Design, Implementation, and Production Experiences of a Global Storage Grid

Phil Andrews, Chris Jordan,
San Diego Supercomputer Center,

Hermann Lederer,
Rechenzentrum Garching der Max-Planck-Gesellschaft
Max-Planck-Institut fuer Plasmaphysik

andrews@sdsc.edu, ctjordan@sdsc.edu, lederer@rzg.mpg.de

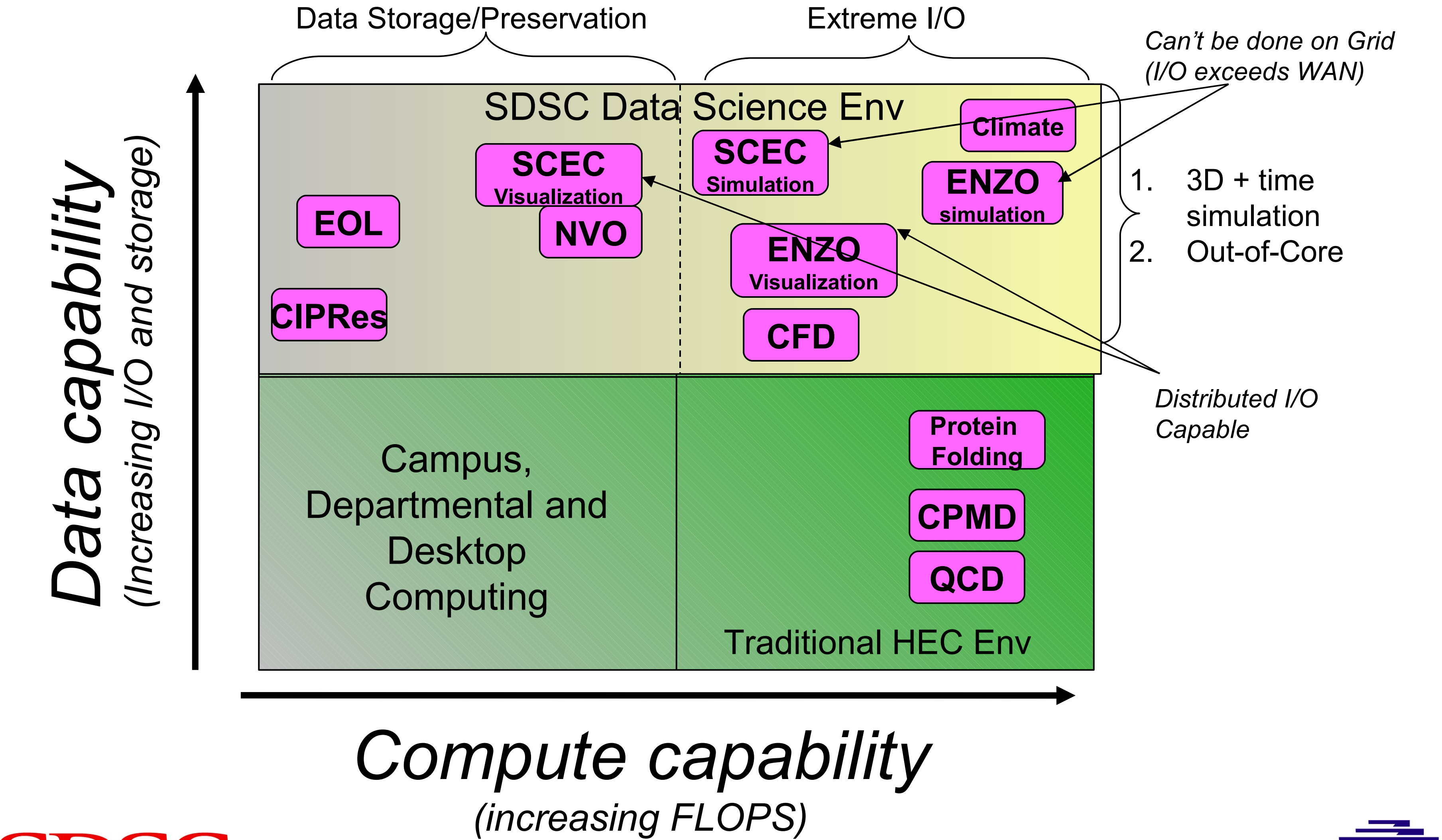
Abstract (1 of 2)

In 2005, the San Diego Supercomputer Center placed in production a large Global File System, consisting of over 500 TB of raw storage. Initial access to this resource was via the NSF TeraGrid, but this was later extended to non TeraGrid sites. In many cases, access rates to this centralized storage were faster than to local storage and authentication was handled by GSI certificates in a true Grid manner. Usage modes were both interesting and different from those anticipated, resulting in a major reconfiguration of the disk resource.

Abstract (2 of 2)

Overall acceptance has been startling, with sustained daily growth rates in the 1-3 TB range. SDSC is working with IBM to closely integrate this GPFS file system with the HPSS mass storage system and to extend GPFS to include local caching. The intention is to provide an apparently unlimited capacity high performance Global Storage Grid for scientific researchers across the US.

Computing: Data is extremely important!



Some Data numbers

- Enzo (Mike Norman) can output >25 TB in a single run at SDSC
- Earthquake simulations can produce > 50 TB
- “The entire NVO archive will contain about 100 terabytes of data to start, and grow to more than 10 petabytes by 2008.” Brian Krebs
- Expect Grid Computing growth to be overwhelming Data driven

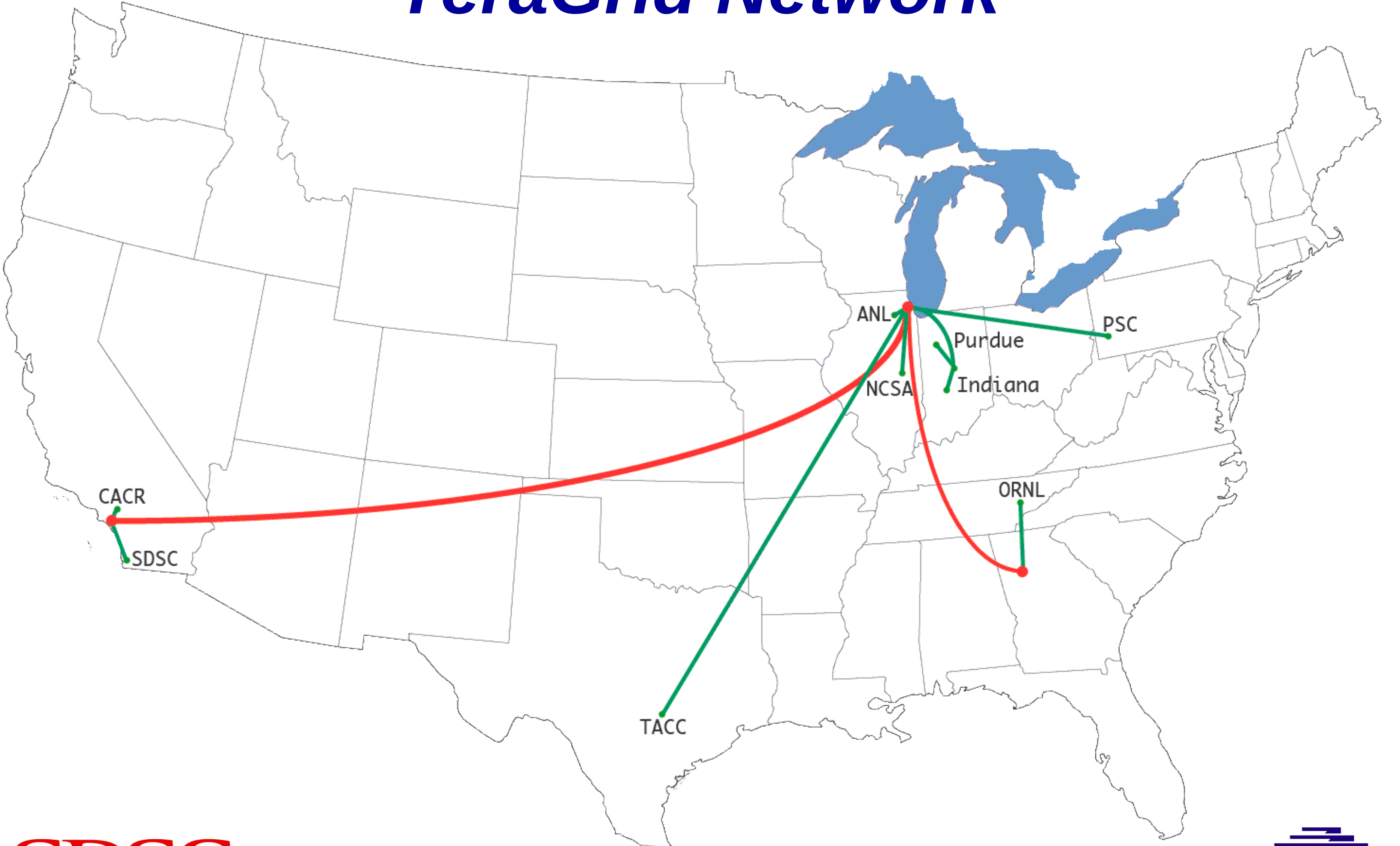
Data has a life cycle!

- **A computation is an event; data is a living thing that is conceived, created, curated, consumed, then deleted or archived, and/or forgotten.**
- **For success in data handling, all aspects must be considered and facilitated from the beginning within an integrated infrastructure**
- **Global Storage Grids are the only way to satisfy all user data requirements: Global File System + Grid Authentication + transparent, integrated Archival systems**

Local Data only Part of the Story

- TeraGrid users are not “Captive”; they move around from site to site and architecture to architecture
- Many users are part of multi-site collaborations, whose major intersection is via common data sets
- Essential to extend data’s reach across the USA (then the world!)

TeraGrid Network



Users ask for Performance- but Demand Convenience!

Global File Systems over WAN

- **Basis for some new Grids (DEISA)**
- **User transparency (TeraGrid roaming)**
- **On demand access to scientific data sets**
  - Share scientific data sets and results
  - Access scientific results from geographically distributed instruments and sensors in real-time
  - No copying of files to here and there and there…
  - What about UID, GID mapping?
- **Authentication**
  - Initially use World Readable DataSets and common UIDs for some users. GSI coming
- **On demand Data**
  - Instantly accessible and searchable
  - No need for local storage space
  - Need network bandwidth

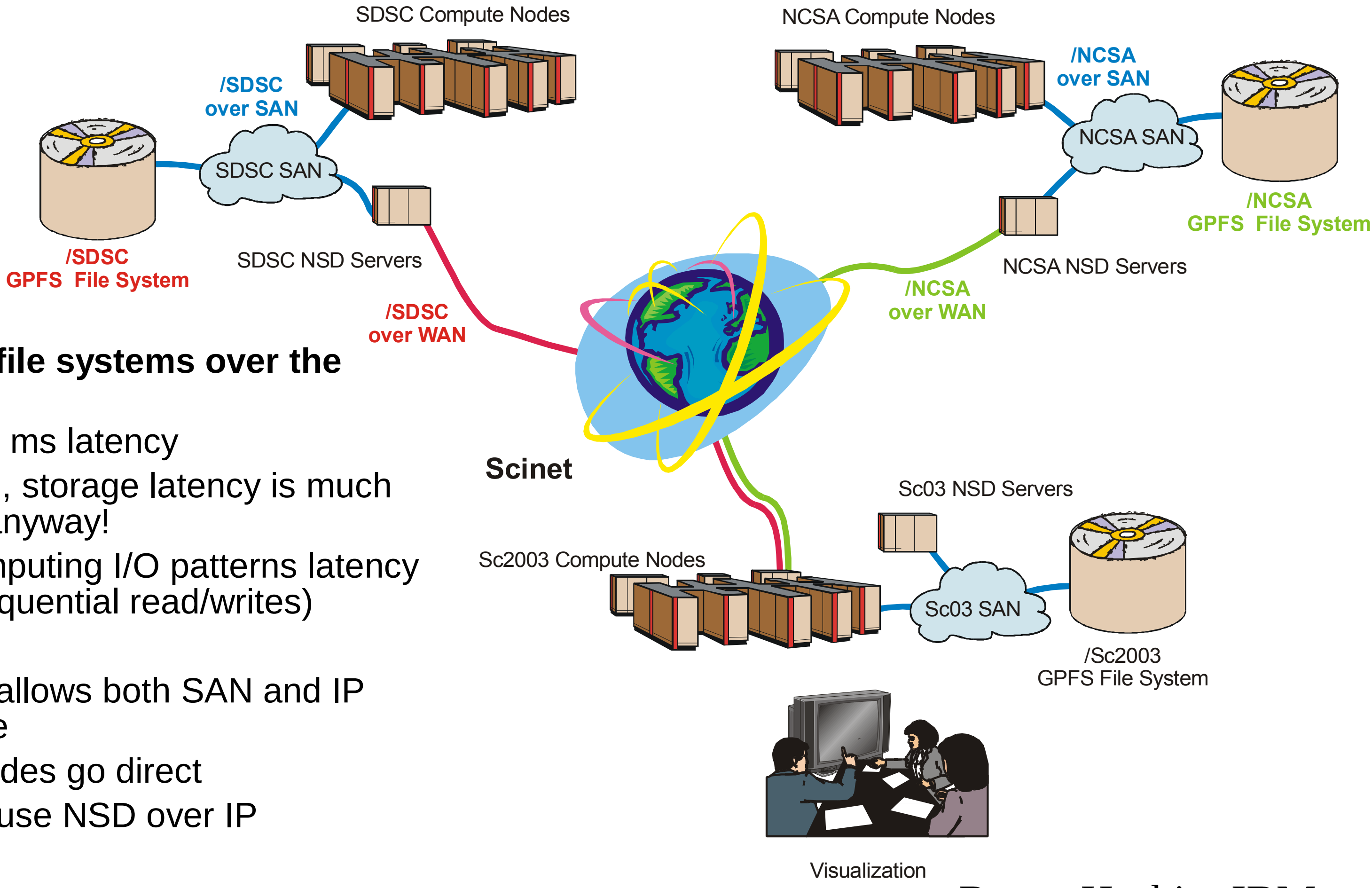
Approach for True User Acceptance

1) Determine the Paradigm
2) Demonstrate the Capability
3) Optimize the Performance
4) Implement in Infrastructure
5) Let the Users loose on it!

Specific Approach Acceptance

1) Forced the Idea at SC'02
2) Vendor Collaboration at SC'03
3) True Prototype at SC'04
4) Production in '05
5) Expanding in '06

Access to GPFS File Systems over the WAN during SC'03



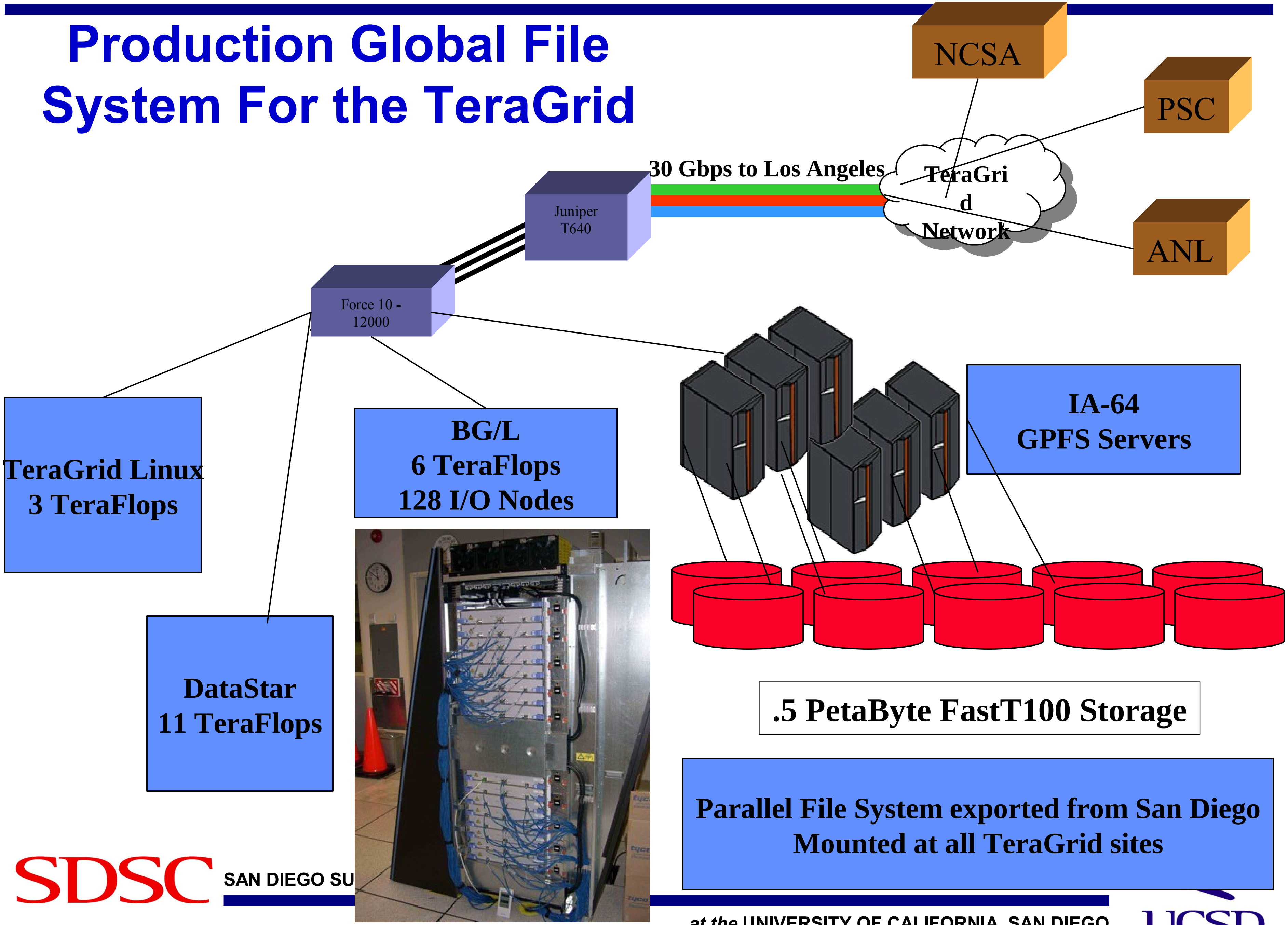
- **Goal: sharing GPFS file systems over the WAN**
  - WAN adds 10-60 ms latency
  - … but under load, storage latency is much higher than this anyway!
  - Typical supercomputing I/O patterns latency tolerant (large sequential read/writes)
- **New GPFS feature**
  - GPFS NSD now allows both SAN and IP access to storage
  - SAN-attached nodes go direct
  - Non-SAN nodes use NSD over IP
- **Work in progress**
  - Technology demo at SC03
  - Work toward possible product release

Roger Haskin, IBM

SDSC now serving 0.5 PB GFS disk

- Initially served across TeraGrid and mounted by ANL and NCSA
- Plan to start hosting large datasets for the scientific community
- One of first will be NVO, ~50 TB of Night Sky information; Read-Only dataset available for computation across TeraGrid
- Extend rapidly with other datasets

Production Global File System For the TeraGrid



TeraGrid Linux
3 TeraFlops

BG/L
6 TeraFlops
128 I/O Nodes

DataStar
11 TeraFlops

IA-64
GPFS Servers

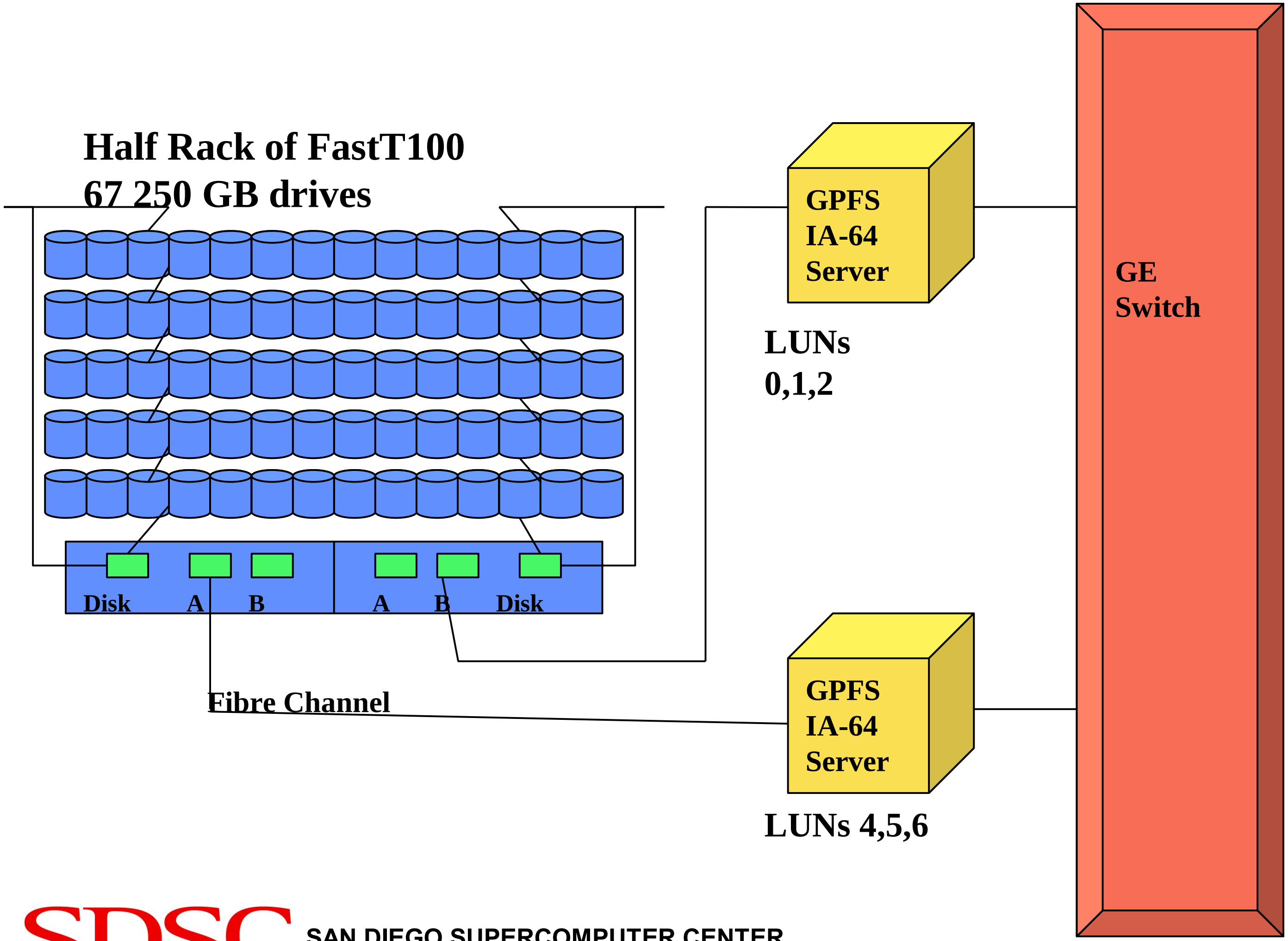
.5 PetaByte FastT100 Storage

Parallel File System exported from San Diego
Mounted at all TeraGrid sites

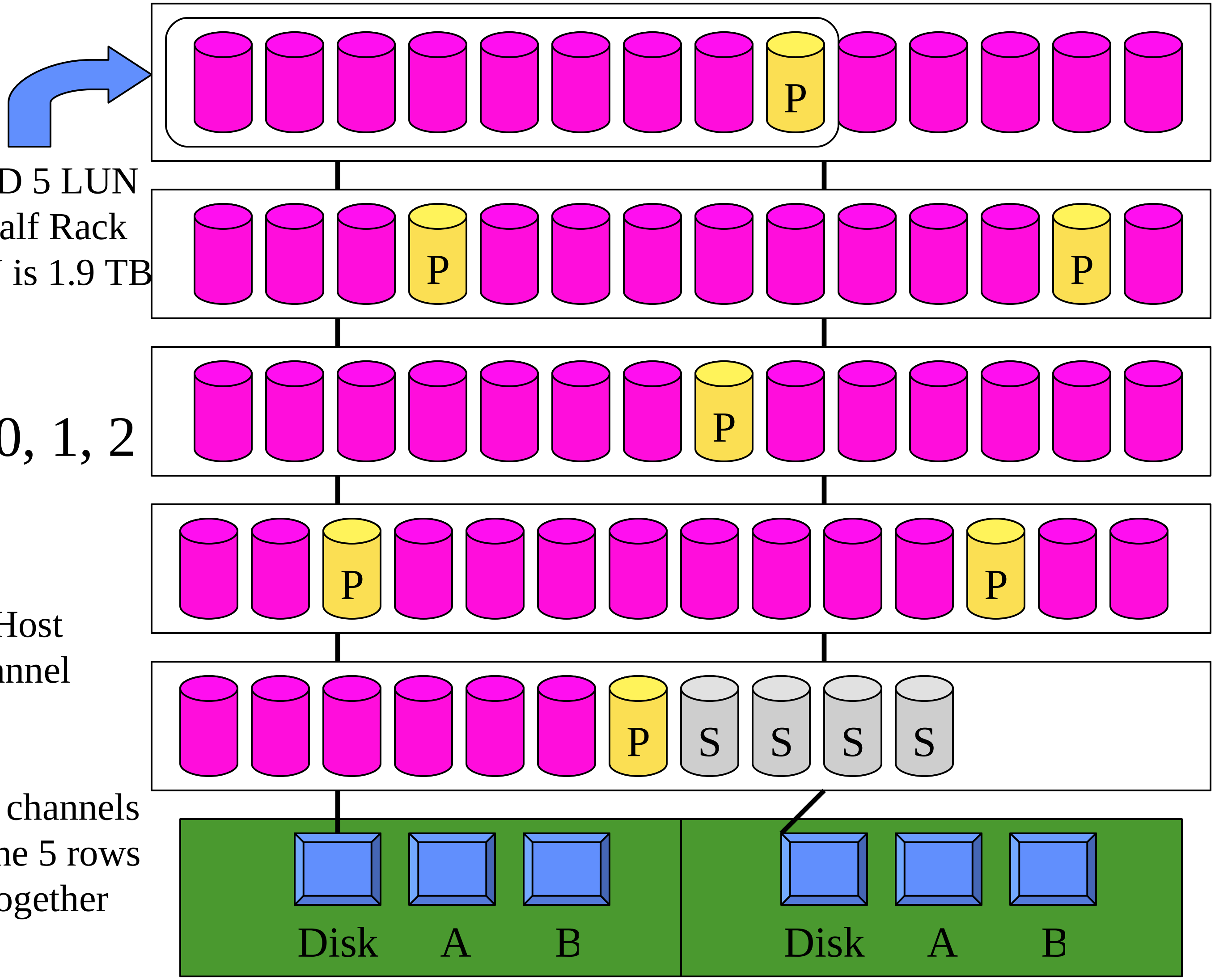
Initial Production Status

- All TeraGrid nodes at SDSC, ANL, and NCSA mounted the Global File System from SDSC
- 1K-2K nodes reported
- Sustained 3 GB/s reads between SDSC and NCSA
- Saw 9.7 Gb/s on a single 10 Gb/s link
- Large number of writes lead to Mirroring reconfiguration for safety

1/32 of GPFS-WAN File System



Initial Configuration of One Half Rack FastT-100 SATA Disk



8 + P RAID 5 LUN
7 LUNs/Half Rack
Each LUN is 1.9 TB

LUNs 0, 1, 2

A, B are Host Fibre Channel Ports

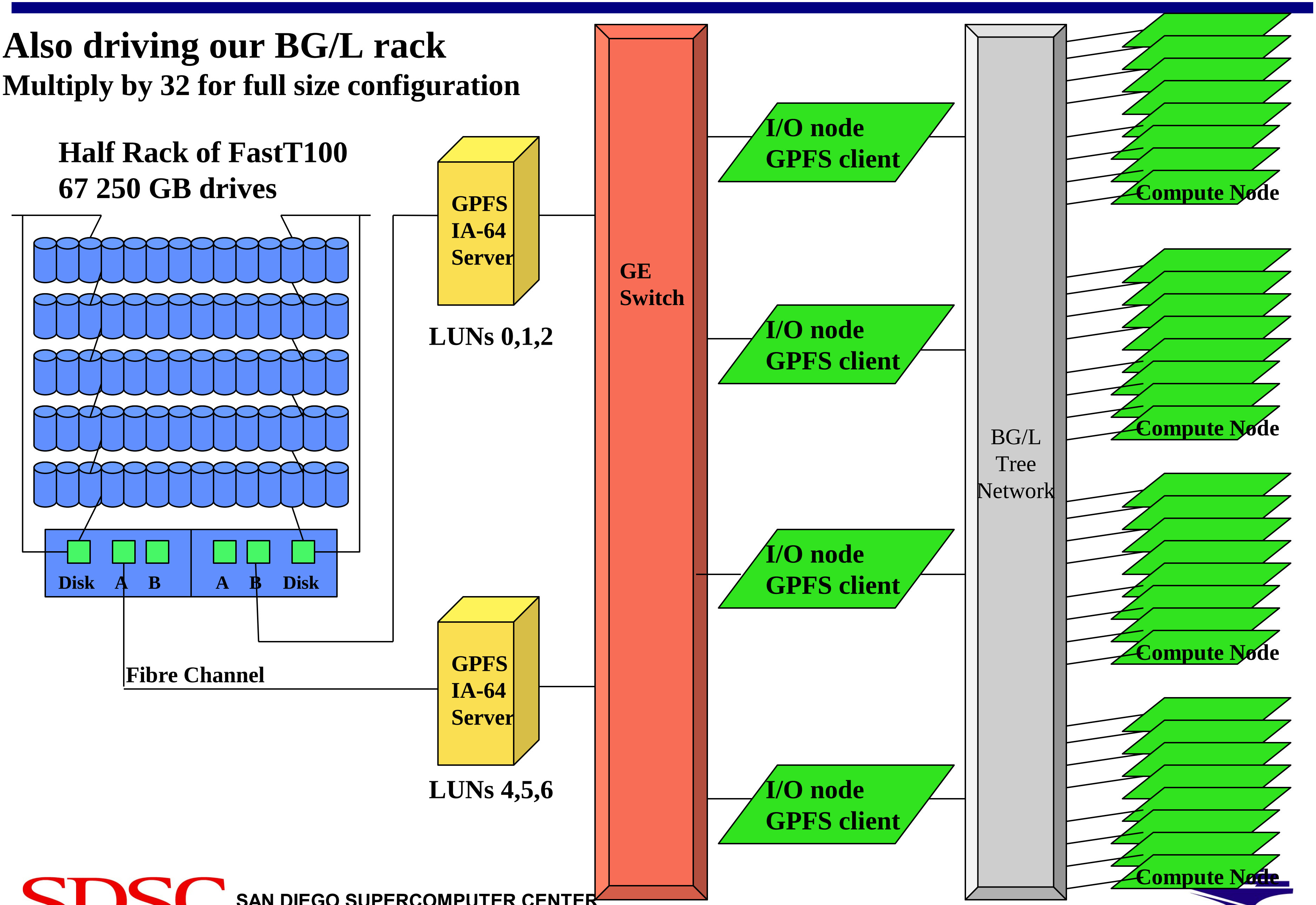
Two disk channels connect the 5 rows of disks together

There are 7 global spares "S" per half rack

There are 7 parity "P" disks per half rack

Also driving our BG/L rack

Multiply by 32 for full size configuration



Learning as We Go…

- **Original assumption was large Read-Only datasets**
  - Optimized File System for Reads
  - RAID5
  - No backups
- **Users Wrote directly to File System from Production Codes!**
  - We Improved write performance
  - Converted to RAID10
  - Took advantage of Hardware Mirroring

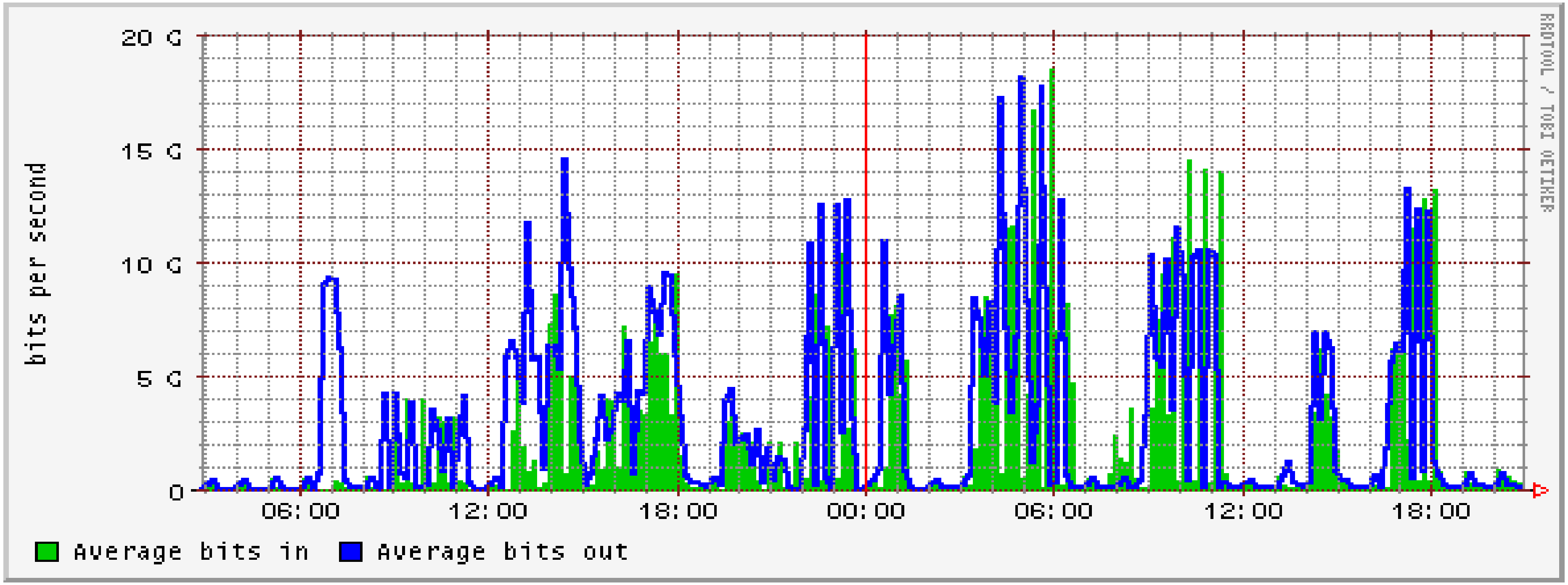
GSI-Based UID Mapping

- **General mechanisms for mapping between UIDs and globally unique names (GUNs)**
- **All Teragrid sites have Globus Distinguished Name to local username mappings in the “grid-mapfile”**
- **For GPFS-WAN, Globus Distinguished Names are used as globally unique identifiers.**
- **Collaboration between SDSC and IBM produced the specific GSI-based UID-mapping implementation.**

Authentication and Authorization

- **Authentication - GPFS 2.3+ uses RSA public/private keypairs, exchanged out-of-band, to establish trust relationships between clusters**
- **Authorization - Access control on a per-cluster, per-filesystem basis; option to remap remote root to non-privileged local UID.**

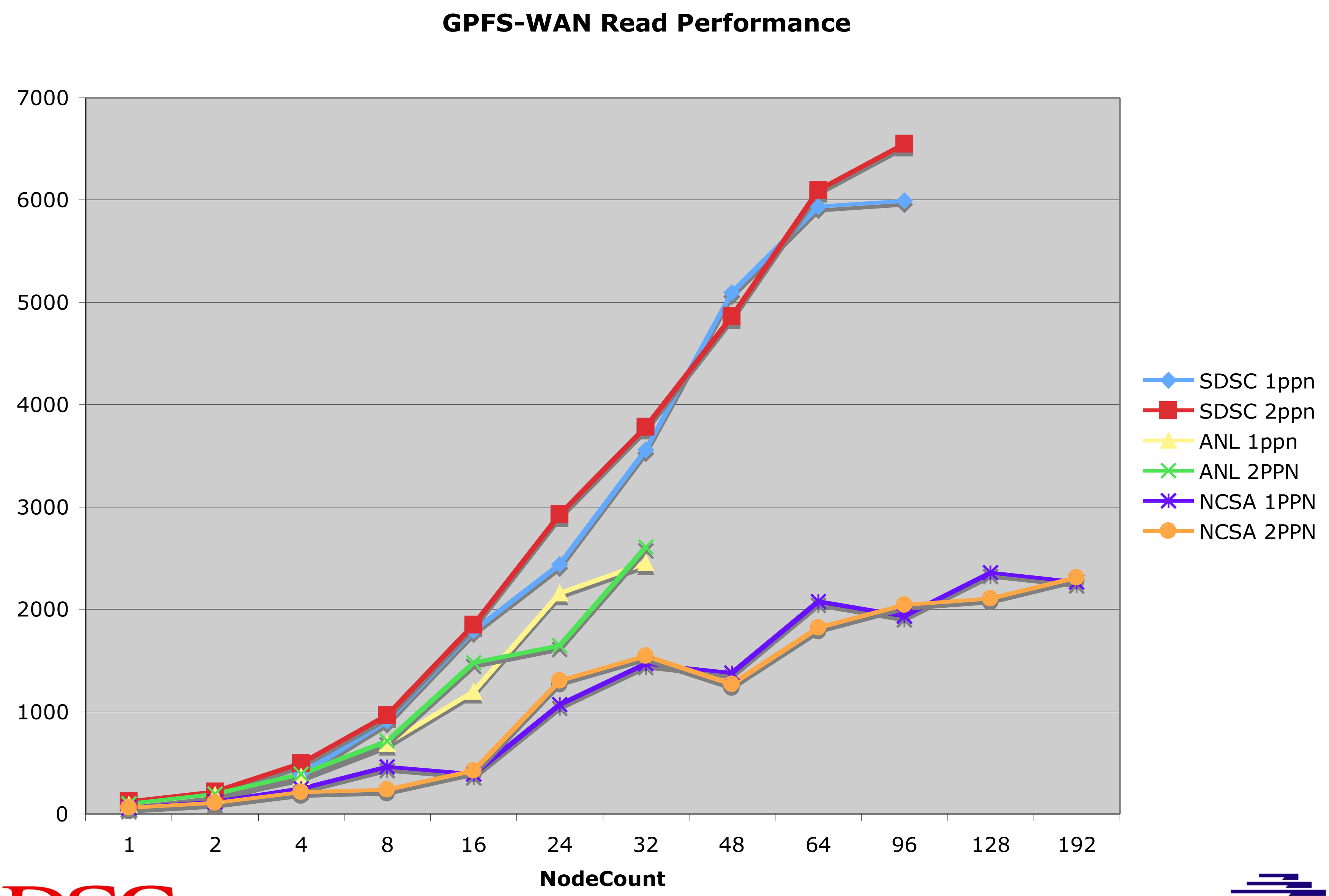
Network traffic during 20 Gb/s tests



Smallest Resolution is 5-minute average

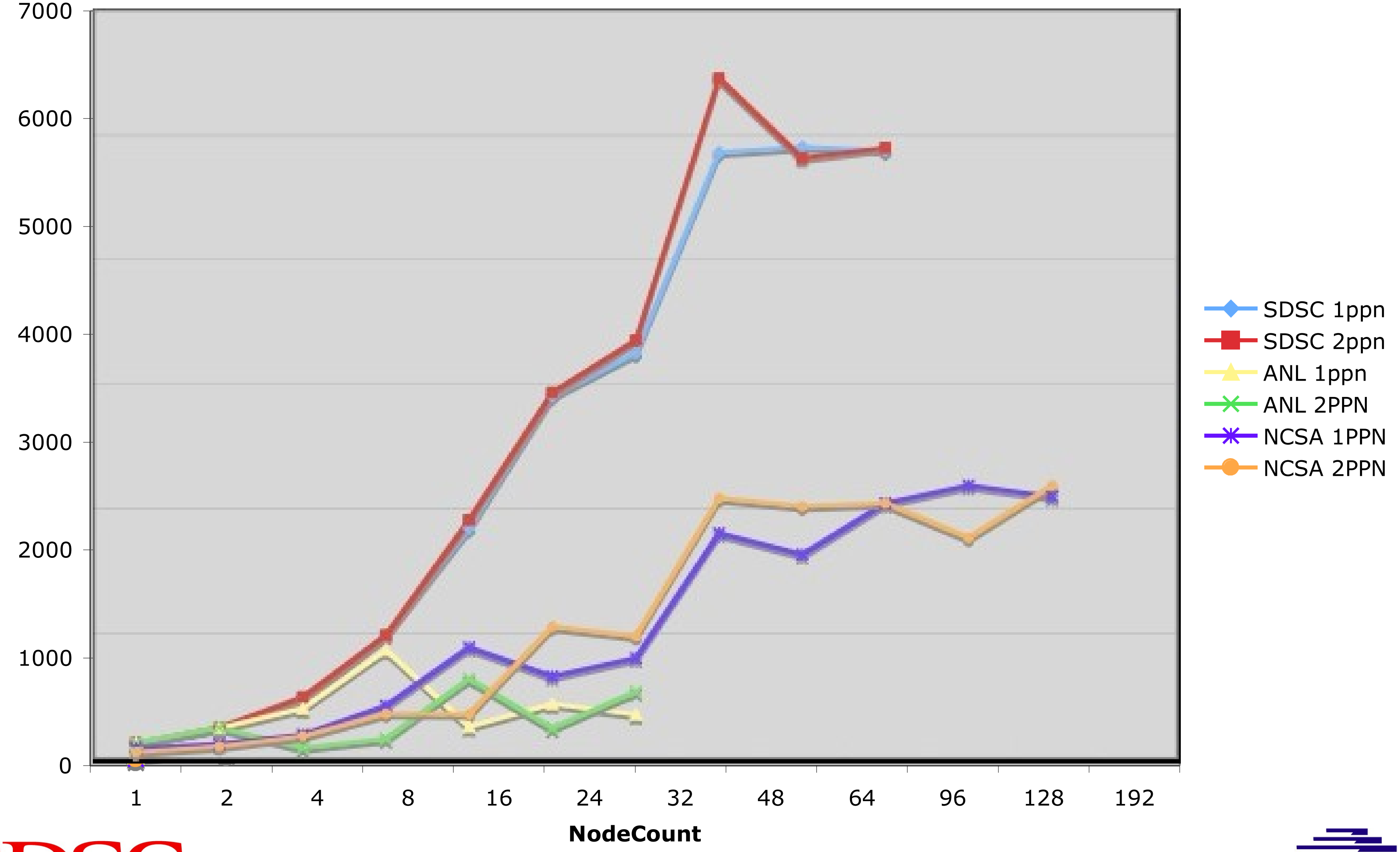
Network Tuning

- By default, GPFS uses the OS default sizes for both TCP send and receive buffers
- Default buffer sizes are often quite low relative to the bandwidth delay product between TeraGrid sites
- Specific GPFS config parameters to control TCP send and receive buffer sizes
- Untuned performance ~20MB/sec/node
- Tuning TCP window sizes allows ~100MB/sec/node


GPFS-WAN Read Performance
7000
6000
5000
4000
3000
2000
1000
0
1
2
4
8
16
24
32
48
64
96
128
192
NodeCount
SDSC 1ppn
SDSC 2ppn
ANL 1ppn
ANL 2PPN
NCSA 1PPN
NCSA 2PPN

GPFS-WAN Write Performance

Now comparable to GPFS-WAN Read Performance



Went into Production Oct ‘05

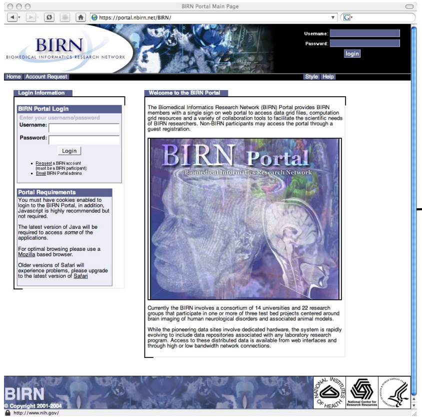
NVO: large read only data set (~50TB, very important)

- **Enzo application: writes to wan-gpfs from both SDSC and NCSA, data mining from multiple sites**
- **BIRN: compute at NCSA, store input and output at SDSC, visualize at Johns Hopkins, all without doing a single explicit network data transfer except to upload the initial dataset from the archives. Improved job throughput by sevenfold.**
- **gpfs-wan growth: 1-3 TB/day!**

ENZO /gpfs-wan performance

	SDSC local	SDSC Wan	NCSA local	NCSA wan
Wall time (s)	5929.7	5708.1	6043.7	7231.5
IO dump (s)	459.7	216.5	562.0	1283.5
IO min (s)	23.9	11.1	26.7	60.2
IO max (s)	26.3	13.3	39.7	119.4
IO mean (s)	25.5	12.0	31.2	71.3
Rate (MB/s)	537	1142	439	192

BIRN and Telescience Advancing Biomedical Research on TeraGrid



BIRN Portal Main Page

https://portal.nbirn.net/BIRN/

BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK

Username:
Password:
login

Home | Account Request | Style | Help

Login Information

BIRN Portal Login

Enter your username/password

Username:

Password:

Login

- Request a BIRN account (must be a BIRN participant)
- Email BIRN Portal admins

Portal Requirements

You must have cookies enabled to login to the BIRN Portal, in addition, Javascript is highly recommended but not required.

The latest version of Java will be required to access *some* of the applications.

For optimal browsing please use a Mozilla based browser.

Older versions of Safari will experience problems, please upgrade to the latest version of Safari

Welcome to the BIRN Portal

The Biomedical Informatics Research Network (BIRN) Portal provides BIRN members with a single sign on web portal to access data grid files, computation grid resources and a variety of collaboration tools to facilitate the scientific needs of BIRN researchers. Non-BIRN participants may access the portal through a guest registration.

BIRN Portal
Biomedical Informatics Research Network

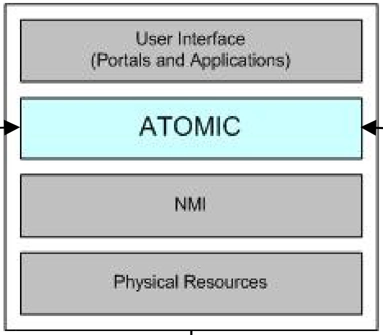
Currently the BIRN involves a consortium of 14 universities and 22 research groups that participate in one or more of three test bed projects centered around brain imaging of human neurological disorders and associated animal models.

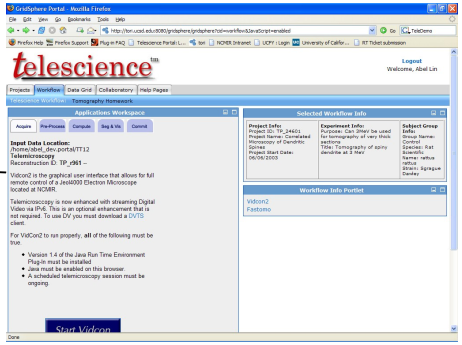
While the pioneering data sites involve dedicated hardware, the system is rapidly evolving to include data repositories associated with any laboratory research program. Access to these distributed data is available from web interfaces and through high or low bandwidth network connections.

BIRN
© Copyright 2001-2004

National Institutes of Health | National Center for Research Resources

http://www.nih.gov/





GridSphere Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://tori.ucsd.edu:8080/gridsphere/gridsphere?cid=workflow&JavaScript=enabled

Firefox Help | Firefox Support | Plug-in FAQ | Telescience Portal: L... | tori | NCMIR Intranet | UCFY : Login | University of Califor... | RT Ticket submission

telescience™

Logout
Welcome, Abel Lin

Projects | Workflow | Data Grid | Collaboratory | Help Pages

Telescience Workflow: Tomography Homework

Applications Workspace

Acquire | Pre-Process | Compute | Seg & Vis | Commit

Input Data Location:
/home/abel_dev.portal/TT12
Telemicroscopy
Reconstruction ID: **TP_r961** --

Vidcon2 is the graphical user interface that allows for full remote control of a Jeol4000 Electron Microscope located at NCMIR.

Telemicrosccopy is now enhanced with streaming Digital Video via IPv6. This is an optional enhancement that is not required. To use DV you must download a DVTS client.

For VidCon2 to run properly, **all** of the following must be true.

- Version 1.4 of the Java Run Time Environment Plug-In must be installed
- Java must be enabled on this browser.
- A scheduled telemicroscopy session must be ongoing.

Start Vidcon

Selected Workflow Info

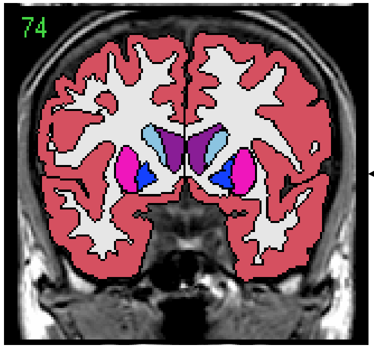
Project Info:	**Experiment Info:**	**Subject Group Info:**
Project ID: TP_24601 Project Name: Correlated Microscopy of Dendritic Spines Project Start Date: 06/06/2003	Purpose: Can 3MeV be used for tomography of very thick sections Title: Tomography of spiny dendrite at 3 MeV	Group Name: Control Species: Rat Scientific Name: rattus rattus Strain: Sgrague Dawley

Workflow Info Portlet

Vidcon2
Fastomo

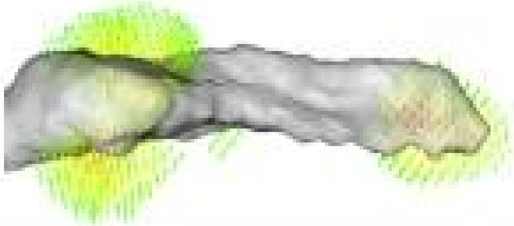
Done

TERAGRID



Expectation/Maximization (EM) Segmentation

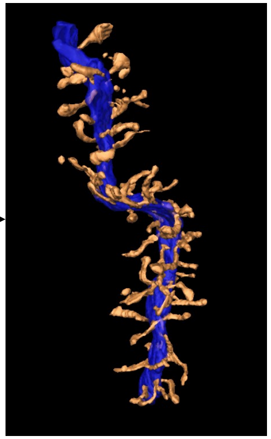
Brigham and Women's Hospital



Large Deformation Diffeomorphic Metric Mapping (LDDMM)

John Hopkins University

(interested in GPFS WAN)



Transform Based Backprojection for Volume Reconstruction (TxBR)

University of California, S

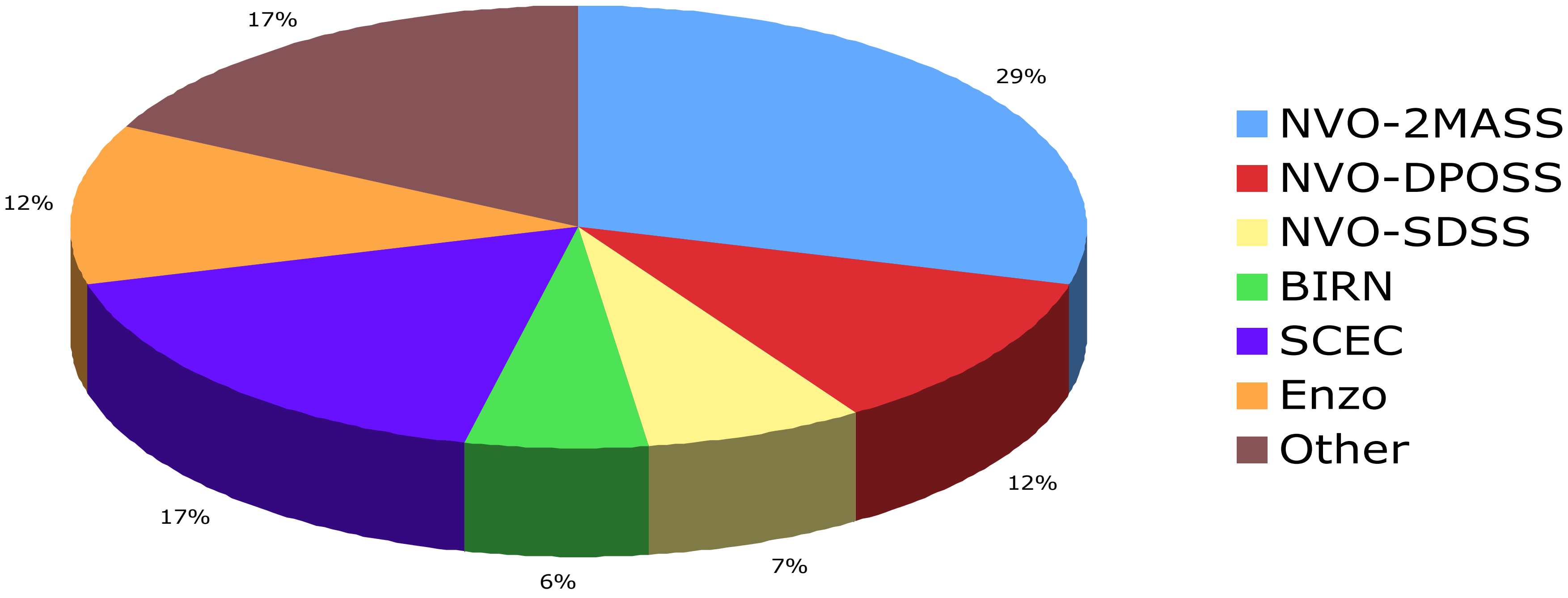
(interested in GPF

Usage Patterns:

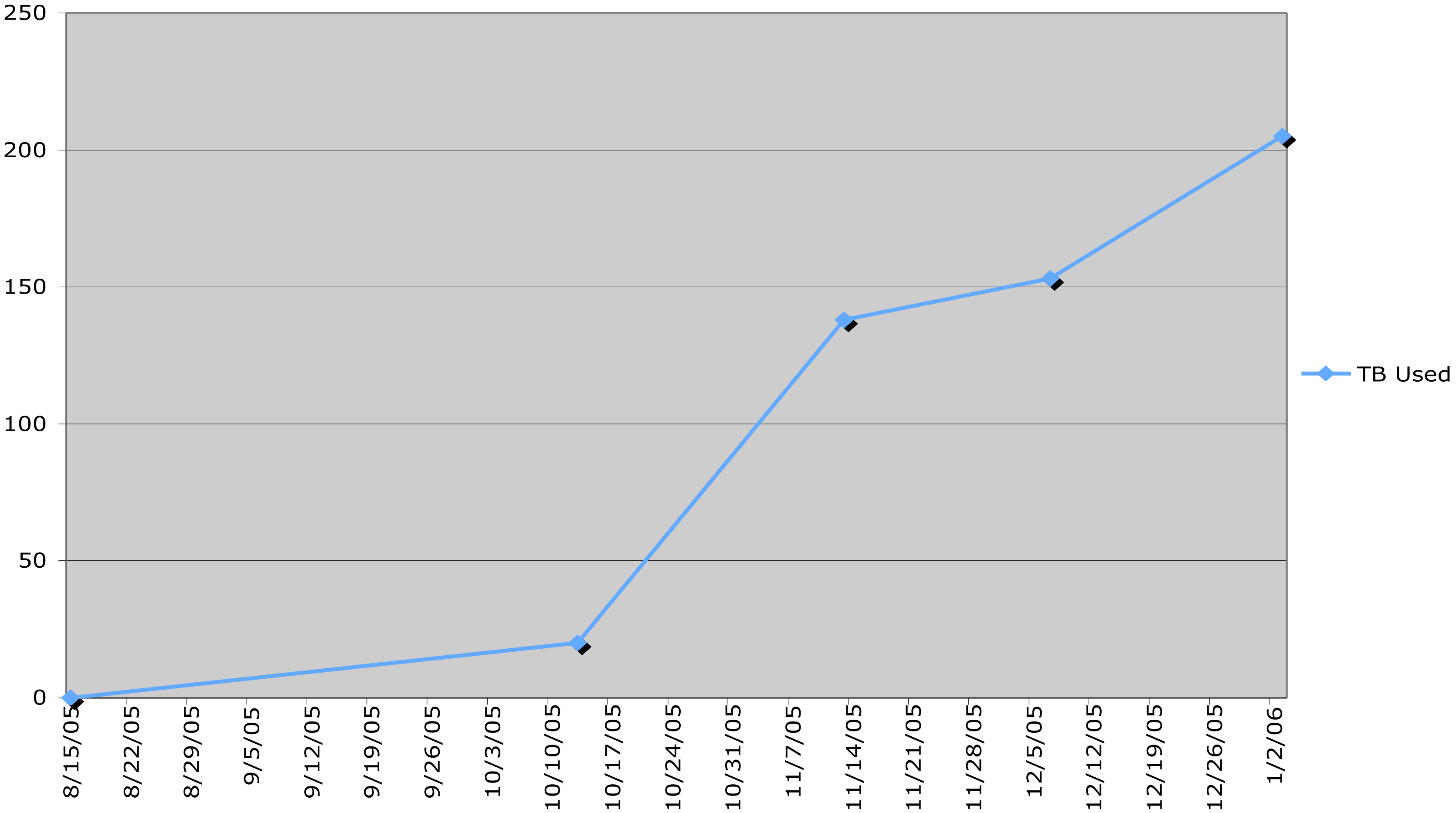
- **Expected usage: Large, Read-Only Datasets; improved efficiency, ubiquitous access, updates. E.g., National Virtual Observatory "Night Sky"**
- **Unexpected usage (1): Community codes writing large datasets from HPC system for data mining & visualization at member sites (Enzo, SCEC)**
- **Unexpected usage(2): "Pipelining" multiple resources used at several sites on same data by a single user. E.g., BIRN observational data->processing-> visualization, 70->1,000+ jobs/day**

Snapshot of Usage by Project

Three distinct usage modes:
1) Large, read-only datasets,
2) writing common datasets used by widespread community,
3) pipelining thru diverse resources



GPFS-WAN Usage (TB) by Date

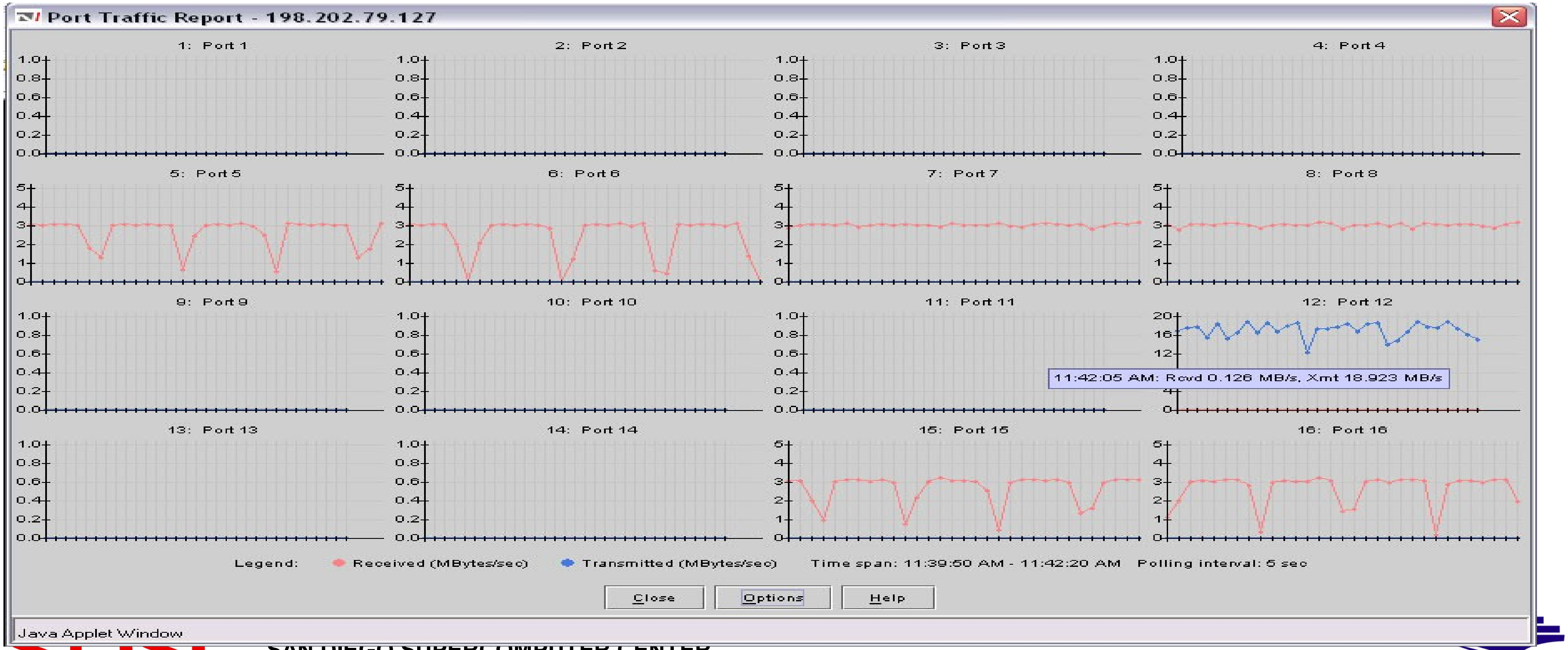


Future needs:

- **Not all applications are seeing performance as high as we would like: need to investigate**
- **Disk filled up an frightening rate: need a) more disk, and b) transparent integration with HPSS archival system: Prototype system running at SDC**
- **Have to deal with Group ID problem**
- **Would like to get away from being a "canary in a coal mine"!**

Pittsburgh Supercomputing Center and SDSC Integrating archives

- Used FC/IP encoding via WAN-SAN to attach 6 SDSC tape drives to PSC’s DMF Archival System
- STK Silo installed at PSC solely for SDSC remote second copies



Port Traffic Report - 198.202.79.127

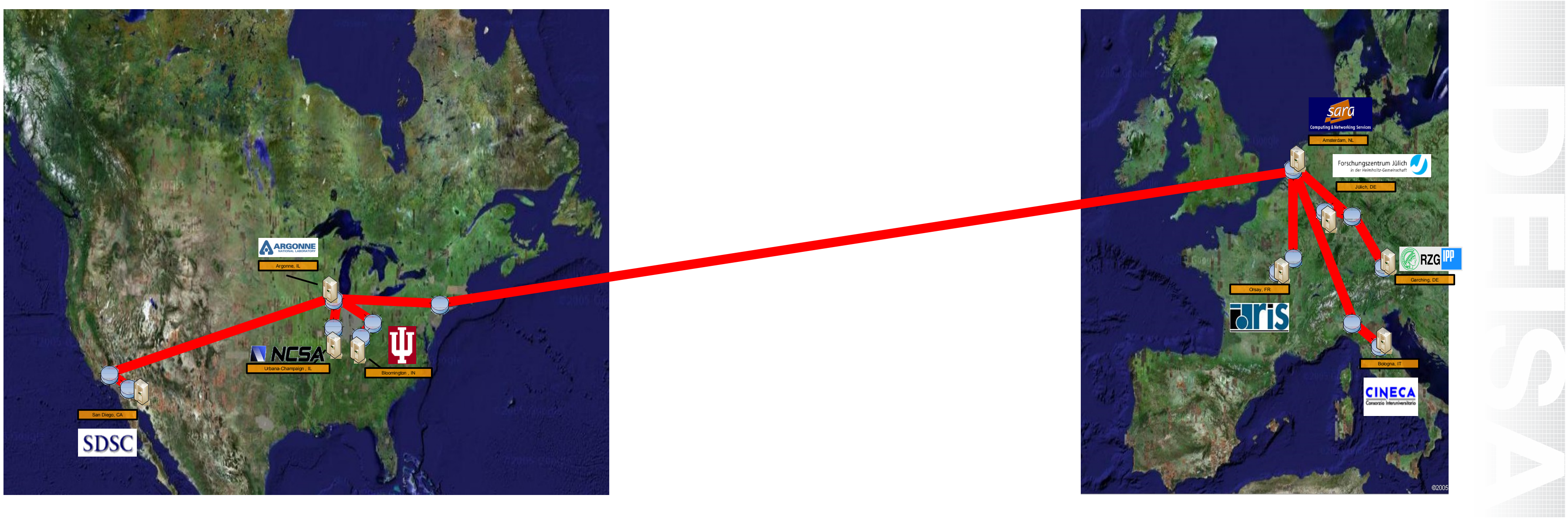
11:42:05 AM: Rcvd 0.126 MB/s, Xmt 18.923 MB/s

Legend: ● Received (MBytes/sec) ● Transmitted (MBytes/sec) Time span: 11:39:50 AM - 11:42:20 AM Polling interval: 5 sec

Java Applet Window

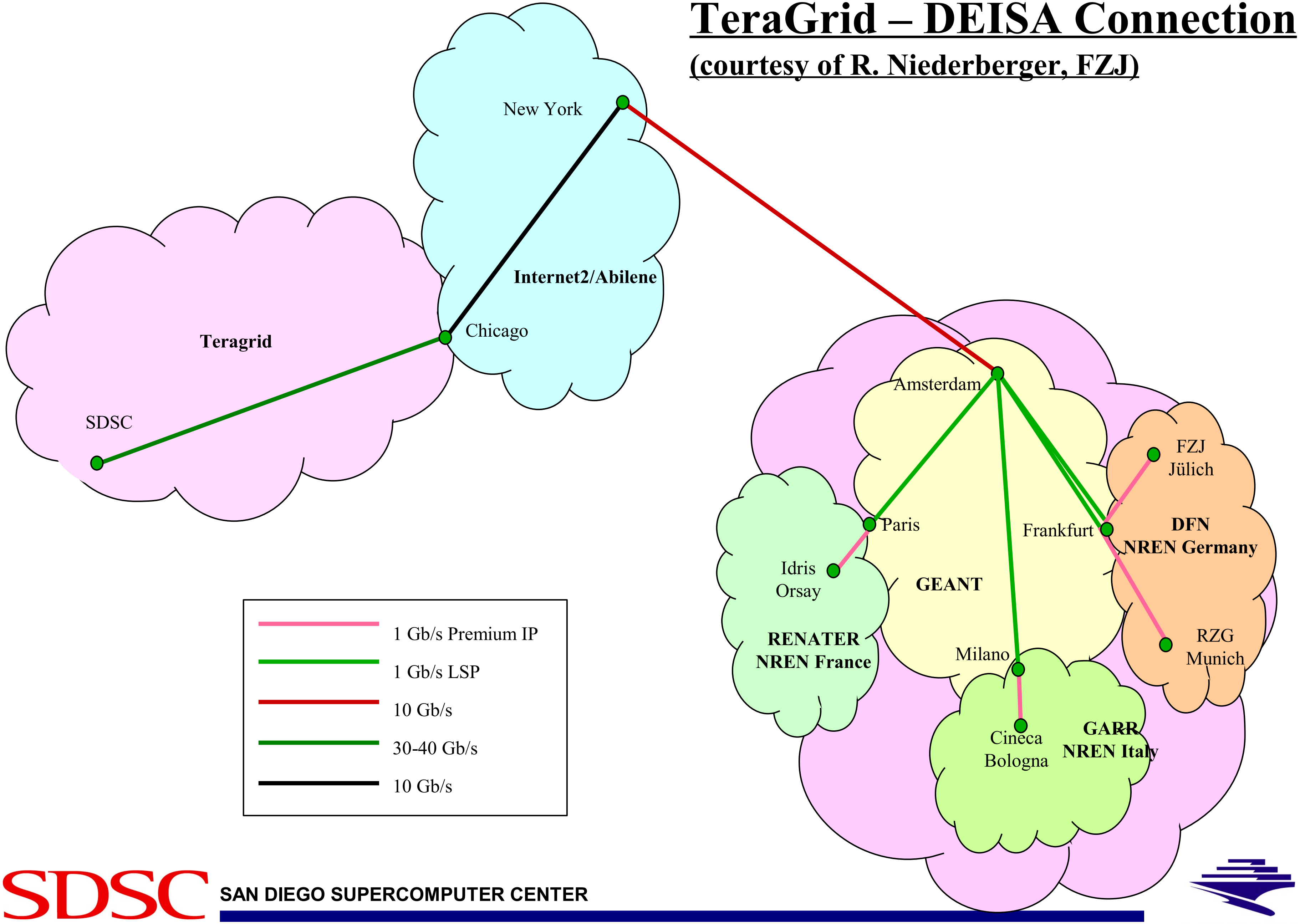
TeraGrid – DEISA GFS connectivity

- DEISA European Grid runs GPFS in Production at 4 sites. For SC'05 all 4 sites mounted wan-gpfs from SDSC, and SDSC mounted all 4 DEISA GPFS file systems

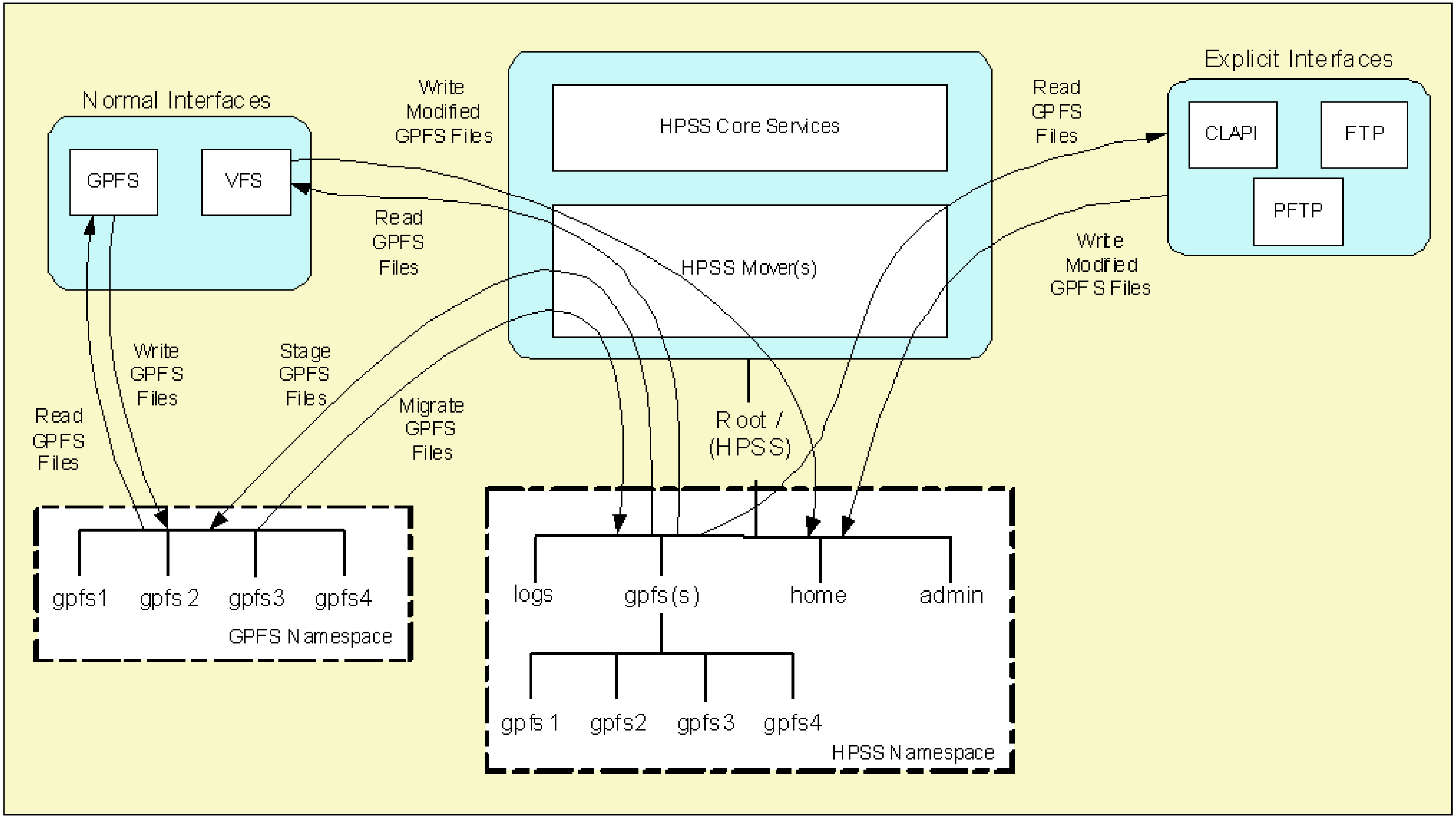

ARGONNE NATIONAL LABORATORY
Argonne, IL
NCSA
Urbana-Champaign , IL
Bloomington , IN
San Diego, CA
SDSC
sara Computing & Networking Services
Amsterdam, NL
Forschungszentrum Jülich in der Helmholtz-Gemeinschaft
Jülich, DE
RZG IPP
Garching, DE
Orsay, FR
idris
Bologna, IT
CINECA Consorzio Interuniversitario
©2005
DEISA

TeraGrid – DEISA Connection

(courtesy of R. Niederberger, FZJ)



GPFS-HPSS Integration (thanks to HPSS/IBM)



Working with IBM/HPSS/NERSC on defining HPSS-GPFS integration parameters

- Expect to have several different migration/file system protocols:
- Automatic migration and global availability for source trees, original datasets
- High-performance, cached, non-migrated temporary storage
- Short-term derived data on (mirrored?) disk
- Long-term derived data migrated to Archival
- Prototypes at SDSC and NERSC