HEWLETT PACKARD

100 Base-T /IEEE 802.12/ Packet Switching

Greg Watson, Mart Molle*
Network Technology Department
HP Laboratories Bristol
HPL-96-58
April, 1996

100Base-T, IEEE 802.12, 100VG-AnyLAN, packet switching, IEEE 802.3, CSMA/CD, demand priority

Three recent LAN technologies look set to satisfy the ever-increasing demand for LAN bandwidth. Two of these technologies are 100Mb/s shared medium LANs: 100Base-T (aka IEEE 802.3 Fast Ethernet) and IEEE 802.12 (aka 100VG-AnyLAN or 100VG). The third technology is packet switching, which is really an extension of existing LAN bridge technology, but which offers excellent performance gains at very low cost. In this paper we describe the three technologies and provide a comparison between the two 100Mb/s LANs. We also present results that compare the measured performance of 100 Mb/s shared medium LANs with switched LANs. Given the "holy wars" between 100Base-T and 100VG-AnyLAN we have worked hard to present the technologies on an equal footing. One of the authors is from HP, and served on the 802.12 committee. The second author is the chairperson of an IEEE 802.3 subcommittee. We hope we have achieved our goal of presenting a balanced view of these technologies.

Internal Accession Date Only
**Computer Science Department, University of California*

100Base-T / IEEE 802.12 / Packet Switching

Mart Molle
Computer Science Department
University of California
Riverside CA 92521 U.S.A.

Greg Watson
Hewlett-Packard Laboratories
Bristol BS12 6QZ U.K.

1 Introduction

In 1990 it seemed that FDDI would be the only standard 100Mb/s shared-medium LAN technology. However, the IEEE 802 project has since developed two new 100Mb/s standards - 802.12 and 100BaseT - which offer equivalent data rates at much lower cost. In addition, packet switching has emerged as a third technology that is driving the evolution of LANs.

The initial motivation for developing a new 100 Mb/s LAN was the realization that a "faster Ethernet" would be easy to design, and much less expensive to build than FDDI. Moreover, if the new low cost LAN were compatible with 10 Mb/s 10Base-T (Ethernet) technology, then the new LAN would be easy to integrate into existing LANs and hence should share in the enormous success of 10Base-T.

Maintaining compatibility between 10Base-T and the new low cost 100 Mb/s LAN was not an easy goal to achieve, however. The easy part was deciding on supporting the same frame format and the same basic star-wiring architecture as 10Base-T. However, the ten-fold increase in speed meant that the CSMA/CD algorithm could only be retained at the cost of reducing the maximum topology by a factor of 10. This is the core compromise: the 802.3u group opted to keep CSMA/CD; 802.12 decided that a new access method was preferable.

In 100Base-T, the network topology was restricted so that the familiar CSMA/CD medium access control (MAC) algorithm from 10Base-T could be retained. However, because users were already using network segmentation to promote security and manageability, the group decided that these topology restrictions were not very important. Thus, the majority of the work in 100Base-T was focused at the physical layer. Three signalling schemes are defined: two for copper cables and one for optical fibre, enabling 100Base-T to take advantage of the high quality cabling now being installed.

In 802.12, the group decided that preserving the ability to handle large 10Base-T topologies without network segmention was an important enough goal for them to develop a new MAC protocol. The new protocol, called Demand Priority, avoids the end-to-end timing constraints imposed by CSMA/CD and also supports two priority classes. In addition, substantial work on new physical layer technologies was carried out so that existing 10Base-T networks can be upgraded to 802.12 without rewiring. The initial choice of frame format was to use the format defined by 802.3. Later, the 802.12 working group extended the definition to also include the use of the larger IEEE 802.5 frames, thus providing a similar upgrade path to the millions of users of 802.5 (Token Ring) technology. At any instant an 802.12 LAN must use either the 802.3 or 802.5 frame format and not both, although equipment can be designed to use either format.

The Demand Priority protocol was initially presented to 802.3 in 1992. In 1993 the IEEE 802 executive committee decided that both CSMA/CD and Demand Priority should be developed as 100 Mb/s LAN standards within Project 802. In 1995 the Demand Priority Access Method, Physical Layer and Repeater specification was published as IEEE 802.12 [8] and 100Base-T was published as supplement 802.3u [7].

Packet switching is not really a new technology because a switch is a multi-port bridge and, like a bridge, will comply with the forwarding and filtering rules that are well specified in IEEE 802.1d [6]. An important difference however is that today's "packet switches" are optimised to forward packets between ports rather than to filter (discard) them. This difference reflects the change in application: old bridges were used to impose traffic isolation between LAN segments whereas modern switches optimise bandwidth and latency between segments to provide high speed connectivity between any two points in the switched LAN.The demand for packet switches has grown enormously over the past two years, and looks like it will continue for years to come. This growth is driven by the fact that substantial performance gains are possible by installing a single high performance / low cost switch, without changing anything else in the network.

These three technologies are presented in the following sections, starting with 802.12 Demand Priority and continuing with 100Base-T. Some comparisons are made between these two 100 Mb/s shared LANs in section 4 while section 5 looks at packet switching and provides some measured performance data.

2 Demand Priority (IEEE 802.12)

2.1 Basic protocol

The MAC protocol used in 802.12 is called Demand Priority. Figure 1 illustrates the basic operation in a simple network consisting of a single repeater and several nodes. Before transmitting a frame a node first sends a Request signal to the repeater. The repeater arbitrates among all requests using a very simple round-robin algorithm, and issues a grant to one of the nodes. The granted node can then transmit a single frame up to the repeater. The repeater examines the destination address within the frame to determine the destination node and begins forwarding the frame to only that node. The repeater sends an IDLE signal to the remaining nodes. Multicast and broadcast frames are sent to all nodes. The repeater forwards the frame the instant that it knows where the frame should go - within a few microseconds of receiving the start of the frame.

A repeater only forwards a unicast frame to the specific destination and not to all nodes. This filtering is possible because the repeater knows the MAC address of all connected nodes. This address, together with other information, is exchanged between each node and the repeater during a training session when the node first starts. Note that an 802.12 repeater has far more intelligence than an 802.3 repeater: the 802.12 repeater actively manages the access to the network whereas an 802.3 repeater is a much more simplistic device which essentially acts only at the physical layer.

Many observers confuse the Demand Priority protocol with a token passing protocol, such as 802.5 or FDDI. This is not correct as Demand Priority has several subtleties beyond token passing. First, there is parallelism and asynchronous operation, since nodes may have many opportunities over the round-robin cycle to send a request signal to the repeater. Second, although the grant signal circulates among the repeaters like a token, each repeater only sends the grant to those of its end nodes, in round-robin order, that have registered a request for this cycle. Thus, an end node is only involved in the protocol if it has a frame to send, and no time is lost in circulating a token to nodes that do not have any frames.

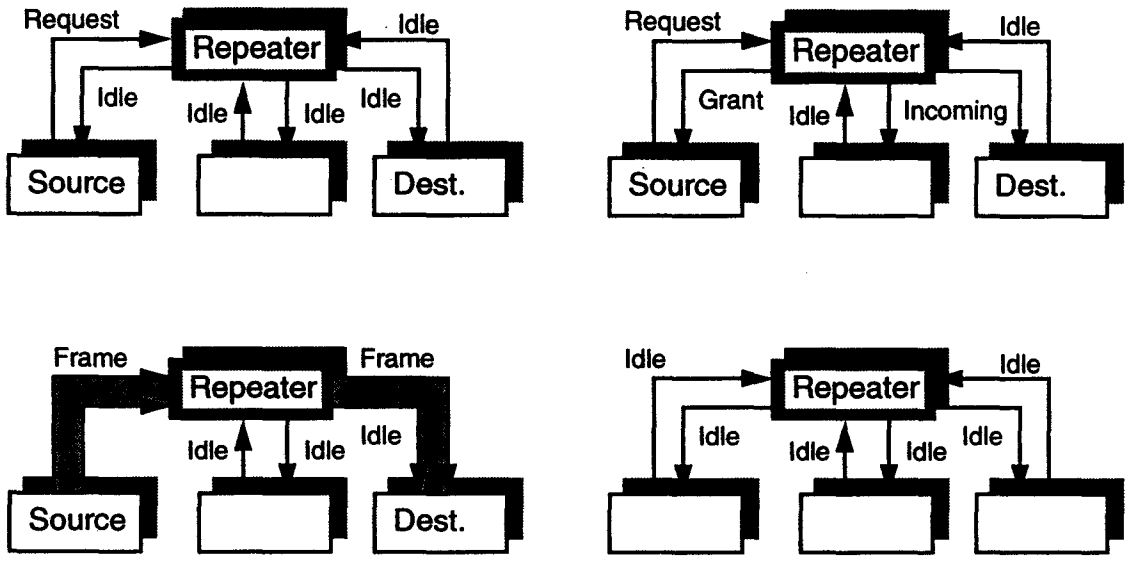

FIGURE 1. Basic operation of the Demand Priority MAC Protocol

2.2 Two Priorities

The Demand Priority protocol enables frames to be transmitted at one of two priorities, either normal or high priority. Normal priority is expected to be used for "normal" data such as file transfers, print jobs, email, etc., whereas high priority might be used for delay sensitive data such as video and voice in a videoconferencing session. IEEE 802.3 does not offer such a service whereas both 802.5 and ANSI FDDI offer multiple priority levels that can be used to distinguish traffic sent by different classes of applications.

A node sends either Request_Normal or Request_High according to the priority that is required. The repeater arbitrates between all requests on the basis that all high priority requests are serviced before any normal priority requests. Within each priority level the requests are serviced on a round-robin basis. So if nodes 1, 3 and 5 each have a single normal priority frame to send, and nodes 2 and 4 each have two high priority frames then the order of service will be 2-4-2-4 (high) and then 1-3-5. Note that when the round-robin cycle of a given priority level resumes after some interruption, it always continues from the point where it left off rather than starting over from the beginning or from the location of the last transmitter of the other priority class.

The Demand Priority protocol ensures that some minimal bandwidth is always available to normal priority. This is achieved by "promoting" a normal priority request to high priority if it has not been serviced within a certain period, currently between 200 and 300ms.

We have conducted some simple experiments that illustrate the benefit, in terms of end-to-end delay, that high priority traffic can offer. Figure 2 shows the experimental configuration. One computer, shown as the Test computer, was used to measure end-to-end delay. Between zero and three other computers were used to impose high priority traffic at 20Mb/s. Four other computers were used to impose normal priority traffic at a total load ranging from zero to one hundred percent of bandwidth. All computers are HP

9000/725 workstations using the HP-UX 9.05 operating system and with 32MB RAM and EISA 802.12 interface cards. All computers were connected to the repeater via 100m of Category 3 cable.

The end-to-end delay (Δt) was measured from the instant the generator program in the Test computer issued a packet transmit function to the instant the packet had been copied from the receiving network card to the consumer program in the same computer. The Test computer was equipped with two 802.12 interface cards, one to transmit and one to receive. By using the same computer for both functions we can avoid timing discrepancies that would arise if we had used two separate computers. The delay includes the time to transfer the packet from memory to the network card and back again to memory, as well as the necessary cache flushes and operating system overhead. There is no transport protocol - it is equivalent to a pure datagram transport such as UDP. Timing inaccuracies were minimised by ensuring that the workstation encountered no other interrupts between sending the packet and receiving it. The Test computer sends packets at a low mean rate - about 0.56 Mb/s - corresponding to constant rate compressed video.

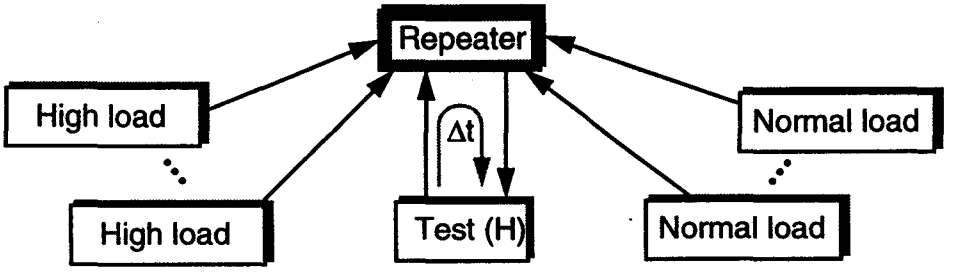

FIGURE 2. Basic operation of the Demand Priority MAC Protocol

Figure 3 shows the measured maximum delay for a total of zero, one, two and three high priority load generators, as well as the observed minimum delay (which is the same under all configurations). The minimum delay is about 300 microseconds. This consists of: 145 microseconds of packet copying (twice) and cache flushing, about 25 microseconds of context switching, and 130 microseconds of packet transmission time and overhead.

The maximum delay, with no other high priority loads, is the minimum plus 130 microseconds - one maximum packet time. This occurs when a normal priority transmission starts just before the Test computer can issue a high priority request. However it is clear that the actual normal priority load has no subsequent effect on the maxmimum end-to-end delay experienced by the Test computer. The remaining curves show that each additional high priority load adds one extra maximum packet time to the maximum delay - this is when the Test computer is the last high priority request to be served in the round-robin sequence.

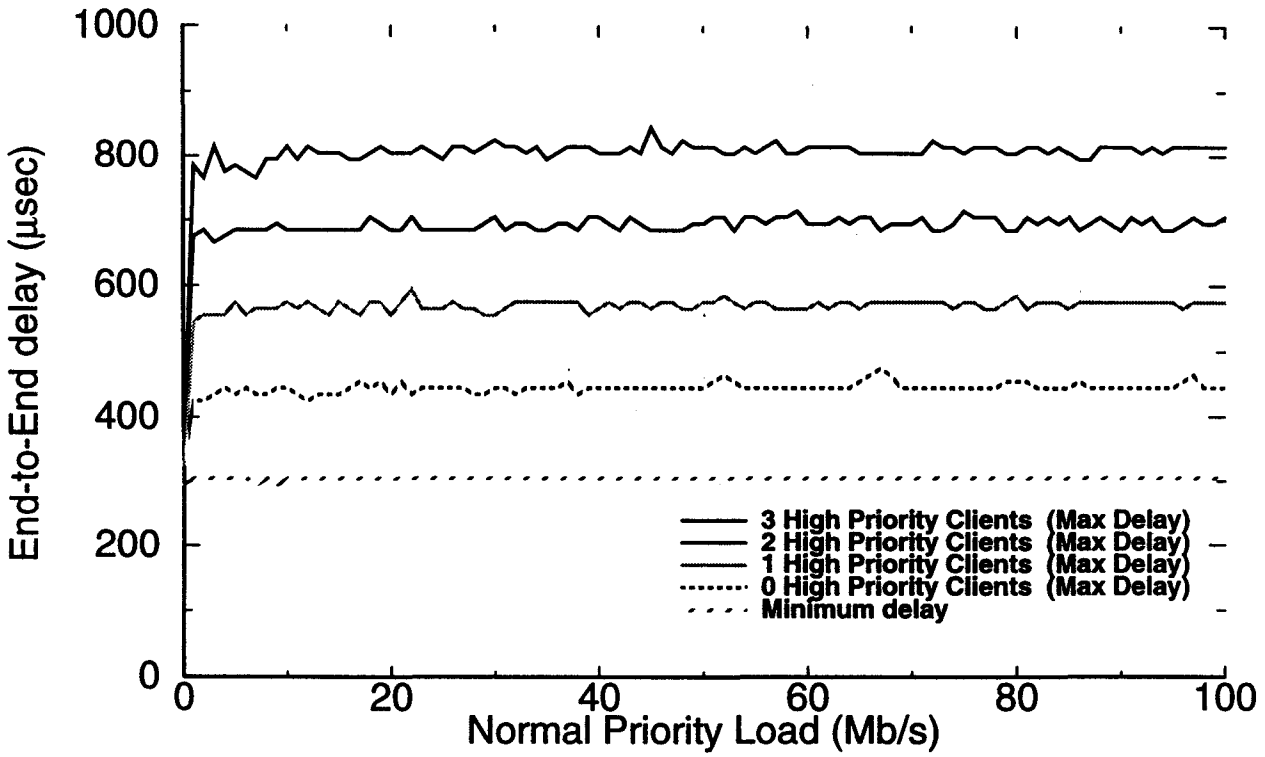

FIGURE 3. End-to-End delay observed by the Test computer sending at High Priorityl

2.3 Topologies

One of the objectives within 802.12 was to define a technology that could operate with all of the topologies that are legal under the 802.3 10Base-T standard. This includes cascaded networks where many repeaters are connected in a tree topology and which may span many hundreds of meters. To meet this objective the basic Demand Priority protocol was enhanced to operate over a large rooted tree topology such as the one shown in figure 4. Each repeater has many "down" links which connect to either a lower repeater or a node, and a single "up" link that connects to a higher repeater (except for the root repeater). In the single repeater topology described earlier the Demand Priority protocol operates a simple round-robin algorithm. In a multi-repeater topology this algorithm is distributed such that all nodes are collectively serviced in a single round-robin domain, maintaining the fairness of the simple topology. So, if all of the nodes in figure 4 have a normal priority request then they will be serviced according to their node number 1-2-3-... etc.

If a repeater receives a high priority request while another repeater is in the process of servicing normal priority requests then the first repeater can effectively interrupt the normal priority sequence in order to service its high priority request. Once the high priority requests have all been serviced then normal priority service will resume at the point that it was interrupted. This ensures that fairness is maintained even in a large topology with many repeaters.

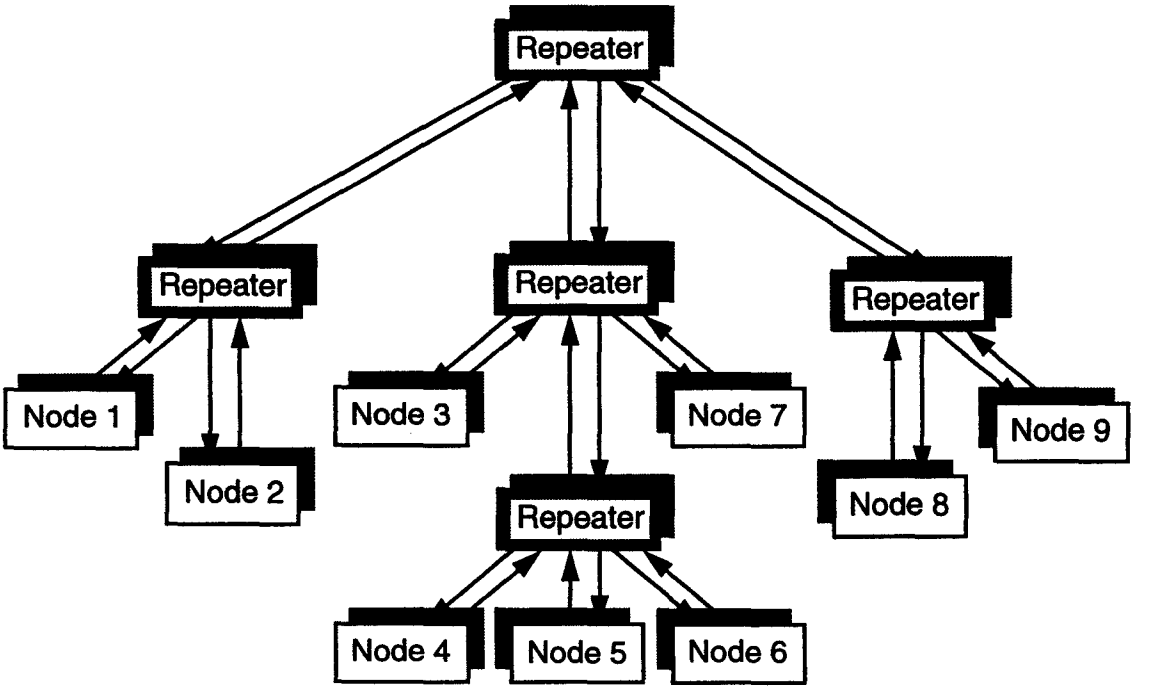

FIGURE 4. A Cascaded topology, showing 802.12's ability to operate over all 10Base-T topologies.

2.4 Physical layers

The 802.12 standard describes a 5B/6B coding scheme in which consecutive groups of 5 data bits are mapped into 6 code bits which are then sent over each of the four pairs. The mapping is carefully chosen to closely bound the imbalance between the number of 1s and 0s in any data sequence which makes it easier to design receivers than if the signal is unbalanced. Furthermore, the data is scrambled before being encoded; this is done to avoid long runs of particular symbols [12] which would otherwise focus the energy of the transmitted signal at particular frequencies and which might cause interference with other devices.

The current 802.12 standard defines several physical layers that support a variety of widely used cables. In particular, the standard supports the use of four-pair Category 3 (voice grade) UTP cable, which is the most widely used cabling for 10Base-T, although each 10Base-T link uses only two of the four pairs. When used over Category 3 UTP cable, the 802.12 packet data is transmitted on all four pairs. Additionally the standard supports Category 4 and 5 UTP as well as 100m of shielded twisted pair (STP) found in 802.5 environments, and also 62.5 micron multimode fibre for links up to two kilometres. See [1] and [2] for more details of the physical layer.

Moreover, 802.12 is currently the only 100 Mb/s standard that supports the use of bundled Category 3 cable, in which 25 pairs are bundled together within a single sheath. Such cables are often found in wiring closets to connect repeaters to patch panels. Bundled cables are more difficult to use at 100 Mb/s than unbundled cable because they exhibit higher crosstalk due to the large number of co-located wires. Consequently 802.12 restricts their use to ensure that only one data stream at a time is present in the bundle: bundled cable cannot be used for repeater-repeater or repeater-bridge links. Further, if the source and des-

tination nodes share the same bundle, and the frame is a multicast or broadcast frame, then the entire frame is buffered in the repeater before it is forwarded on the outgoing link.

3 100Base-T (IEEE 802.3u)

3.1 System Considerations

100Base-T specifies the operation of IEEE 802.3 CSMA/CD networks at 100 Mb/s. We assume that the reader is familiar with the basic operation of CSMA/CD. For our purposes, the key feature is that each node is responsible for scheduling its own transmissions via the optimistic policy of simply sending the frame when it sees that the network is quiet (via carrier sensing), and rescheduling using independent random delays if it sees that its attempt was unsuccessful (via collision detection).

Because of the tremendous popularity of 10 Mb/s 802.3 networks, CSMA/CD is a familiar and well-tested MAC algorithm. However, since the worst-case round trip propagation delay (known as the "slot time") is a critical timing parameter for the CSMA/CD algorithm, the ten-fold increase in data rate necessitated a similar reduction in the maximum diameter of a collision domain in comparison to 10Base-T. Although this limitation in size for repeater-based networks is clearly a drawback, its significance is reduced by the fact that large networks are often segmented via bridges, switches or routers for reasons such as bandwidth management, broadcast management, workgroup partitioning, and security considerations.

In practice the 100m limitation from repeater to hub may not be an important issue because the current ISO/IEC 11801 building wiring standards recommend a star-wiring plan in which the run length of the cable from each data jack to the equipment closet is less than 100 meters.

3.2 Physical Layer Standards

Several distinct physical layer transmission schemes are supported, so a new Media Independent Interface (MII) has been defined between the MAC and physical layers. If the MII is exposed, then it defines a standard way to support changeable media, for example, plug-in modules to allow the device to operate over two-pair (Category 5 UTP or STP) cable, or over four-pair (Category 3, 4 or 5 UTP) cable, or over 2 optical fibres.

3.3 100Base-X

The approach in the 100Base-X family was to take the existing physical layer standard from FDDI and adapt it to Ethernet. Since the FDDI standard supports both optical fibre and copper cable, the 100Base-X family includes 100Base-FX (which runs over 62.5/125 μm. multimode optical fibre, using a duplex SC, MIC or ST connector) and 100Base-TX (which runs over Category 5 UTP cable or STP cable). The basic features of 100Base-X are inherited from the FDDI standard, including the use of its 4B/5B block coding scheme and full duplex signalling. The 4B/5B signal stream is scrambled to ensure EMC compliance, and the scrambled sequence is then coded via an MLT3 coder, producing a three level signal on each pair. Although 100Base-TX requires Category 5 UTP cabling, it uses only two pair and supports full duplex transmission at the physical layer. Most Category 5 cable has four pairs and the standard recommends common-mode termination of the two unused pairs to meet EMC limits, which renders them unuseable for high speed data traffic.

Similar to 10Base-T, the entire 100 Mb/s data stream in 100Base-X is transmitted on one link segment in each direction. This makes collision detection very easy to do via simple digital logic that looks for the arrival of an inbound signal on one link segment at the same time as an outbound signal is being transmitted on the other link segment. It also makes full duplex operation very easy to support. 100Base-X also has one very unusual design feature for a network that uses "carrier sensing" at the MAC layer, namely that each link segment carries a continuous sequence of data symbols at all times. If there is not supposed to be any data present, then a special reserved idle symbol is sent. This helps simplify the hardware design and makes the timing requirements easier to meet.

3.4 100Base-T4

The approach in 100Base-T4 was to develop a new physical layer standard for transporting 802.3 frames over UTP cable. The advantage is that 100Base-T4 requires only Category 3 (voice grade) cable, although Category 4 and 5 cable is perfectly acceptable, of course. The disadvantage is that, like 802.12, all four pairs are required and thus cannot be used on installations that only have two pairs available. An 8B/6T physical transmission scheme is used, which first converts each 8 bit data byte into a block of 6 DC-balanced ternary symbols, and then stripes the consecutive symbol blocks across 3 out of the 4 pairs to create a 25MHz ternary data stream on each active UTP cable. Since one "primary" pair is reserved for each direction, collision detection is easy, but full duplex operation is not possible. Note that, unlike 100Base-TX, there is no signal present on the wire when the network is idle.

3.5 Autonegotiation of Capabilities

Because the same RJ45 connector is used for several different physical signalling schemes, 100Base-T allows the two devices connected to the ends of a UTP link to advertise their respective networking capabilities to each other and then select their highest common operating mode. Autonegotiation works by replacing the 10Base-T link integrity test pulse, by a coded burst of fast link pulses that describes the set of capabilities (10Base-T/100Base-TX/100Base-T4, half vs. full duplex, etc) that this device wishes to advertise now.

Autonegotiation allows developers to build more flexible products (such as dual speed 10/100 Mb/s network adapter cards), which in turn makes life easier for the users, since they don't have to set options manually on every device. However, verification of the quality of the intervening link segment (i.e., that its length is within the allowable limits, that all the required wire pairs are connected, and that the error rate is within specifications) is beyond the scope of the autonegotiation process.

3.6 Repeaters

Repeaters are basically dumb devices that copy bits from the incoming port to all others, or else supply a jam signal to all ports if more than one port has an incoming signal. However, there are still quite a lot of details to take care of inside a repeater, especially when it includes both 100Base-X and 100Base-T4 ports and therefore must translate the data between the sequential 4B/5B code in 100Base-X and the parallel 8B/6T code in 100Base-T4. In addition, repeaters support mechanisms for protecting the rest of the network from problems on one port, including jabber control (to terminate abnormally long input frames) and link partitioning (for disconnecting noisy links to protect the rest of the network from spurious signals). Repeater Types I and II have been defined with different limitations on their respective timing budgets.

3.7 Topologies

The well-known "five segment / four repeater" topology rule for 10 Mb/s 802.3 networks has been replaced by the following guidelines for 100Base-T networks:

(a) 2 nodes directly connected without a repeater;
(b) an N-node star with a single Type I repeater in the middle; or
(c) an N-node star with two Type II repeaters within 10m of each other.

More complex topologies are possible if they avoid using maximum length cables; they must be evaluated individually to make sure they stay within the bit budget. Of course, much larger topologies can be constructed through switching.

4 Comparison of 100Base-T and 802.12

Both 100Base-T and 802.12 provide certain advantages and disadvantages over one another, depending on the particular issues under consideration. It is important to quantify the differences, and compare them to one's needs, to see if they really matter. Moreover, even though there is no clear winner, it is obvious that these two technologies are driving down the cost of 100 Mb/s networking.

- Efficiency. Both approaches offer reasonably good performance in their respective normal operating conditions. As with most shared media networks, performance generally improves with larger frame lengths and degrades with the geographical extent of the network. For example, Cronin, et al., [3] found that for a 210 meter star-wired network (the largest possible topology for 100Base-T), the maximum efficiency of 802.12 ranged from about 46% with minimum sized (64 byte) frames to about 95% with maximum sized (1518 byte) frames, whereas the maximum efficiency of 100Base-T ranged from about 65% to 85% under the same circumstances. Conversely, assuming one wants to build a shared 2.2 km tree-wired network (representing a very large 10Base-T topology), the maximum efficiency of 802.12 ranged from only about 19% with minimum sized frames to 85% with maximum sized frames, whereas 100Base-T is not even applicable without bridging. Of course these figures are theoretical limits whereas the mean packet length encountered on a network is of more practical importance.

Once bridges are introduced, the efficiencies of both networks are governed by the diameter of each segment. Moreover, the relative efficiencies also depend heavily on the number of active transmitters. If one node is attempting a bulk file transfer over an otherwise-quiet network, then CSMA/CD allows the sender to consume all of the available network bandwidth whereas the round-robin scheduling in Demand Priority results in some overhead, especially for large networks. Conversely, if the network is busy because of many active nodes, then CSMA/CD loses efficiency due to collisions and backoff delays whereas Demand Priority saves overhead because the cost of a round-robin cycle can be amortized over more nodes.

- Network delays. Networks rarely operate at very high load because of the inevitable queueing delays. Thus, network delays at light to moderate traffic levels are generally more important than maximum efficiency. Theoretically, 100Base-T has lower latency on a quiet network, for the same reason that a yield sign wastes less time than a traffic light late at night in a quiet residential area. Of course, the resulting delays for both networks are small compared to the corresponding delays under heavy load, so the difference really doesn't matter. However, as the traffic levels on both networks approach their respective maxima, an 802.12 network should have a significant delay advantage over 100Base-T because each node gets regular opportunities to transmit a frame instead of letting all the frames wait until the node captures the network and can transmit a large burst. Thus, although 100Base-T has a higher maximum

efficiency with short frames, the associated delays under those conditions are likely to be unacceptably large so few users will see the improvement. Using BLAM instead of the standard CSMA/CD MAC will result in a significant improvement in the relative performance of 100Base-T in such comparisons, since it allows CSMA/CD networks to operate at higher loads without suffering from unfairness and poor delay characteristics. On the other hand, 802.12 has started to work on defining its own burst mode.

- Cost. Clearly, the least expensive 100 Mb/s network on a per-node basis is a two node 100Base-T network with one link segment and no hub. Such a topology is only possible with 802.12 if at least one node implements the functionality provided by a repeater. As soon as we add a third node, however, a repeater becomes a necessity for both networks and there seems to be no reason to expect a significant difference in cost between 100Base-T and 802.12 for moderately sized networks. If the network diameter exceeds the topology guidelines for 100Base-T, then 802.12 should be less expensive, assuming repeaters continue to be less expensive than bridges. If the network traffic exceeds the capacity of a shared 100 Mb/s system, then both approaches will require bridging, although 100Base-T may need more bridges because of its poorer heavy load performance.

- Priorities. 802.12 offers two priority classes, whereas 100Base-T has only one. The use of a two priority scheme is often either criticised or lauded. The critics claim that a high priority service has no benefit because there are no commonly available application programming interfaces (APIs) that enable programmers to use the feature. While this is true at the moment, there are products such as video servers that are designed to use the high priority feature. In addition the latest network API from Microsoft, Winsock 2.0 [16], will offer the ability to exploit various quality of service features that a particular interface might support and Microsoft's NDIS 4.0 will support LAN priorities. Unlike some other networks with priorities (FDDI and ATM), 802.12 does not currently have a bandwidth allocation protocol to control the total amount of high priority traffic. Despite the lack of an allocation protocol the priority mechanism was included because some people believe that priorities will become increasingly important as APIs are developed. Ongoing work within IEEE 802.1, specifically 802.1p, should provide mechanisms for extending priorities through bridged/switched LANs.

- Ease of Configuration. 100Base-T uses autonegotiation to exchange configuration information across a link segment, from which the attached devices can select their highest common operating mode. Similarly, 802.12 uses the presence or absence of link training tones to choose between 10Base-T or 802.12 operating modes. Having selected 802.12 mode, a node can then exchange configuration information with the repeater. In both standards the negotiation process can be extended to add new capabilities when the need arises. Sadly, there are currently no plans to make these configuration methods interoperate, so a link connecting two 100 Mb/s devices, each of which supports full duplex transmission of 802.3 frames over a UTP cable, would most likely configure itself to 10Base-T operation if one of the devices is 802.12 and the other is 100Base-T.

- Bounded Access Delays. Because 802.12 uses a deterministic round-robin access rule, one can calculate an upper bound on the "access delay" (i.e., the time that a frame spends at the front of its local transmit queue) for a given node in an N-node network as the sum of the worst-case propagation delays and frame transmission times for the other N-1 nodes, assuming there is no high priority traffic. Of course, since CSMA/CD uses random rescheduling after collisions, 100Base-T can offer only some statistical guarantees that a certain fraction of the frames will experience an access delay below some given threshold. These bounded access delays may be quite important for time sensitive applications like voice or videoconferencing, since the source generates data at regular intervals, and for real-time playback of the data we require timely delivery of each frame. Thus, the sender cannot afford to let its transmit queue

grow large, and the access time will be a good indication of the total delay from the generation of a frame to its delivery. The arrival patterns in normal data traffic are far less predictable, however, so that multiple frames may be waiting in the transmit queue at the given node. Since the bound in 802.12 applies only to the time that a given frame spends at the front of the queue, and not its total delay time, the bound has no real significance to a normal data application.

- Best-Effort Delivery. Under CSMA/CD, frames are occasionally dropped with an "excessiveCollisionError" after experiencing 16 consecutive collisions. Since 802.12 has an orderly access protocol with no collisions, 802.12 has no equivalent to this error. However, one must be careful not to jump to the conclusion that 802.12 offers "guaranteed delivery" of frames, while 100Base-T does not, since there are still many ways (bit errors, hardware failures, receiver not ready, buffer overflows, etc) in which a data frame can get lost between the source and destination without an error notification being returned to the source. Thus, if your application really needs to get its data delivered within some bounded time (e.g., safety critical real-time data for controlling a nuclear power plant or the flight of a jet fighter) then you cannot rely on the bounded access delay feature as a delivery guarantee and instead should view it as a performance optimization for handling the normal case [13]. Most other applications should be perfectly happy with the best-effort delivery service provided by 802.12, since almost all their data will be delivered within some given time bound -- the only real question, therefore, is whether the slightly worse-effort service provided by 100Base-T would be equally acceptable.

- Fairness. The goal of any MAC protocol is to provide all nodes with efficient and fair access to the shared medium. However, fairness is a tricky concept, and there isn't universal agreement on what sort of idealised scheduling rule would be the fairest one to use in a network. 802.12 uses round-robin scheduling on a per-node basis. The rationale for this choice is that if there is more demand than the network can sustain, then round-robin throttles every node back to a common value, i.e., its "fair share" of the available bandwidth. In particular, this keeps a misbehaving node from stealing all the network bandwidth for itself. This is obviously much more fair than CSMA/CD, which basically degenerates to last-come first-served scheduling under heavy load. However, BLAM's uniform randomized scheduling is also quite fair, and it even has some advantages over round-robin since each node can expect to get roughly equal bandwidth under heavy load, independent of its choice of packet sizes.

If different nodes generate frames of different lengths, then 802.12 is unfair to the nodes that generate short frames. In this case, the fairness of 802.12 would be improved if round-robin were replaced by a more general algorithm, such as weighted fair queueing [5], or the burst mode in BLAM and now under development in 802.12. However, all of these fairness guarantees get defeated when we split the network with a bridge. In this case, the bridge will become a bottleneck because of its high traffic volume in comparison to its "fair share" of the bandwidth. Moreover, bridging also increases the upper bound on the network access delay from O(N) to $O(N^2)$ because a frame sent to the other side of the bridge might have to wait while each of the N/2 nodes on the other side of the bridge transmits a maximum size frame.

New developments in IEEE 802.3 and IEEE 802.12

Link Failure Recovery:

802.12: In July 1995, work began on defining redundant "up" links. Only one up link is active at any time, with the second link acting as a hot standby.

802.3: Alternate routes require the use of bridges/switches which provide redundant links by the spanning tree algorithm defined in IEEE 802.1d[6].

New Media:

802.12: Work on a new signalling scheme for two pair Category 5 UTP wiring began in July 1995.

802.3: 100BaseTX already supports full duplex operation over two pair Category 5 UTP. However, work begain in March 1995 to define 100BaseT2, which is suitable for two pairs of Category 3 or better UTP. 100BaseT2 uses a sophisticated signalling scheme that splits the data into 50 Mb/s streams for transmission over each cable pair. However, even though both pairs are being used in parallel, full duplex operation is still supported via echo cancellation.

Full Duplex Operation:

802.12: Work on defining full duplex operation on those links that support it began in July 1995. Full duplex is inherently point-to-point, and cannot be used with repeaters. Thus, its importance is in connection with switches.

802.3: Work on full duplex began in March 1995. The project is also developing a flow control scheme that uses in-band transmission of legal 802.3 frames to deliver control messages.

Higher Speed Operation:

802.12: In July 1995, work on technologies for higher data rates began, including 424 Mb/s over UTP and 848 Mb/s over fibre.

802.3: In March 1996, work began on specifying how Ethernet will operate at 1 Gigabit/sec. operation.

MAC Layer Enhancements:

802.12: Along with higher speed operation, Demand Priority is being enhanced to allow a node to send a burst of several frames per grant, so as to improved its efficiency at higher speeds.

802.3: In March 1995, work began on adding the Binary Logarithmic Arbitration Method (BLAM) as an optional enhanced MAC layer protocol. BLAM is designed to work on heterogeneous networks, where some nodes use BLAM and the others use the existing CMSA/CD MAC, while providing a substantial improvement in fairness and delay on heavily loaded networks [11]. BLAM is a symmetric algorith, thereby eliminating an unfairness problem in heavily loaded CSMA/CD networks known as ``capture" [14]. In addition, BLAM allows multipacket bursts to improve its efficiency with short packets. Because of these changes, the delays in a heavily loaded network are much smaller and more deterministic with BLAM instead of the existing CSMA/CD MAC.

5 Packet Switching

A packet switch, or just "switch", is a multi-port bridge which simply forwards packets from a LAN on one port to a LAN on another port. The forwarding decision is based on the destination MAC address at the head of the packet (MAC frame). A switch will ignore a packet that is destined for a node located on the same port as the source node. Switches make their forwarding decisions based on Link layer (layer 2 of the OSI model) information; this is in contrast to routers that forward packets based on Network layer (layer 3) information. In addition, switches do not modify packets as they pass through, whereas routers must change the packet to include the MAC address of the next-hop and may also increment a hop-count field.

Each port in a switch is connected to a different LAN and all of which operate independently and simultaneously. Only those packets that need to pass from one LAN to another are forwarded by the switch, and so a multi-port switch might increase overall bandwidth by many times that of a single shared LAN. Most recent switches can transfer packets at the maximum theoretical rates, at least for 10 Mb/s LANs, and so will not be a performance bottleneck. Later in this section we contrast the performance of some shared and switched LANs.

The widespread use of high performance networked computers, notably personal computers, has driven up the bandwidth requirements for many sites. Switches are a simple way to increase the aggregate bandwidth without needing to recable the site or purchase new network adapter cards for each node. A large 10Base-T LAN, for example, can be segmented into many independent 10Base-T LANs by simply replacing the root repeater with a switch - full connectivity is maintained and the aggregate bandwidth is dramatically increased. However, switches are not a panacea, and careful attention must be paid to traffic flows, as shown later.

Switches are often much simpler, cheaper and faster than routers: they are simpler because switches operate at layer 2, do not modify packets, and do not need to run complex routing protocols such as RIP or OSPF. Switches may be both cheaper and faster than routers because the switching function is often implemented entirely in VLSI rather than being performed by software running on an expensive high-performance processor.

Architectures for switches have been discussed for decades, and while there are many variations in architectures most switches fall into one or a combination of three basic types: input buffered, output buffered and shared memory. A comparison of these architectures is beyond the scope of this paper, and while the literature is inundated with papers on the architectures of ATM switches, e.g. [15], the LAN packet switch vendors have been reluctant to disclose details of their architectures.

5.1 Issues in switching

Although bridges have been around for a long time there are several interesting issues that are currently intensely debated within the industry. We provide a brief discussion of some of these issues. Also see the box on Virtual LANs (VLANs).

Cut-through versus store-and-forward

A switch may choose either to forward a packet after it has been fully received (called store-and-forward) or it may start to forward the packet as soon as it has decided on the output port (called cut-through). Cut-through is only possible for a packet going to a port that operates at the same data rate, or a slower

data rate as the source port. One advantage of cut-through is that it reduces the latency through the switch if the output port has no other packets to transmit and if the medium is idle (no-one else is currently transmitting on that LAN). In practice the performance difference between cut-through and store-and-forward may be minimal due to the way that most modern reliable Transport layer protocols tend to send bursts of packets. This is because the switch receives the second packet at the same time that it forwards the first and so the full packet delay is incurred only once per burst.

A disadvantage of cut-through is that the switch may forward a packet that has been corrupted because the error check is not received until the end of the packet. Genuinely corrupted packets are rare and will not usually cause a problem. Of perhaps greater concern is a switch that propagates "runt" packets which are smaller than the minimum sized packets allowed for the LAN. This is not uncommon in CSMA/CD LANs where a node may start to send a packet and then stop before it has sent a minimum sized packet because it observes a collision with another node. If the switch has already started to forward the packet then a runt packet will be propagated onto the destination LAN, wasting bandwidth. Once again, though, this is unlikely to be a problem in most installations. Many cut-through switches only perform "safe" cut-through whereby they wait until they have received at least the minimum number of bytes in a complete packet before they forward the packet.

Flow control

All switches contain some packet buffers to cope with contention at a port: i.e. if two packets arrive at the same time and are destined for the same output port then one of them must be stored temporarily. However, buffers can only deal with contention that is very short-lived. If contention persists (the switch is congested) then the switch will eventually exhaust its available buffers. The switch must then either discard packets or else it must somehow prevent the sources from sending additional packets. Flow control is a mechanism whereby a receiver can tell a source to wait for some period of time before sending any further packets.

A sub-group within 802.3 is exploring flow control mechanisms for 802 LANs. Their current proposal is to use special MAC frames that are identified as flow control frames, and which contain information that specifies how long the source at the other end of the link should remain silent. A problem with this approach is that the flow control signal can only start or stop an entire link, rather than stopping the true source of the congestion although the 802.3 approach could, in theory, be extended to identify the true source. A consequence of link based flow control is that a congested path might interfere with an uncongested path. Figure 5 shows a switched LAN in which node A sends to B and nodes X and Y send to Z. Congestion occurs in switch 2 because X and Y are continually contending for Z. Consequently switch 2 sends flow control frames to Y and to switch 1, which has the effect of disrupting the packets from A to B even though that path is not congested.

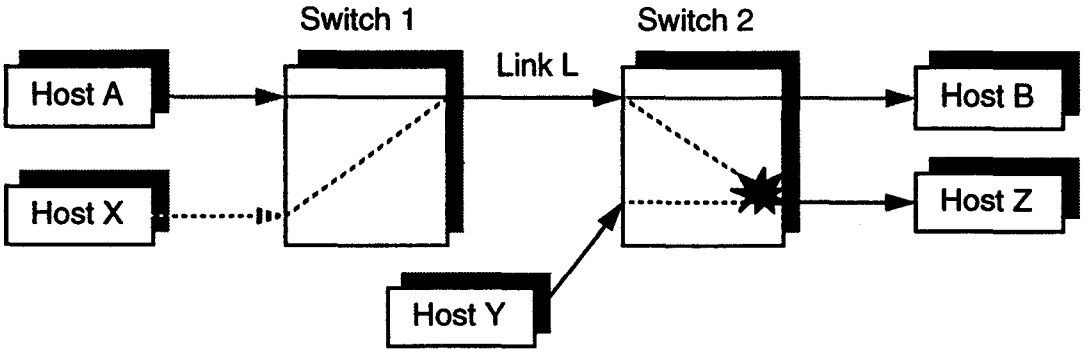

FIGURE 5. By using flow control on Link L, Switch 2 can interfere with an uncongested flow from A to B

The arguments for and against flow control are very interesting: switch vendors find flow control attractive because it might allow switches of a given performance level to use less memory and hence be cheaper to construct. Many protocol experts, however, believe that flow control is really an end-to-end issue which is handled very effectively by Transport layer protocols such as TCP, and the best thing the switches could do is simply notify the sources about congestion. Indeed, link layer flow control may even impede the Transport protocols which rely on packet loss as an indicator that congestion has occurred. Flow control and congestion control are deep problems and the reader is referred to [9] and [17] for many useful references.

Virtual LANs

A VLAN is a subset of the LAN devices (end-nodes, repeaters, switches) within a single bridged (switched) LAN domain. The choice of end-nodes within any subset is defined according to some arbitrary rules specified by the LAN administrator. Thus, for example, all nodes within the marketing department might form a VLAN. Unfortunately within the LAN switch industry the term "VLAN" has many meanings, many of which are incompatible with the description given here.

One important attribute of VLANs is that any member of the VLAN can move to any other location within the physical LAN and retain the same connectivity to peers as before. In contrast, a node on one routed subnet cannot move to another port on the router without requiring some manual re-configuration either within the node or the router.

Another benefit of VLAN technology is that it restricts the coverage of multicast/broadcast traffic to a subset (the VLAN members) of a large switched LAN. This saves bandwidth and means that end nodes only have to process broadcast traffic intended for them.

Participants within IEEE 802.1 are in the process of defining how VLANs are implemented and how VLAN information is distributed within a switched LAN, and standards may appear in 1997 or 1998.

Multicast/Broadcast management

The availability of low-cost, high performance switches has encouraged many network managers to build large bridged (switched) LANs, and move away from the model that is widely used at the moment in

which a router is used to separate LANs into independent subnets. A potential problem with a large switched network is that multicast/broadcast packets are propagated throughout the network. This is not a problem in routed networks because routers do not, in general, forward broadcasts. While the volume of multicast/broadcast traffic on any one LAN is generally small, the aggregate over a large switched LAN consisting of many separate LANs may be quite substantial.

Broadcast packets are of concern because every end node must receive the packet and decide whether it should take any action, potentially wasting a great deal of computing resources. Multicasts are less of an issue at the moment, due to the scarcity of applications that use multicasts, and because most network interface cards can perform some simple filtering of multicast packets. However multicast traffic is likely to increase as applications such as videoconferencing or shared whiteboards become more popular [10]. The problem is likely to be exacerbated in switched networks which have both 10 Mb/s and 100 Mb/s LANs: if nodes on a 100 Mb/s LAN generate many Mb/s of multicast traffic then this could severely disrupt traffic on the 10 Mb/s LANs.

Ideally a switch should only forward multicast/broadcast packets to the ports that have nodes that are interested in receiving such packets. This might be achieved by some form of multicast registration, so that switches can identify the set of ports that participate in a particular multicast group. The IEEE 802.1 group has been working on such a protocol for some time, and the current work exploits much of the knowledge that has been gained by the people who have developed multicasting for the Internet Protocol [4].

5.2 Performance

In this section we provide some measured performance data to indicate how shared LANs perform against switched LANs; it is *not* a performance comparison of 100Base-T and 802.12. These are very simple measurements, and should not be considered a definitive comparison; they are provided only as an *indication* of how the technologies might perform. Furthermore, these measurements do not imply limitations of the LAN technologies, because many other parts of the system, such as disk and operating system overheads will contribute to delay.

Four LAN configurations were measured: a shared 10 Mb/s Ethernet LAN, a shared 100 Mb/s 802.12 LAN, a 10/10 switch, and a 10/100 (100Base-T) switch. Each LAN comprised four clients communicating with a server. For the 10/10 switch configuration all of the computers communicate over 10 Mb/s links. For the 10/100 configuration the server was connected via a 100 Mb/s link but the clients used 10 Mb/s links. The clients were HP Vectra N2 4/66i (Intel 486 processors) PCs with 8MB RAM and the server was an HP NetServer 5/66 LF with 16MB RAM. The operating system was Novell Netware 4.1 with burst mode. The clients used HP 10/100 ISA cards with selectable 10Base-T and 802.12 functionality. The server used either an EISA bus master 10/100 card (for 802.12) or a 3Com PCI 10/100 card for 100Base-T. The 10 Mb/s repeater was an HP J2610A 8 port repeater. The 100 Mb/s repeater was an HP J2410A 15 port 802.12 repeater. The switch used in both cases was the 3Com Linkswitch 1000.

Four tests were run for each configuration. In every test each client performs random accesses (reads or writes) to a file located on the server machine. Each access is either small (64 bytes) or large (32kbytes). A read request packet is always small (112 bytes) and results in a single packet being returned for small acesses or a burst of packets being returned in the case of large accesses. Each test is run for a fixed length of time (about one minute) and the mean data rate is recorded.

The table shows the mean file data rate per client and, in parentheses, the aggregate server data rate.

	Client (Server) bandwidth (Mb/s)			
	10Base-T LAN	**100 Mb/s LAN**	**10/10 Switch**	**10/100 Switch**
Large Reads	2.3 (9.1)	14 (54)	1.9 (7.5)	7.1 (28)
Large Reads/Writes	2.0 (8.0)	10 (40)	2.0 (8.0)	6.7 (27)
Small Reads	0.5 (2.0)	0.6 (2.3)	0.47 (1.9)	0.48 (1.9)
Small Reads/Writes	0.41 (1.6)	0.5 (2.0)	0.4 (1.6)	0.44 (1.8)

There are some important qualitative observations to be made. The first is that a switched LAN is not necessarily faster than a shared LAN. The performance for the 10/10 switched configuration is marginally worse than for the simple 10Base-T (shared) LAN. This is to be expected: the switch introduces delay in the network and, in any event, the server is limited by its 10 Mb/s connection. However, if two servers had been used, with a separate link for each server, then the switched 10/10 configuration would almost certainly have been faster than the shared 10 Mb/s LAN.

The second observation is that 10/100 switching can yield useful performance improvements over shared 10 Mb/s LANs even with just a single server, provided that the server is connected to the switch via the fast link.

The third observation is that a shared 100 Mb/s LAN can perform better than the other three, by a substantial margin, but only for large accesses: when performing small accesses the performance is not limited by the LAN but other components in the system. This improvement for large accesses is also intuitive: all communications take place at 100 Mb/s and there is no delay introduced by a switch, and provided that the aggregate bandwidth does not exceed 100 Mb/s then we would expect it to be faster than the switched 10/100 LAN.

Finally, we see that the use of small reads and writes renders the choice of network almost immaterial as the systems are limited by other factors and not by the network.

The general conclusion is that the choice of switched versus shared or 10 Mb/s versus 100 Mb/s is critically dependent on the traffic patterns between the various computers.

5.3 Packets versus Cells

ATM technology, seen by many as the future of networking, is an inherently switched technology which transfers data in small 53 byte cells rather than the large, variable-sized packets used in 802 LANs. Some of the important differences between these technologies are discussed below.

ATM is a connection-oriented service, unlike the connectionless datagram service provided by packet networks. Consequently its true potential is realised only when ATM is installed from end to end. This requires a substantial investment for most users and while ATM is achieving notable success as a backbone technology its success at the desktop has been less evident.

One immediate distinction to users is the price: packet switches are cheaper than ATM cell switches. 10Base-T switches are especially low cost because the millions of users of 10Base-T do not need to replace the interface cards in their computers or change any of their 10Base-T wiring. Of course a

10Base-T switch does not provide comparable bandwidth to that of a 155 Mb/s ATM switch. In addition, the price of ATM switches does not reflect their cost and prices are likely to fall as competition increases and vendors develop improved technology. Moreover, the recent arrival of 25 Mb/s ATM products should further reduce the difference in price.

An advantage of ATM is that it can potentially provide a sensible solution to the flow control problem. This is possible because the virtual channel identifier in each cell allows a switch to identify each logical flow (data stream) within the sum of flows on a physical link. Consequently, flow control can be imposed on a single logical flow without interfering with other flows on the same link. This is only true provided that the host software is configured such that each logical connection has its own virtual channel.

Another potential advantage of ATM is that it offers a variety of qualities of service, ranging from guaranteed bandwidth through to best-effort. Packet switches, in contrast, provide the rather primitive service defined for 802 LANs which is restricted to best-effort service with simple prioritization of traffic. As yet ATM's potential benefits have not been realised, at least not for LANs. Most ATM switches that are sold into the LAN environment provide LAN Emulation, and offer little more than a packet switch.

In terms of market share, market analysts at Dell'Oro, California, report 25,100 ATM ports shipped in the first half of 1995 relative to 569,000 switched 10 Mb/s ports and 244,000 100 Mb/s shared medium LAN ports (100Base-T, 802.12 and FDDI). These figures are dwarfed by the total 56M LAN ports estimated to have been sold in the whole of 1995 - predominantly shared 10Base-T.

6 Conclusions

IEEE 802.12 and 100Base-T, as well as packet switching, are standardised technologies that can provide substantial performance improvements over the vast majority of installed LANs. Each technology has its own strengths and weaknesses and these must be carefully considered before choosing one against another.

The performance comparison in section 5 shows that 100 Mb/s shared medium LANs can offer exceptional performance at a reasonable cost. However, for some users a better approach will be to opt for a switch, with clients on 10 Mb/s ports and the servers on a 100 Mb/s port. This can be cheaper yet still offer a substantial improvement in performance.

It is clear that 100 Mb/s technologies will continue to evolve, with full duplex links, burst mode support, and even the development of yet higher speeds. However, there are always competing technologies, and ATM may yet achieve success in the LAN marketplace. Hopefully this competition will result in more choice, lower costs and better performance for users.

Acknowledgements

We would like to thank the many reviewers who have contributed so much to this paper. Thanks also to Peter Kim for the 802.12 high priority performance measurements, and to Paul Congdon for performing the measurements reported in the section on switching.

References:

[1] Coles, A.N., et al.: "Physical Signalling in 100VG-AnyLAN," HP Journal, pp.18-26, August 1995.

[2] Coles, A.N., Cunningham, D.G., and Methley, S.G.: "100 Mb/s Data Transmission on UTP and STP Cabling for Demand Priority Networks", IEEE Jnl. on Sel. Areas in Comms., Vol. 13 No 9, pp.1684-1691, December 1995.

[3] Cronin, W.J., Hutchison, J.D., Ramakrishnan, K.K. and Yang, H.: "A Comparison of High Speed LANs," 19th Conference on Local Computer Networks, pp.40-49, October 1994.

[4] Deering, S.E.: "Host extensions for IP multicasting," RFC1112, 1989.

[5] Demers, A., Srinivasan, K., and Shenker, S.: "Analysis and Simulation of a Fair Queueing Algorithm," Internetworking: Research and Experience, Vol. 1, pp.3-26, 1990.

[6] IEEE Standard 802.1d-1993 (ISO/IEC 10038), "MAC Bridging," IEEE, 1993.

[7] IEEE Standard 802.3u-1995, "Media Access Control (MAC) Parameters, Physical Layer, Medium Attachment Units, and Repeater for 100 Mb/s Operation, Type 100BASE-T (Clauses 21-30)," IEEE, 1995.

[8] IEEE Standard 802.12-1995, "Demand Priority Access Method, Physical Layer and Repeater Specification for 100 Mb/s Operation," IEEE, 1995.

[9] Lefelhocz, C., Lyles, B., Shenker, S., and Zhang, L.: "Congestion Control for Best-Effort Service: Why We Need a New Paradigm," IEEE Network Vol.10 No.1, pp.10-19, Jan/Feb 1996.

[10] Macedonia, M.R., and Brutzman, D.P.: "MBone provides audio and video across the Internet," IEEE Computer, Vol.27 No. 4, pp.30-36, April 1994.

[11] Molle, M.L.: "A New Binary Logarithmic Arbitration Method for Ethernet," Technical Report CSRI-298, University of Toronto, April 1994.

[12] Partridge, C., Hughes, J. and Stone, J.: "Performance of Checksums and CRCs over Real Data," ACM SIGCOMM 1995 in ACM Computer Communication Review, Vol.25, No.4, pp.68-76, October 1995.

[13] Saltzer, J.H., Reed, D.P., and Clark, D.D.: "End-to-End Arguments in System Design," ACM Transactions on Computer Systems, vol.2, pp.277-88, Nov. 1984.

[14] Shenker, S.: "Some Conjectures on the Behaviour of Acknowledgement-Based Transmission Control of Random Access Communication Channels," ACM Sigmetrics '87 Conference, pp.245-255, May 1987.

[15] Stephens, W.E., DePrycker, M., Tobagi, F.A., and Yamaguchi, T. (Editors): "Large-scale ATM Switching Systems for B-ISDN," IEEE Jnl. Sel. Areas in Comms., Vol. 9, No. 8, October 1991.

[16] Winsock 2.0 Specification, available from: ftp://ftp.intel.com/pub/winsock2/spec

[17] Yang, C-Q, and Reddy, A.V.S.: "A Taxonomy for Congestion Control Algorithms in Packet Switching Networks," IEEE Network Vol.9 No.4, pp.34-45, 1995.