Progress of ATM Customer Trials at Sandia National Laboratories

John H. Naegle, Sandia National Laboratories and
Nicholas Testi, Central Wyoming College

ABSTRACT: *Sandia National Laboratories has had an extensive ATM TCP/IP research effort in place for about three years, the focus of which is to assess the deployment of the technology in a high speed, compute intensive, engineering environment. The current production environment has been reported at previous CUG conferences. It is based on a client-server model providing supercomputing, file storage, and visualization services to local and remote LANs using FDDI and SMDS/ATM technologies. This environment is shown in Figure 1. A demonstration of the target Sandia ATM environment was shown at Supercomputing '93 in Portland, Oregon which included multimedia applications. This network is shown in Figure 2. Our efforts for the last two years have focused on internal ATM customer trials which are structured to test the ATM technologies in production environments using real world engineering applications. This paper describes one customer trial and the introduction of the Cray J90 into this pure ATM environment.*

Objectives of the Sandia ATM Customer Trials

The objective of the ATM customer trials initiated at Sandia is to introduce ATM technologies into engineering groups that have compute/communication intensive production applications in order to verify performance, identify bottlenecks, and test reliability and management. Figure 2 shows the production network schematic for the Engineering Sciences Center. Engineering desktops machines are Sparc 20 workstations running the Solaris 2.4 operating system which access Sparc 2000 compute and file servers and a pair of IBM SP-2 compute servers via Ethernet. A cisco 7000 router provides access to the Sandia Corporate network which also provides computing capabilities via a Cray YMP and an Intel Paragon and graphics services via an SGI Challenge visualization server running AVS. The ATM version of the ESC was constructed by installing ATM technology into ESC machines which provided a parallel pseudo-production environment. All applications currently utilizing the production Ethernet environment are available to be tested over the parallel ATM environment. The ATM network for the ESC is shown in Figure 4. The testbed features OC-3 (155.51 Mbps) ATM/SONET connections to the client engineering Sparc 20 desktops. Typical workstation configurations would include 64 megabytes of memory. The Sparc 20 machines have ATM NIC cards from Interphase installed, using multimode fiber to connect to ATM switches. The SP-2 server machines each have a single OC-3 ATM/SONET multimode connection using NIC cards from Fore Systems. The SGI Challenge machine is connected using a Fore Systems VME OC-3 multimode NIC. Edge devices in the testbed include the cisco 7000 router with an OC-3 multimode ATM interface to provide connectivity to production legacy LANs. The Cray J90 machine was recently connected to the testbed in order to measure the performance of the machine in a pure ATM environment. The J90 is running Unicos 8.04 which includes the ATM driver code supplied by Cray. The testbed utilizes Bay Networks Lattiscell 16x16 ATM switches, most of which are configured with 14 multimode ports and 2 single mode ports. Future implementations of ATM networks at Sandia will standardize on single mode ports for NNI connections between switches.

At the time of the tests, full interoperability of the vendors' Q2931 signaling code for switched virtual circuits (SVCs) had still not been achieved. Consequently, the testbed utilizes Permanent Virtual Circuits (PVCs) for connections between all devices, including the IP router. As has been pointed out many times, the management task for even a relatively small network such as the testbed is quite extensive, given the size of the PVC matrix required. Since we are utilizing a client-server model in the testbed, we have reduced the size of the PVC matrix by not providing connectivity between the desktops; therefore the matrix is not fully meshed. It is nonetheless an error prone management task to make changes in network connectivity. Each UNI device in the network must have an table entry for every other UNI device to which a connection is desired. This table entry maps a unique VPI/VCI designator pair onto each destination IP address. In addition, switches must have a VPI/VCI designator pair for each port pair in order to correctly route the cells through the ATM switch fabric. This PVC specification is done through a connection management station . In the case of the Lattiscell switches, only the source and destination ports must be referenced, as opposed to having to configure

PVCs through intermediate switches when multiple switch hops are required.

Some Comparative Memory-to-Memory Metrics

In order to establish some benchmarks for the ATM performance of the J90, the familiar TTCP memory-to-memory metric code was used. Network speeds for all the other computers in the testbed had been established previously using the TTCP code, and all TCP/IP connections were tuned for maximum performance by using a 64 Kbyte window size. Table 1 shows a throughput breakdown by client-server(s).

Table 1. TTCP throughput by Machine Pairs in Megabits Per Second

	Sparc 2000 (r)	IBM SP-2 (r)	SGI Challenge (r)	Cray J90 (r)
Sparc 20 (t)	50	118	80	60

All measurements were done on dedicated machines, i.e., no other substantial processing or network traffic was occurring at the time of the measurements. We believe we have achieved maximum throughput from these machines, based on TCP/IP tuning experience and we expect that these numbers are repeatable by other testing sites if all experimental parameters are the same for any given pair of machines in the table. Deviation from these figures could occur from differences in machine configuration (memory, etc.), differences in load on the machines, or differences in the operating system or driver code versions. The maximum throughput achieved in the testbed is 124 Mbps between the SP-2 machines. This is approaching the theoretical limit (signaling rate minus overhead) of the OC-3c rate. The low rate for the Sparc 2000 relative to the Sparc 20 is due to the CPU differences in the machine; the Sparcs are model 61 machines while the Sparc 2000s are model 40 machines.

TCP/IP Application Metrics

These memory-to-memory tests result in throughputs that are much faster than are achievable using common TCP/IP applications, such as NFS, which is the most used application in the ESC. For example, the Sparcs utilize NFS Version 2 which only supports up to an 8 KByte window, resulting in a throughput maximum of only 3 Mbps. NFS version 3 under Solaris 2.4 utilizes a 64 Kbyte window which results in throughputs of 27 Mbps. We have not had the opportunity to test NFS on the J90 or the Challenge, but the general observation that larger window sizes increase throughput for TCP/IP applications will be applicable to these machines as well. This will be empirically verified as time permits.

The FTP application uses the system default window size of the host workstation, which means the default window size must be set to the maximum the host operating will allow in order to achieve maximum throughputs. At that point, the host disk speed becomes the limiting factor. We have observed FTP throughputs that range from 28 Mbps on the Sparcs, 48 on the Challenge and the J90, up to 74 Mbps on the SP-2.

A Sample Engineering Application

We have included a test scenario which represents a likely production use of the testbed. Our customer group in the testbed typically execute compute intensive simulation codes, such as finite element analysis which models the behavior and physical properties of various materials subject to physical stress. The results of these simulations are typically displayed using products like AVS, allowing the engineer to view the model at different times on the simulation clock. This application was demonstrated using the SP-2 as the compute server but will be repeated in the next battery of tests using the J90 as the compute server. In our application, the simulation data file was transferred from the Sparc file server to the SP-2 compute server which executed the finite element simulation code. The data resulting from the simulation code was then transferred to the AVS server, which rendered the data and transferred the images back to the engineering desktop machine for display. The engineer controls the simulation clock and is able to continuously spin the visualized model in order to view the model from all sides. This desktop workstation received a continuous 32 Mbps of data from the AVS server. This same experiment dies in place when attempted on the Ethernet network, as the transmission speeds required are greater than 10 Mbps.

Summary

The customer trial network represents a complete supercomputing environment utilizing ATM technologies from the engineering desktop to the server machines, including the J90, at OC-3 rates. The client-server model implemented in this trial works well in the pure ATM environment. The maximum memory-to-memory performance is approaching the theoretical OC-3 limit (approximately 130 Mbps) on the IBM SP-2s but varies widely across the other machines in the testbed. In general, we have found that NICs and drivers from the various vendors in the ESC ATM network are extremely sensitive to platform, operating systems, and patch level differences. For example, we have observed a 30% increase in throughput for the Sparc machines after migrating from Solaris 2.3 to Solaris 2.4. The real-world application performance is still limited by protocol stack and disk speed on all machines in the testbed. The J90 beta driver we tested was easy to configure and has been stable. The throughputs achieved for the J90 are slow relative to other machines in the testbed but we have observed that most vendors of ATM drivers are capable of optimizing the code for higher throughputs in subsequent versions. Future work in the customer trials include the introduction of the Intel Paragon and the High Performance Storage System (HPSS) via ATM/OC-3 interfaces.

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

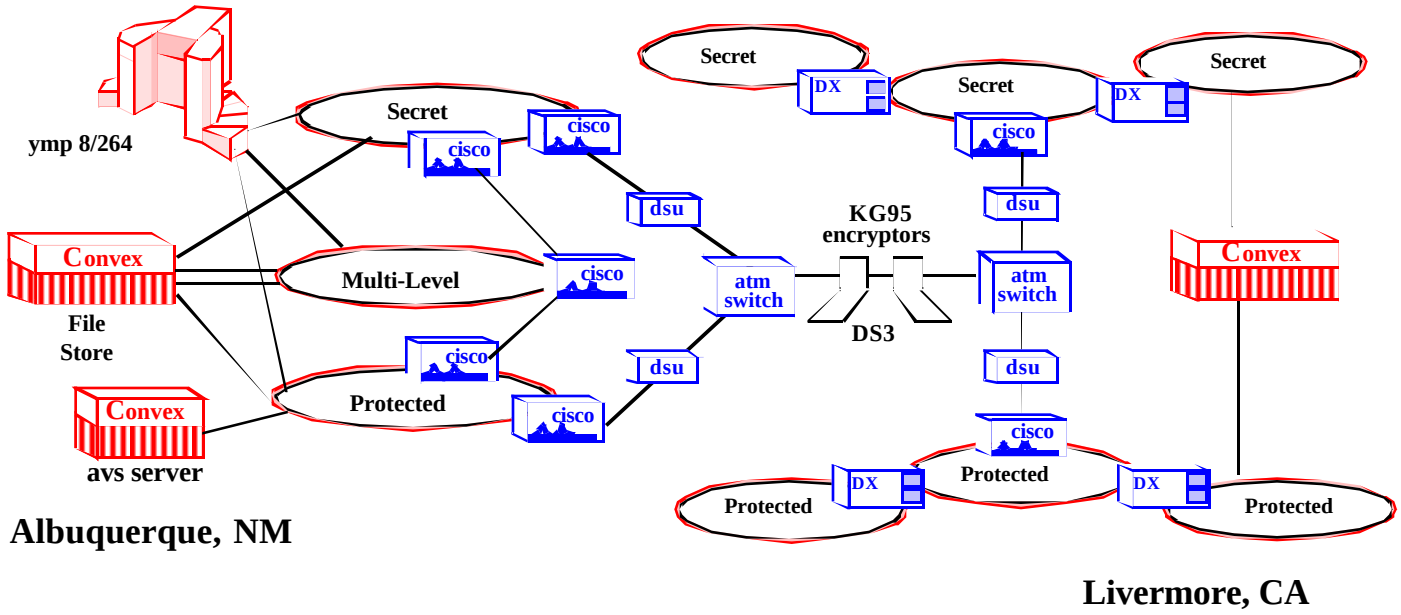

Figure 1: Supercomputer Consolidation Production Network.

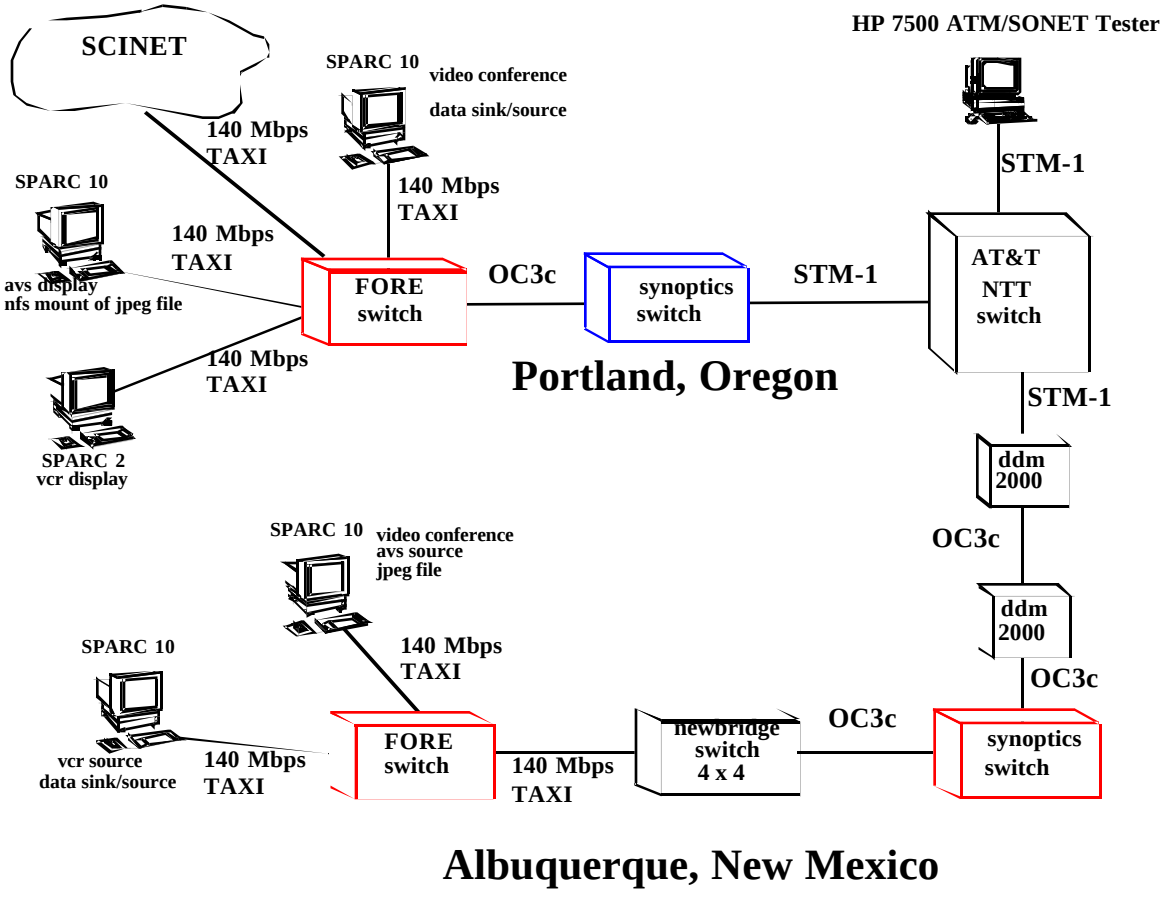

Figure 2: SC '93 Network.

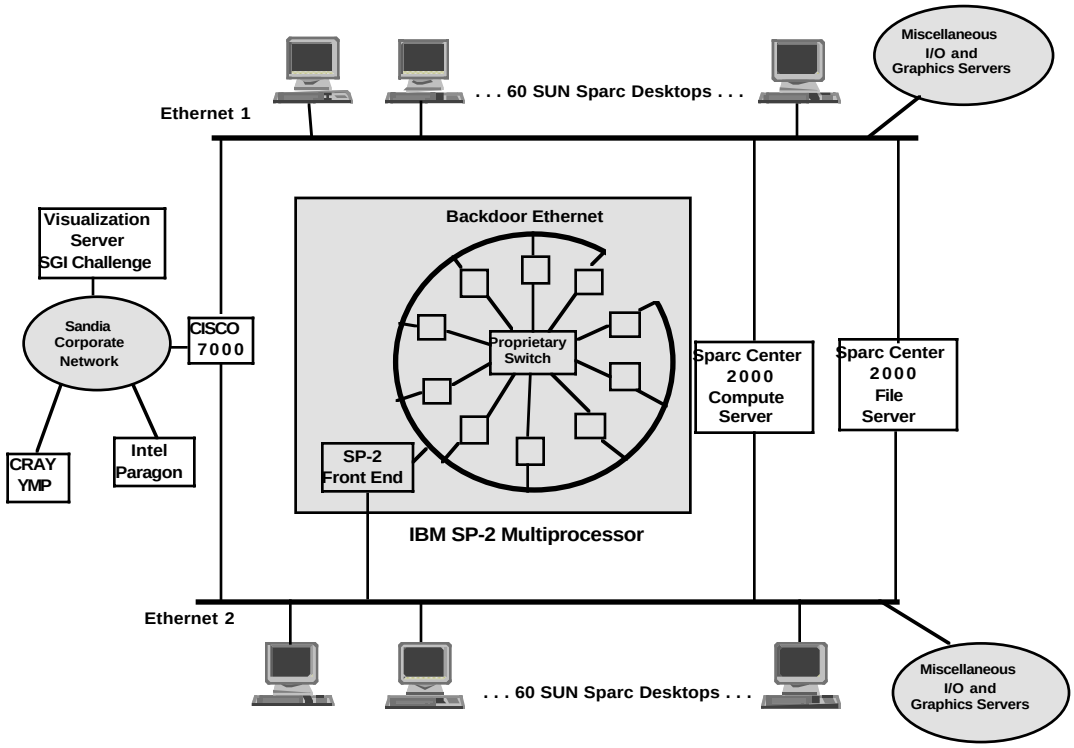

Figure 3: Engineering Sciences Center (ESC).

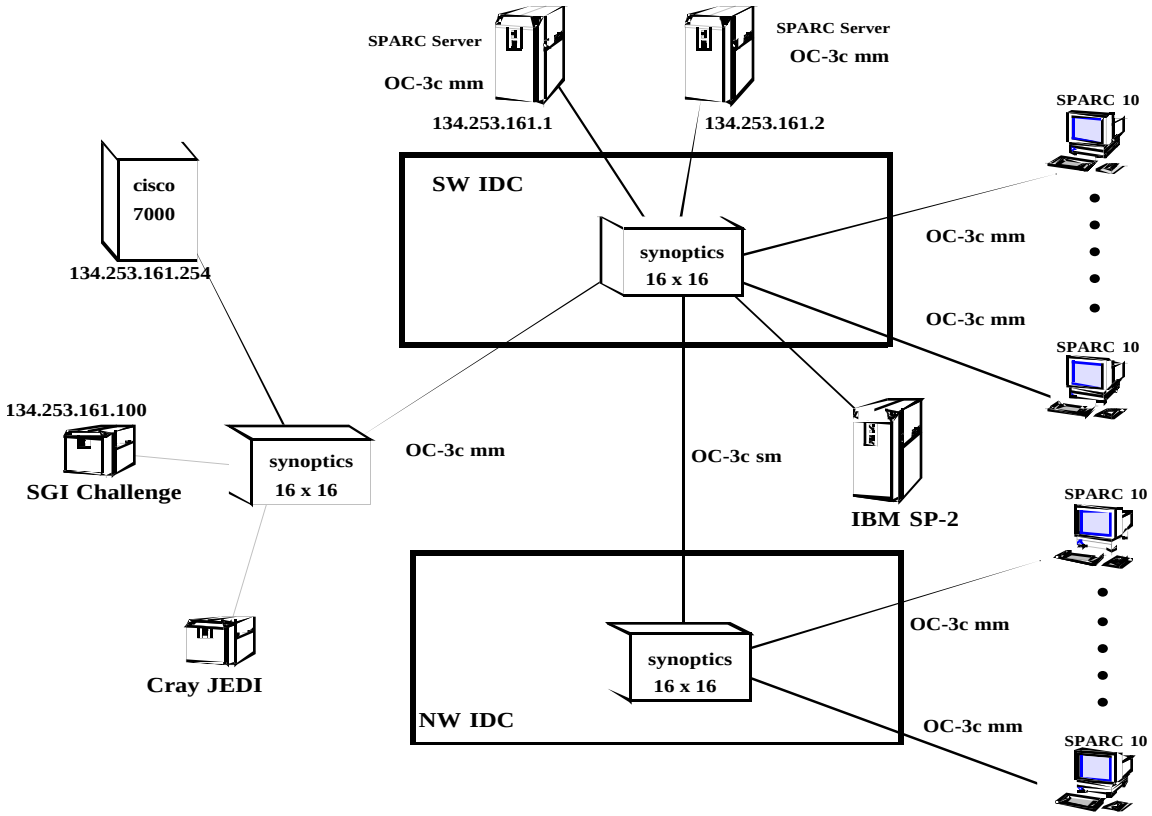

Figure 4: Complete ESC ATM Environment.