R-Aloha model is not exactly representative of the use of minislots in HAM; a station continues to generate messages which might be transmitted even when another of the station's messages is being transmitted or is backlogged. This is approximated by assuming a very large number of users generating the same total traffic.

ADS is the average delay, in minislots, for a reservation request to be successfully transmitted into an available minislot.

$$ADS = Bs + NZP + 1 + \frac{Cm\ X + (TFZ-X)2}{2\ TFZ}, \text{ where}$$

Bs is the average backlog time per stream reservation request, in minislots. It is calculated like Bb, but increased by a factor of 1+(ARD/X)(TFZ-X)/X, which represents the delay of waiting for the appropriate (minislot) portion of the frame for each frame over which retransmission is delayed. ARD/X is the ratio of the average number of minislots between a collision and message retransmission, to the number of minislots available in a frame.

NZP is the number of minislots for propagation delay; NZP=0. The single minislot is the initial reservation transmission time. Cm is the average additional delay, a fraction of a minislot, between minislots. When an integer number of minislots completely fills a slot, Cm=0; otherwise Cm = TSZ/NZO, where TSZ is the slot duration in minislots and NZO is the number of minislots per slot. If the slot and minislot size are not well designed, Cm can be a large fraction of a minislot. TFZ is the frame duration expressed in minislots and X is the expected number of minislots available for requests.

$$X = NZF - \frac{ASU}{E(y)} - Lq + (CMS)(TFO), \text{ where}$$

ASU is the average number of slots used and Lq is the expected number of unsatisfied requests held in the queue of reservation minislots; both are calculated in (II). In the steady state, arrivals equal departures so (CMS)(TFO) is the number of minislots freed per frame by departures.

The overall message delay must be increased for the HAM model. Additional delay is incurred after a minislot is reserved and before transmission can begin. First, a new successful reservation is not effective until the next frame. All users must see the requests and calculate what slots, if any, will be used to satisfy it. Second, the delay Wq is incurred waiting for slot(s) to become available. Third, there is a one-time delay between the minislot for reservation and the first slot for transmission. Last, delay is incurred for propagation and transmission of the slot(s) in the first frame. The remaining transmission time is TFS-1 frames.

The average delay, in minislots, for the stream reservation requests is

$$Ds = ADS + TFZ + Wq + \frac{NZF}{2} + [ASU+E(y)]TSZ+(TFS-1)TFZ$$

where Wq is the expected waiting time in the queue of minislots, calculated in (II). TFZ + Wq + NZF/2 + (ASU)(TSZ) is the average time spent from the successful reservation until transmission can begin in the stream slots. [E(y)](TSZ) is the average time for transmission in the first frame. TSZ is the slot duration expressed in minislots. (TFS-1)(TFZ) is the average transmission time for the remainder of the message, expressed in minislots. TFS is the average number of frames duration of a message.

Input to the S-Aloha model for stream traffic analysis must be modified to account for the effect of continuing reservations. The steady state probability of a successful new request in a minislot (excluding the effect of controls) is S = CMS/X. S is greater than CMS/NZF because on the average only X and not NZF minislots are available for new arrivals. The quantity ASU/E(y) of the minislots in use corresponds to the ASU stream transmissions underway. Average retransmission delay and delay waiting for the reservation portion of the frame are also increased by the ASU minislots in use. The average rate of requests is the steady state throughput rate of new requests, as calculated with the above value for S, plus the steady state message throughput, ASU/TFS.

With these modifications, the R-Aloha model is representative of the reservation process. The throughput of reservation minislots is comprised of new requests plus previously successful requests for which service is not yet complete.

General Results

The major network parameters are channel bandwidth and network distance. Given unlimited buffering by users and the ability of users to transmit more than one message each frame, total traffic load and the relative percents of bursty and stream traffic are the significant input parameters.

Frame size and slot size are significant design parameters. They must be chosen to allow:

- multiple reservation minislots in a slot
- multiple slots per frame to be available for reservations and for stream and bursty transmissions
- large enough slots for efficient bursty data transmission
- small enough slots for efficient utilization in voice transmission without excessive buffering delay
- properly sized slots for stream transmissions to use one or more slots efficiently.

Stream transmissions are assumed to be at 1.544 Mbps, which is equivalent to the AT&T T1 carrier, and at 6.312 Mbps, which is equivalent to the T2 carrier. The offered stream load is assumed to be divided equally between these in the example results. Table 2 shows several cases of the effect of varying operating bandwidth and distance on slot and minislot requirements. It is the increasing size of the propagation overhead that reduces the number of minislots per slot as operating bandwidth or distance increases.

The number of slots per frame is fixed at 38 for convenience of comparison. The number of information bits per slot is kept constant. Stream traffic of 1.544 Mbps requires one slot per frame, 6.312 Mbps requires three, and their average requires two for several sets of network parameters. (If the number of slots per frame is increased, while frame duration is kept constant, slot size must decrease. Then, even the number of slots per frame for bursty data varies. If the number of slots per frame is increased while slot duration is kept constant, frame length increases. Then, as operating bandwidth or distance varies, the number of information bits per slot -- and the number of slots per frame for bursty data -- varies also.)

Table 2: The Effect of Bandwidth and Distance on Slot and Minislot Requirements

Operating Band width (Mbps)	Distance (km)	Propagation Overload Bits	Slot Size (secx10-3)	Slots/Frame for 6.312 Mbps Stream	Mini-slots/ Slot
100	0.50	230	0.0127	3	3
	0.75	345	0.0138	4	2
	1.00	460	0.0150	4	2
	1.25	575	0.0161	4	2
	1.50	690	0.0173	4	2
	1.75	805	0.0184	5	1
125	0.50	288	0.0106	3	2
	0.75	432	0.0117	3	2
	1.00	575	0.0129	3	2
	1.25	719	0.0140	4	1
150	0.50	345	0.0092	3	2
	0.75	518	0.0104	3	2
	1.00	690	0.0115	3	2
	1.25	863	0.0127	3	1
175	0.50	403	0.0082	2	2
	0.75	604	0.0094	3	2
	1.00	805	0.0105	3	1
200	0.50	460	0.0075	2	2
	0.75	690	0.0085	2	2
	1.00	920	0.0098	3	1

Table 3: The Effect of Bandwidth and Distance on Efficiency

Band-width (Mbps)	Distance (km)	Percent Utilization of: Information for 1.544	Information for 6.312	Bits/Slot for Voice	Total Bits/Slot for Data
100	0.50	0.717	0.977	0.260	0.818
	0.75	0.782	0.799	0.255	0.750
	1.00	0.847	0.866	0.246	0.693
	1.25	0.912	0.932	0.265	0.643
	1.50	0.977	0.999	0.248	0.600
	1.75	0.521	0.852	0.265	0.563
125	0.50	0.600	0.817	0.261	0.782
	0.75	0.665	0.906	0.265	0.706
	1.00	0.730	0.995	0.265	0.643
	1.25	0.795	0.813	0.260	0.590
150	0.50	0.541	0.711	0.265	0.750
	0.75	0.567	0.800	0.255	0.667
	1.00	0.651	0.888	0.260	0.600
	1.25	0.717	0.977	0.260	0.546
175	0.50	0.466	0.912	0.270	0.720
	0.75	0.531	0.723	0.270	0.631
	1.00	0.596	0.812	0.260	0.563
200	0.50	0.424	0.866	0.261	0.693
	0.75	0.489	0.999	0.266	0.600
	1.00	0.554	0.755	0.261	0.530

Even with the number of slots per frame fixed, the number of slots per frame for a specified stream bandwidth decreases as operating bandwidth increases, and increases as network distance increases. The reason is that changing slot duration changes frame duration and the number of stream information bits per frame. Table 3 shows the effect of varying operating bandwidth and distance on slot utilization for streams at 1.544 and 6.312 Mbps, for voice at 56 Kbps buffered 5 msec, and for data packets of 1036 bits.

Parameters for these example results are: 1036 message bits per packet; 168 bits per reservation request; (50%) 1.544 Mbps and (50%) 6.312 Mbps stream bandwidths, requiring 1 and 3 slots per frame, respectively; 60 sec per message mean duration, exponentially distributed; 0.1 msec maximum between retransmissions; 0.021 - 0.149 message arrivals per second; 150 Mbps network bandwidth; 1 km network distance; 4.6 microsec/km propagation delay; 38 slots per frame; 2 minislots per slot; 25 maximum slots per frame for stream traffic; 18 minislots per frame; 0.437 msec per frame; 0.6 ratio of information bits to total bits per slot; (50%) 0.652 and (50%) 0.888 utilization of information bits for streams; and 0.999944 - 0.99821 probabilities of sufficient minislots for all requests (assuming M/M/oo queuing system).

Delay generally decreases with increasing bandwidth and increases with network distance; this is influenced by the number of minislots available per frame and the number of slots per stream requests.

The number of slots per frame for minislot reservations is set at nine and the maximum number of slots per frame for stream transmissions is set at 25 for this comparison. Thus, a minimum of 4 slots are available for bursty traffic only. The stream traffic delays, excluding transmission times, are shown in Figure 3 as a function of utilization of the maximum number of stream slots. Delay is calculated using each analytic formulation and simulation over a full range of utilizations.

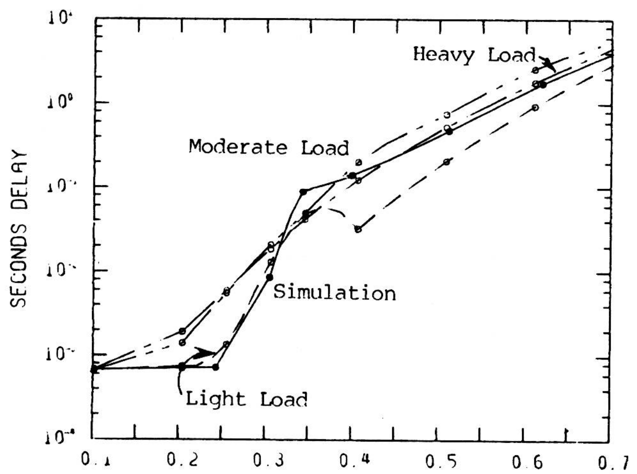


FIGURE 3: UTILIZATION OF STREAM SLOTS

The increase in delay when network distance increases is a direct result of the increase in slot duration, which must include propagation time. The effect of distance within the range where slots per frame and minislots per slot are constant is only noticable at low utilizations, where queuing time does not dominate. Changing the number of minislots per slot has little effect on delay if the total number of minislots still holds almost all requests.

Results for the analysis of bursty traffic are straightforward. As the load of bursty traffic increases while the average number of slots remains the same, or as the number of slots available decreases while bursty load remains constant, delay for bursty traffic increases as shown in Table 4. The network parameters are as above except that the minimum number of slots for bursty traffic is 16; the maximum number of slots per frame for stream traffic is 13. Simulation and analytic results are almost identical for bursty traffic.

Table 4: Changes in Delay for Bursty Traffic

Utiliz. of Slots for Bursty[1]	Stream	Seconds Delay for Data x10-3	Voice x10-3	Average Slots Available	
6.8	35.3	0.050	2.550	24.42	
18.1	35.3	0.062	2.562	24.42	held
48.1	35.3	0.097	2.597	24.42	fixed
25.7	47.0	0.077	2.577	22.89	
37.6	62.7	0.100	2.600	20.85	
52.0	78.3	0.129	2.629	18.82	

[1] Relative to the maximum utilization of 1/e.

Conclusion

The hybrid protocol, HAM, is a significant development for integrating data, voice, and image traffic on a wideband local network. HAM uses a contention access method to serve bursty traffic and a reservation access method for stream traffic; the number of users and channel bandwidth are assumed to be large. Channel time is divided into frames, and each frame is divided into time slots so that a stream transmission can have the same relative slot positions each frame. These divisions also allow simple user-network interface devices for bursty traffic because the same set of relative slot positions is available for access each frame.

Analysis provides the number of stream traffic reservation requests in the minislots, the time a request is delayed (backlogged) at the interface device, the time a request is queued in the minislots waiting for service slots, the number of slots used for stream traffic and, thus, the number of additional slots available for access by bursty traffic, and bursty traffic delay.

References

1 Pickens, A., and K. Hanson, "Integrating Data, Voice, and Image Traffic," Chapter 17 in Computer Communications, Vol. II, Edited by W. Chou, Prentice-Hall, Inc., NY, to be published in 1982.

2 Coviello, G. J., and P. A. Vena, "Integration of Circuit/Packet Switching by a SENET (Slotted Envelope Network) Concept," Nat. Telecommun. Conf., Dec. 1975, 42-12 - 42-17.

3 Abramson, N., "The Aloha System -- Another Alternative for Computer Communications," AFIPS Conf. Proc., 37, 1970, 281-285.

4 Binder, R., "ALOHA Packet Broadcasting -- A Retrospect," AFIPS Conf. Proc., 44, 1975, 201-215.

5 Abramson, N., "Packet Switching with Satellites," AFIPS Conf. Proc., 42, 1973, 695-702.

6 Kleinrock, L., and S. S. Lam, "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation," IEEE Trans. Commun., COM-23(4), Apr. 1975, 410-422.

7 Kleinrock, L., and F. A. Tobagi, "Packet Switching in Radio Channels: Part I-Carrier Sense Multiple-Access Modes and Their Throughput Delay Characteristics," IEEE Trans. Commun., COM-23(12), Dec. 1975, 1400-1416.

8 Hansen, L. W., and M. Schwartz, "An Assigned-Slot Listen-Before-Transmission Protocol for a Multiaccess Data Channel," IEEE Trans. Commun., COM-27(6), June 1979, 846-857.

9 Metcalfe, R. M., and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. ACM, 19(7), July 1976, 395-404.

10 Shoch, J. F., and J. A. Hupp, "Measured Performance of an Ethernet Local Network," Commun. ACM, 23(12), Dec. 1980, 711-721.

11 Shoch, J. F., "A Brief Note on Performance of an Ethernet System Under High Load," Computer Netw., 4, 1980, 187-188.

12 Tobagi, F. A., and V. B. Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Proc. Local Area Computer Netw. Symp., May 1979, 217-245.

13 Watson, W. B., "Configuration-Dependent Performance of a Prioritized CSMA Broadcast Network," Computer, Feb. 1981, 51-58.

14 Lam, S. S., "A Carrier-Sense Multiple Access Protocol for Local Networks," Comput. Netw., 4(5), Oct./Nov. 1980, 21-32.

15 Tobagi, F. A., and L. Kleinrock, "Packet Switching in Radio Channels: Part IV -- Stability Considerations and Dynamic Control in Carrier Sense Multiple Access," IEEE Trans. Commun., COM-25(10), Oct. 1977, 1103-1119.

16 Tobagi, F. A., "Multiaccess Protocols in Packet Communications Systems," IEEE Trans. Commun., COM-28(4), Apr. 1980, 468-488.

17 Lam, S., "Multiple Access Protocols," Chapter 4 in Computer Communications, Vol. I, Edited by W. Chou, Prentice-Hall, Inc., NY, to be published in 1982.

18 Crowther, W., R. Rettberg, D. Walden, S. Orstein, and F. Heart, "A System for Broadcast Communication: Reservation-ALOHA," Proc. 6th Hawaii Internat. Conf. Syst. Sciences, University of Hawaii, Honolulu,

19 Lam, S. S., "Packet Broadcast Networks -- A Performance Analysis of the R-Aloha Protocol," IEEE Trans. Comput., C-29(7), July 1980, 596-603.

20 Roberts, L. G., "Dynamic Allocation of Satellite Capacity Through Packet Reservation," AFIPS Conf. Proc., 42, 1973, 711-716.

21 Binder, R., "A Dynamic Packet-Switching System for Satellite Broadcast Channel," Internat. Conf. Commun., San Francisco, CA, June 1975, 41-1 - 41-5.

22 Jacobs, I. M., R. Binder, and E. V. Hoversten, "General Purpose Packet Satellite Networks," Proc. IEEE, 66(11), Nov. 1978, 1448-1467.

23 Pan, A., "Integrating Voice and Data Traffic in a Broadcast Network Using Random Access Scheme," Internat. Conf. Comput. Commun., 1978, 551-556.

24 Johnson, D. H., and G. C. O'Leary, "A Local Access Network for Packetized Digital Voice Communication," IEEE Trans. Commun., COM-29(5), May 1981, 679-688.

25 Shoch, J. F., "Carrying Voice Traffic Through an Ethernet Local Network -- A General Overview," IFIP WG 6.4 Internat. Workshop on Local-Area Computer Networks, Zurich, Aug. 1980.

26 Green, L., "Queues in which Customers Require a Random Number of Servers," Mgmt. Sci., 27(1), Jan. 1981, 65-74.

27 Weinstein, C. J., M. L. Malpass, and M. J. Fisher, "Data Traffic Performance of an Integrated, Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. Commun., COM-28(6), June 1980, 873-878.

28 Bially, T., A. J. McLaughlin, and C. J. Weinstein, "Voice Communication in Integrated Voice and Data Networks," IEEE Trans. Commun., COM-28(9), Sep. 1980, 1478-1490.

29 Gruber, J. G., "Delay Related Issues in Integrated Voice and Data Networks," IEEE Trans. Commun., COM-29(6), June 1981, 786-800.

30 Kim, S. S., M/M/s Queueing System where Customers Demand Multiple Server Use, Ph.D. Thesis, Southern Methodist University, Texas, Aug. 1979.

31 Nilsson, A. A., J. Seraj, and W. J. Stewart "The Analysis of Multiprocessor Systems by a Diffusion Approximation," Ninth Internat. Teletraffic Conf., Oct. 1979.

32 Green, L., "A Queuing System in Which Customers Require a Random Number of Servers," ORSA, 28(6), Nov.-Dec. 1980, 1335-1346.

33 Hanson, K. W., Integration of Voice, Data, and Image Traffic on a Wideband Network, Ph.D. Dissertation, Operations Research, North Carolina State University, Raleigh, Dec. 1981.

* Work supported in part by the National Science Foundation Grant Number ECS 77-24110.