SAND--98-1154

SANDIA REPORT

SAND98–1154
Unlimited Release
Printed May 1998

Sandia's Network for SC '97: Supporting Visualization, Distributed Cluster Computing, and Production Data Networking with a Wide Area High Performance Parallel Asynchronous Transfer Mode (ATM) Network

Thomas J. Pratt, Luis G. Martinez, Michael O. Vahle, Thomas V. Archuleta, Vicki K. Williams

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

SAND98-1154
Unlimited Release
Printed May 1998

Sandia's Network for SC '97: Supporting Visualization, Distributed Cluster Computing, and Production Data Networking With a Wide Area High Performance Parallel Asynchronous Transfer Mode (ATM) Network

Thomas J. Pratt, Luis G. Martinez, and Michael O. Vahle
Advanced Network Integration

Thomas V. Archuleta and Vicki K. Williams
Telecommunications Operations Department I
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0806

Abstract

The advanced networking department at Sandia National Laboratories has used the annual Supercomputing conference sponsored by the IEEE and ACM for the past several years as a forum to demonstrate and focus communication and networking developments. At *SC '97*, Sandia National Laboratories (SNL), Los Alamos National Laboratory (LANL), and Lawrence Livermore National Laboratory (LLNL) combined their *SC '97* activities within a single research booth under the Advance Strategic Computing Initiative (ASCI) banner. For the second year in a row, Sandia provided the network design and coordinated the networking activities within the booth. At *SC '97*, Sandia elected to demonstrate the capability of the Computation Plant, the visualization of scientific data, scalable ATM encryption, and ATM video and telephony capabilities. At *SC '97*, LLNL demonstrated an application, called RIPTIDE, that also required significant networking resources. The RIPTIDE application had computational visualization and steering capabilities. This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations support Sandia's overall strategies in ATM networking.

Contents

DISCLAIMER

1 Introduction

The Advanced Networking Department at Sandia National Laboratories has used the annual Supercomputing conference sponsored by the IEEE and ACM for the past several years as a forum to demonstrate and focus communication and networking developments. The string of participation began in Minneapolis at the 1992 conference with a demonstration of the prototype Switched Multi-megabit Data (SMDS) and Synchronous Optical Network (SONET) technology that Sandia subsequently used in its consolidation of supercomputers [1]. As a direct result of this participation, the National Information Infrastructure Testbed (NIIT) was born. At the 1993 conference in Portland, Sandia focused on the interoperability of emerging ATM technologies and their efficacy in providing high quality video and multimedia capability [3]. This conference resulted in an early pilot of an interconnection of the three DOE Defense Program (DP) National Laboratories over a capable wide area network. The need for this type of network has continued to expand and current efforts in this arena include the DOE Laboratories SecureNet project. At *SC '95*, in San Diego, Sandia demonstrated that broadcast quality video could be passed across a shared ATM network. Also at this show, a high performance OC12 interface for a Paragon supercomputer was demonstrated. This interface later won an R&D100 award and was the prototype for a VIA implementation [8]. This use of an enterprise network is just now beginning to be utilized by the scientific visualization community. In Pittsburgh in 1996, Sandia and ORNL combined to demonstrate that large scale high performance computing could be done on nationally distributed platforms by connecting them together across a national shared communication network [9]. Potentially this model for distributed supercomputing will be adopted by the DOE communities as the way to do production large scale computing.

In all cases, the significant contributions of Sandia's technical partners made the results possible and added significantly to the accomplishments.

Some common themes and benefits at each of the conferences have been:

- partnering with industry to gain early access to new technology;
- focusing current projects and activities by planning and preparing challenging demonstrations;
- engendering new and evolving partnerships with industry, academia, and the other government labs and agencies;
- highlighting the synergy that results from the tight coupling of networking and communication technologies and organizations; and
- providing a stage to professionally interact with colleagues and associates from other organizations in order to challenge and validate our current thinking.

In 1997, the *Supercomputing* conference changed its name to *SC* to reflect the conference leadership's desire to include the entire high performance computing paradigm within the conference's auspices. For *SC '97* in San Jose, Sandia chose to demonstrate:

- the capability of the Computation Plant, a high performance cluster, to demonstrate visualization of scientific data;
- scalable ATM encryption; and
- ATM video and telephony capabilities.

Also within the ASCI booth, Lawrence Livermore National Laboratories demonstrated an application, called RIPTIDE, that required significant networking resources. The RIPTIDE application demonstrated computational visualization and steering capabilities.

This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations support Sandia's overall strategies in ATM networking [4,5,6]. Additionally, it describes the construction of a network to support the DOE's Defense Programs National Laboratories at the conference.

2 *SC '97* Networks

At *SC '97*, for the second year running, the three DOE DP Laboratories built a single integrated research booth. The three Laboratories, Sandia National Laboratories, Los Alamos National Laboratory, and Lawrence Livermore National Laboratory teamed under the ASCI rubric. While all the laboratories contributed equipment and personnel to the effort, Sandia once again led the effort to provide the communication network needed to support the booth demonstrations.

The combined booth was divided into 5 sections. These sections were defined as LLNL, SNL, LANL, ASCI Wall, and the Networking Area see Figure 1. The individual laboratories controlled their particular section of the booth. The three labs shared the ASCI Wall section to show ASCI demonstrations and videos. The Networking Area section was used to house the networking equipment and network monitoring equipment as well as a network available printer.

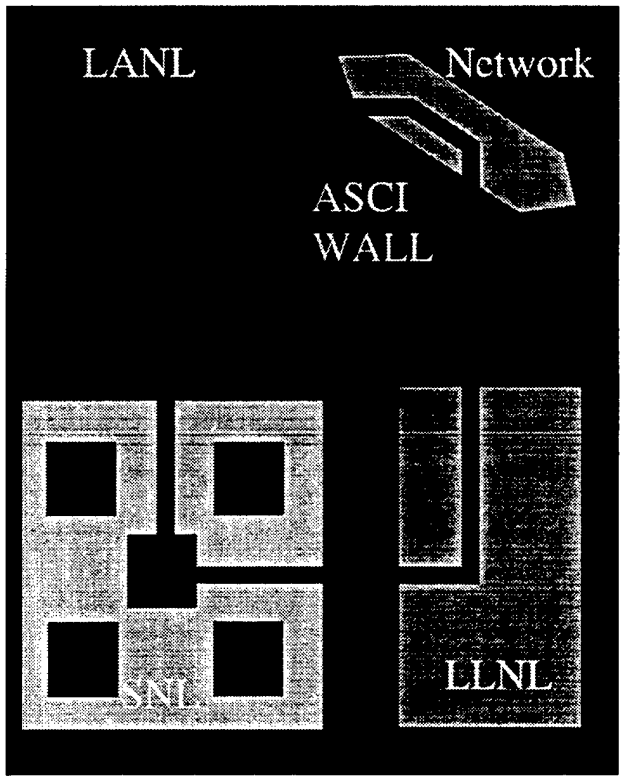

Figure 1 ASCI Booth Layout

3 Network Design

The network for the ASCI booth was designed to meet the needs of the demonstrations and research activities those exhibitors from SNL, LANL and LLNL planned to showcase at *SC '97*. A secondary goal was to demonstrate the current state-of-the-art in production networking through the use of promising new communication technologies. ATM and Ethernet communication technologies were heavily used within the

ASCI booth's network. The Ethernet being utilized included both the switched and shared versions at speeds of 10 and 100 Megabits per second. The ATM technology used in the booth included OC12 (622 Megabits/second), OC3 (155 Megabits/second) and 25 Megabits per second Network Interface Cards and switch ports. The ATM classes of services used to support the booth demonstrations were CBR, VBR, VBR-RT, and UBR. Various policing policies were utilized to ensure a fair sharing of network resources across the competing applications. Some applications required permanent virtual circuits while others use ATM signaling protocols for dynamic circuit setup. The ATM data services CLIP and LANE were used to support data applications that required dynamic connectivity.

Myrinet, a low-latency high-bandwidth system area network switching technology, was used in the Computational Plant demonstration. The Myrinet network connected individual workstations together to build an integrated high performance parallel computing platform.

The design of the network made extensive use of switching elements. Both ATM and Ethernet switches were positioned to minimize congestion within the local networks thereby maximizing throughput both within the LAN and through the external connections. The booth's total switching capacity was greater than 10 gigabits per seconds.

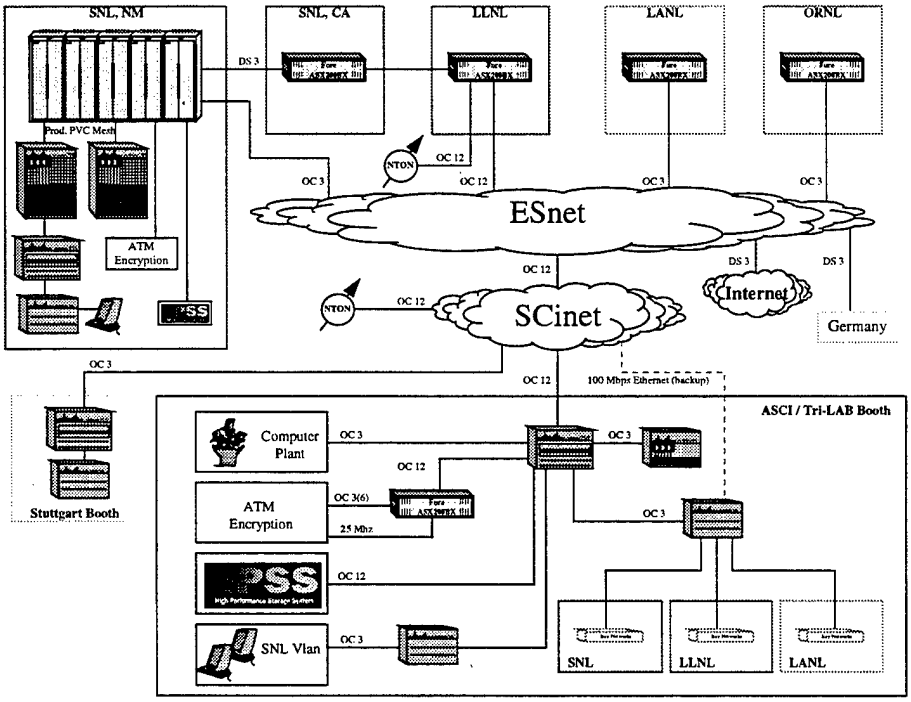

Figure 2 ASCI Booth Network

A router was included in the design to provide the users within the booth isolation from problems that have tended to crop up in the past with the show network (SCINET[1]). The connection to SCINET was an OC12 622 Mbps ATM connection; this connection provided all the networking access outside the ASCI booth. An additional 100 Megabits/second Ethernet connection was provided as a backup in the event that the SCINET's ATM network might break. The ATM connection connected the ATM interfaces within the booth to the show network.

[1] SCINET is the high performance network built each year by volunteers for the supercomputing conference.

The booth's external networks consisted of multiple ATM networks. These networks connected the ASCI booth to each of the participating laboratories' home sites, and to our *SC '97* partners. The ESNET[2] network was used in parallel with Sandia's corporate network to provide high bandwidth to Sandia's New Mexico site. This combined network provided a usable bandwidth greater than 10 megabytes per second. The National Transparent Optical Network (NTON[3]) and ESNET was utilized in parallel to provide an additional 135 megabytes per second of bandwidth to LLNL and Sandia's California site. On the *SC '97* show-floor the ASCI booth network was connected to SCINET via an OC12 ATM circuit. This path provided the ASCI booth with all of its external connections and provided connectivity to the Fore Systems booth and the University of Stuttgart booth on the *SC '97* exhibit floor. A Gigabit Ethernet connection to the NASA Ames Laboratory's booth completed the ASCI booth's external connections.

The network design goal was to create a high-performance production data network that met, through a single data connection to the external resources, all the competing performance requirements of the demonstrations. The design used network switching to limit the possible congestion points in the path to the external data connection. The design needed to be flexible to meet any additional requirements that might appear during the show. The initial physical and logical requirements collected from the Labs are shown in Table 1.

Los Alamos National Laboratories (LANL)	Lawrence Livermore National Laboratory (LLNL)	Sandia National Laboratories (SNL)
6 IP addresses 2 10BaseT Ports	8 IP addresses 6 10BaseT Ports 1 OC3 ATM	12 IP addresses 16 10BaseT Ports 10 100baseT 2 OC12 ATM 6 OC3 ATM
Internet Access Via SCINET	Internet Access Via SCINET Access to NTON Via SCINET	Internet access via SCINET Access to NTON Via SCINET Access to ESNET Via SCINET
2 machines located on the ASCI wall	ATM LAN extension of a LLNL network	ATM LAN extension of a SNL-CA network ATM LAN extension of a SNL-NM network

Table 1: SC '97 Networking Requirements

In addition to the connectivity requirements some of the applications had additional performance requirements. SNL's CPLANT requested that 480 Mbps be provided to SNL in California. The ATM encryption demonstration requested 32 Mbps to be divided between four circuits. These circuits had to be connected from the ASCI booth to ORNL's *SC '97* booth, Fore Systems' *SC '97* booth, SNL's building 880 room C4 in New Mexico, and ORNL's communication laboratory in Tennessee. The MPEG Scientific Visualization demonstration required a 40 Mbps service to SNL in California. While the RIPTIDE demonstration didn't formally request bandwidth, its peak demand was expected to be around 20 Mbps. Through negotiation, the CPLANT requirements were reduced to 400 Mbps.

4 A Parallel Wide Area Network Paradigm

At *SC '97,* Sandia had the opportunity to experiment with parallel wide area networking. The location of the conference in San Jose placed the conference in close proximity to both ESNET and Sandia's corporate

[2] ESNET is the DOE's Energy and Scientific network, operated by LBL, connecting to many national and international sites.

[3] NTON The National Transparent Optical Network, is an all Optical Network, operated by LLNL, connecting to many sites in northern California.

networking resources. To reach both of these wide area networks SCINET employed the NTON, an optical wave division multiplexing research network that has a large communication system that spans much of the greater San Francisco Bay Area. LLNL is a major participant in the NTON network. Prior to *SC '97*, NTON had been extended into the San Jose Convention Center, SJCC, to provide services to other conferences. NTON was also already extended to both the LLNL

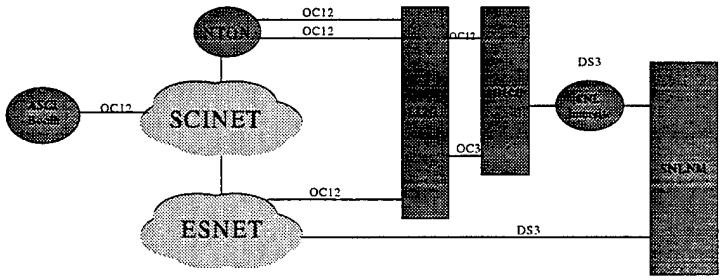

Figure 3 The Parallel Networks

and SNL California campuses. During *SC '97*, the NTON network was connected to ESNET at the Burlingame, California central office of Pacific Bell. The NTON network, the Sandia corporate Intersite network, and ESNET provided the ASCI booth with parallel physical networks to both of SNL's campuses in New Mexico and California. To use this parallel resource, the demonstrations, which were communicating to SNL in New Mexico, were divided between the two parallel circuits to achieve a total of 90 Mbps into New Mexico. The decision as to how to divide the traffic between the two 45 Mbps circuits to New Mexico was made primarily on the Quality of Service (QOS) required for the proper performance of the demonstration.

The demonstrations that required ATM CBR QOS were slotted to run across SNL corporate Intersite network, since ESNET didn't support this QOS. The rest of the traffic was divided in order to limit the use of the production networks to around 50% of their available bandwidth. The end result was that when all the New Mexico demonstrations in the ASCI booth was operating at full utilization, about 10 Mbps of the SNL intersite bandwidth and 20 Mbps of the ESNET bandwidth would be in consumed. Both of these levels were within the acceptable range for the networks to easily support. The partition of *SC '97* demonstrations across the parallel network limited the impact of that *SC '97* had on the production traffic running on both ESNET and Sandia's intersite network. In addition, it was also planned that the parallel network could be employed to off-load the NTON network of all traffic, other than CPLANT, if the CPLANT demonstration was capable of using the entire OC12 bandwidth. It would also have been possible to re-provision the network, if any other performance problems with the OC12 circuit became apparent during the show.

The contingency plan was not needed because the CPLANT application wasn't tuned to use the OC12 circuit with the 1-3 millisecond of networking latency that existed between SJCC and SNL California. However, the ASCI networking team did experiment with switching demonstrations between the two parallel networks to see what effects that switching would cause. The SNL-NM encrypted video demonstration was switched off of its regular data path onto the backup. This caused the video to freeze for about 30 seconds while the virtual cross connects were being made. Then the video restarted looking and feeling the same as before the network was changed. Similarly, the MPEG visualization demonstration was moved off of its original circuit when that demonstration was having difficult running on its initial physical

network. The experimenter thought that the data stream was being interfered with by other traffic from CPLANT running across the same network. After the circuit was changed, the demonstration ran exactly the same as before. This additional observation allowed the researcher to re-evaluate the problem and later find the actual problem. In addition to the parallel WAN experiments, the CPLANT demonstration used multiple OC3 interfaces to do parallel communication.

In these demonstrations, a problem in achieving a generalized parallel IP solution was noted. Local Area Network (LAN) protocols can commonly support many-to-many parallel data paths; however, once into the area of internetworking the path for the data is serial in nature. Therefore, exploiting multiple IP physical paths simultaneously from one LAN to another caused difficulties. Separating the common LAN senders and receivers into different logical LANs allowed physical parallelism to work. Unfortunately, this approach breaks the existing LAN parallelism that works in the local area. The current routing protocols allow for some parallelism to provide load balancing and redundancy across multiple physical paths. In the future, these routing protocols will also provide some inverse multiplexing capabilities. Unfortunately, the use of parallelism features of routers to build parallel data paths between routers doesn't provide an optimum solution to building parallel networks between endpoints. These methods will just lead to routers that keep getting larger and more expensive, while the router will remain a bottleneck to parallel applications. The key to an optimum solution is to use the router to get addressing information, and then to schedule parallel data transfers directly between the communicating machines.

In summary, our experiences with parallel networks suggest:

- they can be used to increase production bandwidth;
- they can help minimize the impact of large applications on other production services;
- they can assist in isolating problems;
- they work in a LAN environment; and
- they have limitations with the currently implemented Internet protocols.

5 Extending the Local Environments

To provide high performance networking service to demonstrators in the ASCI booth several of the demonstrators home LAN environments were extended onto the *SC '97* exhibit floor. The RIPTIDE demonstration required that a LLNL local environment be extended. The CPLANT demonstration required that an SNL California local environment to be extended. These local environment extensions were accomplished using Classical IP over ATM or CLIP. To provide support for a storage technology demonstration and to experiment with ATM LAN emulation a local SNL New Mexico network was also extended. The LAN emulation extension was not successful because of limitations of the ILMI and PNNI protocols to operate in a multiple hierarchical network. This limitation should be addressed in future releases of these protocols.

LAN extension appears at this time to be one of the great benefits of a ubiquitous ATM network. This capability allows the network to make use of multiple physical paths between locations. LAN extension also provides the mechanism to build a geographically dispersed network, while maintaining a common network security architecture. Furthermore, LAN extension can be used to reduce the complexity of running a mobile computer network. Without this ability, the IP network setup becomes difficult to manage and maintain. In the past, mistakes made trying to do this complicated type of networking have led to service outages caused by routing loops.

6 Network Traffic Summary

The test period ran from 12:47 PM MST November 18, 1997 to 5:45 PM November 20, 1998. The data taken consisted of the dumping of the Virtual path and virtual circuit cell counters in the LS1010 switch. Some data was lost during the periods due to resetting of circuits and troubleshooting. The data presented here is a conservative view of the data taken. More data could have been passed but not less. Over the

testing period 1.331 terabytes of data was passed between the ASCI booth network and its remote partners. 639 Gigabytes of data came into the ASCI booth and 692 Gigabytes. The average per second traffic rate for the entire period was 23.64 Mbps incoming and 23.65 Mbps outgoing. The final day, November 20, had the highest average traffic rate of 67 Mbps outgoing and 26 Mbps incoming. The first day, November 18, had the highest incoming data rate, which was 31 Mbps. The peak incoming traffic rate for all of the 15-minute test period was 53 Mbps traffic. The peak outgoing traffic rate for all of the 15-minute test period was 129 Mbps traffic. The graph in Figure 4 shows the profile of this traffic.

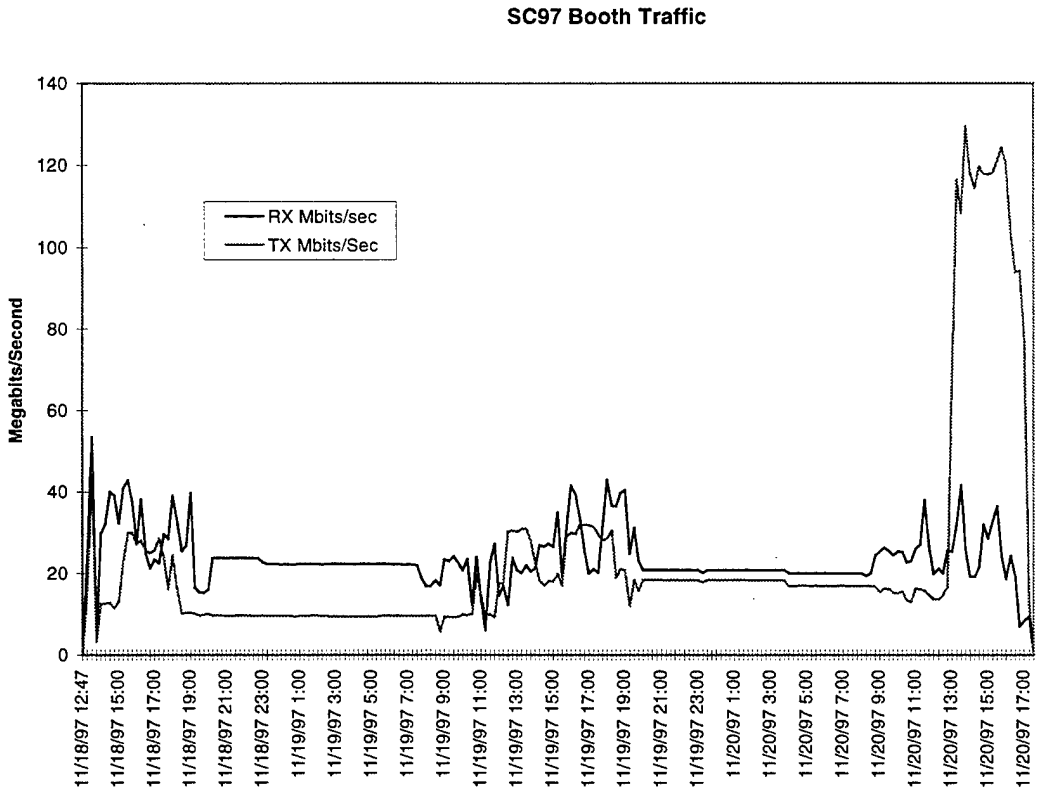

Figure 4 SC '97 Booth Traffic Profile

7 Computational Plant

This demonstration used CPLANT clusters located at SNL in New Mexico and California and at *SC '97* together to compute a solution to a Smooth Particle Hydrodynamics (SPH) code. To enable the demonstration it was necessary to have both ATM and Myrinet communications working together to provide a complete communication system that wasn't distance limited. Prior to *SC '97*, a large effort was undertaken to get the ATM and Myrinet adapter to coexist on the PCI bus of a DEC Miata workstation and to get ATM drivers for the Linux operating system written and functioning. The DEC Miata is the base unit of the FY97 CPLANT cluster. All of this was accomplished just prior to the conference. Once the ATM communications driver was in place, it was demonstrated that a 4 Miata workstation cluster could sustain 400 Mbps of data throughput to a second 4 workstation cluster across an ATM network. This result was obtained when the clusters were locally connected.

The CPLANT demonstration was the third largest user of ASCI booth bandwidth during *SC '97*. During the testing period the demonstration passed 178 Gigabytes. The data was primarily unidirectional; 177 Gigabytes transmitted out of the ASCI booth. The demonstration didn't begin running successfully until the last day of the conference. It was by far the largest user of the network during the last day. The application averaged 89.7 Megabytes per second when it was running. The application was run with the default TCP/IP settings. Therefore, the bandwidth utilization results were limited by the 1-3 millisecond latency of the network.

The logical network to support the CPLANT *SC '97* demonstration consisted of an ATM CLIP network extension of a New Mexico SNL LAN. The LAN was extended to both SJCC and SNL California. The links between SNL California and SJCC came up without any problems. The link to New Mexico exhibited instability during large data transfers. This problem was traced to an interoperability problem between the Cisco LS1010 ATM switches and the ATM adapters in the CPLANT. A team in New Mexico rebuilt the local CPLANT network so that the first switch in the path was a Fore System switch. This solved the problem. The main effort during *SC '97* was to get the application to run. The graphs in Figures 5 and 6 show the data traffic generated by the CPLANT demonstration.

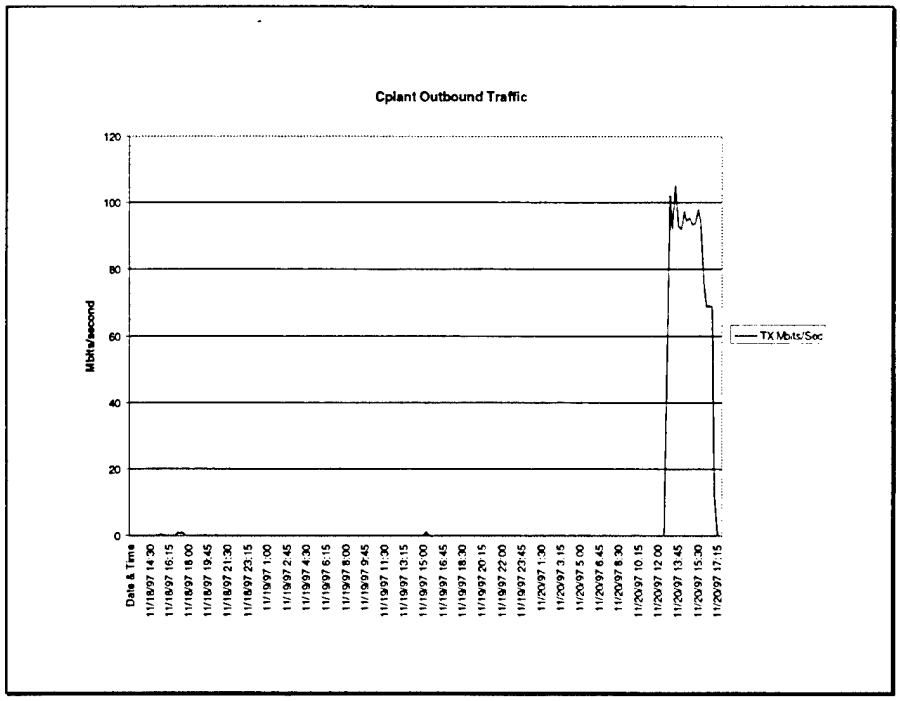

Figure 5 CPLANT's Outbound Traffic Profile

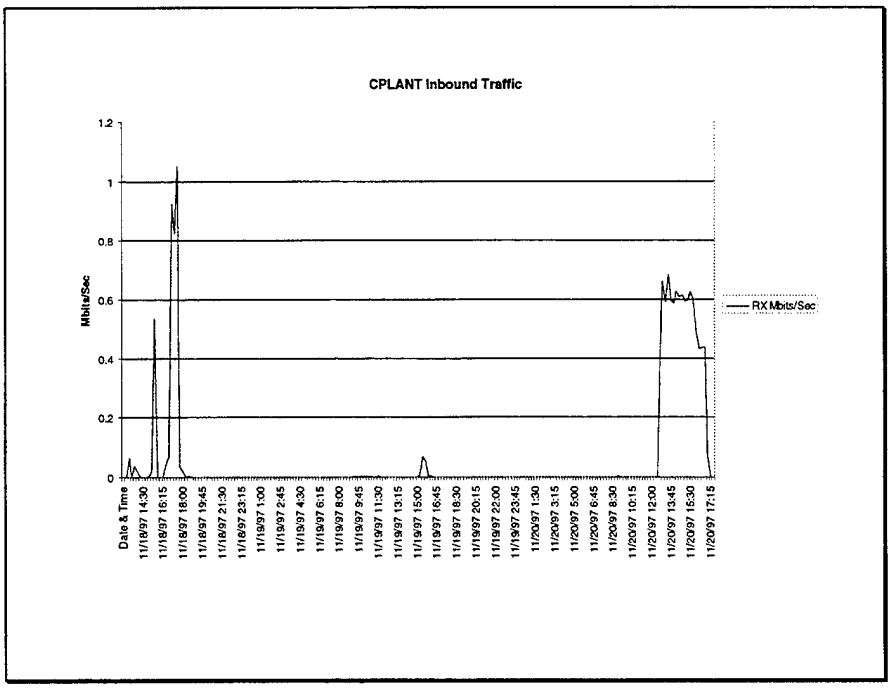

Figure 6 CPLANT's Inbound Traffic Profile

8 ATM Scalable Encryption, ATM Video, and ATM TELEPHONY

This demonstration was the largest single user of the network going into and out of the ASCI booth. Over the monitored period, 395 Gigabytes were transferred. The demonstration sent 296 Gigabytes out of the

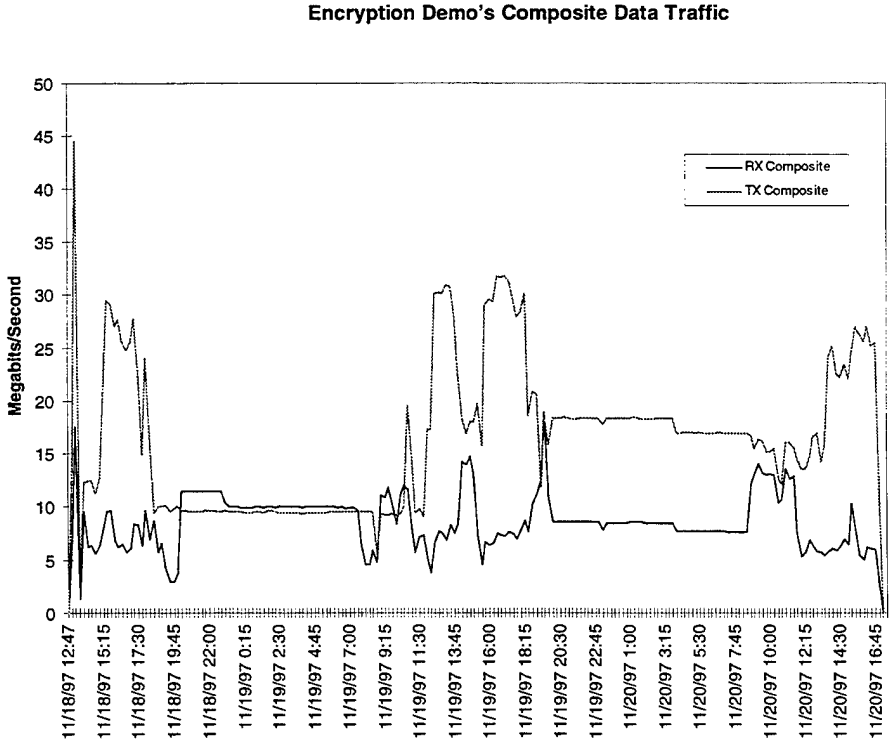

Figure 7 Encryption Traffic Profile

ASCI booth and received 99 Gigabytes from outside of the ASCI booth. The demonstration was a conglomeration of ATM real time applications that included interactive video, telephony quality voice and full speed DES ATM cell encryption. The graph in Figure 7 shows the demonstration network traffic over the testing period.

The video demonstration had *SC '97* exhibit hall partners in the Fore Systems Inc. booth and the ORNL booth. The remote partners for the video demonstration were SNL New Mexico and ORNL. A maximum of four video streams could be simultaneously generated and received. The network utilization of a single stream was limited to a peak bandwidth of 8 MBPS or 32 MBPS total for the combined four streams. The demonstration ran continuously for most of *SC '97*. A video connection server that served as the control platform for the networks was located in the ASCI booth. The quality, full frame rate VHS quality picture, was much better than the more common business video units that exist within SNL corporate video conferencing rooms. The demonstration used Fore Systems AVA and ATV video encoding units to create ATM cells that contain the video and sound signals. These units appeared to solve the audio feedback problem, which earlier ATM videoconference units have had, by using echo cancellation technology. The AVA video encoding unit took in NTSC analog video and audio signals, converted it into a Motion Picture JPEG digital video, and then segmented it into ATM cells and passed the cells into the network. The ATV accepted cells from the network, rebuilt the NTSC analog signal, and passed it to an NTSC receiver. The Fore System's proprietary implementation of the MJPEG audio and video could also be played directly on a Sparc workstation that contained a Fore System ATM NIC card. The video connection server accepts a user's request for a video connection and communicated with the AVA, ATV, and ATM switches to create point to point and multicast video channels.

The ATM telephony portion of the demonstration used Spherical^{TM} a proprietary ATM telephony product from Sphere Inc. The demonstration provided three remote telephone connections to the resident 5ESS telephone switch in New Mexico. There were two trunk lines connecting the 5ESS to Sandia's ATM Telephony Testbed. This provided 12 Plain Old Telephone Service (POTS) connection from the testbed to the 5ESS. The quality of the connection was equal to the voice quality of POTS. The telephony demonstration used network bandwidth only when the phone calls were being made on the demonstration phones. From a data communication perspective, the amount of bandwidth used per call was very small. The system required that an addition number (9) be dialed to access the 5ESS. This is the same manner as when a PBX is used to connect to the public switched network. Typically, the extension of a corporate telephone switch across a shared data network comes at a risk of having the switch resource stolen from within the shared network. In this case we prevented this type of attack by using a DES ATM cell encryptor to protect the corporate resources. The SECANT OC3 DES encryptor accomplished the encryption portion of the demonstration. These encryptors, one in New Mexico and one in the ASCI booths were capable of encrypting multiple ATM streams simultaneously. The Secant encryptor could encrypt and decrypt at the full OC3 line rate.

9 MPEG Visualization Tools Using IP Cut Through Routing

The primary goal of this visualization demonstration was to demonstrate an operable interactive remote visualization environment. The demonstration made use of several leading edge technologies; a high end graphic server, scientific visualization applications, an IP Cut-through switching router, ATM QOS networking, and an MPEG-2 video encoder and decoder. This demonstration was the second largest user of the ASCI booth network. During the testing period, the demonstration transported 295 gigabytes. Most of the visualization was inbound traffic to the ASCI booth with only 413 megabytes of this demonstration's traffic passing out of the ASCI booth. The graph in Figure 8 shows the video demonstration network traffic over the testing period.

This demonstration encountered some interesting networking effects. The output of the demonstration was an interactive video window on an engineering workstation. The demonstration initially ran pretty well with an occasional glitch. However, when the show opened the demonstration fell apart. The demonstration coordinator had requested a VBR QOS for the application that gave it a 20 Mbps circuit. The burst

characteristic of the demonstration wasn't specified. SCINET and SNL California networking personnel set

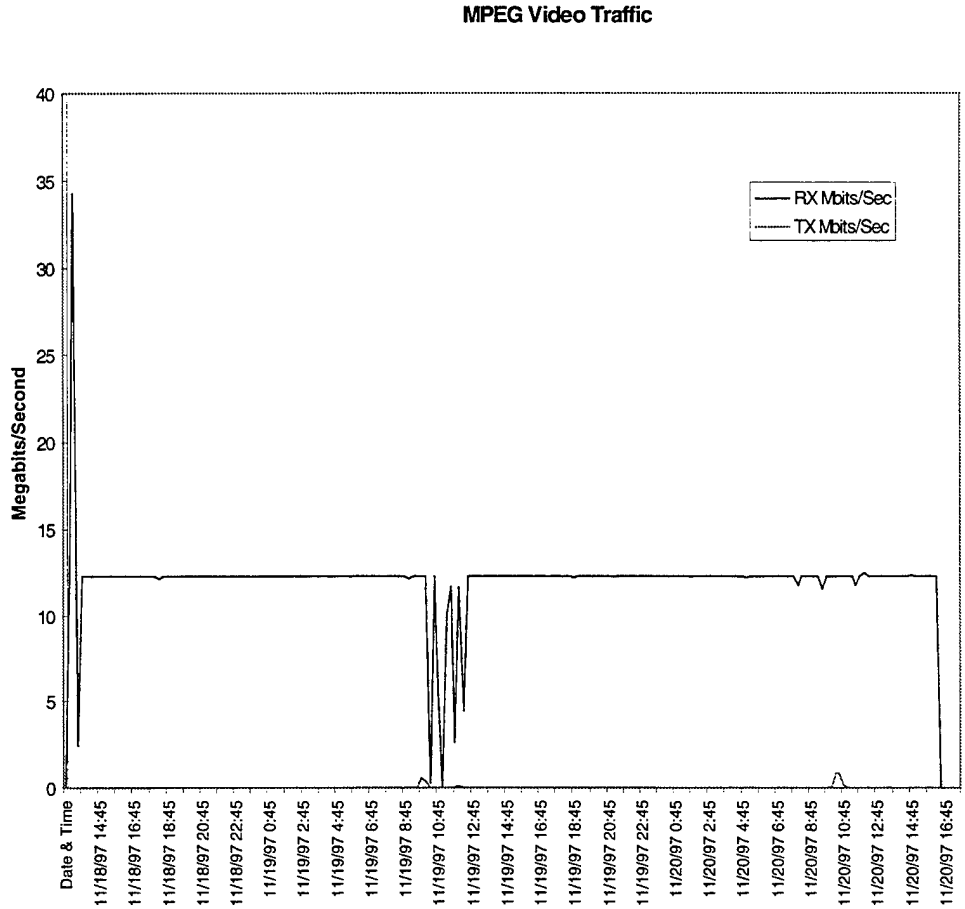

Figure 8 Video Traffic Profile

up the VBR circuit using the default 8 kilobytes burst size. The first attempt to improve the networking performance was to boost the burst size to cover a 16 kilobyte UDP packet size. The demonstration's visualization expert believed that the UDP packet size wouldn't exceed 13 kilobytes. This allowed the video output to be useable, but there were still some glitches. The network was examined to see if the burstiness of the CPLANT demonstration or some other network problem was causing the glitches. The ATM traffic was routed onto a backup virtual path that took a different physical path between SJCC and SNL California. The change didn't result in any improvement, so the application was examined and it was determined that it wasn't VBR in nature. Therefore the network tagged some of the application's data as nonconforming to the VBR traffic contract. Because the switches in use at both SNL California and SCINET were Fore System switches, data was dropped prior to network congestion being seen. Fore System switches are by default configured to drop tagged cells under moderate network load conditions. There were ways to reconfigure the switches to be better at passing tagged cells but that involved design tradeoffs that need to be fully understood in the particular network that was in question. Since that wasn't an option, the QOS for the demonstration was changed to UBR. The demonstration was solid for the remainder of the show. The bottom line was that to boost a UBR application network priority the application has to be well understood. If the traffic profile of the application isn't well behaved, near VBR, and local traffic shaping can't be employed, then don't attempt to upgrade its status.

10 RIPTIDE

RIPTIDE was an interactive visualization demonstration constructed and run by LLNL personnel. During the testing period, the demonstration passed 77 gigabytes. Most of the visualization was inbound traffic to the ASCI booth with only 1.1 gigabytes of this demonstration's traffic passing out of the ASCI booth. The demonstration consisted of the RIPTIDE visualization server, an ONYX2000 at LLNL and the output device, an SGI Octane workstation, in the ASCI booth. A LLNL network was extended to provide the highest level of performance for this demonstration. The demonstration used SGI's GLR software to convert the Infinite Reality engine video output to IP datagrams. The graph in Figure 9 shows the network

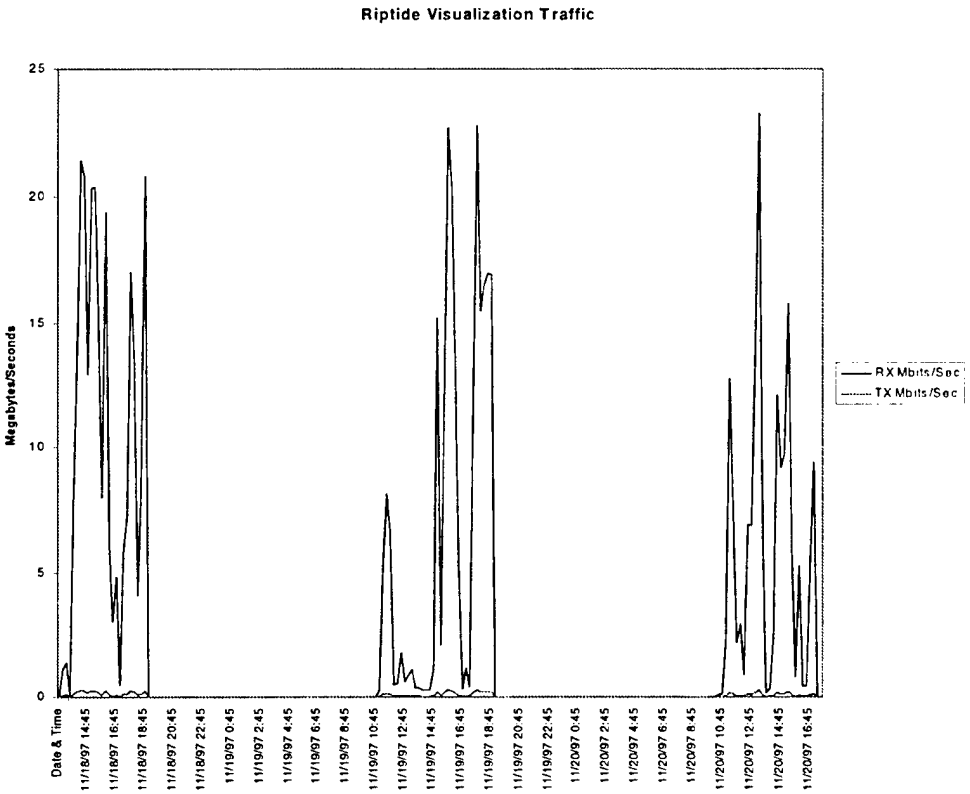

Figure 9 Riptide's Traffic Profile

traffic for this demonstration over the testing period.

11 File transfer and Production traffic

There was a considerable amount of traffic associated with the rest of the activities in the booth. This traffic represents the ASCI booths production networking traffic One of the main sources of the traffic was file transfer between the booth and the demonstrators' home sites. Over the testing period, 3.5 Gigabytes of this type of data was passed between the ASCI booth network and remote sites. 2.9 Gigabytes of data came into the ASCI booth and 704 megabytes went out. The graph in Figure 10 shows the network traffic for the basic networking services over the testing period.

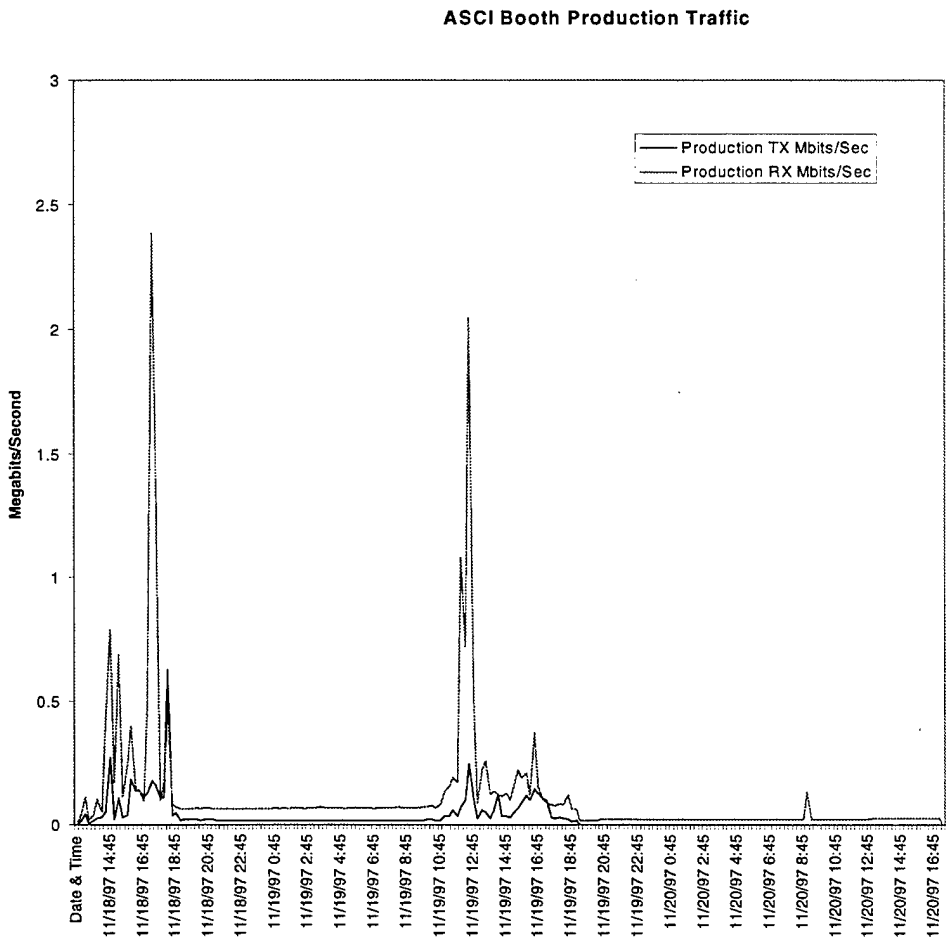

Figure 10 Production Traffic Profile

12 Lessons Learned

The booth's infrastructure planning was inadequate. Early information indicated that a trough system would be used in the construction of the booth. This choice would have greatly simplified the networking cabling installation. Later in the design phase the trough was replaced with a pipe. This piping initially was scheduled to be in place by Saturday afternoon. That turned out to be optimistic, as the piping didn't get put into place until Sunday night at 9:00 PM. This change to overhead piping as the network conduits caused substantial delays in getting the network operational. Since the change to an enclosed pipe wasn't known, pre-made cable bundles weren't available and the paths through the piping weren't planned. By putting the majority of cables into the pipe, which was the last one to be put into place, the most difficult mechanical connection was almost impossible to make. Better cable routing could have been made if we had a prior warning about the change to an enclosed pipe. Because of the delay in getting the piping put up, the networking team had to work late into Sunday night. By noon on Monday all but one application had been made operational.

One demonstration didn't come up during the conference. That demonstration was unable to get served because the partner didn't understand ATM networking. The application required multiple Virtual paths be used. The vender equipment wasn't capable of using any path other than path 0.

13 Conclusion

By all measures the conference proved successful for Sandia. The conference provided a forum for Sandia to feature a wide variety of state-of-the-art networking and communications technologies along with associated applications. The demonstrations showed the importance of networking to the future of hig-performance computing. The introduction of the concept of parallel networking show that more important networking research needs to be accomplished. The CPLANT Cluster demonstration showed that distributed computing is the next major networking challenge. Distributed computing is the killer application for parallel networking. Three different methods of visualizing scientific data were demonstrated. These demonstrations underscore the importance of the visualization to the scientific data in the overall problem-solving environment. The fact that three separate methods were demonstrated also highlights the fact that no single solution solves all the difficulties. The data taken and presented in this paper shows that while the transporting scientific visualization data is not trivial, the traffic for the most part is fairly constant in nature and should be easy to engineer into existing ATM networks. The performance advantage of ATM LAN extension was demonstrated. Using ATM enables applications to bypass traditional network bottlenecks.

These successful demonstrations were the culmination of work accomplished by many people both within and outside of Sandia. In order to meet the challenging goals of the state-of-the-art networks, many teams were formed that crossed corporate and organizational boundaries. The conference also provided an opportunity to identify future goals and plan joint activities. The conference provided a key platform to exchange theoretical networking research information with network engineering reality. The teamwork amplified the accomplishments and achievements of all the participants. Similarly, the conference provided many Sandians an individual opportunity for professional growth, friendly competition, and professional association. Still, on another level, the conference challenged its participants to take stock of their individual projects and to focus them for the demonstration. In all these ways, Sandia benefited from its participation in *SC '97*.

14 Acknowledgments

To put together the activities surrounding the Supercomputing conference takes a large number of talented and dedicated individuals. Without their efforts, Sandia couldn't have accomplished the demonstrations that were done at the conference. We would like to thank the following individuals for their efforts in making Sandia *SC '97* networking efforts a success.

SNL	-	Herb Blair, Jim Brandt, Arthurine Breckenridge, Joseph Brenkosh, Debbie Brunty, R. Michael Cahoon, Cindy Caton, Helen Chen, John Eldridge, Stephanie Fellows, Rich Gay, Steve Gossage, David Greenburg, Rena Haynes, Tan Chang "Richard" Hu, Jeff Jointner, Joseph Maestas, John Naegle, Lyndon Pierson, Cassandra Shaw, Leonard Stans, Tom Tarman, Bruce Whittet, Pete Wyckoff, and Dan Zimmerer
SCINET	-	The Whole SCINET Team.
LANL	-	Alice Chapman, Jerry Delapp, and Steve Tenbrink
LLNL	-	Wayne Buteman, and Mary Zosel
NTON	-	Bill Lennon and R. 'Lee' Thombley
ORNL	-	Arthur "Buddy" Bland, Lawrence MacIntyre, Philip Papadopoulos, Tim Sheehan and Bill Wing
ESNET	-	Mike Collins, Jim Gagliardi, Kevin Oberman and Jim Leighton

15 References

[1] Arthurine Breckenridge and Michael O. Vahle. An account of Sandia's research booth at *Supercomputing '92*: A collaborative effort in high performance computing and networking. Technical Report SAND 93-0224, Sandia National Laboratories, Albuquerque, New Mexico, March 1993.

[2] Steven A. Gossage. Delivery of very high bandwidth with ATM switches and SONET. Technical Report SAND92-1295, Sandia National Laboratories, Albuquerque, New Mexico, October 1992.

[3] Steven A. Gossage and Michael O. Vahle. Sandia's research network for *Supercomputing '93:* A demonstration of advanced technologies for building high performance networks. Technical Report SAND93-3807, Sandia National Laboratories, Albuquerque, New Mexico, December 1993.

[4] Nicholas Testi, John H. Naegle, and Steven A. Gossage. ATM: Sandia user case study. Telecommunications, February 1994.

[5] Nicholas Testi, John H. Naegle, and Steven A. Gossage. Combining ATM and SONET at the LAN. Business Communications Review, pages 47-50, January 1994.

[6] Nicholas Testi, John H. Naegle, Steven A. Gossage, Michael O. Vahle, and Joseph H. Maestas. Building networks for the wide and local areas using asynchronous transfer mode switches and synchronous optical network technology. IEEE Journal on Selected Areas In Communications, 1995.

[7] Michael O. Vahle, Steven A. Gossage, and Joseph P. Brenkosh. Sandia's network for *Supercomputing '94*: Linking the Los Alamos, Lawrence Livermore, and Sandia National Laboratories using Switched Multimegabit Data Service. Technical Report SAND 94-3096, Sandia National Laboratories, Albuquerque, New Mexico, December 1994.

[8] Thomas J. Pratt, Michael O. Vahle, and Steven A. Gossage. Sandia's network for *Supercomputing '95*: Validating the Progress of Asynchronous Transfer Mode (ATM) Switching. Technical Report SAND 96-0820, Sandia National Laboratories, Albuquerque, New Mexico, April 1996.

[9] Thomas J. Pratt, Luis G. Martinez, Michael O. Vahle, and Thomas V. Archuleta. Sandia's network for *Supercomputer '96*: Linking Supercomputers in a Wide Area Asynchronous Transfer Mode (ATM) Network. Technical Report SAND 97-0748, Sandia National Laboratories, Albuquerque, New Mexico, April 1997.

Distribution

Jim Verrelle, Lucent Technologies
Oscar Suarez, U.S. West
0103 R. J. Detry, 12100
0159 L. D. Bertholf, 4500
0318 A. Breckenridge, 9215
0318 G. S. Davidson, 9215
0321 A. L. Hale, 9224
0321 W. J. Camp, 9204
0622 J. F. Jones, Jr., 4600
0630 M. J. Eaton, 4010
0660 W. D. Swartz, 4619
0661 M. H. Pendley, 4612
0741 S. G. Varnado, 6200
0801 M. J. Murphy, 4900
0806 C. D. Brown, 4621
0806 D. C. Jones, 4411
0806 J. H. Naegle, 4616
0806 J. M. Eldridge, 4616
0806 J. P. Brenkosh, 4616
0806 L. B. Dean, 4616
0806 L. F. Tolendino, 4616
0806 L. G. Martinez, 4616
0806 L. G. Pierson, 4616
0806 L. Stans, 4616
0806 M. O. Vahle, 4616 (10)
0806 S. A. Gossage, 4616 (50)
0806 T. C. Hu, 4616
0806 T. D. Tarman, 4616
0806 T. J. Pratt, 4616
0807 B. C. Whittet, 4417
0807 I. C. Alexander, 4417
0807 M. A. Schaefer, 4417
0807 R. M. Cahoon, 4418
0807 S. D. Nelson, 4417
0807 T. V. Archuleta, 4417
0807 V. K. Williams, 4417
0812 F. J. Castelluccio, 4914
0812 G. A. Yonek, 4914
0812 M. R. Sjulin, 4914
0819 A. C. Robinson, 9231
0820 P. Yarrington, 9232
0826 J. D. Zepper, 9111
0828 J. H. Biffle, 9103
0841 P. J. Hommert, 9100
0874 P. J. Robertson, 1716
08812 R. L. Adams, 4914
1002 P. J. Eicker, 9600
1111 S. S. Dosanjh, 9221
1207 R. Moya, 5908
1212 J. M. McGlaun, 5903
1221 J. S. Rottler, 05400
1361 W. F. Mason, 4506
9003 D. L. Crawford, 5200
9011 H. Y. Chen, 8910
9011 J. A. Hutchins, 8910
9011 J. C. Berry, 8930
9011 J. M. Brandt, 8910
9011 P. W. Dean, 8910
9011 R. E. Palmer, 8901
9012 C. L. Yang, 8920
9012 F. T. Bielecki, 8930
9012 J. A. Friesen, 8920
9012 J. E. Costa, 8920
9012 R. D. Gay, 8930
9018 Central Technical Files, 8940-2
0899 Technical Library, 4916 (2)
0619 Review & Approval Desk, 12690
For DOE/OSTI, (2)

Report Number (14) SAND--98-1154

Publ. Date (11) 199805

Sponsor Code (18) DOE/DP, XF

UC Category (19) UC-700, DOE/ER

ph

DOE