

SAN, HPSS, Sam-QFS, and GPFS technology in use at SDSC



Bryan Banister, San Diego Supercomputing Center

bryan@sdsc.edu

Manager, Storage Systems and Production Servers

Production Services Department

Big Computation

- **10.7 TF IBM Power 4+ cluster (DataStar)**
 - 176 P655s with 8 x 1.5 GHz Power 4+
 - 11 P690s with 32 x 1.3 and 1.7 GHz Power 4+
 - Federation switch
- **3 TF IBM Itanium cluster (TeraGrid)**
 - 256 Tiger2 with 2 x 1.3 GHz Madison
 - Myrnet switch
- **5.x TF IBM BG/L (BagelStar)**
 - 128 I/O nodes
 - Internal switch

Big Data

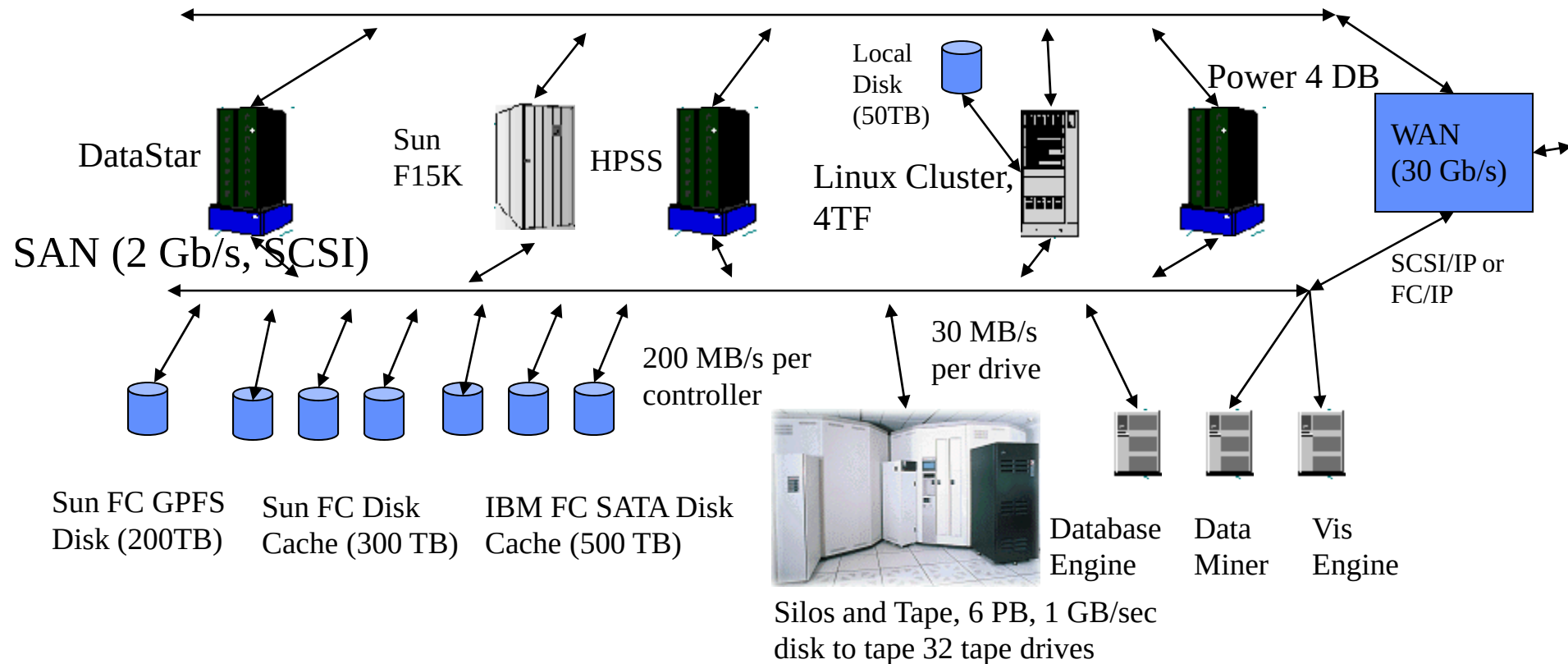
- **540 TB of Sun Fibre Channel SAN-attached Disk**
- **500 TB of IBM DS4100 SAN-attached SATA Disk**
- **11 Brocade 2 Gb/s 128-port FC SAN switches (over 1400 ports)**
- **5 STK Powderhorn Silos (30,000 tape slots)**
- **32 STK 9940-B tape drives (30 MB/s, 200 GB/cart)**
- **8 STK 9840 tape drives (mid-point load)**
- **16 IBM 3590E tape drives**
- **6 PB (uncompressed capacity)**
- **HPSS and Sun SAM-QFS Archival systems**

SDSC Machine Room Data Architecture

- Philosophy: enable SDSC configuration to serve the grid as data center

LAN (10 GbE, multiple GbE, TCP/IP)

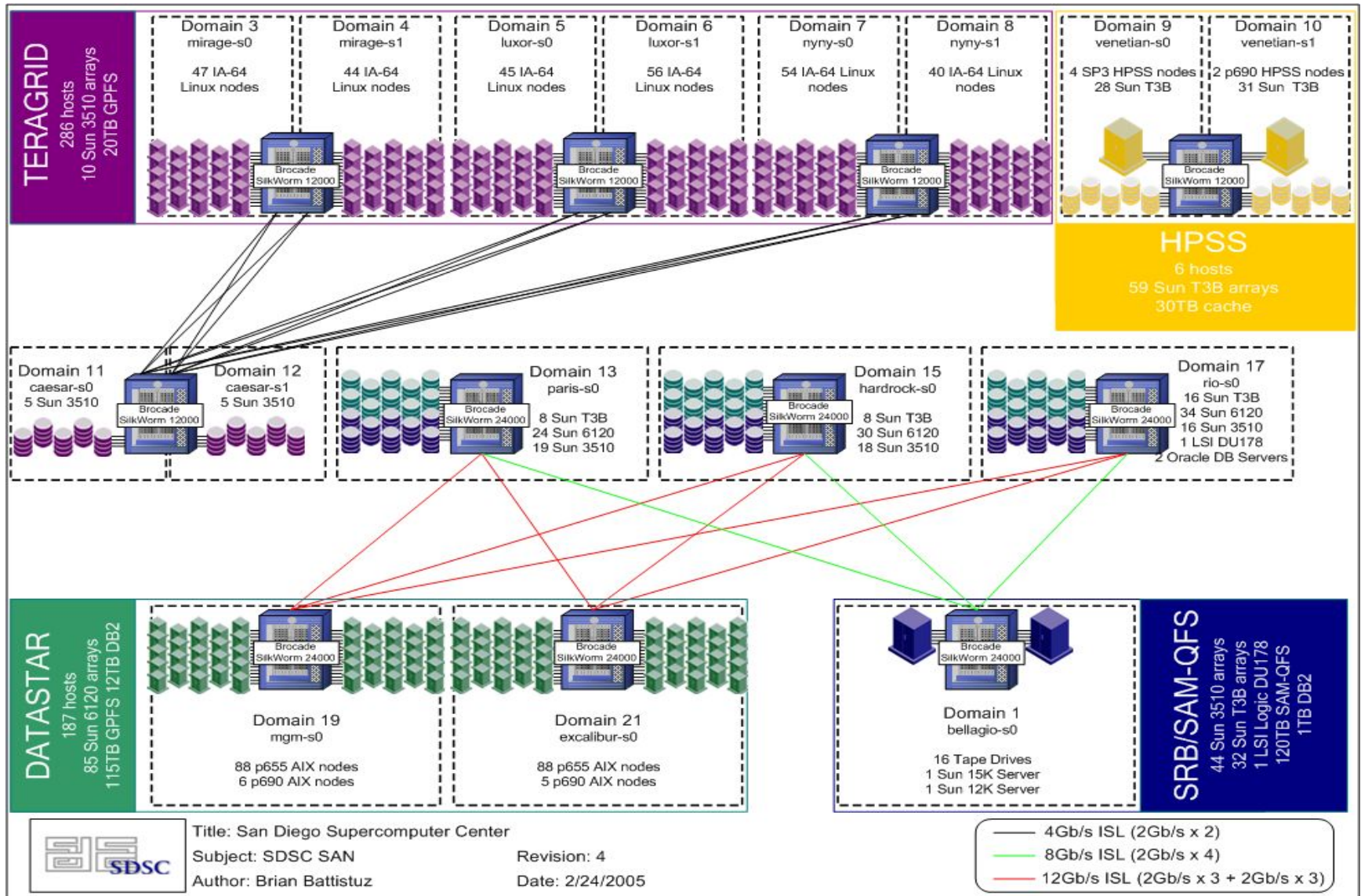
- 1 PB disk (500 FC and 500 SATA)
- 6 PB archive
- 1 GB/s disk-to-tape
- Optimized support for DB2 /Oracle



SAN Foundation

- **Goals:**
 - One fabric, centralized storage access
 - Fast access to storage over FC
- **It's just to big**
 - 1400 ports and 1000 devices
 - All systems need coordinated downtime for maintenance
 - One outage takes all systems down
 - Zone database size limitations
 - Control processor not up to the task in older 12K units

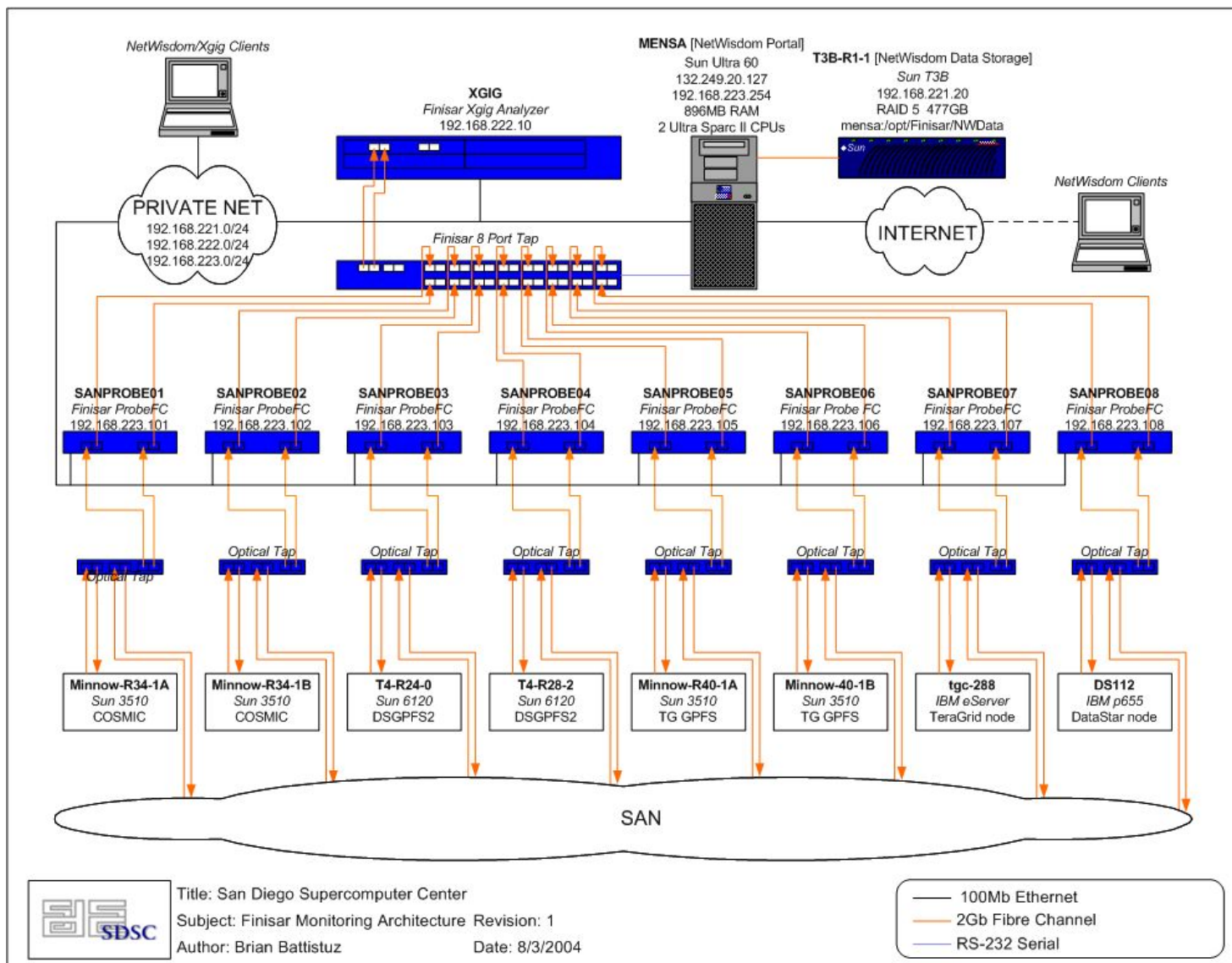
Storage Area Network at SDSC



Finisar XGig Analyzer and Netwisdom

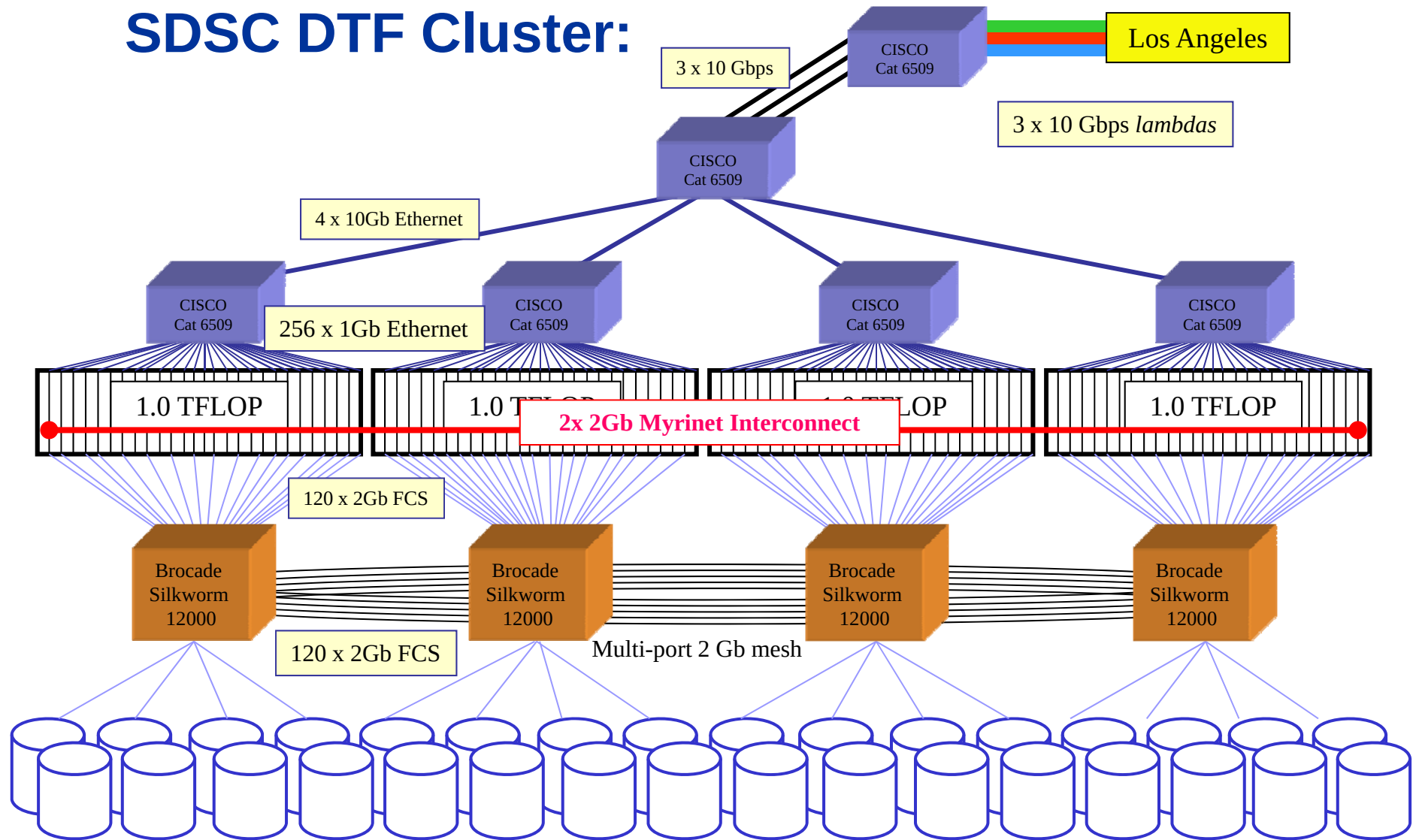
- **Large SAN fabric**
 - Many vendors: IBM, Sun, QLogic, Brocade, Emulex, Dot Hill...
 - Need both failure and systemic troubleshooting tools
- **XGig**
 - Shows all FC transactions on the wire
 - Excellent filtering and tracking capabilities
 - Expert software makes anyone an FC expert
 - Accepted in industry, undeniable proof
- **NetWisdom**
 - Traps for most FC events
 - Long term trending

Finisar setup at SDSC



GPFS

SDSC DTF Cluster:



~30 Sun Minnow FC Disk Arrays (65 TB total)

NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

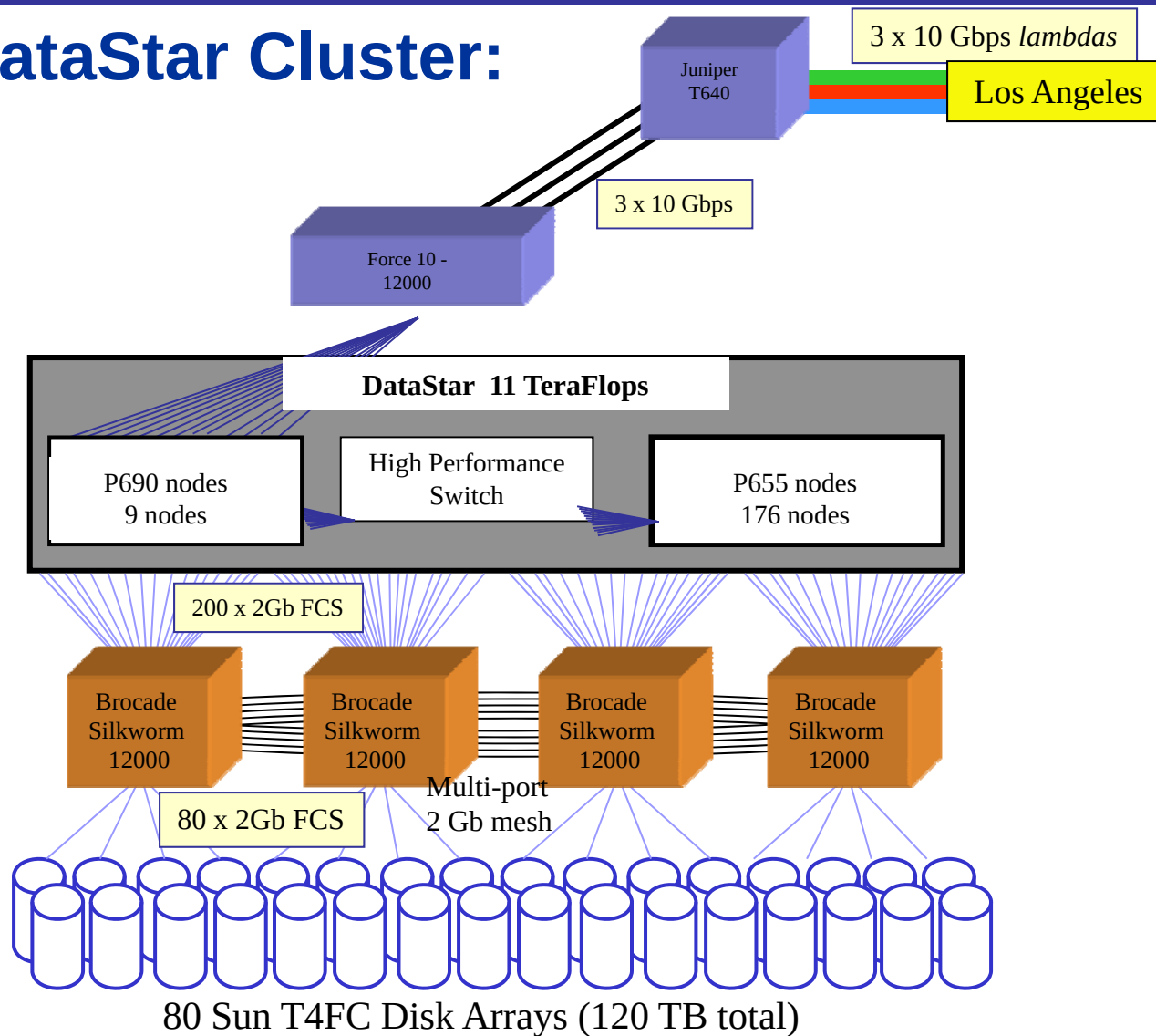


SAN DIEGO SUPERCOMPUTER CENTER



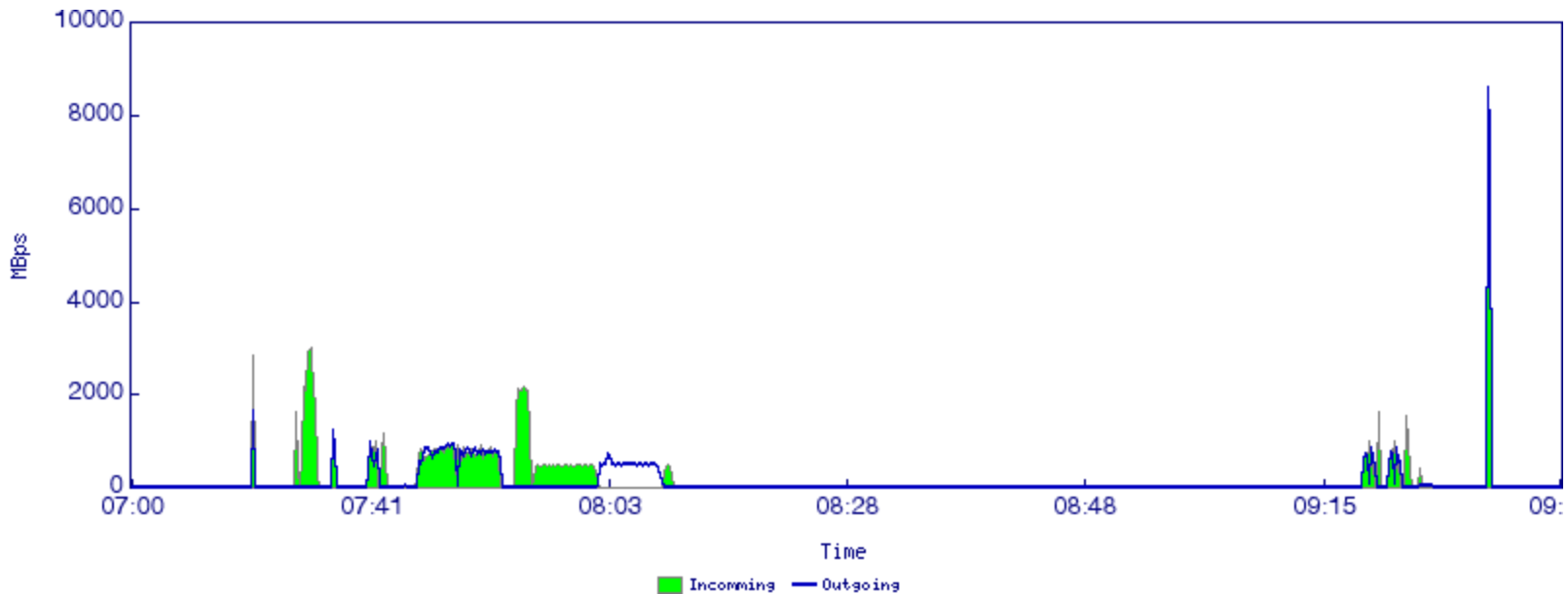
GPFS Continued

SDSC DataStar Cluster:



GPFS Performance Tests

- **~ 15 GB/s achieved at SC'04**
 - Won SC'04 StorCloud Bandwidth Challenge
 - Set new Terabyte Sort record, less than 540 Seconds!



Data and Grid Computing

- **Normal Paradigm: Roaming Grid Job will GridFTP requisite Data to its chosen location**
- **Adequate approach for small-scale jobs, e.g., Cycle Scavenging on University Network Grids**
- **Supercomputing Grids may require 10-50 TB Datasets!**
- **Whole Communities may use Common Datasets: Efficiency and Synchronization are essential**
- **We propose a Centralized Data Source**

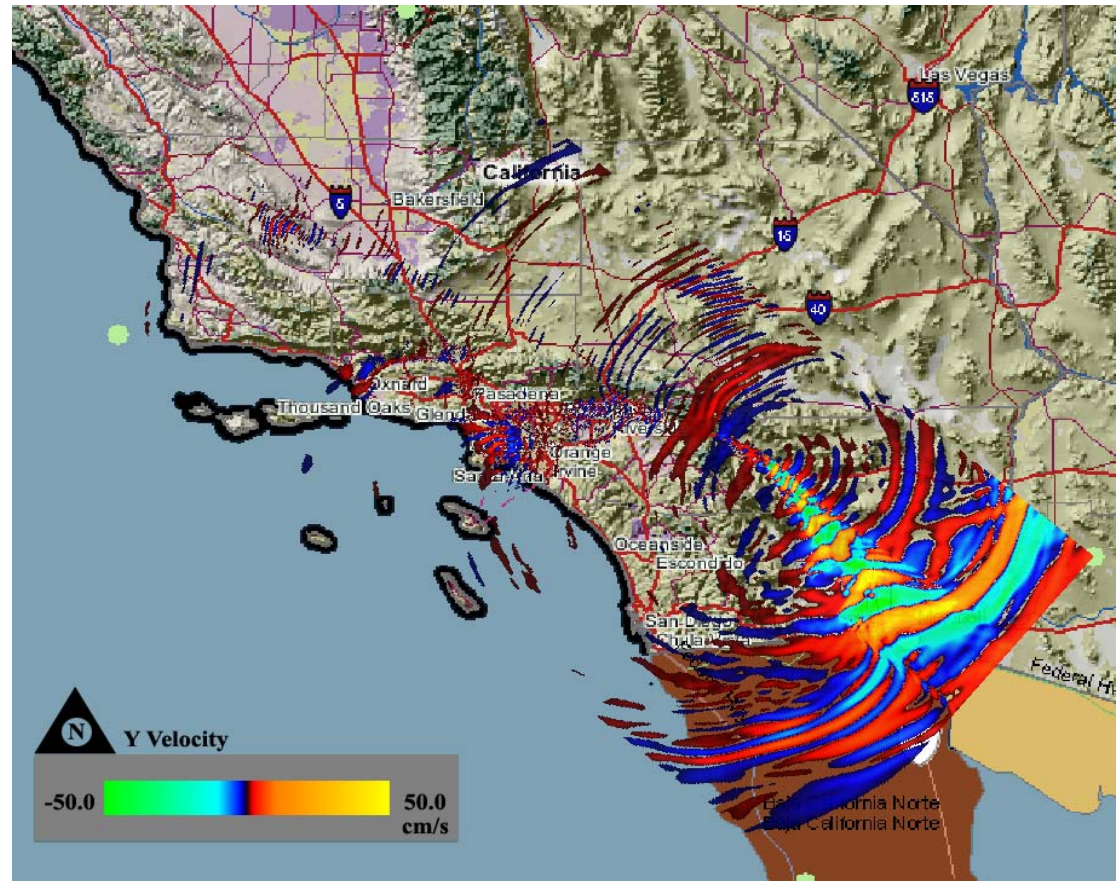
Example of Data Science: SCEC

Southern California Earthquake Center / Community Modeling Environment Project

Simulation of Seismic Wave Propagation of a Magnitude 7.7 Earthquake on San Andreas Fault

–PPIs: Thomas Jordan, Bernard Minster, Reagan Moore, Carl Kesselman

– This was chosen as a **SDSC Strategic Community Collaborations (SCC)** project and resulted in intensive participation by SDSC computational experts to optimized and enhance code both MPI I/O management and checkpointing



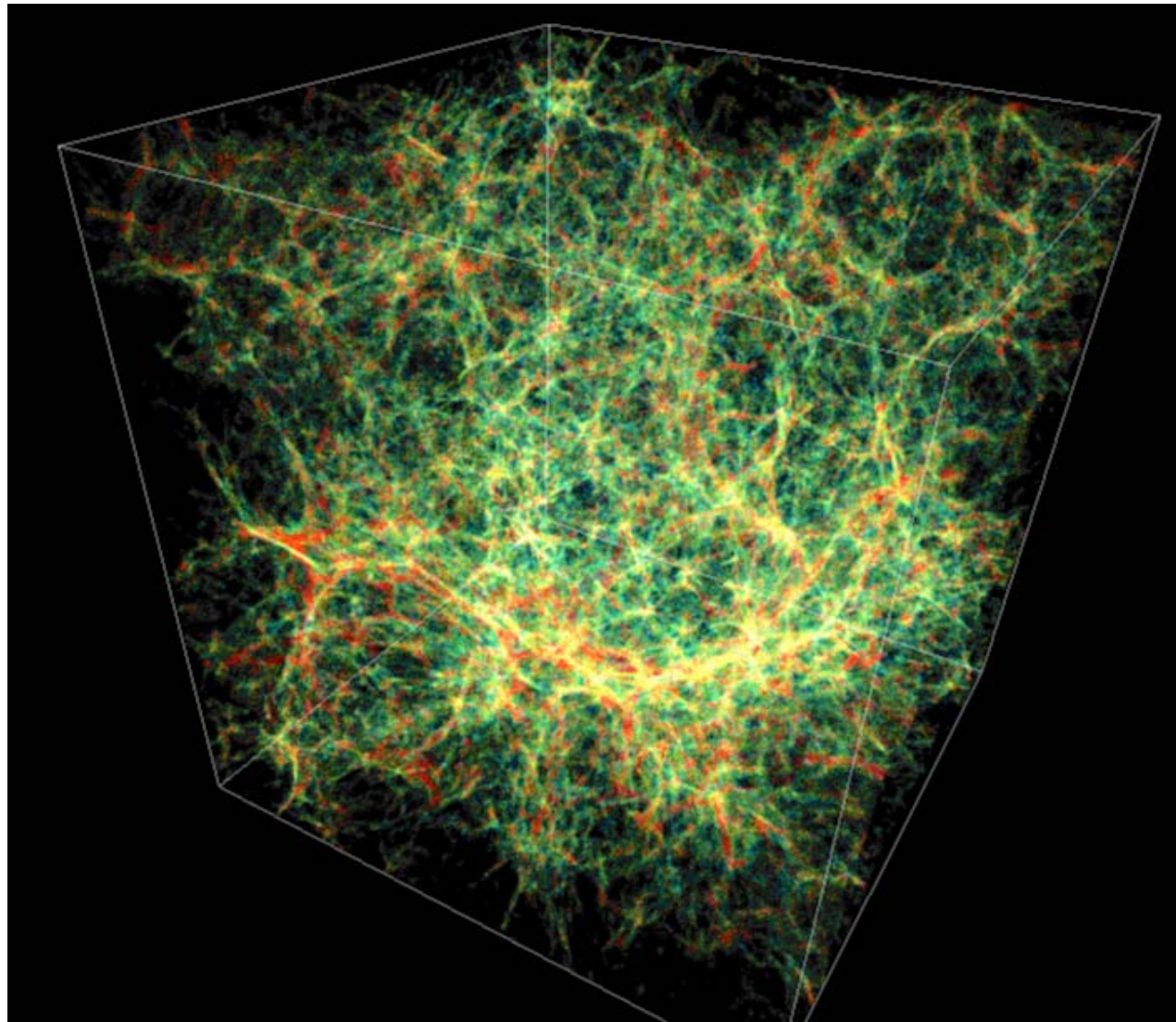


Example of Data Science (cont))

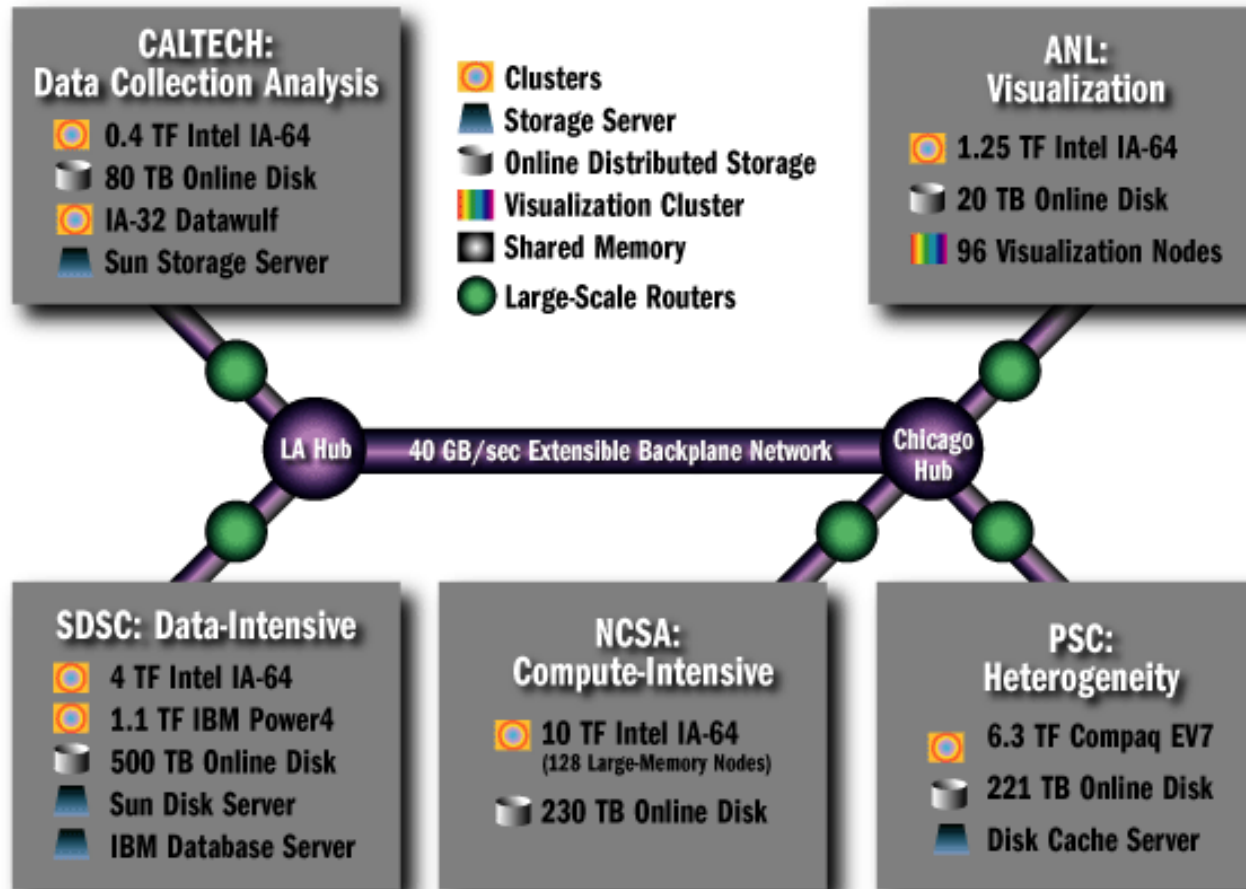
- SCEC required “*Data-enabled HEC system*” – could *not* have been done on “*Traditional HEC system*”.
- Requirements for first run
 - 47TB results generated
 - 240 processors for 5 days
 - 10-20TB data transfer from file system to archive per day
- Future Plans/Requirements
 - Increase resolution by 2 -> 1PB results generated
 - 1000 processors needed for 20-days
 - Parallel file system BW of 10GB/s needed in NEAR future (within a year) and significantly higher rates in following years
- Results have already drawn attention from geoscientists with data intensive computing needs

ENZO

- UCSD
- Cosmological hydrodynamics simulation code
- “Reconstructing the first billion years”
- Adaptive mesh refinement
- 512 cubed mesh
- Log of overdensity
- 30,000 CPU hours
- Tightly coupled jobs storing vast amounts of data (100's of TB), performing visualization remotely as well as making data available through online collections



TERAGRID



Prototype for CyberInfrastructure

Extending Data Resources to the Grid

- **Aim is to provide apparently unlimited Data Source at High Transfer rates to whole Grid**
 - SDSC is designated Data Lead for TeraGrid
 - Over 1PB of Disk Storage at SDSC in '05
- **Jobs would access data from Centralized Site mounted as local disks using WAN-SAN Global File System (GFS)**
- **Large Investment in Database Engines (72 Proc. Sun F15K, Two 32-proc. IBM 690)**
- **Rapid (1 GB/s) transfers to Tape for automatic archiving**
 - Multiple possible approaches: presently trying both Sun's QFS File System and IBM's GPFS

SDSC Data Services and Tools

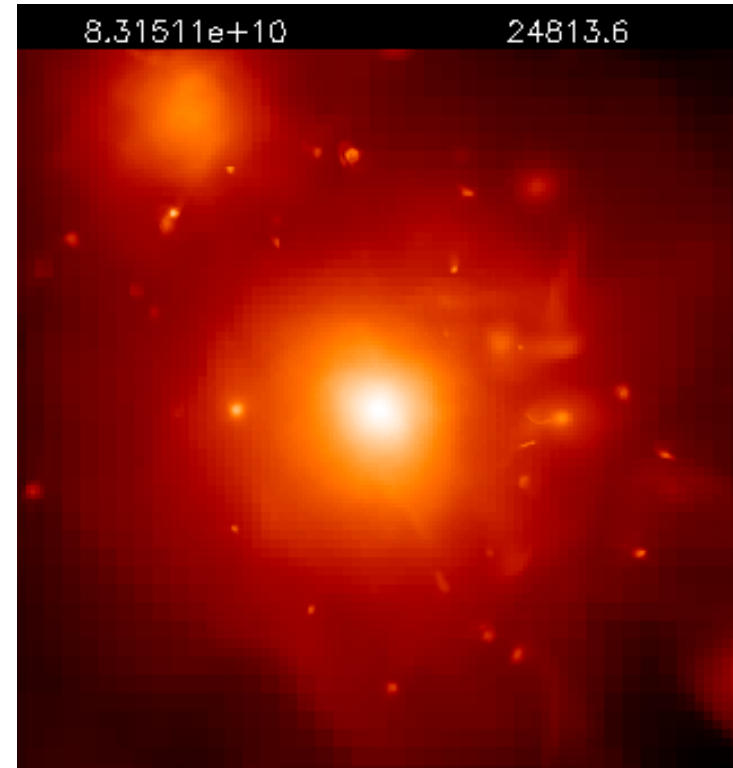
- **Data management and organization**
 - Hierarchical Data Format (HDF) 4/5
 - Storage Resource Broker (SRB)
 - Replica Location Service (RLS)
- **Data analysis and manipulation**
 - DB2, federated databases, MPI-IO, Information Integrator
- **Data and scientific workflows**
 - Kepler, SDSC Matrix, Informnet, Data Notebooks
- **Data transfer and archive**
 - GridFTP, globus-url-copy, uberftp, HIS
 - Soon Global File Systems (GFS)

Why GFS: Top TG User Issues

- **Access to remote data for read/write**
 - Grid compute resources need data
- **Increased performance of data transfers**
 - Large Datasets and Large network pipe
- **Easy of use**
 - TCP/Network tuning
- **Reliable File Transport**
 - Work Flow

ENZO Data Grid Scenario

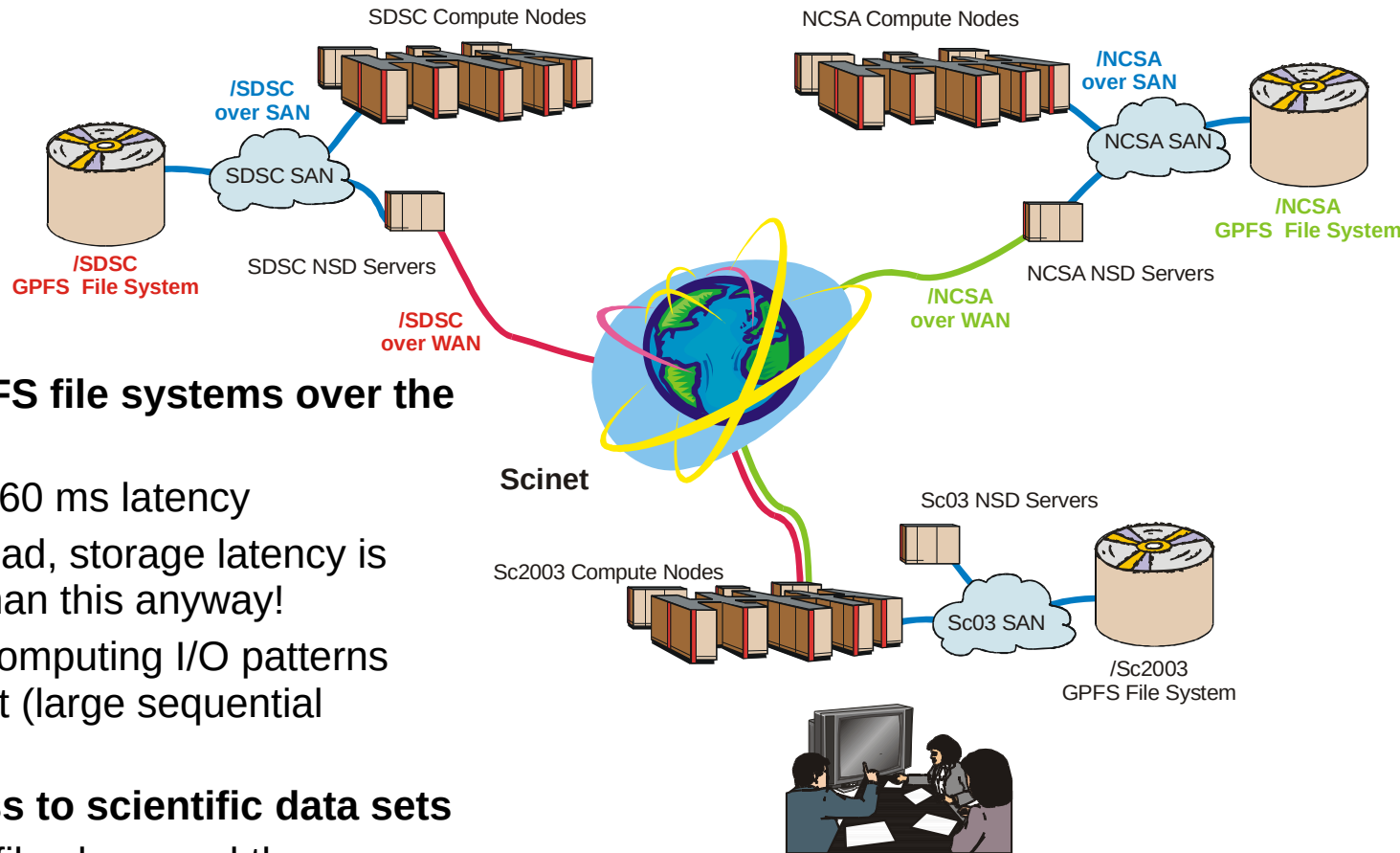
- ENZO
- Workflow is...
 - Use computational resources at PSC or NCSA or SDSC
 - Transfer 25 TB data to SDSC with SRB or GridFTP **or access directly with GFS**
 - Data is organized for high speed parallel I/O with MPIIO with GPFS
 - Data is formatted with HDF 5
 - Post processing the data at SDSC large shared memory machine
 - Perform visualization at ANL, SDSC
 - Store the images, code, raw data at SDSC SAMFS or HPSS



Projected x-ray emission
Star formation

Patrick Motl, Colorado U.

Access to GPFS File Systems over the WAN



- **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)

- **On demand access to scientific data sets**

- No copying of files here and there

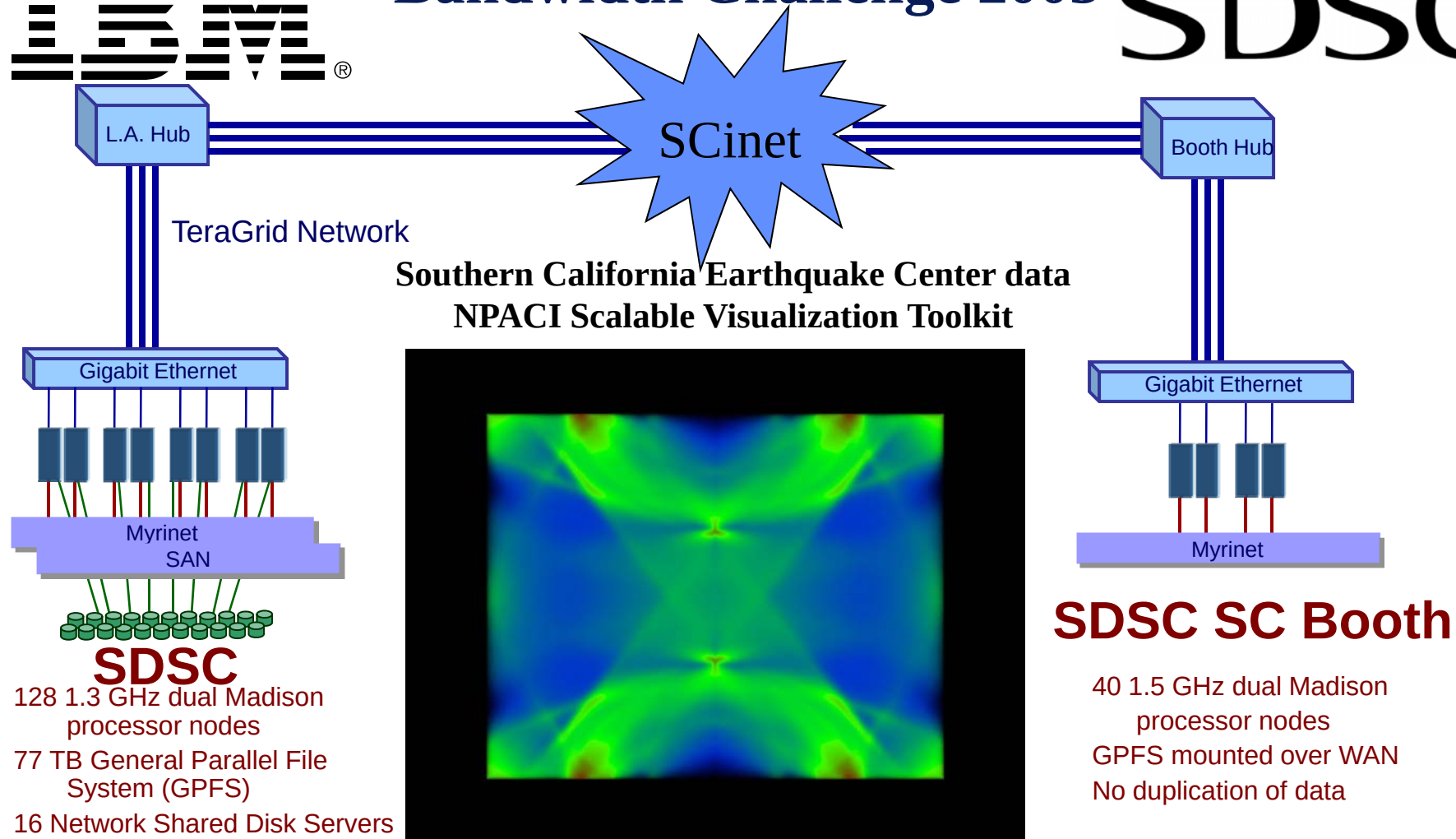
Visualization Roger Haskin, IBM

On Demand File Access over the Wide Area with GPFS

IBM®

Bandwidth Challenge 2003

SDSC



NPACI

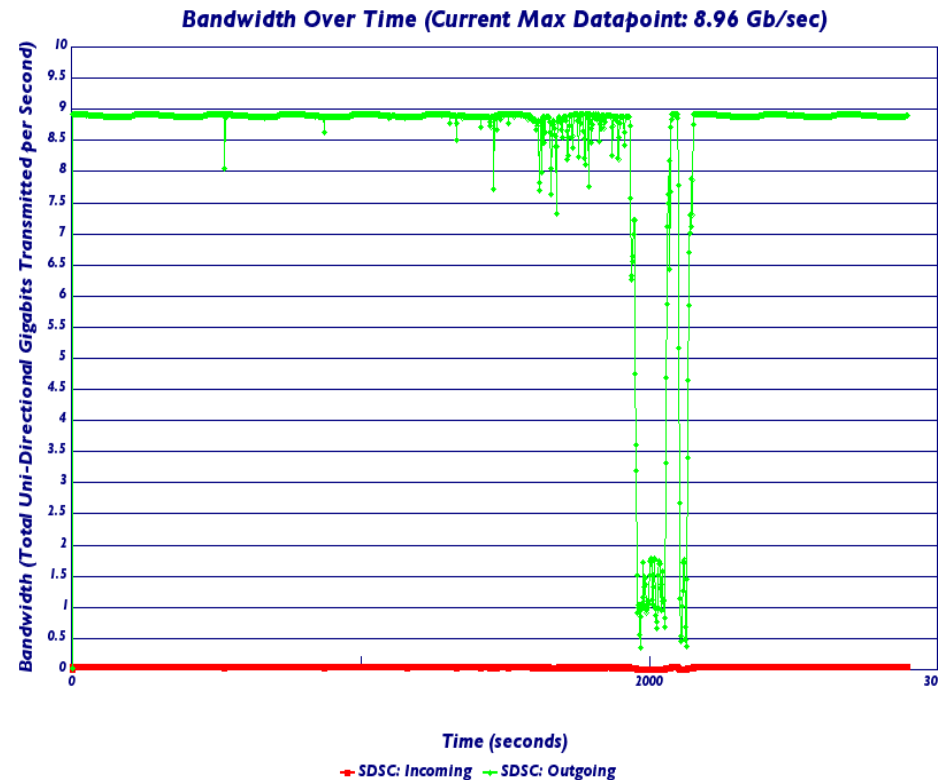
NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

SAN DIEGO SUPERCOMPUTER CENTER

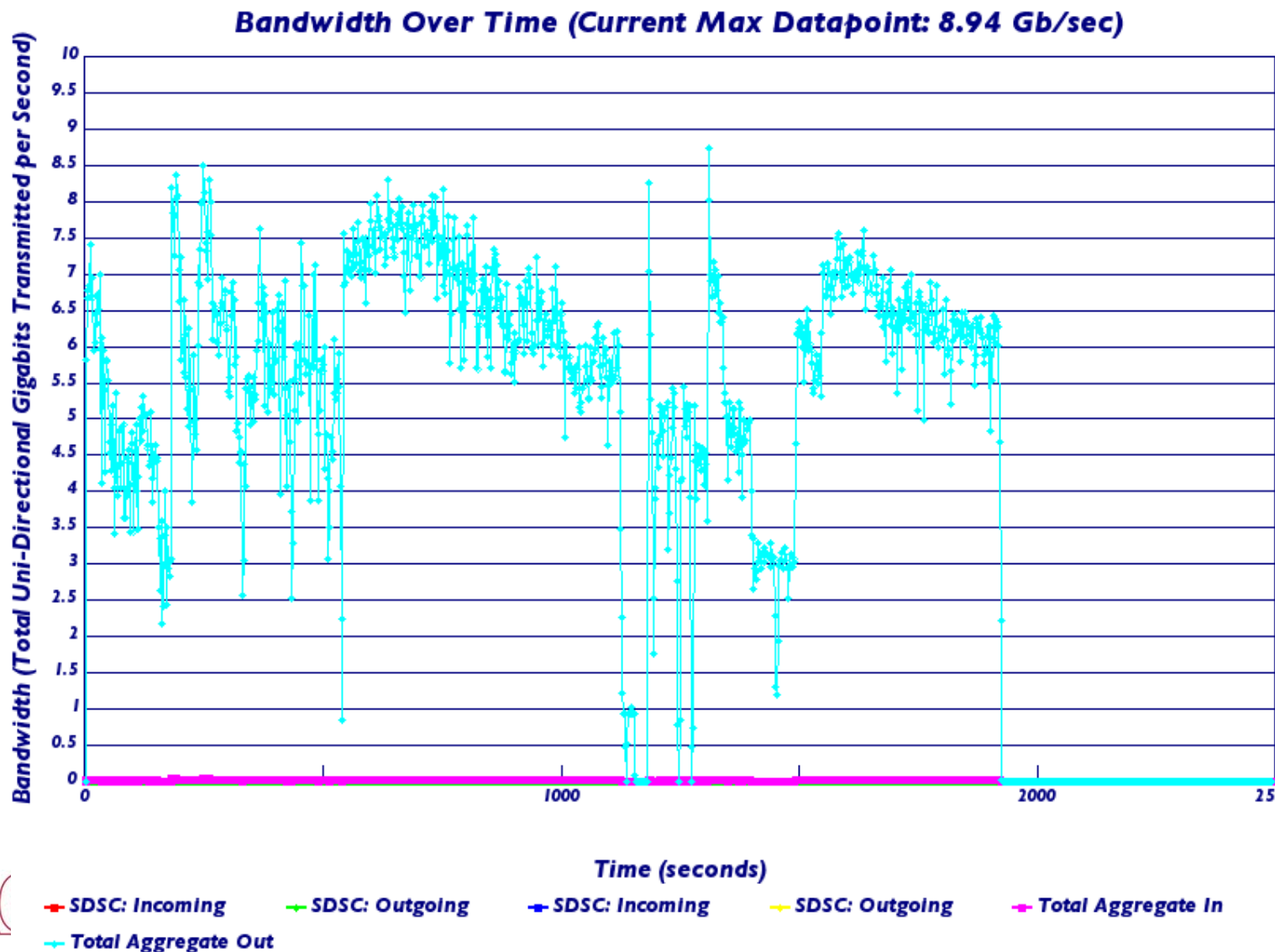


Proof of Concept: WAN-GPFS performance

- Saw excellent performance: over 1 GB/s on a 10 Gb/s link. Almost 9 Gb/s sustained: won SC'03 Bandwidth Challenge

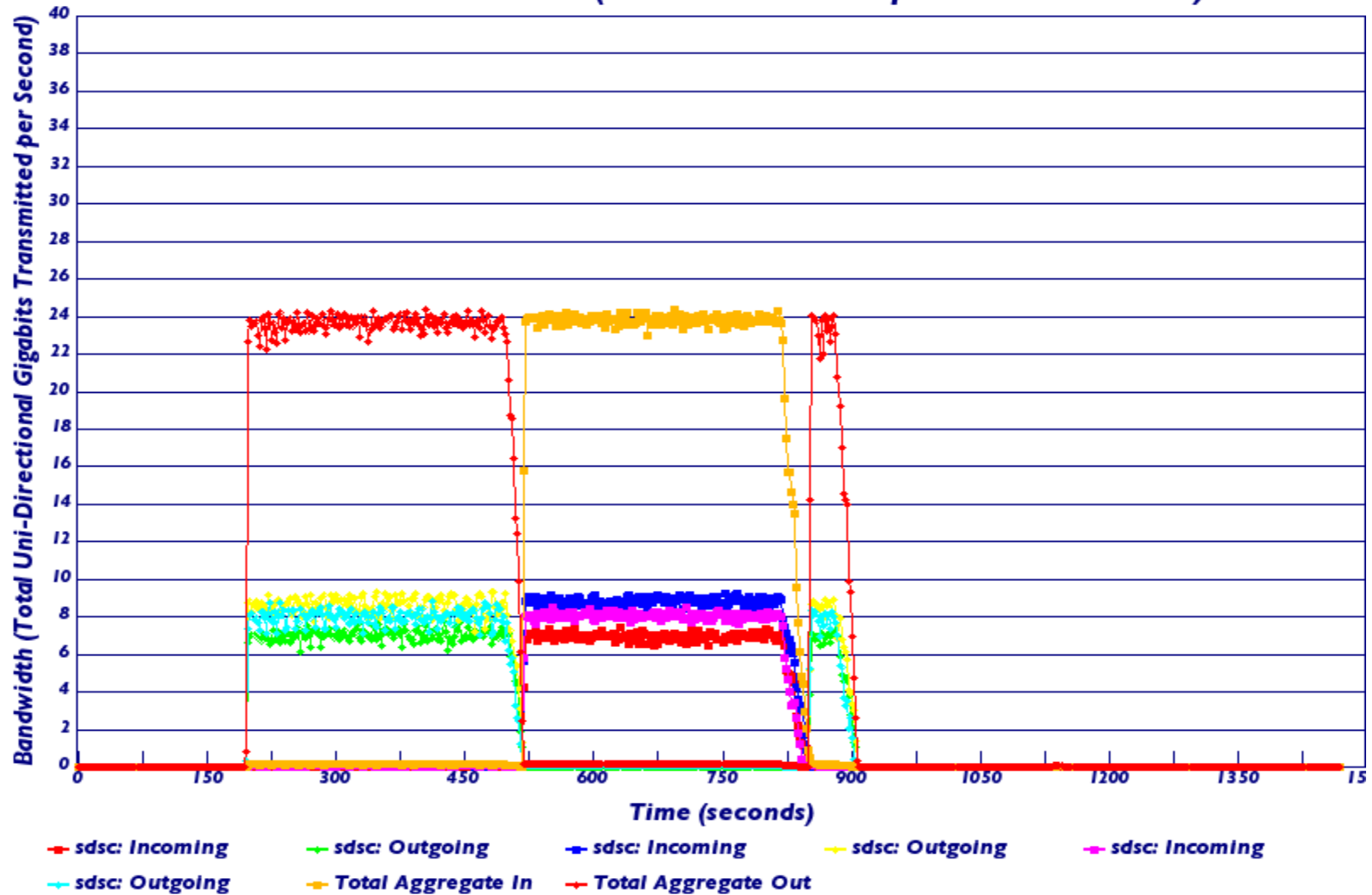


High Performance Grid-Enabled Data Movement with GridFTP



WAN-GFS POC continued at SC'04

Bandwidth Over Time (Current Max Datapoint: 27.36 Gb/sec)



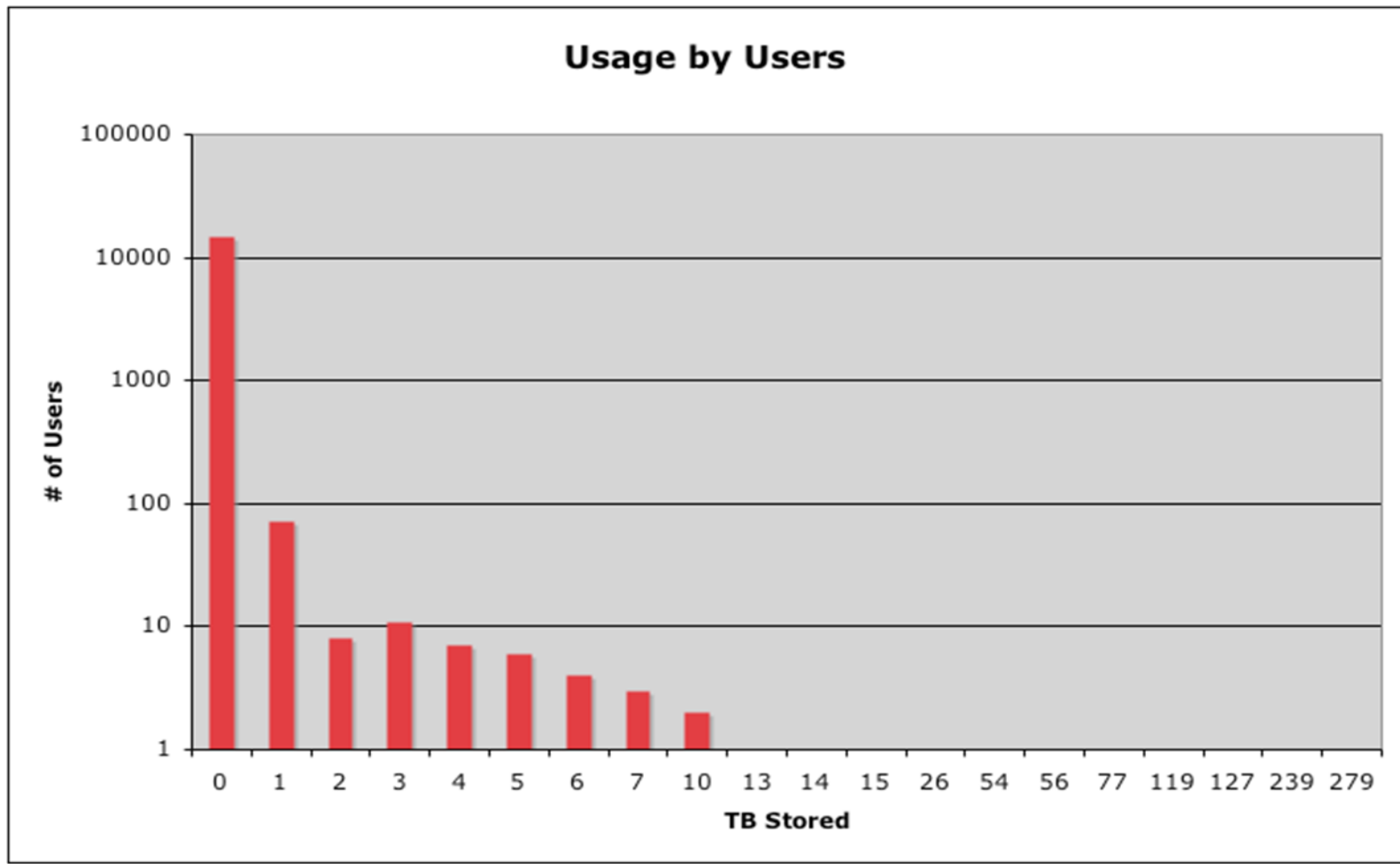
Configuration of GPFS GFS Testbed

- **20 TB of Sun storage at SDSC**
 - Moving to 60 TB of storage by May
- **42 IA-64 Servers at SDSC serving filesystem**
 - Dual QLogic FC cards w/ multipathing
 - Single SysKonnnect GigE network interface
 - Two servers with S2io 10GbE cards
 - Moving to 64 servers by May
- **20 IA-64 Nodes at SDSC**
- **16 IA-64 Nodes at NCSA**
- **2 IA-32 Nodes and 3 IA-64 Nodes at ANL**
- **4 IA-32 Nodes at PSC**
- **2 IA-32 (Dell) Nodes at TACC**

HPSS and SAM-QFS

- **High Performance HSM essential, need ~1GB/s disk to tape**
- **HPSS from IBM/DoE and SAM-QFS from Sun**
- **SAM-QFS and HPSS combined stats:**
 - 2 PB stored (1.45 PB HPSS, .45 PB SAM-QFS)
 - 50 Million files (25 Million HPSS, 25 Million SAM-QFS)

Current HPSS Usage

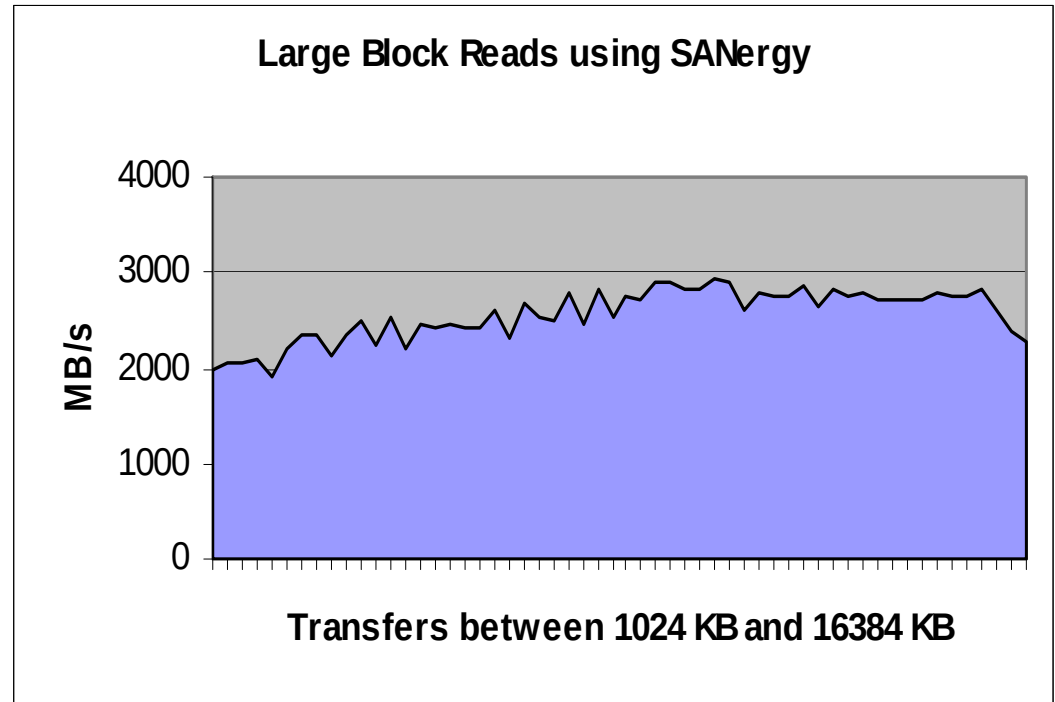


SAM-QFS at SDSC

- **What is SAM-QFS?**
 - Hierarchical Storage Management System
 - Automatically moves files from disk to tape and visa-versa
 - Filesystem interface
 - Tools to manage what files are on disk cache
 - “Pinned” datasets
 - Direct access to filesystem through SAN from clients
- **Hardware**
 - Metadata servers on two Sun Fire 15K domains
 - Over 100 TB of FC disk cache for filesystem
 - GridFTP and SRB server on Sun Fire 15K domains w/ 7 GbE interfaces

Proof of Concept: QFS and SANergy

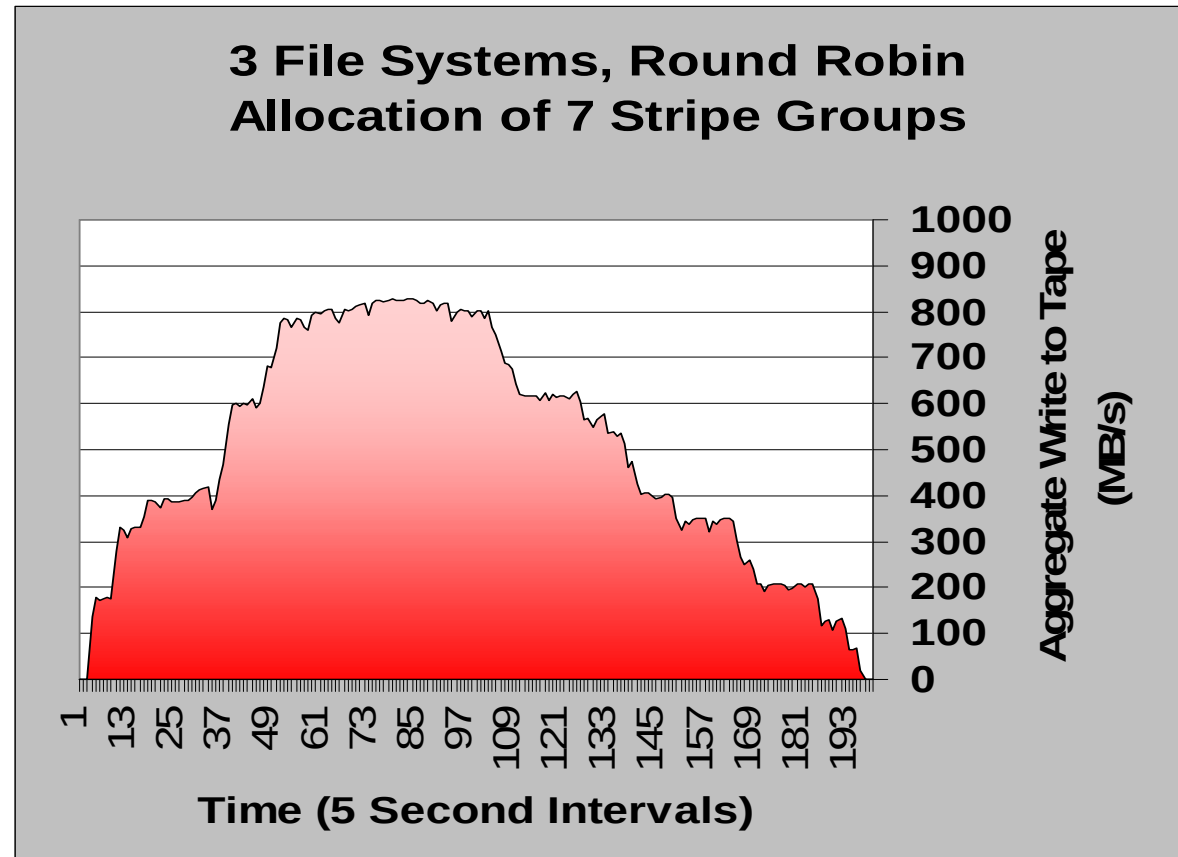
- **Sun QFS File System**
- **32 Sun T3B storage Arrays**
- **Sun E450 Metadata Server**
- **Sun SF 6800 SANergy Client**
 - 16 Qlogic 2 Gb FC adapters
- **Four Brocade 3800 2 Gb FC switches**



Proof of Concept: SAM-QFS

Archive Testing

- **Sun Fire 15K**
- **18 Qlogic 2310 Fibre Channel HBAs**
- **Brocade 12000**
- **25 STK 9940B FC Tape Drives**
- **14 Sun T3B FC Arrays**
- **All Tape/Disk drives active**
- **828.6 MB/s peak**

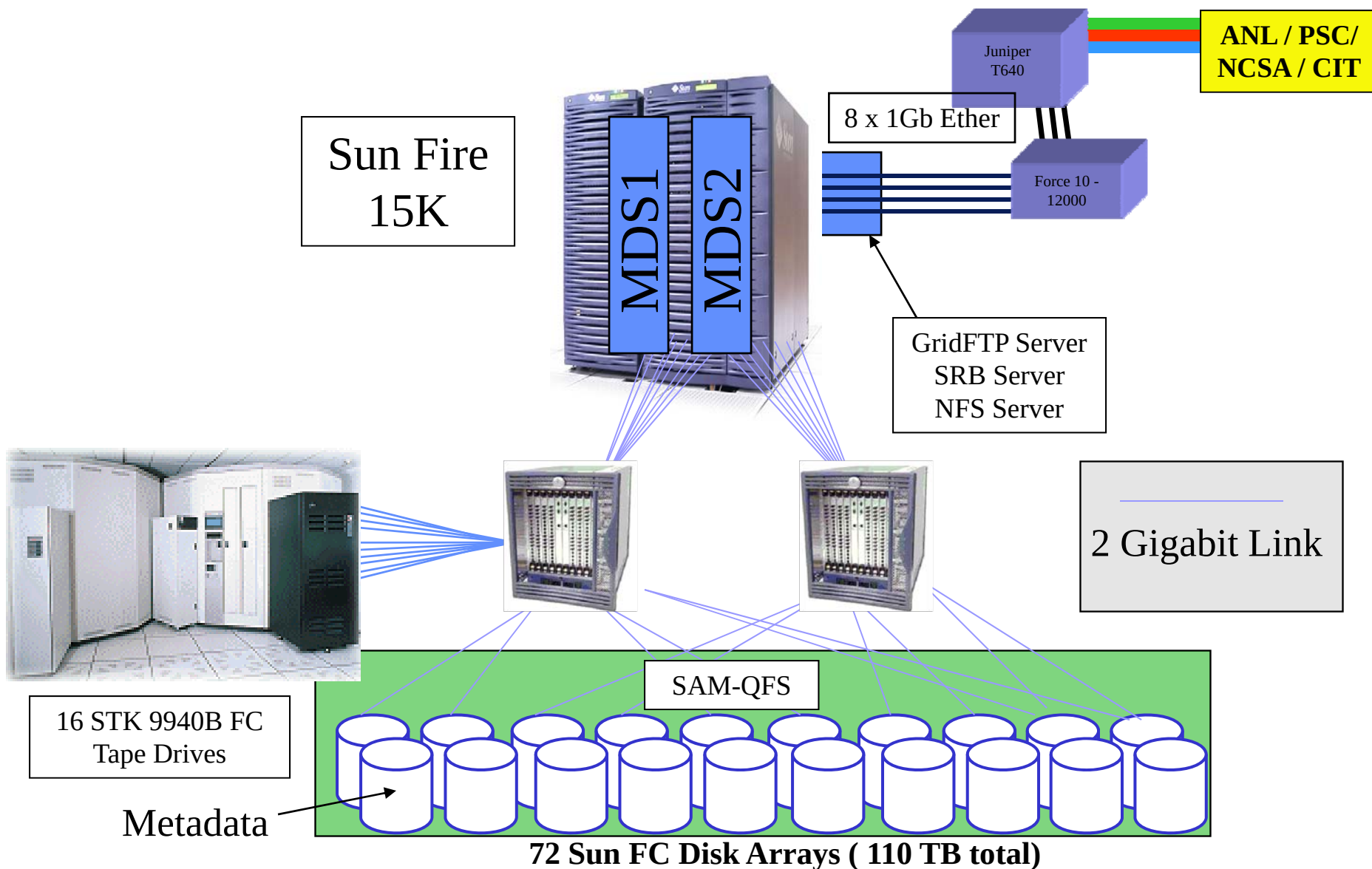


TeraGrid SAM-FS Server

- Next generation Sun server
- 72 (900 MHz) processors
- 288 GB shared memory
- 38 Fiber channel SAN interfaces
- Sixteen 1Gb Ethernet interfaces
- FC-attached to 540 TB SAN
- FC-attached to 16 STK9940B Tape Drives
- Primary applications
 - Data management applications
 - SDSC SRB (Storage Resource Broker)



SAM-QFS Architecture on Sun Fire 15K

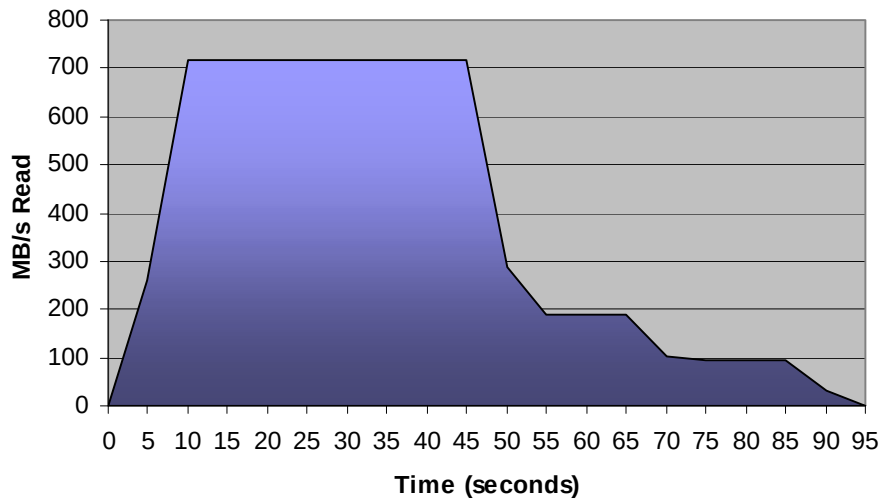


Proof of Concept: WAN-SAN Demo

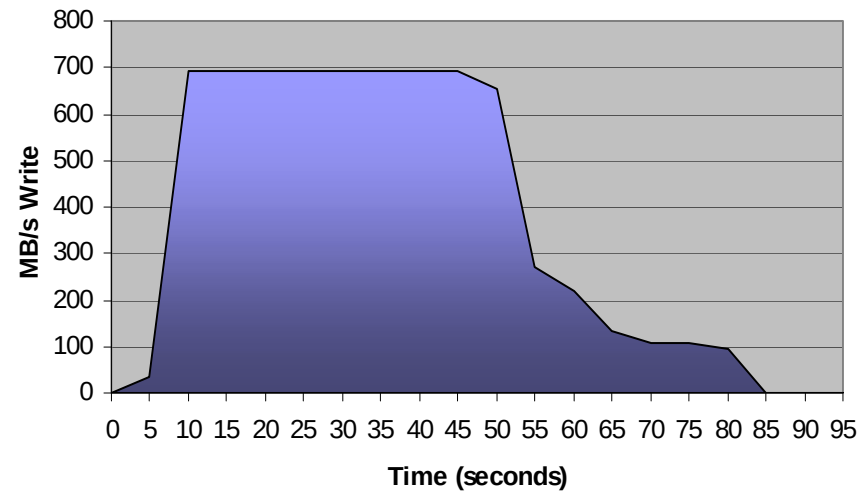
- **Extended the SAN from the SDSC Machine Room to the SDSC booth at SC'02**
- **Reads from Data Cache at San Diego to Baltimore exceeded 700 MB/s on 8 X 1 Gb/s links**
- **Writes slightly slower**
- **Single 1 Gb/s link gave 95 MB/s**
- **Latency approx. 80 ms round trip**

Performance Details: WAN-SAN

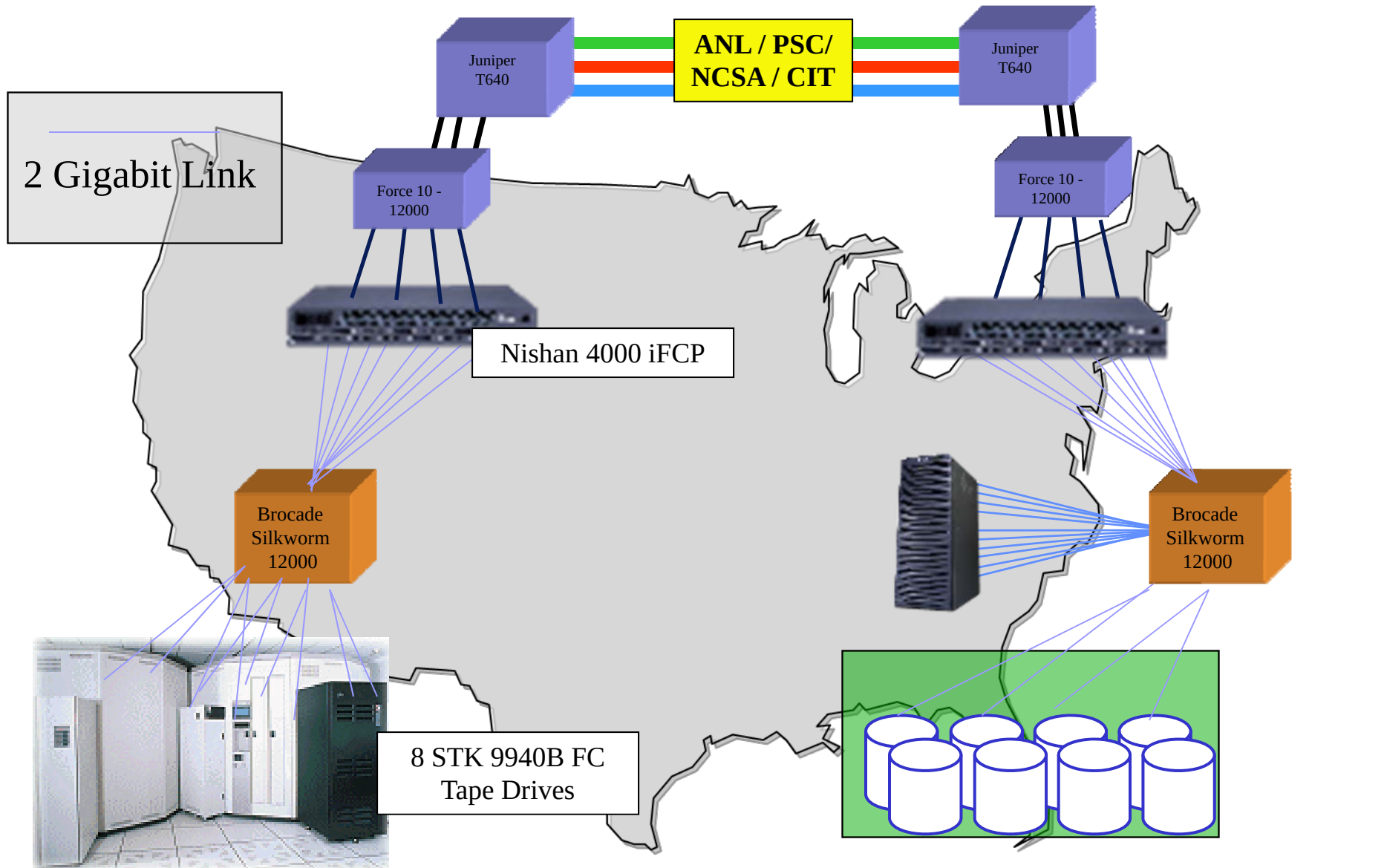
Read Performance: SC02 from SDSC



Write Performance: SC02 to SDSC

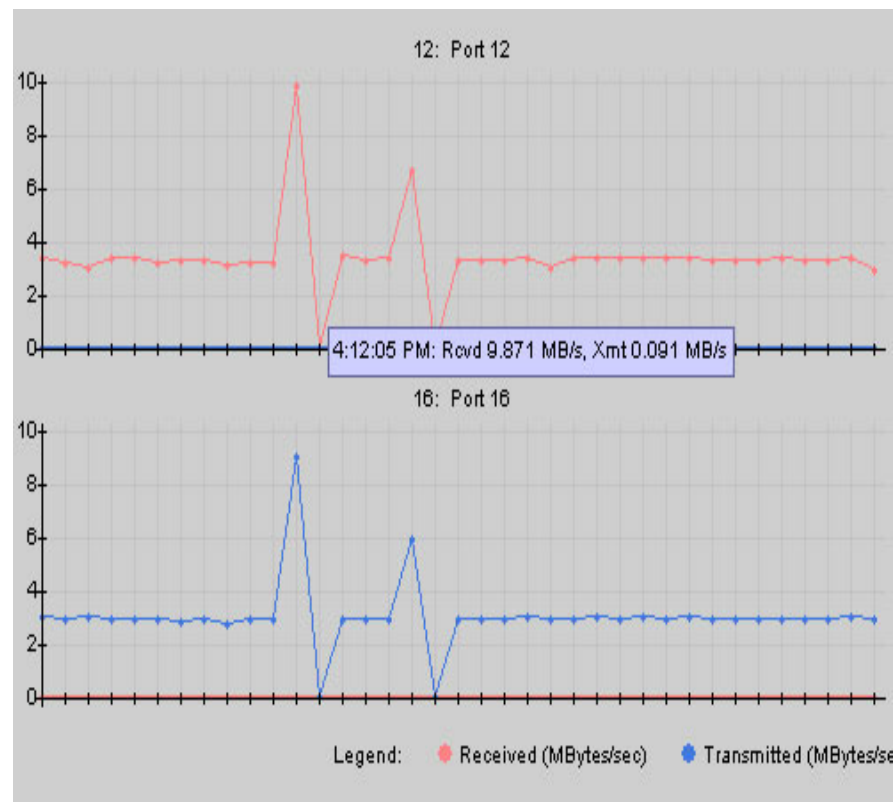


Remote Fibre Channel Tape Architecture

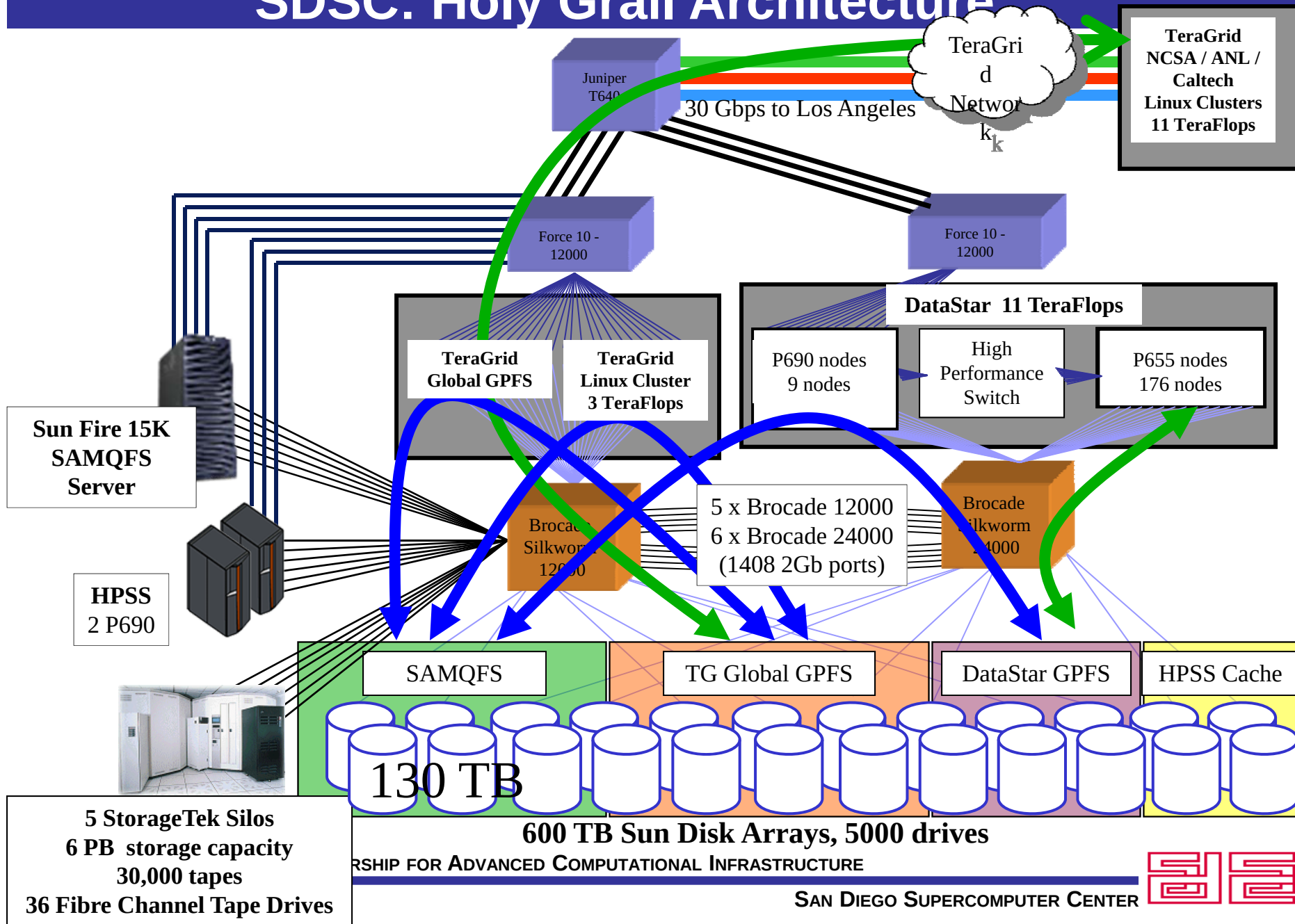


Also Interfaced PSC DMF HSM with SDSC tape archival system

- Used FC/IP encoding to attach SDSC tape drives to PSC's Data Management Facility
- Automatic archival across the Grid!



SDSC: Holy Grail Architecture



Lessons Learned, Moving Forward

- **Latency is unavoidable, but not insurmountable**
- **Network performance can approach local disk**
- **FC/IP can utilize much of raw bandwidth**
- **WAN-Oriented File Systems equally efficient**
- **FC/SONET requires fewer protocol layers, but FC/IP easier to integrate into existing networks**
- **Good planning and balance required**
- **File Systems are Key!**

Moving Forward: SAM-QFS

- **SATA disk hierarchy**
 - FC QFS filesystem
 - SATA Disk Volumes
 - FC Tape
- **Native Linux Client**
 - Direct access by SDSC IA-64 Cluster
 - Massive transfer rates between GPFS and SAM-QFS
- **Primary Issues**
 - Data residency
 - Archival rates
 - Mixed large & small files
 - Large number of files