Scaling a Global File System to the Greatest Possible Extent, Performance, Capacity, and Number of Users

Phil Andrews, Bryan Banister, Patricia Kovatch, Chris Jordan
San Diego Supercomputer Center
University of California, San Diego
andrews@sdsc.edu, bryan@sdsc.edu, pkovatch@sdsc.edu, ctjordan@sdsc.edu
Roger Haskin
IBM Almaden Research Center
roger@almaden.ibm.com

Abstract

- *At SC'04 we investigated scaling file storage to the very widest possible extents using IBM's GPFS file system, with authentication extensions developed by the San Diego Supercomputer Center and IBM. Close collaboration with IBM, NSF, and TeraGrid*

- *The file system extended across the United States, including Pittsburgh, Illinois, and San Diego, California, with the TeraGrid 40 Gb/s backbone providing the Wide Area Network connectivity. On a 30Gb/s connection, we sustained 24 Gb/s and peaked at 27 Gb/s*

- *Locally, we demonstrated 15 GB/s with new WR sort times*

- *Implementation of the production Global File System facility is underway: over 500 TB will be served up from SDSC within 2 weeks.*

- *Usage to come; plan to go to 1 PB by end of year*

Background: National Science Foundation TeraGrid

- **Prototype for CyberInfrastructure**
- **High Performance Network: 40 Gb/s backbone, 30 Gb/s to each site**
- **National Reach: SDSC, NCSA, CIT, ANL, PSC, TACC, Purdue/Indiana, ORNL**
- **Over 40 Teraflops compute power**
- **Approx. 2 PB rotating Storage**

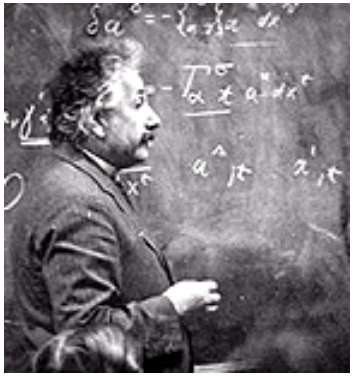
TeraGrid Network Geographically

What Users Want in Grid Data

- **Unlimited data capacity. We can almost do this.**
- **Transparent, High Speed access anywhere on the Grid. We can do this.**
- **Automatic Archiving and Retrieval (maybe)**
- **No Latency. We can't do this.**

(Measured 60 ms roundtrip

SDSC-NCSA, 80 ms SDSC-Baltimore)



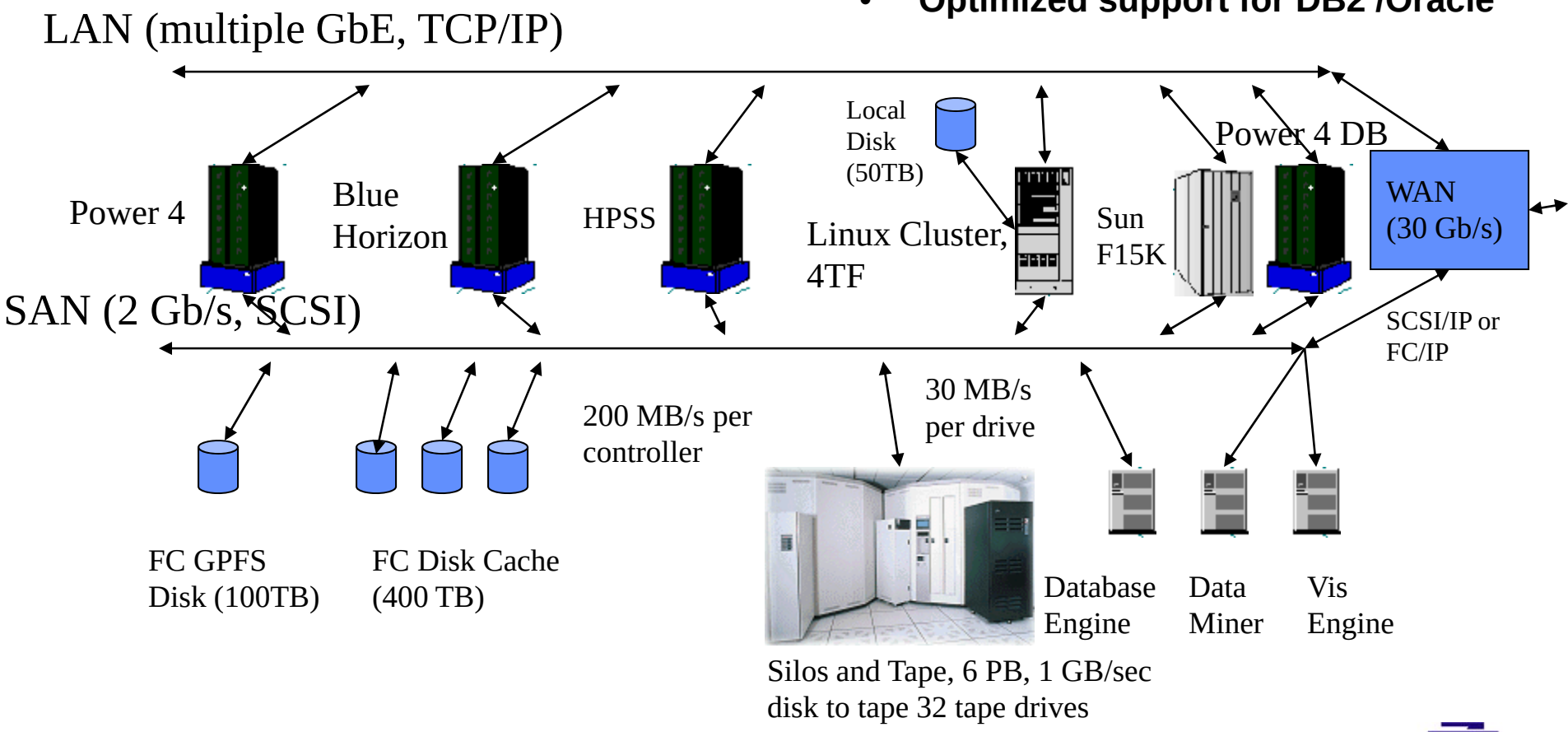
Proposing Centralized Grid Data Approach

- **SDSC is designated Data Lead for TeraGrid**
- **Over 1 PB of Disk Storage**
- **Large Investment in Database Engines (72 Proc. Sun F15K, Two 32-proc. IBM 690)**
- **32 STK 9940B Tape Drives, 28 other drives, 5 STK silos, 6 PB Capacity**
- **Storage Area Network: Over 1400 2Gb/s ports**
- **Close collaboration with vendors: IBM, Sun, Brocade, etc.**

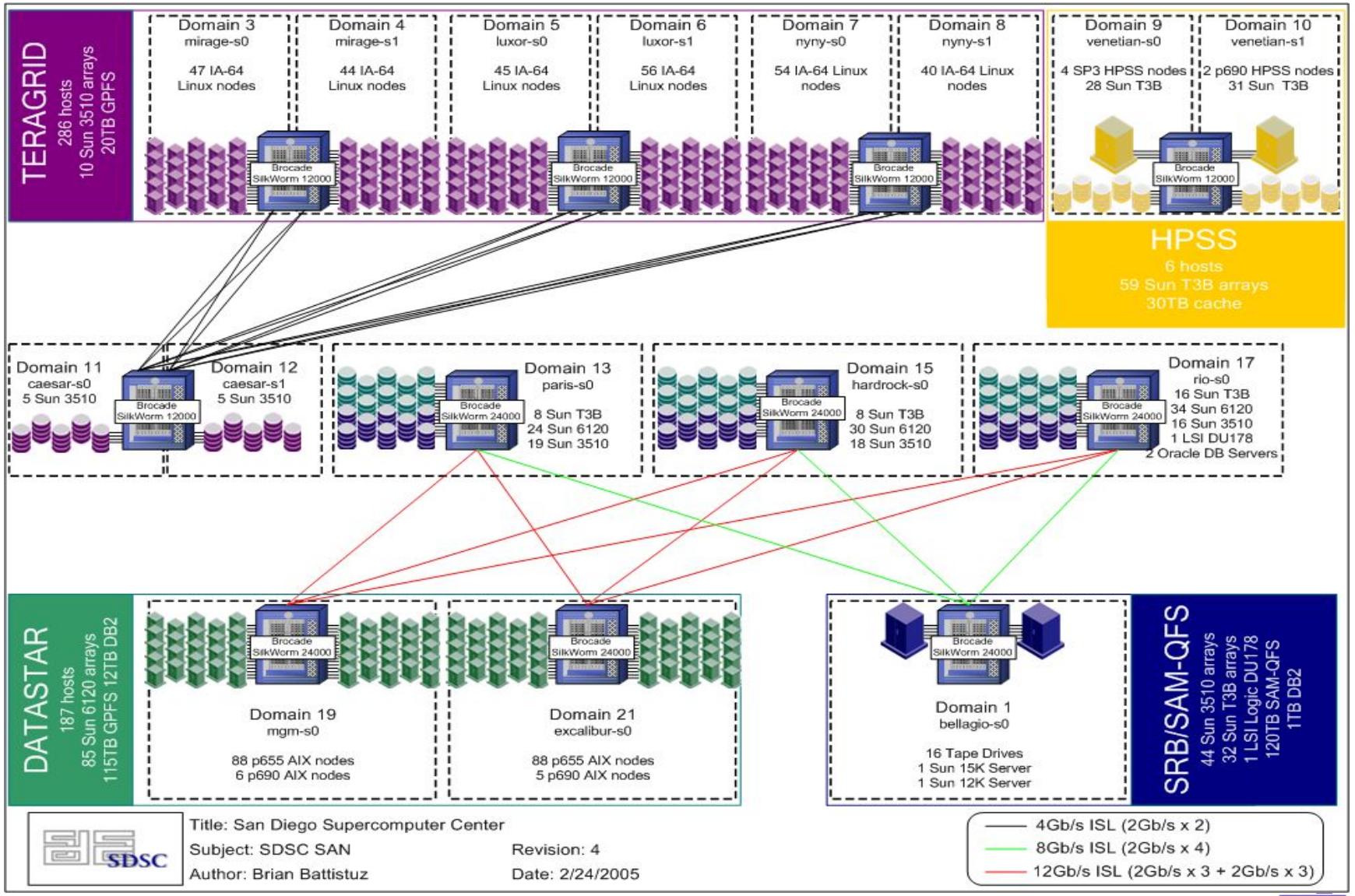
SDSC Machine Room Data Architecture

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

- **1 PB disk**
- **6 PB archive**
- **1 GB/s disk-to-tape**
- **Optimized support for DB2 /Oracle**



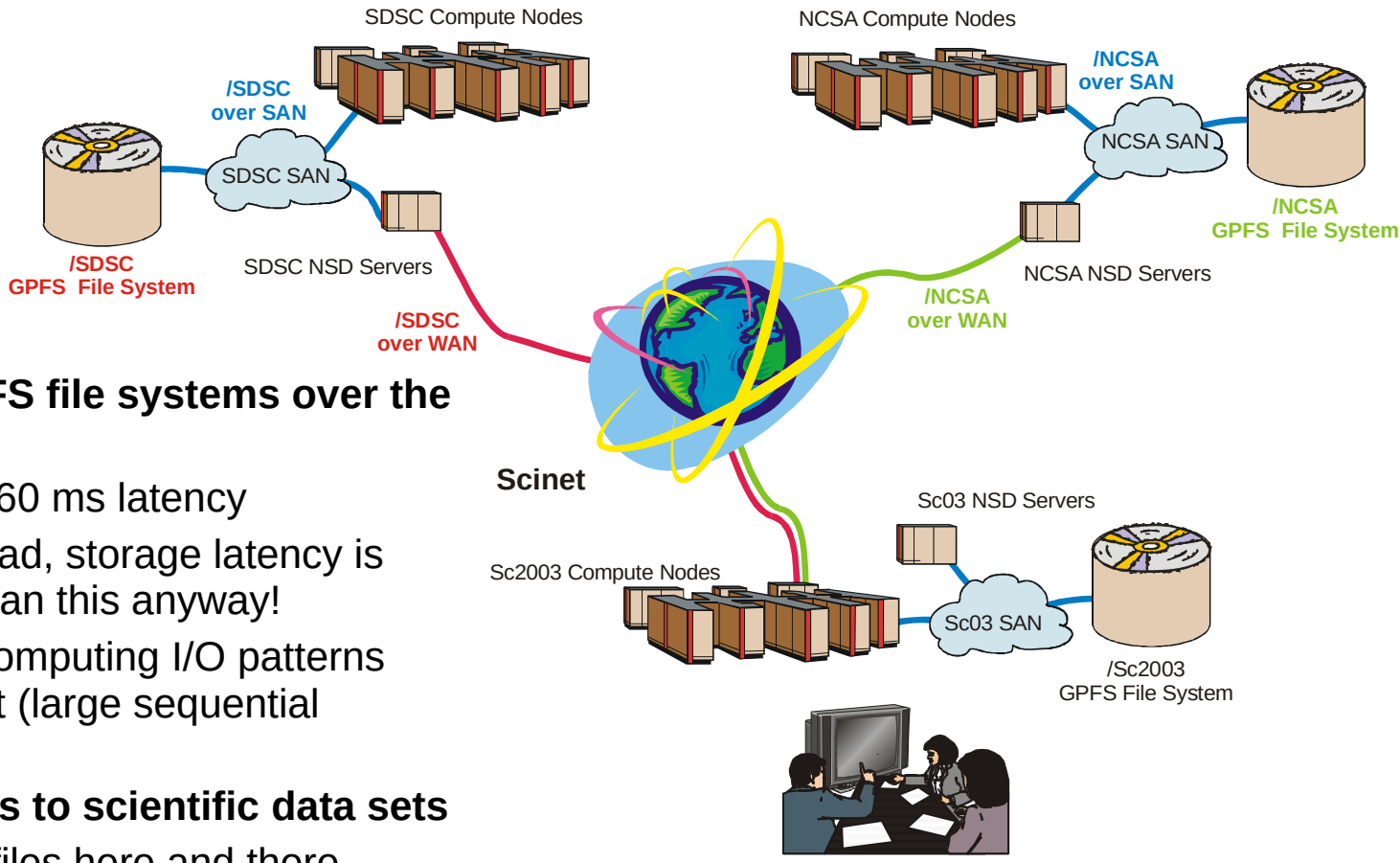
Storage Area Network at SDSC



What do we use the TeraGrid Network for?

- **We expect data movement to be major driver**
- **Conventional approach is to submit job to one or multiple sites, GridFTP the data there and back**
- **SDSC/IBM & TeraGrid are also exploring Global File System approach**
- **First demonstrated with a single 10 Gb/s link and local disk at SC'03 (Phoenix)**
- **Exported at high speed to SDSC and lower speed to NCSA**

Access to GPFS File Systems over the Wide Area Network (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)
- **On demand access to scientific data sets**
  - No copying of files here and there

Roger Haskin, IBM

On Demand File Access over the Wide Area with GPFS

Bandwidth Challenge 2003

IBM®

SDSC

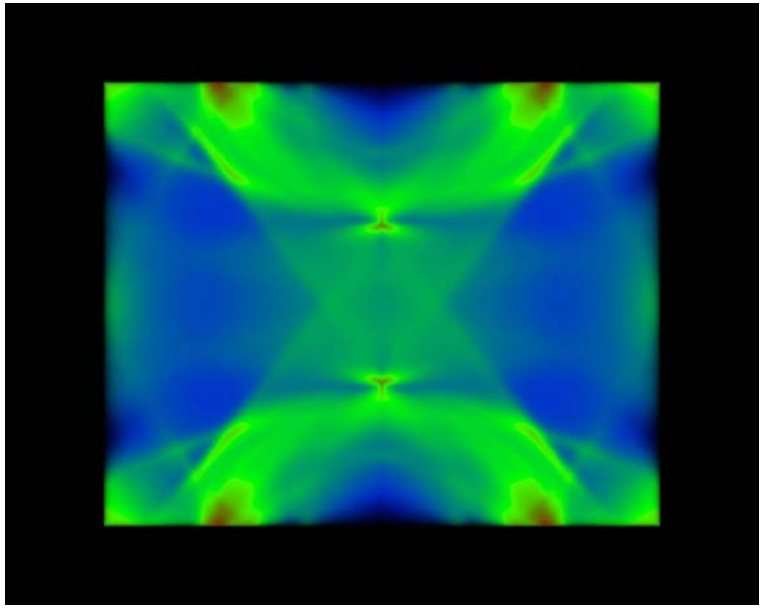
L.A. Hub

SCinet

Booth Hub

TeraGrid Network

Southern California Earthquake Center data
NPACI Scalable Visualization Toolkit



Gigabit Ethernet

Myrinet SAN

SDSC

128 1.3 GHz dual Madison processor nodes

77 TB General Parallel File System (GPFS)

16 Network Shared Disk Servers

Gigabit Ethernet

Myrinet

SDSC SC Booth

40 1.5 GHz dual Madison processor nodes

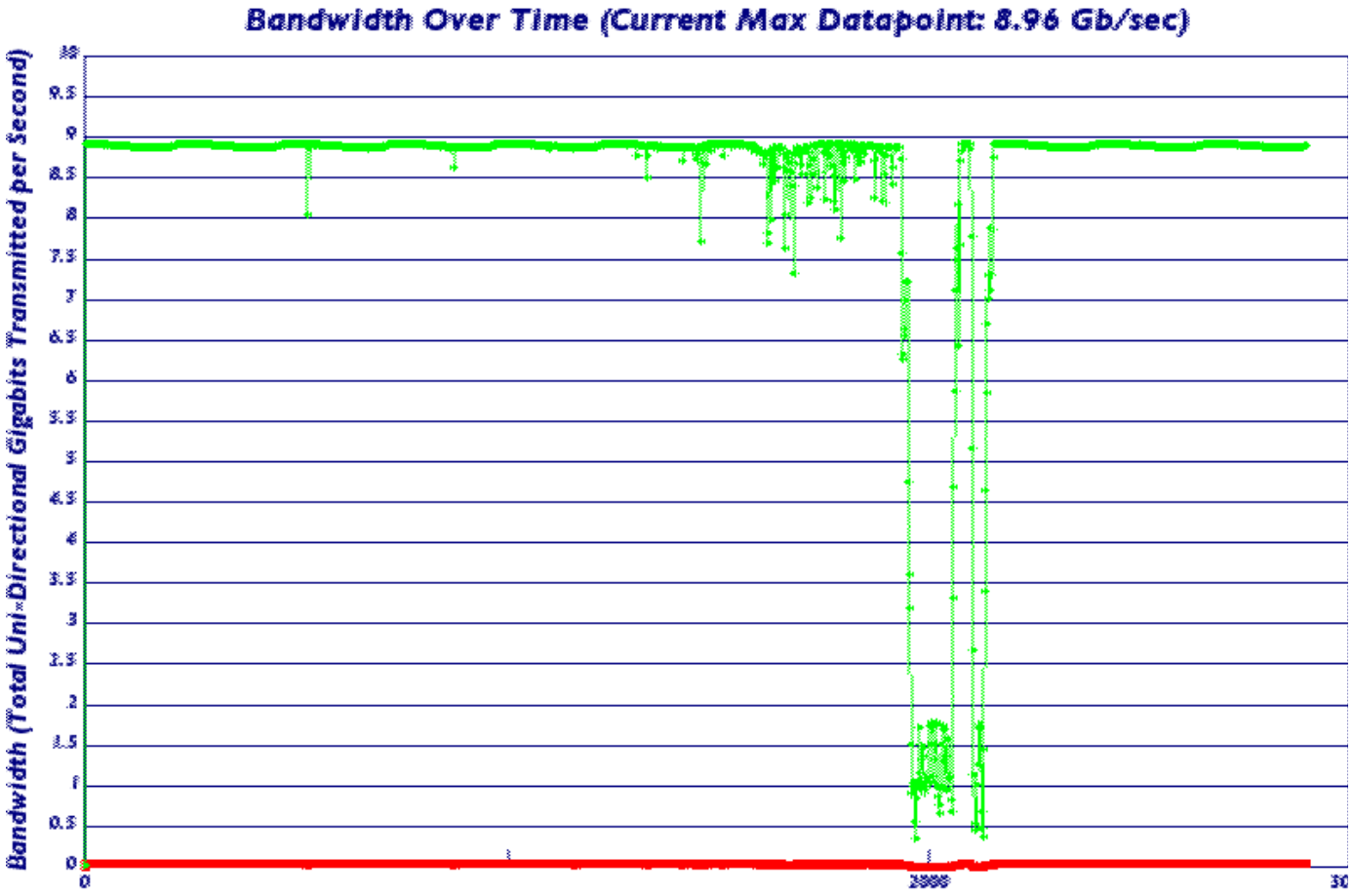
GPFS mounted over WAN

No duplication of data

First Demonstration of High Performance Global File System over standard TCP/IP

- **Visualized application output data across 10 Gb/s link to SDSC from show floor at Phoenix SC'03**
- **Saw approximately 9 Gb/s out of 10 Gb/s link**
- **Very encouraging, paved way for full prototype at SC'04**
- **If successful, planned to propose production GFS facility for TeraGrid**

Global TG GPFS over 10 Gb/sWAN (SC'03 Bandwidth Challenge Winner)



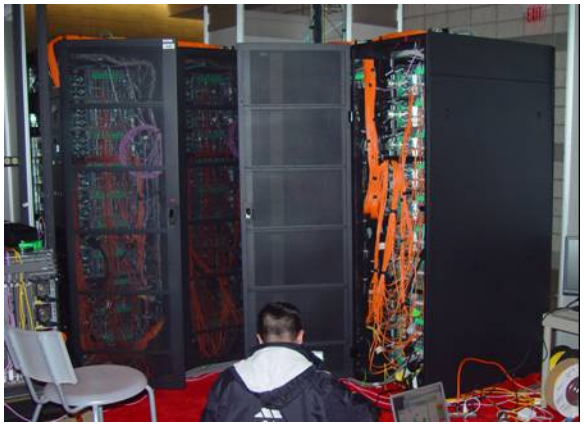
SC'04 SDSC StorCloud Challenge Entry

- **160 TB of IBM disk in StorCloud booth**
- **Connected to 40 Servers in SDSC booth via 120 FC links**
- **Saw 15 GB/s from GPFS on show floor**
- **Exporting GPFS file system across TeraGrid to NCSA and SDSC using SCinet 30 Gb/s link to TeraGrid backbone**
- **Saw 3 GB/s from show floor to NCSA & SDSC**
- **These are all direct file system accesses**

StorCloud Demo

StorCloud 2004

- **Major initiative at SC2004 to highlight the use of storage area networking in high-performance computing**
- **~1 PB of storage from major vendors for use by SC04 exhibitors**
- **StorCloud Challenge competition to award entrants that best demonstrate the use of storage (similar to Bandwidth Challenge)**



40-node GPFS server cluster

Installing the DS4300 Storage

SDSC-IBM StorCloud Challenge

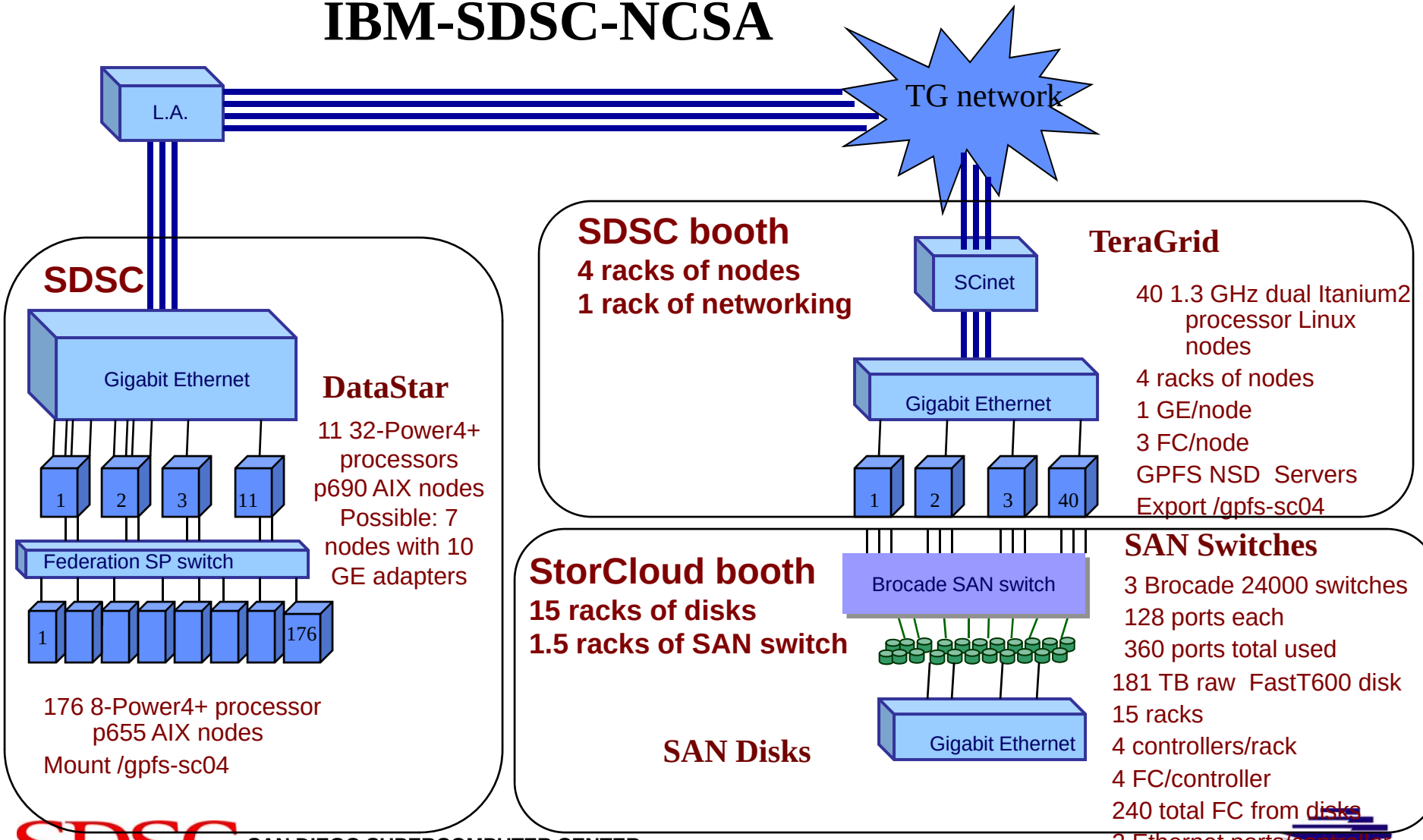
A workflow demo that highlights multiple computation sites on a grid sharing storage at a storage siteIBM computing and storage hardware and the 30 Gb/s communications backbone of the Teragrid

IBM DS4300 Storage in StorCloud booth

SC'04 demo purposely designed to prototype possible future production

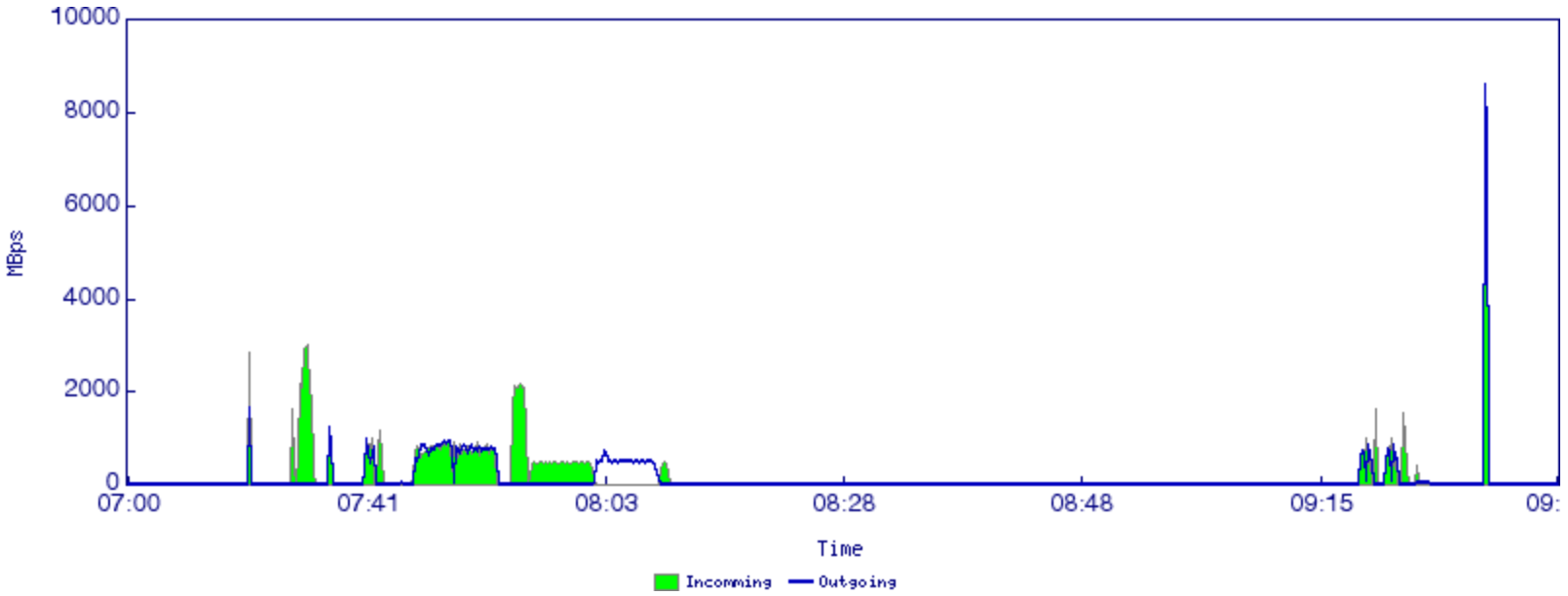
- **30 Gb/s connection to TeraGrid from Pittsburgh show floor identical to production setup**
- **GPFS mounted on IBM StorCloud disks on show floor to both SDSC and NCSA**
- **Local sorts demonstrated File System capability**
- **Production Application (Enzo) ran at SDSC and wrote to Pittsburgh disks (~ 1TB/hour)**
- **Output data visualized at SDSC and NCSA (~3GB/s), was real production data**

SC '04 Demo
IBM-SDSC-NCSA



GPFS Local Performance Tests

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

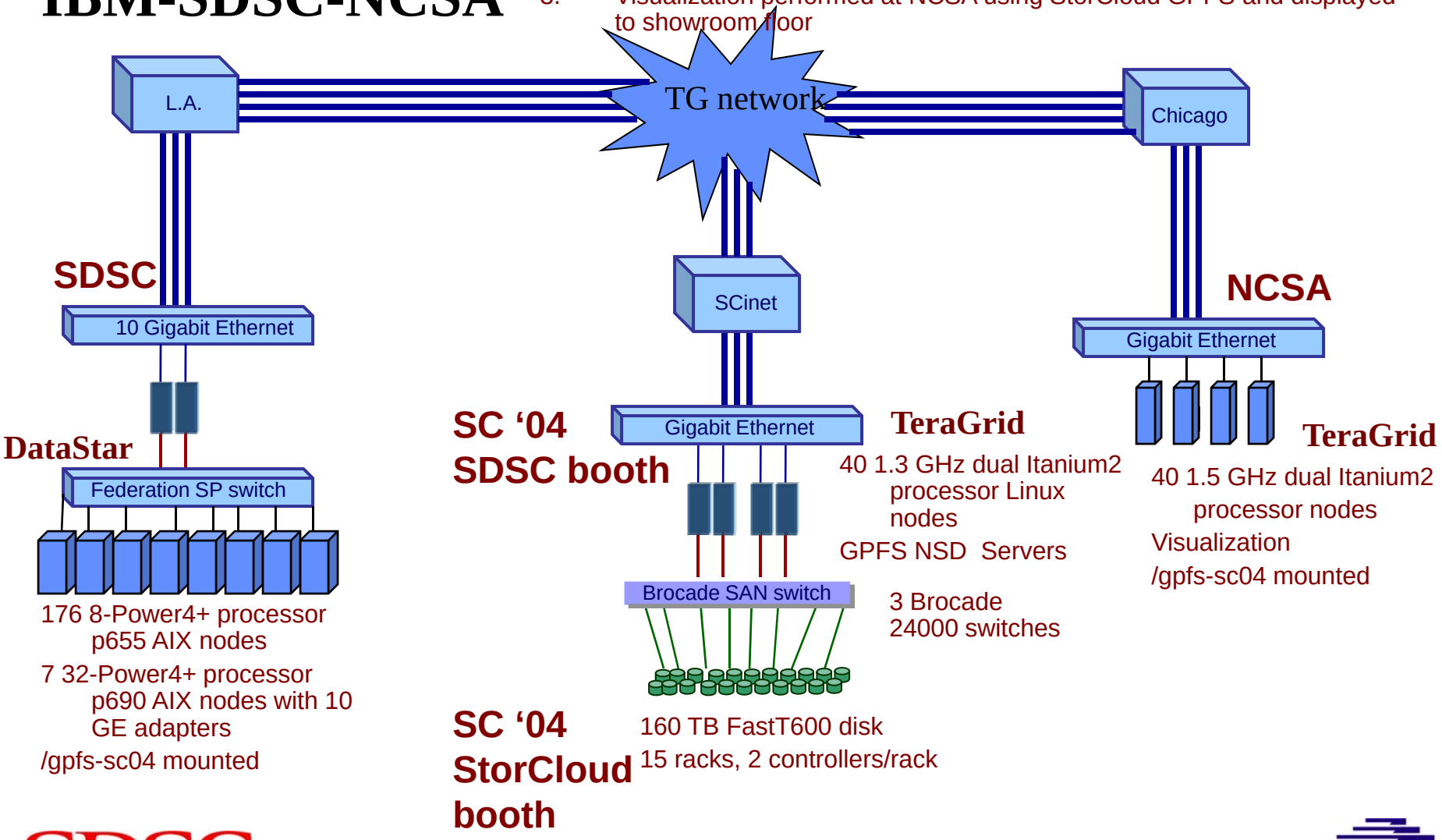


Global File Systems over WAN

- **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. Initial GSI authentication in use (joint SDSC-IBM work)
  - On demand Data
  - Instantly accessible and searchable
  - No need for local storage space
  - Need network bandwidth

SC ‘04 Demo IBM-SDSC-NCSA

1. Nodes scheduled using GUR
2. ENZO computation on DataStar, output written to StorCloud GPFS served by nodes in SDSC’s SC ‘04 booth
3. Visualization performed at NCSA using StorCloud GPFS and displayed to showroom floor



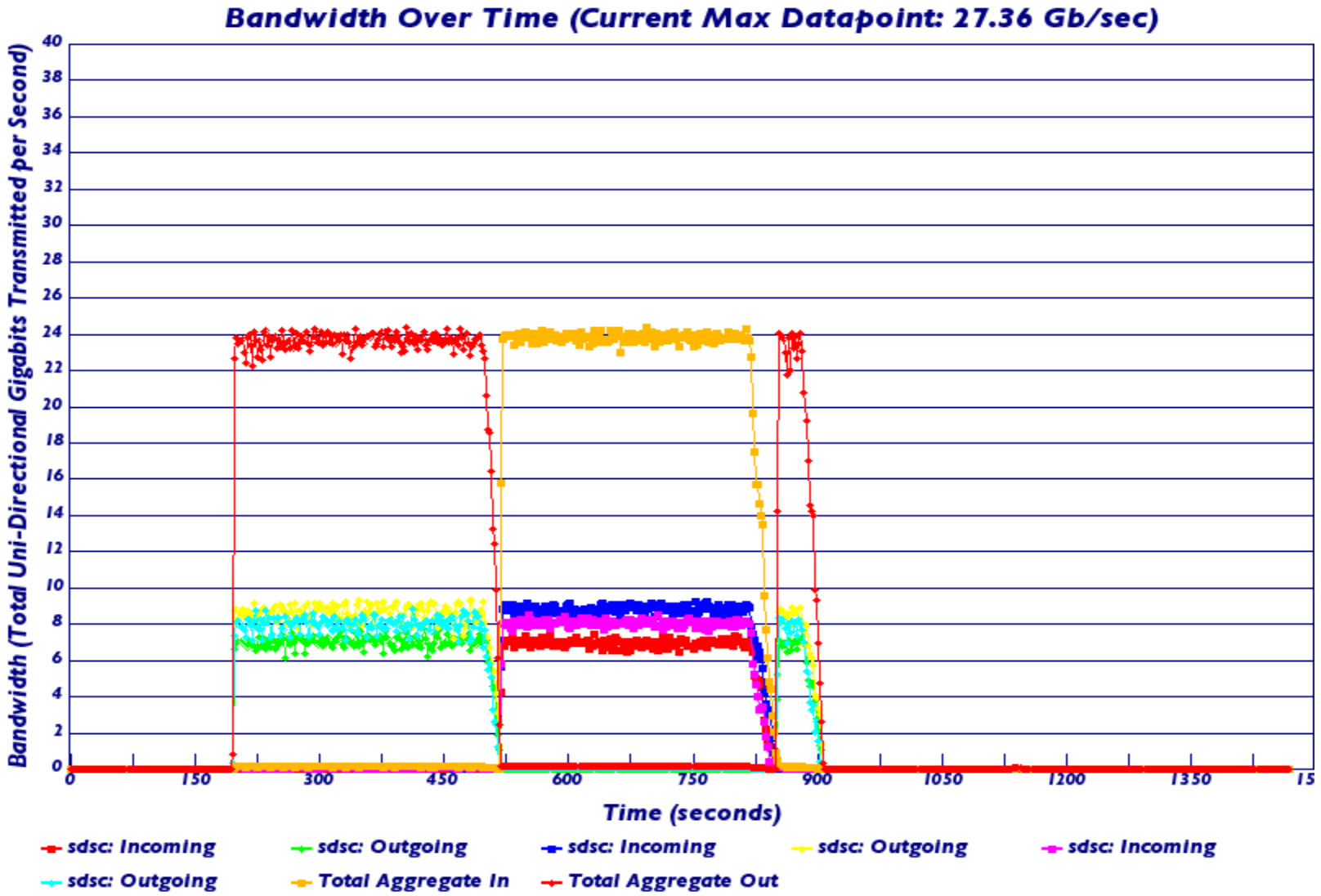
GSI Authentication Essential

- Conventional authentication is via UID
- Several resource providers combine on a grid
- Only very tightly coupled grids have a unified UID space
- On general grids, users will have multiple UIDs across the Grid
- Need to recognize GSI certificates to handle data ownership, etc., across Grid
- IBM & SDSC worked on GSI-UID mapping

Applications demo at SC'04/StorCloud

- **Large (1024 processors) Enzo application running at SDSC in San Diego**
- **Writes ~50 TB directly to SC'04 StorCloud disks mounted as Global File System across Country**
- **Data analyzed and visualized at NCSA in Illinois and at SDSC**
- **True Grid application, running across USA, saw ~3 GB/s**

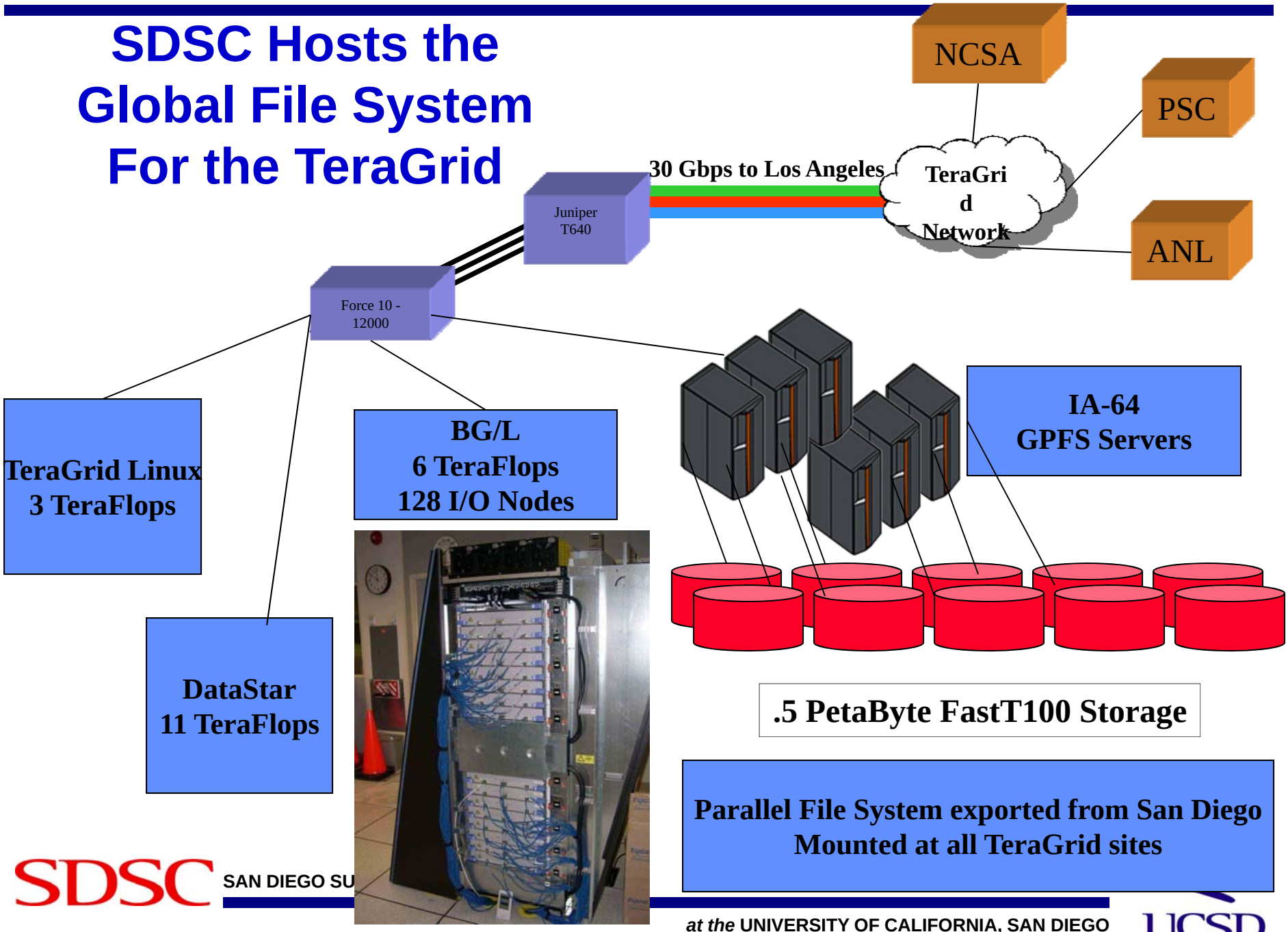
WAN-GFS-ROC continued at SC'04



SDSC installing 0.5 PB of IBM disk now

- Plan to start hosting large datasets for the scientific community
- First will be NVO, ~50 TB of Night Sky information; Read-Only dataset available for computation across TeraGrid
- Extend rapidly with other datasets
- Planning on 1 PB before end of year

SDSC Hosts the Global File System For the TeraGrid



Lessons Learned

- **For real impact, close collaboration with vendors is essential during both development and implementation**
- **There is a natural migration of capability from experimental middleware to entrenched infrastructure**
- **Middleware providers should welcome this, and respond with new innovations**

Predictions (of the Future)

- **Many sites will have lots of cheap disk**
- **Relatively few will invest in significant Hierarchical Storage Management systems**
- **DataSets will be locally cached from central GFS/HSM sites across the Grid**
- **Replication becomes part of Grid infrastructure**
- **Data integrity and ultimate backup will be the province of a few “copyright library” sites**
- **Data is the birthright and justification of the Grid**