SANDIA REPORT
SAND2001-3625
Unlimited Release
Printed November 2001

The ASCI Network for SC 2000: Gigabyte Per Second Networking

Thomas J. Pratti, John H. Naegle, Luis G. Martinez, Tan Chang Hu, Marc M. Miller, Marty Barnaby, Roger L. Adams, and Ed Klaus

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.

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

SAND2001-3625
Unlimited Release
Printed December 2001

The ASCI Network for SC 2000: Gigabyte Per Second Networking

Thomas J. Pratt, John H. Naegle, Luis G. Martinez,
Tan Chang Hu, Marc M. Miller,
Advanced Network Integration

Marty Barnaby
Scientific Computing Systems

Roger L. Adams, Ed Klaus
Telecommunications Operations Department II
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0806

Abstract

This document highlights the Discom^2's Distance computing and communication team activities at the 2000 Supercomputing conference in Dallas Texas. This conference is sponsored by the IEEE and ACM. Sandia has participation in the conference has now spanned a decade, for the last five years Sandia National Laboratories, Los Alamos National Lab and Lawrence Livermore National Lab have come together at the conference under the DOE's ASCI, Accelerated Strategic Computing Initiatives, Program rubric to demonstrate ASCI's emerging capabilities in computational science and our combined expertise in high performance computer science and communication infrastructure. DISCOM2 uses this forum to demonstrate and focus communication and networking developments within the program. At *SC 2000*, DISCOM demonstrated a pre-standard implementation of 10 Gigabit Ethernet, the first gigabyte per second data IP network transfer application, and VPN technology that enabled a remote Distributed Resource Management tools demonstration. Additionally a national OC48 POS network was constructed to support applications running between the show floor and home facilities. This network created the opportunity to test PSE's Parallel File Transfer Protocol (PFTP) across a network that had similar speed and distances as the then proposed DISCOM WAN. The SCINET SC2000 showcased wireless networking and the networking team had the opportunity to explore this emerging technology while on the convention exhibit floor. We also supported the production networking needs of the booth. This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations supports DISCOM overall strategies in high performance computing networking.

CONTENTS

List of Figures

1 Introduction

SC2000 marked the twelfth year for this IEEE high performance computing and communication conference. The conference was held from November 4th - 10th in Dallas, Texas. The conference is sponsored by the IEEE and ACM. While the three defense program laboratories, Sandia, LANL and LLNL have all participated in the conference for a decade and for the last four years the three laboratories have come together at the conference under the DOE's ASCI, Accelerated Strategic Computing Initiatives, Program rubric to demonstrate ASCI's emerging capabilities in computational science and our expertise in high performance computer science and communications. The DISCOM2 communication team uses this forum to demonstrate and focus communication and networking developments within the program. Many notable accomplishments have been achieved during these four years that the ASCI has lead our efforts starting with sc96 in Pittsburgh, Pennsylvania. These accomplishment include, The first OC12 connected remote clusters (1997) telephony and video over ATM, (1997), 10 gigabit per second networking over a 16 wavelength DWDM system (1998), and Terabit Routing(1999).

In the planning for *SC 2000*, the DISCOM2 team was challenged to create a gigabyte per second data movement application. The genesis of the challenge arose from the ASCI curves that require ten gigabit per second networks and application to be deployed in the early 2002 time period. We also were inspired by the inaugural SCINET Netchallenge. To achieve this level of performance require every section of the data movement problem had to be optimized. It required parallelism to be extended end to end. We partnered with CPLANT, SGI, and CISCO SYSTEM to create this demonstration.. By doing the demo we extended our knowledge base for building parallel networks. We extended the CPLANT model for data movement from a cluster. We increased our knowledge of the SGI cluster file system CXFS. A segment of the network supporting this demo consisted of a pre-standard implementation of 10 Gigabit Ethernet. To create and move the data for this demo a gigabyte per second cluster based data transfer application was developed. The details of this demonstration are presented in section five of this document

Other networking activities in the booth included deployment of VPN technology to support a remote Distributed Resource Management tools demonstration. The details of the VPN network are cover in section four of this document. A Sandia developed RGB video extender was also demonstrated at SC2000. Our colleagues at Sandia California in conjunction with Avici System, Lawrence Livermore Lab, Lawrence Berkeley Labs, the National Transparent Optical Network Consortium, NTON, and QWEST provided a OC48 WAN platform to run application and experiments between the convention center and home facilities. The DISCOM team members from LLNL used the opportunity to test PSE's Parallel File Transfer Protocol (PFTP) across this shared network. The result was a five month head start on understanding the wide area effects on Parallel File Transfer Protocol (PFTP) performance. For more information on the demos run across the WAN you can refer to the website at: http://www.avici.com/pr110800.html

This year demonstrations highlighted DISCOM2's and PSE's accelerating pace for providing performance to the ASCI scientific community. The ASCI booth this year was build on the theater scheme. The central feature of the booth was a two by three Powerwall theater. The presentations on the powerwall were videotaped and those tapes are available through the ASCI office. The ASCI theme of the booth for this year was Curves and Barriers. A large poster showing the ASCI roadmap was a huge success. It created a lot of interest in the conference attendees. An added bonus for the networking team was that they got to explore wireless networking. Wireless networking was featured by SCINET this year. SCINET built an extensive wireless network that covered the convention complex and adjacent hotels. The ASCI booth networking team had time to try out this technology while at the convention and we uncovered vulnerabilities in the current technology that will need to be addressed before wide scale use of this technology within the laboratory networks will be made. We believe that it is possible to build a production networking structure around wireless but it requires more than just the out of the box wireless hardware and

software. This exposure is a typical example of how the conference provides a forum to explore communication technique and technologies on a wider stage.

Joint participation from Sandia, Los Alamos, and Lawrence Livermore in the planning and support of the booth's production networking continued this year.

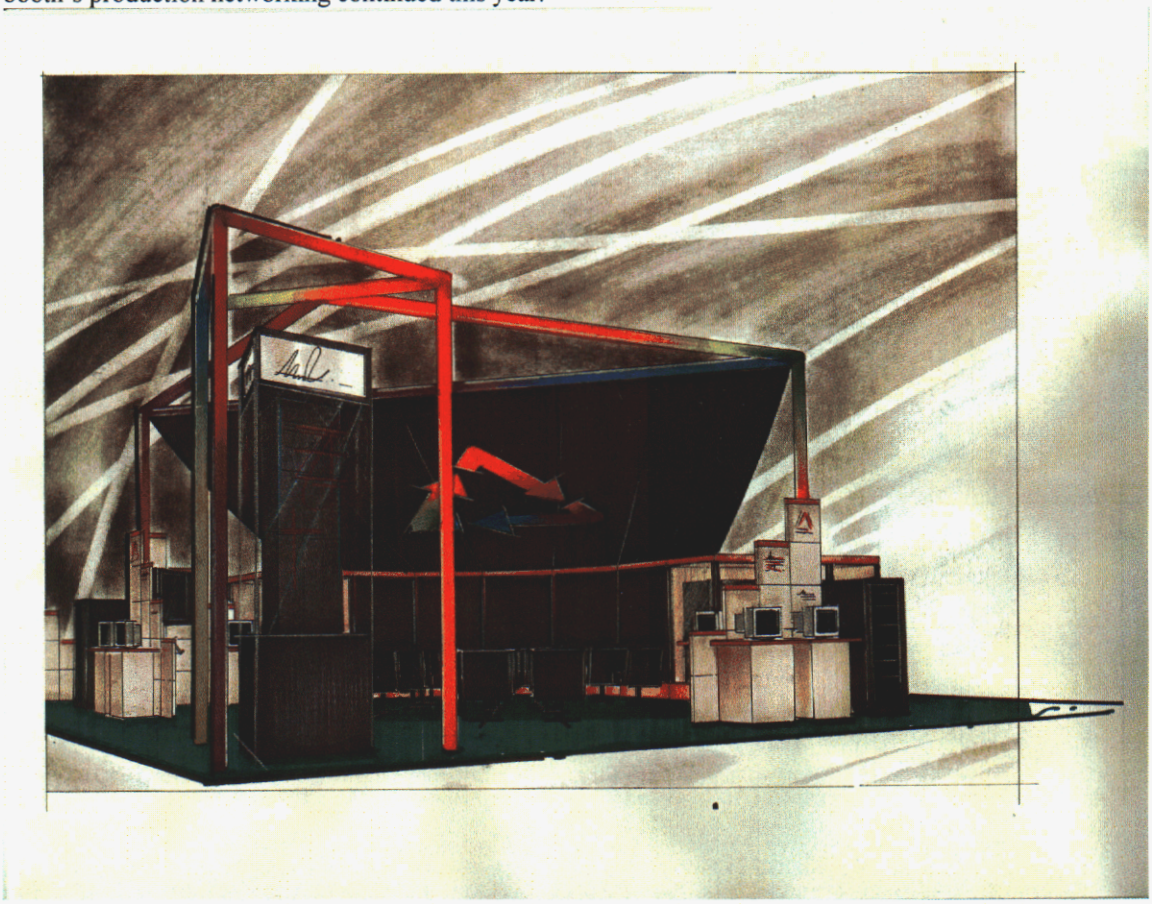


Figure 1: The SC2000 ASCI Booth

Many ASCI's technical and commercial partners made the significant contributions that these demonstrations represent possible. The industrial partners at this year show included Cisco, Avici, Nortel, Compaq, SGI, and Sprint. The Visualization Theater within the ASCI booth was provided by the University of Minnesota's Laboratory for Computational Science and Engineering, LCSE. SCINET experimental networking group XNET was key partner for the ASCI booth this year.

Some of the enduring themes and benefits of this conference are:

- partnering with industry to gain early access to new technology,
- focusing current projects and activities through 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 and organizations,
- providing a stage to professionally interact with colleagues and associates from other organizations in order to challenge and validate our current thinking.

This paper documents those accomplishments, discusses the details of their implementation, and describes how these demonstrations supports ASCI's strategies in networking

2 Equipment and Booth Layout

The ASCI booth at SC2000 was nearly identical to the previous year booth. The booth was organized around a theater concept with a 3X2 powerwall at the center of the theater. Extending out from the theater into the booth was two area for individual demos, while presentation on the powerwall are schedule to run a few minutes per speakers the demo area are always available for conference attendee to view. On both sides of the theater a large display space was created behind clear Plexiglas windows. This year the network demonstration was set up in the display space to the left of the theater.

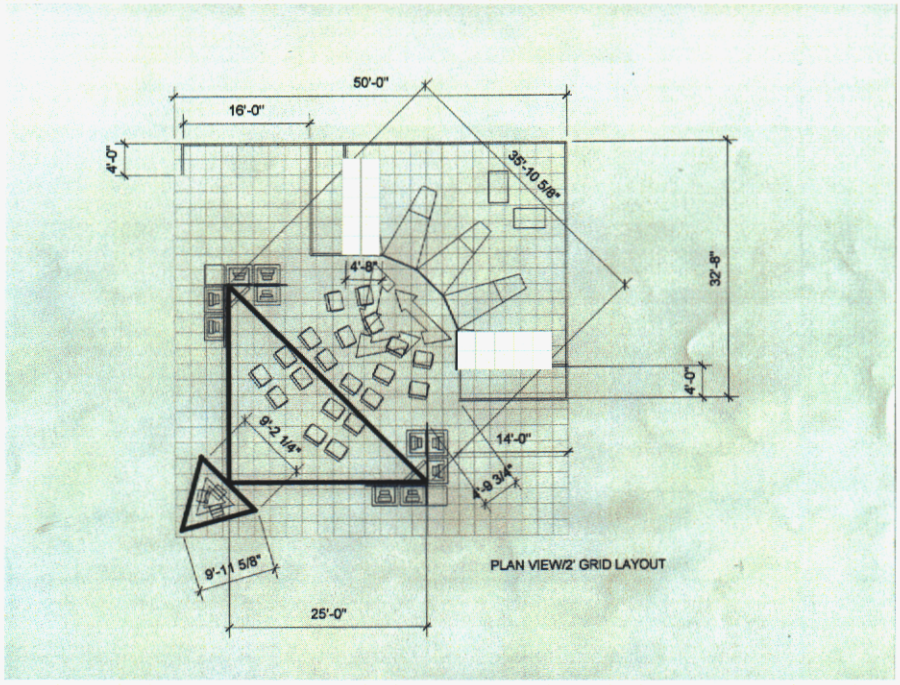


Figure 2: ASCI Booth Layout

Figure 3 ASCI's Booth Networking Demonstration Area

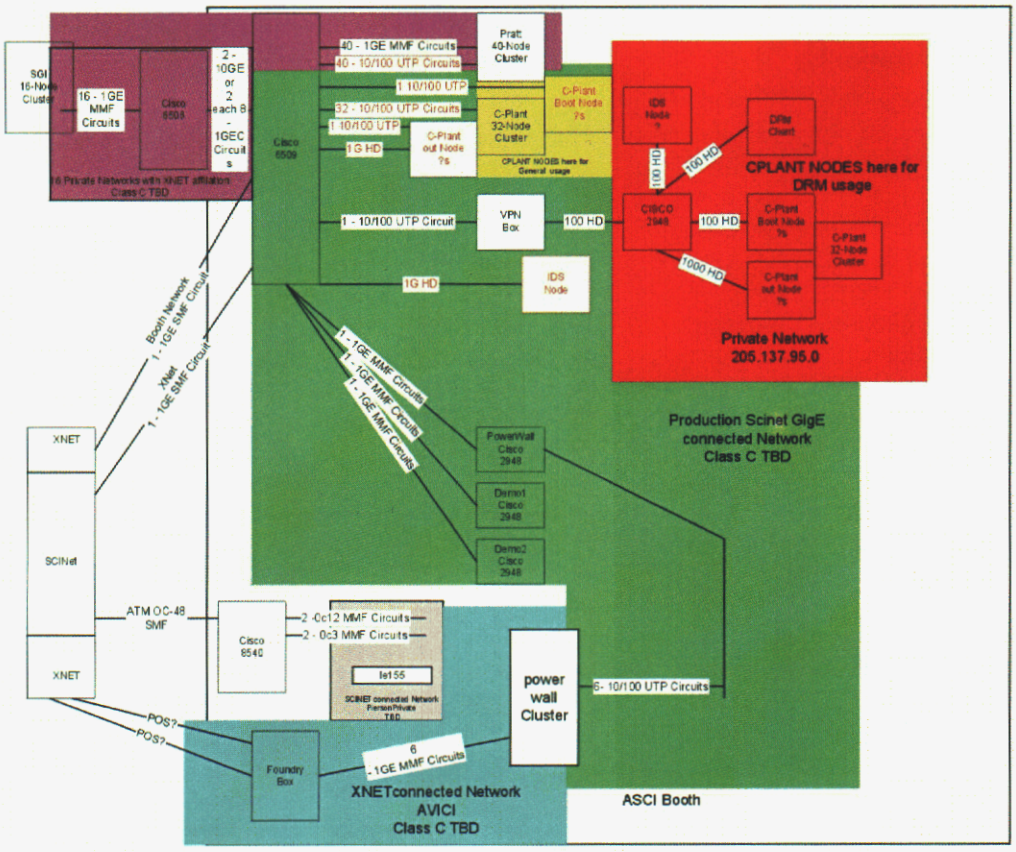
The Discom advance networking demonstration this year focused on creating an IP network capable of performing at a gigabyte per second level. The network centered on a CISCO 6509 Ethernet switch that contained two prestandard 10Gigabit Ethernet ports. To drive the network two racks of 33 Alpha Linux workstations were deployed. One of these racks was configured as a CPLANT cluster with a Myrinet networked backplane. The other rack was constructed with gigabit Ethernet as the network connection. This rack approximated CPLANT output nodes. A stack of SMARTBIT Ethernet testers were added to allow the demonstrator the ability to test the 10 Gigabit networks for performance and function. Additionally to support this demo two SGI 9400 disc arrays, a SGI Origin 2600, and a Cisco 6506 was deployed in the SGI booth at SC. Interconnecting the ASCI booth to the SGI booth was two special fiber array cables and eight single mode fiber cables. The single mode fiber bundle was put into place as a contingency plan in case problem with the beta 10Gigabit Ethernet failed. Once again our contingency plan paid off.

Also in the booth network for the second year in a row was an AVICI terabit router. The AVICI terminated the OC48 POS, Packet Over SONET network. This year the AVICI was front-end by a Riverstone, formally Cabletron, Ethernet switch router. This year the booth equipment included a portable air conditioner. The cool air was ducted so that the Linux clusters would not overheat. Additional ducting was provided so that the powerwall equipment and support personnel got some cooling also. This year we had used wireless workstations to monitor the production networks within the booth. This provided several benefits to the booth. Support personnel didn't need to compete for the space within the booth and the working conditions outside the booth was much more pleasant.

3 *SC 2000* Networks

At *SC 2000*, for the fourth 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. The design and operation of the production network was a joint effort of all three of the Defense Program Laboratories. 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 SC2000 forum to validate new communication technology while increasing the understanding of the high performance networking technologies available to the ASCI communities. 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 highlight the use of parallel networks and to maximize the performance of data movement between network connected computer clusters and file systems.

The design of the network for SC2000 was structured to provide reliable production communication to the majority of booth participants, while providing a platform for high speed communication and networking research. At the core of the booth network four communication protocols was deployed 10 Gigabit Ethernet, Gigabit Ethernet, ATM, and Packet Over SONET (POS). Gigabit Ethernet, 10 gigabit Ethernet and POS equipment were integrated inside the booth. The ATM network utilization was limited to a single demonstration. It was solely used to accommodate the RGB video extender demonstration. This was a major shift from earlier conferences. In addition CISCO System equipment that traditionally has provided the production networking equipment, this year also provided highest performing network research equipment for the booth. The design utilized twenty-two virtual networks that provided the separation between the production networks and the research networks.



SC2000 SCHEMATIC
ED KLAUS and ROGER ADAMS
tjp revised 09/19/00

Figure 4: SC2000 Asci Network

4 Extending the TeraOps Network to Super Computing 2000 using Virtual Private Networking

As part of the Super Computing 2000 show in Dallas Texas, the Sandia National Laboratories super computer network (TeraOps LAN) was extended to the show floor using Lucent Bricks. The Brick is a commercial grade Firewall and Virtual Private Network (VPN) system.

A demonstration of the DRM toolset available on the Sandia National Laboratories super-computing network was desired as part of the SC2000 floor exhibition. The agreement to permit access from the show floor to the TeraOps network, however, was contingent on providing secure communications to the exhibitors without degrading the existing security environment. To meet this goal, it was elected to provide a Virtual Private Network connection from the TeraOps to the show floor using the Lucent Brick. The Brick is a commercial, state-full firewall capable of providing an encrypted network path in conjunction with access controls.

The design of the TeraOps extension was bounded by the following criteria.

1. The client hosts in the exhibitor's booth require the same access as the standard TeraOps host, i.e. the same network servers available to the TeraOps network need to be available to the exhibitor systems.
2. The TeraOps and Sandia Open Network (SON) production network environment should not be compromised by the VPN extension, i.e. the network path between the exhibition and the TeraOps network shall exclude any third party not authorized to utilize it.
3. Sandia should be able to monitor and report on the network extension during the week of the show.
4. The TeraOps and SON network security will not be adversely impacted by the extension, i.e. changes to the TeraOps network Access Control List (ACL) and to the SON routing should be at a minimum.

A VPN fit criteria 2 and 3 and a firewall meets criteria 1, 3 and 4. The equipment available on hand from Network Alchemy and Lucent could meet all five criteria. The two equipment types, however, have different placement philosophies that impact the VPN design. The Lucent Brick was designed to be placed as an in-line bridge and by default both the in and out interfaces are in the same logical subnet. The Network Alchemy acts more like an in-line static router with the in and out interfaces on two different logical subnets. In both, the VPN tunnel endpoints are terminated with a pseudo address that is an Internet routable address. A management station is required for either piece of equipment. As the Lucent Brick management station was already built as part of another project and had better reporting capabilities, it was elected to use the Lucent Brick as the primary and the Network Alchemy as the backup.

The exhibition termination point was simple to design as there were no legacy requirements to account for and the network requirements could be met with 100BaseT. The number of clients to be attached was limited to a maximum of three. This isolated network was easily attached to the existing SC2000 network. The schematic for the SC2000 termination appears in Figure 1:

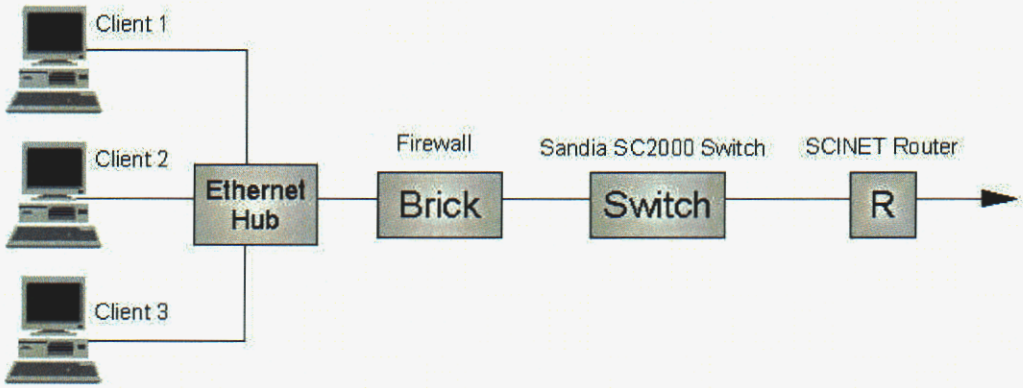


Figure 5 SC2000 Termination Schematic

The TeraOps termination point, however, was more complex. The initial design was to place the Lucent Brick in line with the TeraOps network. This would mean no changes to the SON production network and minimal changes to the TeraOps ACL. The Lucent Brick would be able to use the same logical subnet address space for the VPN termination point and the in/out interfaces. It would require the addition of VPN services to the TeraOps ACL. Physical media incompatibility between the Lucent Brick running 100BaseT and the TeraOps network running OC-3 ATM killed this option.

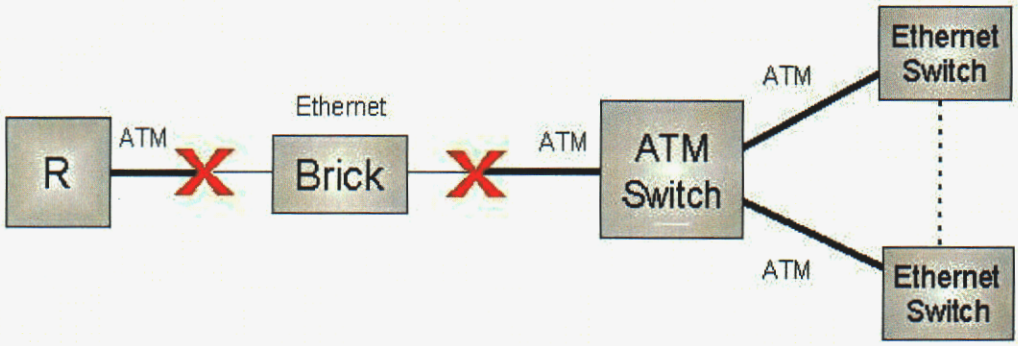


Figure 6 Ethernet/ATM Mis-Match

The next option was to place the Lucent Brick behind the Ethernet switches as shown in Figure 3. This circumvented the media mismatch. Unfortunately, as the Lucent Brick acts like a bridge, this configuration created a loop. The loop amplified broadcasts through the Ethernet switch and resulted into a broadcast storm that momentarily interrupted the TeraOps network during testing.

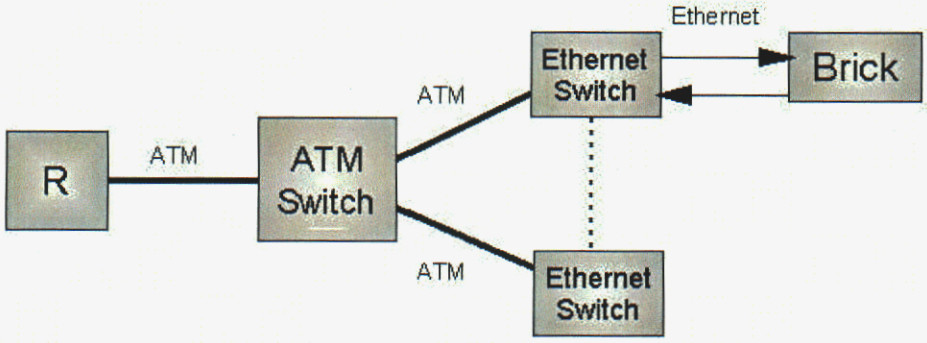


Figure 7 Loop Condition

The next attempt, shown in Figure 4., was to place the brick in parallel with the TeraOps network and use two separate subnet addresses on the in and out interfaces. The input interface would be connected to the Sandia router and the output interface would be connected into the TeraOps network

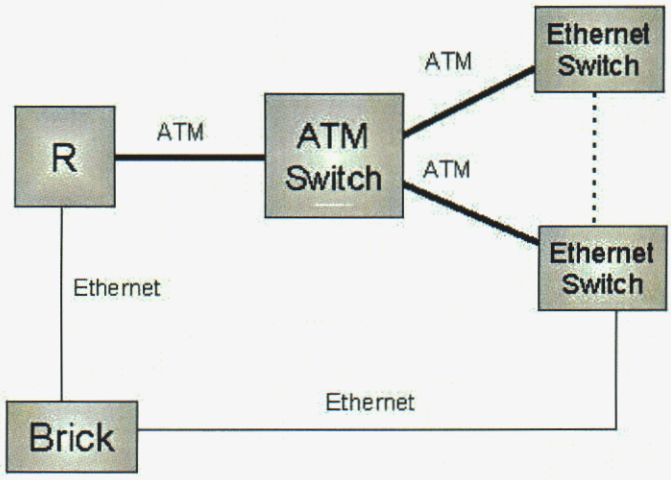


Figure 8 Parallel Configuration Schematic

The design was implemented in the laboratory using two end hosts, two Lucent Security Management Server (LSMS), two Lucent Bricks, a single router, and one Ethernet switch. Private address space was used to emulate the different subnets. Two LSMS were used to simplify the management issue. A single LSMS would require in-band management through the Internet up to the SC2000 network, and this would have increased the complexity of the Lucent Brick configurations. It was found, however, that although the Brick acts like a bridge at the physical layer, it actually filters at the transport layer. In other words, once the VPN tunnel is created and the packets decoded at the tunnel termination point, the Brick could not forward the decoded packets to the client subnet.

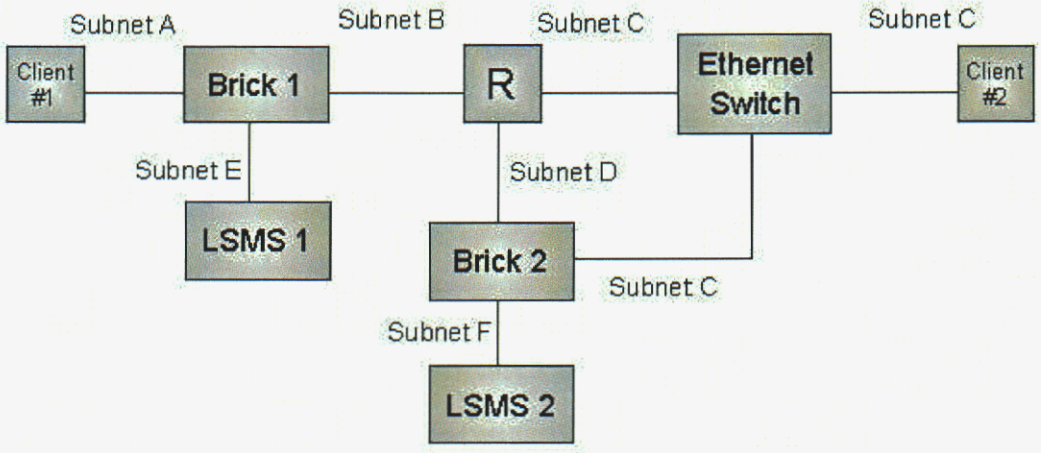


Figure 9 Parallel Configuration Detail

Using the laboratory setup shown in Figure 5 as reference, this is what happened:

1. Client #1 initiates a connection designated for client #2. The source address of the packet is from subnet A and the destination address falls into subnet C.
2. The Lucent Bricks (1 and 2) negotiate a VPN tunnel and set it up.
3. The packet from client #1 is encrypted, encapsulated, and sent over the VPN tunnel. The source address of the encapsulated packet falls within subnet B and the destination address falls into subnet D.
4. The client #1 packet is de-encapsulated and decrypted.
5. Lucent Brick #2 now has a packet destined for subnet C and does not forward it although it has its output interface assigned to subnet C. The packet gets dropped.
6. The same happens for traffic going to the other direction.

The LSMS did not indicate any error nor was there any caveats in the Lucent documentation regarding this behavior. This behavior was only deduced by observing the packet count from the Lucent Brick console.

To implement the TeraOps extension we had to place the VPN endpoint external to the TeraOps network due to the physical network limitation and the Brick functional behavior (see Figure 6). The design opted for SC2000 had the VPN terminating outside the TeraOps packet filtering firewall. As a result, data traveled unencrypted output of the TeraOps to the Brick. This was deemed an acceptable risk since the data was within the Sandia controlled network environment. Explicit routes were placed into the Sandia SON routers to limit the traffic destined for the exhibitor floor to the exhibitor's client systems. Entries were made to the TeraOps firewall to permit the client systems to access the TeraOps networks. All connections between the TeraOps network and the SC2000 network were monitored by the LSMS.

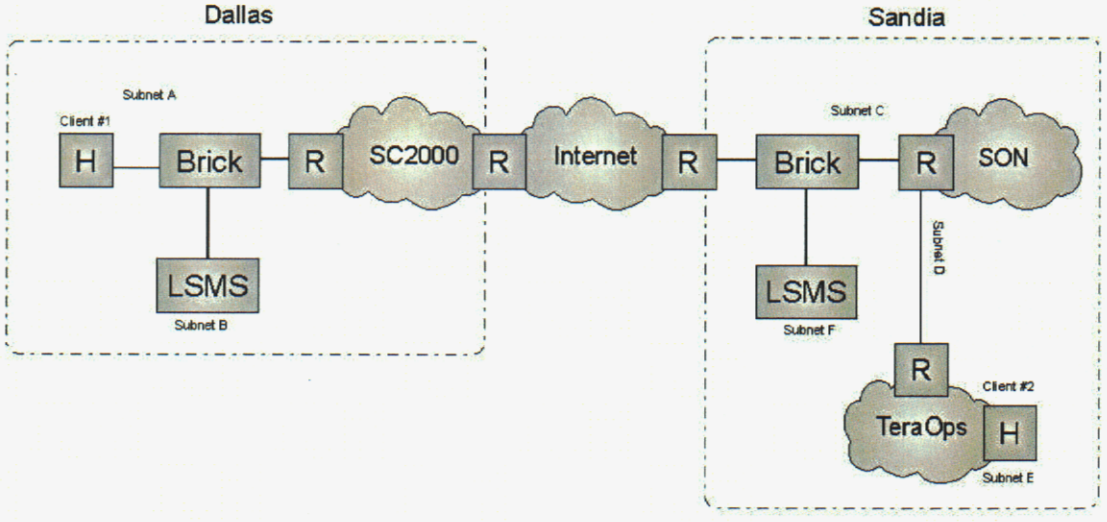


Figure 10 Initial SC2000 VPN Configuration

This implementation worked for both the Lucent Brick and for the Network Alchemy equipment with the exception of different subnet addressing schemes of the in and out interfaces.

The last issue encountered as part of the SC2000 extension was an operational one. The VPN security association was configured manually and was suppose to last for the four days of the show. It was therefore a surprise when it was found that the VPN would not re-establish itself after disconnecting the Dallas LSMS from the Brick overnight. The LSMS was removed and transported back to Albuquerque once the VPN was established to prevent theft of the LSMS laptop. To address this unforeseen issue, a second LSMS was built

at the Sandia site for in band management of the show floor Brick (a configuration we initially tried to avoid due to the added complexity). The existing Sandia LSMS could not be used for managing the Dallas Brick as it was utilizing private address for its interface that was not routable on the Internet. A new Brick configuration was built at the Sandia site and sent electronically to SC2000. The image was written to a floppy with the help of the booth personnel and used to update the SC2000 Brick. Once the SC2000 Brick was able to exchange information with its management LSMS back at Sandia, it was able to maintain the VPN tunnel for the remainder of the show. The final configuration is shown in Figure 7.

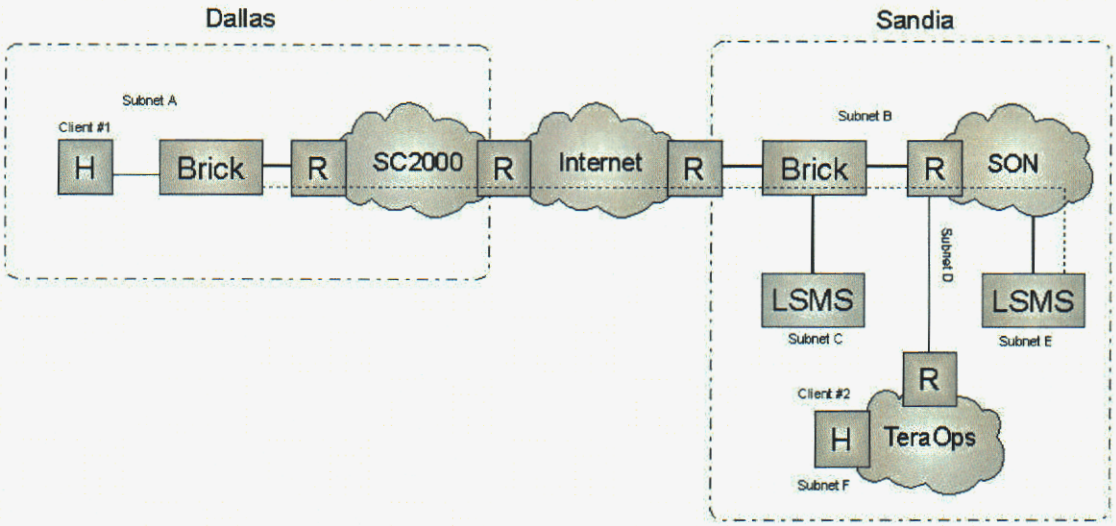


Figure 11 Final SC2000 VPN Configuration

The extension of the TeraOps network to SC2000 proved to be feasible but technically and operationally challenging. Once the design issues were ironed out, the VPN tunnel behaved as advertised. The VPN equipment functional limitation, physical media, legacy network design, and the location of the management station all impacted the implementation. The experience gained from this endeavor expanded the VPN knowledge base at Sandia.

5 Gigabyte per second Linux-cluster-memory to file-storage via parallel TCP streams

The hardware for this demonstration included the Linux cluster, an SGI Origin 2400, four SGI TP9400 disc array, a Cisco Systems' 6509, and a Cisco Systems' 6506. The Linux cluster consisted of 33 Compaq DS10L. Each DS10L had a 466 MHz 21264 (EV6) microprocessor, 256 MB ECC SDRAM, 4 MB L3 cache, and 1 gigabit Ethernet card. The SGI Origin 2400 had 32 processors, 16 gigabit Ethernet cards and 16 fibrechannel cards. Each TP9400 had 4 fibrechannel controllers. The Cisco 6509 consisted of SupII supervisory cards, two prestandard 10 Gigabit Ethernet ports, and 96 gigabit Ethernet ports. The Cisco 6506 consisted of SupII supervisory cards, two prestandard 10-gigabit Ethernet ports, and thirty-two gigabit Ethernet cards. All of the Cisco network cards were fabric enable. Originally the plan was to connect the 6509 to the 6506 with the two prestandard 10 gigabit Ethernet networks. We had a problem with connector alignment on two of the 10 gigabit Ethernet cards and we replaced that pair with eight single mode gigabit Ethernet ports. The 6509 was also the center of the ASCI booth production network. No impact on the production network was seen during the large data movement tests from this demonstration.

Sixteen subnets were used in the demonstration. Each subnet consisted of two Linux gigabit Ethernet attached nodes and one Origin 2400 gigabit Ethernet node. The sixteen subnets were combined into two groups of eight. Each group was configured to use a different 10 gigabit trunk. Because one of the 10 gigabit trunks was inoperable the second group of eight was reconfigured so that each subnet would use a separate gigabit Ethernet trunk between the cluster and the SGI file server.

The concept and design for the software that provided the top layer of our demonstration was modeled on a portion of the Cplant ENFS, Extended Network File Service created at Sandia. While the data moving application is not representative of a complete application for moving scientific or visualization data from a cluster, it is a semblance of one segment of a compound arrangement that would delivers the appearance and functionality of a filesystem dynamically to computational executions on very large computational clusters like Sandia's CPlant

With respect to the ENFS path, the gap that we isolated is normally bridged by regular NFS. However, this protocol would not support our desire of maximizing network throughput, so it was replaced with SGI's Bulk Data Server (BDS). This is proprietary software construct, in which SGI has made some investment, and it appears to be a novel, leveragable mechanism precisely suited for our objective. In SGI documentation BDS is described as sitting on top of NFS, though it would more specifically be described as being integrated within the client-side of the NFS, and operates with an exclusive, persistent (and non-secured) daemon on the server-side. In a special agreement, Sandia was given access to most of the source code for the original BDS version 1 package, with the understanding that Sandia would only be experimenting with porting the client to Linux.

Sandia never got the piece that incorporated the BDS option in the NFS client, but for our demonstration this was adequate because we would be running a Linux-compiled deviation of BDS developed by Sandia. Since we were already partnering with SGI, with them providing networking, computing, and filesystem hardware and software for the service end, we could take advantage of their high-performance, network-to-disk daemon. This allowed SGI to take full responsibility of their side of the demonstration, including this item that they held proprietarily. Prior to SC00, tests were run that indicated that Marty's revisions to the server for an SGI platform made no difference in performance of the server, so the final software configuration was comprised of Marty's client side code in Linux cluster, delivering data to the SGI's stock daemon.

The simple application itself, distributed uniformly across all the Linux nodes, directed, of the daemon on the remote-end and in the appropriate protocol, the same physical file, on the remote-end filesystem, to be opened from each, respectively, and wrote to it in the same block sizes. The only difference across the nodes was that the offset of each write was staggered, with respect to that node, so that each block from every one had its own place in the file; in a normal parallel I/O fashion. We did not achieve maximum performance until we used all 32 Linux nodes writing with blocks of 16 megabytes, which means the total amount of data written in one complete, parallel write was 512 megabytes—a very large chunk (maybe unrealistically so) to have in one operation.

We started the demonstration by testing the network performance using ttcp a network performance tool. The ttcp test showed that the network was capable of achieving data rates of 1.7 gigabytes per second. Next we used the devnull mode, a standard option for the BDS daemon and documented in the man page. This allowed us to isolate the application

performance from the disc and filesystem performance, which, again suited our purposes, as the intent is to overwrite the data, buffers with no actual disk storage. With this tactic, we were able to see total data rates of 1.3 gigabytes per second, in continuous, sustained aggregate, using 32 nodes on the client side and all 16 gigabit Ethernet addresses on the SGI machine. SGI's graphical utility pmchart provided a nice real-time picture of the data transfer, automatically performing the aggregation function. We next attempted to store the data to a single file on the disk system that SGI provided (their new CXFS). This resulted in a short peak, 10-20 seconds of data transfer that was greater than 500 megabytes per second after which the server crashed. By reducing the sources on the Linux cluster side we were able to reliably transfer files between the clients and servers. However the performance dropped to 330 megabytes per second. Using this set up we were pleased with the ability to create, transfer and store a terabyte of data in 55 minutes.

In the four days of the exhibit, we work to find the cause of the server crash. Internally the filesystem sustained 900 megabyte per second transfer rates. Several SGI technicians and engineers took note of the problem and confirmed that many aspects of their machine OS and filesystem connection were behaving other than as expected or desired. The Server was shared with other demonstrations, however, it didn't appear that they caused the problem directly or indirectly.

6 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
Marty Barnaby	Sandia National Laboratories
James Brandt	Sandia National Laboratories
Wayne Butman	Lawrence Livermore National Laboratory
Helen Chen	Sandia National Laboratories
Martha Ernest	Sandia National Laboratories
Parks Fields	Los Alamos National Laboratory
Jason King	Lawrence Livermore National Laboratory
Ed Klaus	Sandia National Laboratories
Brian Lawver	Lawrence Livermore National Laboratory
Luis Martinez	Sandia National Laboratories
Mark M.. Miller	Sandia National Laboratories
John Naegle	Sandia National Laboratories

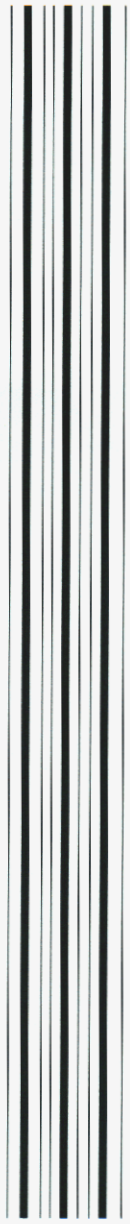
We would like to thank the following individuals for their efforts in making our ASCI's DISCOM *SC 99* networking efforts a success.

SNL - Jim Ang, Esther Baldonado, Forrest "Herb" Blair, Authurine Breckenridge, Joseph Brenkosh, R. Michael Cahoon, Eli Dart, Gary Evans, Stephanie Fellows, Rich Gay, Jerry Gorman, Steve Gossage, Michael Hanna, Richard Hawkins, Steven L. Humphreys, Wilbur Johnson, John H. Naegle, Ron Olsberg, Diana Perea, Lyndon Pierson, Jill Schwegel Tom Tarman, Tim Toole, Steve Valdez, Alan Williams, Ed Witzke, Eli Dart

LANL - Alice Chapman, Jerry Delap, Ann Hayes, Steve Tenbrick, Steve Turpin

LLNL	-	Jean Shuler, Joe Slavic, Dave Wiltzius, Mary Zosel
LBL	-	Wes Bethal, Steve Lau
Y12	-	Rhonda Macintyre
SCINET/XNET	-	Paul Daspit, Jim Deleskie, William "Bill" Wing
ESNET	-	Jim Lieghton, Kevin Oberman
CISCO	-	Mark Bleth; Brad Irwin
Corp Comm.	-	Bob Dobinski
Compaq	-	Ira Grollman
Avici	-	Glen Yallaly, Hank Zannini
Spirent	-	John Clem, Les Kouke, Erik Plesset,
Univ. of Min	-	Ben Allen and his team of video wizards
NTON	-	William Lennon

Appendix A: Parallel FTP Performance in a High-Bandwidth, High-Latency WAN



UCRL-MI-142491

Parallel FTP Performance in a High-Bandwidth, High-Latency WAN

Jason S. King

November 10, 2000

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Background

Within the Accelerated Strategic Computing Initiative (ASCI) program, the Distance and Distributed Computing program, or DisCom^2, is charged with ensuring that Laboratory scientists have the best possible access to computing resources no matter where those resources may be located. Network file transfer has been identified as an important part of this effort. The three weapons labs, Lawrence Livermore, Sandia, and Los Alamos, have been working for some time on plans for a secure, high-speed, low-latency Wide Area Network (WAN) spanning the sites in Livermore, Albuquerque, and Los Alamos. The proposed file transfer tool for this new network is the parallel file transfer protocol (FTP) client as distributed with the High Performance Storage System (HPSS).

This tool was chosen because it is a mature code that has many of the desired capabilities identified by the DisCom^2 program. Among these capabilities are parallel file transfer, compatibility with HPSS, and backward compatibility with "standard" FTP as specified by RFC 959.

Testing at LLNL in early September 2000 showed that, although the parallel FTP code meets the desired capability requirements, its performance in the Local Area Network (LAN) is not quite optimal. Examination of the architecture revealed that a large portion of the performance penalty is directly related to the code's support of the mover and pdata protocols needed to communicate with HPSS. These protocols basically introduce a lock step for each block of data sent across the network. While this did not appear to significantly reduce performance in the LAN, the higher latency found in the WAN would likely result in greatly decreased performance. At least one of these two protocols will always remain necessary for communication with HPSS. They are not, however, required when communicating between two non-HPSS systems, for example, between ASCI White and an SGI visualization platform.

LLNL has demonstrated that, in the absence of HPSS, parallel file transfer can be accomplished with much less overhead and higher performance, even in the LAN. A modification was made to the parallel FTP client and server (non-HPSS server based on the public domain wuftpd code and parallel modifications from M. Barnaby at Sandia) that essentially removed the mover and pdata protocols in favor of a much simpler protocol with 16 bytes of overhead per parallel stripe, per file transfer, with no lock-step mechanism.

Considering the importance of file transfer performance to the upcoming DisCom^2 WAN, it was decided that SC2000, held in early November, would provide a great opportunity for testing the two versions of parallel FTP. For purposes of the discussion in the rest of this paper, the standard HPSS version of parallel FTP, which includes the mover and pdata protocols, is referred to as the "PFTP-hpss," while the modified version is referred to as "PFTP-simple."

SC2000 Network Topology

Source and sink hosts were identified that adequately represent the actual platforms in use on ASCI networks today. Those hosts were an SGI Onyx2 located on the SC2000 show floor in Dallas and an IBM Nighthawk-1 SP node located in Livermore. Each host was connected to the network with four Gigabit Ethernet adapters. Each adapter was placed in a separate Virtual LAN (VLAN), and all traffic was carried across a 2.5 Gb/s OC48c between Livermore and Dallas. The network topology is shown in Figure 1.

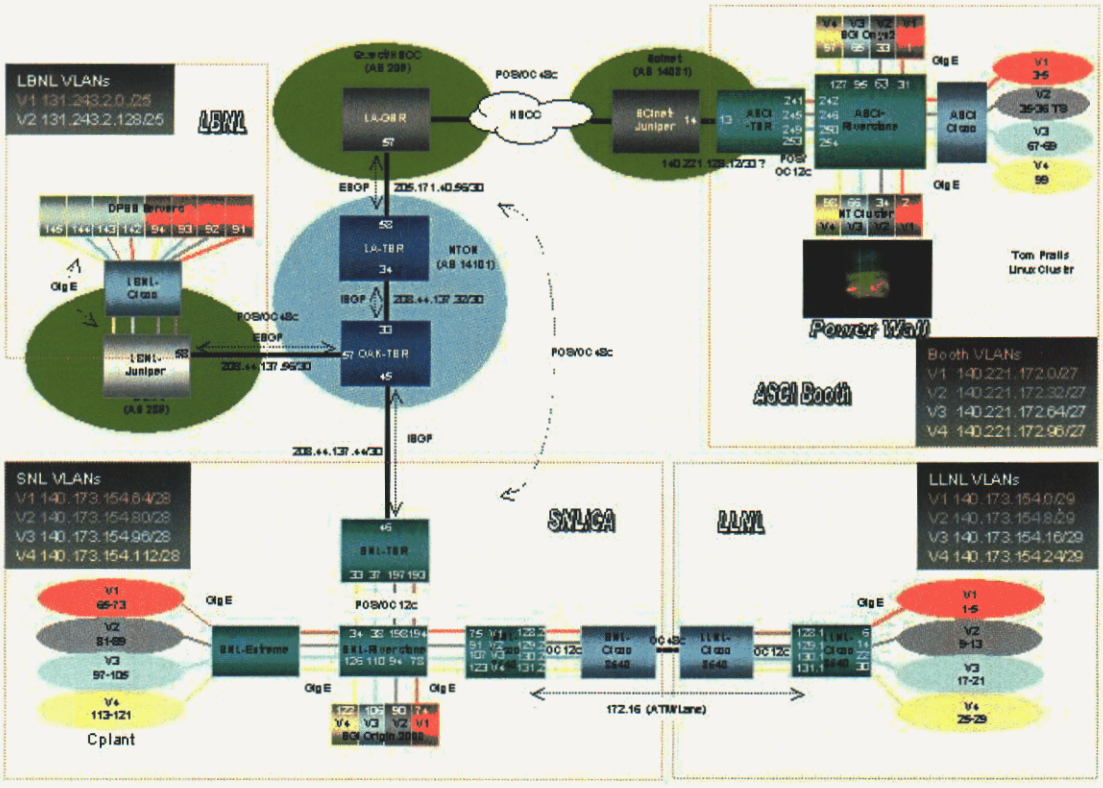


Figure 1. Network topology for testing two versions of parallel FTP.

Tests

The testing methodology was first to use the netperf benchmark tool (http://www.netperf.org) to establish a performance baseline for the network between the source and sink. Because it does not perform disk I/O and strictly measures memory-to-memory copies over TCP and UDP, netperf is a good tool for determining maximum network performance between two hosts. Then, equipped with the knowledge of what was theoretically possible on the network, we looked at how the two different parallel transfer methods performed in comparison.

Iterations of netperf were used to determine the optimal TCP window size to tune the network for maximum performance. With the hosts configured for maximum performance, each transfer method, PFTP-hpss and PFTP-simple, was run through iterations of different parameters, including block sizes and stripe widths. Packet traces of a typical transfer were generated for each method for additional analysis.

Results

The architecture of the network connecting LLNL to the SC2000 show floor was such that, to take full advantage of the bandwidth available, it was necessary to have a minimum of four TCP streams, each destined for a different subnet on the remote end. This allowed traffic to be spread across each of four OC-12 ATM links. Since the round-trip delay was 50 ms and OC-12 bandwidth is 622 Mb/s, the bandwidth delay product of this network is 3.88 MB. Under optimal conditions, the bandwidth delay product should yield the optimal TCP window size, but in this case it did not. Increasing the TCP window size beyond

2 MB had an adverse effect on performance. Unfortunately, the results above 2 MB are not available, but Figure 2 shows that performance plateaus at a 1.5 MB window size. It is believed the poor performance beyond the 2 MB window is directly related to TCP's slow start and congestion control algorithms. Because this network would not run at full bandwidth without losing packets, anytime the TCP window started to approach the network's maximum speed, a packet was lost, TCP's congestion control algorithm activated, and performance declined. TCP appeared to take a very long time to reopen the window once packet loss occurred.

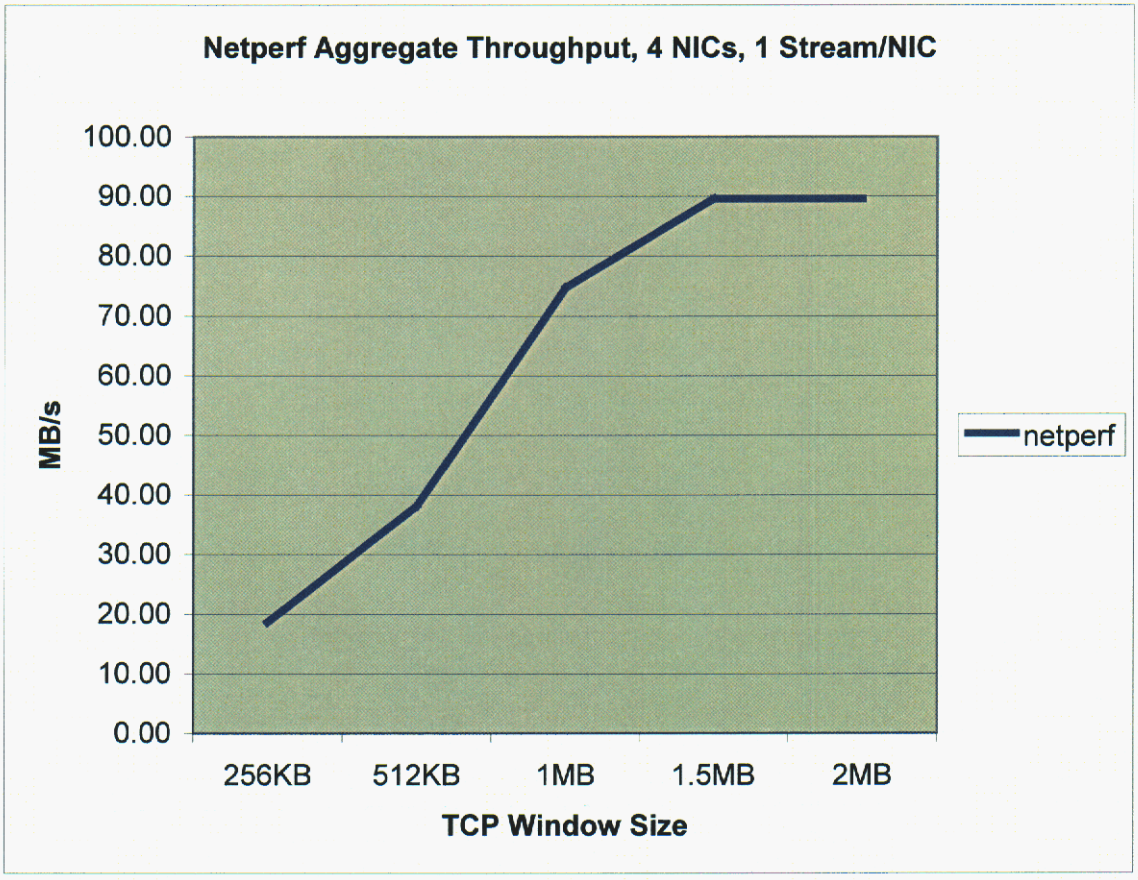


Figure 2. The netperf aggregate throughput performance.

Based upon the results shown in Figure 2, both hosts were configured with socket buffers of 2 MB for the remainder of the tests. Figure 3 illustrates the performance of netperf and the two parallel FTP methods when using 2 MB socket buffers. The performance of the PFTP-hpss method, in general, was half that of the PFTP-simple method.

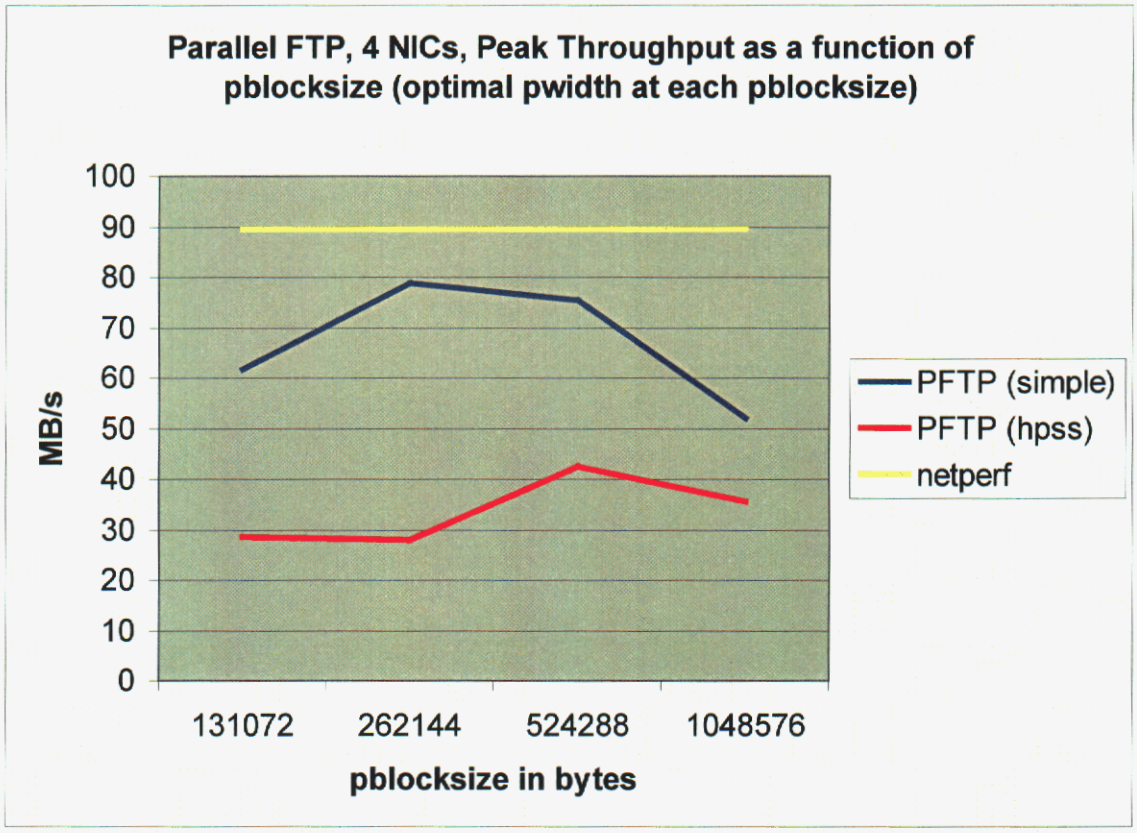


Figure 3. The throughput performance of netperf and the two parallel FTP methods.

Discussion

Figures 4 and 5 show why the performance of PFTP-hpss is sub-optimal. Figures 4 and 5 represent a one second snapshot of throughput from one of four total streams. Figure 4 is a good illustration of the impact that the mover protocol has in the WAN. Because PFTP-hpss operates in a lock-step fashion—sending a mover protocol message, then awaiting the acknowledgement before sending data—large gaps are created where throughput drops to zero. These gaps should be roughly the same size as the round-trip delay of 50 ms. Examination of the packet data bears this out.

Figure 5 shows the same data for PFTP-simple. Note that it also fluctuates greatly over time, but in a more rapid fashion, and average throughput is roughly twice that of PFTP-hpss. The relatively small TCP window used likely explains the rapid fluctuations. Because the chosen window size was small enough to avoid packet loss, it does not allow the network to be fully utilized. Figure 5 seems to support this theory in that slightly less than 50% of the total time shown on the graph is spent idling. Because 2 MB is roughly 52% of the optimal window size of 3.88 MB, we would expect to see approximately 48% of the time spent idle.

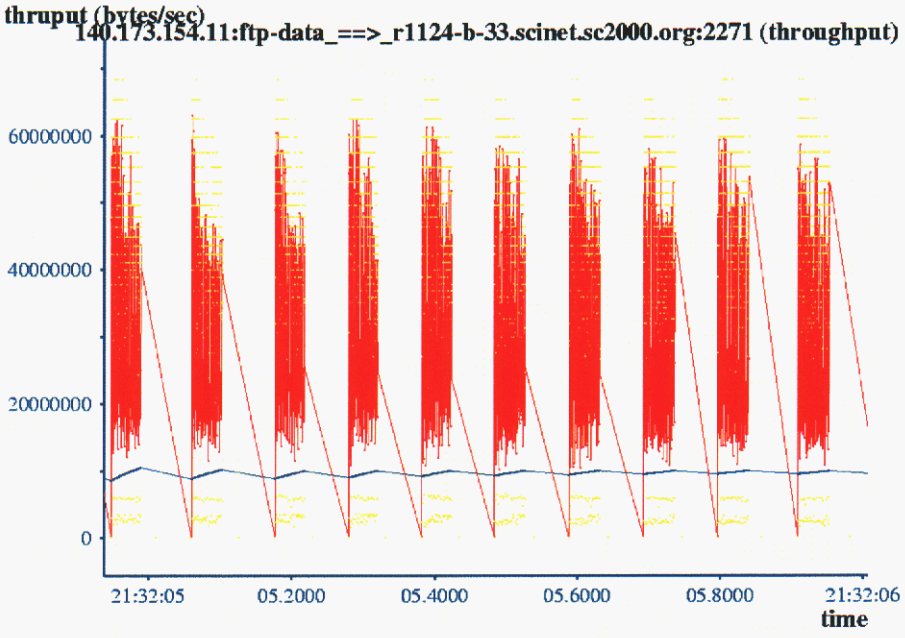


Figure 4. PFTP-hpss, pwidth 4, pblocksize 256 KB.

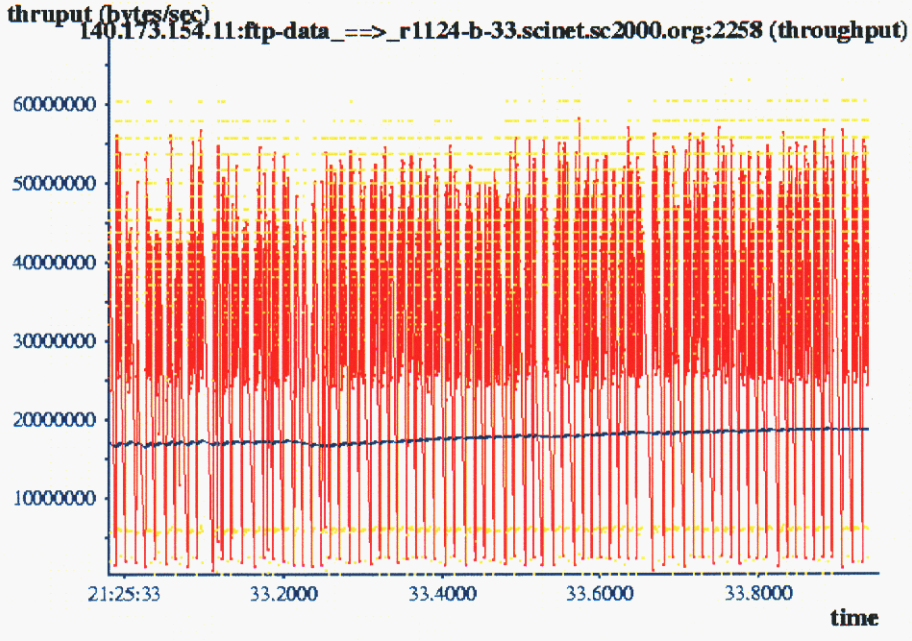


Figure 5. PFTP-simple, pwidth 4, pblocksize 256 KB.

Conclusion

At the least, this work shows that packet loss in the WAN can severely limit throughput. It also shows that there is great room for improvement in our chosen method of file transfer in the WAN and, at least for the

moment, that the largest performance gains in the WAN will likely come from work on the protocols in use, rather than from work on disk or system I/O issues. We need protocols capable of high performance in the WAN before we can expect to fully utilize increasingly high bit rate networks.

Acknowledgements

Special thanks to Bryan Lawver (LLNL) and to Helen Chen and Jim Brandt (SNL/CA) for all their work in setting up this network and babysitting it through a hectic week at SC2000. Also, thanks to the many people at SNL/NM who supplied the SGI and helped with administrative issues.

Appendix B: The SC Conference Networking Cookbook

or

What do you need to know to put on the show?

Introduction

We have been doing remote networking at SC conferences since 1991. The conference and networking have evolved greatly in these ten years. Somethings like getting close interaction with the demonstrators that are doing innovative communications and placing insiders in the SCINET structure are as importance today as they were then. Of course many other things do change through the years. For instance, I don't believe that we will ever see another HIPPI or Hyperchannel demonstration at a later SC conference. First you should understand the general SC conference structure along with the ASCI Structure that put together the booth. For SC, the SC Executive Committee runs the show. Executive committee usually have Sandia Representation. Dona Crawford is currently on this committee. The work of the executive board is broken down into several working groups Of these groups the networking effort is led by the Information Architecture (SCINET) group. The SC2000's Information Architecture (SCINET) group leads are listed below. I've highlighted the people on the committee that have been useful resource in the past.

INFORMATION ARCHITECTURE

CONFERENCE VICE-CHAIR
Bill Kramer, Conference Vice-chair, Lawrence Berkeley National Laboratory / NERSC

Information Architecture Chairs

Eli Dart, Network Security Chair, Sandia National Laboratories
John Dugan, Wireless Chair, National Supercomputing Applications Center/Univ. of Illinois
Rex Duncan, Committee Networking Chair, Oak Ridge National Laboratory
Chuck Fisher, Production Chair, Oak Ridge National Laboratory
Ian Foster, Application Evangelist Chair, Argonne National Laboratory
Doug Luce, Information Management / Customer Support Chair, Aaronsen Group
Jeff Mauth, Physical Infrastructure Chair, Pacific Northwest National Laboratory
Martin Swany, Network Management/Monitoring Chair, University of Tennessee, Knoxville
Tim Toole, Deputy Chair, Sandia National Laboratories
William "Bill" Wing, Experimental Network Chair, Oak Ridge National Laboratory

Information Architecture Committee

Zaid Albanna, MCI
Greg Almes, Internet2
Warren Birch, Army Research Laboratory
Bryan Bodker
Pascal Boudreau, Terabeam Networks
Roberta Bourcher, Lawrence Berkeley National Laboratory
Steve Corbato, Internet2 UCAID
David Crowe, Oregon State University
Paul Daspit, Nortel Networks
Patrick Dorn, National Supercomputing Applications Center
John Dugan, National Center for Supercomputing Applications
Adam Duke, Florida State University
Tom Dunlop, DCC
Larry Dunn, CISCO
Hal Edwards, Nortel Networks

Stacy Eubanks, DCC
Larry Flourney, Internet2, Texas A&M University
Greg Goddard, University of Florida
Jalal Haddad, Oregon State University
Jason Hasse, CISCO
Wendy Huntoon, Pittsburgh Supercomputing Center
John Jamison, Juniper Networks
Steve Jones, CEWES
Wesley K. Kaplow, Qwest
Ed Kempe, Dallas Visitor's Bureau
Tom Kile, Army Research Laboratory
Dave Koester, MITRE Corporation
Bill Lennon, Lawrence Livermore National Laboratory
Paul Love, Internet2
Rick Mauer, Sandia National Laboratories
George Miller, MCI
Bill Nickless, Argonne National Laboratory
Kevin Oberman, Lawrence Berkeley National Laboratory
Willard Ostrander, TeraBeam Networks
James Patton, Caltech
Ben Peek, GST Telecom
Jim Rogers, CSC/Nichols
Jim Ross, Sandia National Laboratories
Philip Schreher, Qwest
Glen Smith, Qwest
Robert Spenser, Qwest
Janae Tinsley, Smart City Networks
Howard Walter, Lawrence Berkeley National Laboratory
John West, WorldCom
David Wheeler, National Supercomputing Applications Center
Linda Winkler, Argonne National Laboratory.

The Experimental Network, XNET, is a new group within SCinet that can be a resource for services that fall outside the production offering of SCINET. Last year Bill Wing, ORNL ran the XNET effort for SCINET. If you are doing network experimentation it is useful to partner with this group.

The ASCI booth is designed and built under the control of a tri-lab, Sandia, LANL, and LLNL committee. A different laboratory assumes the lead role each year. Last year LANL's Alice Chapman had the role. Next year Jean Shuler, LLNL is the lead. In any case a Sandia representative will be on the committee. The committee effort leads to a conceptual design. The conceptual design is taken by a commercial booth design organization that creates the plan from the booth and actual arranges the construction of the booth. For the last several year Corporate Communication has been the company who has done this work. My contact there is:

Bob Dubinski
Corporate Communication Inc,
33 Ship Ave
Medford, MA
(781) 391-1994

The lead role in networking has also be used to assist the booth designer with getting the power needs of the whole booth. In the conference business there are always cost penalty for late requirements. In most years you will never know the whole scope of power requirements when the deadline for the info comes. Be

prepare to estimate the needs of the booth. I always inflated this number. Currently the booth design is based on having a theater with about 9 working demos. The theater, the Discom demo, and networking equipment usually are the major consumer of power. The theater is also the primary customer of the booth's network. The theater is built and manned by the University of Minn. LSCE. Last year the lead on the effort was:

Ben Allen
612) 626-9224
benjamin@lcse.umn.edu

A timetable of the show.

The show usually runs the week prior to Thanksgiving. In the pased this data has varied slightly. It has been moved forward by a week. It has moved backward by two weeks.

2 year prior to the show	Location and Date decided. Lead roles on the conference committees assigned
12 months prior	SCINET initial meeting & Wide Area commitments
6 month prior	SCINET/XNET start planning deployment
1-3 month prior	SCINET/XNET hot stage
July	Show website up
August	LAN and power requirement due from exhibitors to SCINET
October	Network design Complete
2 week before opening	Installation of network infrastructure
Saturday before opening	network drop made to exhibitors

A timetable for the ASCI booth

June	Send out call for participation Monthly meeting General Booth Theme **Network team committed**
July	Graphic started
Aug	Tri-Lab Weekly meeting **Power requirements to booth designers and network request to SCINET**
September	User graphics due - You should have a good idea of what will be presented **Meet with booth designer to ensure Space and Power needs can be met**
Ship Date	Date dependent on conference location from 1 week to 2 weeks prior to show opening **Network Equipment enroute to show**
Wed/Thurs before show opens	Booth construction **Booth Network infrastructure deployed** **Heavy Equipment in place**
Friday	**Implement the networking plan for the booth**
Saturday	**ASAP Get drop request to SCinet ASCI network available by end of day** Hardware exhibitors arrive
Sunday	Software Application exhibitors arrive
Monday 12:00	Construction finished - hall cleared for cleaning
Monday 1700	prepare to open show

Monday 1800	VIP tours
Monday 1900-2100	General Opening Gala
Tuesday-Wednesday 1000 1800	Exhibit Open
Thursday 1000-1600	Exhibit Open
Thursday 1600	Tear down Exhibit
Friday	Trunk depart to return equipment to Labs

I have had requested to include ideas on how to structure the DISCOM networking demo. I have been attempting to get a year ahead of the ASCI networking performance curve. An end goal of ASCI is a 100Gigabyte per second network connecting large machines. Getting equipment to support a demo in this range has provide a inspiration. Choosing to feature a different communication layer each year has kept the demo fresh. Finding willing partners can be difficult initially and you need to trust your partner will commit to the show. If commitment aren't in by August or early September it is time
to worry. You should be willing to ask for assistance from all parties.
As far as advice on what demo to do I can only suggest that you focus on challenges that are of interest to you and your team.

Distribution

10 - 0139 M. O. Vahle, 9900
1 - 0139 P. J. Wilson, 9902
1 - 0310 A. L. Hale, 9220
1 - 0310 J. A. Ang 9220
1 - 0310 P. Yarrington, 9230
1 - 0318 G. S. Davidson, 9212
1 - 0318 P. D Heermann, 9227
1 - 0321 W. J. Camp, 9200
1 - 0328 D. J. Zimmerer, 2612
1 - 0421 R. J. Detry, 9800
1 - 0429 J. S. Rottler, 2100
1 - 0451 S. G. Varnado, 6500
1 - 0785 R. L. Hutchinson, 6516
1 - 0801 M. R. Sjulin, 9330
1 - 0801 W. F. Mason, 9320
1 - 0802 R. A. Haynes, 9227
1 - 0805 W. D. Swartz, 9329
1 - 0806 C. D. Brown, 9332
1 - 0806 J. H. Naegle, 9336
1 - 0806 J. M. Eldridge, 9336
1 - 0806 J. P. Brenkosh, 9336
1 - 0806 L. F. Tolendino, 9336
5 - 0806 L. G. Martinez, 9336
1 - 0806 L. G. Pierson, 9336
10 - 0806 L. Stans, 9336
1 - 0806 M.J. Ernest, 9336
1 - 0806 S. A. Gossage, 9336
1 - 0806 T. C. Hu, 9336
1 - 0806 T. D. Tarman, 9336
25 - 0806 T. J. Pratt, 9336
1 - 0812 B. C. Whittet, 9334
5 - 0812 E. J. Klaus, 9334
1 - 0812 M. J. Benson, 9334
5 - 0812 R. L. Adams, 9334
1 - 0813 R. M. Cahoon, 9327
1 - 0822 A. Breckenridge, 9227
1 - 0826 J. D. Zepper, 9143
1 - 0835 J. M. McGlaun, 9140
1 - 0841 T. C. Bickel, 9100
1 - 1221 P. J. Eicker, 15100
1 - 9003 K.E. Washington 8900
1 - 9003 P. W. Dean, 8903
10 - 9011 H. Y. Chen, 8910
1 - 9011 J. A. Hutchins, 8910
5 - 9011 J. M. Brandt, 8910
1 - 9011 T. J. Toole, 8910
1 - 9012 J. A. Friesen, 8990
1 - 9012 J. N. Jortner, 8990
1 - 9012 R. D. Gay, 8930
1 - 9037 J. C. Berry, 8935
1 - 0188 LDRD Office, 4001
2 - 0899 Technical Library, 9616
1 - 9018 Central Technical Files, 8945-1
1 - 0612 Review and Approval Desk, 9612
for DOE/OSTI