The Demand Priority MAC Protocol

Greg Watson, Alan Albrecht, Joe Curico,
Dan Dove, Steve Goody, John Grinham,
Michael Spratt, Pat Thaler
Networks and Communications Laboratory
HP Laboratories Bristol
HPL-94-58
June, 1994

IEEE 802.12, MAC protocol, demand priority, multimedia

We present the motivation behind the development of a new 100Mbps LAN protocol that is currently being standardized as IEEE 802.12 demand priority MAC protocol. We describe the protocol in detail and explain how it provides a number of advantages over alternative protocols such as IEEE 802.3 (CSMA/CD), not the least of which is the ability to provide guaranteed services to applications.

Internal Accession Date Only
Submitted to the *IEEE Network Magazine*

1 Introduction

There is great demand for 100Mbps to the desktop, generated by the availability of high speed LAN servers for personal computers. File servers are often equipped with multiple LAN cards in order to prevent the network from becoming the system bottleneck; unfortunately this results in the need for more complex and sophisticated network management than would be required if a single network card could be used. However, a new technology needs to offer more than just 100Mbps. To succeed in the LAN marketplace a new LAN technology must be very cost competitive with the established LANs, such as Ethernet and Token Ring, while also providing backwards compatibility with existing network software. In this paper we describe a new 100Mbps LAN technology which has these characteristics. This technology is being defined as an open standard within the IEEE Project 802.12 Demand Priority[1] group which was formally established in July of 1993.

Two important objectives were established for this LAN technology: first, it should be able to use the unshielded twisted pair (UTP) wiring found in a large number of installations and, in particular, to use the same wiring as defined for use in 10Base-T[1]. This objective was later extended to encompass support for the shielded twisted pair (STP) used for IEEE 802.5. This will enable the majority of current LAN users to benefit from their enormous investment in cable plant. The second objective was that the network should support new applications, such as video conferencing and remote training, while also providing backwards compatibility with the massive installed software base.

Both objectives have been met. The network will support UTP Categories 3 (voice grade),4 and 5 (data grade), as well as STP and fibre. The Demand Priority MAC protocol provides two priority levels in order to support delay-sensitive applications. The 802.12 Draft Standard defines the use of two frame types, either IEEE 802.3 or IEEE 802.5, although it is likely that any particular network would use only a single frame format. In addition, large networks can be constructed by supporting a topology in which multiple hubs can be cascaded in a tree structure without the need for bridges to be placed between them.

In section 2 we describe the Demand Priority protocol. The basic protocol is extended in section 3 so that multiple hubs can be interconnected to form a single logical LAN. Section 4 presents two ways that guaranteed bandwidth and bounded delay can be provided using the Demand Priority protocol, and section 5 explains the results from some simulations of a Demand Priority LAN.

2 The Demand Priority MAC protocol

The Demand Priority protocol has been optimised to support the hierarchical wiring structures that are widely installed. Typically, cables are run from individual desktops to a wiring closet on each office floor. These wiring closets are then interconnected to another closet which is connected to the public network. This wiring scheme provides greater flexibility and security as well as better fault isolation capabilities than the distributed wiring of 10Base5 or 10Base2 (aka thick-net and thin-net). The enormous success of the 10Base-T version of IEEE 802.3 is attributed to its use over this wiring scheme.

An initial proposal for a 100Mbps network was to use the CSMA/CD protocol which is used in IEEE 802.3 and the original Ethernet. There are two reasons why CSMA/CD is not appropriate at 100Mbps:

---

1. The technology is also known as 100VG-AnyLAN (previously 100Base-VG)

- The CSMA/CD protocol allows two sources to send simultaneously but requires that any collision between frames is detected. This requirement limits the physical scope of the network to a few hundred meters because the minimum frame size is just 64 bytes (and much less for 802.5 frames), which has a transmission time of just five microseconds at 100Mbps. Consequently, if a 10Base-T LAN consisting of five hubs were adapted to use 100Mbps CSMA/CD, then several bridges would be needed to maintain the connectivity.
- CSMA/CD is non-deterministic and does not support multiple priorities. Consequently it is impossible for a CSMA/CD network to provide bandwidth or delay guarantees to an application.

Instead, the Demand Priority protocol was proposed. With Demand Priority a station issues a request to its local hub when it has a frame to transmit. (The hub is probably located close to the wiring closet.) The hub checks for requests from its attached stations and indicates to one station that it may transmit a frame. Each hub has a number of ports and each port may be connected to a station.

A simple signalling scheme is used to control access to the network. Figure 1 shows the sequence of events that occur when a station sends a frame. The example assumes that the link from hub to station is four UTP cables, in which case all four pairs carry data while a frame is being transmitted, but the hub and station each use two pairs for exchanging control signals between frames. Each arrow represents two pairs.

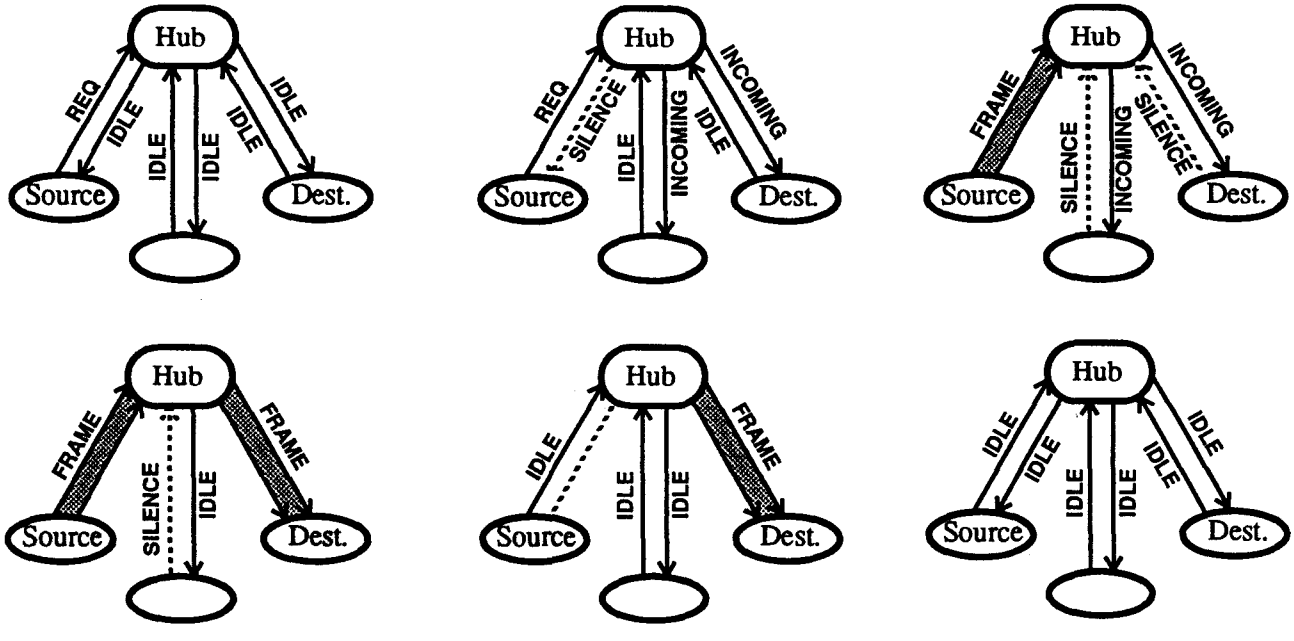

FIGURE 1. The Demand Priority MAC protocol

The sequence of events is as follows: initially the network is idle and so the stations send IDLE and the hub responds with IDLE.

A station sends a REQUEST to the hub, indicating that it wants to send a frame. The hub notes any requests and arbitrates between them on a round-robin basis. The hub then signals to the successful station that it may transmit. On a four UTP link the hub signals this by turning off its transmitters on the two pairs it was using to send IDLE. The station observes the silence on the two pairs and so can now use these two pairs without contending with the hub. (All four pairs are used so that each pair only carries 5B6B-encoded information at 30 Mbaud, which maintains electrical emissions and susceptibility to noise at acceptable levels.)

At the instant the hub selects the source station, the hub also sends INCOMING to all other stations to warn them that they might receive a frame. When the stations observe INCOMING they switch off their transmitters which enables the hub to transfer the frame on all four pairs.

The source station transmits its frame to the hub. The hub quickly receives enough of the frame to determine the destination station. Having determined the destination, the hub transfers the frame to the destination only and sends IDLE to all other stations. The hub does not operate in store-and-forward mode, but passes the frame on as soon as possible (also called cut-through), thus allowing high efficiency. Also, note that the hub will filter a unicast frame so that it is only received by the intended destination (and any stations that have indicated to the hub that they wish to receive all frames).

When the source completes the transmission it may send IDLE or it may issue another REQUEST if it has another frame to send. When the hub has finished transferring the incoming frame to the destination station it can immediately select the next station.

This protocol is very efficient because, unlike token based protocols, there is no token propagating around the network. The hub simply waits until it observes a REQUEST and then services it; the hub does not actively poll the end stations to see if they might have a frame to send.

2.1 The Physical layer

While not part of the MAC, the Physical layer is very important because it specifies the encoding scheme, as well as the types of cable that can be used. The first objective was to define a physical layer for category 3 (voice-grade) UTP. This has been met through the use of a 5B/6B code in which five data bits are encoded as six transmission bits. Because all four pairs are used this means that each pair carries only 30Mbaud which limits the degree to which the cables radiate energy and the degree to which they are susceptible to external noise.

The 5B6B coding scheme is used in conjunction with quartet signalling. The idea behind quartet signalling is that the 6-bit encoded values are skewed in time across the four pairs. This has certain properties which, when used with a carefully chosen 5B6B scheme, enable the stringent IEEE 802 error detection requirements to be met and exceeded.

2.2 Fairness

The Demand Priority protocol is fair in that the hub arbitrates between requests using a round-robin schedule. Hence no station will be able to send two frames before all other stations have had the opportunity to send at least one frame.

The hub makes its decision at the instant that it completes the transmission of the previous frame (or when it receives the first REQUEST if the network is idle). At the same instant the hub forgets all other requests. This is done so that a spurious request, whether caused by noise on the link or a station that has 'changed its mind', will not cause a problem. Stations that are not the destination of the current frame will quickly reassert their request. This works well for those stations which do not receive the frame, but the destination station has a potential prob-

lem: if the destination is continuously receiving frames then it may not obtain the opportunity to issue a REQUEST. Figure 2 shows two time-space diagrams that illustrate both the problem and its solution.

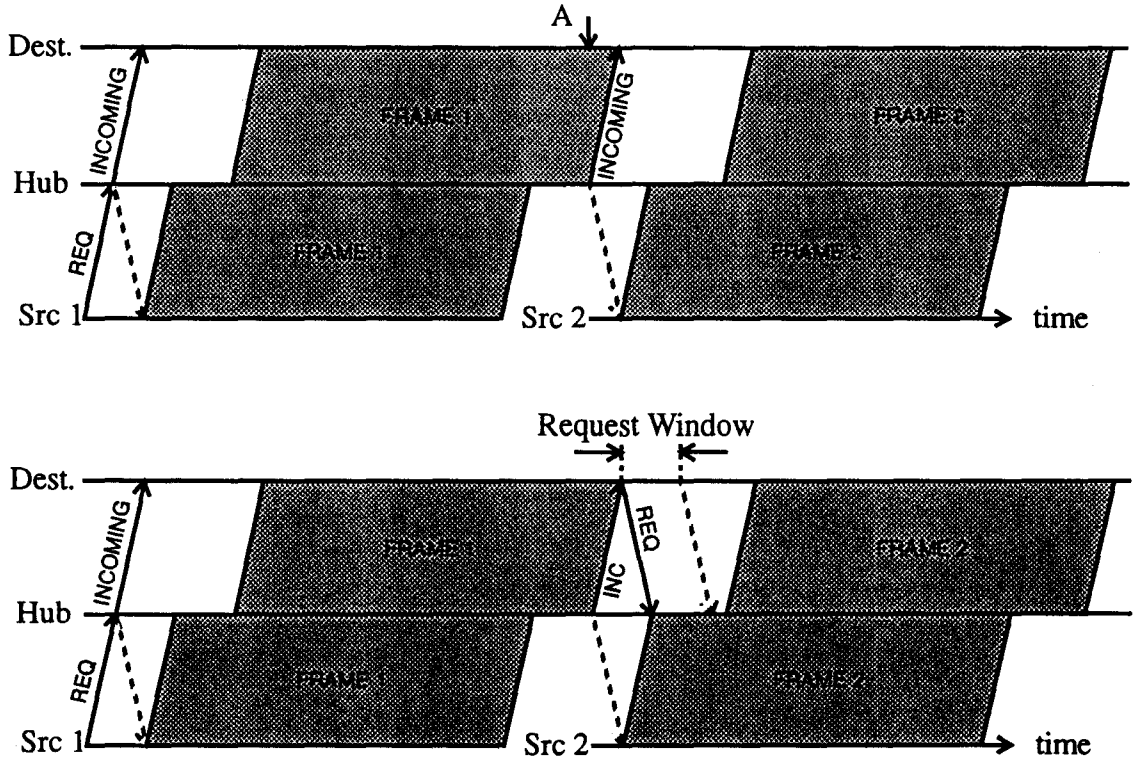

FIGURE 2. The destination must be able to issue a REQUEST for the period marked as the Request Window in order to ensure fairness.

Source 1 sends frame 1 to the destination. The frame is delayed slightly as it passes through the hub so that the hub can decode the destination address. The frame is then passed on to the destination. At time A the hub has completed the transmission of the frame and selects source 2. The selection process occurs very rapidly and so the hub sends out INCOMING almost immediately after the preceding frame. This does not give the destination sufficient time to send its REQUEST.

The solution is for the destination station to ignore INCOMING for a short period of time called the Request Window. This gives the destination enough time to send REQUEST so that it will be detected reliably by the hub. This does not reduce the efficiency of the network because the hub must ensure a minimum inter-frame delay to allow for the maximum variance in link length (and thus delay) between stations. This inter-frame delay is longer than the Request Window.

2.3 Two priority levels

It is very simple to extend the protocol to support two priorities by providing two REQUEST signals: a normal priority request (REQ_N) and a high priority request (REQ_H). We propose that normal priority is used for current applications such as file transfers and remote print spooling, hence the term 'normal'. The high priority should be used by delay-sensitive traffic such as voice or video.

The two priorities are absolute - a hub will always service a high priority request before a normal priority request, even if the high priority request is always from the same source. The hub services each priority level in round-

robin order, and it remembers the port that was last served at any given priority so that access to the network is fair. One situation that may arise is a station might inadvertently overload the network with high priority traffic, effectively preventing the normal priority traffic from gaining access. The current proposal to overcome this is to promote a normal priority request to a high priority request if it has waited longer than some fixed amount of time such as 250ms. This ensures that normal priority traffic will always obtain at least a minimum amount of bandwidth. Note that the overload condition is considered abnormal and should not occur if a bandwidth allocator is used (see section 4).

Consider an eight port hub which services ports in ascending numerical order. The hub is servicing port 3 at normal priority and has another normal priority request waiting on port 7. The sequence of transmissions is shown in figure 3. During the transmission of the frame from port 3 a high priority request arrives at port 5. Once the transmission from port 3 is finished the hub will service the high priority request from port 5 - the high priority does not preempt the normal priority transmission. While port 5 is transmitting there is a normal priority request from port 1. At the end of the high priority transmission the hub sees normal priority requests at ports 1 and 7. Because the last normal priority transmission was from port 3 then the hub continues the round-robin cycle and services port 7.

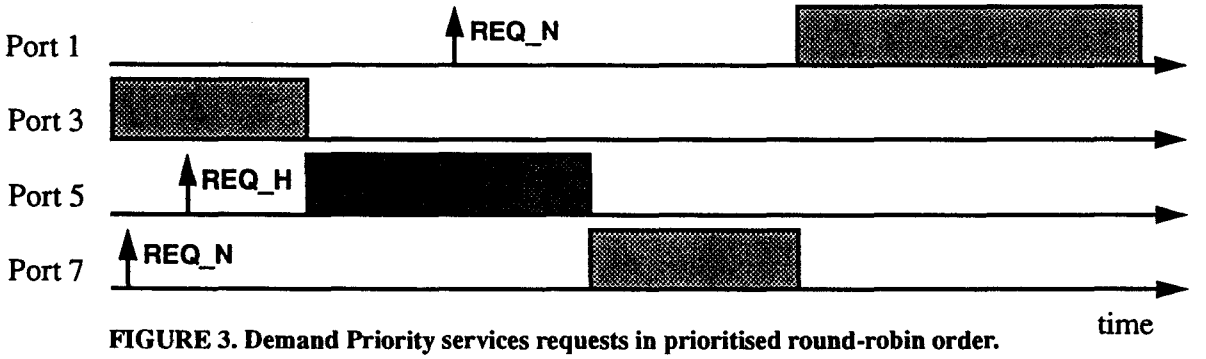

FIGURE 3. Demand Priority services requests in prioritised round-robin order.

The order in which the hub observes requests is not relevant - the decision as to which port is served next is determined solely by the list of outstanding requests and the current value of the round-robin pointer. This means that if, during a frame transmission from port 1, requests arrive from ports 5, 3 and 2, in that order, then the ports will be served in round-robin order, namely 2, 3, and then 5.

2.4 Training

An important feature of the network is that a hub will only forward a frame to its intended destination, and not to all stations, which provides improved network security over IEEE 802.3 and IEEE 802.5 LANs. To do this the hub must store the MAC address of each station. The MAC address, and other information, is exchanged between the station and the hub during the training process.

Training is performed when the station first joins the network and may be repeated later if either the hub or the station detects a problem. Training achieves two results. First, it enables both the hub and station to be confident that the link meets a minimum quality level. The link is deemed to be useable after the station and hub exchange a number of special training frames in succession and without error.

The second result is that training is used to exchange information between the hub and the station. One element of information exchanged is the station's 48-bit MAC address. In addition, the station can indicate whether it is a LAN bridge, whether it wishes to receive all frames (promiscuous mode), and what type of frame it will use: IEEE 802.3 or 802.5. The hub responds to the stations request and may grant the request or indicate that some service cannot be provided. While a station may request to receive all frames, the hub may refuse this request, perhaps

because that it is the policy established by the network administrator. The hub may also detect that the station is attempting to use a MAC address that is already in use, and can indicate this to the station.

Training frames have a destination address of zero so they will not be received by other stations. A station receives only training frames until it has successfully completed the training operation.

3 Cascading multiple hubs

Many users will want to interconnect a number of hubs together to form a single, extended LAN. To provide this facility we have extended the core Demand Priority protocol to enable hubs to be cascaded in a tree structure, as shown in figure 4. A special cascade port is added to hubs. This cascade port can *only* be used to connect to an upper level hub, and never to a station. When a hub receives a frame from an end station the hub always sends the frame to its upper level hub and to any attached hubs. Hubs may be interconnected by long fibre links, potentially up to 2km.

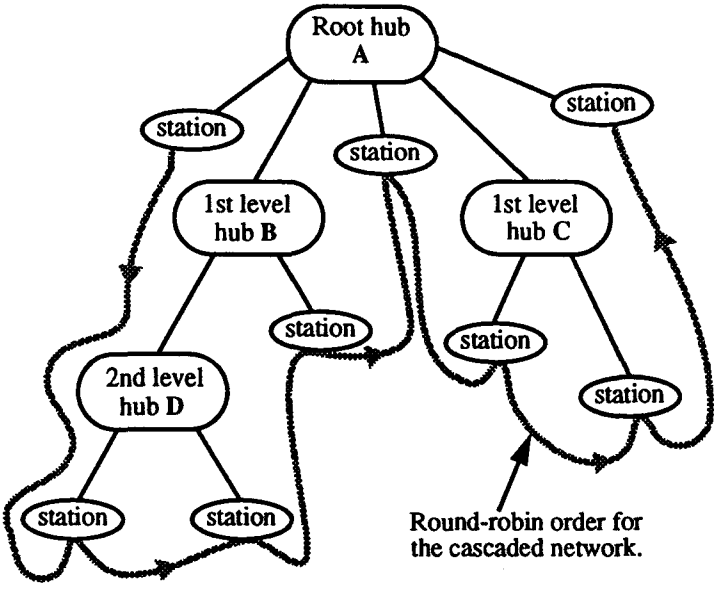

FIGURE 4. Multiple hubs can be interconnected to form an extenbded LAN.

Instead of a single local round-robin domain there is now a single domain that encompasses every node. To obtain the order in which stations are serviced we simply 'draw' around the network as shown by the shaded line in the figure. The order in which the stations are touched by the line is the order in which they will be serviced (assuming they have outstanding requests). Notice that in this case the stations attached to the second level hub will be serviced before the station attached to the first level hub 'B'.

In order to create a single round-robin domain we introduce the notion of 'control'. A hub is said to have control if it is servicing an attached station. If the entire network is idle then the root hub has control. In order to acquire control a lower level hub must send a request to its upper level hub. So, when a hub sees a request on a port the hub then reflects that request up the cascade link to the upper hub. A hub always makes a request at the level of the highest priority request currently observed on its lower ports.

A hub knows whether any given port is connected to a station or to another hub - this information is exchanged during training. Consequently, a hub knows whether to expect just one frame (from a station) or potentially several (from a lower hub). A hub can only perform one cycle of the round-robin sequence before returning control to the upper hub.

So the cascaded network behaves in the same way as though it were a single very large hub. The interesting case arises when a hub with a high priority request needs to preempt another hub that is currently servicing low priority requests. Consider a hub H which observes a high priority request while a hub N has control and is servicing a number of normal priority requests. In this situation hub H with the high priority request will simply propagate the high priority request to its upper hub. This request will continue upwards until it reaches some hub X that earlier passed control to hub N that is servicing its normal priority requests. Hub X will then issue a PREEMPT signal to hub N, perhaps via some intermediate hubs. When hub N observes the PREEMPT it will return control to its upper hub at the end of its current transmission and it will indicate whether or not its normal priority round-robin service was complete.

The critical aspect of the preemption situation is that once the high priority requests have been serviced then normal priority service should resume at the point where it was preempted. If this is not done then stations can be denied access to the network. Normal priority service resumes at the preempted hub because the hubs keep their normal priority round-robin pointers set to the last port that made a normal priority request. In addition, hubs remember if the last port was preempted, in which case they do not advance the normal priority pointer but instead they resume service at the same port.

With this scheme the two priority round-robin scheme can be extended over an arbitrarily large network.

4 Guaranteed service

With two priority levels the network can provide a service that guarantees bandwidth and bounds the access delay[2]. There are two different ways by which such a service can be provided[3]. The first is to exploit knowledge of a given network configuration in conjunction with two properties of the protocol. The second method is to introduce a bandwidth allocator and to use some form of access control. We expand on these two approaches below.

4.1 Using the protocol

The prioritised round-robin ordering used by Demand Priority has two useful properties. The first is that the available bandwidth will automatically be shared evenly among all stations currently active at the highest priority. If high priority traffic does not use all of the bandwidth then the remaining bandwidth is shared evenly among stations with normal priority traffic. This can be used to determine the bandwidth available to a station using high priority traffic. For example, a network with 32 stations could guarantee to provide a minimum of about 3Mbps to each station. We note that it is important to always retain some bandwidth for normal priority traffic so a more sensible configuration might restrict 32 stations to 2Mbps each, leaving about 36 Mbps of bandwidth for normal priority traffic.

The second useful property of the protocol is that it is easy to calculate the worst case access delay for a station. For a hub with N stations using high priority traffic the worst case access delay for one of those stations will be $N.F_{max}$ where F_{max} is the transmission time of a maximum size frame. One might expect the worst case delay to be $(N-1).F_{max}$ but the worst case occurs when a high priority frame is generated at every station just as the hub starts to service a normal priority request. For a hub with 32 stations and 1500 byte frames this worst case delay would be four milliseconds.

This method of providing a guaranteed service can be very attractive to users who want to use a local video server, for example. The video server and its clients would use high priority while other applications use normal priority. One attraction of this method is that there is no need for any additional infrastructure such as a bandwidth allocator. Those stations which only have applications using normal priority traffic do not need to be changed at all, and will operate totally unaware that a high priority service is provided to some other stations. The disadvantage of this method is that the bandwidth and delay depend on the actual network size and configuration.

4.2 Bandwidth allocator and access control

The second method requires the presence of a bandwidth allocator somewhere on the network. The bandwidth allocator provides a centralised bandwidth management service to which an application can apply for bandwidth. The allocator will either grant or refuse the request depending on how much bandwidth is still available. However, a bandwidth allocator is not sufficient and every station with high priority traffic must limit the amount of traffic it sends in any specified period. For this we propose a mechanism we call the Target Transmission Time, or TTT.

The TTT specifies a period over which all stations will honour their bandwidth allocations. As such, the TTT must represent the smallest delay needed by any application, similar to the FDDI Target Token Rotation Time[4], and a figure of 10ms seems reasonable for many LAN-based multimedia applications. So, if a station is granted 10Mbps and the TTT is set to 10ms then the station may not send more then 100 kbits of high priority traffic in *any* period of 10ms[2]. The TTT is, in a sense, a contract: if stations do not exceed their bandwidth allocation in any period TTT then they will always be able to transmit within the period TTT.

We propose that the TTT is enforced by the network driver in each station. The driver must maintain knowledge of when frames were transmitted in the previous period TTT in order to calculate when the next frame can be transmitted. If a station has not used its allocation then it can send a new frame immediately. If the station has exhausted its allocation then it must wait for a period less than TTT before it can send the next frame. In practice the driver controls the time when frames are submitted to the MAC protocol for transmission, and not when they are actually transmitted. We have implemented the TTT algorithm on a workstation in order to provide strict bandwidth control for an Ethernet driver. The code is quite short (100 lines or so) and imposes only a very small overhead.

One attraction of the TTT method is that it operates at the driver level and thus above the MAC layer, and hence is independent of the particular network. Consequently the TTT method could work equally well with IEEE 802.5 or FDDI.

5 Performance

In this section we examine the performance of high priority traffic as we increase the normal priority traffic. The simulator was produced using the Verilog hardware description language. The various simulations were run for some 100,000 frames with results taken only from the central portion of each run in order to avoid the anomalies that occur at start-up and shut-down.

The network configuration is shown in figure 5. There are three 15-port hubs and all ports are used. The stations are connected to their hubs via 100m links whereas the hubs are interconnected using 200m links. Six stations, three on each of hubs B and C, are allocated high priority bandwidth such that they send a block of eight maximum size frames every 10ms. This is equivalent to about 9.7Mbps and substantially more than MPEG[5] encoded video. This gives a total of about 58Mbps of high priority traffic. All stations, including those with high priority traffic, send some maximum size frames at normal priority. We vary the total amount of normal priority traffic and measure the access delays.

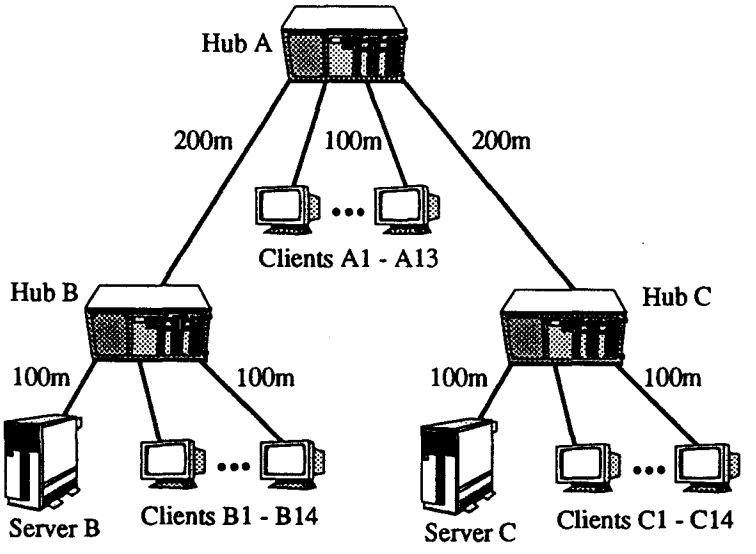

FIGURE 5. The network configuration used in the simulations.

Figure 6 shows the mean and maximum access delays that were observed for both high and normal priority traffic. The access delay is the time that a frame spends at the head of its transmission queue awaiting access to the network. It is clear from figure 6 that the delays encountered by the high priority traffic are almost independent of the normal priority traffic. Even when the total offered load on the system is 93Mbps (35Mbps of normal priority) the high priority traffic still has a mean access delay of less than half a millisecond and a maximum access delay of less than 0.8ms. These delays are, of course, dependent on the number of stations that have high priority traffic, but it is clear that a number of high priority streams can be supported with a *guaranteed* low delay.

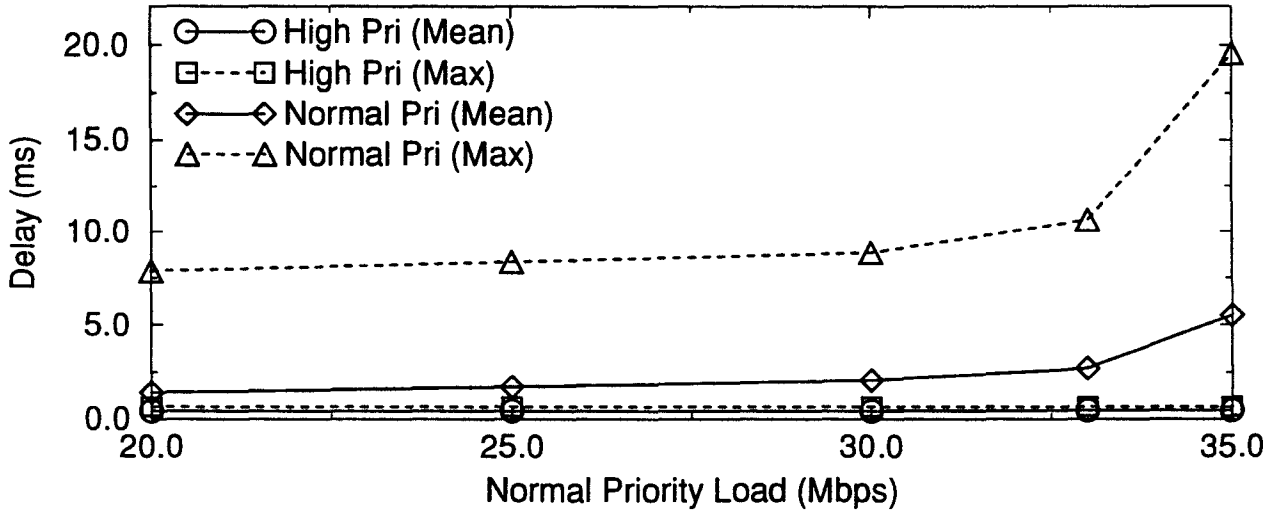

FIGURE 6. Access delays for normal and high priority traffic

6 Future challenges

The Demand Priority protocol provides services that meet the demands of delay sensitive traffic, but there is some work to be done before these services can be fully utilised. Part of this work is to develop the TTT (or some other mechanism) that can provide a simple and consistent interface that will enable higher layer protocols to access the services provided by LANs such as IEEE 802.12 and 802.5 as well as ANSI FDDI. Such work has been started by a group within IEEE 802.1. This group will also address the issues raised by sending multimedia traffic through bridges, such as how to maintain the service or priority level of a frame that passes from a LAN of one type to a LAN of a different type. The control of multicast traffic is another such issue.

There is a great deal of interest and activity in providing new services across the Internet - the Mbone[6] is an excellent example of this. Many researchers are addressing the problems associated with resource reservation in routers, and maintaining the quality of data flows across a large network. Our work is a step towards making the same facilities available at the local area network level so that services will be available end-to-end.

7 Conclusions

The Demand Priority MAC protocol, currently being standardised in IEEE 802.12, offers substantial benefits over the CSMA/CD protocol of IEEE 802.3. In particular, Demand Priority can provide guaranteed bandwidth and predictable access delay so that delay-sensitive applications can operate correctly regardless of the behaviour of other applications. In addition, the protocol enables multiple star-wired hubs to be interconnected to form a larger single network. By preserving both the current wiring infrastructure and investment in software, and by using the very simple Demand Priority MAC protocol, we expect that 100Mbps LANs will soon be as low-cost as 10Base-T is today.

[1] B.F. Gearing, "Building Wiring Standards and their impact on LANs," SPIE Vol. 1577, "High-Speed Fibre Networks and Channels", pp. 28-32., 1991.

[2] J. Grinham and M. Spratt, "IEEE 802.12 Demand Priority and Multimedia," Proceedings of the 4th International Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 75-86, Nov. 1993.

[3] J. Grinham, M. Spratt, "IEEE 802 Tutorial on Multimedia and 100Base-VG," presentation at IEEE 802 Plenary, May 1993 (contact mps@hplb.hpl.hp.com).

[4] Fiber Distributed Data Interface (FDDI) - Media Access Control, ISO 9314-2, 1989.

[5] D. Le Gall, "MPEG: A Video Compression Standard for Multimedia Applications", CACM VOl. 34, No. 4, pp. 46-58, April 1991.

[6] M.R. Macedonia and D.P. Brutzman, "Mbone Provides Audio and Video Across the Internet," IEEE Computer magazine, Vol. 27, No. 4, pp. 30-36, April 1994.