M.I.T. Laboratory for Computer Science

Request for Comments No. 254
March 23, 1984

New Architectures for Data Network Communications

by David D. Clark

Abstract: This report proposes that a research effort be undertaken to develop new approaches to data communications over shared networks. The central thesis of this new research is that some forms of circuit switching are more suitable than packet switching for some communications situations. Suitable circuit switching and packet switching mechanisms can be united to provide a more general form of network facility than now exists. This facility will provide higher bandwidth, better resource allocation, and better application level utility than the traditional technology of today.

1. Introducion

As is well known, there have been great advances in the computer and communications technology of the last decade, and these advances show every sign of continuing. New communications-technology, such as fiber optic lines, directly suggest new possibilities and problems in data networking. Reductions in processing and memory cost affect communications in two ways. First, these reductions change the economics of the network switching nodes. It is now reasonable to include large amounts of buffering, for example. Second, this reduction changes the nature of the computer equipment to be interconnected together. The last few years have seen the terminal being displaced by the personal computer, which is changing the whole nature of interactive network traffic. In addition, specialized communications requirements for digital transmission of speech, images, and video change the requirements that the communications system must need.

2. Packet Switching

The most common technique today for the management of shared data communications is packet switching. Packet switching has proved extremely successful; it provides a mechanism for highly dynamic sharing of bandwidth on a demand basis among a number of simultaneous users. However, the limitations of packet switching are beginning to be felt.

First, the overhead of dealing with each packet separately provides a limit to the speeds that switching nodes and hosts can achieve for data communications. This is apparent in hosts, in the great difficulty generally found in utilizing even a fraction of the bandwidth of today's high speed local area networks, and it is seen in packet switches, which can easily represent substantial bottlenecks. As an example, DARPA has funded the development and deployment of a new multi-processor packet switching gateway for the DARPA Internet; even that may not have the throughput to support the 10 mbps local nets now in use.

Second, packet switching is not a good tool for close control of resource allocation. The regulation of excessive requests for communication bandwidth (congestion control) has been a hard problem in packet switching, which has not yet been well solved after ten years of effort. The typical flow-control mechanism, windows, leads either to artificial bandwidth limitations or very large buffer requirements. This result can be attacked in certain ways, for example by permitting the non-sequenced delivery of packet, but is a basic artifact of packet switching.

If packet switching has proved to be somewhat limited, so has the traditional paradigm of transport level communications semantics, the end-to-end virtual circuit. The whole focus of this abstraction is that the sender and receiver shall be part of a simultaneous transfer operation. But, looking at the DARPA Internet as an example, most of the traffic is computer mail, which is a very poor fit onto the virtual circuit. Most importantly, mail is not delivered in real-time end to end, but is staged through a number of relay points. Since the addressing structure of the virtual circuit is not sufficient to support this relaying, the mail application had to invent a higher level address structure. This two-level global address structure has not worked well in practice. The mail application similarly was forced to develop mechanisms for error detection and recovery, detection of duplicate messages, timeout of undeliverable messages, and so on. In short, mail became a high level transport protocol, and other applications, such as FTP, or remote job submission, have begun to use mail in this way.

The advent of personal computers has strengthened the argument in support of non-real time, staged delivery. Personal computers are often powered off, are often transported away for their network connection, and often run dedicated applications that do not support network activity. (MS-DOS, which is very limited in this respect, is liable to be with us for some time.) In the military environment, the need for operation in the tactical environment, with communications not always available, strongly suggests the need for a mode of data delivery based on staged transport.

The above arguments are not intended to suggest that either the virtual circuit or packet switching should be abandoned. Rather, it suggests that applications are being developed which stretch the limits of these approaches. What is needed is a new, broader view of the communications function, within which a wider variety of applications can be supported.

3. An Alternative For Communications

Packet switching is a technique for allocating bandwidth to a particular user on a very dynamic basis. It is this dynamics that causes problems with control of resource allocation. There are a variety of other allocation techniques with a different mix of control against dynamics. Frequency devision is very stable, but so static that some supplementary technique such as frequency agile modems is needed to add dynamics. The traditional circuit switching of telephones has also proved somewhat too static in certain cases, so TASI (and packet switching) are used to add dynamics. Fast setup circuit switching is now being proposed as a technique with good balance.

There are bandwidth reservation techniques with a more dynamic nature than those above. The stream reservation mechanism of the DARPA satnet and wideband satnet, and the slot structure of some LAN's (in particular, the Cambridge Ring and most of the new high speed LAN's) are good examples.

A number of these switching techniques can be used to provide a service which this report will call "reservation switching." (We avoid the use of more familiar terms to prevent unnecessary associations.) The general characteristics of reservation switching is a reservation setup phase, assumed to occur in time proportional to delay across the net (e.g. a few milliseconds to a few seconds), a data transfer phase, and a disconnect phase. Reservations can be of two forms, bounded and unbounded. A bounded reservation is for a known quantity of bits (a particular file, for example) while an unbounded reservation is for an indefinite number of bits, for example a speech conversation.

Some examples of reservation switching will help show the range of applications over which this technique can be applied. Current packet switching techniques across a local area network are hard pressed to achieve 10^6 bps, and often achieve no more than 10^5 bps. At 10^5 bps, transfer of a 10^6 bit image (e.g. the screen of a modern APA display) would take 10 seconds, and the backup of a 10 Mbyte disk, 800 seconds. All of these numbers are much too long, and even at a data rate of 10^6 bps would be marginal. A data rate of 10^7 bps would be much more consistant with the need of the application. Initial work on 10^8 bps LAN's (which are almost all reservation switching) suggest that 10^7 bps per flow could easily be achieved if the system is properly engineered.

Across a broadcast satellite network, the situation is exactly the same, scaled down by one or two orders of magnitude. Real time speech seems to require reservation switching to work well over a satellite network, and the present DARPA Internet effort includes a reservation switching facility(the ST protocol) to support speech. Other higher bandwidth applications, such as image transfer will require reservation switching, either implicit or explicit, to operate properly.

Perhaps the most important application of reservation switching will be in the development of new medium and long-haul networks constructed from high speed point-to-point fiber optic links. Over the time in which this research project is envisioned, it is reasonable to expect fiber optic links which provide 10^7 bps or even higher data rates. Present experience suggests that it will be difficult to build packet switches that can cope effectively with these data rates. Some form of reservation switching will almost certainly be necessary.

4. Specific Research Efforts

The preceding discussion has identified two fundamental ideas, reservation switching for resource allocation and staged rather than real time delivery of messages. In certain respects, these ideas can be explored independently. However, one of the important ideas that motivates this research effort is the belief that these two ideas, working together, can simplify some of the major difficulties associated with this effort.

The sections below outline a number of specific research projects which can be done as part of this overall effort.

4.1. Basic Media Research

The supposition of this effort is that long-haul bandwidths of 10^7 bps and short-haul bandwidths of 10^8 or 10^9 bps can be made available by fiber optic technology. Research in this area is, of course, currently being carried out to a substantial extent. This research effort need concern itself with basic technology only to the extent that certain sorts of devices, for example couplers to be used as part of host interfaces, may seem particularly important as part of the overall architecture development.

4.2. Long-Haul Network Architecture

Topologically, the long-haul and medium-haul networks of tomorrow will resemble the networks of today, an interconnection of point-to-point links and switching nodes. The important design decisions are the way in which the bandwidth is to be shared among the various users, and the way in which the switching elements are to be designed. Obviously, these two decisions interact very strongly.

If the switching strategy is packet switching, then the future switching node will very much resemble a speeded up version of current-day packet switches, exploiting parallelism and high speed logic where appropriate.

This proposal is based on the assumption that some other form of bandwidth sharing, most probably a more static time division strategy such as slotting, will lead to more effective bandwidth allocation and simpler switch design. Specifically, some form of fixed slot structure on the fiber would permit a de-multiplexing scheme in the switch, easily supported in a special interface unit, which splits the traffic on the fiber into a number of separate streams, each of which can be handled at a lower data rate. This approach will lead to a very high speed, highly parallel, very low cost switching unit.

One normally thinks of a switching node of this sort as having very little buffering. However, the cost of memory today would permit the addition of very large quantities of buffering to a machine like this, sufficient even to buffer very large messages, on the order of megabytes. Such buffering may potentially permit new approaches to route recovery after line outage, transient excessive bandwidth demand, and other temporary disruptions which can be expected within a network.

4.3. Local Area Network Architecture

Future local area networks can be expected to differ from future long-haul nets in approximately the same manner that the present networks of this type differ. The local nets will have substantially higher bandwidth available than long-haul nets, which will make design of switching elements and interfaces even more difficult, and local area networks can make use of specialized topologies, such as rings and stars, to simplify the design of the network.

It is a reasonable supposition that future high speed local area nets will use some form of reservation switching, because the current very high speed local nets being developed today already have adopted some form of this strategy. Almost all of the hundred megabit fiber optic rings now in development use some form of slotted time division multiplexing as a strategy for reducing the host interface. Even some earlier local area nets, such as the Cambridge Ring, use a slotted strategy for bandwidth allocation.

The principal challenge in the devlopment of very high speed local area nets is in the design of a cost-effective host interface. The host interface should ideally have the following characteristics. First, the design should permit an effective trade-off between speed and cost of the interface. Second, the interface should transfer as much of the processing associated with network management out of the host processor. Third, the interface should permit the interleaving of several transfers, so that a single longswitched data element will not lock out short messages of high priority.

The design of local area nets to support 10^8 or 10^9 bps media is an extremely challenging area which deserves attention now, to permit the architectural concepts to mature at the same time that the transmission medium itself evolves.

4.4. Host Interfaces

The previous section alluded to the necessity for the development of new forms of host interfaces. The chief difficulty with current local area networks, such as ethernet, is that the network is already as fast or faster that the I/O bandwidths of some computers. Increasing the bandwidth of communication networks can only make this problem worse. For this reason, new approaches will be required in the development of host interface strategies. While the increasing bandwidths cause difficulties in this area, current technology developments offer great opportunity. In particular, the

ability to develop very complex special purpose VLSI chips mean that great computing power can be included within the host interface unit, if it can be harnessed effectively.

However, packet switching does not seem to lend itself well to the effective utilization of computing power within the host interface. It is certainly possible to off-load the processing associated with the protocols of packet switching, but experience with moving these functions into the interface suggests that there is no great advantage in doing so. Packet processing, by its nature, is intrinsically serial, and thus requires for its execution nothing more than a general purpose processing chip. It makes little difference whether this is the main processor of the computer, or an auxiliary processor provided for the purpose.

Reservation switching has the great advantage that it permits the splitting of the incoming data stream, at a very early stage in its processing, into a number of regular data flows of known rate, which can then be processed in parallel by separate processing elements. These parallel elements can then perform many of the functions which require that each byte of data be manipulated, such as error detection and delivery of data to the main memory of the computer.

As in so many areas of current architectural development, parallelism seems the most effective strategy for increasing the raw speed with which the various operations can be done.

4.5. Development of Internetwork Architectures

The current DARPA internetwork effort has made it quite clear that any future networking development effort must, from the very beginning, pay close attention to the necessity for providing a communications path which is composed of a number of different sorts of network technology, connected together as part of a total network architecture. In the packet switching version of internetworking, the distinction between packet switch and internetwork gateway, is that the gateway is expected to cope with great variation in speed and functionality of the various nets to which it is connected. Similarly, in a reservation switching environment, the internetwork switch would be required to deal with networks that have very different strategies for reservation, and very different available bandwidth. Clearly, there are a number of different approaches which could be taken to the design of a reservation internetwork switch.

4.6. Transport Service Interface

In current packet switching, there are two traditional transport services provided by the network to its users, connectionless or datagram service, which permits the transport of isolated fragments of data, each mapping directly into a packet, and connection or virtual circuit service, in which there is a set-up phase, a data transfer phase of unbounded length, and a connection closing phase. For reservation switching, the transport service will presumably be somewhat different. Unbounded reservation service, in which the amount of data to be sent is not know in advance, somewhat resembles the virtual circuit, except that the dedicated bandwidth means that this service will be of primary interest to applications such as speech and video, which have a continuous predictable requirement for data rate. Bounded reservation, in which the amount of data is known in advance, is somewhat similar to datagrams, except that the data elements may be very large, perhaps megabytes, and will be used for the transfer of data files, images, and non-real time speech units. The transport service interface must deal directly with these very large data aggregates, because only by making these large units known to the transport service can the transport service make the proper reservations for them.

From an architectural point-of-view, such a transport service interface is straightforward to define. However, from an implementation point-of-view, most operating systems will not support this interface very effectively. The problem to be solved is that most operating systems do not provide a structure in which any guarantees can be made about the rate at which data can be generated or received. Of course, specialized devices, such as those for digital speech or imaging, can provide precisely this service, but that suggests that the network interface should be directly connected to these special devices, and not as a general purpose device to the system as a whole. The development of this sort of network interface will require rethinking of hardware architecture, as well as rethinking the relationship between the application program and the operating system, and the relationship between the network and the rest of the operating system facilities, such as the file system.

In fact, reservation switching lends itself to integration into an operating system more than packet switching ever did. The data abstraction provided by a file system, for example, is normally very stream oriented. That is, the data bytes of a file are thought of as occuring in sequence. To transfer a file using packet switching, it was necessary to take the sequence of bytes, break it into data units whose size was determined, not by the file system but by the limitations of the packet switching

architecture, surrounding each data unit with an arbitrary header, and then send these units. Reservation switching permits a much simpler operation. The data transfer reservation is made, and then the data bytes of the file are simply transferred from the storage medium into the network interface. The network must be sophisticated enough to cope with a file system which can not reliably deliver the data at precisely the reserved speed. However, the problem of padding missing data units and buffering data units when the interface falls behind at the receiving end can be solved entirely inside the network interface, which makes the network appear to be a very simple device to the rest of the host operating system.

4.7. Approaches for Staged Delivery

As was discussed above, many applications, most notably mail, have made it clear that the objective of the transport layer is not to deliver the mail item in real time from the sender to the receiver, but rather to move the data item in one or more staged transfers. Note that a mail item is a good example of the element which might be sent as a single bounded reservation. This suggests that a very important transport service, is the delivery, not in real time but staged, of bounded reservation data elements. The development of a non-real time delivery service will require rethinking a number of issues, most specifically addressing, end-to-end error detection and recovery, and flow control.

In today's packet switching, it is interesting to compare a mail item to a datagram. Both of them have unique ID's, addresses, a time to live (sometimes implicit rather than explicit for mail), and a level of reliability characterized by best effort, together with a notification of non-delivery which uses the same mechanism as the delivery mechanism itself, i.e. another datagram or mail item respectively.

Just as the datagram is the building block out of which more complicated transport services can be built, the mail system is now being used as a building block for a more general sort of transport service, in support of such things as file, transfer, and remote job entry. None of these require delivery in real time. In fact, the emphasis on real time delivery comes mostly from the use of networks to support remote login, a function which will become less important as the personal computer replaces the terminal as the user interface device.

What is needed is a service which can be "as real time as possible" but which degrades to staged

delivery, with a well-defined uppertime limit, as necessary. Such a service could be used for a real time data base query, where the maximum delivery time would be a small number of seconds, to mail, where the maximum delivery time would be a small number of days. In neither case, however, is a direct connection from the sender to the receiver required, of the sort traditionally associated with remote login.

A substantial amount of experimentation with staged delivery can be done using existing packet switched technology, but the interaction between staged delivery and reservation switching is very important. The data element which is staged is also precisely the data element for which a bounded reservation is made. Thus, the transport service need only deal with one form of data abstraction, in order to provide these two important services.

4.8. Application Development

Any service facility can only be understood by putting it to use. Thus, if reservation switching and non-real time delivery are to be understood, they must be evaluated by the development and use of appropriate applications. Development of applications must be planned from the very beginning as a part of the project.

The most obvious application is a new mail system. A new mail system has a number of advantages. First, it directly fits into the paradigm of staged reservation switching. Second, it can be enhanced into multi-media mail by the addition of graphics, image or speech. These facilities can be selectively added to stress the development of new host interface software and transport bandwidth. Third, a mail system will cause the network builders to use the facility they are building. This is critical for any successful system development.

A more sophisticated application than mail would be a multi-media teleconferencing system. The distinction we mean to imply between mail and teleconferencing is that teleconferencing system is a more organized and tightly controlled conversation between a number of communicants on one particular topic. Ideally, a teleconferencing system should permit a real time conversation, if all of the participants are simultaneously available. However, if either the participants or the network conductivity do not permit real time communication, then the teleconferencing system should smoothly shift in to a near-real time or fully staged mode of operation. An experiment to determine how this transition between real time and non-real time should be realized will be very helpful in

understanding what functions the transport service should provide, and what functions the application layer should provide.

A third possible application is database query. A database query and its associated reply are each examples of data items which can be sent using a bounded reservation. As mentioned above, database queries are normally thought of as occuring in real time; however, strictly speaking the time limit is usually a small number of seconds, rather than milliseconds. Thus, small degree of staging is acceptable in the implementation of database query. This semi-real time mode makes this application interesting.

The application which will provide the most stress on the ability of the network to deliver raw bandwidth is the storage, retrieval, and transmission of digital images. Digital imaging systems are liable, over the next ten years, to become a very important part of computing. A good human interface to an image system requires the ability to deliver images at very high speed. Images are thus an obvious candidate for transmission by reservation switching. The development of an image work station, and an image store, connected over a reservation switching internetwork, would be a demonstration of a substantially new technology, not available today by any economical means.

5. Conclusion

The development of a demonstration reservation switching inter-network is a project which would take at least five and perhaps ten years. The overall project should thus be thought of as a framework within which a number of these individual studies can be carried out, some in parallel, some sequentially. There are certain efforts, such as the development of the high speed networks, which depend on the availability of high bandwidth fiber optic technology. Some parts of the project, such as the host interface design and implementation, could be prototyped now, simulating the fiber optic links with very high speed direct connection.

In the number of recent research projects, a major problem has been to work far enough ahead of industrial pressures to permit the research concepts to mature sufficiently. I believe that this project, if undertaken now, may just have that lead.