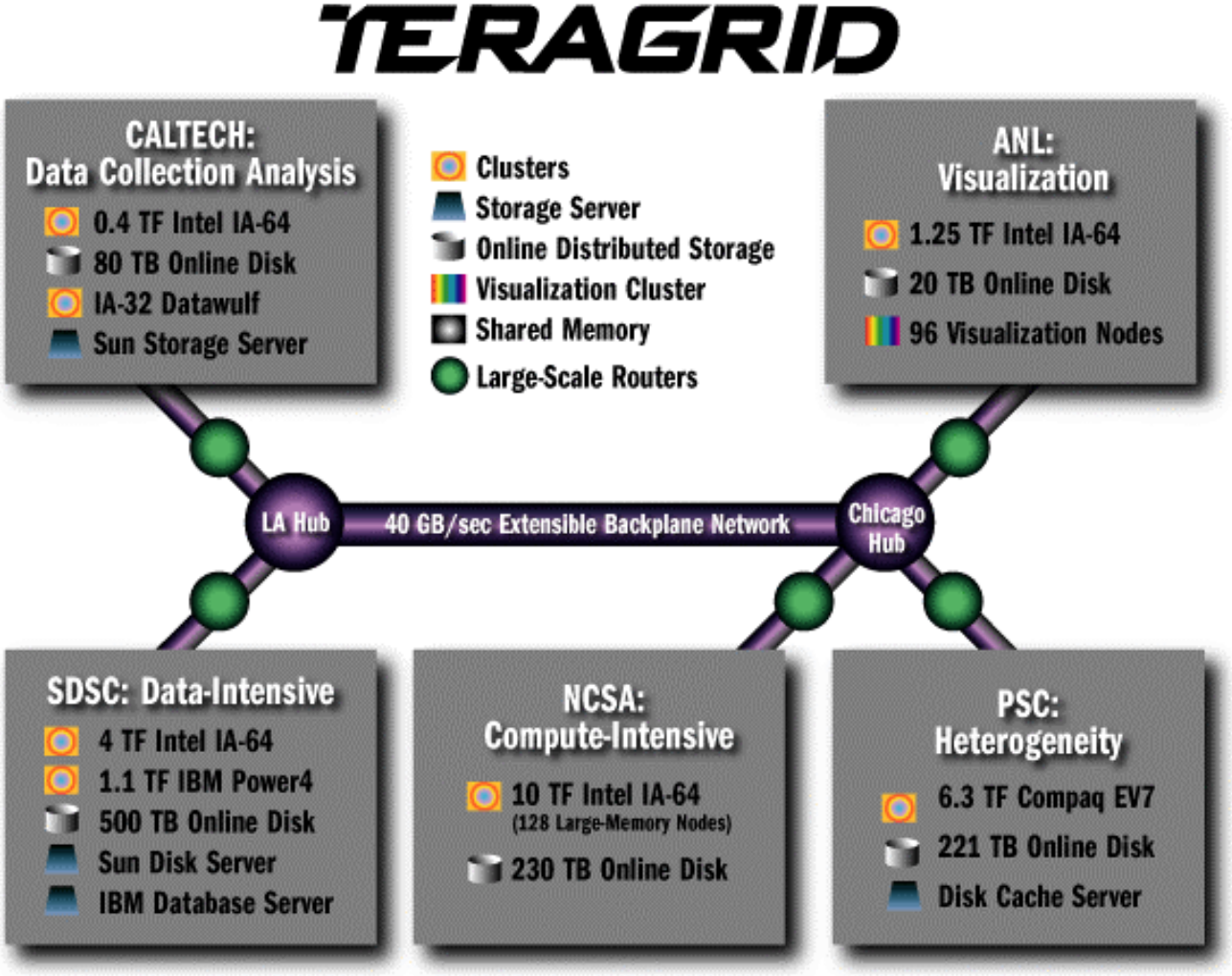
A Centralized Data Access Model for Grid Computing

–

Phil Andrews, Tom Sherwin, and Bryan Banister
San Diego Supercomputer Center
University of California, San Diego
La Jolla, Ca 92093-0505
andrews@sdsc.edu, sherwint@sdsc.edu, bryan@sdsc.edu

Data and Grid Computing

- **Normal Paradigm: Peripatetic 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**

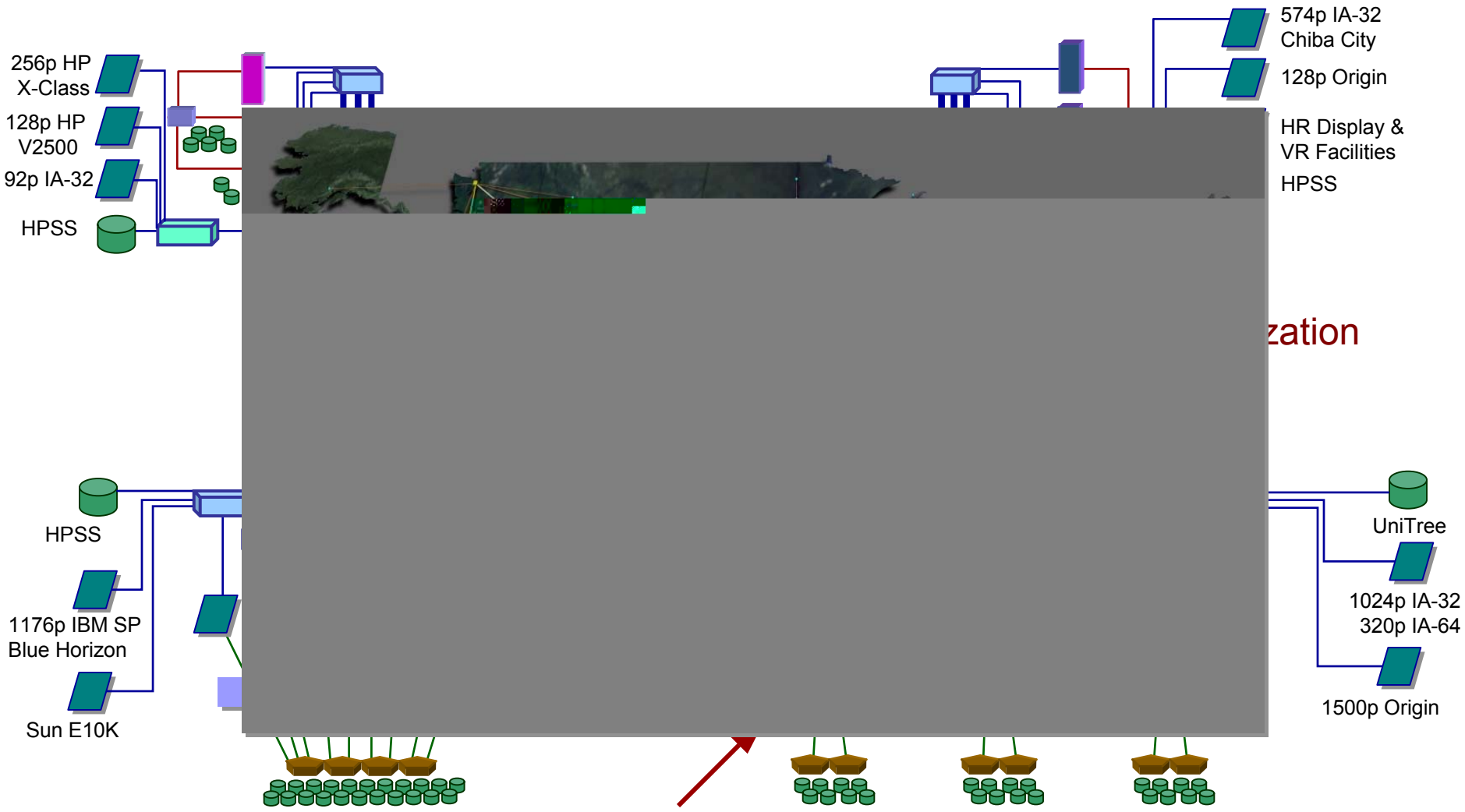


Prototype for CyberInfrastructure

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
- Over 20 Teraflops compute power
- Approx. 1 PB rotating Storage
- Extending by 3-4 sites in ‘03

TeraGrid


574p IA-32
Chiba City
128p Origin
HR Display &
VR Facilities
HPSS
256p HP
X-Class
128p HP
V2500
92p IA-32
HPSS
zation
HPSS
UniTree
1024p IA-32
320p IA-64
1176p IBM SP
Blue Horizon
1500p Origin
Sun E10K
NCSA: Compute-Intensive

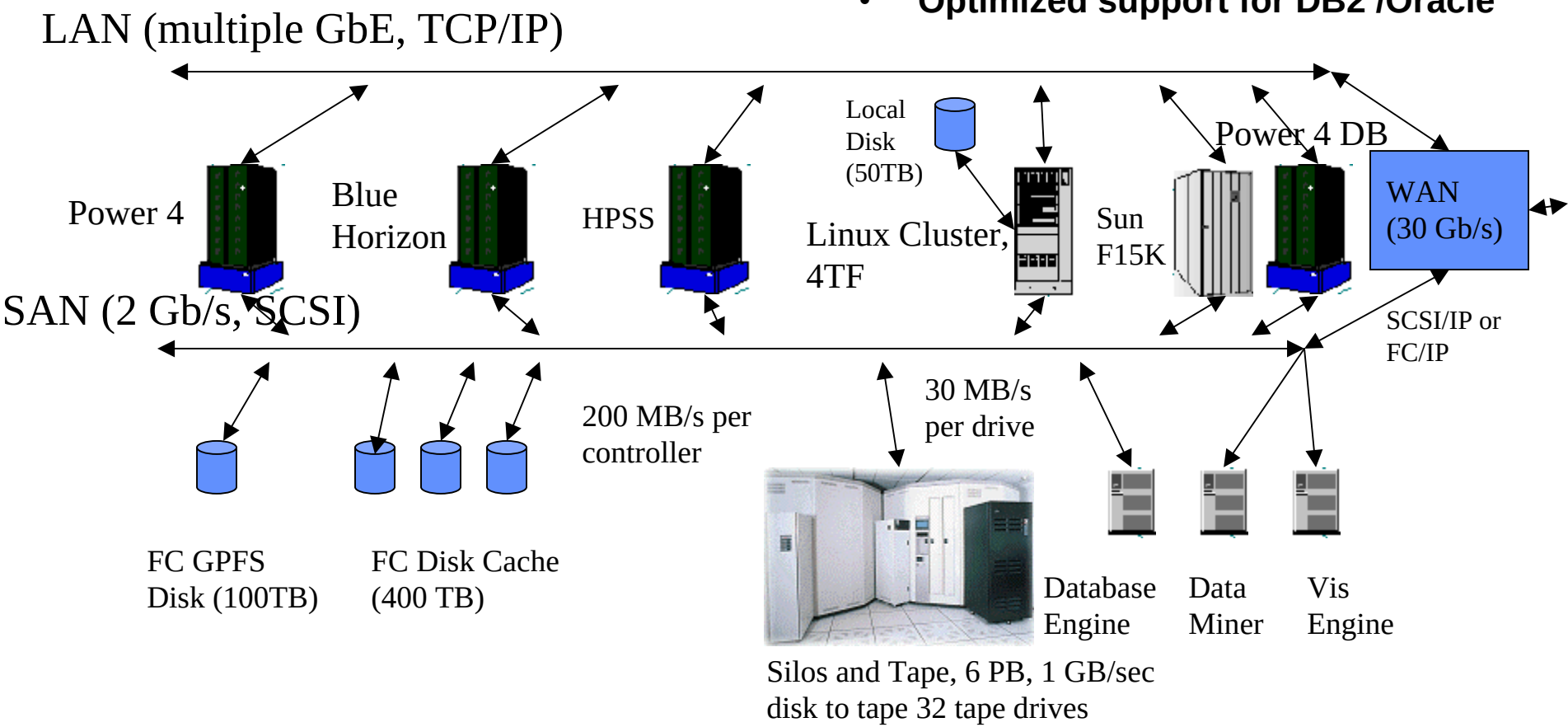
Alternate, Centralized Data Approach

- **SDSC is designated Data Lead for TeraGrid**
- **Over 500 TB of Disk Storage at SDSC in ’03**
- **Large Investment in Database Engines (72 Proc. Sun F15K, Two 32-proc. IBM 690)**
- **32 STK 9940B Tape Drives, 6 PB Capacity**
- **Storage Area Network: Over 500 2Gb/s ports**
- **Almost all disk on SAN: not direct attached**

SDSC Machine Room Data Architecture

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

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



Extending Data Resources to the Grid

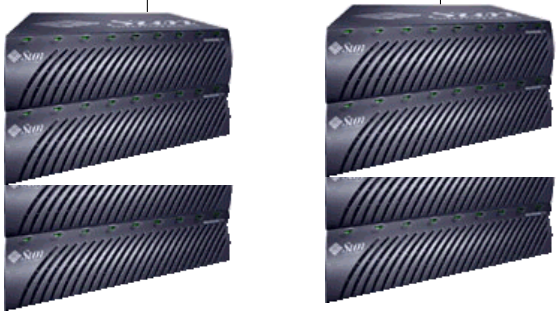
- **Aim is to provide apparently unlimited Data Source at High Transfer rates to whole Grid**
- **Jobs would access data from Centralized Site mounted as local disks using WAN-SAN**
- **Use very large Data Cache (~400TB)**
- **Rapid (1 GB/s) transfers to Tape for automatic archiving**
- **Multiple possible approaches: presently using Sun's QFS File System and SAM-FS HSM**
- **Used Prototype system for investigation**

SAM-FS Prototype Configuration

SAM-FS manages disk-tape transfers, presents apparently infinite resource to users

6 PB capacity, 828 MB/s

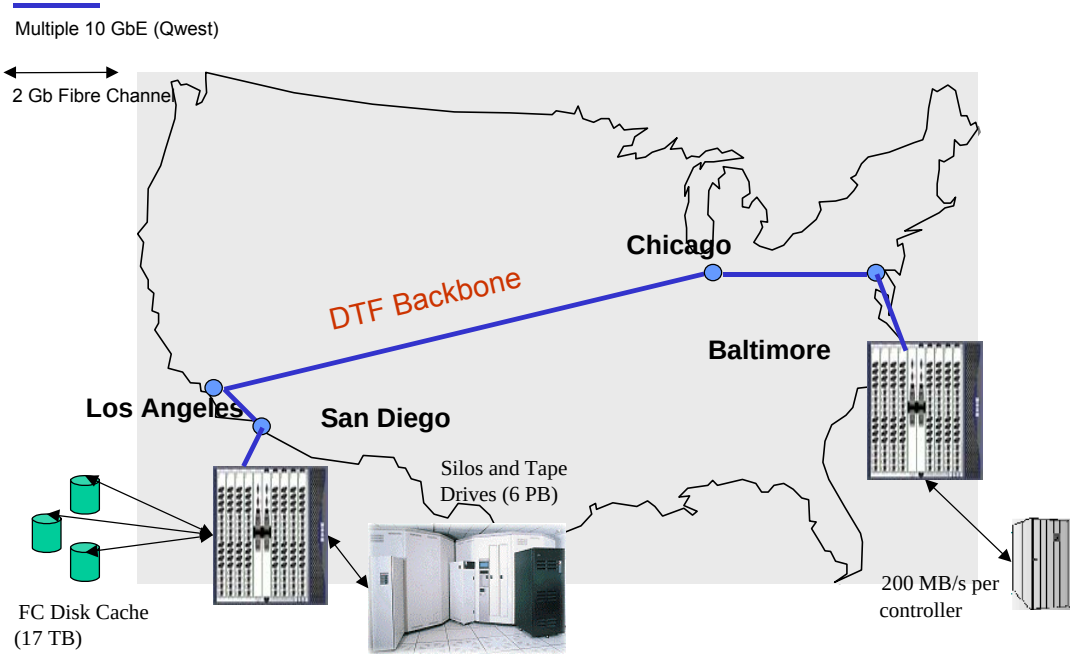
30 TB
3.2 GB/s



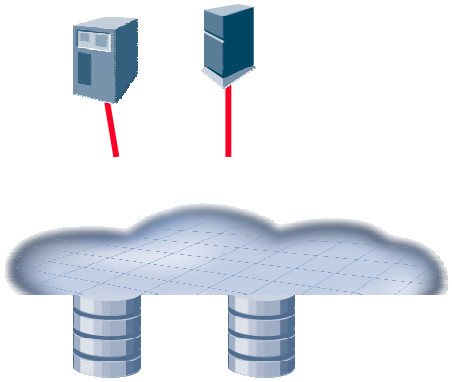
Building up a Test Data Configuration

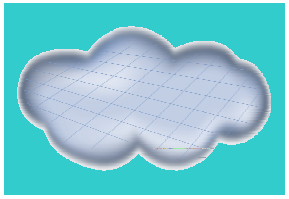
- **Used 30 TB of Sun T3B disk**
- **QFS File System and SAM-FS HSM on Sun F15K**
- **24 STK 9940B tape drives**
- **Demonstrated 3.2 GB/s reads from File System**
- **Writes/Reads from Tape at 800+ MB/s using QFS/SAM-FS HSM**
- **Needed to test Latency effects in realistic setting**

Networking for SC'03 Demo



SAN-SAN Interconnect over IP Networks



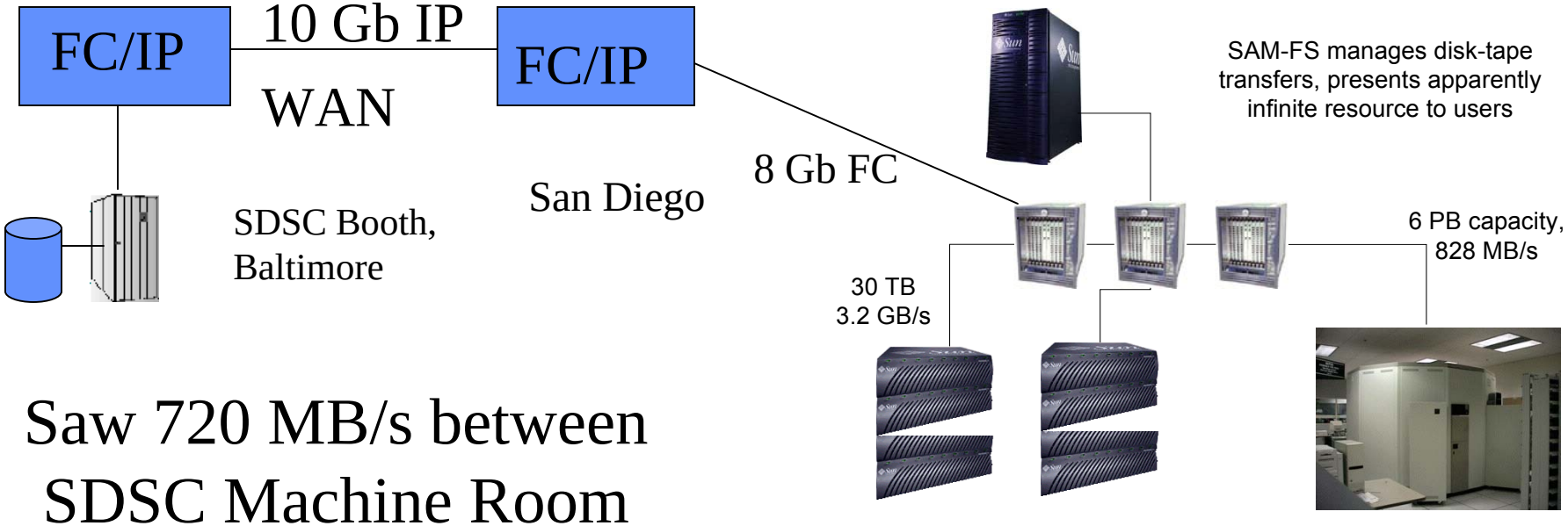


Transferring Data

- If data source works, what is preferred transport mechanism?
- Try FC/IP (Encoding Fibre Channel Frames within IP packets) for transparent SAN extension
- WAN-SAN approach allows remote Data System to appear local across Grid
- Requires hardware FC/IP encoding/decoding
- 8Gb/s gear provided by Nishan Systems
- Used 10 Gb/s link between San Diego and Baltimore

SDSC SC'02 TeraGrid Demo

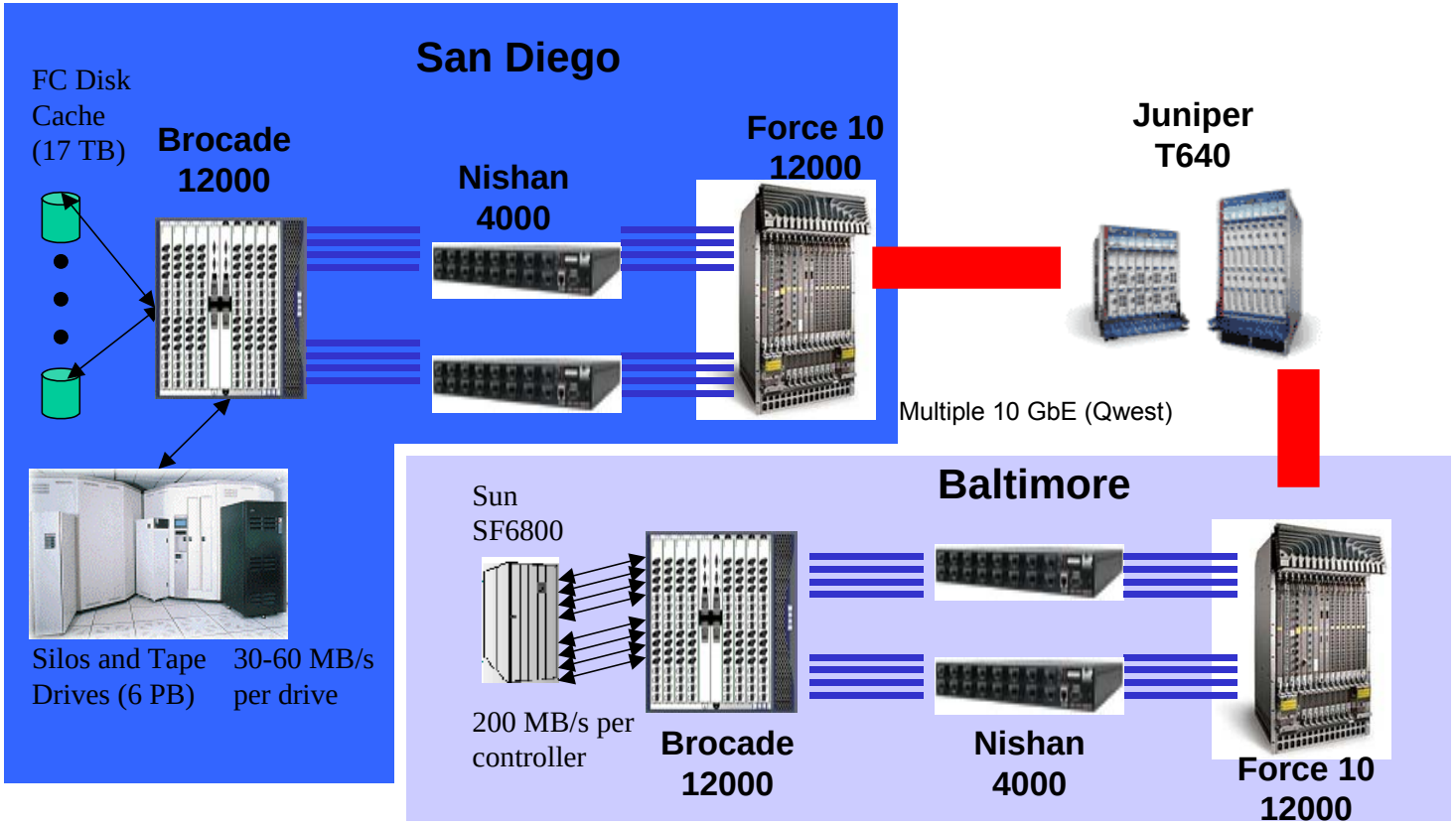
Baltimore



Saw 720 MB/s between SDSC Machine Room and SDSC Booth

San Diego

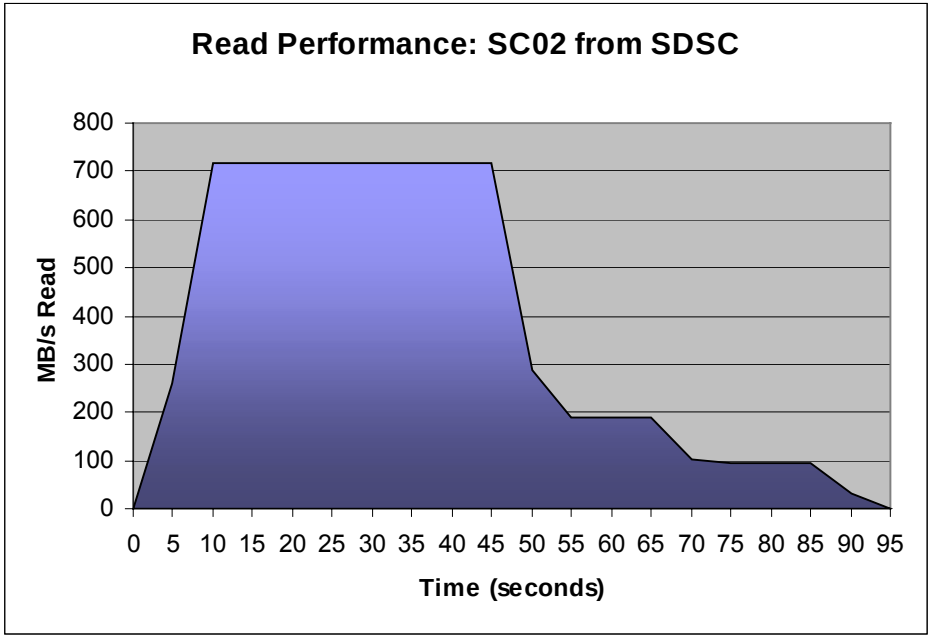
Networking Details:

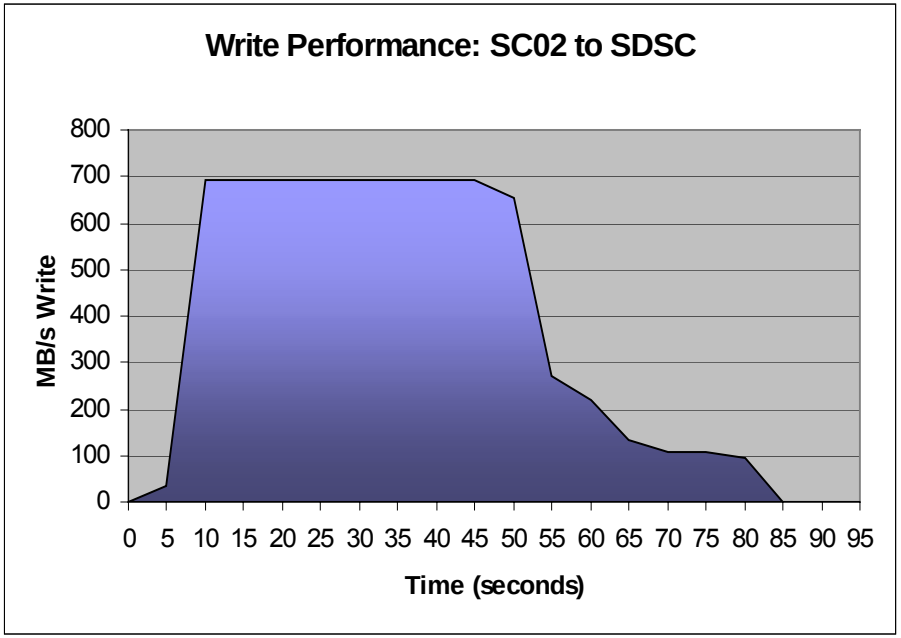


Results from SC'02 WAN-SAN Demo

- **Extended the SAN from the SDSC Machine Room to the SDSC booth at SC'03**
- **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:

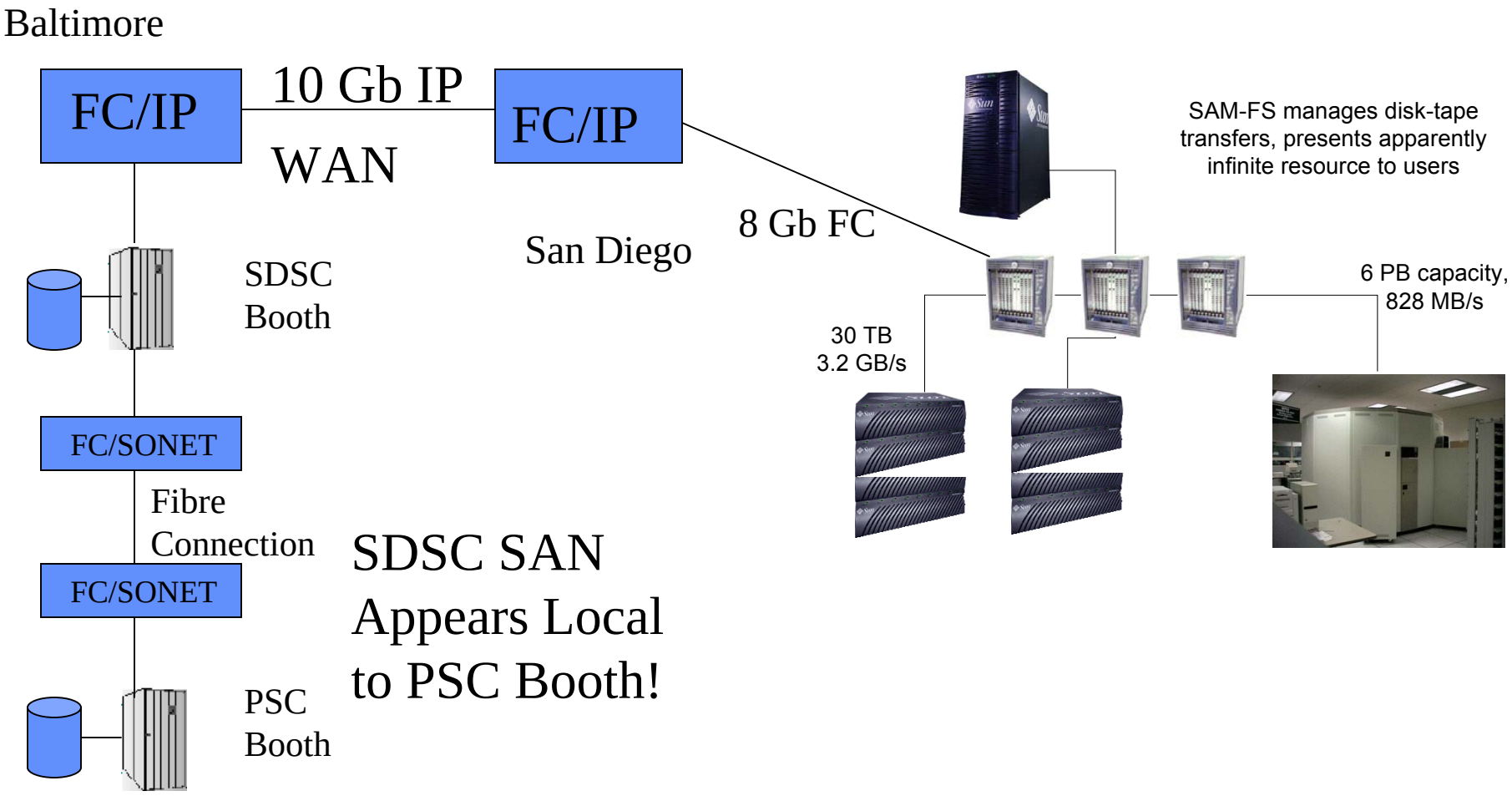




Also Used FC over SONET encoding

- **Extended the SAN from the SDSC Booth to the PSC booth at SC'03 using dedicated Fibre**
- **Able to access Tape Drives in SDSC Machine room from PSC booth using WAN-SAN**
- **Not enough equipment to push the connectivity**
- **SAN protocol survived double encode/decode (FC/SONET, FC/IP)**

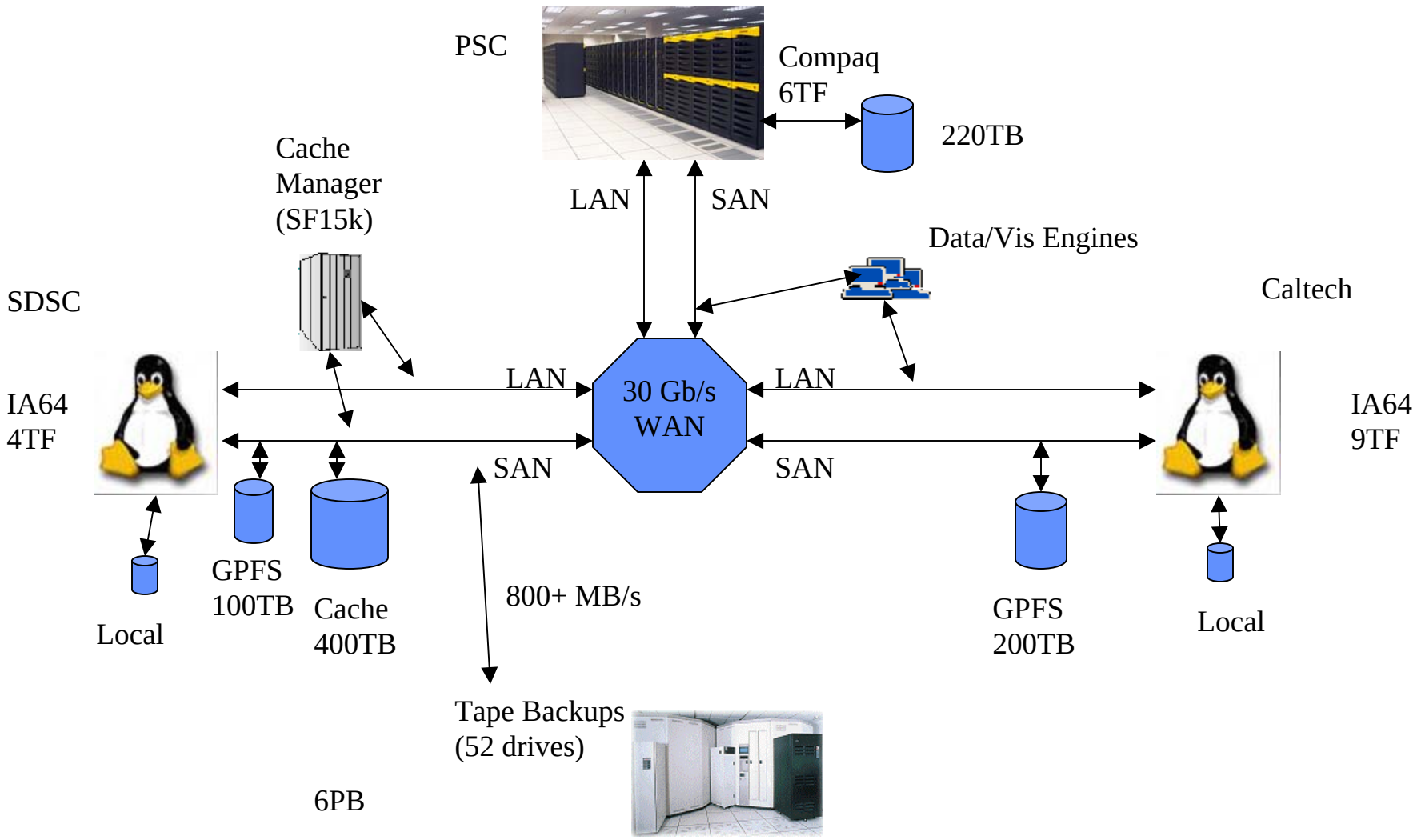
SDSC-PSC TeraGrid Demo



Lessons Learned, Moving Forward

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

Possible TeraGrid Grid Data Architecture



Working On or Need:

- **Must be able to share File Systems: working with Sun to port QFS client to other architectures**
- **Impractical to require coordinated UIDs across Grid: need Certificate based authentication**
- **First use will be Large, Read-Only Datasets**
- **Latency must be designed for**
- **Continuing to examine other File Systems (GPFS, PVFS, Lustre, etc.)**

Next Steps:

- **Obtained equipment for permanent WAN-SAN between SDSC and Caltech**
- **Initially will split HSM (SAM-FS) so that Disk Cache for Caltech is local, but use large tape drive facility at SDSC**
- **May be model for more HSMs in future as Tape storage consolidates at large sites**
- **Should be latency tolerant**
- **Looking for use with other TeraGrid partners**

Eventually:

- **Very large (semi-infinite) data resource at SDSC**
- **All other Grid sites have direct access as if local**
- **Important Scientific Datasets made publicly available to all Grid Users**
- **Data access removed as problem for Grid**
- **Universality of Data Access becomes defining capability for Grid Computing**