CMU-ITC-86-056
AFIPS Press

INTERCONNECTION OF LOCAL AREA NETWORKS AT CARNEGIE-MELLON UNIVERSITY

John Leong
Carnegie-Mellon University
Pittsburgh, Pennsylvania

INTRODUCTION

During the past 5 years, the computing environment at Carnegie-Mellon University, CMU, has been undergoing major evolution from the central data center model to one of distributed processing. Until the late 70's, with few exceptions such as the Computer Science department, most departments relied on the computing resources provided by the Computing Center. Since then, with the decreasing cost of computing hardware, individual departments started acquiring their own machines. The pace of the computing power distribution accelerated with the introduction of high powered work stations and personal computers. Currently, there are over 150 DEC VAX 750/ 780/ 785 class of computers on campus together with approximately 200 SUN's, 150 RT-PC's, 300 microVAX'es, numerous DEC PDP-11's and miscellaneous other work stations from vendors such as HP, Xerox, and Symbolics. Additionally, there are roughly 1,000 IBM PC's and 500 Apple Macintoshes around the campus. Of the collection of machines above, only 6 DEC-20's, 3 VAX 780's, a 3083 and a number of PC and Macintosh clusters are controlled by the Computing Center.

While the distributed approach has significant advantages over the more traditional computing center model in terms of growth, it does have the potential risk of leading to inefficient use of resources and causes fragmentation in the campus computing environment. These risks can be substantially reduced if the machines can be made to co-operate with each other through the efficient and effective use of networking.

THE NEED FOR HIGH SPEED LOCAL AREA NETWORK (LAN)

In the past, most communication traffic is typically generated by the users explicitly. The most common application is host access from terminals, followed by file transfer and mail. These types of applications can normally be served by the standard serial line network especially if it operates at the relatively high speed of 9.6K. Even when the user requests the transfer of a moderately large file, the delay is generally tolerable since the transfer is explicitly requested and

the user is psychologically primed to wait. For those infrequent occasions when huge files have to be transferred, the old and generally reliable method of using magnetic tapes still remains acceptable in most cases.

Over the years, CMU has built up a significantly large serial line network based on 3 Micom data switches providing connections for roughly 2,000 lines into 1,000 ports of various computers.

Recently, a different distributed applications profile began to gather momentum. In the case of the large mainframe, each costly computing station is generally self sufficient in resources. This is usually not true in the world of work stations and the low cost PC's. For that type of installation, it is economically desirable for the stations to be able to share common resources across the network. In this profile, the remote access of the resource is typically system initiated and as such, is done behind the user's back. Any delay will show up as an irritating fluctuation in performance. Depending on the type of coupling between the workstation and the resource provider(s), the communication requirement can range from very servere to moderate. We have experience with applications in both categories : disc-less workstations paging across the network and stations accessing files remotely.

The central focus of CMU's distributed processing environment being developed and deployed today is called Andrew [1]. Through a campus-wide system of interconnected high speed local area networks, a logically central but physically distributed file system is made available to a large number of workstations. The stations will typically have just enough local disc storage to provide for efficient system operations such as paging and file caching. Space for the users' files is provided by the file servers across the network. This model offers a number of advantages besides cost. A user can access his or her files from any station in the network -- users are not restricted to a specific machine; the files will be backed up regularly by a centralized professional organization; data sharing and program control will be much easier.

Another interesting reason why the traditional serial line configuration will not gracefully satisfy the requirement of large scale distributed processing is the fact that they have a mostly point-to-point circuit switched type of configuration. With this approach, a server machine typically has a limited number of physical access ports. Users will have to contend for one of those ports in order to gain access. Due to the typically high overhead in circuit establishment, connections are likely to be held for a long duration with generally poor port utilization. The contention problem and circuit establishment overhead is further aggravated if a distributed transaction involves multiple servers. In the Andrew distributed file system, a station would have to go to an authentification server to get clearance before it can approach the servers for resource accesses. Furthermore, since the main design goal of the Andrew system is that it must be easily scalable to provide service for over 5,000 stations, a substantial amount of the computing is off loaded from the servers to the stations. When a station addresses an inappropriate server for resource, the server will not use inter-server communication to co-operate with other servers and work on behalf of the requestor. Instead, the requestor will be told who it should address the request to. In the point-to-point model, the attempt to connect to an alternative server may fail due to blocking. In fact, the probability of a transaction failure is proportional to the number of servers involved. With most local area networks, this low level type of port contention problem does not arise since they tend to operate on a datagram rather than a circuit switch model.

LOCAL AREA NETWORKS AT CMU

CMU has acquired a lot of experience with a number of different LAN technologies. We have been using Ethernet [2,3] for over five years, starting with the experimental 3 megabit Ethernet. Three years ago, various departments, at their own initiative and funding, started installing 10 megabit Ethernets. Most of these activities were not co-ordinated centrally. Currently, there are 16 separate Ethernets with over 30 segments on the campus. More than 600 high performance work stations and hosts of various types are attached directly onto those networks. Since the beginning of this year, we have been actively deploying IBM token rings [4,5]. This LAN technology has a lot of network maintenance properties built into its design and has, therefore, good potential for really large scale deployment. We have currently 4 rings in operation with roughly 80 attached nodes. The number of rings and nodes will both be increasing substantially over the coming year. Another general purpose LAN that we have in use is Pronet. Although it does not conform to any standard, it has provided us with interesting operational experience with the ring type of network prior to the appearance of the IBM token ring. However, we will not be aggressively expanding in this direction in the future. Due to the popularity of Apple's Macintosh, particularly among students, there is a strong demand for the support of the low cost AppleTalk [6] network. Currently, this demand has yet not been adequately addressed at CMU.

In general, from an operational and maintenance point of view, we would like to see only one LAN technology on campus; in practice, we will be required to support multiple types of LAN's. Therefore, in order to provide a coherent networking environment for the campus, one of the required tasks is the interconnection of dissimilar LAN's.

COMMUNICATIONS PROTOCOLS

Assuming we have the means to interconnect all the heterogenous LAN's together, meaningful communication between machines is still not assured since a variety of machines exist on campus and they use different protocols.

At CMU, the most popular protocol family in use is IP, Internet Protocol. It is the protocol supported by the Department of Defense and is available under UNIX 4.2 - a relatively machine independent operating system. There are over 400 stations on campus that use IP as the native protocol. These stations are attached to a variety of Ethernets, ProNets and IBM token rings. The campus IP internet is also connected to the ARPANET through a packet switch node, PSN or IMP, operated by the Computer Science department. The second most popular protocol on campus is DECNET. It has a population of 100 or more nodes connected together with Ethernets, ProNets and high speed point-to-point links. The CMU DECNET is, in turn, connected to DECNETs at a number of other academic institutions. Other protocols in use are PUP, XNS and AppleTalk. PUP is being phased out and the usage of XNS on campus is relatively limited. As mentioned in the previous section, we intend to provide AppleTalk support particularly for students.

An interesting question that can be asked is why should one want to interconnect all the

dissimilar machines or logical networks. Some of the reasons are : (a) people on different machines would like to exchange mail and access common electronic bulletin boards; (b) exchange data and program source files written in relatively machine independent languages such as C, Pascal and Fortran; (c) access some centrally administrated file and data base servers. Furthermore, it is not uncommon for a user to have accounts on more than one machine. It will be very convenient, particularly in a system with windowing capability, if the user can access multiple machines from the one he or she is currently logged on.

There are three approaches to solving the multiple protocols interconnection problem. They are : (a) provide every machine with the capability to handle all other protocols in use besides its native set; (b) provide protocol translation machines; and (c) select a standard protocol and ensure all machines can handle this protocol in addition to the native set.

The first approach is quickly discounted as impractical. Protocol translation can be achieved either by implementing a set of any-to-any protocol translators or switching through an intermediate protocol. The latter is, by far, preferable. Even then, protocol translation can be very difficult and slow, particularly when it is done at the lower level transport services. In the last approach, if one can find a widely available protocol, no translation will be necessary. The station can simply select the protocol set appropriate to its communication peer. We have focused our attention on approach (c) and use (b) as the fall back.

Our main selection criterion for the standard protocol is that it must have strong support, i.e. implementation available for most if not all machines. The protocol families we have short listed are : IBM's SNA, CCITT/ ISO and DARPA's IP protocols. Because of IBM, every manufacturer has tried to provide SNA interconnection capabilities for their machines - particularly the relatively new peer-to-peer LU6.2 protocol. However, since none of the machines on the campus currently support this rather complex protocol set, we have decided against its introduction. The international standard CCITT/ ISO protocol has the support, among others, of the PTT's which have monopolistic control of all communication services in some countries. However, this protocol set is currently incomplete - particularly at the application level. While a mail protocol standard has emerged recently in the form of X400, the file transfer, access and management (FTAM) and terminal access protocol is still pending. We have, therefore, also decided against this protocol set at this time. The DARPA's IP protocol is a very complete set of protocols and is required for all machines destined for the Department of Defense (DOD). As mentioned earlier, it is the native protocol set for the popular Berkeley UNIX 4.2 operating system and has wide spread usage in CMU. We have, therefore, selected this as our campus standard protocol. Implementation of this protocol is available to a wide range of machines which have other native protocols. Example of these are DEC/ TOPS-20, VAX-VMS, 3083/ 4341-VM, IBM-PC and Macintosh.

INTERCONNECTION OF HETEROGENOUS LAN'S

Given that we have a number of different local area networks on campus, it will be highly desirable to interconnect them together. Physically, LAN's in different buildings can be joined together using the large fibre optic plant at CMU. We will discuss the choice of using fibre optic as an inter-building "back-bone" medium in a later section. For Ethernet, we can use fibre repeaters from companies such as DEC, American Photonics or Ungermann-Bass. However,

physically connecting networks together at OSI, Open Systems Interconnection reference model [7], level 1 is not very desirable. We will end up summing the traffic of the networks and, in some situations, there is the security aspect to be considered. The ideal approach is to connect the networks together logically using a relaying node. The relaying element will typically operate at the OSI level 1.5 MAC (Medium Access Control) layer or at level 3, the network layer. We use equipment that fall into both categories and generally refer to the level 1.5 relaying element as a bridge and to the level 3 element as a router or gateway.

The LANbridge is an Ethernet MAC layer, level 1.5, selective relaying element from DEC. It sits between 2 Ethernets, examines every packet on the networks and decides whether or not to relay it to the other side based on a route table. The route table is generated dynamically through observing the source address of all the packets. Because the bridge has to handle and examine every packet on both networks, very high speed processing, particularly in the area of table look up, is required. The big advantage of this device is that it is completely higher level protocol independent. Hence it can be used to interconnect networks supporting DECNET, IP, XNS and other protocols. There are a number of small disadvantages. Currently at least, the bridge can only be used for the interconnection of Ethernets. Furthermore, each bridge can only be used to interconnect two networks. This makes the cost per connection quite high. Since it is a MAC layer device, network control based on information available at higher level protocols is not possible. Furthermore, one cannot have loops in the topology. Loops, or multiple paths can be very useful for the purpose of load sharing as well as redundancy.

The IBM token ring bridge operates very differently from the Ethernet LANbridge. It is not strictly a MAC layer entity. It relies on the higher level protocol to carry out destination address discovery using broadcast or issuing explicit "resolve" MAC frame requests. A path information consisting of a list of bridges in the path will be returned by the destination station. All subsequent packets between the stations will be sent with a path or route information field (RIF). When a bridge gets a packet containing its ID in the RIF, it will relay the packet accordingly. Hence, unlike the DEC LANbridge, it is not strictly transparent to the stations software.

In general, we deploy bridges for connection of networks which have a heavy mix of protocols such as DECNET and IP packets. For most of the other networks, we use IP routers which are network level relaying elements. The router is developed by the Computer Science department [8] and is specifically designed to handle IP protocols internetting.

The following is a quick tour through the router algorithm :

In the DARPA world, each machine has an assigned IP address. It is a network layer (OSI level 3) address. While the physical address of the machine may change depending on the interface board used, the IP address typically remains associated with the station. In order for an IP machine (IP1) to send a packet to another IP machine (IP2), it must discover the physical address (HW2) of the recipient. If the sender does not already know the mapping, it will broadcast an Address Resolution Request (ARP request) [9]. The ARP request essentially says "I am IP1 at HW1; Will IP2 please let me know your hardware address ?" If IP2 is in the same net, it will hear the request and will reply with its physical address in the form "Hello IP1 at HW1, I am IP2 at HW2"

The above method of discovering the logical to physical address mapping was a DARPA standard and is designed primarily for operation within a single LAN. We extended this for the multi-LAN environment. In that case, when the router hears an ARP request broadcast, it will log

the fact that IP1 has a hardware address of HW1 and then will relay the request to all connected nets as "IP1 at HWR, looking for IP2". Note that the router is replacing HW1 with its own hardware address and is essentially telling the connected nets that it is the agent for station IP1. If IP2 resides on one of the connected nets, it will reply to the router, thinking that it is IP1. The router will relay the reply back to IP1 , again, after substituting its address for IP2's hardware address. From then on, all messages between IP1 to IP2 will be addressed to the router. Note that, unlike the LANbridge described earlier, it does not need to examine every packet in the networks. Instead, it will only need to examine and possibly relay packets addressed to it. This reduces substantially the processing load of the machine.

The advantage of the algroithm is its simplicity. The orginal implementation was done for the PDP-11. Most of the routers currently deployed use lower cost, higher performance 68000 multibus or PC-AT based hardware. Since each router can typically support the interconnection of 3 to 4 networks, the per net cost is significantly lower than that of the LANbridge. Furthermore, the router can support the interconnection of a variety of LAN types including Ethernet, ProNet, IBM token ring, 56K and 9.6K synchronous lines. We are currently implementing a derivative router that can handle a relatively large number of very high speed synchronous lines cost effectively. This type of router is for connecting home based work stations back to the campus using the experimental low cost 64K synchronous services to be provided by our local Bell operating company.

We have been using the routers for almost two years and they have been very reliable. Furthermore, we have built an extensive set of monitoring and network control capability into the routers that can be remotely invoked. This plays a very significant role on our overall network management scheme.

There are three shortcomings to our current routing scheme that we would like to address. First, the algorithm depends on ARP which is native only to the DARPA protocol set. Hence it will not support the interconnection of DECNET or other non-IP stations. Since all machines will support the DARPA campus standard protocol as described earlier, it is not a serious problem. For networks with a high level of protocol mixture, we can fall back to the bridge approach. The second shortcoming is that the ARP requests results in all nets broadcast. Since broadcast traffic has to be handled by every station, it can be very wasteful in CPU cycles particularly for low power stations. While it is not a serious problem at this stage, it can become expensive when our networks contain thousands of stations. Our current version of the router will heuristically relay most of the ARP requests directly to stations instead of using broadcast. The third problem is that the algorithm, similar to the LANbridge, will not allow loops in the topology. This means no alternative paths for either redundancy or load sharing. While it has not been a problem for us since the reliabilty of the deployed routers has been very high, having multiple paths support is still desirable.

We are in the process of developing a second generation router that is sub-net based [10]. It will be able to support multiple paths. In the IP world, the address is composed of a net ID, a subnet ID and a host ID. Each institution is typically assigned a unique net ID. Hence all machines at CMU have the same net number but each different physical LAN on campus will have a different subnet ID. If a station wishes to communicate with another station on the same subnet, it will use the ARP protocol to discover the IP to physical network address mapping as described earlier. However, if the destination station resides on another subnet, the sending station will forward the packet to a router instead. If the router chosen is not the appropriate one either

because of path or load, the station will be informed of an alternative. Meanwhile, the routers will co-operate with each other to determine the best path to get from one IP net to another. This inter-router information exchange and best path determination is still being evolved. There are some moderately good algorithms available [11]. Meanwhile, research is ongoing for more optimum solutions. While this approach will provide multiple paths support as well as elminate ARP broadcast traffic from passing through to non-local subnets, it does make station mobility less easy. Since the IP address of a station is now dependent on the LAN it is attached to, the address has to be changed when a station is moved to another LAN. This is not necessary with our current routing scheme.

It is our goal to provide the minium number of router "hops" between any two nets in our internet configuration. We are currently using a 10 megabit Ethernet as a switching backbone net. Hence the number of hops between most networks should be 2. We are paying close attention to this net and monitoring it frequently for utilization as part of our network management function. Once this net exceeds a certain threshold (currently set at 15% load), we will take steps to segment the net into two.

At the time of writing, we have all our 24 LAN's of various types interconnected. This internet currently supports over 800 nodes. The most significant users of the internet are the 200 high power work stations scattered across campus accessing the Andrew distributed file system. We are anticipating the load to increase as the number of Andrew nodes escalates. The target for Fall 1986 is at least another 800 nodes.

THE BACKBONE NETWORK

For the inter-building outside plant, the main contenders are broadband or fibre optic systems. We selected the fibre optic approach since it offers better noise immunity and is less susceptible to interference from lightning. It is generally configured in star topology, which is much better from a network maintenance point of view. In terms of research and development, there is much more activity in the fibre optic area. A fibre system also fits well with our desire to support PABX service at a later date. Currently, we have a large 50 micron fibre optic plant with over 150 cables reaching most buildings on campus. Over the past three years, we have developed a substantial amount of expertise and positive experience with this media. Therefore, we are enhancing our existing fibre plant instead of installing new broadband cables as the main inter-building trunking system. One change we will be making, however, is the type of fibre to be installed. While the current fibre plant is based on 50 micron cable, most of the data equipment manufacturers have engineered their product for 100 micron operation. The high insertion and connection loss have been a problem on a number of occasions. However, since most of the fibre and equipment manufacturers seems to be converging towards the new AT&T standard of 62.5 micron, we will be installing that in the future.

NETWORK MANAGEMENT

We cannot conclude this paper without at least touching on the subject of network management. For a network of the size we have at CMU, the ability to manage operations and

growth is very important. We have essentially divided network management into two separate but closely related sub-tasks. They are operations and evolution management. Operations management deals with problem determination and isolation. Problem determination consists mainly of real time network monitoring and periodical active probing of co-operating stations. It also consists of longer term "soft" or self recoverable error statistics gathering to assist in anticipating points of failure. To this end, the logical design of the IBM token ring lends itself well to the task. We also have built up a significant collection of tools for this purpose. Besides determining the source of a problem, the ability to quickly localize and isolate it without significantly reducing the overall capability of the network is also important. This calls for careful network configuration. We have tools for monitoring traffic profiles in the network. This is part of our evolution management tool kit. Its function is to gather and analyze network performance related information to help us with planning. It also contains a collection of network data bases. Besides enhancing our management tool sets, we are working towards better integration of the tools and designing a hierarchical set of management interfaces for different categories of operational staff. The use of expert systems to assist in network management is also being investigated. A detailed description of this area of activities will be described in a future paper.

CONCLUSION AND ACKNOWLEDGEMENT

Over the past two years, the local area networking activities at Carnegie Mellon has evolved from a somewhat experiemental enterprise to a large scale coherent and integrated system. The fact that the internet is functioning relatively smoothly today is the result of a substantial amount of technical and administrative co-operation from virtually every department on campus, particularly from the Computer Science and Robotic Institute; Electrical and Computer Enginnering; Computing Center; and the Information Technology Center, which is a joint venture between CMU and IBM. We have also obtained a lot of valuable information and assistance from various other academic and research institutions through the ARPANET.

REFERENCES

[1] Morris, Satyanarayanan, Connor, Howard, Rosenthal, Smith "Andrew - a Distributed Personal Computing Environment" Communications of the ACM, March 1986

[2] CSMA/ CD, IEEE Std 802.3-1985 (ISO/ DIS 8802/ 3)

[3] Shoch, Dalal, Redell, Crane "Evolution of the Ethernet Local Computer Network" IEEE Computer, August 1982

[4] Stole "A local Communications Network Based on Interconnected Token Access Rings : A Tutorial" IBM J. Res. Develop. Vol. 27. No. 5. September 1983

[5] Token Ring Access Method and physical Layer Specifications, IEEE 802.5-1985 (ISO/ DP 8802/ 5)

[6] Inside AppleTalk, Apple Computer Inc.

[7] International Standards Organisation, Data Processing "ISO DP7498 : Open Systems Interconnection - Basic Reference Model"

[8] Accetta "DARPA Internet protocol Service on the CMU Local Area networks" internal paper, Computer Science department, Carnegie-Mellon University

[9] Plummer "An Ethernet Address Resolution Protocol" RFC826, November 1982

[10] Mogul, Postel "Internet Standard Subnetting Procedure" RFC950, August 1985

[11] Xerox System Integration Standard : Internet Transport Protocols, Chapter 4 "Level two : Routing Information Protocol"