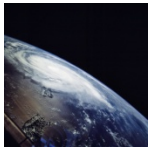


NCCS

NCCS Wide Area CXFS Test Bed Project

NASA Center for Computational Sciences (NCCS)
Scientific Computing Branch
Goddard Space Flight Center
April 2005

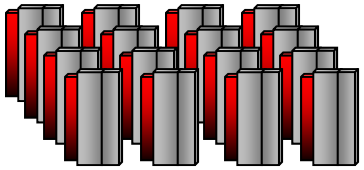
Enabling exciting advances in Earth and space sciences research through state of the art High End Computing as only NASA can.

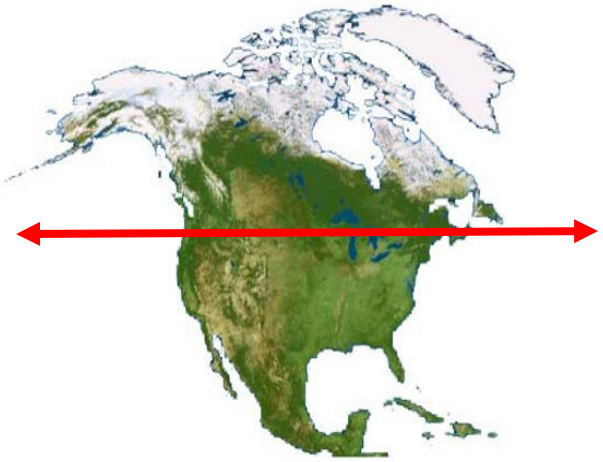
Driving Requirements

NCCS

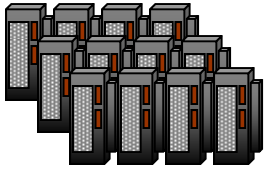
NASA Center for Computational Sciences

NASA Ames
Moffett Field, CA





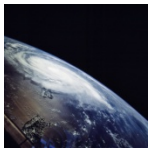
NASA Goddard
Greenbelt, MD



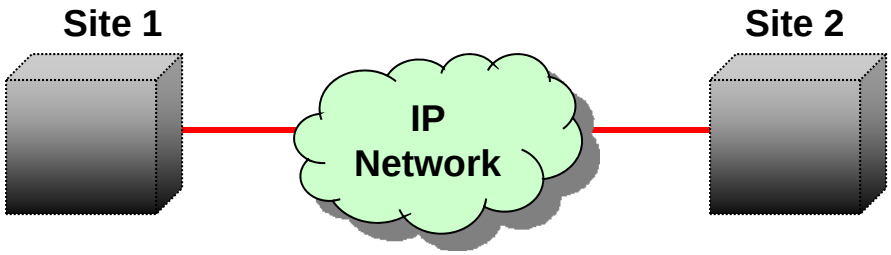
- Project Columbia
  - SGI Altix (60+ TF Peak)
  - CXFS & DMF
  - 16 PB Capacity
- NASA Goddard Resources
  - SGI Origin 3800 & HP SC45
  - Small SGI Altix (and growing)
  - CXFS & DMF
  - ~450TB (2+PB Capacity)

Science mission applications will be running at both centers.
Maximizing distributed resources requires data accessibility.

Wide Area CXFS Test Bed Conceptual View



NCCS



- **What are the benefits?**
  - Seamless access to data regardless of its physical location.
  - File systems at both sites can be seen as standard file systems by the other site.
  - Enhanced workflow and high productivity as users can concentrate on applications rather than moving data.

- **Each site basically has**
  - A piece of the CXFS domain
  - Metadata Servers
  - Storage and Fibre Channel (FC) network
  - Clients (Computational Engines)
  - FC to IP converters

- **What are the concerns?**
  - Security
    - IT
    - Data Access Issues
  - Performance
  - Availability
  - Inter-site cooperation and coordination

The concerns drive the need for a test bed within NASA.

Why a test bed?

- SGI and others have shown that this technology has worked for other customers
- Not yet proven this is ready for Earth and space science applications within NASA
- Proof of concept
- Continued partnership between NASA and SGI to drive technology to help meet our user's requirements
- Lots of questions about this technology:
  - Security
  - Performance
  - Reliability

Sample CXFS Installation

NCCS

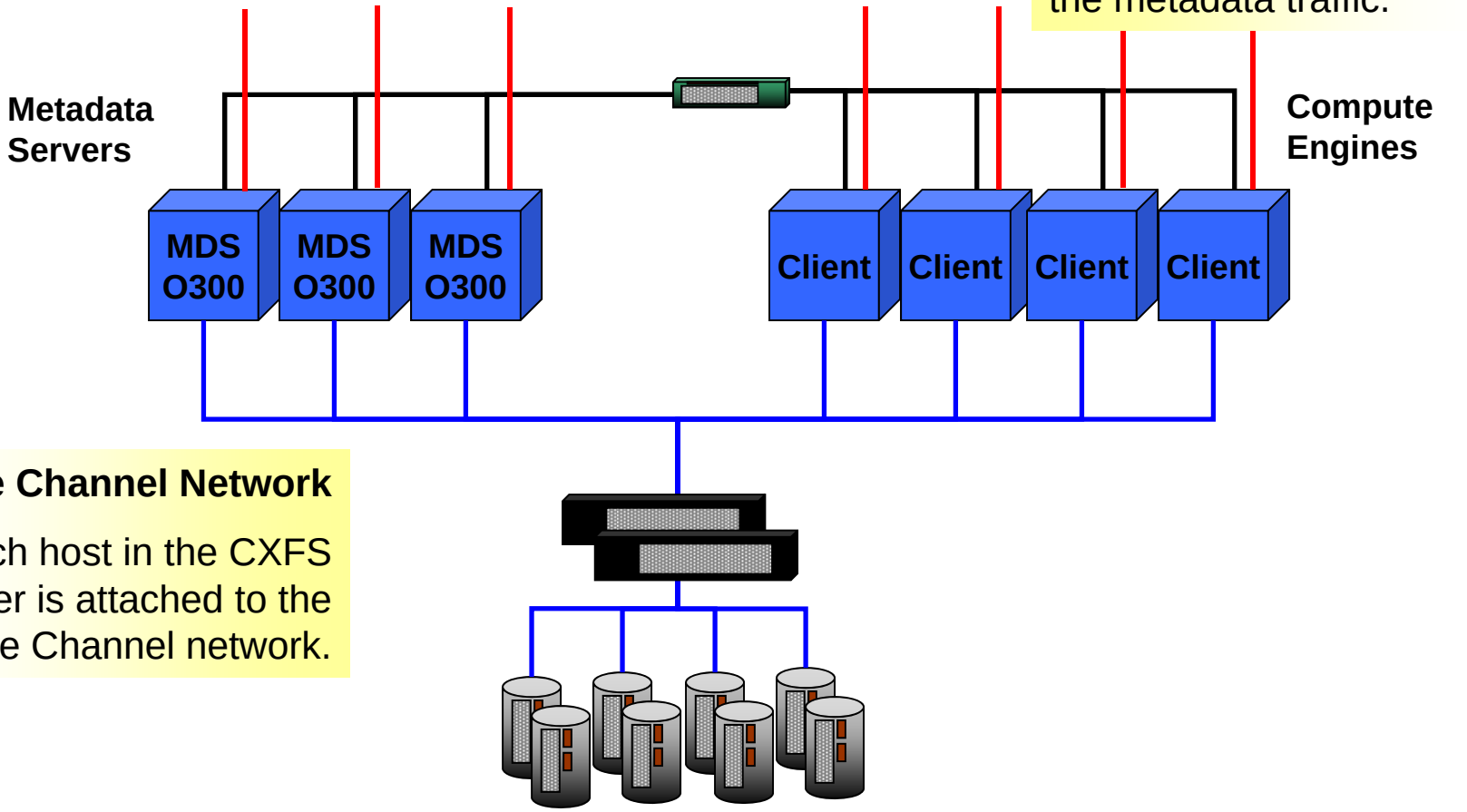
Metadata Network (1 Gb)

Fibre Channel (2 Gb)

External Network (Ethernet or 1 Gb)

CXFS Metadata Network

All members of the cluster have private interfaces for the metadata traffic.

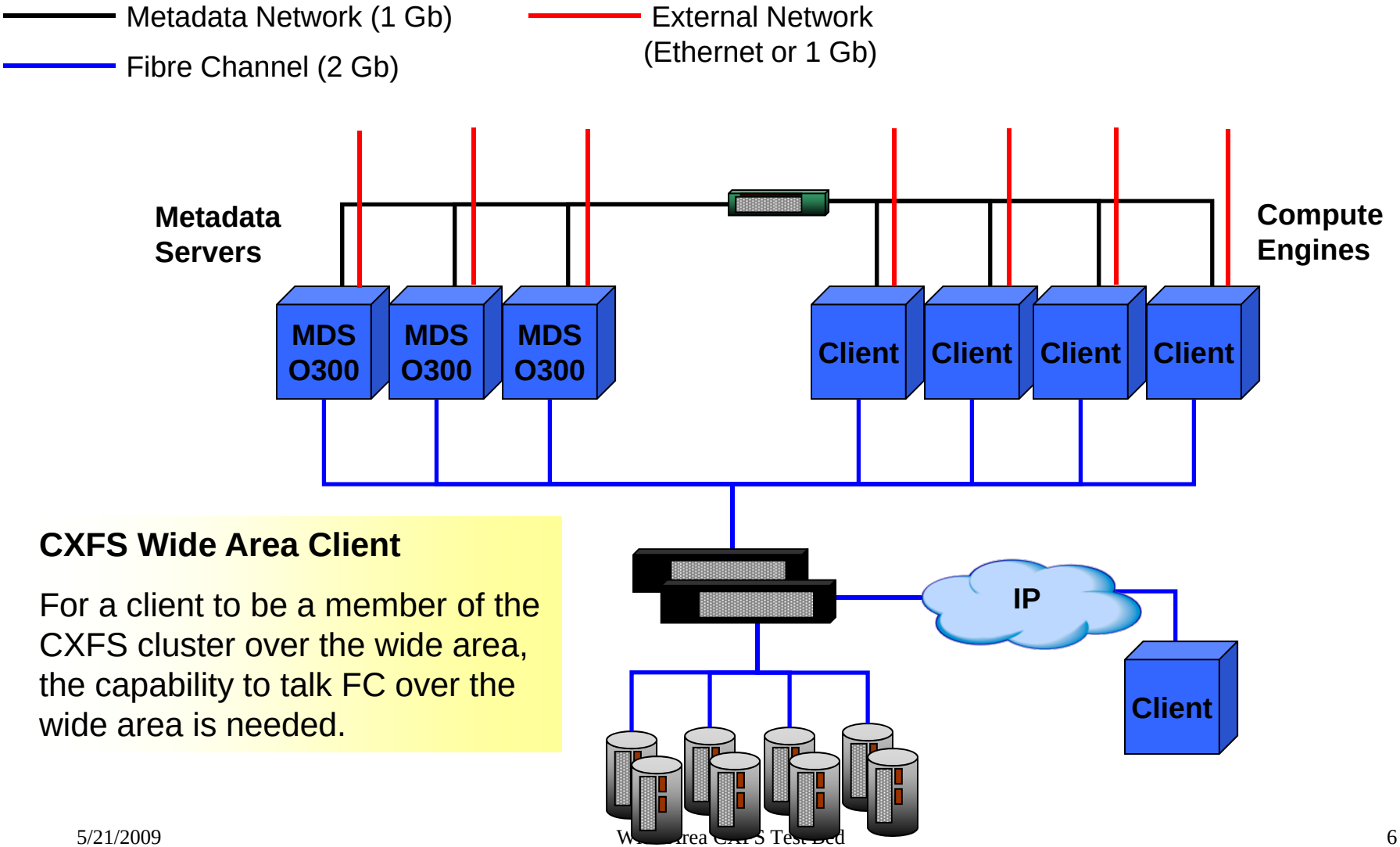


Fibre Channel Network

Each host in the CXFS cluster is attached to the Fibre Channel network.

How can clients mount CXFS file systems over a WAN?

NCCS



CXFS Wide Area Client

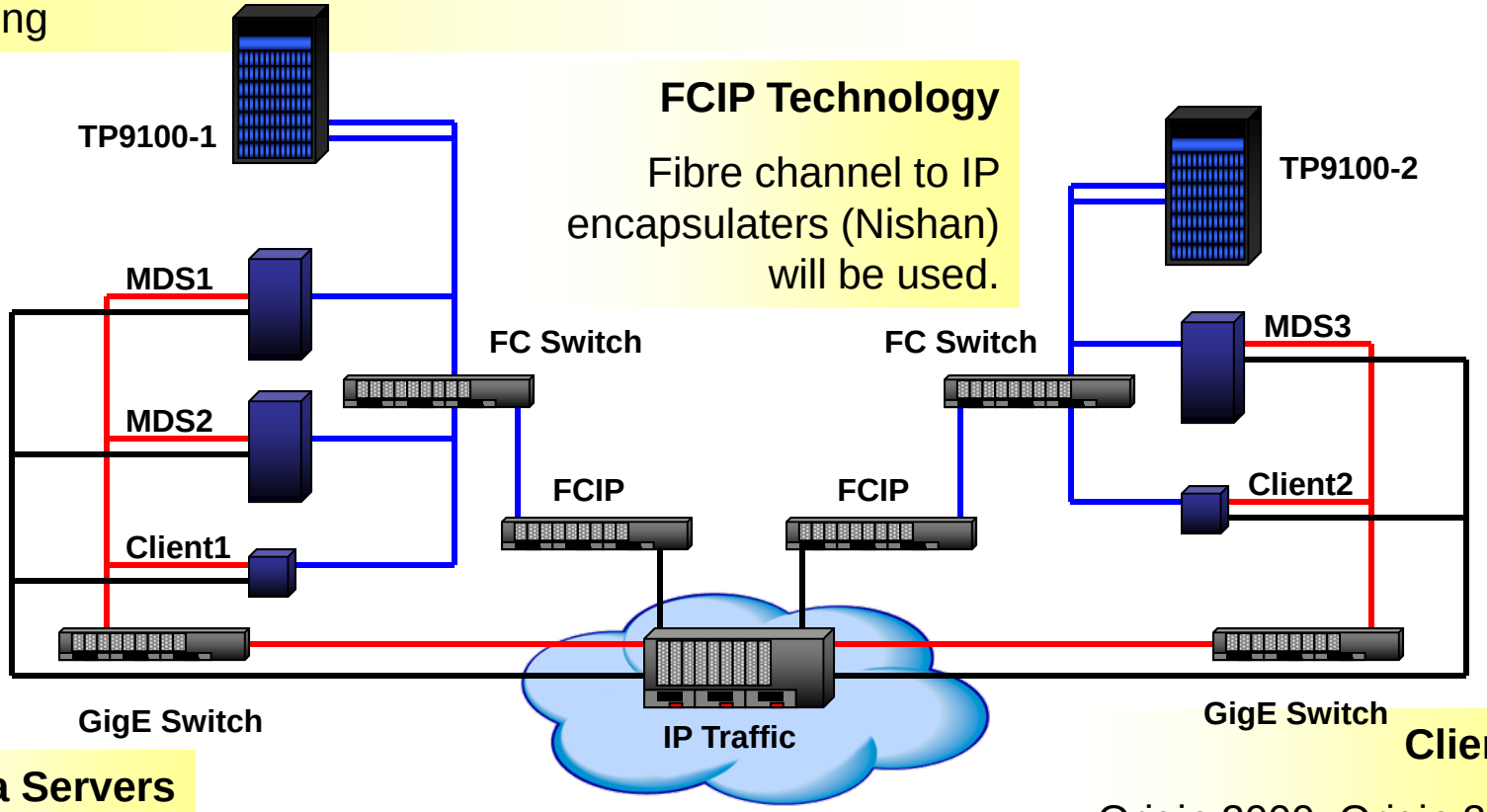
For a client to be a member of the CXFS cluster over the wide area, the capability to talk FC over the wide area is needed.

Phase One Notional Architecture

NCCS

Test Bed

- Built within NCCS Environment (private network)
- Inject latencies to simulate distances of a wide area
- Initial Testing



FCIP Technology

Fibre channel to IP encapsulaters (Nishan) will be used.

Metadata Servers

Origin 300

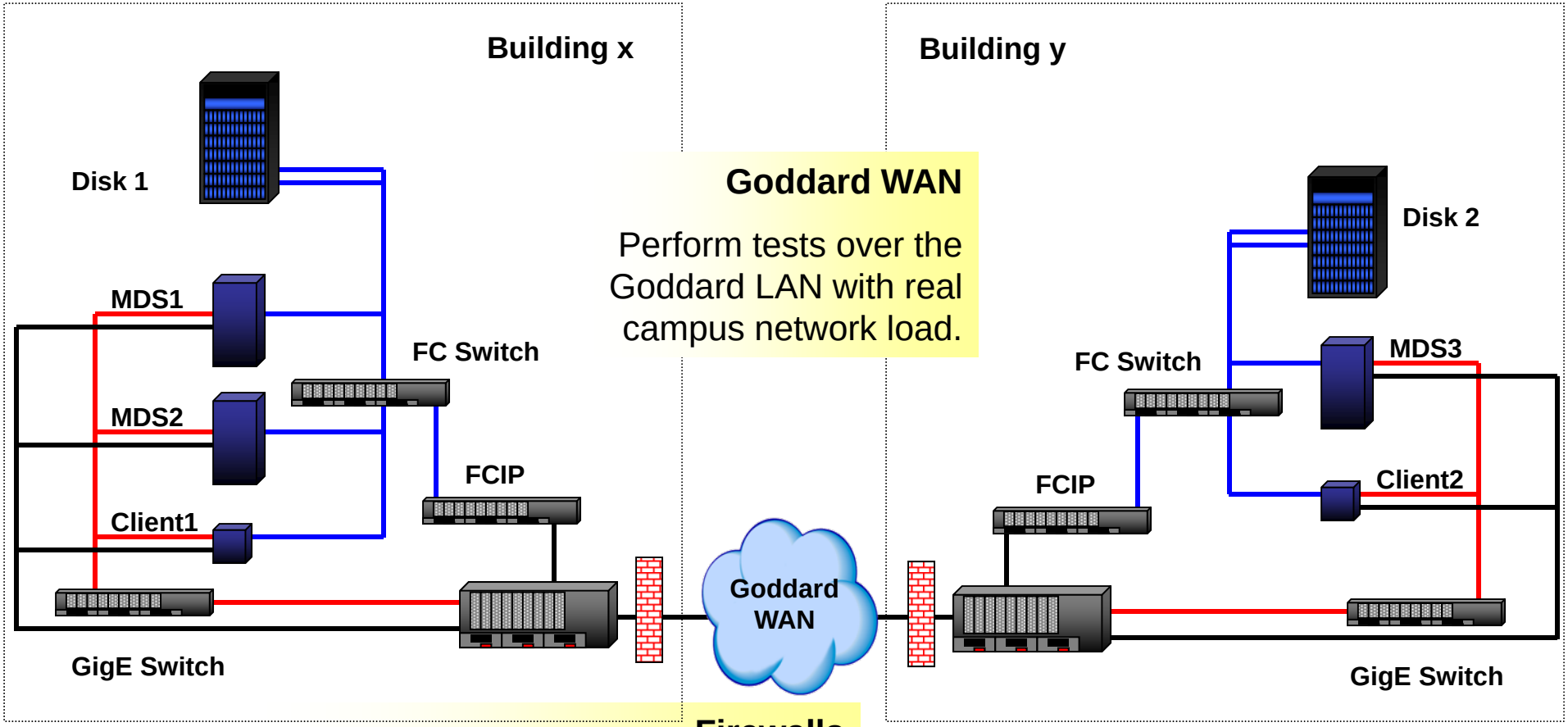
Clients

Origin 2000, Origin 300, and Linux

Phase Two Notional Architecture

NCCS

NASA Center for Computational Sciences



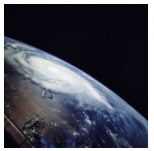
Goddard WAN

Perform tests over the Goddard LAN with real campus network load.

Firewalls

Introduction of firewalls to further make test bed closer to real centers.

Assembling the Test Bed



NCCS

- Steps involved in putting the test bed together

  1) Assemble all hardware components and cables

  2) Upgrade Irix on Metadata Server and Clients

  3) Upgrade firmware on FC switch's & Nishan Switch's

  4) Configure hardware components without Nishan switch's, verify CxFS cluster functioning properly and run performance tests to the file systems using Iozone benchmarking software

  5) Configure Nishan switch's, verify CxFS cluster functioning properly and run performance tests to the file systems using Iozone benchmarking software

- Lessons learned putting in the Nishan switches

  1) All hardware components need to be at recommend firmware levels for compatibility

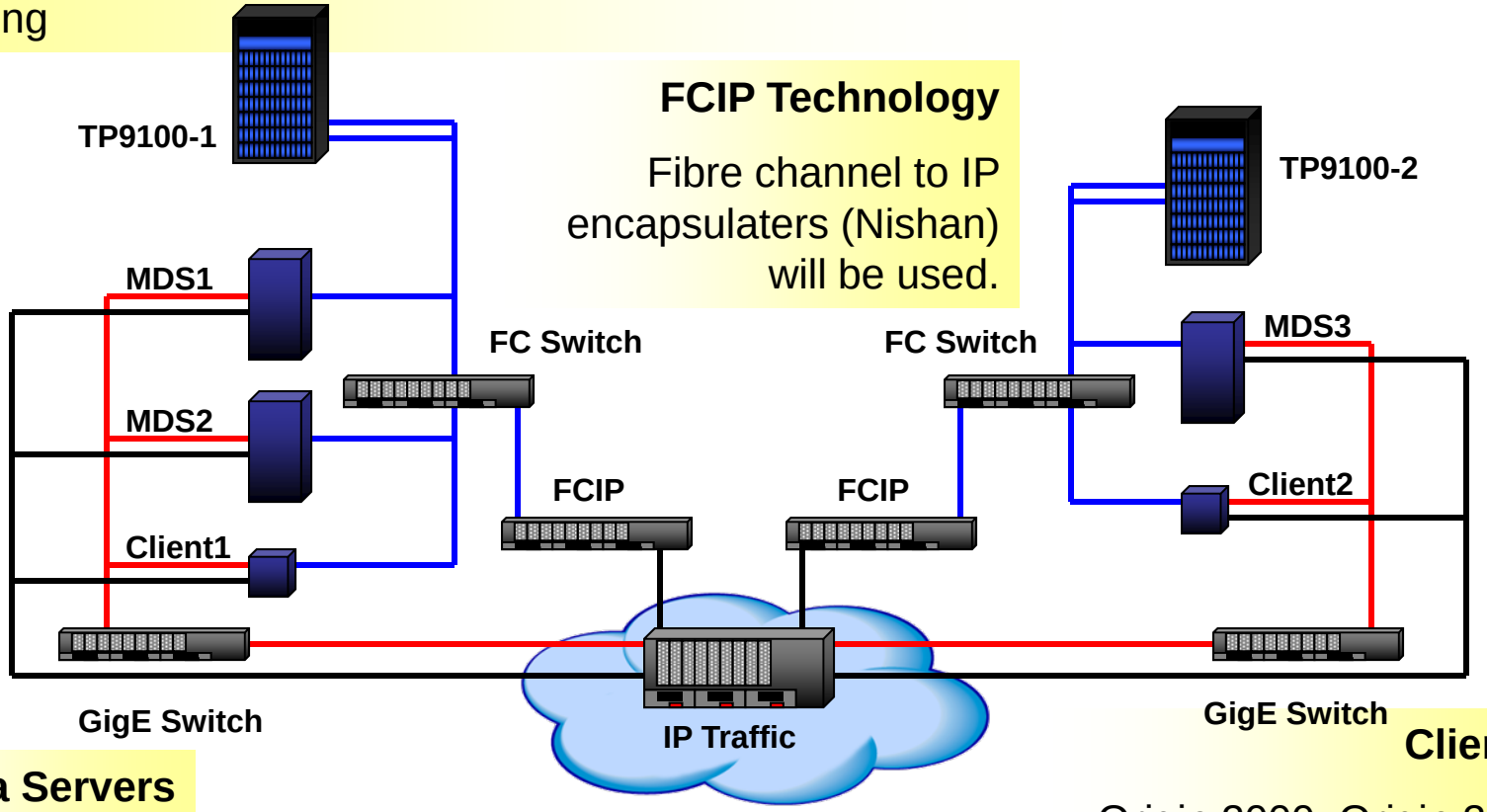
Actual Test Bed Components

NCCS

NASA Center for Computational Sciences

Test Bed

- Built within NCCS Environment (private network)
- Inject latencies to simulate distances of a wide area
- Initial Testing



FCIP Technology

Fibre channel to IP encapsulaters (Nishan) will be used.

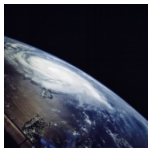
Metadata Servers

Origin 300

Clients

Origin 2000, Origin 300, and Linux

Actual Test Bed Architecture



NCCS

- Metadata Servers (3)
  - Origin 300 (4 CPUs with 4GB of memory)
  - Irix 6.5.24
  - CXFS 3.2
- Clients or Compute Hosts (2)
  - Origin 2000 (32 CPUs with 16GB of memory)
  - Irix 6.5.24
  - CXFS 3.2
- Disk
  - TP9100
    - 7 x 181G Drives
    - 2 Luns
    - 2 Controllers
    - 128M Cache
    - 7.75 firmware
- FC Switches
  - Brocade 2800 (2.6.1c firmware)
  - Brocade 3800 (3.1.1 firmware)
- Nishan Switches (2)
  - IPS-3300's

Remote Client

Currently, one of the compute hosts is configured as a wide area client.

All subsequent test results are run from the remote client over the wide area to the disk.

I/O Zone Benchmark

- ## I/O Zone Benchmark

  IOzone is a file system benchmark tool. The benchmark generates and measures a variety of file operations and is useful for determining a broad file system performance analysis.

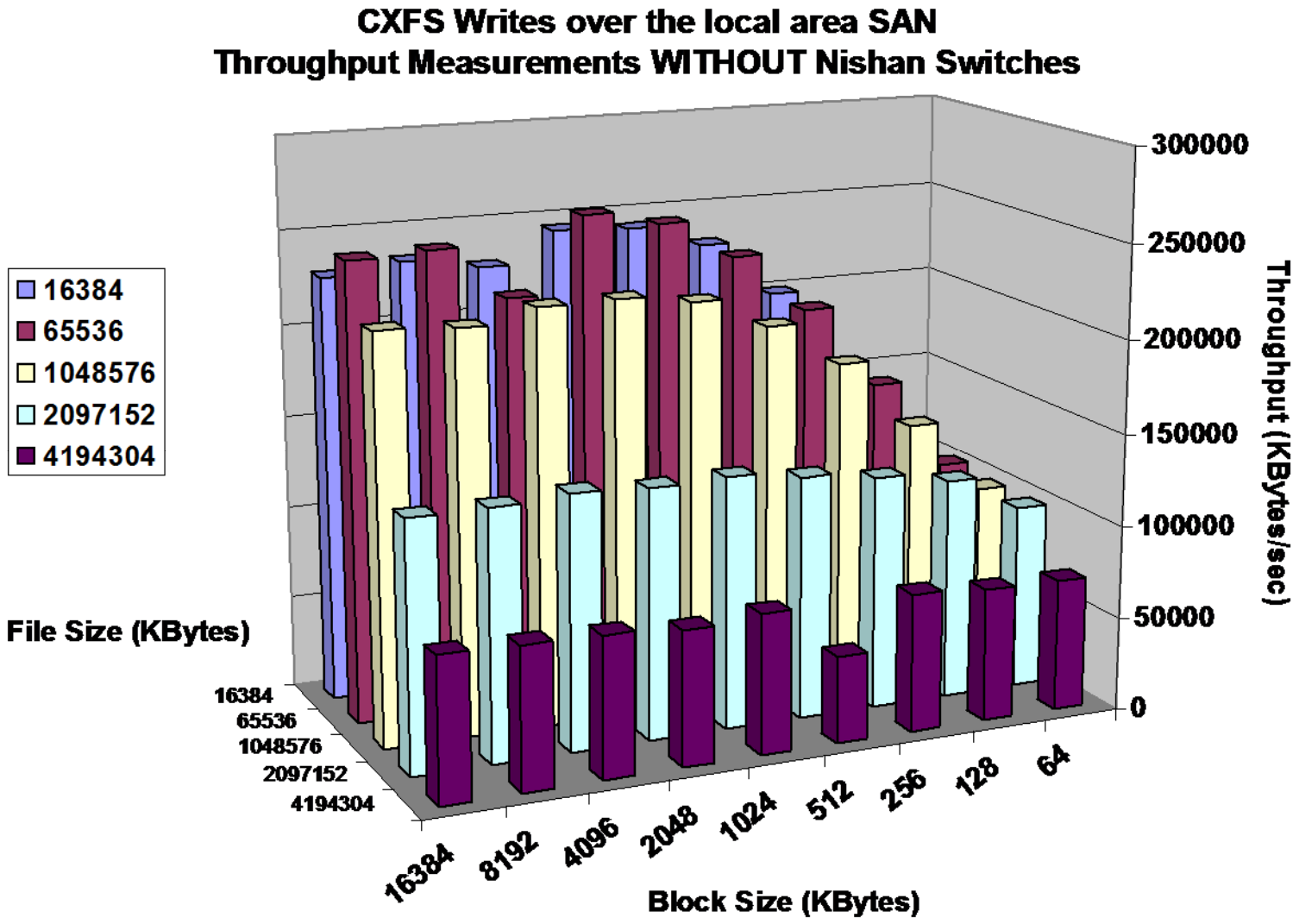
- ## Test's Performed

  - I/O Zone was run from the remote client talking to the metadata servers and the disk over the wide area
  - CXFS Test: Reads/Writes over the Wide Area SAN throughput Measurements WITH and WITHOUT Nishan Switches
  - Example Command: iozone -i 0 -i 1 -t 8 -r 1m -s 8g

    -i 0 = Write –i 1 = Read -t 8 = 8 threads –r 1m = record size –t 8 = # of Threads –s 8g = File size

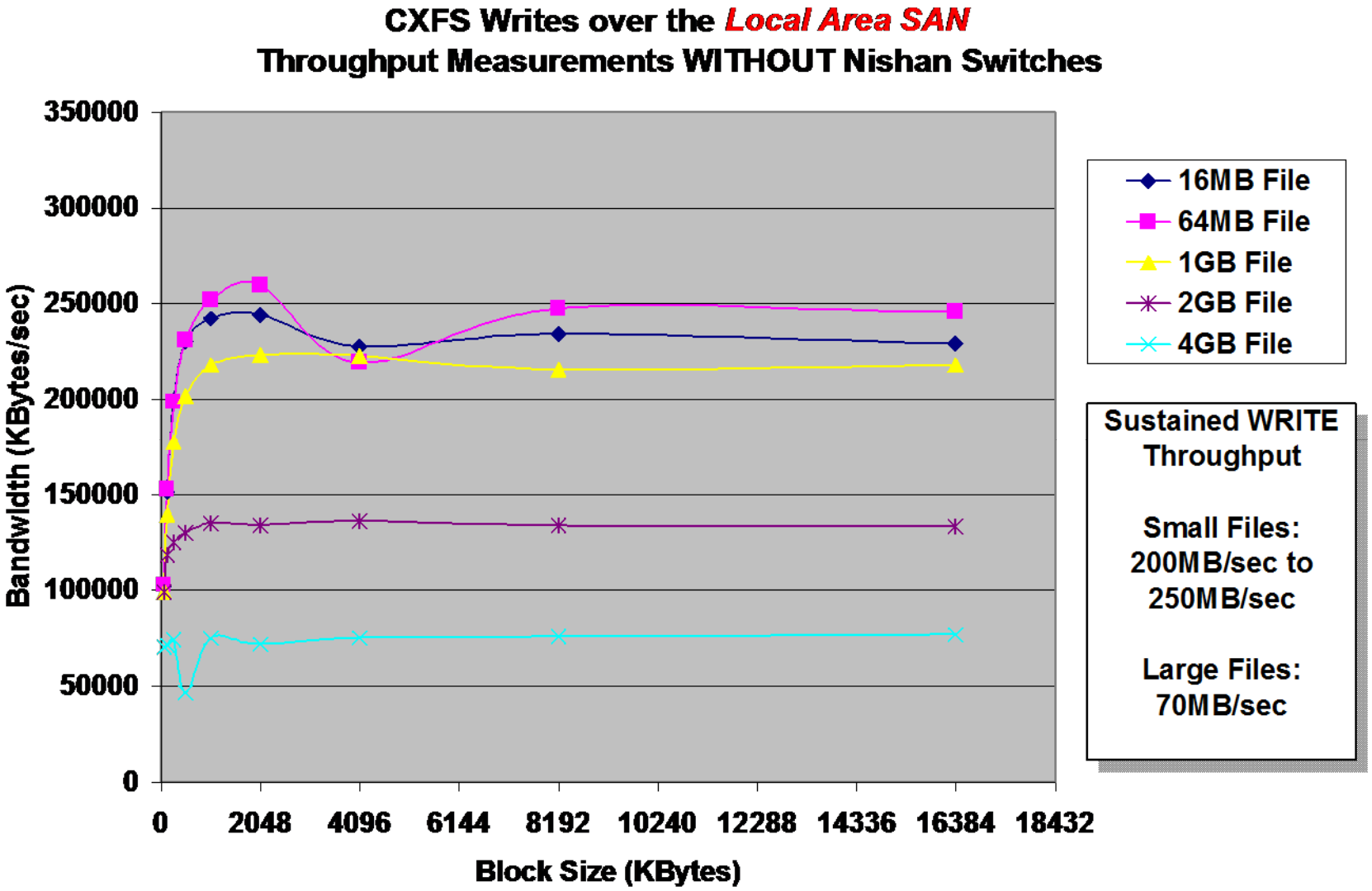
CXFS *Local Area SAN* Writes (1)

NCCS



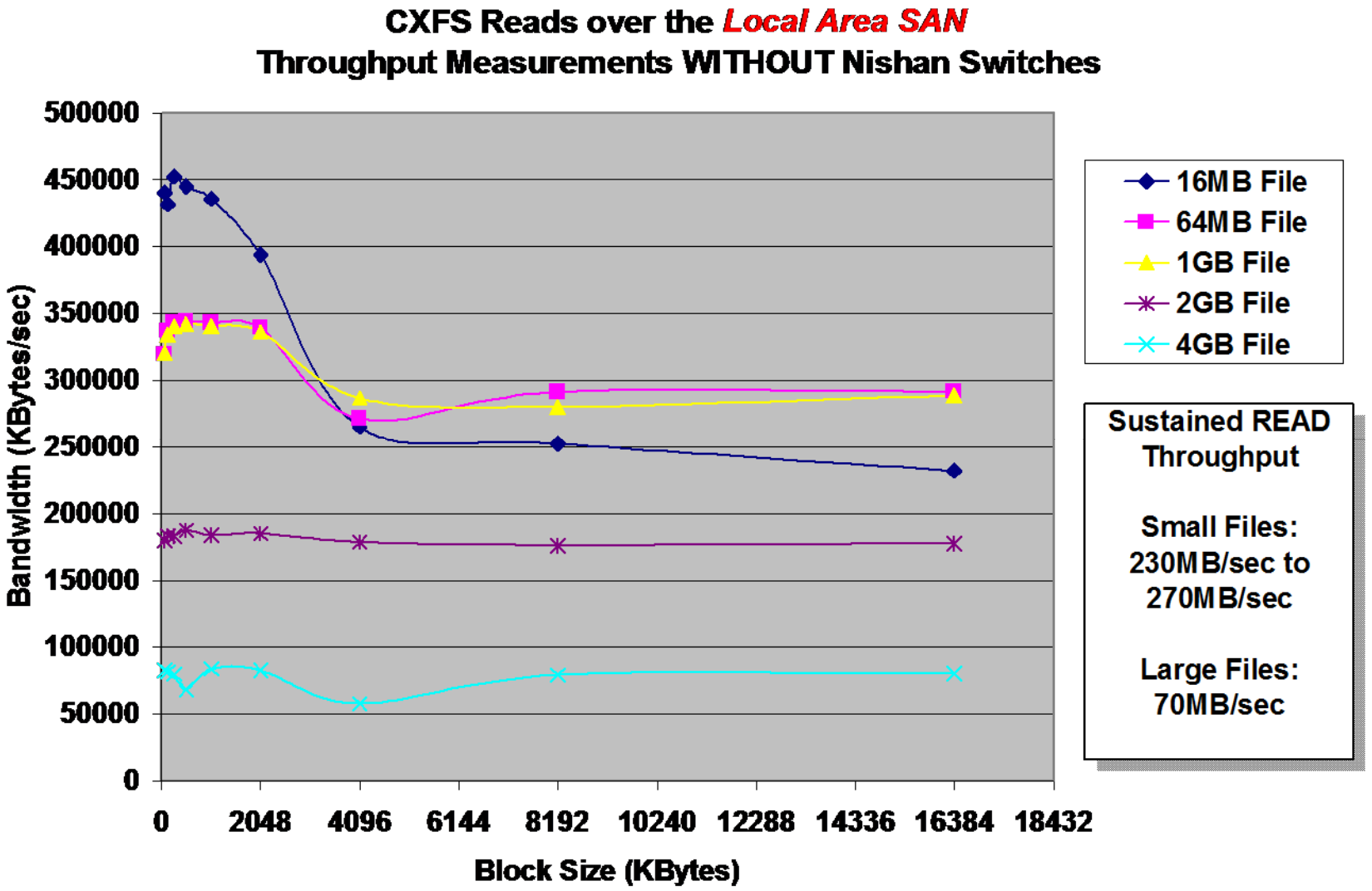
CXFS *Local Area SAN* Writes (2)

NCCS



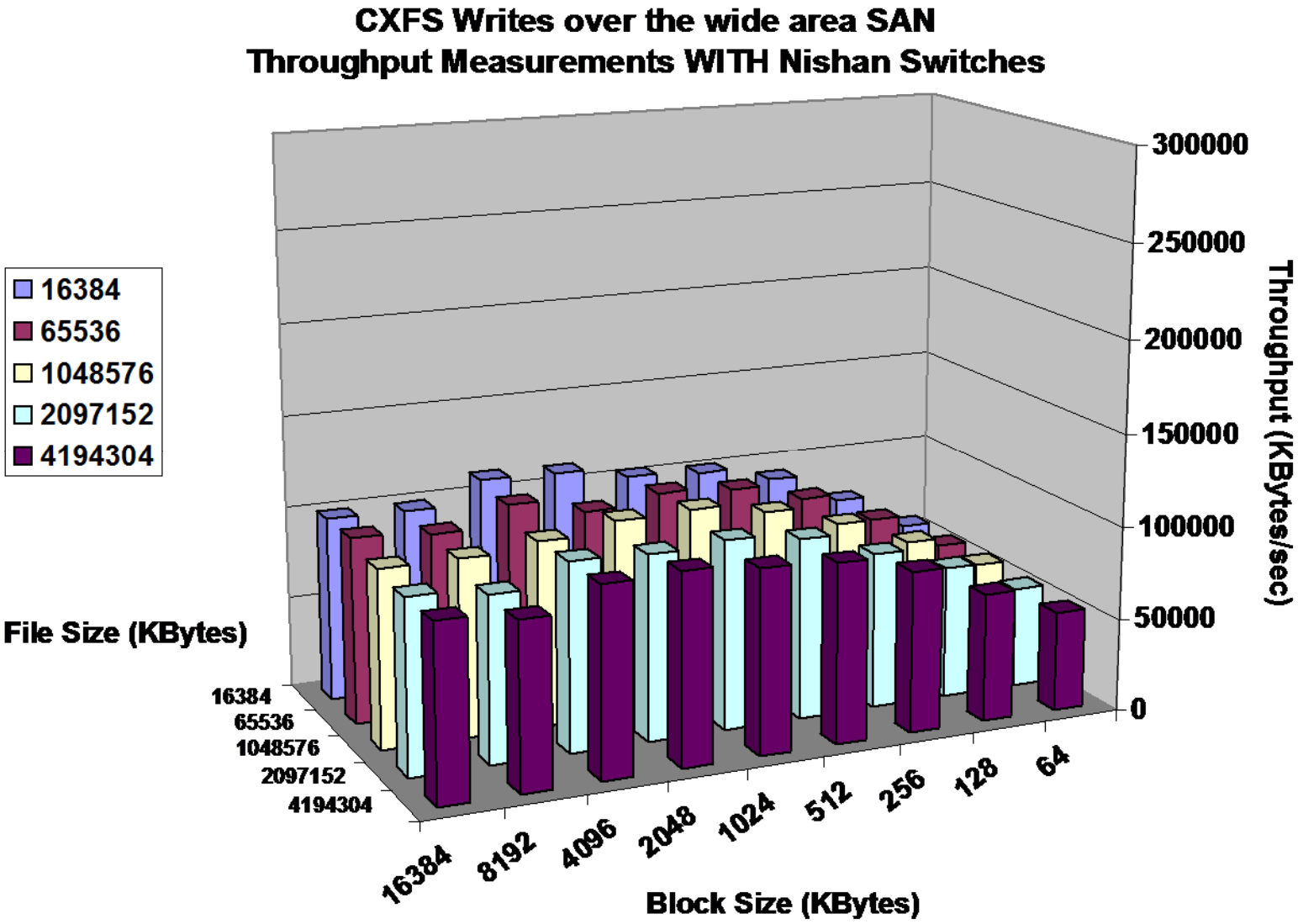
CXFS *Local Area SAN* Reads

NCCS

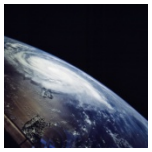


CXFS *Wide Area SAN* Writes (1)

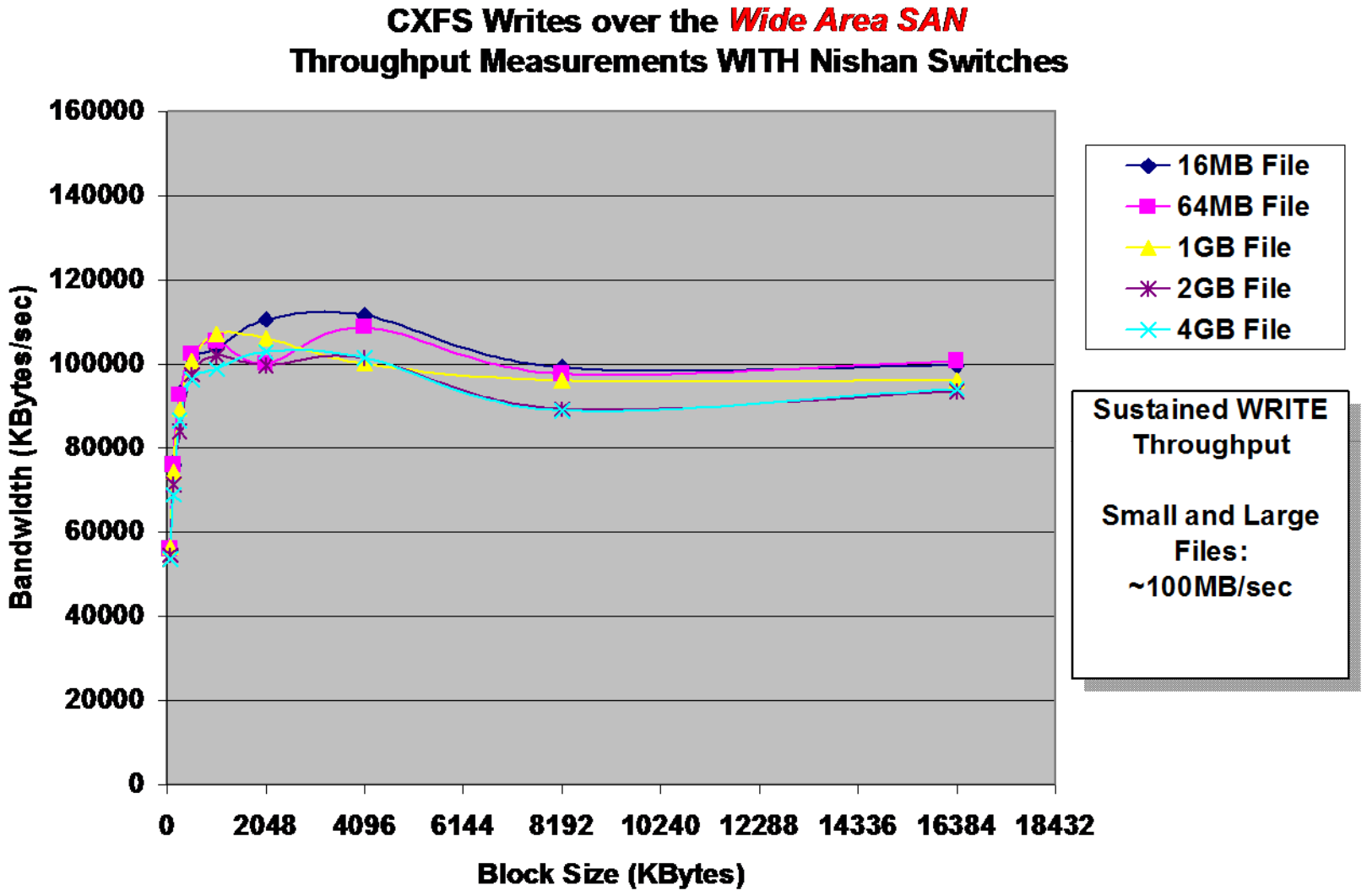
NCCS



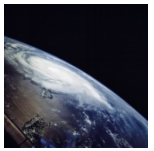
CXFS *Wide Area SAN* Writes (2)



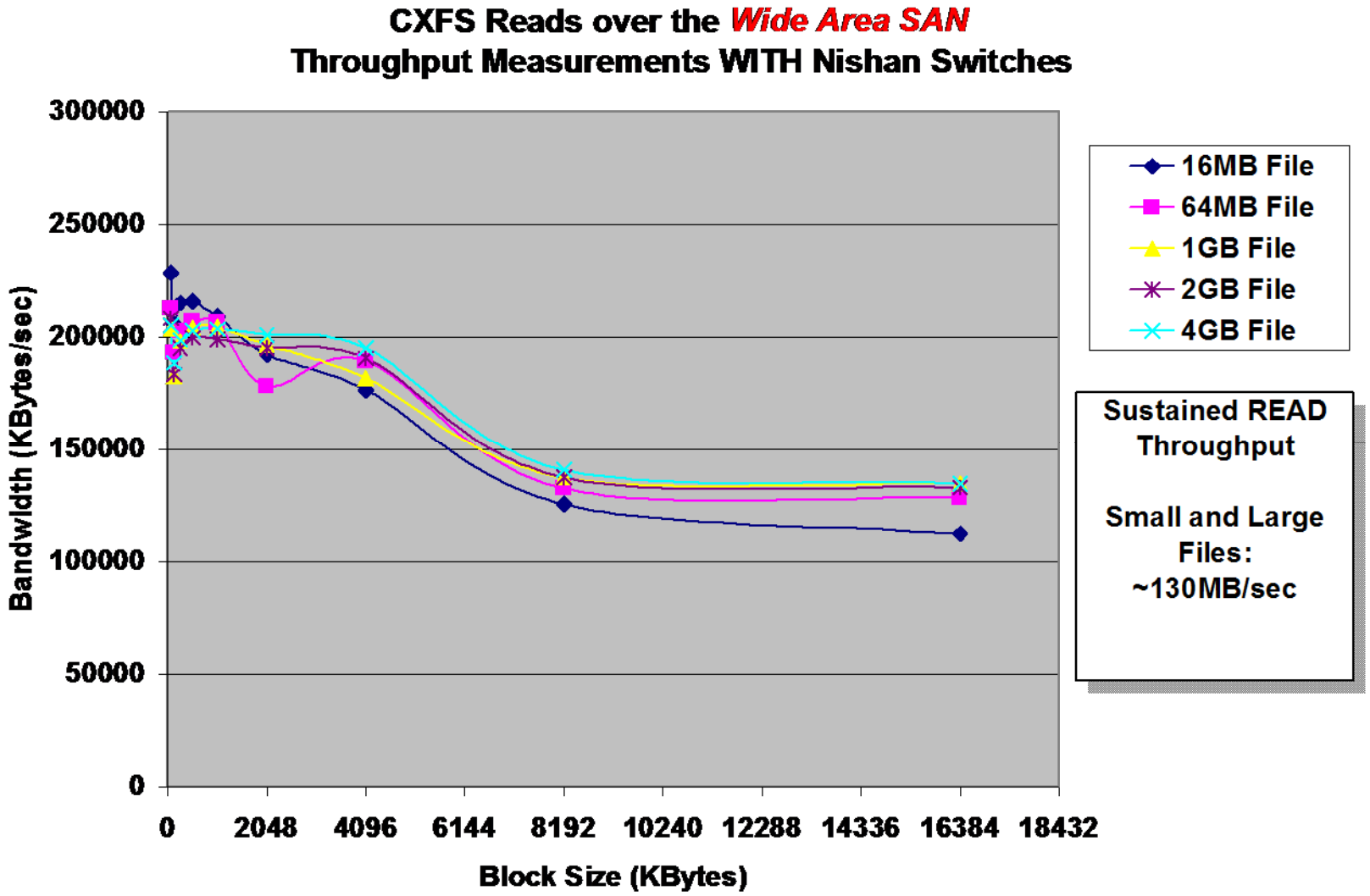
NCCS



CXFS *Wide Area SAN* Reads



NCCS



Initial Test Results

- Conclusions from Performance Test:

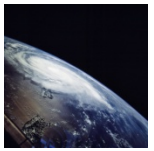
-Proof of concept - providing block-level data from Fibre channel attached disks using Nishan IP to FC network switches was successful.

- Latency and slower performance were experienced when using the Nishan switches as expected.

- Further testing is being done to find out how much network latency is acceptable within the CxFS cluster.

- Further testing is being done to increase performance to the application while dealing with the network latency.

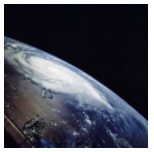
Future Projects



NCCS

- Continue to expand the test bed and evaluate the SAN for use within the NASA HPC environment
- Create an abstraction layer to the data stored the mass storage archives in order to
  - Provide easier distributed access to data without making additional copies
  - Easier methods for providing access to data holdings to a wider community
  - More robust data management tools

Wide Area CXFS Test Bed
Potential Future Testing Environment



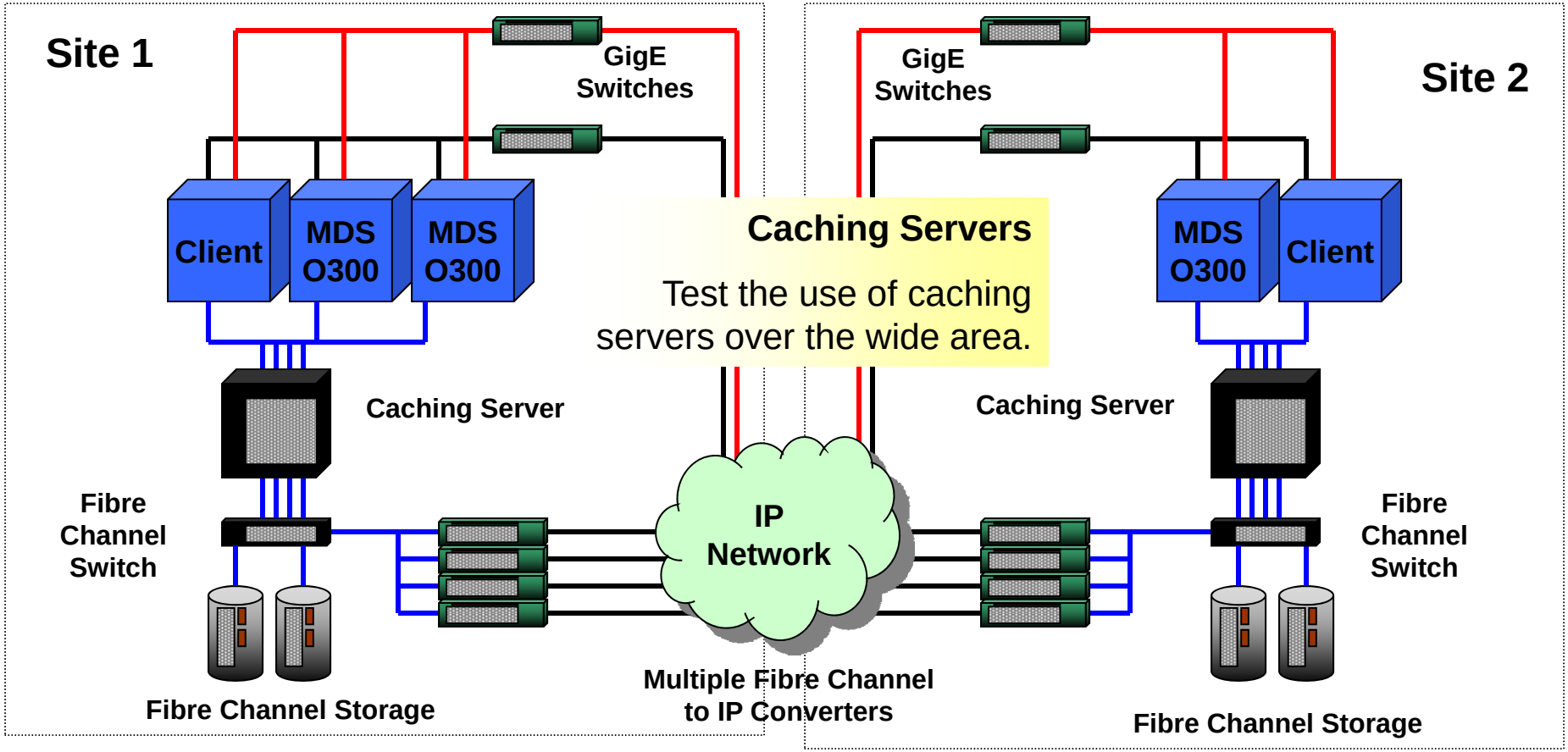
NCCS

NASA Center for Computational Sciences

Metadata Network (1 Gb)

External Network (Ethernet or 1 Gb)

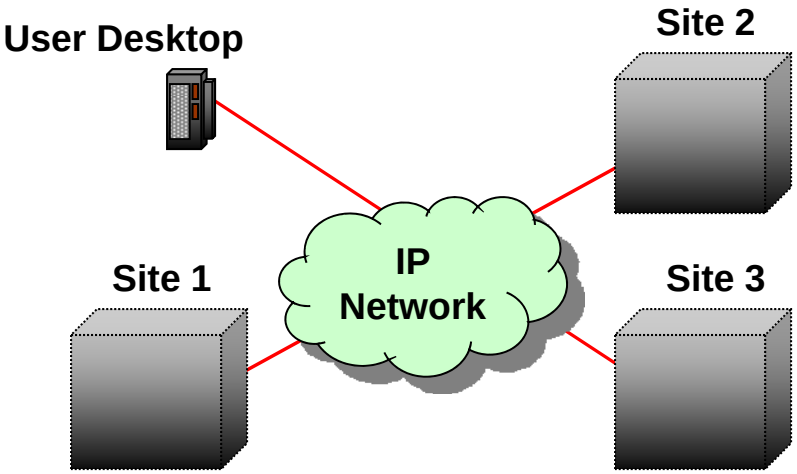
Fibre Channel (2 Gb)



Data Intensive Computing in a Distributed Environment

NCCS

NASA Center for Computational Sciences



- **Data Abstraction Layer**
  - Secure access to data over IP networks
  - User does not need to know where the data resides or about the storage system
  - Can capture data about the data (metadata)
  - Don't have to make multiple copies
  - Simplifies the user's view of distributed data

- **NASA User's Compute in a Distributed Environment**
  - Data is being generated everywhere
  - Need to analyze and bring information back to the user's desktop
  - Need methods to share generated data in an easy and secure manner

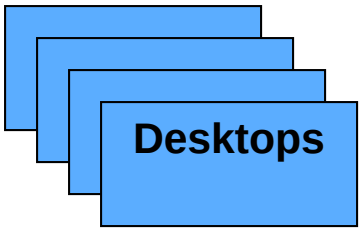
- **What are the concerns?**
  - Security and access control to data
  - Performance and Reliability
  - Synchronization of metadata and actual data stored

NASA has been using the Storage Resource Broker (SRB) to provide much of this type of capability.

Special Project: Distributed Data Environment (DDE)

NCCS

NASA Center for Computational Sciences

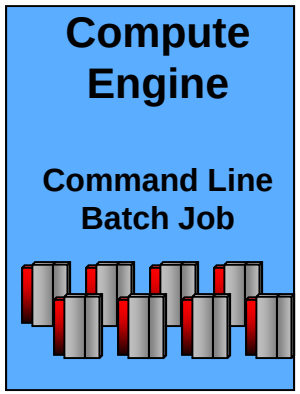


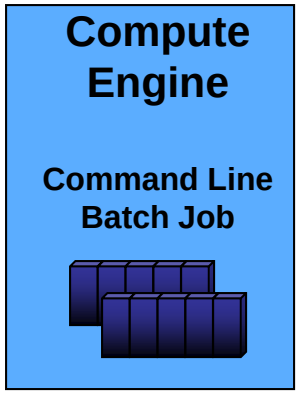
DDE/DODS can manage a local disk cache of read only SRB objects

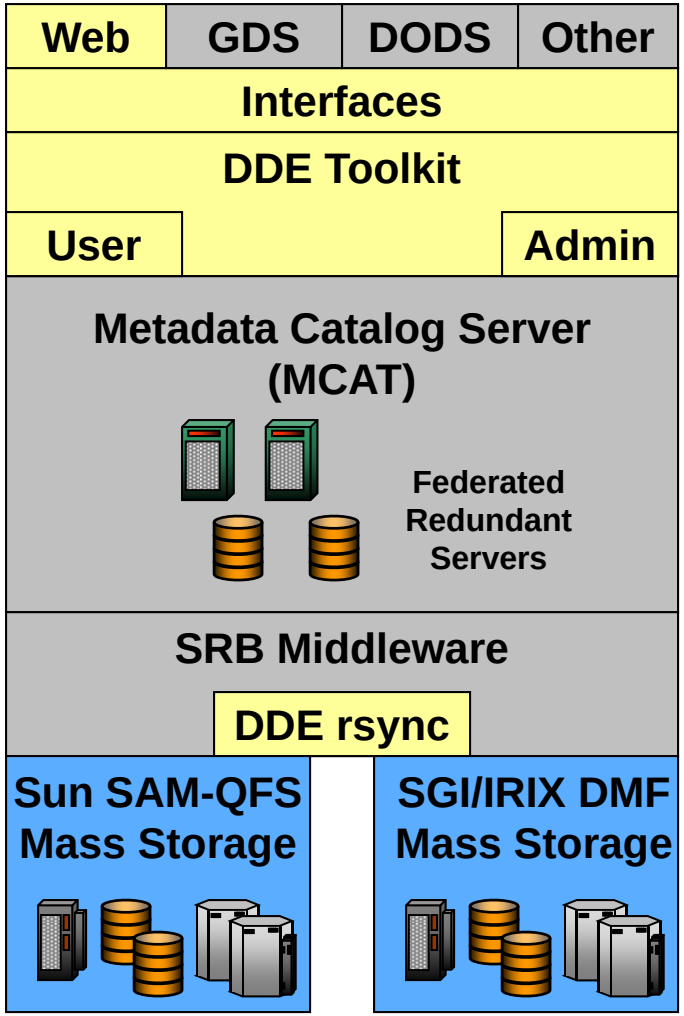
User and administrative commands can ingest information into and query the MCAT

SRB commands can be used from the command line or batch jobs to interface to mass storage

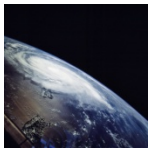
Users maintain the underlying access to mass storage systems from the compute engines







The Team



NCCS

NASA Center for Computational Sciences

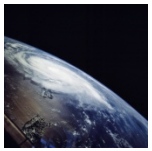
Tom Schardt, J. Patrick Gary, Bill Fink, Ben Kobler
Computational Information Sciences and Technology Office (CISTO): Code 606
NASA Goddard Space Flight Center
Greenbelt, Maryland 20071
tom.schardt@nasa.gov

Nicko Acks, Vaughn Noga, Matthew Whitehead, Mike Rouch, Hoot Thompson, Daniel Duffy
Computer Sciences Corporation
Sanz
Patuxent Technologies
Computational Information Sciences and Technology Office (CISTO): Code 606
NASA Goddard Space Flight Center
Greenbelt, Maryland 20071
daniel.q.duffy@gsfc.nasa.gov

Mike Donovan, Jim McElvaney, Kent Kamiscky, Pam Kennedy
Silicon Graphics Incorporated
Computational Information Sciences and Technology Office (CISTO): Code 606
NASA Goddard Space Flight Center
Greenbelt, Maryland 20071
miked@nccs.nasa.gov

Special thanks to SGI for loaner hardware and software!

For More Information



NCCS

- http://esdcd.gsfc.nasa.gov/
- http://nccs.nasa.gov/
- http://www.nas.nasa.gov/About/Projects/Columbia/columbia.html
- *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.
- http://www.npaci.edu/DICE/SRB/
- http://www.unidata.ucar.edu/packages/dods/
- http://gmao.gsfc.nasa.gov/
- *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.
- http://www.iozone.org
- http://www.yottayotta.com/

Standard Disclaimers and Legalese Eye Chart

NCCS

- All Trademarks, logos, or otherwise registered identification markers are owned by their respective parties.

- Disclaimer of Liability: With respect to this presentation, neither the United States Government nor any of its employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

- Disclaimer of Endorsement: Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. In addition, NASA does not endorse or sponsor any commercial product, service, or activity.

- The views and opinions of author(s) expressed herein do not necessarily state or reflect those of the United States Government and shall not be used for advertising or product endorsement purposes.