CONF-960116--1

SAN095-1103C

Implementing and Testing ATM in a Production LAN

John Naegle,
MS0806, PO Box 5800
Sandia National Laboratories
Albuquerque, NM 87185-0806,
(505)844-8044, jhnaegl@sandia.gov, Fax (505)844-2067

Nick Testi
MS0806, PO Box 5800
Sandia National Laboratories
Albuquerque, NM 87185-0806,
(505)844-4896, ntestil@sandia.gov

Larry Tolendino
MS0806, PO Box 5800
Sandia National Laboratories
Albuquerque, NM 87185-0806,
(505)845-8587, lftolen@sandia.gov

John Zepper
MS0827, PO Box 5800
Sandia National Laboratories
Albuquerque, NM 87185-0806,
(505)845-8421, jdzeppe@sandia.gov

Abstract:

Asynchronous Transfer Mode (ATM) technology is currently receiving extensive attention in the computer networking arena. Many experts predict that ATM will be the future networking technology for both the Local Area Network (LAN) and the Wide Area Network (WAN). This paper presents the results of a collaboration between Sandia National Laboratories' Advanced Networking Department and Engineering Sciences Center to study the implementation of ATM in one of Sandia's most heavily loaded production networks. The network consists of over 120 Sun Sparc 10s and 20s, two SparcCenter 2000s, a 12 node parallel IBM SP-2, and several other miscellaneous high-end workstations.

The existing network was first characterized through extensive traffic measurements to better understand the capabilities and limitations of the existing network technologies and to provide a baseline for comparison to an ATM network. This characterization was used to select a subset of the network elements which would benefit most from conversion to the ATM technology. This subset was then converted to equipment based on the latest ATM standards.

With direct OC-3c (155 Mbps) host connections for the workstations and the file and compute servers, we demonstrated as much as 122 Mbps throughput (memory-to-memory TCP/IP transfers) between endpoints. Flow control in the classical many-to-one client server environment was also investigated.

Throughout all of our tests, the interaction of the user applications with the network technologies was documented and possible improvements were tested. The performance and reliability of the ATM network was compared to the original network to determine the benefits and liabilities of the ATM technology.

This work performed at Sandia National Laboratories supported by the U. S. Department of Energy under contract DE-AC04-94AL85000.

DISCLAIMER

Portions of this document may be illegible electronic image products. Images are produced from the best available original document.

I. Introduction

The Advanced Networking Integration Department (ANID) at Sandia National Laboratories is responsible for evaluating and integrating new networking and communications technologies into the Sandia infrastructure. As part of this assignment the ANID has investigated Asynchronous Transfer Mode (ATM) technology for over 5 years and has determined several benefits that ATM and SONET provide in our corporate communications infrastructure. In order to evaluate and develop ATM and SONET technologies for deployment throughout Sandia's communications infrastructure, we have operated an ATM testbed since the first ATM equipment was available. As the technology matured, we effectively demonstrated the capabilities of ATM beyond that of legacy network technologies. These efforts in the Local Area Network (LAN) and Wide Area Network (WAN) environments have been reported in several publications[1,2,3]. However, the testbed environment has always been tightly controlled and limited in scope. In order to find the pitfalls and bottle necks of ATM technology deployment, a more demanding environment was required. To test ATM in the demanding LAN environment, the ANID has collaborated with the Engineering Sciences Center (ESC) to trial ATM in their production network.

The ESC performs fundamental research in such diverse areas as fluid dynamics, hypervelocity airflow, structural dynamics, and high-energy physics. Their research and investigations make extensive use of powerful computer simulations that both model and give insight into real world systems and physical phenomena. This work requires state-of-the-art, high powered computer systems that allow the engineers to visualize and analyze simulation results. To meet these requirements, the ESC maintains a large and sophisticated network of desktop workstations, file and application servers, and several powerful compute and visualization servers. In order to investigate ever more complex systems and phenomena, the performance and capacity requirements of this network continually grow. The ESC network managers have investigated several technologies to provide greater network bandwidth for their LAN. This ATM trial provided a handful of ESC workstations with high speed ATM access to network resources prior to general ATM availability at Sandia. This early ATM deployment provided the ANID the opportunity to

study the issues involved in configuring and maintaining a state-of-the-art ATM LAN. This paper presents the results of the ATM trial deployment and the lessons learned.

II. Current State of the Engineering Sciences Network

The Engineering Sciences Center employs several applications to accomplish the simulation of complex systems. Some of these applications, such as Patran and the Advanced Visualization System (AVS) are commercial codes but many others are developed in-house to meet specific requirements. These applications are computationally intensive and some require a large amount of network bandwidth. In order to support the computation and network requirements of their applications, the ESC maintains state-of-the-art computers and networks. The communication protocol in this network is TCP/IP. Dynamic NFS is used to mount file systems whenever required and X11 is used extensively for remote viewing of graphics and application user interfaces. This section gives a summary of the current configuration and performance of the ESC network.

Configuration

Each of the roughly 120 engineers in the ESC has a powerful workstation (a Sun Sparc 2, 10, or 20) with enhanced graphics capabilities at their desk. See figure 1 for a simplified diagram of the production Ethernet based network. To balance the system management requirements and the network load, each computer has a local disk for the operating system, but uses a Sun SparcCenter 2000 for file and application services. Depending on their requirements, the engineers have several choices for computational resources. Smaller jobs can be run locally on their desktop while larger jobs can be run on a local SparcCenter 2000 compute server, a local 12 node IBM SP-2, or several supercomputers and massively parallel machines which are available from Sandia's central computing facilities. There are also several local SGI and HP computers available for visualization and compute services. Other services available on the network include input/output devices such as high-speed black and white printers, several full color printers, high-density scanners, and CD writers.

The production networks that interconnect all of these devices consists of two main Ethernets and a few back-door Ethernets for connecting special systems. The desktop computers and I/O devices are

split between the two main Ethernets while the large compute and file servers are attached to both networks. A Cisco 7000 router routes traffic between the two Ethernets and the rest of Sandia's networks.

Traffic Measurements

Two Ethernet Sniffers from Network General were used to determine the traffic characteristics on the production networks. Network utilization statistics were summed over a 10 second interval and then written out to disk. This allows reasonably fine-grained statistics to be collected over a long time period. We gathered these statistics for 2 weeks to ensure that our data was not dominated by a short term anomaly. Although this is by no means an exhaustive characterization of the network, several interesting observations can be made from our data.

The data measured during the 10 second sample period was too variable to get a general view of the load on the networks. The plots were so jagged that it was impossible to detect the trends. We therefore smoothed the data by averaging over a 5 minute time period. Figure 2 shows a typical plot of the smoothed data over a full workday (7:30 AM to 5:00 PM). This was not at all what we expected to see! This network has over 60 desktop workstations and several compute servers. Each pair of workstations is capable of using 90% of the Ethernet in a single FTP session. Yet, in the aggregate, average usage was only 15%. This demonstrates the power of statistical sharing of limited resources. Since Ethernet is a fixed network resource, minimizing network bandwidth requirements has been the goal of application and system architects. If a network is well designed, with sufficient caching and local resources, the amount of traffic that traverses the network can be minimized. These networks demonstrate that even with 60 engineers using high-powered applications and resources, the average network usage can be kept within Ethernet's capabilities.

On the other hand, even when viewing the smoothed data, the very bursty nature of network usage is evident. Several peaks of over 30% usage for longer than 5 minutes indicate that there are applications using a significant portion of the Ethernet over extended periods of time. If we analyze the 10 second average raw data, we see that within almost every 5 minute period, a large peak occurs. Figure 3 is a histogram of network usage. In a collision based technology such as Ethernet, it is difficult to determine what level of network traffic reduces performance for users. The delay caused by the Ethernet

back off algorithm and the throughput reduction due to sharing the Ethernet bus will eventually cause the users to feel that the network is slowing down. The ESC users report that the response of the network is generally good, but that there are times when response drops off considerably. This agrees well with the histogram, which indicates that the network is sustaining loads over 50% for 4.65% of the time and that the network is over 30% loaded for 11.22% of the time.

There are several reasons why these networks are not experiencing much heavier network congestion. The most important is careful design and management of the overall computer and network systems. The network managers have investigated several different network environments such as diskless and dataless clients, Xterms, and diskfull standalone workstations. Their choice of putting the core of the operating system on a local disk, putting most of the applications and user data on a central server, and installing large amounts of memory in the desktop workstations has been very effective in balancing the demands of network bandwidth utilization, user performance, and administrative operations.

Basing the applications on TCP/IP reduces network bandwidth requirements when compared to other common protocols. Both Novell IPX and AppleTalk provide superior automated functionality at the expense of increased bandwidth requirements. In order to provide this superior functionality, both protocols rely on excessive broadcasts messages as well as other service descriptor messages. These messages announce available services and the current network state while greatly increasing network traffic. IP, and the applications it supports, use broadcasts sparingly and rely on preconfiguration to minimize network traffic.

NFS packets comprise the bulk of the traffic on the ESC network. The small maximum buffer size (8 KBytes) used in the current NFS[4] implementation limits the achievable throughput and also helps to reduce peaks in network usage. The maximum NFS throughput achievable between a Sparc 20 and the SparcCenter 2000 is only 3 Mbps, or about 30% of the Ethernet. The effective throughput that users typically experience is probably much less. Therefore, several typical concurrent disk accesses can be sustained within the available Ethernet bandwidth without congestion being noticeable to the users.

Any application that is found to require large amounts of network bandwidth becomes unpopular very quickly. Not only do the other users on the network complain, but the performance is so poor that the application is not effective. Therefore, current applications artificially limit their functionality to perform within the constraints of the Ethernet environment. If more network bandwidth were widely available, engineers would use all the functionality that exists, but is not currently effective, and the application programmers would quickly add more functionality. For example, a system composed of AVS rendering on an SGI Challenge system with the display transmitted to a SUN Sparc 20 desktop has the ability to rapidly render large wire-frame and shaded models to help the scientist visualize physical objects and simulation results. This process can easily use 32 Mbps of network bandwidth on an ATM network. If this process is attempted on the Ethernet, the poor performance makes it ineffective and the rest of the network users suffer severe degradation in their network operations.

In summary, the existing ESC Ethernets today are performing quite well considering the number of users and the sophisticated applications that are in use. The network does experience regular congestion periods where the users perceive a degradation in performance. There are also applications and functions that are available today that are not being used because of their ineffectiveness over Ethernet and the severe network congestion they cause. Several new applications, such as enhanced visualization, distance learning, distributed computing, and desktop video-conferencing will be enabled by greater network bandwidth. The network bandwidth must be increased well beyond shared Ethernet to take advantage of the capabilities of these emerging applications.

III. Options for network upgrades

The network managers for the ESC network have considered several alternatives for increasing the bandwidth available to the users. This section discusses the benefits and limitations of these options.

Segregation

One method of increasing aggregate network bandwidth is to segregate networks into ever smaller groups and interconnect them with bridges or routers. The ESC network has already started this process by creating two networks and routing between them. Segregation by routing is practically limited

by the inefficient use of network address space and the cost of router ports. There are some IP based and many non-IP based applications that do not operate across a router. Another option is to dedicate a full Ethernet to each desktop and bridge between each of these dedicated Ethernets. This method is commonly called switched Ethernet and has become very popular as the technologies to bridge large numbers of Ethernets has become reasonably priced. Bridging allows all of the nodes to appear to the applications as if they were on the same network, which allows all of the applications designed for a shared bus topology to operate correctly. It also increases the aggregate bandwidth of the network dramatically by giving each attached device a dedicated 10 Mbps Ethernet pipe without changing the host interface. As mentioned in the previous section, there are few applications available today that can effectively use more than 10 Mbps.

The client-server model of computing, which is fundamental to the ESC network and most networks in general, makes network segregation less effective and impractical in many cases. If all the network stations are competing for the resources on just one file server, then all of the network traffic is competing for the single Ethernet dedicated to the server. Obviously, the Ethernet attached to the server will become the bottleneck and no net gain in bandwidth is achieved. There are also applications that require more than a dedicated Ethernet can provide. The solution to this problem is to connect the server and top-end users to higher speed networks such as 100 Mb Ethernets or an FDDI rings. However, this solution introduces another set of problems such as overflowing of buffers (in the bridges or routers) in the many-to-one case and the high-speed to low-speed case, implementing and integrating multiple network technologies, replacing server interface cards, and scaling to very large switched networks.

Faster Bus Based Technology

Another option to gain more network bandwidth is to upgrade the complete network to a faster bus based technology. Since there are no routers or bridges involved, this is a much simpler architecture. There are no buffers in the network so packet loss due to congestion is not a problem. But this method requires a complete swap of equipment infrastructure and host interfaces. Each host would require a more expensive, faster interface even though it may not be able use a significant portion of the available

bandwidth. Eventually, even the faster bus will be congested as applications and the number of users grow and the same network upgrade options will have to be faced again.

Switching

A more efficient solution is to build a network where each host has a network interface matched to its capabilities and requirements. As greater network speeds are required in the servers and high-powered desktops, their interfaces are upgraded while the core network technology and the rest of the hosts remain the same. By using switching technology, similar in function to the switched Ethernets mentioned above, virtual networks can be configured to provide network separation while allowing more flexibility for moves, adds, and changes than the current router segregated architecture allows. By requiring that all packets that traverse the core of the network be a fixed 53 byte size, the core infrastructure could also carry traffic that is very sensitive to latency, such as isochronous voice and video. Obviously we are describing an ATM network. While ATM does have all of these and several other benefits, it also has some significant problems. The issues of congestion management mentioned above for switched Ethernet is even more difficult in switched ATM given the wide disparity in available interface rates. Since ATM standards for LAN implementations are relatively young, there are several issues such as dynamic connection setup, addressing, broadcasting, and interoperability with upper layer protocols that are still being defined in the standards bodies. The rest of this paper describes the problems we have encountered and the benefits we have achieved in our deployment of ATM in the ESC network.

IV. Integrating ATM into the Engineering Sciences Network

The configuration of the ESC ATM trial is shown in figure 4. Six engineers with high bandwidth requirements and an interest in advanced applications agreed to convert their desktop workstations to ATM. Both of the SparcCenter 2000s, an SGI Challenge, two of the SP-2 nodes, and the router to the rest of Sandia's network also had ATM interfaces installed. The existing Ethernet interfaces in these machines were left in place for a backup if the ATM network failed. This allowed us to create an environment where all of the current functionality of the production Ethernet was duplicated and new capabilities could be added as performance allowed. The ATM network was spread across three ATM

switches since the hosts were distributed across several wiring closets. All of the ATM interfaces were optical OC-3c (155 Mbps). Four of the switch-to-switch interfaces were single mode, while all of the other interfaces were multimode. The remainder of this section discusses some of the interoperability issues encountered and the raw ATM performance achieved.

Interoperability

The ESC trial network was intentionally created using equipment from multiple vendors to learn more about the standards and interoperability issues. The central switching equipment was from Bay Networks, the host Network Interface Cards (NICs) came from Interphase, Efficient, and FORE, and the router interface was Cisco's AIP card.

In a previous paper[1], we have reported that ATM interoperates very well from the SONET through the application layers when using Permanent Virtual Circuits (PVCs). We had hoped that by the time this paper was written that we would have a complete network using the ATM Forum LAN Emulation (LANE) standard for Switched Virtual Circuit (SVC) connection setup. Unfortunately, all of the vendors did not supply interoperable LANE code in time for this paper. Therefore, the results in this paper are based on PVC connections. As described in our previous paper[1], all of the various ATM vendors could communicate with each other by using the TCP/IP protocol and by ensuring that the IP encapsulation methods matched.

We learned one crucial lesson as we moved from the ANID testbed to the ESC production environment. All of the hosts used in the ANID testbed use generic operating systems where patch levels and detailed version numbers were not carefully controlled. As we attempted to install the same ATM NIC drivers into the ESC environment where all of the latest versions and patches were strictly maintained, we encountered several problems. Since none of our NICs came directly from the host vendors, there is a time lag from the time the operating system (OS) is modified to when the NIC vendors are able to fix any new problems that may have been introduced by the OS modification. It is understandably difficult for third-party vendors to deal with all of the subtleties and interactions of a vendor's OS and hardware. Much better coordination between the third-party NIC vendors and the host vendors will be required to provide the robust, manageable environment required for production networks.

Raw ATM Performance

ATM network performance can be measured at many different levels and has several limiting factors. This section will deal strictly with host-to-host, memory-to-memory performance using the full TCP/IP stack. The two performance characteristics that we will discuss in this section are the maximum throughput and the round trip delay. The throughput tests in this section were performed using the test TCP (TTCP) program, available from the Internet, to open a standard TCP socket between two machines and send raw data using the full TCP guaranteed delivery mechanisms. While this process uses the same protocol stack as most applications, throughput is not limited by disk speeds or application processing time. Since we were interested in measuring the maximum performance of ATM interfaces, the full TCP/IP window size of 64 KBytes was used for all reported results. To measure round-trip delays, we wrote an application that uses the entire protocol stack and the TCP Echo port to echo several packets and record the average round-trip time. We must mention one note of caution when we report ATM performance measurements. All of the ATM measurements reported here are currently repeatable in the ESC network. However, as we have progressed with our testbed experiments, we have watched the performance of ATM interfaces steadily increase, with occasional decreases as new drivers are implemented. Therefore, even we may not be able to reproduce these exact results in a few months.

Echo testing was performed on a pair of SUN Sparc 20 desktop workstations which were connected via Ethernet and ATM. Two packet sizes were utilized and the echo test results are shown in the following table.

Echo Packet Size in bytes	Ethernet Time in msec	ATM Time in msec
64	2.1	1.75
1500	4.1	2.1

The echo test times for small packets are virtually identical because the packet processing time dominates the test results. Regardless of the network speed, most of the echo test time is derived from the time necessary to process the packet through all the layers of the TCP/IP protocol stack. On the other

hand, round trip transmission times become a significant factor in the echo tests for large packets. While the effects of delay can be mitigated by using a large TCP window size[5] for bulk data transfers, delay can be a serious limiting factor in distributed computing environments where many small messages are transmitted.

Although the clocking rate of OC-3c is 155.52 Mbps, the actual SONET bandwidth available to the user is 151.488 Mbps. Since 5 out of every 53 bytes in the ATM cell is overhead, that leaves only 137.19 Mbps of usable bandwidth. When using the standard 9180 byte MTU size for ATM, the AAL5 and TCP/IP overhead for maximum size packets is less then 2%. Therefore, the maximum throughput possible over OC-3c, ATM interfaces is roughly 135 Mbps. While the early ATM equipment we tested could not achieve throughputs close to this data rate, the equipment in the ESC network has come very close to the maximum theoretical throughput. Between two nodes in the SP-2, we were able to sustain 122 Mbps, even under production (100% CPU usage) operation. We have not been able to schedule dedicated time on the SP-2 yet, but the rate we have demonstrated is very impressive. This throughput is almost as high as the 125 Mbps throughput achievable with TCP/IP over the proprietary interconnect switch supplied with the SP-2. Each of the SP-2 nodes is an extremely powerful IBM model 590 operating at 67 MHz.

There were a few interesting variations in the performance from the Sparc 20s to the SparcCenter 2000s. While the CPU in the Sparc 20s and the SparcCenter 2000 compute server were model 50s, the SparcCenter 2000 file server had model 41 processors. When the file server transmitted to the Sparc 20, the throughput was 37 Mbps, but when the compute server transmitted to the Sparc 20, the throughput was 45 Mbps. Although the process of segmentation and reassembly (SAR) is done on the ATM NIC, the CPU processing time of the TCP/IP stack still appears to be a bottle neck. Another test demonstrated how the efficiency of the network stack can greatly affect the throughput. The tests mentioned above were performed with all of the systems running Solaris 2.3. When the compute server was upgraded to Solaris 2.4, the performance increased to 55 Mbps. We are told that the main difference between the network stacks is that Solaris 2.3 performs 3 memory to memory copies of a data packet as it progresses through

the protocol stack while 2.4 only performs 1 copy. This both verifies the CPU bottleneck and indicates that the network software must be carefully optimized to reach peak performance.

The interfaces in the SGI and Cisco router were also capable of sustaining very high throughput. From the SGI to a Sparc 20 with a model 61 processor, we measured 82 Mbps, and in the reverse direction, 64 Mbps. So far, we have only been able to push the Cisco interface to 60 Mbps, but we suspect that we are seeing the limit of the hosts we used in the tests, and not the router interface itself.

One last performance variation we will discuss is caused by the different encapsulation methods defined in the standards. All of the measurements reported above are for the SNAP and the NULL encapsulations. When we tested new drivers using the LANE encapsulation, we have seen as much as a 20% degradation in throughput, even after ensuring that the MTU size is the same. Since there is very little difference in the amount of overhead between the various encapsulation techniques, we suspect that this difference is caused by the lack of optimization in the new drivers. We hope that the performance of the new drivers will increase to match the performance of the more mature drivers in the near future.

V. Application Performance and Bottlenecks

This section will discuss the results of our tests of ATM with applications used in the ESC network. These particular applications were chosen to represent the current, evolving, and future production applications that are most important to the ESC engineers.

NFS

The vast majority of the traffic on the ESC network will continue to be NFS. From our test results in the ESC network, it appears that performance of the currently available NFS is severely limited by the small NFS READ/WRITE buffer size (8K bytes) and not the physical network capabilities. By simple wall-clock timings, we measured roughly 3 Mbps throughput using NFS over Ethernet between a Sparc 20 and the file server. The same test performed over ATM only resulted in 3.2 Mbps throughput. Although the ATM network is an order of magnitude faster, there was no significant improvement in the NFS throughput. Even with the much more powerful SP-2 processors, NFS to the SparcCenter 2000 compute server increased to 10.6 Mbps, just beyond Ethernet capability. A larger buffer would improve

network utilization by allowing more data to be pipelined, analogous to the operation of the TCP window size[5]. Although we have not been able to perform NFS tests between two Sun Solaris 2.4 machines yet, Sun has claimed that their NFS performance should improve by 26%. The performance of NFS should improve much beyond the capability of Ethernet when Version 3, which will allow for 64 KByte buffers[6], is released.

FTP

FTP performance in the Ethernet production network was limited to roughly 9.2 Mbps by the inherent data rate of Ethernet. In the ATM network, disk speeds became the limiting factor. The disks in the desktop Sparc 20s were standard 1 GByte SCSI with a throughput of roughly 16 Mbps and the disks in the file server are fast and wide SCSI-2 models capable of 96 Mbps throughput. The FTP throughput from the file server to the Sparc 20s was limited by the slower disks to 14 Mbps. The SP-2s are also equipped with fast and wide SCSI interfaces and are capable of achieving 73.5 Mbps between two SP-2. As with the TTCP tests, it is very important to make sure the TCP window size is large for these tests. Since FTP does not currently have the ability to set the TCP window size, the system default is used. Care must be taken to ensure that the system has been configured with a window size larger than 32 KBytes to achieve our throughput results.

Advanced Visualization System (AVS)

The AVS application is used extensively for visualizing simulation results and physical models. A full 3-D view can be rendered and displayed on a high-resolution, full color display. AVS is written to be highly modular so that its different functions can be distributed across different network hosts. This gives the network manager several configuration choices. The rendering process requires a tremendous amount of computation that can be implemented in special hardware. Traditionally, high-performance visualization has required special graphics engines such as the SGI Challenge machines with Reality Engines to render the model and display the view on a local console. The performance capabilities of ATM will enable a change in the configuration such that the rendering can still be performed on centralized specialty hardware but would then use X11 to remotely display the final view on the engineers desktop. This allows the very large model databases required for the rendering to be moved to the

graphics engine once, thereafter only display information traverses the network. As previously mentioned, transmitting the X11 display information can use large amounts of bandwidth. Even with the current software rendering on the SGI Challenge, we have measured a sustained data rate of 32 Mbps as an engineer continuously rotated his model to see the simulated system from all angles. This function is unacceptably slow and disruptive in an Ethernet environment. Functionality such as this, enabled by the ATM network, helps the engineers understand and interpret the results of their simulations much more quickly and intuitively.

Parallel Computing

The engineers in the ESC use parallel programming to speed up their simulation applications. Their local 12 node SP-2 cluster is one of their main compute resources. Although the cluster has a very high speed, proprietary switch for communication among the SP-2 nodes, there is still a serious bottleneck when communicating with the rest of the network. This is a common problem shared by most parallel machines. Massively parallel machines are the future hope for solving the huge computational problems that are beyond the capabilities of the current scalar machines. These problems usually require extremely large data sets for input and generate even larger data sets as output. The inter-node communication in current parallel machines is generally very good, but the complete system problem of getting the data in and out of the machines is still a bottle neck. The SP-2 in the ESC normally uses NFS over Ethernet to get and store the data sets. As mentioned above, the NFS throughput is only 3 Mbps and causes the SP-2 to run jobs much slower than the engineers expect. Even when FTP is used over Ethernet to move the data to a local SP-2 disk, the transfer time can be a significant part of the job run time. With direct ATM between the file server and the SP-2, NFS still performs poorly, but FTP throughput is 30 Mbps. With the full TCP/IP stack, the performance of the proprietary SP-2 switch and the ATM network was almost the same. If the performance of ATM without the TCP/IP stack matches the performance of the proprietary switch, both the inter-node and external I/O could be accomplished using a single technology. This would significantly simplify and improve the overall system for configuring, maintaining, and running parallel codes on the SP-2.

The ATM network can also be used to tap the vast store of CPU power that sits idle on the engineers' desktops. The interest in parallel computing and the availability of excess CPU cycles on most workstations has led the industry to generate several standards to allow heterogeneous distributed computing. The Message Passing Interface (MPI) standard allows programmers to write parallel code that runs transparently on most massively parallel machines as well as distributed computing environments. We used a benchmark program distributed with the MPI application that tests the effective throughput of inter-node message passing over whatever network is used by simply measuring the time it takes to pass several messages between nodes. We used this as a relative test to compare the different network technologies rather than try and obtain absolute network bandwidth usage. On the Sparc 20 workstations with 4 nodes participating in the message passing, ATM was somewhat erratic, but still roughly 30% faster than the Ethernet. This agrees well with results published by Lin, et al[7] . They also found little difference between Ethernet and ATM in small distributed computing environments. We agree with their assertion that the benefit of ATM will be in the scalability of distributed computing. Many parallel algorithms synchronize communications such that all nodes pass data to neighbors at the same time. Therefore, all of the nodes will attempt to access the Ethernet at roughly the same time, resulting in severe congestion. In the switched environment of ATM, each pair of workstations can communicate without effecting other traffic. This will allow ATM to scale to very large distributed computing environments while an Ethernet can not. Bypassing the TCP/IP stack and using native ATM will reduce the network latency and further improve performance.

VI. Maintenance and reliability of the ESC ATM network

In setting up an heterogeneous ATM network of this size and capability, we have learned that the maintenance and reliability issues are much more complex for ATM than for Ethernet. Ethernet is a very mature, plug-and-play, reliable technology. ATM still requires careful configuration and monitoring to ensure reliable operation.

Connection setup

The setup of virtual circuits (VCs) between computers is the most difficult part of the ESC ATM network. The goal is to be able to plug an interface into the ATM backbone and have all connection management operations performed automatically. Unfortunately, the standards to allow this functionality are not available from all of our vendors and we must manually configure PVCs for all connections. Since the number of PVCs required to fully interconnect n hosts is $O(n^2)$, we did not even attempt to fully interconnect all of the ATM hosts on the ESC network. Some ATM switching implementations require a PVC entry in each switch. Since we had a common connection management system (CMS) for all of our ATM switches, we only had to enter the PVC once and the interconnect path through the switching network was automatically configured. These PVCs were entered into a batch file that was run whenever required. Although having a single CMS greatly simplified the configuration of the network, it also introduced a single point of failure that could bring the whole network down. Whenever any switch lost communications with the CMS, all of the PVCs were cleared when the communications were restored. Since there was no mechanism for automatically restoring the PVCs, manual intervention was required whenever failures occurred. This has been the most common cause of user down-time in the ATM network. Along with the CMS, each host had to set up its PVC information before it could communicate to any other ATM attached hosts. Automating this process was required since most of the hosts attempted to NFS mount parts of their operating systems as they boot up. This was a difficult maintenance operation since each NIC vendor had a different method for setting up PVCs during the boot process. Eventually we learned to manage each of these difficulties and the connection management has been very stable. We have only experienced a few of minutes of unplanned user down-time in the 6 weeks that we have been in full pseudo-production. Due to the effort required to configure the PVC tables, managing the moves, adds, and changes in the ESC still requires a significant amount of effort. We will not attempt to managing a large LAN environment based on PVCs with the current tools at our disposal.

Cell Loss

Once the VC has been established, there are several possibilities for ATM packets to be corrupted and dropped. All of our equipment can display the SONET statistics, and we have not seen any cells

dropped due at the SONET physical layer. Once within the switching network, the cell may be misdelivered, corrupted in a switch, or dropped due to congestion. Our internal switch statistics report that we have never experienced corruption in a switch.

Cell loss due to congestion in the ATM network is very easy to create. The only congestion control mechanisms available in our current ATM equipment is large switch output buffers and static rate control at the host interface. We are currently not enforcing any rate control on the ATM host interfaces. Since the hosts on the ESC network are capable of using almost the full ATM interface bandwidth, even a simple TTCP test of two hosts transmitting to one receiver causes severe network congestion and cell loss in the buffers[8]. But in monitoring the ESC network, we have not seen any of the users' applications put enough load on the network to threaten congestion cell loss. As mentioned in previous sections, we do not expect to see user applications that push the ATM network to it limits for some time. We believe that the congestion control standards being developed in the ATM Forum will be available before the applications are capable of causing significant congestion problems. If the applications that cause congestion do arrive before the standard is implemented, we may have to use the static rate control on the host interfaces to limit maximum bandwidth usage. This limitation will still provide aggregate network performance well beyond what is currently available on the ESC Ethernets.

VII. Conclusion

The collaboration between the ANID and the ESC to trial ATM in a production environment has been very successful. The ANID has learned valuable lessons about deploying ATM in production networks that could never be discovered in a testbed environment. The utilization of ATM using the powerful systems in the ESC network is much higher than we could achieve with testbed facilities. In-depth analysis of the network usage patterns of a large, sophisticated production Ethernet has helped to guide our plans for general deployment of ATM in Sandia's LAN environment. The limited requirements of most current user applications gives us some time to trial ATM in production networks without having to deal with congestion control issues. Most of the current Ethernet congestion could be greatly reduced

by deploying bridged Ethernet devices attached to the ATM backbone. Servers and high-end users (such as the ESC engineers) will be converted to ATM while users who do not need more than a dedicated Ethernet will not require any changes. Maintaining the ESC network has also validated our original plan to not offer ATM as a general service in the LAN environment until standard SVCs are available from all of our ATM vendors.

The ATM network has been reliable and the improved network performance has enabled the ESC engineers to investigate applications that are beyond the capabilities of their current Ethernets. These and future applications promise to improve the efficiency and capabilities of the engineers by allowing them faster access to their network resources. As their overall system for computation and analysis improves, they will be able to investigate phenomena that are beyond their current resources. The results of the trial ATM network have convinced the ESC network managers that ATM will be able to satisfy their future network requirements and that intermediate technologies need not be considered.

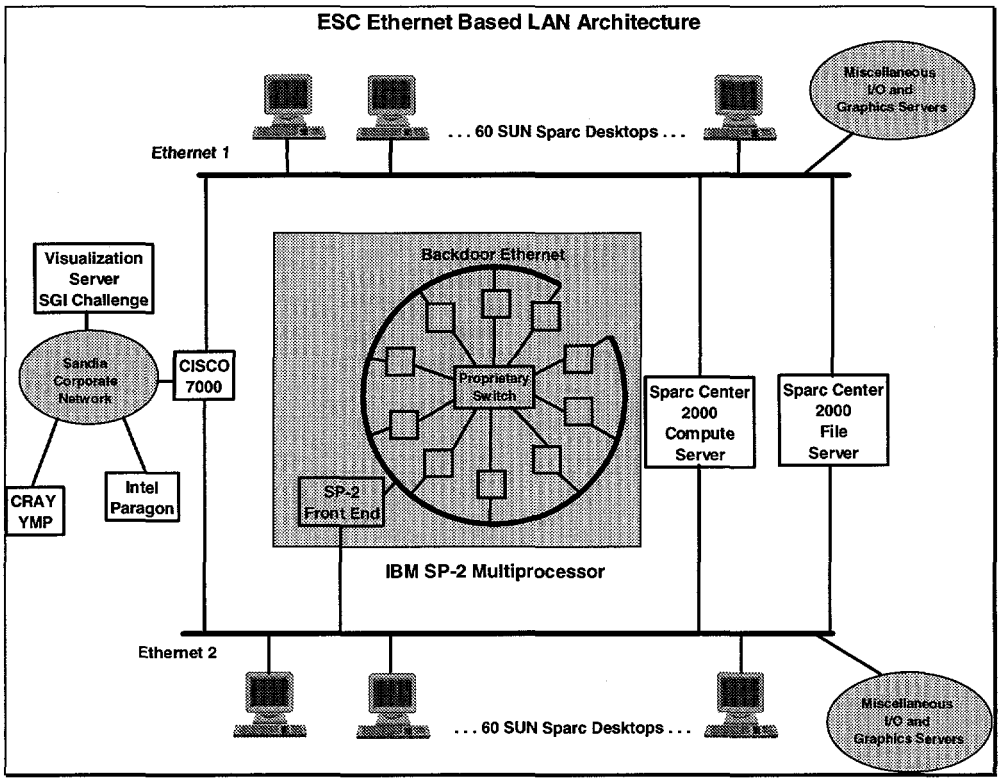
ESC Ethernet Based LAN Architecture
Miscellaneous I/O and Graphics Servers
. . . 60 SUN Sparc Desktops . . .
Ethernet 1
Visualization Server SGI Challenge
Backdoor Ethernet
Sandia Corporate Network
CISCO 7000
Proprietary Switch
Sparc Center 2000 Compute Server
Sparc Center 2000 File Server
CRAY YMP
Intel Paragon
SP-2 Front End
IBM SP-2 Multiprocessor
Ethernet 2
. . . 60 SUN Sparc Desktops . . .
Miscellaneous I/O and Graphics Servers

Figure 1

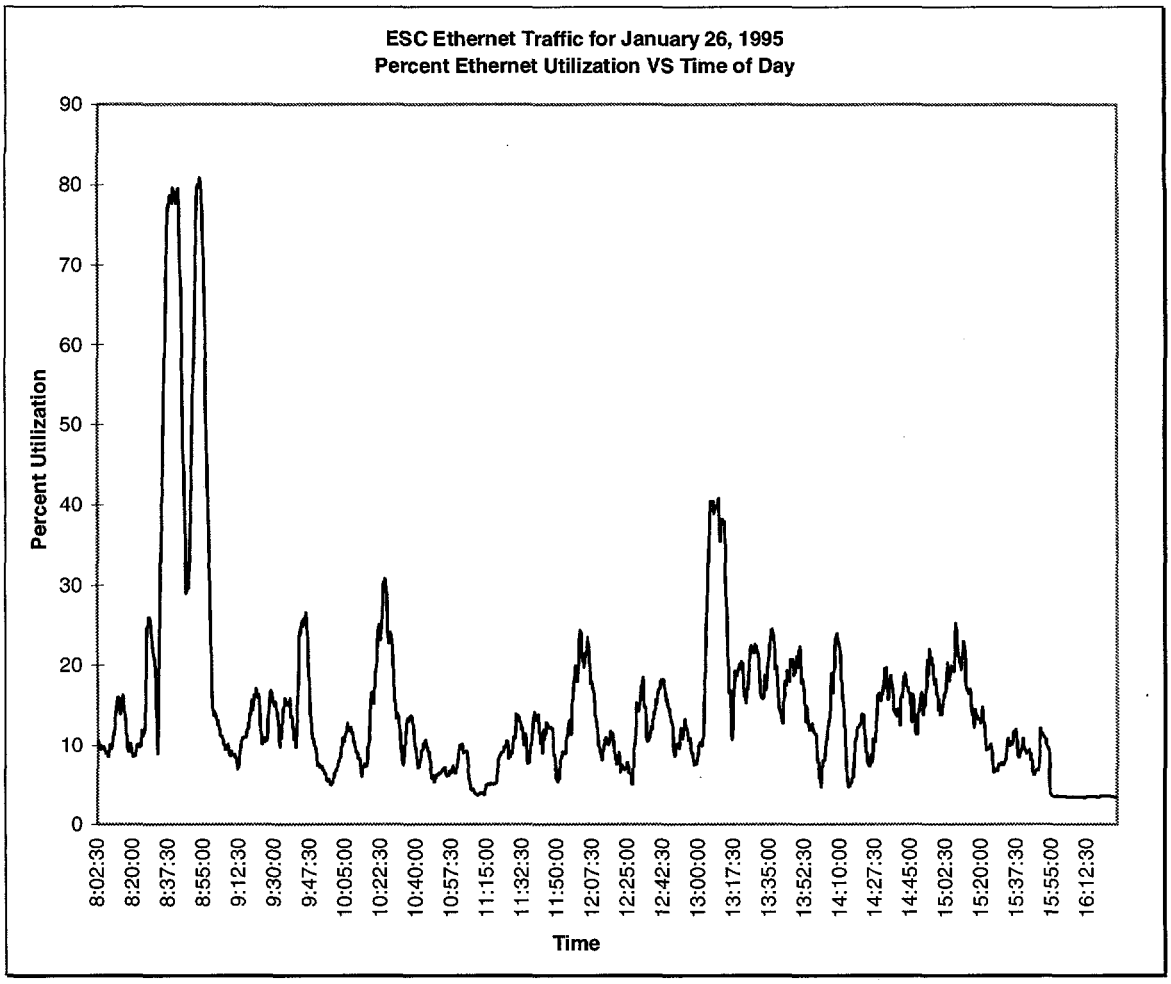
ESC Ethernet Traffic for January 26, 1995
Percent Ethernet Utilization VS Time of Day
Percent Utilization
90
80
70
60
50
40
30
20
10
0
8:02:30
8:20:00
8:37:30
8:55:00
9:12:30
9:30:00
9:47:30
10:05:00
10:22:30
10:40:00
10:57:30
11:15:00
11:32:30
11:50:00
12:07:30
12:25:00
12:42:30
13:00:00
13:17:30
13:35:00
13:52:30
14:10:00
14:27:30
14:45:00
15:02:30
15:20:00
15:37:30
15:55:00
16:12:30
Time

Figure 2

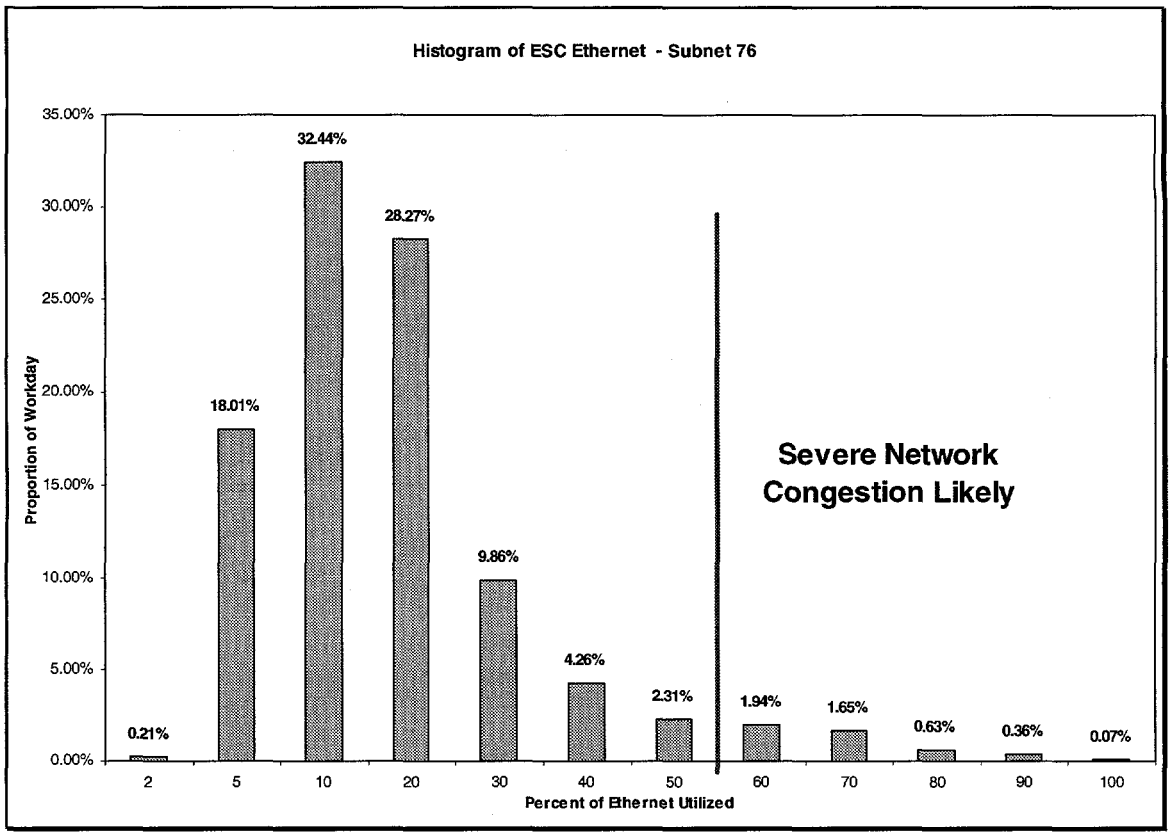
Histogram of ESC Ethernet - Subnet 76
Proportion of Workday
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
0.21%
18.01%
32.44%
28.27%
9.86%
4.26%
2.31%
1.94%
1.65%
0.63%
0.36%
0.07%
Severe Network Congestion Likely
2
5
10
20
30
40
50
60
70
80
90
100
Percent of Ethernet Utilized

Figure 3

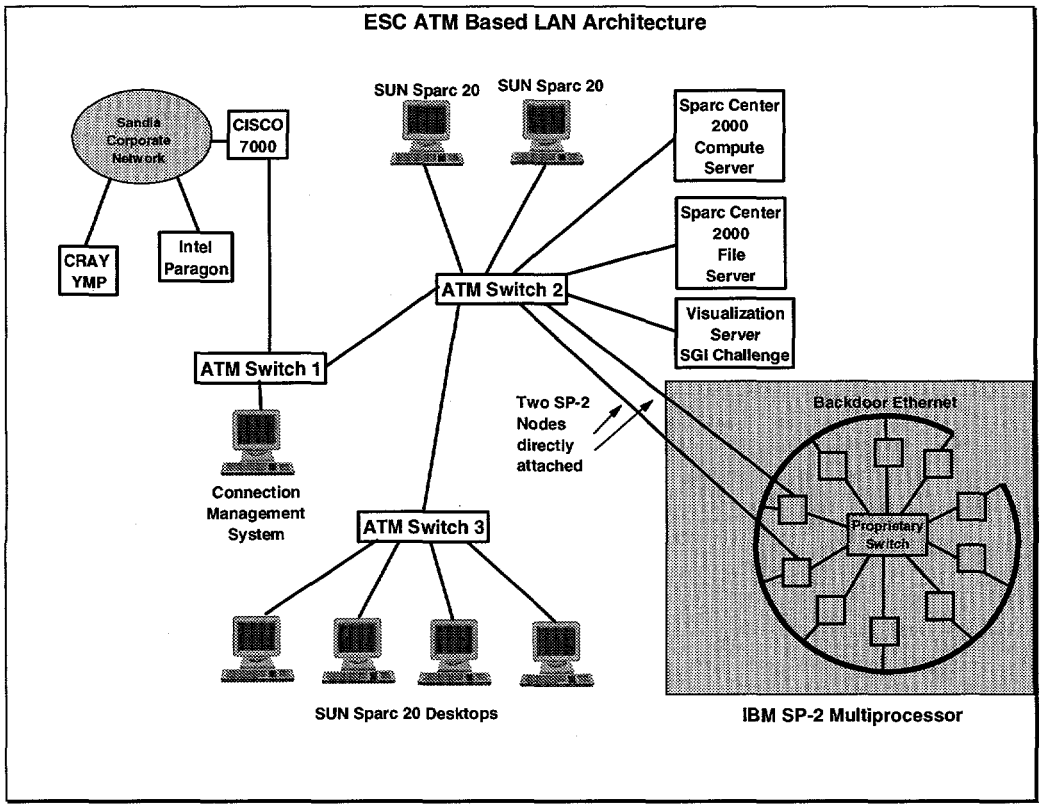

Figure 4

---

[1] Gossage, S. A., "Delivery of Very High Bandwidth with ATM Switches and SONET", SAND92-1295, November 1992.

[2] Naegle, J. H., Testi, N., and Gossage, S. A., "Developing an ATM Network at Sandia National Laboratories", Telecommunications, Vol. 28, No. 2, February 1994, pp. 21-23.

[3] J. H. Naegle, *et. al.*, "Building Networks for the Wide and Local Areas Using ATM Switches and SONET," IEEE J. Select. Areas Commun., vol. 13, no. 4, pp. 662-672.

[4] Sun Microsystems, "NFS: Netowrk File System Protocol Specification," IETF RFC-1094, March 1988

[5] Eldridge, J. "Modeling Data Throughput on Communication Networks", SAND93-0817, November 1993.

[6] W. R. Stevens, *TCP/IP Illustrated, Volume 1: The Protocols*. Reading, MA: Addison-Wesley, 1994

[7] M. Lin, *et. al.*, "Distributed Network Computing over Local ATM Networks," IEEE J. Select. Areas Commun., vol. 13, no. 4, pp. 733-748.

[8] A Romanow, S. Floyd, "Dynamics of TCP traffic over ATM Networks," IEEE J. Select. Areas Commun., vol. 13, no. 4, pp. 633-641