SANDIA REPORT

SAND2000-1812
Unlimited Release
Printed July 2000

The ASCI Network for SC '99: A Step on the Path to a 100 Gigabit Per Second Supercomputing Network

Thomas J. Pratt, Thomas D. Tarman, Luis G. Martinez, Marc M. Miller,
Roger L. Adams, Helen Chen, Jim Brandt, and Pete Wyckoff

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
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

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

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2000-1812
Unlimited Release

Printed July 2000

The ASCI Network for SC '99: A Step on the Path to a 100 Gigabit Per Second Supercomputing Network

Thomas J. Pratt, Thomas D. Tarman, Luis G. Martinez, Marc M. Miller
Advanced Network Integration

Roger L. Adams
Telecommunications Operations I

Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185

Helen Chen, Jim Brandt, and Pete Wyckoff
Security & Networking Research

Sandia National Laboratories
Livermore Ca. 94551-0969

Abstract

This document highlights the Discom^2's Distance computing and communication team activities at the 1999 Supercomputing conference in Portland, Oregon. This conference is sponsored by the IEEE and ACM. Sandia, Lawrence Livermore and Los Alamos National laboratories have participated in this conference for eleven years. For the last four years the three laboratories have come together at the conference under the DOE's ASCI, Accelerated Strategic Computing Initiatives rubric. Communication support for the ASCI exhibit is provided by the ASCI DISCOM^2 project. The DISCOM^2 communication team uses this forum to demonstrate and focus communication and networking developments within the community. At *SC 99*, DISCOM built a prototype of the next generation ASCI network, demonstrated remote clustering techniques, demonstrated the capabilities of the emerging Terabit Routers products, demonstrated the latest technologies for delivering visualization data to the scientific users, and demonstrated the latest in encryption methods including IP VPN technologies and ATM encryption research. We also coordinated the other production networking activities within the booth and between our

demonstration partners on the exhibit floor. This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations support Sandia's overall strategies in ASCI networking.

CONTENTS

List of Figures

1 Introduction

SC99 marked the eleventh year for this IEEE high performance computing and communication conference. The ASCI DISCOM project 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 tri-lab networking activities at this conference have their roots in the Supercomputing 91 conference which was held in Albuquerque NM during which the DP laboratories took on the building of an advance network infrastructure to support high end computing. Over the course of the last nine year several networking firsts have been recorded at this conference. At SC93 in Portland. A HIPPI SONET Gateway demonstration, between Portland and Beaverton, Oregon showed the potiential of passing computational data across SONET telecommunication equipment. Also at SC93, ATM technology was introduced with the first national OC3 SONET transport provided by USWEST and by the presenting of the Bell Lab's developed NTT Oc48 ATM Switch. Starting at SC96 in Pittsburgh, Pennsylvania Sandia, Lawrence Livermore and Los Alamos national laboratories have jointly participated in the conference under the ASCI banner. Working together at the conference has further the ASCI vision of tightly coupling the High Performance Computing(HPC) activities within the national labs. This year continued

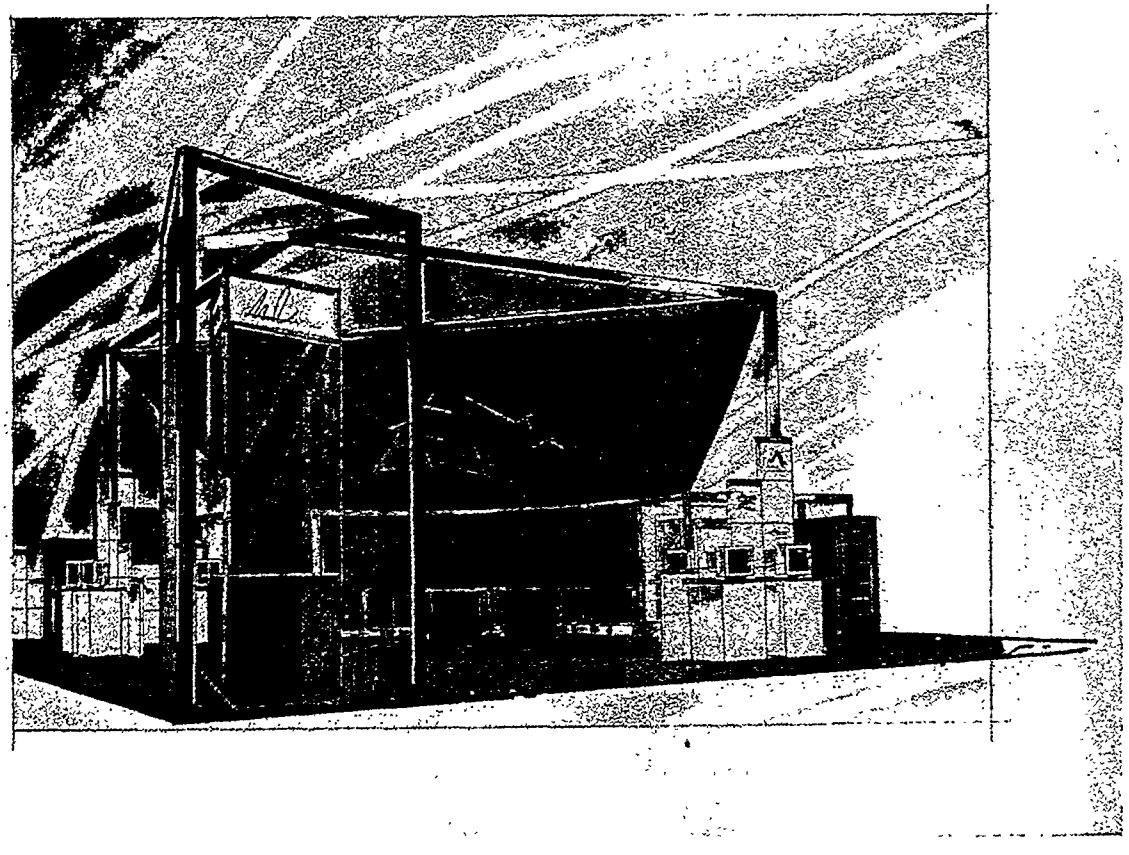

Figure 1: The SC99 ASCI Booth

the cooperative effort with participation from Sandia, Los Alamos, Lawrence Livermore, and Y12 in the planning and support of the booth's production network.

The networking design this year was developed to support the visualization theater concept that was the central theme of the SC99 ASCI booth. This concept closely correlates with ongoing ASCI and DISCOM efforts to support the output side of the scientific simulation and analysis problem. The network contained all of the high-speed commercial networking technology currently commercially available that are of interested to the ASCI community. Packet Over SONET (POS) made it SC debut this year. For the first time Gigabit Ethernet played a central role in the booths network design. The booth's edge networks contained 10/100 Ethernet, OC3 ATM, OC12 ATM, and Gigabit Ethernet interfaces. The network core was built from POS, ATM, and Gigabit Ethernet devices. For each of these different networking technologies, a separate core network was constructed. These core networks were each interconnectioned using IP routing to provide the necessary physical translations. At Sc99, the DISCOM team demonstrated the very latest in routing technologies. These routers can deliver hundreds of gigabit per second of routing capacity today and they appear to be able scale to terabit per second performance. DWDM, was supplied by the XNET, A new structure within SCINET to provide a platform for experimental networking at the SC conference. DWDM wavelengths were used to reach show floor partners with gigabit per second connections. The DWDM service also was used to connect to Sandia's California campus. The ASCI booth's external connections included three Gigabit Ethernet connections, an OC48 POS and an OC48 ATM connection. The aggregate data transport rate for these five services exceeded 7 Gigabit per second. Tripling the SC98 record of 2.1 Gigabit per second. Some other demonstration that are worthy of mention in this report were a Sandia developed OC48 ATM DES encryption hardware, Cluster computing across a WAN, Virtual Private networking, and Remote Scientific Visualization. This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations supports DISCOM's strategies in networking

Many ASCI's technical and commercial partners made the significant contributions that these demonstrations represent possible. The industrial partners at this year show included Lucent, Cisco, Fore Systems/Marconi, Avici, Compaq, SGI,. Netcom and Adtech. The Visualization Theater within the ASCI booth was provided by the University of Minnesota's Laboratory for Computational Science and Engineering, LCSE

Some of the themes and benefits of this conference continue to be:

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

2 *SC 99* Networks

At *SC 99*, for the forth year, the three DOE DP Laboratories, Sandia National Laboratories, Los Alamos National Laboratory, and Lawrence Livermore National Laboratory, put together a single integrated research booth. This year for the first time the design and operation of the production network was a joint effort of all three of the Defense Program Laboratories. Also for the first time a Defense Plant also participated in the networking activities with the addition of networking personnel from the Y12 plant. The joint team of network engineers worked together to provide networking services to the ASCI booth demonstrations, as well as, presenting the DISCOM^2 advance networking demonstration to the attendees of the conference. The advance networking demonstration within the ASCI booth was designed to test and display the latest available networking technology while serving the booth's communication needs. The

Distance Computing and Distributed Computing Program (DisCom^2) provides the network design for the ASCI booth. Discom^2 intention is to deliver key computing and communications technologies that complement the ASCI vision. DisCom^2 implements the technologies to efficiently integrate distributed resources with high-end computing resources both locally and at a distance. Discom^2 uses the SC99 forum to validate new communication technology while increasing the understanding of the high performance networking technologies available to the ASCI communities.

The networking demonstration within the ASCI booth consisted of many large switching components. The core of the booth network consists of three parts, each of which individually constitutes a high performance network. We have designated these individual networks by the technology being used. Gigabit Ethernet, ATM, and Packet Over SONET (POS) are the three different technologies that are at the core of the booth's network. All of the switching elements in the network's core were capable of peak performances ranging from twenty gigabits per second to terabits per second. Each of the core networks contains external connections equal to or greater than one gigabit per second. The total external bandwidth from the booth was greater than 7 gigabits per second. All three of the technology networks were interconnected within the booth. One of the key elements of this networking demonstration was to demonstrate the interoperation of these high performance technologies.

The Network Equipment

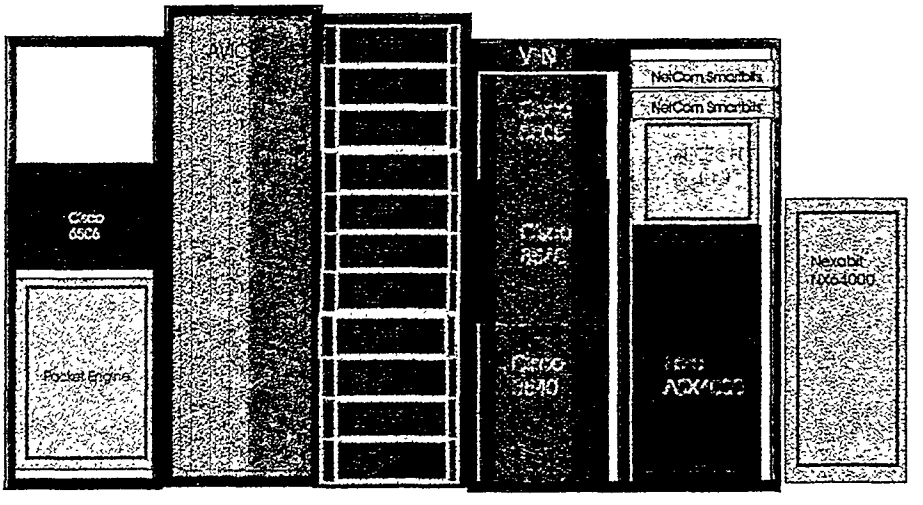

Figure 2: Equipment In Communication Racks

The network demonstration equipment occupied a twelve by five foot space. Sandia provided three production communication cabinets. Avici, Lucent, and Compaq provided the racks for their loaned equipment. Almost all of the network equipment was loaned to ASCI for the exhibition. Over 2 million dollars worth of equipment was used in for this demonstration. Starting on the left side of the exhibit, See Figure 2, is a Cisco Systems 6506 Ethernet Switch. This switch is the core element of the booth's Gigabit Ethernet network. This switch is capable of thirty gigabits per second. This switch currently support only Ethernet interfaces. Below the 6506 is the Packet Engines 5200. This switch is the interconnection point between the Gigabit Ethernet network and the POS network. It also provided connection between the computer cluster and the Video Powerwall in the ASCI booth to the computation clusters in the COMPAQ booth. This connection is provided via the experimental network (XNET). The Packet Engine switch is

capable of fifty-two gigabits per second. It currently supports Ethernet and POS interfaces. In the next cabinet is the AVICI Terabit Switch Router (TSR). This equipment was the core for the POS network. It was also the external entry point for SC99's XNET. The TSR appears to be capable of scaling to thousands of gigabit interfaces. Currently the TSR support OC48 and OC12 POS interfaces. In the next cabinet was an ALPHA cluster, the cluster consisted of ten DP10 workstations interconnected to the booth network by twenty, hundred megabit per second Ethernet interfaces. This cluster was also connected to clusters in Sandia National Laboratories in Livermore Ca. and to computation cluster in the Compaq booth network. The next cabinet contains three network switches. The two switches at the bottom of the rack are Cisco's 8540s These switches are the choice to be the equipment that will pass data from the large computing platforms at Lawrence Livermore National Laboratory and Los Alamos Laboratory to the DISCOM Wide Area Network. The 8540s support Gigabit Ethernet and ATM Interfaces. They are rated at twenty gigabit per second switches with an additional ten gigabit per second of redundant backplane. The last switch in this rack is a Cisco 5505. This switch was used to support network separation for the Virtual Private Network (VPN) running between the booth and Sandia National Laboratories in New Mexico. The 5505 support 10/100baseT Ethernet, OC3, and OC12 ATM interfaces. This switch is rated at three gigabits per second. The equipment at the top of the rack was the Network Alchemy's CryptoCluster 2500. This box implements the IPSEC Triple DES encryption and firewalling functions required by a VPN. The VPN provided connectivity to support the ATM remote visualization demo in the booth. The next cabinet contains a Fore System ASX4000 ATM switch. This switch is the core element in the booth's ATM network. The ASX4000 support OC3, OC12, and OC48 ATM interfaces. It is rated at 40 gigabits per second. The ASX4000 is Sandia 's DISCOM FY00 wide area network point of presence(POP). Above the ASX4000 was an Adtech AX4000 Analyzer/Generator. This testset supports POS, ATM and Gigabit Ethernet testing. Line interface at OC48, GigE, Oc12, and Oc3 are supported. Above the Adtech were two Netcom Smartbits testers. These testers support GigE, ATM, POS, and 10/100 Ethernet testing. The Smartbits and Ax4000 roles in the booth were to test the performance capabilities of the ASCI network equipment. The cabinet on the far right of the networking area contained a Lucent NX64000 Multi-Terabit Switch/Router This router currently supports OC192 POS, OC48c POS, OC12c POS, OC3c POS, OC12c ATM, OC3c ATM, and DS3 frame interfaces with line-speed QOS features. The NX64000 provided the booth with POS to ATM interoperability. The NX64000 switching fabric is rated at 6.4 terabits per second. The largest consumer of network bandwidth within the ASCI booth was the distributed visualization demos running on the powerwall. Add to this booth's network, computation resources capable of tens of teraOPs, storage capacity growing to pitabytes, and hundreds of scientific and engineering users spread through the United States and you have the a model of the ASCI network for the year 2000.

3 Network Design

The ASCI booth layout is shown below in figure 3. The ASCI booth at SC99 consisted of a video theatre(A), the network demonstration area (B) two staffed demonstration areas(C&D), the platform kiosk(E) and a web-based interactive station(F). Most of the networking connections were in areas A and B; these two areas were on a raised floor. CAT 5 cables and Multimode fiber connections were provided as need to the Clusters and workstations located in these areas. There were no network connection to platform kiosk Area E. Cables were run under carpeting from the network area to all of the other areas to support connectivity. To support the staffed demo areas' Ethernet needs a Cisco 2900 Ethernet switch was placed into each area. The switch was connected to the core Ethernet network by a GigE SX connection. In Demo Area C in addition to the multimode GigE SX connection, additional multimode and single mode fibers were run to provide physical infrastructure for the ATM circuits that the demos in the area were using and for the analog video needs of the Lightpipe equipment. Two CAT 5 cables were deployed to support the separations requested by the VPN team. Several additional fibers were added to the bundle to provide for growth and redundancy in that area. The network infrastructure to the demonstration area D was just three pairs of multimode fiber. One of these was used to connect to the GigE SX port in a Cisco 2900 Ethernet switch. The others were reserved to provided for growth and redundancy. Three CAT5

cables supported area F. The web-based interactive application platform used one while the other two were made available to the network monitoring effort. We labeled all cables that were run to the demo areas and documented them. Infrastructure labeling and documentation was a lesson learned from past SC shows and proved very helpful this year. We did run into problem managing access to the console ports of all the equipment in the networking area. The access to the front of the networking equipment was limited this year. The front of network area was enclosed with Plexiglas. To access the front panels required the removal of the panels. Glass suction cups pads where used to handle the large panels.

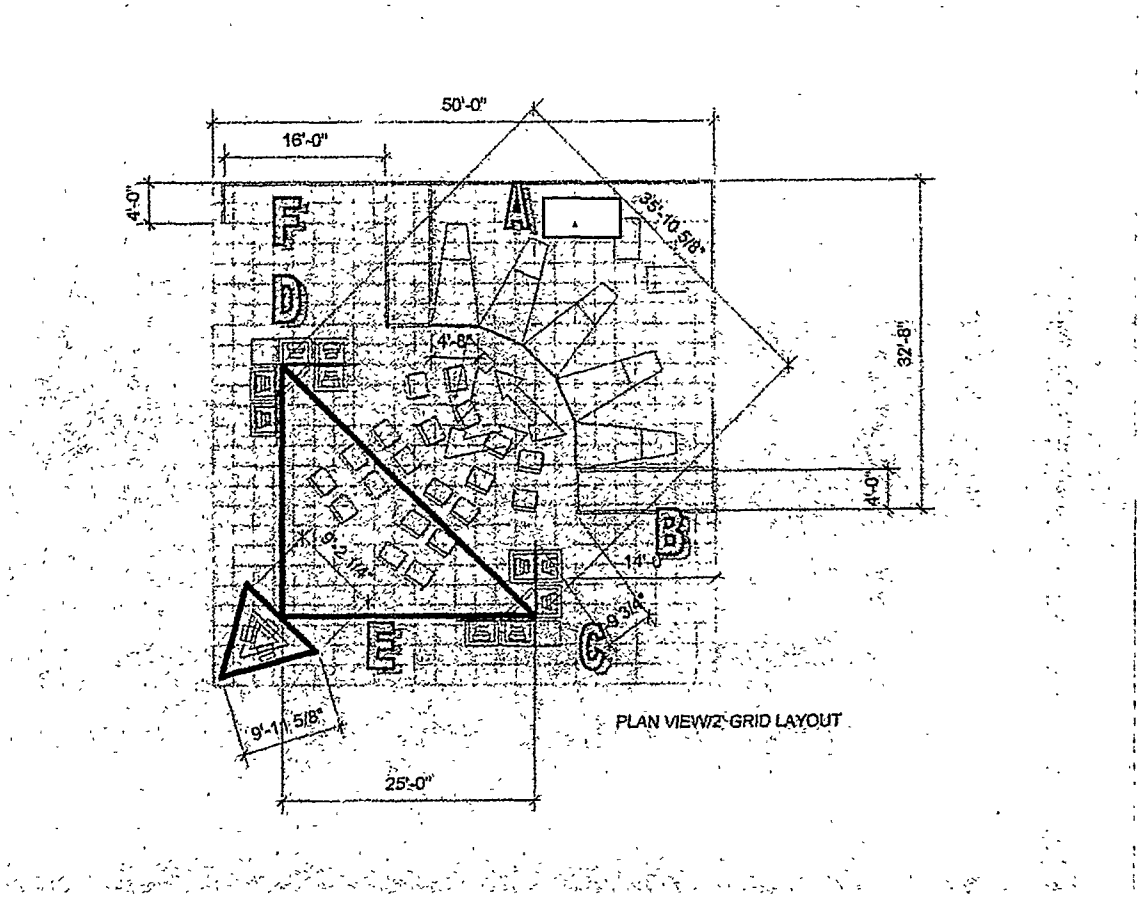

Figure 3: The Booth's Layout (Area B was where the core network equipment was located)

The SC99 booth had several external connections. This year, as in years passed, the network was segmented to provide a level of production network capacity that was separated from the research activities in the booth and in the XNET/SCINET network. This year a gigabit Ethernet connection to SCINET was considered to be the primary production network connection. An OC48 ATM network connection to SCinet also share in this distinction by providing production native ATM services. An OC48 connection from NTON via XNET was brought into the booth to provide connection to the remote networking research demonstrations. An additional OC48 POS connection was also brought to the booth to provide the SCINET XNET connection to the XNET POS router capacity. These service were supplied to the booth via XNET's DWDM system. The XNET DWDM system provided sixteen optical wavelengths to the show floor. In addition to the wavelengths that supported the POS traffic, two more of the

wavelengths were used to reach across the SC showfloor to the Compaq booth. We terminated these two wavelengths into GigE optical sources and sinks. Figure 4 shows the network connectivity.

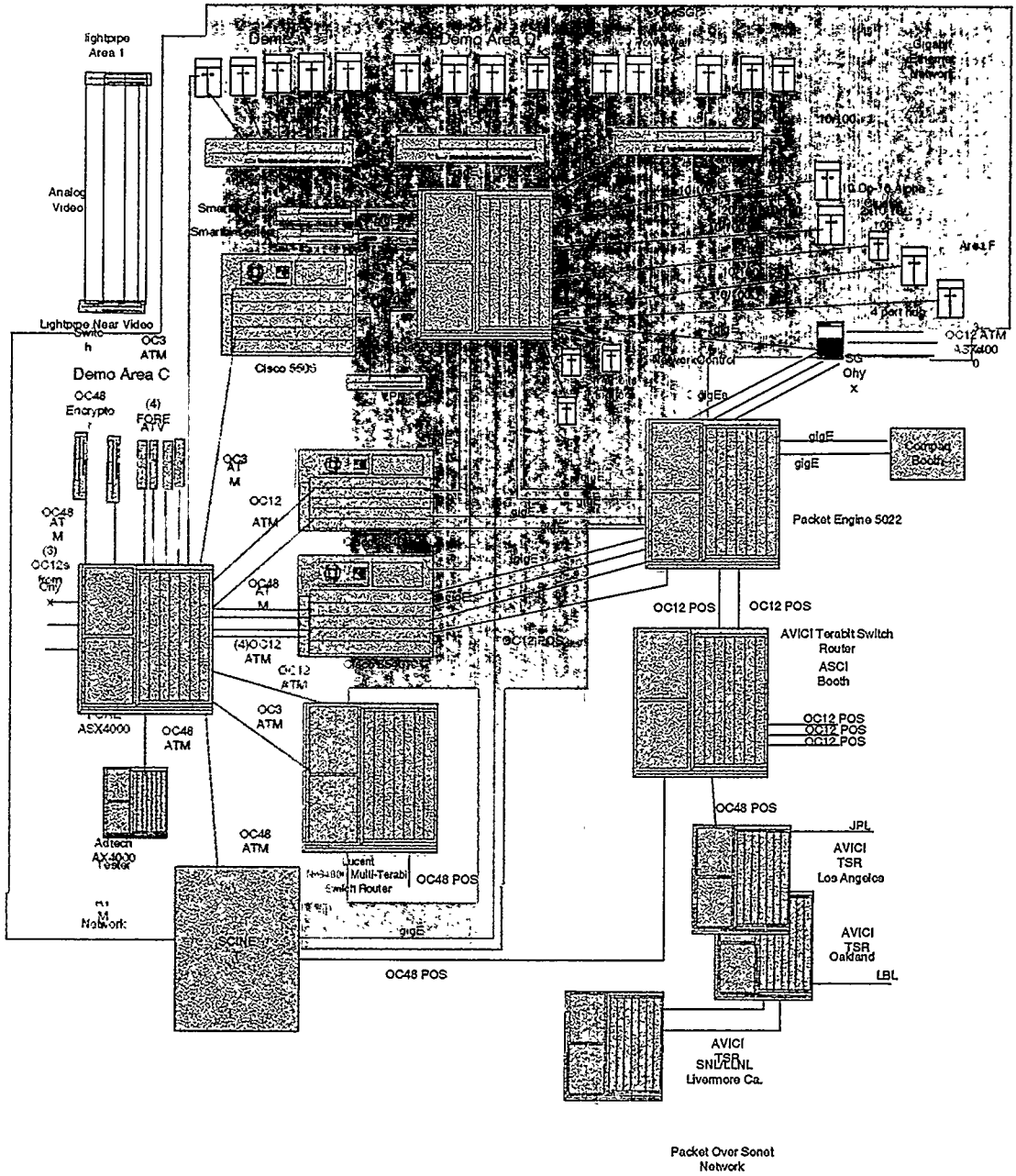

Figure 4: The SC99 ASCI network

4 ASCI WAN at SC99

Leveraging on the advances made by the National Transparent Optical Network (NTON) [1] dense wave division multiplexing (DWDM) [2] infrastructure, we prototyped a leading edge wide area network (WAN) to meet the bandwidth demands of large-scale ASCI applications. This testbed used state-of-the-

art packet-over-SONET (POS) [3] links of the Avici Terabit Switch Routers (TSR) to connect Sandia/CA, LBNL, JPL, and the ASCI booth. Local area network (LAN) resources were accessed via Gigabit Ethernet, Fast Ethernet, as well as ATM switches. To demonstrate the performance benefit of this testbed, the Image Based Rendering Assisted Volume Rendering (IBRAVR) [4] application developed at LBNL was run. This application performed parallel 2D-image compositing on Sandia-CA Cplant, using data stored on the Distributed Parallel Storage System (DPSS) [5] at LBNL. Resulting 2D-images were then transmitted to a SGI graphics engine and assembled into 3D-images for display at the Portland ASCI booth.

Figure 5 depicts the NTON POS testbed we constructed for the Super Computing 99 Conference at Portland, OR. As shown, its backbone consisted of four TSR's [6] from Avici Systems, connecting Sandia-CA, the GST Oakland and Los Angeles PoP's, and the Portland Show floor. The Avici TSR was selected because it is designed to scale along both the bit-rate and the port-count axes. To achieve this goal, the TSR switch fabric adapts a direct connection network architecture, where each network line card is also a switching node. Line cards are connected to form a dual 3D torus, using twelve 20 Gbps fabric links. This architecture allows incremental growth to achieve up to 364 Tbps of aggregate backplane bandwidth in order to meet the demands of the exponential growth in the Internet.

While some of the backbone links still relied on the SONET overhead to ensure high network availability, the connection between LA and Portland adopted Nortel Network's Optera system [7] to implement an all-optical network. This DWDM-based system included a maintenance channel within each of its wavelengths in order to provide end-to-end visibility of signals being transported, thereby ensuring high availability transports. Additionally, Optera is bit-rate- and protocol-independent, and is, therefore, a passive device.

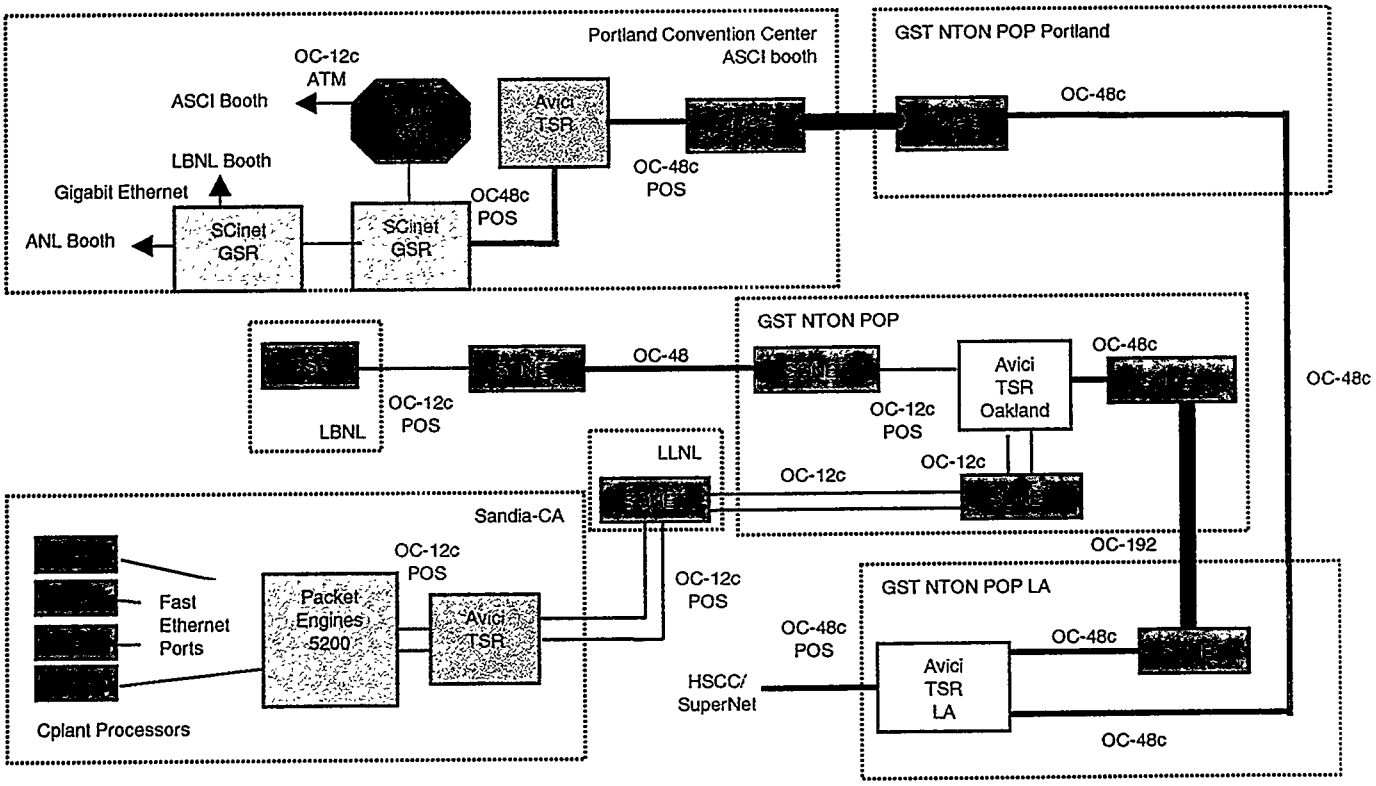

Figure 5 The NTON POS Testbed

As such, the Optera system had to acquire clocking signals from both the LA and the Portland TSR's in order to provide SONET synchronization; it compensated for clock skew between the two TSR's using

internal buffers. The bandwidth provisions on the POS testbed were one OC48c (2.5 Gbps) link each for the Oakland to LA and the LA to Portland connection, two OC12c (2 x 622 Mbps) links to Sandia-CA, one OC12c link to LBNL, and one OC48c link to the Portland conference network.

The IP representation of the NTON POS testbed is illustrated in Figure 2. As shown, the Avici TSR's used standard Interior Border Gateway Protocol (IBGP) [8] to exchange routing information with each

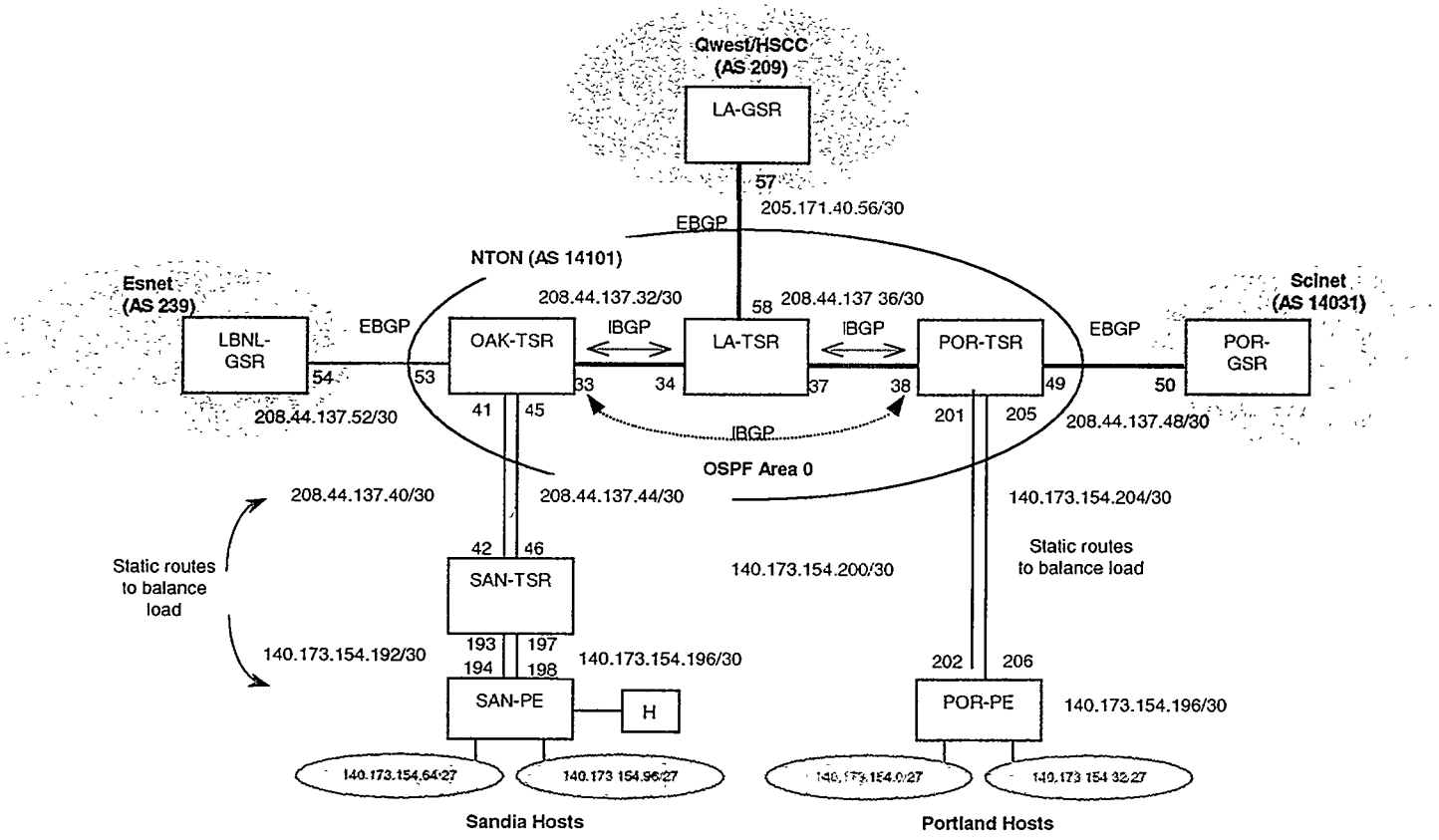

Figure 6 IP representation of the NTON POS testbed

other and Exterior Border Gateway Protocol (EBGP) [8] to exchange routing information with the Cisco GSR at the conference and at LBNL. Static routing was employed to balance traffic between the OC12c links to compute nodes at Sandia-CA and the compute as well as graphics servers at the SC99 booths.

At SNL-CA, 32 Alpha processors from our Computational Plant, Cplant [9], were allocated for this demonstration. These nodes were connected using a Packet Engine PowerRail 5200 switch/router with Fast Ethernet interfaces. These nodes accessed the POS testbed via two OC12c POS interfaces configured on the same switch/router. The PowerRail 5200 has 52 Gbps aggregate bandwidth on its backplane and can support up to 72 Gigabit and 240 Fast Ethernet ports. When fully populated, the PowerRail 5200 will experience blocking in its switching fabric. The PowerRail 5200 was used because it was the only switch/router that offered OC12c POS and Gigabit Ethernet connectivity at the time.
At LBNL, we used the Distributed Parallel Storage System data block server to provide high-performance data handling for our demonstration application. The DPSS architecture builds high-performance storage systems from low-cost, commodity hardware components. This technology has been quite successful in providing an economical, high-performance, widely distributed, and highly scaleable architecture for caching large amounts of data that can potentially be used by many different users. LBNL connected the

DPSS data block servers directly to its Cisco GSR using Gigabit Ethernet links. The Cisco GSR was connected to the NTON POS testbed via an OC12c POS port.

Over this testbed, a high-resolution, interactive remote visualization technique developed at LBNL under the NGI Combustion Corridor project was demonstrated. From eight of the Cplant nodes at SNL-CA, the IBRAVR application loaded raw volume data of a combustion calculation from the DPSS at LBNL. This operation used the Message Passing Interface (MPI) library to achieve high I/O rates through parallelism. Upon load completion of one time-step of data, the eight Cplant nodes performed 2D-image compositing in parallel, each with a small subset of the overall data volume to achieve scalability. Resulting 2D images were then transmitted, using parallel TCP/IP sockets, over the NTON testbed to a rendering engine at the ASCI or the LBNL booth. The rendering engines in turn assembled the received 2D images into 3D representations for display either on the ASCI Powerwall or the ImmerseDesk at the LBNL booth. The ASCI Power Wall had an OC12c ATM interface running Classical IP over ATM. It accessed the NTON testbed through two SCinet switch/routers, a Fore ASX 4000 and a Cisco GSR. The ImmerseDesk at the LBNL booth had a Gigabit Ethernet connection. It also traversed two SCinet switch/routers to reach the NTON testbed, a Cisco Catalyst 5500 and a Cisco GSR.

The demonstration was designed to deliver the 2D images from the Sandia-CA Cplant nodes either to the ASCII booth Power Wall or the LBNL booth ImmerseDesk. Because only a one half-hour per day of

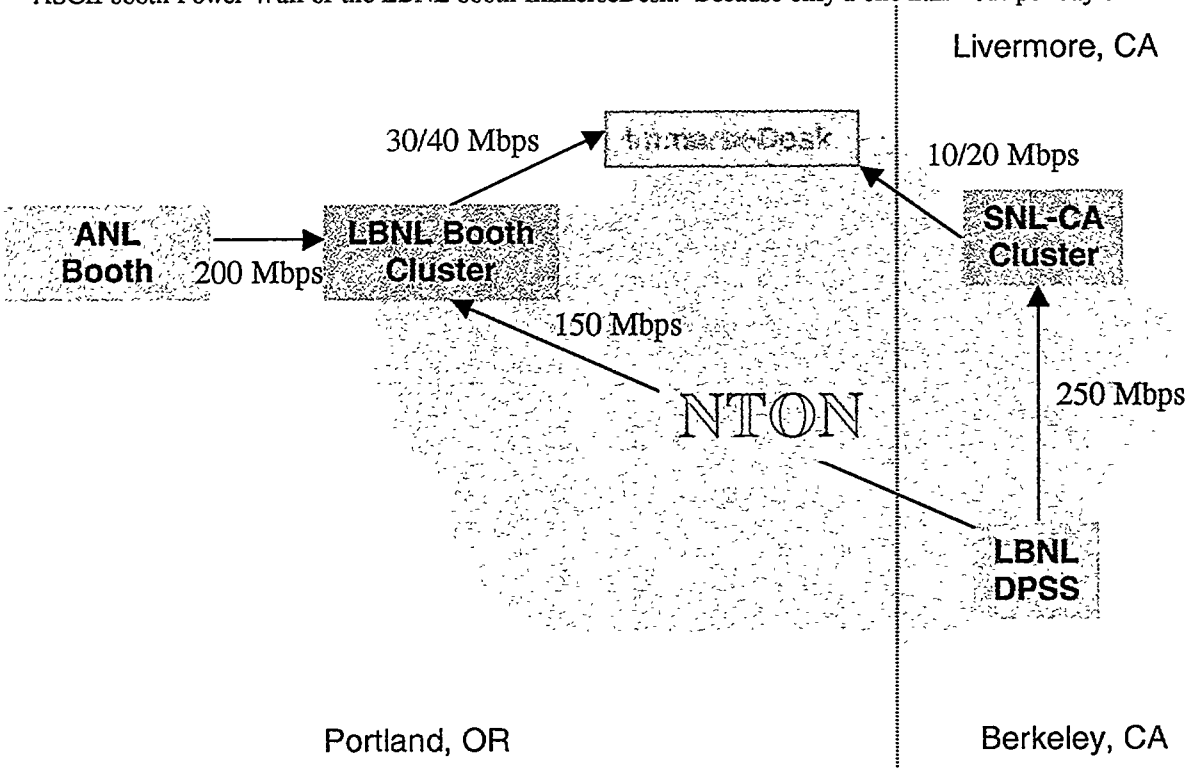

Figure 7 IBRVR throughput statistics

access to the ASCII Power Wall was given to this demonstration, all measurement data was collected during visualization at the LBNL booth where the demo was running most of the time. Additional measurements were made using local compute and/or local storage with respect to the graphics server at the LBNL booth. These measurements contrasted performance characteristics of local vs. remote data retrieval and visualization.

First, the performance of remote vs. local DPSS access was examined. As shown, the throughput of LBNL-DPSS to SNL-CA-Cluster was 250 Mbps, which is less than half of the path's bottleneck bandwidth (622 Mbps) at the LBNL-to- NTON hop (Figure 1). Since we have tuned the TCP's window to cover the link's bandwidth-delay product, the performance bottlenecked is elsewhere and will be the topic for future study. The ANL booth-DPSS and the LBNL- booth-Cluster were local to each other; they were connected via a Cisco Catalyst 5500. The lower throughput between them, 200 Mbps, is believed to be network bandwidth limited; the Catalyst 5500 is a slower device and it had to handle competing SCinet traffic. The LBNL-DPSS to LBNL-booth-Cluster experienced the worst performance of the three, 150 Mbps. Again, the bottleneck was thought to be in the SCinet network. The network traversed two SCinet routers, a Cisco GSR and the same Catalyst 5500 (Figure 1) each with competing traffic from other booths.

Next the throughput performance between the LNBL-ImmerseDesk and the back-end parallel processors locally at the LBNL-booth and remotely at SNL-CA was looked at. Figure 3 shows that the aggregate throughput of transferring the eight 2D images was 10-20 Mbps from the SNL-CA Cluster and 30-40 Mbps from the LBNL-booth Cluster. The communication here involved the transfers of small amounts of texture data (of the 2D images) interleaved with control messages needed to synchronize distributed cooperating components. Therefore, its performance is delay sensitive, and thus the low bandwidth especially in the WAN case. A new release of the IBVAVR application will remove all of the control messages between the viewer and the back-end processing elements, thereby allowing the TCP window to open up and the throughput to increase.

The DOE ASCI project is chartered with the simulation of science and engineering problems of unprecedented size (100's of terabytes) and complexity, often requiring the use of human and compute resources that are in geographically distributed locations. As such, very high bandwidth networks as well as innovative remote visualization software tools are essential to ensure the success of the ASCI program. This prototype testbed explores leading network technologies available to the ASCI communities. the LBNL IBRAVR application was run to validate these technologies because it demands high-speed access of data, compute, and visualization resources from distributed locations.

5 Compressed Remote Visualization Consoles at SC99

The Supercomputing '99 Remote Visualization Console exhibit demonstrated the ability to remotely access the console displays on visualization servers located at Sandia/NM, Sandia/CA, Lawrence Livermore, and Los Alamos. This system used video conversion compression gear located at the servers' sites and at the client site (SC '99 exhibit booth), with the various sites connected via the ESnet wide-area ATM network. Access to these servers from the clients was performed via traditional ssh, and user input on the client (i.e., keyboard and mouse) was redirected to the servers via Irix's Network Dual-headed Software Daemon (ndsd) command.

The system configuration for the remote visualization console demonstration is shown in Figure 8. This demonstration made use of the 4-way quadrant technique to re-direct the output of each visualization console to the SC '99 show floor. An SGI O2 provided client-side input, and output was switched between a flat panel display in the booth demonstration area, or to a powerwall in the presentation area.

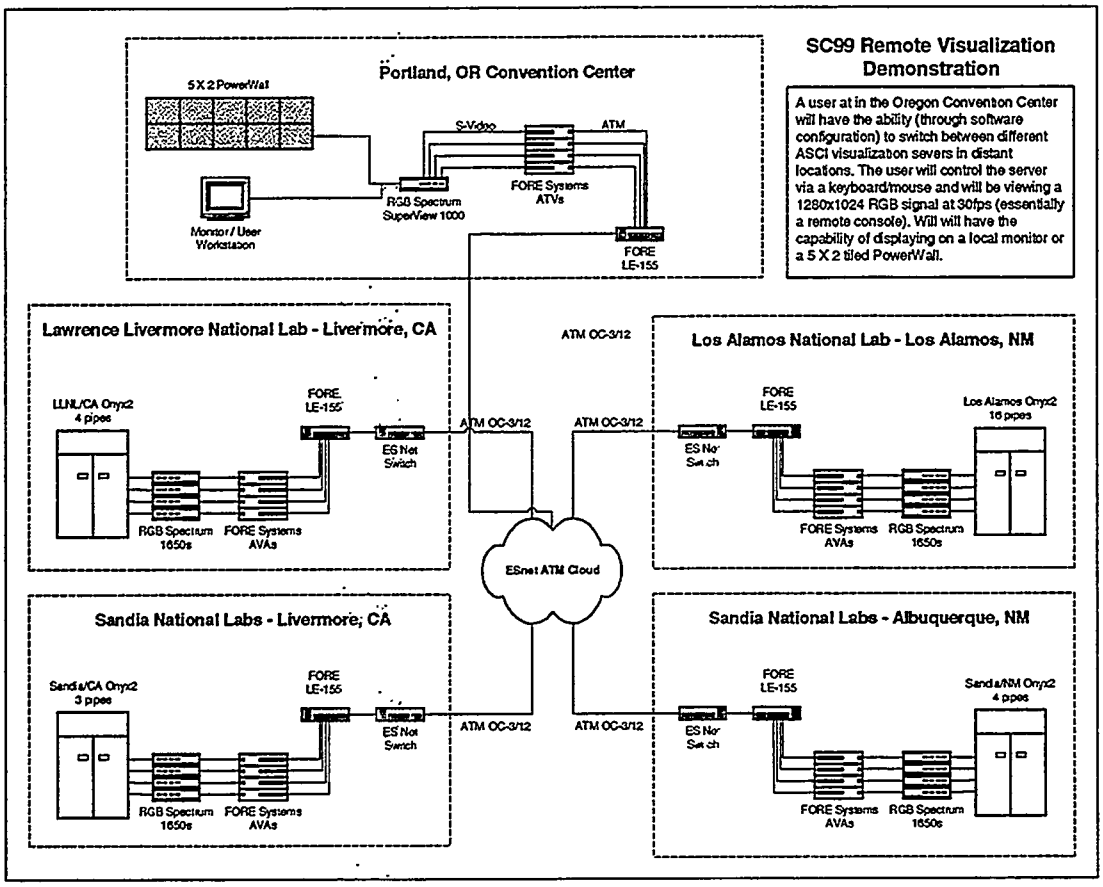

Figure 8: SC '99 Compressed Video System Configuration

To implement the quadrant technique with all servers, each server site was provided with four RGB Spectrum 1650 scan converters to convert the servers' RGB signals to NTSC S-video. The S-video cables were connected to four FORE Systems AVA-300 video encoders to convert the S-video to ATM cells. Finally, the AVAs were connected to a FORE Systems LE-155, which was connected to each site's existing ATM equipment and to ESnet. The cost to bring each server "online" is estimated at $35 K.

To bring the client on line, the incoming video from the servers' console displays are switched out to four FORE Systems ATV-300s video decoders via a FORE Systems ASX-4000 switch. Each of these S-video signals were combined in an RGB Spectrum Superview 1000, and either displayed on a local flat panel, or sent to the powerwall. A Lightwave console extender device was used to extend the display about 100 feet from the demonstration area to the powerwall. The cost to bring the client on line (excluding the ASX-4000 and the Lightwave extender) is estimated at $30 K.

Early in the demonstration setup, ESnet problems precluded fine-tuning of the display settings. In fact, our first server (godzilla at SNL/CA) did not come up until Wednesday morning. However, we re-used configuration settings that were determined *a-priori*, so once the servers came on line, the amount of fine-tuning required was minimal. With sufficient overscan of each quadrant (~10 pixels), alignment of all quadrants proceeded without much difficulty. In addition, although NTSC S-video was used, we were lucky in the sense that color matching between quadrants also went well (although in the future we should use PAL).

Although each quadrant could be "synchronize" in terms of alignment and color balance, synchronizing the frame timing of all four quadrants' video streams could not be accomplished. When showing the system with rapidly moving graphics, this lack of synchronization was noticed when rapidly moving lines

that crossed multiple quadrants appeared to be broken at the quadrant boundary. At first this seemed to be a mis-alignment problem. However, when the rapidly moving object became stationary, the object no longer appeared broken. This problem is due to the fact that the video streams from some quadrants arrived before others. This problem would be solved if there were a way to send video synchronization information across the network, allowing AVAs and ATVs to be genlocked to a common timing source. However, this feature is not provided in the current hardware release.

At the client side, an SGI O2 was used to ssh into each server, and provide console input. Access methods to each site differed significantly, so several hurdles needed to be jumped to ensure not only access to the machine, but also to ensure an unbroken ssh chain (to allow end-to-end transmission of X). Many times (especially early in the week), network or host instabilities severely degraded the availability of this unbroken chain, causing many re-logins.

Since the purpose of the exhibit was to show this system's high interactivity, it was important to show this system with something that required fast turnaround between user input and user output. Initially, this system was shown by grabbing objects from the various SGI "powerflip" demonstrations. After doing this for a while, however, we stumbled upon the "vroom" application. This application required users to click a mouse to cause a vroom "car" to "change lanes" before crashing into other cars. In order to successfully use this application, low latency between images of impending collision and the user's response (mouse clicks) is required. Since the video compression devices are in this path, low encoding latency is required. This application was extremely effective in illustrating this feature of the system.

The network for the remote visualization console demonstration is shown in Figure 9.

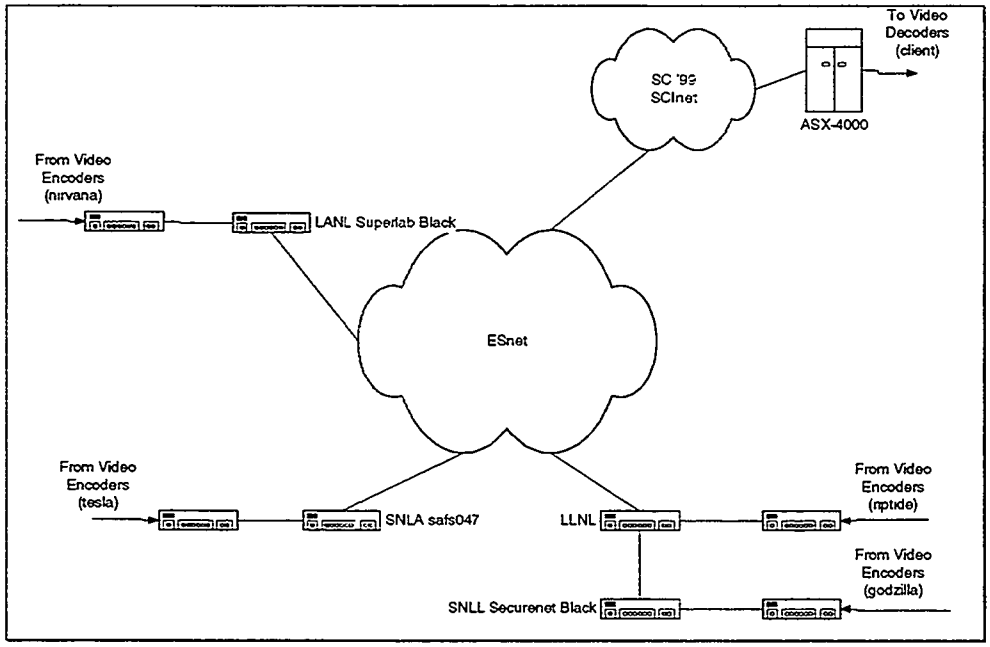

Figure 9: SC '99 Wide-Area Remote Visualization Network

To simplify configuration for the ESnet administrators, four virtual paths were established for this demonstration, terminating in the LLNL switch and in the ASX-4000 on the SC show floor. Each path corresponded to video virtual circuits from each of the server sites, which facilitated experiments such as VP loading response and cell drop tests. To switch cells from each site onto their respective SC '99 VP, the existing SecureNet VP mesh was used, with the switching occurring in the SNLL "SecureNet Black" switch. In the LANL case, additional switching was required in the SNLA "safs047" switch, to switch cells from LANL to SNLL, for additional switching onto the SC PVP to Portland. All virtual circuits and

virtual paths that were configured for this demonstration were UBR virtual circuits. This was done to ensure fair sharing with existing users of these links.

During the show, VP loading tests were performed to examine the response of the video circuits to cell drop. Since all circuits were UBR circuits, no priority was given to the video traffic. Therefore, UBR overloading of links resulted in dropped cells from the video streams. As a result of cell discard, the video circuits went completely blank. No partial information (not even noise) was evident on the video displays.

6 Connecting to Sandia National Laboratories using VPN technology

The remote visualization demonstration at SC99 needed to access the Sandia's high end visualization resources in Sandia's protected intranet. To provide the needed service a Virtual Private Network (VPN) was established. This was the first non-testbed VPN implementation by the Sandia's advance networking group. The SC99 effort allowed us to increase our experience with this rapidly emerging technology. The VPN was established over the Internet between Portland and Albuquerque. The VPN was built using CryptoCluster 2500 hardware from Network Alchemy.

On the Thursday prior to the show, a VPN node was placed at the edge of Sandia's protected network, providing the gateway into the protected network. Because this VPN was intended to be one-time, short-lived, and specific implementation, this node was configured such that the internal interface was terminated in the same LAN as the computer that was to communicate over the VPN (by way of a virtual LAN). This eliminated the need to establish routes in the internal network to correctly point VPN return packets to the VPN device. Instead, the computer that was to communicate through the VPN simply would send those packets to the VPN device on the LAN. A disadvantage of this arrangement is that the VPN management software was then unable to communicate with the VPN node through the internal network. This required placing the VPN management node on the external network.

Establishment of the VPN proved to be challenging, as the SCINET network was very unstable during the weekend prior to the show. Connectivity would come and go as the VPN nodes were being configured. On the Sunday prior to the show, when the Sandia folks at SC'99 were ready, the VPN connection was attempted. The SC'99 node had to be first configured manually in Portland by entering the inside and outside IP addresses and the router. Once this was completed, the node was connected to the network. The management node in Albuquerque then was able to successfully connect to the VPN node in Portland in order to complete the installation. With the installation complete, the next step was to establish a VPN tunnel. This is where the instability of the SCINET network apparently caused problems for the VPN. The establishment of the tunnel seemed to go together without a problem, but no communication was possible through the tunnel. Ping was used to test communication from a laptop in Portland to the destination computer in Albuquerque, but to no avail. Tunnels had been established several times in the lab in Albuquerque over the proceeding weeks without problem, so something was introducing an unknown element in connecting to Portland. Due to Sandia's networking group's inexperience with VPNs, and especially with a real VPN implementation, debugging skills were few. Attempts were made to measure connectivity between the VPN nodes to determine if this was causing the tunnel to fail. The SCINET network instability suggested that perhaps the VPN nodes were having trouble communicating with each other, and yet the management node was able to communicate with each node. Pinging and tracerouting between the VPN nodes showed that indeed the network was having problems, with routes changing and significant delays appearing. However, even during times of stability, the tunnel would not communicate.

With the show underway, and pressure on to get the VPN operational, Network Alchemy's support line was contacted for assistance. It was quickly determined that everything was properly configured. The problem lied elsewhere. However, as advised by Network Alchemy, further debugging required that data be presented to the tunnel to determine where the failure was occurring. Experiences in the lab never

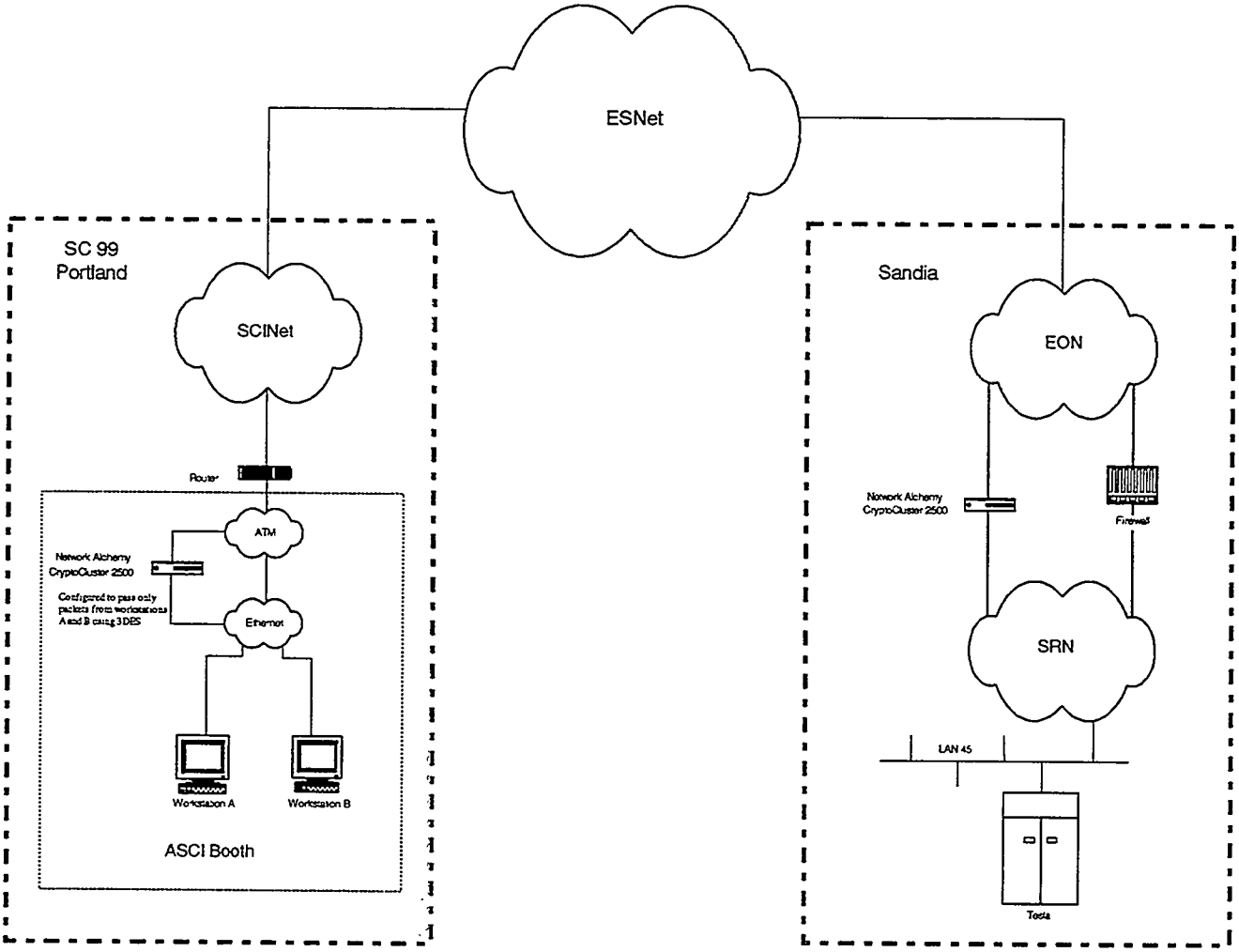

Figure 10 SC99 VPN Design

included a non-functioning tunnel; therefore, tunnel-debugging processes were unknown. Network Alchemy engineers suggested that if the management node are able to communicate with both VPN nodes, then the path between nodes should be fine and attention should be paid to the tunnel only. A tunnel can only be debugged with data presented to the tunnel, such as a simple ping. This is when a tunnel will report activity, such as the establishment of the Security Associations for the IKE and IPSec sessions.

Due to the difficulty in contacting anyone in Portland to have them start some data flow, contact with Network Alchemy's support line was discontinued. However, their assistance was greatly appreciated and very helpful. We had concurred that with no configuration problems identified, the next step would be to drop and re-establish the tunnel. This was done, but could not be tested again due to being unable to contact anyone in Portland.

The good news came later Tuesday from Portand that the VPN was working. Ping was working through the tunnel. This was sufficient to conclude the tunnel was operational. Looking at the node statistics then showed that packets had indeed been transmitted successfully, including Encapsulating Security Payload packets statistics that demonstrate packet encryption.

The bad news came on Wednesday when it was learned that the VPN was not being utilized. The person running the demonstration that was to utilize the VPN was very sick on Monday and Tuesday, and additionally, during the time the VPN was down, alternative paths were explored to bypass the VPN in case it was never established. An alternative path was discovered and was being utilized.

On Thursday, when there was some available time, the Portland folks ran some throughput tests through the tunnel. Lab tests had shown TCP throughput of over 20 Mbps, and UDP over 90 Mbps. Much lower values for a tunnel running over a public network were anticipated, and that was the actual case. The TCP throughput measured only 2.2 Mbps, but interestingly enough, this same throughput was observed when measuring non-VPN paths. Therefore, the throughput appeared to be throttled by some mechanism other than the VPN: perhaps packet delays as well as other traffic. The UDP throughput was 62 Mbps, which is not bad, but not to be taken too seriously due to the unreliable nature of UDP. When the VPN tunnel is pushed hard with UDP, many of the packets never make it through the tunnel because they are dropped.

The lessons learned from this VPN experience will be valuable as we proceed to implement production VPNs. Successful connectivity between the management node and remote VPN nodes should be a sufficient indicator of successful connectivity between VPN nodes. Therefore, attention should be paid to tunnel operation, which requires a data stream for debugging. A non-functioning but properly configured tunnel, should be torn down and rebuilt (a simple procedure). Once communications is established through a tunnel, protocol statistics in either the end nodes or the VPN nodes will indicate if further problems exist. A stable network also helps!

7 Lessons Learned

Internet Access at SC99 was troublesome over the weekend. External connectivity (from the booth network to the Internet) is always an area of concern. SCinet builds a state of the art network in about 5 days. The short time frame and the use of the latest equipment available can cause network instabilities. This year, as in most years, the network stabilized later in the week. The VPN effort and the Breckenridge's movies were impacted by network outages. The goal of the booth networking team is to establish the best working relationship with SCinet and strive for the best booth Internet access possible while still pursuing the networking research side of networking.

The theater concept work well this year. Most of the presentations were canned. While these presentations didn't fully utilize the simulation capability of the scientific environment that was available, the general audience accepted these presentations just as well as the more difficult live visualization displays. We have always taken the viewpoint that these challenges are what make the ASCI booth different from other SC booths. We take chances and these challenges push the envelope.

In the booth we had problems with maintaining console access. The cabling of the consoles ports on each piece should have all been installed and labeled before buttoning up the exhibit. The routing plan for the booth should have been shared with the XNET demonstration before the setup. Socialization of the plan with our counterparts might have made it easier to pull together the networking demos.

8 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. I would like to acknowledge the first class networking team that supported the booth network and the networking demonstrations

Roger Adams	Sandia National Laboratories
James Brandt	Sandia National Laboratories
Wayne Butman	Lawrence Livermore National Laboratory
Helen Chen	Sandia National Laboratories
Martha Ernest	Sandia National Laboratories
Lawrence Macintyre	Y12
Luis Martinez	Sandia National Laboratories
Leonard Stans	Sandia National Laboratories
Pete Wykopp	Sandia National Laboratories

We would like to thank the following individuals for their efforts in making our ASCI's DISCOM *SC 99* networking efforts a success.

SNL	-	Jim Ang, Forrest "Herb" Blair, Authurine Breckenridge, Joseph Brenkosh, Judy Beiriger, Gary Evans, Stephanie Fellows, Rich Gay, Jerry Gorman, Steve Gossage, Rena Haynes, Tan Chang "Richard" Hu, Steven L. Humphreys, Wilbur Johnson, John H. Naegle, Ron Olsberg, Diana Perea, Lyndon Pierson, Tom Tarman, Tim Toole, Steve Valdez, Alan Williams, Ed Witzke,
LANL	-	Alice Chapman, Ann Hayes, Mitch Sukalski, Steve Tenbrick
LLNL	-	Don Patterson, George Pavel, Jean Shuler, Joe Slavic, Dave Wiltzius, Mary Zosel
LBL	-	Wes Bethal, Steve Lau
Y12	-	Rhonda Macintyre
SCINET/XNET-		Paul Daspit, Jim Deleskie
ESNET	-	Jim Lieghton, Kevin Oberman
CISCO	-	Mark Bleth; Brad Irwin
Corp Comm.	-	Bob Dobinski
Fore Systems	-	Dave DeBhur, Frank Yeh,
Compaq	-	Ira Grollman
Avici	-	Glen Yallaly, Hank Zannini
Lucent	-	Elaine Jones, Ken Evanchik, Judy Meester
Adtech	-	John Clem
Univ. of Min	-	Thomas M. Ruwart and his team of video wizards
NTON	-	William Lennon

9 References

[1] http://www.llnl.gov/NTON

[2] R. Ramaswami and K. N. Sivarajan, "Optical Networks – A Practical Perspective", Morgan Kaufmann Publishers, Inc. 1998.

[3] J. Manchester, J. Anderson, and R. Wilder, "IP over SONET", IEEE Communications Magazine, Vol. 36 No. 5, May 1998, pp. 136-142

[4] http://rockford.lbl.gov/projects/ibravr/

[5] http://www-didc.lbl.gov/DPSS/Overview/DPSS.handout.fm.html

[6] W. J. Dally, "Scaleable Switching Fabrics for Internet Routers", Computer Systems Laboratory, Stanford University and Avici Systems, Inc. July 1999.

[7] http://www.nortelnetworks.com/products/library/collateral/56009.25-09-99.pdf

[8] ftp://ftp.isi.edu/in-notes/rfc1771.txt

[9] http://www.cs.sandia.gov/cplant/

[10] http://www.sc99.org

[11] http:// www.lucent.com/ins/products/nx64000)

[12] http://www.avici.com/

Distribution

0139 P. J. Wilson, 9902
0318 R. A. Haynes, 9215
0318 D. J. Zimmerer, 9215
0318 P. D Heermann, 9215]
0318 A. Breckenridge, 9215
0318 G. S. Davidson, 9202
0321 W. J. Camp, 9200
0321 A. L. Hale, 9224
0321 J. A. Ang 9024
0421 R. J. Detry, 9800
0429 J. S. Rottler, 2100
0449 R. L. Hutchinson, 6236
0469 M. J. Murphy, 2900
0622 D. C. Jones, 9011
0630 P. J. Vandevender, 9010
0660 W. D. Swartz, 9313
0741 S. G. Varnado, 6200
0801 M. O. Vahle, 9300 (10)
0801 W. F. Mason, 9302
0806 C. D. Brown, 9321
0806 J. H. Naegle, 9316
0806 J. P. Brenkosh, 9316
0806 L. F. Tolendino, 9316
0806 L. G. Martinez, 9316(5)
0806 L. G. Pierson, 9316
0806 L. Stans, 9316 (10)
0806 S. A. Gossage, 9316
0806 T. J. Pratt, 9316(25)
0806 T. D. Tarman, 9316
0806 M.J Ernest, 9316
0806 T. C. Hu, 9316
0806 J. M. Eldridge, 9316
0812 M. R. Sjulin, 9314
0812 B. C. Whittet, 9314
0812 R. L. Adams, 9314 (5)
0813 R. M. Cahoon, 9311
0820 P. Yarrington, 9232
0826 J. D. Zepper, 9111
0828 R. K. Thomas, 9102
0841 T. C. Bickel, 9100
0866 I. C. Alexander, 1904
0899 Technical Library, 9616 (2)
1002 P. J. Eicker, 15200
1201 J. M. McGlaun, 5903
9003 K.E. Washington 8900
9003 J. Costa, 9903
9003 D. L. Crawford 9900
9011 H. Y. Chen, 8910 (10)
9011 J. A. Hutchins, 8910
9011 J. M. Brandt, 8910 (5)
9011 P. S. Wyckoff, 8910 (5)
9011 P. W. Dean, 8903
9011 E. D. Dart, 8910
9012 J. A. Friesen, 8990
9012 T. J.Toole, 8990
9012 J. N Jortner, 8990
9012 R. D. Gay, 8930
9018 Central Technical Files, 8940-2
9037 J. C. Berry, 89301

0612 Review and Approval Desk, 9612
for DOE/OSTI