THE PERFORMANCE OF LAN PROTOCOLS

J.L. Adams

Rapporteur: A. Linton

THE PERFORMANCE OF LAN PROTOCOLS

J.L.Adams

Notes of a lecture given to the International Seminar on the Teaching of Computing Science at University Level, University of Newcastle Upon Tyne, 3rd - 6th September, 1985

ABSTRACT

When a local area network (LAN) is used purely for local computer networking the main performance measures of interest are the overall throughput and the point-to-point bandwidth obtainable on the LAN, together with the expected message delay time. Recently, however there has been increasing interest in using LAN's to carry voice, leading to the multi-service mode of operation. This gives rise to new constraints and problems relating to performance. In particular, the LAN must operate efficiently at a much higher transmission rate, and due account must be taken of the delay-critical nature of the voice service. These additional criteria are being used to re-examine and further characterize the various bus and ring LAN architectures, particularly the latter. The discussion of this topic concludes with some remarks on current activity in the area of high-speed LAN standardisation, and recent British Telecom work in this area.

1 INTRODUCTION

A local area network (LAN) may be defined as a network linking users within a limited area which supports some type of communications processing and transparent information transfer. A general characteristic of most LAN's is that a mixture of services with different bit rates can share the available bandwidth through the operation of an appropriate media access protocol. The choice of which access protocol to use is an important consideration, and remains an active study area taking into account such factors as:-

-the network topology

-the range of service bit-rates

-the access delay tolerance of individual services

-the error tolerance of individual services

-the total load to be carried

-the degree of imbalance in the load distribution

-the requirements of overload control

-the efficiency of bandwidth utilisation

-the number of attached nodes

The number of factors involved reflects the large range of uses which have been identified for the LAN. At one end of the range are LAN's which are isolated from other networks and which are required to handle purely delay-tolerant data connections with modest bit-rates. Such LAN's may utilise transmission rates which are of the order of 1 Mbit/sec. Increasing the transmission rate by an order of magnitude implies that a larger number of terminals can be attached to the network and a wider range of service bit-rates can be accommodated. It also appears feasible for such LAN's to carry a limited amount of voice. With these capabilities it is therefore more likely that a bridge or gateway connection to another network is required,giving rise to a potential long-term imbalance in the load distribution. Increasing the transmission rate once again to around 100 Mbit/sec takes us into the realm of high-speed LAN's for which the principal interests are in high point-to-point bandwidth availability and in handling a high proportion of voice connections.

The motivation for studying LAN access protocol performance is to determine the area of applicability of a particular protocol to the range of above-mentioned user requirements. A number of access protocols are of interest,namely:-

-polled access protocols

-random access protocols

-dedicated access protocols

The subdivisions of these 3 broad categories contain the major access techniques which are of interest:-

(i) Polled access protocols can be implemented on token and slotted systems wherein transmissions are polled either through reception of a token, which is passed serially around all nodes,or through the reception of an empty slot.

(ii) Random access protocols can also be subdivided depending on what limits are placed on each node's right to transmit. These range from pure ALOHA, where nodes have a permanent right to transmit,through various schemes like CSMA/CD and limited contention protocols which successively reduce that right.

(iii) Dedicated access may be associated with a hybrid scheme wherein the LAN bandwidth is divided into two areas which are accessed using different protocols. On one side access is allowed via either a polling or random access protocol, whereas on the other side connections are assigned fixed parts of the bandwidth for the duration of the call.

From the list of user requirements we can now go on to examine several performance aspects of interest and determine how each of these very different access techniques behaves,progressing towards the case of general mixed services.

2 EFFICIENCY CONSIDERATIONS

Once bandwidth has become free an interval of time will elapse before the protocol has established which node can transmit a further packet. This delay limits the amount of useful load that can be carried and leads to a measure of the efficiency of the protocol, which can be expressed as:-

$$E = \frac{\text{TRANSMISSION TIME}}{\text{TRANSMISSION TIME + ACQUISITION TIME}} \qquad (1)$$

In this expression, the 'acquisition time' refers to the expected time required to transfer an entitlement to transmit from one active node to the next. The 'transmission time' refers to the expected time required to complete each transmission.

For random access protocols, the acquisition time lengthens as the number of nodes waiting to transmit increases and this produces a corresponding deterioration in efficiency. This loss of efficiency becomes significant if the acquisition time approaches a value similar to the transmission time. For example, with the Ethernet system it can be shown that, when many nodes have packets to send, the acquisition time is approximately 5.4L,where L is the end-to-end propagation delay [1]. Thus, for a given packet length of P bits, the acquisition time and transmission time will be equal when:-

$$\frac{P}{C} = 5.4L \qquad (2)$$

where C is the channel transmission rate in bits/sec. This formula shows that efficiency problems become significant on high-speed LAN's and on LAN's which have a long cable length.

In contrast, for polled access protocols the acquisition time reduces as the number of nodes waiting to transmit increases. This is true of both slotted and token systems. For example,consider a token ring where N nodes each transmit a packet which is of length P bits including the token. If the ring is of length R bits the token cycles once round the ring in a time S given by:-

$$S = \frac{NP + R}{C} \qquad (3)$$

However, the fraction of this time spent transmitting the N packets is just NP/C. Hence the efficiency E is:-

$$E = \frac{NP}{NP + R}$$

ie $$E = \frac{P}{P + R/N} \qquad (4)$$

The second term in the denominator represents the acquisition time and exhibits the characteristic that it reduces with increasing node activity. However, its dependence on R shows that efficiency again becomes a significant problem on high-speed rings and on long rings which, in either case, gives rise to an increase in the number of bits contained on the ring circumference.

For both random access and token access protocols equations (2) and (4) show the further characteristic that an increase in P or, more generally, an increase in the transmission length,causes an improvement in efficiency. Consequently, where efficiency is a consideration and especially on high-speed LAN's, it is necessary that each node be allowed to transmit several packets before releasing the channel to other nodes. However, this introduces a certain amount of deliberate 'hogging' into the protocol,ie a node is allowed to continue transmitting further packets despite the fact that it delays other nodes which are waiting to transmit. The success of this technique is therefore limited by delay considerations. This aspect will be examined in more detail in the next section.

For slotted rings the above approach is not feasible since transmissions are limited to the length of just one mini-packet. Hence, scaling slotted rings towards high-speeds requires further examination. The most significant efficiency problem posed by higher speeds is the occurrence of multiple slots on the ring. With the Cambridge ring protocol a node must wait for a slot which it has filled to return before any further transmission can take place. Although this simplifies the retransmission of rejected mini-packets it reduces slot utilisation because of the enforced reduction in node activity. The effect becomes significant when the number of slots on the ring is similar to the number of active nodes.

Scaling slotted rings towards the higher speeds is therefore accompanied by attempts to remove many of the restrictions on slot access. For example, it is not considered necessary to retransmit voice packets because voice at 64 kbit/sec exhibits a relatively high degree of error tolerance. This encourages the removal of the above mentioned access restriction of the Cambridge ring in situations where voice is the dominant load. The consequent gain in efficiency can be further reinforced by taking advantage of another means that slotted architectures have for improving throughput. Using the concept of 'destination deletion' the slot can be emptied may be attached to a node usually makes an allowance for the fact that, typically, only 20% of telephones are off-hook at any one time. This achieves economies in terms of the number of nodes required. However it also assumes that, on nodes where there is a heavy demand, extra bandwidth can be both found and retained. Otherwise a given user's perception of performance would be adversely affected if he happens to be attached on a node where there are heavy users. A significant reduction in the number of attached telephones is necessary on such LAN's to avoid this situation.

3.3 Token Ring and Slotted Ring Implementations

The principal that each node must have the opportunity to transmit in every interval T means that the required access protocol is essentially non-random. Both slotted and token polled access techniques are suitable. With token-passing, the technique is to ensure that the token comes back at least once in each interval T. This time, called the token rotation time (TRT), is controlled by setting the maximum length of time that the token can be held by each node. The latter time interval can have a different value on each node according to demand, ie it is set long enough to allow all of the delay-critical packets which have accumulated in an interval T to be transmitted. Load can be added and the maximum time the token is held by a node can be increased provided this does not cause the threshold T to be exceeded.

This is, in outline, the proposal before the American National Standards Institute for the high-speed LAN standard, the so-called Fibre Distributed Data Interface (FDDI). There may also be an attempt to standardize the slotted ring proposal [2] for high-speed LAN's. This achieves the same control of delay by allowing each node to transmit up to 'd' mini-packets, where 'd' varies from node to node. After all nodes have had the opportunity to transmit, the 'd' counters are reset by a special slot which travels once round the ring. Again the purpose is to ensure that the counters are reset at least once in each interval T by controlling the magnitude of the values 'd'. Thus load can be added and 'd' subsequently increased on a node provided this does not cause T to be exceeded.

An important performance difference between the slotted and token high-speed ring proposals is the access delay time,ie the time which an active node must wait before it is allowed to transmit. For any given value of T, the limiting case for the token ring is just one re-appearance of the token every T milli-seconds. Each node's transmissions are then spaced T milli-seconds apart. For the slotted ring, the limiting case of one reset every T milli-seconds does not imply that transmissions are spaced T milli-seconds apart. Instead, transmissions occur throughout the interval whenever empty slots arrive at a node. The result is that, for small packets (ie the packet length is not bigger than the slot length) the mean packet transfer time is significantly reduced below the value T/2, which is the approximate limiting value of the mean delay on the token ring.

Figure 2 illustrates simulation results of the delay experienced by voice packets on a slotted ring with a channel rate of 34 Mbit/sec and using the protocol defined by [2]. The mean time between successive resets of the 'd' counters is 385 microseconds whereas the mean access delay time (the time that an assembled packet has to wait until the start of its transmission) is only 26 microseconds. Less than 0.5% of all packets were delayed by more than 150 microseconds.For a mean token rotation time of 385 microseconds, the mean packet delay time would be approximately 190 microseconds. The reduced delay experienced on the slotted ring is important in wide area network applications where the connection crosses several LAN's. To keep within a given end-to-end delay for voice connections, the slotted access method is the best choice.

A further disadvantage of the FDDI token ring is that packets have a minimum overhead of 160 bits of control and routing information. Thus for voice packets, in which the data field only has around 128 bits for an acceptable packet assembly delay, the overhead is over 50% of the total transmission. Partly because of these performance issues the specification of the high-speed LAN standard has moved towards a hybrid arrangement with circuit-switched voice. However, another powerful reason is that the local network must interface to a circuit-switched public network for its voice connections. This provides a strong incentive to keep the transmission of voice in the same format on the local network.

4 SUMMARY

We have seen that Ethernet and token-passing rings have very similar performance characteristics at low channel rates (ie for channel rates of approximately 1 Mbit/sec). They are both ideally suited to situations where the load consists of long packets. However, polled access protocols in general are better suited to situations where there are a large number of attached nodes. At the higher transmission rates, the need to take into account a sustained high-bandwidth demand from a node favours polled access. Both slotted and token polled access methods can be applied. However, the slotted ring is better suited to situations where the dominant load consists of short packets, especially voice, where it offers a significant reduction in access delays. For these reasons, the standardisation of high-speed token rings has moved towards hybrid systems. The slotted ring, on the other hand, remains a promising area for the development of LAN's which provide for the full integration of services.

It can be seen that there are many complex factors to be considered by the engineer whose job it is to recommend a particular LAN system for a particular environment. Because of the wide area of applicability of LAN's to data processing, process control and the 'office of the future', it is likely that these decisions will have to be faced many times by many individuals. Therefore a thorough training in this subject is of considerable importance in producing engineers who can make informed judgements on these issues.

5 REFERENCES

1. TANENBAUM, A.S., 'Computer Networks' Englewood Cliffs, N.J., Prentice-Hall, 1981

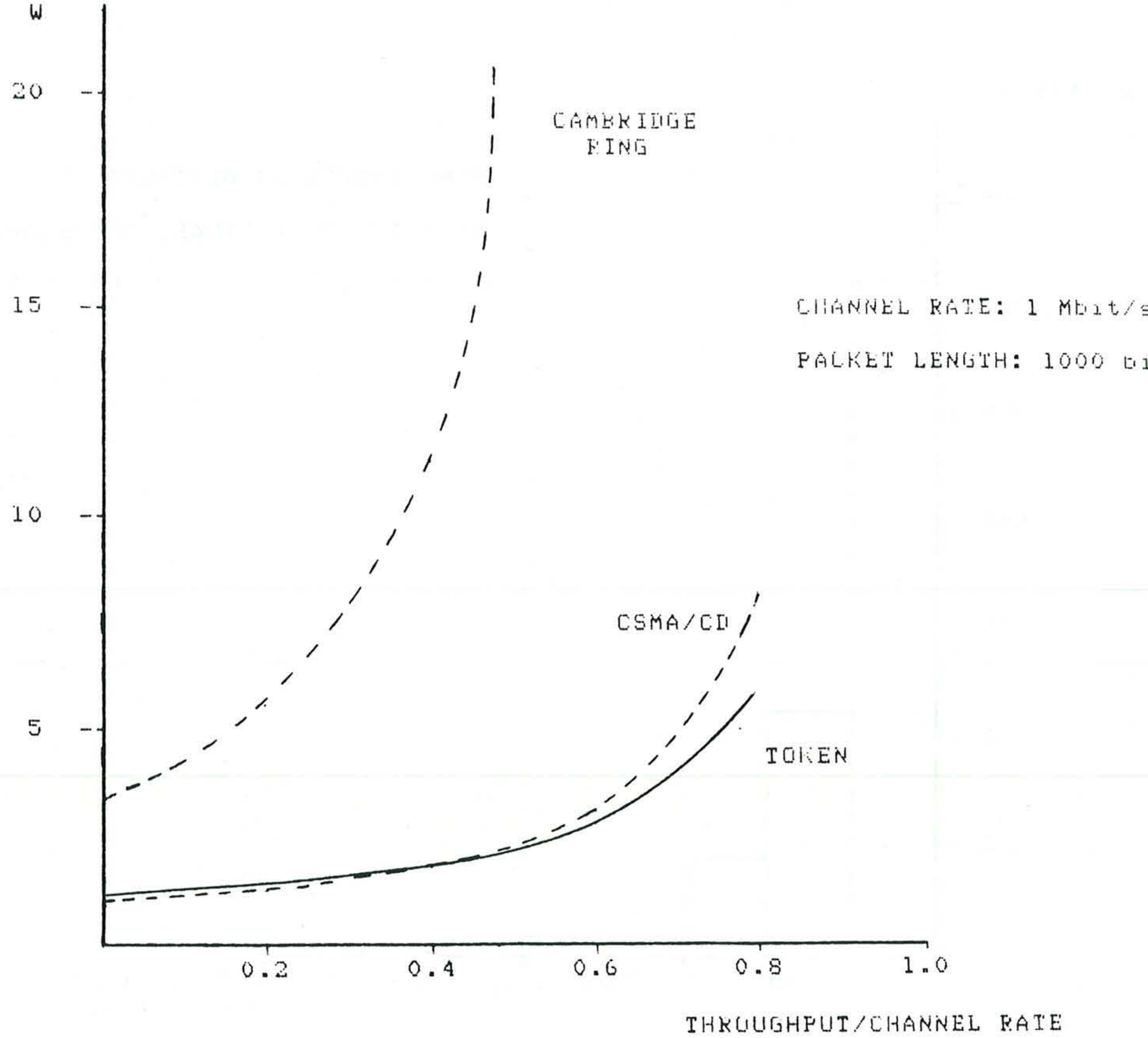
2. ADAMS, J.L., and FALCONER, R.M. 'Orwell: A Protocol for Carrying Integrated Services on a Digital Communications Ring' Elec. Letters, vol 20, pp 970-971, November 1984

3. BUX, W. 'Local-Area Subnetworks: A Performance Comparison' IEEE Trans. Commun., vol COM-29, pp 1465-1473, October 1981

6 GLOSSARY

The principal symbols used in this paper are summarized:-

C	The channel rate in bits/sec
d	A parameter of the Orwell slotted ring protocol [2]. Up to 'd' mini-packets may be transmitted by a node without any restrictions being imposed on its access to empty slots. A counter is supplied on each node to count the number of mini-packets transmitted. If this counter is not reset before it reaches the permitted maximum value 'd', the node may not transmit any further mini-packets until it is reset. A reset is performed by a specially marked slot travelling round the ring.
E	The efficiency of bandwidth utilisation.
L	The end-to-end propagation delay on the Ethernet cable.
N	The number of attached nodes on the LAN which are active, ie requiring to transmit.
P	The mean packet length in bits.
R	The number of bits contained in the ring circumference of a token ring LAN.
T	A threshold controlling the time between successive transmissions on any node. A node is allowed to transmit at least once in each interval T.
W	The mean time to transfer a packet from the source to the destination node measured as a scalar multiple of the mean time to transmit a packet.



CHANNEL RATE: 1 Mbit/s

PACKET LENGTH: 1000 bi

FIGURE 1: TRANSFER DELAY CHARACTERISTICS FOR THREE DIFFERENT TYPES OF ACCESS PROTOCOL

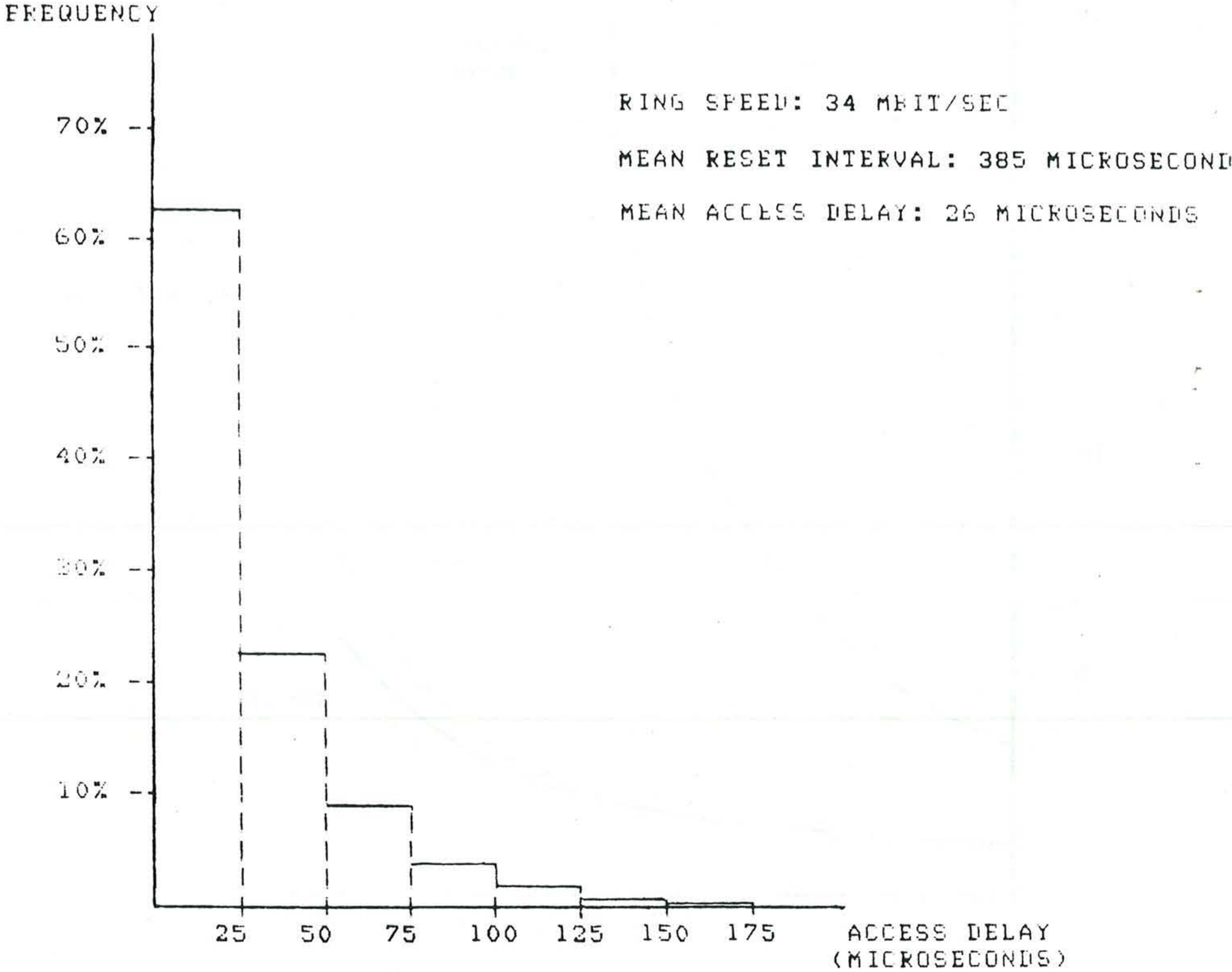


FIGURE 2: VOICE PACKET DELAYS ON THE ORWELL SLOTTED RING