

Beyond the Storage Area Network: Data Intensive Computing in a Distributed Environment*

Daniel Duffy, Nicko Acks, and
Vaughn Noga
Computer Sciences Corporation
daniel.q.duffy@gsfc.nasa.gov

Tom Schardt, J. Patrick Gary,
Bill Fink, and Ben Kobler
NASA Goddard
tom.schardt@nasa.gov

Mike Donovan, Jim McElvaney,
and Kent Kamischke
Silicon Graphics Inc.
miked@nccs.nasa.gov

Abstract

NASA Earth and space science applications are currently utilizing geographically distributed computational platforms. These codes typically require the most compute cycles and generate the largest amount of data over any other applications currently supported by NASA. Furthermore, with the development of a leadership class SGI system at NASA Ames (Project Columbia), NASA has created an agency wide computational resource. This resource will be heavily employed by Earth and space science users resulting in large amounts of data. The management of this data in a distributed environment requires a significant amount of effort from the users. This paper defines the approach taken to create an enabling infrastructure to help users easily access and move data across distributed computational resources. Specifically, this paper discusses the approach taken to create a wide area Storage Area Network (SAN) using the SGI CXFS file system over standard TCP/IP. In addition, an emerging technology test bed initiative is described to show how NASA is creating an environment to continually evaluate new technology for data intensive computing.

1. Introduction

The Computational and Information Sciences and Technology Office (CISTO)[1] at the Goddard Space Flight Center (GSFC) in Greenbelt, MD, provides networking, security, high performance computing, and information technology research and development for the Earth and space science community. In particular, this Office provides High Performance Computing (HPC) services on a wide variety of platforms (HP, SGI, Linux

clusters, Sun) and petabyte-capacity data storage to a diverse set of user groups. The mission of CISTO is to enable Science Mission research by providing its user community with access to state of the art HPC and data systems for computational modeling.

The NASA Center for Computational Sciences (NCCS)[2] operates and manages the HPC and mass storage archives for CISTO. The major compute systems include:

- HP Alpha Server ES45 with 1,392 processors,
- SGI Origin 3800 complex with a total of 640 MIPS processors configured into multiple single system images, and
- SGI Altix Bx2 with 128 Itanium2 processors.

The total peak computational capability achievable by the NCCS resources is over 4 TFlop/s. Each of the HP and the SGI systems are tightly coupled through a Storage Area Network (SAN) to their own fibre channel-attached storage arrays. In addition, certain SGI file systems are NFS mounted to the HP for user convenience.

2. NCCS Strategic Initiatives

Due to the very large data sets required for modeling in Earth and space science and the requirement to readily share this data, the NCCS has developed the following strategic initiatives.

- *Application Portals and Expertise on Leadership Class Systems:* Provide NASA owned and managed systems to allow application development and porting to other leadership class systems, such as Project Columbia.[3] While maintaining current proficiency in Earth and space science applications, this initiative will continue to grow valuable application expertise within NASA.

* This research was sponsored by the NASA Center for Computational Sciences, Computer Sciences Corporation (CSC), and Silicon Graphics Inc. (SGI). The views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of NASA, CSC, or SGI.

- *Capacity Computing*: Provide platforms to satisfy the Science mission modeling initiatives that require sheer bulk capacity. These types of applications are typically more latency or bandwidth tolerant and do not necessarily scale to large numbers of processors.
- *Large Computational Platforms (Capacity Computing)*: Strive to maintain the necessary computational capacity for large scale Earth and space science applications. In addition, the NCCS can be a business continuity center for mission critical applications currently running on Project Columbia.
- *Operational Support*: Strive to continue to support missions and projects for Earth and space sciences that require highly available resources. These types of missions levy additional requirements on the IT infrastructure that go well beyond those resources found in production computing.
- *Data Management and Data Fabric*: Endeavor to facilitate distributed computing and analysis through the creation of a data fabric from a NASA enterprise wide perspective.

This paper focuses on the current projects within the NCCS to achieve the data management and data fabric strategic initiative. The approach to achieving this goal is to encompass the compute engines with both an underlying enabling infrastructure and an overarching service layer. In the following sections, sample technologies for both the service and infrastructure layer will be discussed.

3. Distributed Data Environment (DDE)

Over the past year, the NCCS has been actively prototyping a Distributed Data Environment (DDE)[4] to provide users with the beginnings of an abstraction layer for mass storage (see Figure 1). The basis of the DDE is the Storage Resource Broker (SRB)[5] along with custom created tools.

The system allows users to maintain the underlying access to mass storage systems over the SAN or through well understood and standard protocols (like secure copy). After files are saved into the Hierarchical Storage Management Systems (HSMS), users can register files into a Metadata Catalogue (MCAT). Information beyond the inode can be saved as metadata about the data, which may include information about how to read the data or even, for example, data expiration dates.

Once registered into the MCAT and access controls set at the MCAT level, data sets may be accessed through a variety of web/grid based tools. Currently, the NCCS has

enabled files to be accessed through a simple web interface and Distributed Oceanographic Data System (DODS).[6]

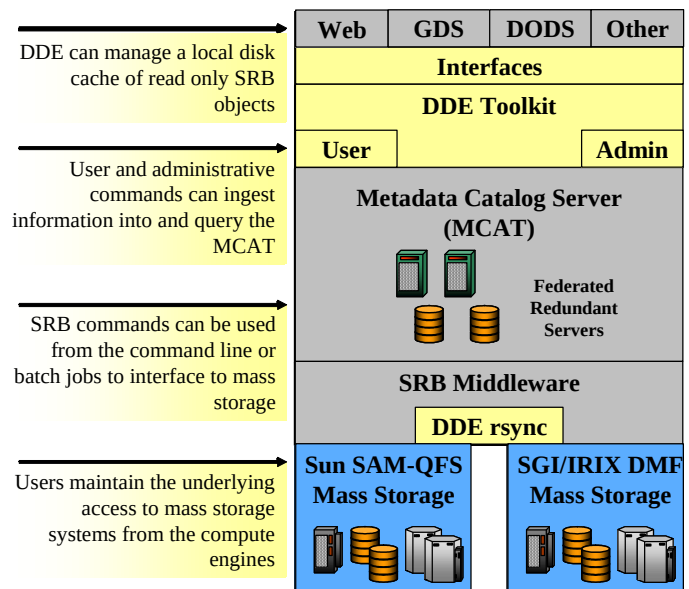


Figure 1. Distributed data environment architecture.

As an enhancement to the SRB set of tools, the NCCS is currently exploring the installation of a Distributed Data Server (DDS). This platform will have a set of trusted services running that will be able to reach into the NCCS mass storage systems to retrieve register files. For example, a data set stored on one of the mass storage systems can be registered into the MCAT. If the appropriate access controls are set, users from outside the NCCS may be able to access these data sets through the DDS.

At the first request for a data set, the data is retrieved from the HSMS and stored locally on the DDS. Subsequent requests for this data set are served directly from the local cache. A synchronization tool has been created to make sure that the data set requested has not been modified on the underlying HSMS. If the data has been modified, a new transfer is made.

In addition to enhancing user accessibility to data while minimizing the impact on the HSMS resources, the metadata catalogue can be used to provide administrative tools to manage data within an environment. By storing metadata within a database, database queries are used to recover various metrics about the mass storage holdings. As an example, database queries can be used for anything from accounting purposes to queries about the actual data holdings in order to find relevant data sets. The NCCS is currently prototyping expiration dates set for files within the MCAT. As an expiration date for a file is reached, a given policy for that data file can be implemented. This

could mean anything from the simple deletion of the file to sending notifications to the owner.

4. Storage Area Networks

In addition to the abstraction layers to assist in the accessibility and management of data, additional technologies can be leveraged at lower levels to facilitate a data fabric.

Originally, the SGI storage systems were direct attached to each of the system images and NFS mounted out to the rest of the cluster. However, the performance, reliability, and availability of NFS were not sufficient to meet the operational requirements for the NCCS users on the SGI systems. In the summer of 2004, the file systems on the SGI cluster were reconfigured into a SAN using the Clustered Extensible File System (CXFS) from SGI. A simplified CXFS installation diagram is shown in Figure 2.

CXFS uses Metadata Serves (MDS) to control the necessary locking and unlocking of blocks on the disk arrays for the file systems under CXFS management. The MDSs are installed on Origin 300 systems and store all file inode information for any file managed by CXFS, and multiple servers are used for fail over and redundancy. CXFS clients are then installed on the compute platforms to allow these platforms to mount and use the file systems. Whenever a client requests access to a file in a CXFS file system, the client first contacts the MDS for information as to the physical location of the file and its properties. If necessary, the MDS locks the file for write by a single client and sends a response back to the client. Once the response has been received, the client can directly read or write to the designated blocks in the file system.

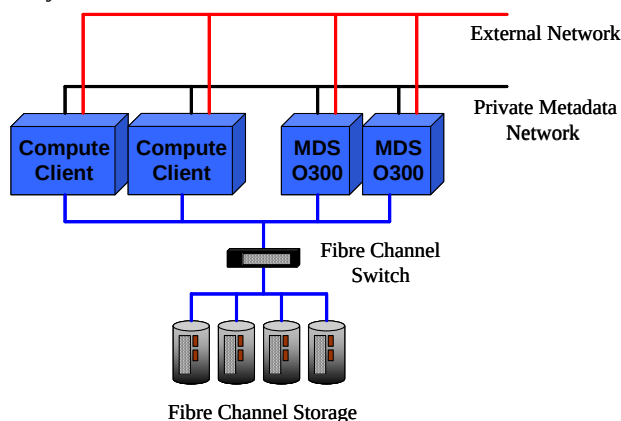


Figure 2. Simplified view of the current CXFS environment.

This direct access by all clients over a fibre channel network provides a flat performance profile for all clients in the cluster. Moreover, the performance for all clients

attached to the SAN is comparable to all clients being direct attached to the disk subsystems, minus the added latency levied by the MDS. Furthermore, CXFS eliminates the need for file duplication in cluster environments which often occurs when slow NFS speeds are not sufficient. CXFS file systems have large file support (64-bit scalability) and supports files sizes up to 9 million terabytes and file systems up to 18 million terabytes. Clients are available for IRIX, Solaris, Windows, and Linux.

In addition to serving out the file systems in the SGI cluster at the NCCS, CXFS is also serving the Data Management Facility (DMF) mass storage archive, which is available to all clients in the SGI cluster. DMF, an SGI application, is being used to store the large amount of data within the NCCS created by the GSFC's Global Modeling and Assimilation Office (GMAO)[7] within the NCCS. The DMF presently houses over 400 terabytes of data and about 15 million files.

4.1. Wide Area Requirement

In 2004, NASA announced the success of Project Columbia at the NASA Ames Research Center based at Moffett Field, CA. The result of this project was the deployment of a large cluster of SGI Altix systems (around 60 TFlop/s peak) with over 500 TB of file storage served by CXFS and a complementary DMF mass storage system. This cluster is currently running some of the largest Earth and space science applications in the world and generating some of the largest data sets ever managed. Since most of the users of Project Columbia are geographically distributed, users typically either move the entire data set or a subset locally in order to perform adequate analysis. With the largest component of the Earth and space scientists physically located at Goddard, the NCCS is exploring ways to facilitate the movement of data between Goddard and Ames.

Current access methods of secure copy or FTP are insufficient either in bandwidth or security in order to be able to satisfy all the requirements for moving data between the two centers. A different method for transferring files is needed. How can two CXFS file systems, one at the NCCS and one at Ames, separated by thousands of miles be connected over a wide area network in order to be able to serve file systems to multiple hosts in a secure, highly performing, and convenient manner?

4.2. Potential Solution

The convenience of providing storage devices over existing IP networks is very popular, and many centers and programs are using network-attached storage (NAS) devices. These storage devices allow the movement of

files between clients and storage devices over existing TCP/IP networks. Typically, there is no need to alter the storage application in order to take full advantage of NAS devices.

However, transferring block-level data from either a direct attached storage device or a fibre channel SAN over IP is not straight forward and requires the use of new protocols. Three such protocols are currently being used within storage networks: iSCSI, FCIP, and iFCP.[8] Each allow for centers to virtually extend the concept of a direct attached storage device or a SAN over a wide area network. This work focuses only on the FCIP and iFCP protocols that allow the fibre channel SAN traffic to be routed over IP based networks.

4.3. Pilot Project

Figure 3 shows the setup for the pilot project for the WAN CXFS testing. Multiple MDS and clients will be configured with the appropriate fibre channel and network connectivity to install the CXFS software. In addition, the system will have at least two fibre channel storage devices. Roughly half of the cluster will be located within one building at the GSFC and the other half in a second building within GSFC. Hence, the fibre channel traffic will be routed over the Goddard IP network using the fibre channel to IP converters or gateways.

The purposes of this test bed are to explore and present results for the following areas of interest:

- *Proof of Concept:* While CXFS is successfully being used in both the NCCS and Ames centers, CXFS over a WAN is relatively new. Much work has been done in the storage community to prove the benefits and usefulness of network attached storage devices, but the concept of using CXFS over a wide area remains to be proven in the NASA environment.
- *Security:* Even though the system may be usable, the security may not be sufficient. This work will document the lessons learned with respect to securing a SAN over a wide area.
- *Fibre Channel to IP Protocols:* The pilot project calls for testing out hardware devices that implement both the FCIP and iFCP protocols. This work will explore the benefits and drawbacks of both types of devices with respect to reliability, security, etc.
- *Performance:* While proof of concept and security are major priorities, if the system's performance is not sufficient, users will not see the advantage of this solution. Therefore, latencies will be introduced into the pilot system

as files are read and written in order to simulate the distance between two centers. The pilot project will document the benchmark tests, performance, and the ability for CXFS to serve files over a long distance.[9]

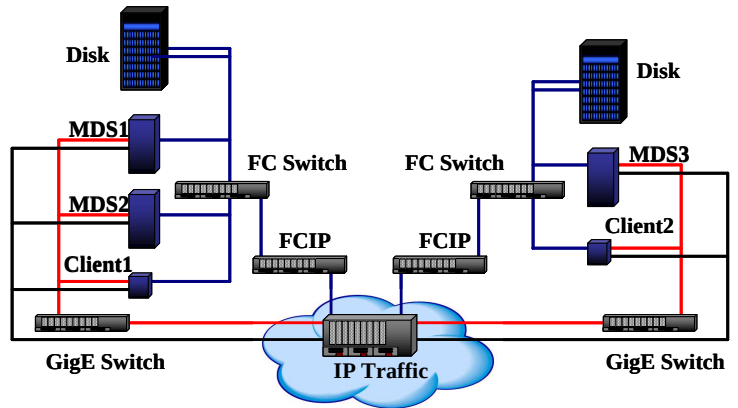


Figure 3. Architecture for the wide area CXFS SAN.

4.4. Future Steps

As the project matures and the test bed proves stable and reliable with adequate performance, further tests are being considered. While the pilot project will simulate distances, the next logical step is to physically extend the test bed to other sites, like NASA Ames. Implementation of the WAN extension is proposed initially using existing networks and potentially a dedicated 10 gigabit per second channel of the National Lambda Rail. In addition to extending the SAN externally, additional clients (mainly Linux) need to be tested for stability and performance. This will allow the NCCS to more closely link other assets, such as the Linux clusters and the GSFC's Scientific Visualization Studio (SVS), to the data served by the CXFS cluster while taking full advantage of existing IP networks. Finally, caching servers[10] are being suggested to create mirror images of file systems on either side of the wide area SAN. These cache servers only transfer the modified blocks of a file, and make it seem like they can exceed the theoretical peak of the network bandwidth.

5. Data Intensive Computing/Emerging Technology Test Bed

In addition to evaluating the emerging technology in support of wide area SAN's, the NCCS has joined a multi-agency effort in the creation of an interagency Emerging Technology Test Bed. Through partnerships with industry, this environment will provide a "real"

world testing environment for technology that can enable data intensive computing.

The creation the test bed provides a collaborative environment for active participation by multiple agencies in technology evaluation and research activities. This consortium across agencies maximizes investment dollars while minimizing gaps and duplication of research efforts. In addition, it creates a strong alliance to influence industry to better meet the requirements for data intensive computing.

Figure 4 shows a schematic view of the test bed environment. Hardware representative of what is currently found in production HPC centers will be installed at three geographical distributed sites. More specifically, each arm of the test bed will have a small Linux cluster, mass storage system, and be in an isolated Kerberos realm for security purposes. After an initial phase of seeding the test bed with equipment, pioneer users will be allowed to explore the environment to better understand how to perform data intensive activities in a distributed environment.

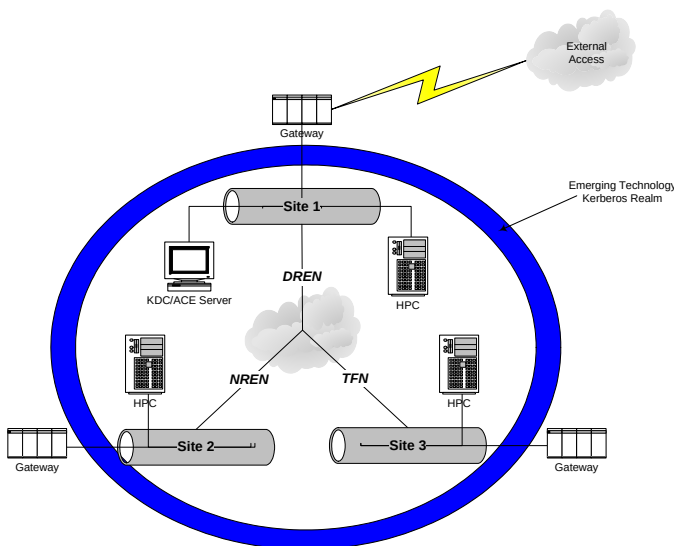


Figure 4. Notional view of the emerging technology test bed.

The test bed is not designed solely for hardware evaluation purposes. Both vendors and researchers within NASA will be encouraged to install software products relevant to data intensive computing over a wide area. As an example, the on-going work of the DDE prototype services will be installed on this test bed.

As milestones are reached and the partnerships with vendors solidified, an environment of open testing will be created. Vendors will be encouraged to offer new solutions to be placed into the environment for full and open testing.

References

- [1] <http://esdcd.gsfc.nasa.gov/>
- [2] <http://nccs.nasa.gov/>
- [3] <http://www.nas.nasa.gov/About/Projects/Columbia/columbia.html>
- [4] See for example *Data Management as a Cluster Middleware Centerpiece*, Jose Zero, et al., Proceedings of the Twenty-First IEEE/Twelfth NASA Goddard Conference on Mass Storage Systems and Technologies, April 2005.
- [5] <http://www.npaci.edu/DICE/SRB/>
- [6] <http://www.unidata.ucar.edu/packages/dods/>
- [7] <http://gmao.gsfc.nasa.gov/>
- [8] *SAN and Data Transport Technology Evaluation at the NASA Goddard Space Flight Center*, H. Thompson, Proceedings of the Twenty-First IEEE/Twelfth NASA Goddard Conference on Mass Storage Systems and Technologies, April 2005.
- [9] The pilot project will use the Iozone (<http://www.iozone.org>) file system benchmark for initial testing.
- [10] <http://www.yottayotta.com/>