Some Early Experiences with DCE/DFS and HPSS at Sandia National Laboratories/NM

Barbara Jennings, *Glenn Machin*, *Pat Moore*, and *Walter Vandevender*,[1] Sandia National Laboratories

ABSTRACT: *With the Accelerated Strategic Computing Initiative (ACSI), there is an increasing demand to share information and computing resources among the DOE National Laboratories. DCE and DFS provide enhanced security access controls and data integrity features that seem ideal for this networked computing environment. Using a DCE test cell, the authors present some experiences with providing both DCE/DFS and High Performance Storage System (HPSS) services to internal and external users of some Sandia National Laboratories computing resources.*

[1] This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

1 Introduction

Sandia National Laboratories (SNL), a US Department of Energy laboratory, works with the US Defense Department, the National Science Foundation, NASA, and industry to develop High Performance Computing (HPC) technologies and apply them to nationally important problems. Its mission includes national security, industrial competitiveness, energy resources, and environmental quality. At Sandia, computers are being used to design and optimize materials ranging from catalysts to optoelectronics, and simulations are replacing tests and experiments that are environmentally unacceptable or prohibitively expensive.

2 ASCI

The US commitment to ending underground nuclear testing, constraints on non-nuclear testing, and loss of production capability call for new methods of verifying the safety, reliability, and performance of the US nuclear stockpile. One of these methods is compute-based virtual testing and prototyping of nuclear weapon systems. The Accelerated Strategic Computing Initiative (ASCI) is one element of DOE's Stockpile Stewardship Program, designed to advance DOE/Defense Programs computational capabilities to help meet the future needs of stockpile stewardship. ASCI will create the leading-edge computational modeling capabilities that are essential for maintaining the safety, reliability, and performance of the US nuclear stockpile in the absence of underground testing.

3 Networking Infrastructure

Sandia embraced asynchronous transfer mode (ATM) as a networking technology several years ago and has been very active in deploying and promoting this technology for both the wide-area and local-area networks. As the only distributed DoE laboratory, with facilities in Albuquerque, New Mexico, and Livermore, California, Sandia requires long-distance, high-speed secure communications. The need to conduct large-scale simulations over very long distances has driven work in high-speed encryption/decryption. Moreover, Sandia is a member of AT&T's experimental university network and the Bay Area Gigabit Testbed, and has partnered with communications companies to help form the National Information Infrastructure Testbed.

Figure 1 shows a high-level view of the networks at Sandia, New Mexico. The central facility, which will house SNL's newest supercomputer (Teraflop), the High Performance Storage System (HPSS), and other major equipment, is connected to an AT&T GlobeView high capacity ATM switch, national networks, and Sandia's own internal networks. These internal networks are connected through technical control centers (TCC) via single-mode fiber and synchronous optical

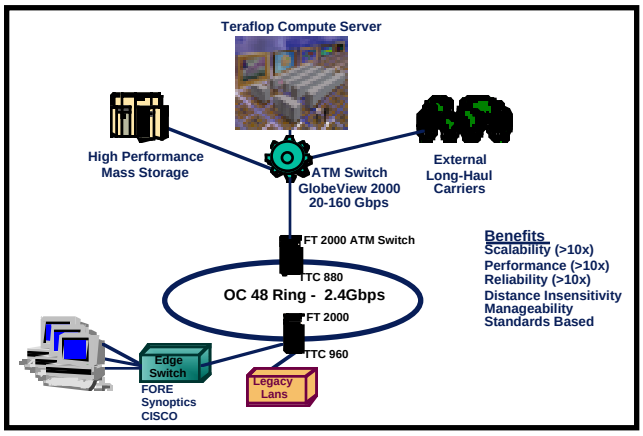

Figure 1: Wide- Local-Area Networks

network (SONET) switches. The entire Sandia campus, which spans about 7 miles, is connected via the TCCs. Each TCC provides connections to the backbone and a variety of ATM edge switches. The edge switches, from a number of different vendors, provide connectivity to devices with ATM interfaces and legacy FDDI/ethernet local area networks.

Sandia's open network is divided into the external open environment (EON), and the internal restricted environment (IRN). The IRN is behind a screening router firewall. Most SNL workstations are located in the IRN for protection against hacking from the Internet. All sensitive and proprietary data at Sandia is in the IRN. ASCI resources must be available to hosts both inside and outside the IRN/EON firewall.

4 SecureNet Network

The SecureNet computing network provides connectivity among various participating DoE sites to enable the sharing of computer resources and classified data in order to facilitate research and development efforts. Using Motorola NES encryption units attached to the DOE Energy Sciences Net (ESNet), this network has received DOE approval for classified use.

5 Distributed Computing Environment

As the Laboratories move from a facilities-based test environment to one with much greater dependence upon computational simulation, the need for an integrated computer environment that provides a single system image is most important. This single system image will provide: 1) a user-friendly environment where one-time authentication propagates throughout the campus, 2) a robust network queuing system where jobs are routed to the appropriate HPC resource, 3) a mechanism that enables critical applications to obtain shared resources, 4) a network infrastructure that enables compute, visualization and storage servers to scale as the complexity of the applications scale and 5) security measures designed to provide a high degree of interoperability while meeting DoE requirements.

DCE, the Distributed Computing Environment from the Open Software Foundation (OSF), is a software environment for supporting distributed computing. It is a combined set of tools, protocols, and methods that support the use and maintenance of distributed applications in a heterogeneous networking environment. DCE is supported on most major workstation platforms as well as supercomputers, midrange processors, desktop systems, and personal computers. DCE version 1.1 is being used for this project.

DCE is considered a middleware technology. Applications are written using a rich set of Application Program Interfaces (API's) to take advantage of the DCE services. DCE is composed of a set of services that can be used, separately or in combination, to form a secure distributed environment. Each of these services is based upon a secure remote procedure call (RPC) capability. The primary services for DCE are a threads service to process multiple simultaneous RPC requests, a Security Server or DCE Security registry based on Kerberos, a directory service called the Cell Directory Service (CDS), and a time service (DTS) used to synchronize clocks. The data-sharing services of the Data File System (DFS) are built upon the secure core services. It enables end users to access data transparently from heterogeneous platforms and to control access to their data via Access Control Lists (ACLs) and groups.

5.1 DCE Cells and Platforms

The ASCI project is supported among the three national labs: Los Alamos National Laboratory (LANL), Lawrence Livermore National Laboratory (LLNL), and SNL using DCE 1.1 or later revision in a multicell configuration. In DCE, each administrative domain /namespace is called a cell. Each site will have its own primary DCE cell (dce.*sitename*.gov).

SNL has one DCE cell in production on its open network, and one in the process of being established for its classified network. SNL's "Open" DCE cell configuration currently consists of two security servers and four CDS servers, running on 5 systems. (Figure 2) These systems also act as global time servers, obtaining their time from Sandia's NTP providers. The two security servers are running under HPUX 9.05 using DCE1.1 provided by HP. The CDS servers are running under HPUX 9.05, Solaris 2.5 and AIX 4.1, using DCE1.1 provided by HP, IBM, and Transarc.

5.2 Restricted Access

SNL's DCE cell spans both the EON and IRN. For this to work, all DCE systems are required to run DCE services using UDP ports in a restricted range. DCE services can be forced to use UDP ports within a restricted range through the settings of two environment variables. RPC_SUPPORTED_PROTSEQS can force DCE RPC's to use UDP, while RPC_RESTRICTED_PORTS can define both UDP and TCP port ranges. The key to using these restrictions is proper configuration of systems and login accounts to automatically set these environment variables. Slow response and time-outs resulted when systems were not properly configured.

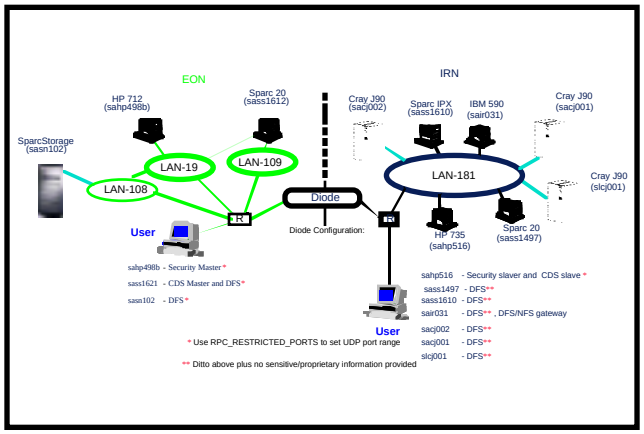

Figure 2: SNL "Open" DCE Cell

During testing it was found that neither of the vendors of DCE for Windows NT (Digital Equipment Corp. and Gradient Technologies, Inc.) correctly implemented their runtimes to adhere to the restricted ports environment variable. Gradient has logged a defect report and special (but temporary) firewall modifications have been found that appear to support the Transarc DFS client on top of Digital's DCE for NT.

5.3 Cross Cell Access

As part of the ASCI computational environment, it is desirable that users authenticate in their local DCE cell but be allowed to access resources controlled by other DCE cells. In order to implement this functionality, cross cell trust has been established between the cells at SNL, LLNL, and LANL, allowing each site to permit or deny access to cell resources by principals authenticated at remote sites. Other cells that will be participate in cross-cell trust relationships include: Argonne, NERSC, Allied Signal, Oak Ridge National laboratory. Pantex and others in the ESNet community will be active in establishing trust with the three ASCI Lab sites.

5.3.1 Cell Names

To facilitate intercell communications, cells at each site (e.g. dce.sandia.gov) must have an associated Domain Name Server (DNS) machine resource record in a local DNS nameserver. This record must contain the hostname of the system where the CDS server resides.

5.3.2 Network Names to Local Name Mapping

DCE account information contains identity information such as the user's Unix UID. However, when a user crosses from one cell and uses resources in another cell, it is possible that conflicts in Unix UIDs may exist. Therefore some mechanism needs to be in place for a "foreign" user to be assigned a Unix UID and username for an account in the local cell. At SNL, DCE Extended Registry Attributes (ERA) are used for such assignments. Each user at a foreign cell who wishes to have access to a Sandia resource is assigned a local account which reflects the foreign user's network name. For instance, the user bob at the foreign cell dce.lanl.gov will have an account name in the dce.sandia.gov cell of "bob%dce.lanl.gov". A Unix UID is assigned for this account, and an ERA is created, LOGINNAME, which is set to be this user's name for any system within the dce.sandia.gov cell. This account is configured so that bob will not be able to perform a *dce_login* (or *kinit*) directly into the bob%dce.lanl.gov account.

Finally the *telnetd*, *ftp* daemons and *SSH* daemons have been enhanced so that when a user attempts access using Kerberos credentials, or by logging in without a fully qualified network name (bob@dce.lanl.gov or bob%dce.lanl.gov), the daemon verifies with the DCE registry the existence of this cross cell account and logs the user in under the name assigned in the LOGINNAME ERA. This enforces the DCE model, where users authenticate within their own cells and are authorized by Access Control Lists (ACL's) that apply to foreign users.

5.4 Distributed File System (DFS)

DFS (a file system created for the DCE infrastructure) is a Distributed File System that enables file sharing across platforms. Users access files from any DFS client using a consistent global directory structure. Because of local caching, repetitive file access is often available without network delay. DFS provides POSIX semantics for easy file access. The user logs into a DCE cell, then accesses DFS files in the same way the user would access other files on that machine's native file system. Existing applications need no modifications to access DFS files. The user can protect DFS files appropriately using ACL's.

At SNL, with the use of restricted ports, DFS servers within the IRN can be made inaccessible outside of the IRN simply by not setting the RPC_RESTRICTED_PORTS environment variable. This is ideal for those servers handling sensitive information.

SNL currently has four production DFS systems, two Sparc20 machines running Solaris 2.4, one SparcIPX running Solaris 2.5, and one IBM 590 running AIX 4.1. During the fall of 1996, several Cray J90 computers will be added to the production DFS systems. A machine in the EON will include public filesets visible to both the open and to the restricted environment. Another machine will contain private filesets visible only to hosts inside the IRN.

5.5 Kerberos 5.2 to DCE Transition

Sandia has employed the Kerberos authentication/verification system for several years. This system provides a rudimentary mechanism for multi-system access. However, Kerberos was an afterthought applied to our then-current network design. With DCE, Kerberos is integrated into all aspects of DCE services that require authentication. SNL expects to be able to provide access to all services for every customer via a single DCE login associated with that customer. This will greatly simplify customer access to Sandia DCE computing resources and improve security by requiring customers to retain only one password per network (unclassified and classified network will remain completely separate).

SNL is currently using enhanced Kerberos V5.2 in the open and secure environments, and will continue to use these servers and clients in parallel with DCE 1.1 security servers. In order to operate in this parallel mode, SNL has modified the Kerberos *kadmind* and *krb5_edit* to update the DCE registry at the same time that the Kerberos Key Distribution Center (KDC) is updated. Plans are under way to migrate SNL Kerberos V5.2 clients to Kerberos V5.6. This will offer better cross cell capabilities and fit better into a DCE environment.

6 High Performance Storage System (HPSS)

Sandia, in partnership with other DoE labs, NASA laboratories, and IBM has developed software systems to manage the next generation of mass-storage technology. Sandia's High Performance Storage Facility (figure 3), HPSS, was on-line in the 4Q of 1995. The initial system consists of an IBM 3494 tape archive (27 TBytes capacity without compression) and 8 IBM 3590 tape drives. The system is controlled by 4 IBM RS6000/5xx workstations with OC3c (155 Bit/sec) interfaces. As a network attached device, the HPSS system was connected to the Paragon with OC12 (622 Mbit/sec) and HiPPI interfaces. During the fall of 1996, two additional HPSS systems will be added to provide storage for the Teraflop supercomputer.

HPSS is a new generation storage system that provides storage of extremely large amounts of data (petabytes = 10^{18} bytes) and provides access to this data at very high data rates (10s to 100s Mbytes/sec). The HPSS system supports standard access to files via FTP and NFS and allows higher speed accesses through a parallel FTP HPSS utility. In addition, DCE-based machines can use POSIX compliant API to develop their own specialized access to the HPSS data.

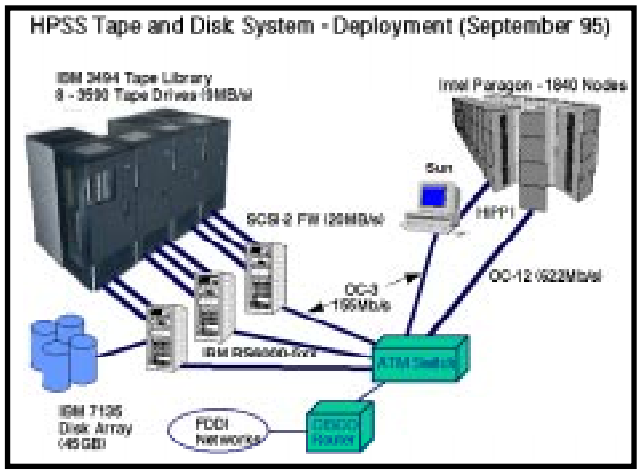

Figure 3: Initial High Performance Storage System