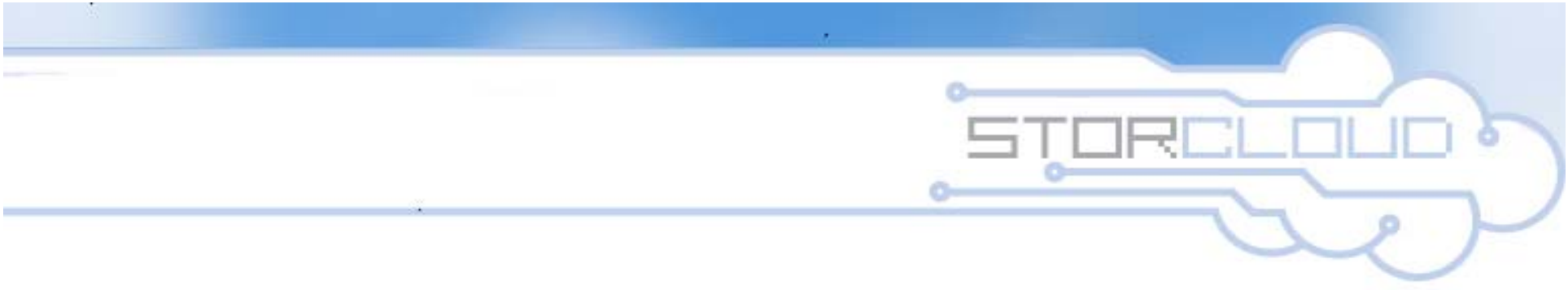


SC04 StorCloud Tutorial

A Peta-byte Heterogeneous Storage Initiative

IEEE/NASA MSST2005
April 11, 2005

Michael Knowles (Raytheon, ARL/MSRC), ‘04 Architecture
Virginia To (Sandia National Laboratories), ’04 Chair

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

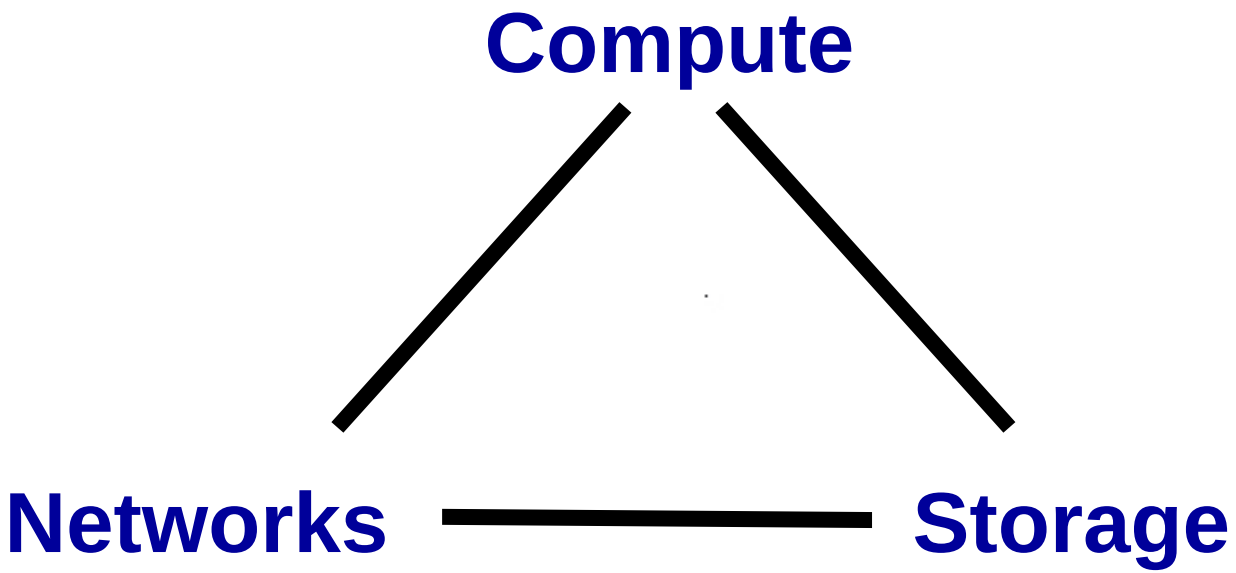
SC-Conference Series

- Formally: “Super Computing XY”
- Formally: “SCXY: High Performance Computing and Networking Conference”
- IEEE/ACM Sponsored Annual Event
- Different Location Yearly
- Different Executive Committee Yearly
- ~ 8000 Attendees
- Exhibitors: (1/3 HPC Equipment & Network Vendors, 1/3 Research Labs & Academia, 1/3 Other (Storage, SW, Interconnect, Integrators))

SC-Conference Series

- International Flavour (~1/3 non US)
- Special Initiatives (SCGlobal ’03 (Phoenix), Storage ’04 (Pittsburgh), Storage and Analytics ’05 (Seattle))
- SCInet
  - 8+ Years Building Usable High Speed Networks During Show
  - StorCloud Model – Usable Showcase
- Good Natured Research Booth Competition
  - Bandwidth Challenges
  - Non-Vendor Related, Application Oriented
- Significant Growth in Last 3 Years

SC2004 Brings Storage to SC Series



SC2004 Brings High Performance Storage to "SCXY: High Performance Computing, Networking and Storage Conference" Series

From Concept to Reality

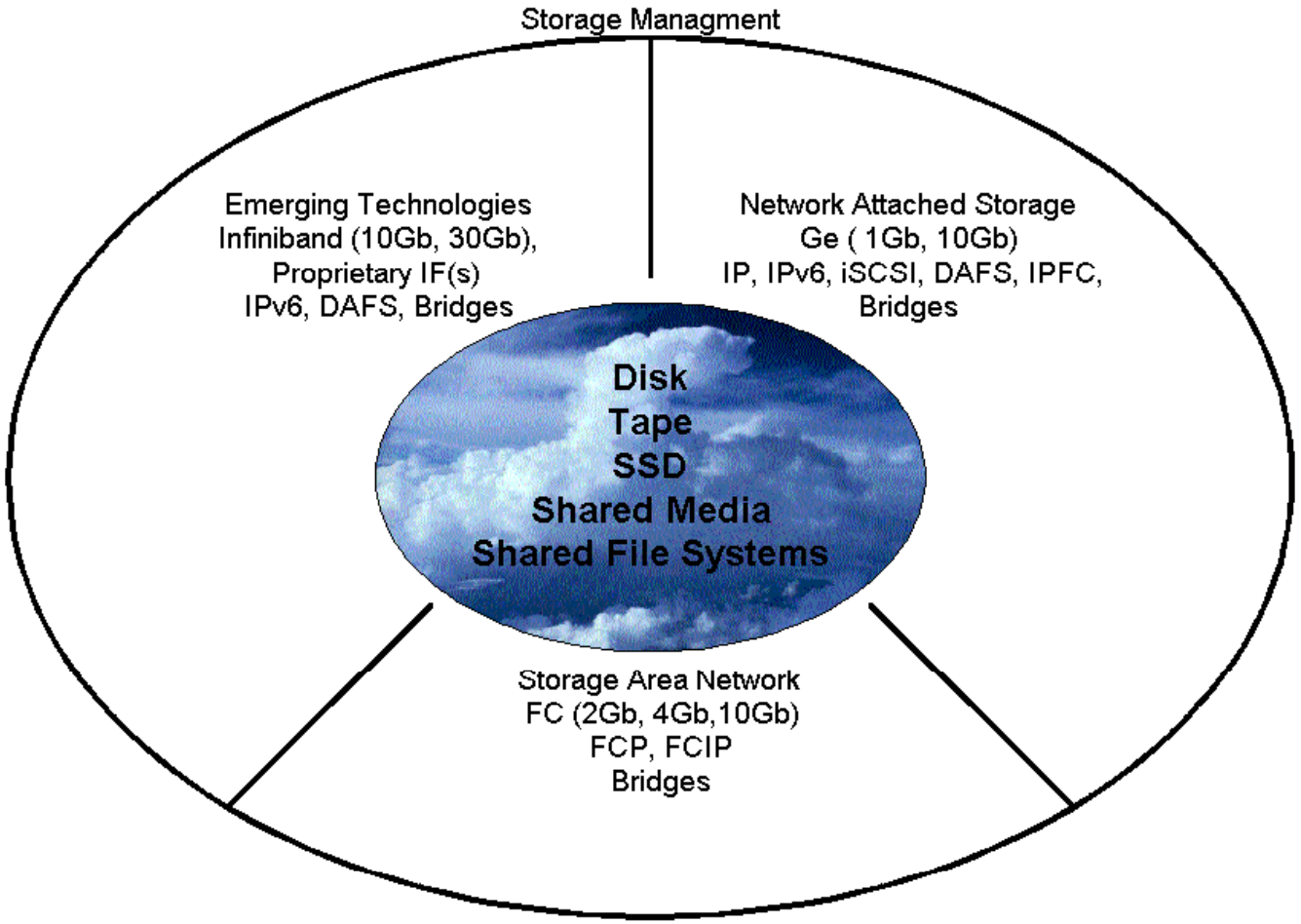
- Two years of work from concept to reality
  - StorCloud idea conceived at SC2002, Baltimore over a very good Chilean Merlot and several Napkin Diagrams
  - Presented initial proposal to SC2004 Executive Committee – Nov 2002
  - Developed and presented formal proposal – Jan 2003
  - Executive committee approval - March 2003
  - Vendor solicitations began - March 2003
  - Applications solicitation began - Nov 2003
  - Applications commitments – July 2004
  - Finalized architecture - Sept 2004
  - Pre-staging and Staging – Sept to Nov 2004
  - StorCloud Debut – Nov 2004

Objectives

- **Objectives** – showcase high performance data storage
  - Storage Technologies, Access, Management
  - Emergent technologies: Topologies/Interconnects/Protocols
  - StorCloud Challenge, an applications competition
  - Available/Usable to SC2004 Research Exhibitors

- **Goals**
  - Expand SC Conference focus to include storage
  - Proposed targets
    - 1 PetaByte multi-vendor rotating storage
    - 1 TeraByte/sec bandwidth to/from storage
    - 1 GigaByte/sec backup capability
  - Enhanced support for exhibitors with HPC storage needs
  - Raise the storage intensive applications performance bar
  - Enhanced recognition for storage vendors

Architectural Concept



Partnerships

- Research Booths (Virtually All Academic and Government Sponsored)
  - Provided Applications
  - Provided Design Requirements
  - Provided Storage Drivers
- Storage, Networking & HPC Vendors
  - Contributed HW/SW
  - Contributed Configuration, Installation and Configuration Expertise
- Executive Committee
  - Sponsored & Funded StorCloud Special Initiative
- StorCloud Committee
  - Designed Booth & Baseline Architecture
  - Provided Environmentals/Infrastructure
  - Provided Logistics/Security/Manpower
  - Allocated Resources to Exhibitors
  - Coordinated Events/Press

Partnership took StorCloud from concept to reality

Partners:
Research Booths

- The Challenge Applications expected to stretch StorCloud to its limits
  - Sandia National Labs (SNL)
  - Larence Livermore National Laboratory (LLNL)
  - Pacific Northwest National Laboratory (PNNL)
  - San Diego Supercomputer Center (SDSC)
- Non-challenge Application exercise other dimensions of StorCloud
  - National Center for Data Mining
  - Marconi
  - SDSC
  - Ohio Supercomputing Center (OSC)
  - SNL
  - Multi-vendor (Cisco, McData, Brocade) interoperability application
  - National Institute of Advance Industrial Science and Technology (Japanese Research Consortium)

Partners: Storage, Networking & HPC Vendors

- 3ParData
- Brocade
- Cisco
- Cluster File Systems
- Data Direct Networks
- Dell
- Finisar
- IBM
- IBRIX
- Intransa
- Linux Networx
- Marconi
- McDATA
- Mellanox
- Network Appliance
- Panasas
- Qlogic
- RedHat
- SGI
- StorageTek
- Systimax
- Terrascale
- Veritas

Partners:
SC04 Executive Committee

Conference General Chair, Jeff Huskamp (University of Maryland)
Conference Vice Chair, Beverly Clayton (Pittsburgh Supercomputing Center)
Conference Deputy Chair, Bill Kramer (NERSC, Lawrence Berkeley National Laboratory)
Executive Director, Virginia To (High Performance Technologies, Inc.)
Finance Vice Chair, Sandra Huskamp (East Carolina University)
Communications Vice Chair, Vivian Benton (Pittsburgh Supercomputing Center)
Program Vice Chair, John Grosh (Department of Defense)
Research Exhibits Co-Vice Chairs, Bruce Loftis (NCSA), Nancy Wilkins-Diehr (SDSC)
Industry Exhibits Co-Vice Chairs, Becky Verastegui (ORNL), Eric Sills (North Carolina State University)
Industry Liaison Vice Chair, Ray Paden (IBM)
Education Vice Chair, Joyce Williams-Green (Winston-Salem State University)
Minority Serving Institutions Outreach Vice Chair, Stephenie McLean (NCSA)
SC Global Vice Chair, Jackie Kern (National Center for Supercomputing Applications)
SCinet Vice Chair, Charles (Chuck) Fisher (Oak Ridge National Laboratory)
Special Initiatives Vice Chair, Ernie Marshburn (East Carolina University)
Conference Arrangements Vice Chair, Elvira Prologo (Pittsburgh Supercomputing Center)
Logistics Vice Chair, Eleanor Schroeder (Naval Oceanographic Office)
Student Volunteers Co-Vice Chairs, Ken Hackworth (PSC), Barbara Horner-Miller (ARSC)
IEEE Computer Society Representative, Anne Marie Kelly
ACM SIGARCH Representative, Donna Baglio

Partners:
StorCloud Committee

- **StorCloud Chair:** Virginia To (HPTi)
- **Applications Team**
  - **Chair:** Ken Washington (SNL)
  - Helen Chen (SNL)
  - Cary Whitney (NERSC)
- **Challenge Team**
  - **Chair:** Phil Andrews (SDSC)
  - Steve Louis (LLNL)
  - Roger Haskin (IBM)
  - Ruth Aydt (NCSA)
  - Thomas Ruwart (U of Minnesota)
  - Steve Waterhouse (Sun)
- **Architecture Team**
  - **Chair:** Mike Knowles (Raytheon)
  - Mike McCraney (MHPCC)
  - Dov Cohen (SNL)
  - Bryan Bannister (SDSC)
- **Integration Team**
  - **Chair**: Tom Kendall (ARL MSRC)
  - Ian Lumb (Platform)
  - Tom Kile (ARL MSRC)
  - Bill Nickless (ANL)
  - Jennifer Hare (U of Maryland)
- **Student Volunteers**

Partners:
Core Committee

Partners:
Committee Organizations

- Argonne National Labs
- ARL MSRC
- DoD High Performance Computing Modernization Program
- East Carolina University
- High Performance Technologies Inc
- Lawrence Livermore National Labs
- Maui High Performance Computing Center
- NCSA
- NERSC
- Raytheon
- Sandia National Labs
- SDSC
- University of Delaware
- University of Maryland
- University of Minnesota

Partners:
Student Volunteers

Every cloud has a silver lining

STORCLOUD

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

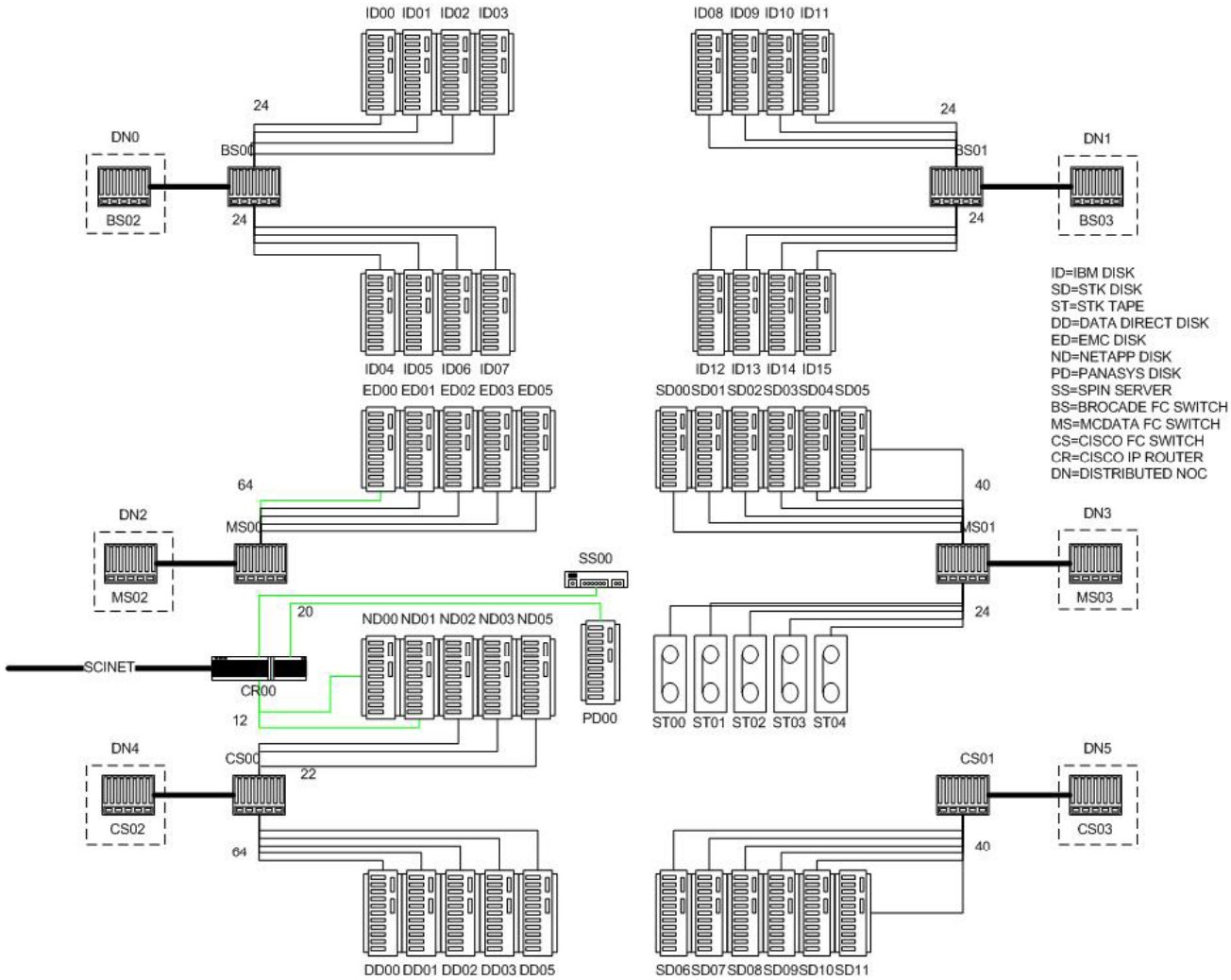
Architecture Guidelines

- Comprised of Various Storage Technologies
- Comprised of Heterogeneous Equipment & Vendors
- **Vendors Contribute/Ship/Install/Maintain/Monitor Their Own Equipment**
- **Equipment/Capacity/Connections Allocated By SC Committee to Conference**
- Few Inter-Vendor Connections (Interoperability)
- SC Committee Provided:
  - Power/Cooling/Showcase (Central SC Site)
  - Booth Drops for Applications/Challenge & SC Related Vendor Contributions
  - Allocation & Coordination of Resources
  - Special Events
  - Advertising
- Applications/Challenge/Vendors Control Their Own Allocated Resources
- SC Resources Available to Applications/Challenge Users Sunday Evening
- Production Oriented Services

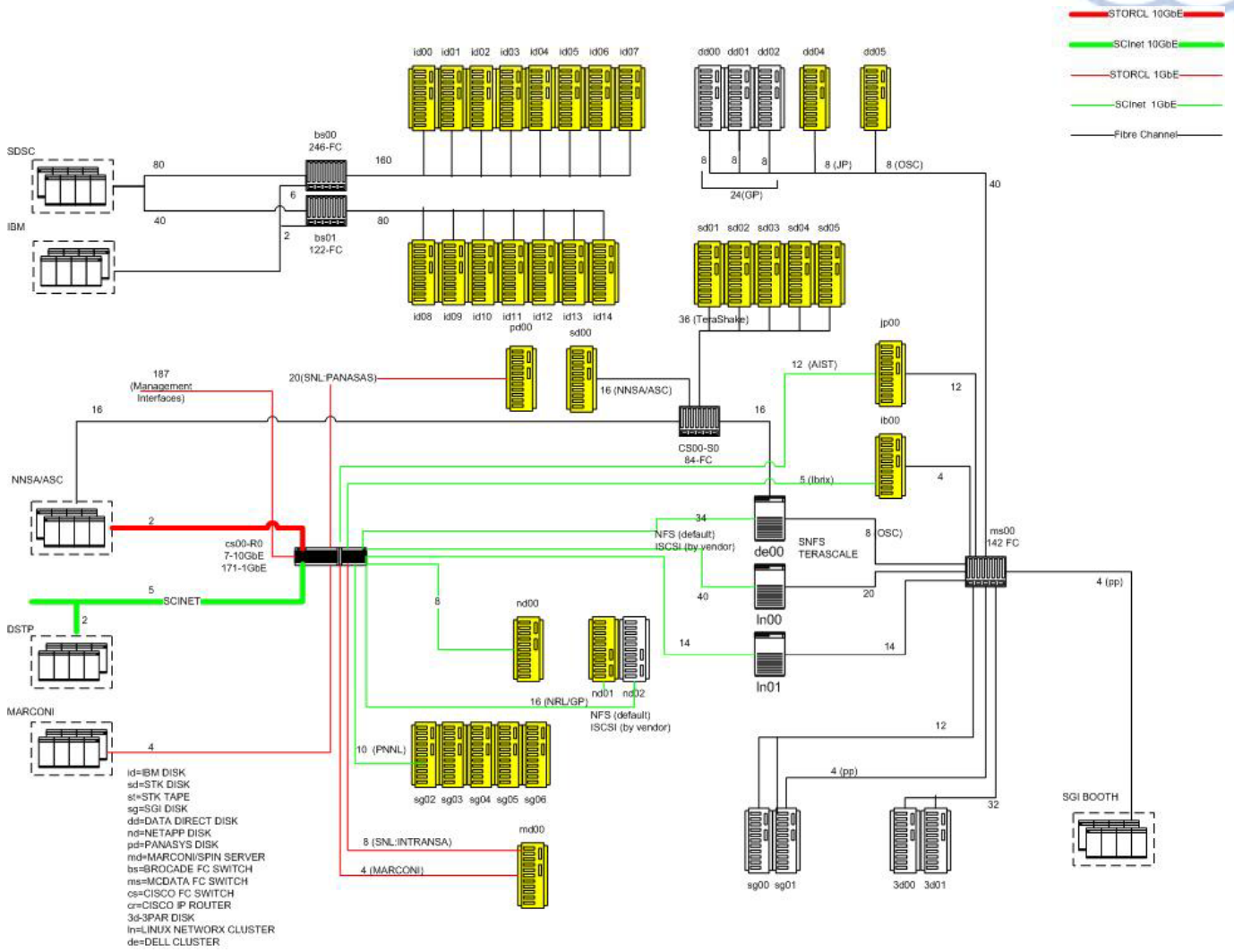
Resource Allocation Policy

- Allocation Profile
  - First Allocation to Contributing Vendors (20-30%)
    - Raw Resources
    - Vendor File Systems, Demos,...
    - Teaming with Applications Encouraged (Subject to SC Committee Priorities)
  - Second Allocation to Application/Challenge Users (60-70%)
    - Raw Resources
    - SC Committee Will Work with Applications and Vendors to meet Requirements
    - Teaming with Vendors Allowed (Subject to SC Committee Priorities)
  - Third allocation to Conference at Large (10%)
    - Formatted File Systems (NAS)
    - Allocations by IP Address
    - Quota Controlled
    - Distribute Conference and SC Info & Events

Logical Diagram 3/21/04 (Iteration 3)



Logical Diagram
3/21/04 (Iteration 333)



Networks

- Fibre Channel Networks
  - External – To Vendor & Research Booths
    - Remote/Booth Hosts to Systimax Patch Panel In Booth, to Patch Panel in StorCloud
    - 4-72 Fibre Multi-Fibre Spools
    - 2-24 Fibre Multi-Fibre Spools
    - 2-24 Fibre Multi-Fibre Spools
  - Internal – Within StorCloud Booth
    - Used for External and Internal Host Access to Storage
    - Patch Panels to Switches, Switches to Storage
    - Internal File Servers to Switches, Switches to Storage
- IP Networks
  - Public - Provide Conference & World Access
    - SCInet Allocated Addresses
    - StorCloud Internal File Servers
    - StorCloud Network Attached Storage (NAS)
  - Requested - Requested Networks by Research Booths
    - StorCloud Allocated Addresses
    - Private NAS
    - StorCloud Internal File Servers
  - Private – Non-Routed Management Network
    - StorCloud Equipment Management
    - Accessible Through SCInet but Heavily Filtered (IP to IP)

Fibre Channel Network(s)

- External StorCloud Fibre Channel Network
  - Primary Storage Access Mechanism For Applications & Challenge Booths
  - Fibre Channel & Storage Allocations By Request
  - Fibre Routed from Application Booth to StorCloud Booth, Through Switch, then to Preallocated Storage
  - Allocations/Access Controlled by SRM SW at Switches
  - Coordinated & Implemented by StorCloud Committee
  - 2Gb, 50/125 MM Fibre Patch Cables in Booth to StorCloud Provided Patch Panel
  - Traffic Across Pre-run Systimax Multi-Fibre Cables to StorCloud Patch Panel
  - 156 FC Pairs
- Internal StorCloud Fibre Channel Network
  - Provided Storage Access For Internal and External File Servers Through Switches
  - Internal Storage Allocated/Accessed Through StorCloud Servers Only (NFS and ISCSI)
  - External Storage Allocated Through Patch Panel Connected Fibre
  - Allocations/Access Enforced by SRM SW and Servers
  - Coordinated & Implemented by StorCloud Committee
  - 624 FC Pairs

FC Switch Resources

- All Access Via 5 Core Switches
  - 624 Ports 2Gb (SFP)
  - No Edge Switches (For StorCloud Usable Storage)
  - Zones Used Across Intra-Vendor Switches to Control Access
  - 3-Brocade 12000
    - All Ports Dedicated to SDSC and IBM Collaboration
    - 40 External Hosts with 3 HBA Each
    - Minimal ISL(s) (1-2Gb Link Per Switch)
  - 2-McData 6140 (10Gb ISL)
    - Ports Used for Internal and External Host Access
    - 2 10Gb ISL for Inter-Switch Traffic, Resources Allocated Across
    - 32 Dedicated to Applications
  - 1-CISCO MDS-9000
    - Ports Used for Internal and External Host Access
    - 84 Ports Dedicated to Applications

FC Disk Resources

- FC Disk Arrays
  - 260 Ports 2Gb (SFP)
  - ~1946 Disks
  - ~377 TB
  - Most 73 & 146 GB FC some SATA
  - IBM
    - 30-FASTT 600 Controllers, All FC Disk
    - 15 Racks
  - Storage Tek
    - 13-D280 Controllers, All FC Disk
    - 6 Racks
  - Data Direct
    - 8-S2A8500 Controllers, FC and SATA Disk
    - 5 Racks
  - 3par
    - 1-InServ S800 Controller, All FC Disk
    - 2 Racks
  - SGI
    - 1-TP9500 (FC Disk), 1-TP9300 (SATA Disk, 400GB)
    - 2 Racks

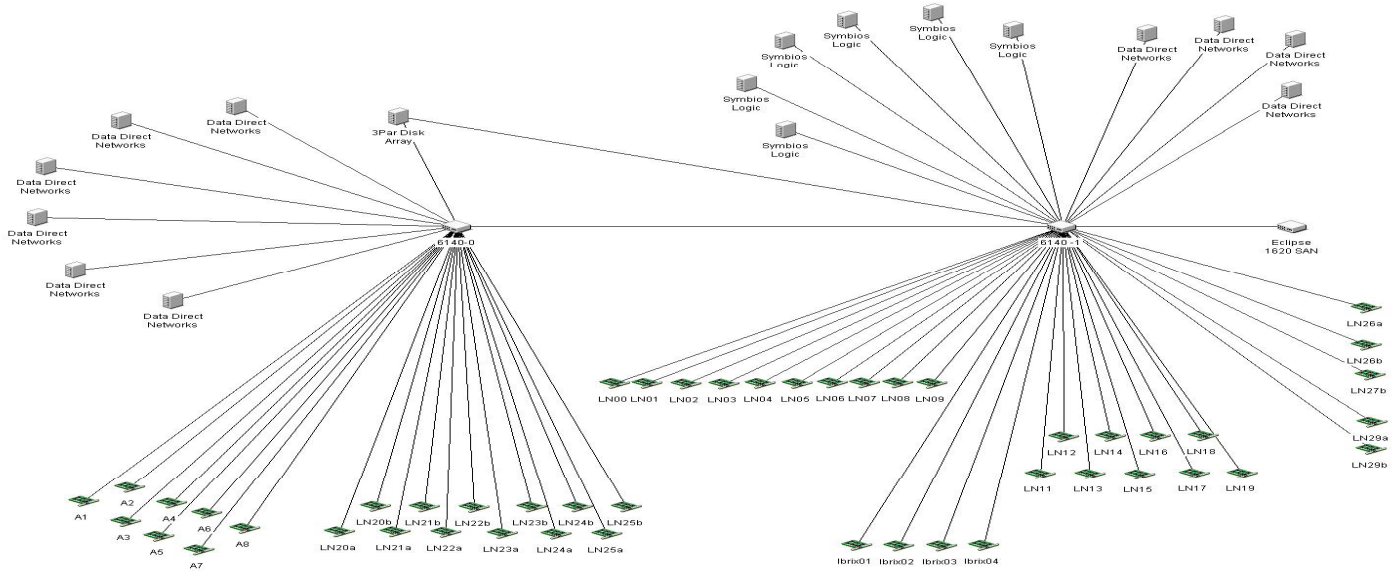
FC Host Resources

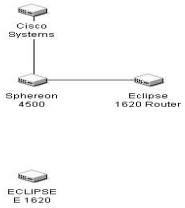
- 48 Internal Hosts
  - All Linux File Servers
  - All Qlogic 2300 HBAs
  - Supporting Several File Systems
  - 9 Dedicated to Applications
  - 39 Supporting General Access
- Dell
  - 16-1850 Nodes, 2 Xeon@3.4GHz, 4 GB Memory
- Linux NetworX
  - 20-Evolocity II, 2 Xeon@3.2GHz, 2 GB Memory
  - 7-2u Storage Nodes, 2 Xeon@2.4GHz, 4 GB Memory
- IBRIX
  - 5-Xeon Based Servers, 2 Xeon@3.2GHz
- Advanced Industrial Science & Technology (AIST, JAPAN)
  - 12-Intel Xeon Based, Damaged in Shipping

FC Based File Systems

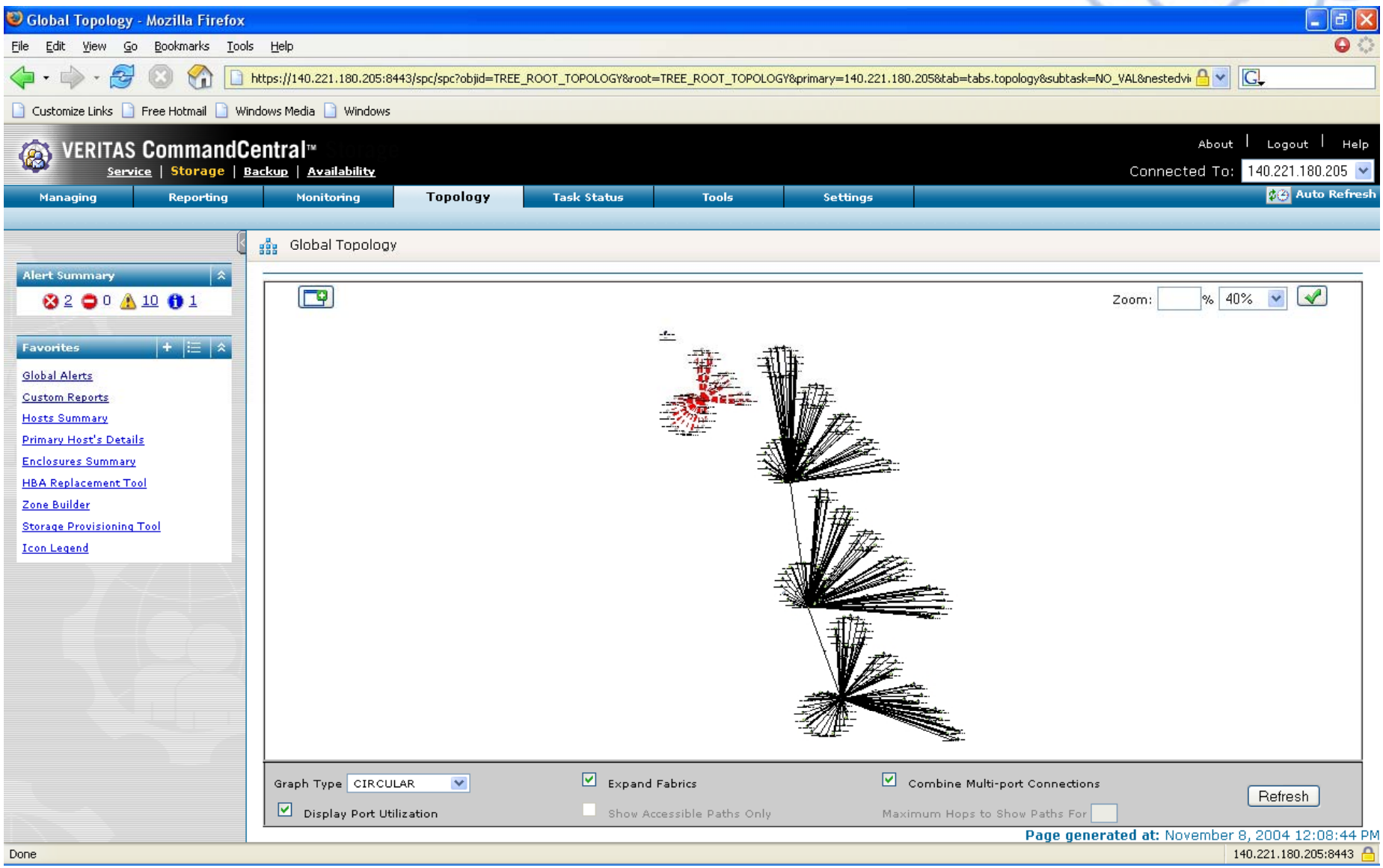
- GPFS
  - 40 SDSC External Hosts, 3 HBA(s) Each (Dedicated, SDSC/IBM)
  - 120 FC Links, IBM FT600 Disk, ~ 90 TB, ~1300 Disks
- PVFS
  - 8 StorCloud Dell Hosts (Dedicated, OSC)
  - 8 FC Channels, DDN FC Disk, ~40 TB
- IBRIX
  - 20 StorCloud Linux NetworX and 5 IBRIX Hosts (General Purpose)
  - 32 FC Channels, DDN FC & ATA Disk, ~80 TB
- STORNEXT
  - 7 StorCloud Linux NetworX Hosts (General Purpose)
  - 16 FC Channels, 3PAR Disk Array, ~20 TB
- TERRASCALE
  - 8 StorCloud Dell Hosts (Shared)
  - 8 FC Channels, STK D280 Disk, ~40 TB
- CXFS
  - ? SGI Hosts (Dedicated/Private)
  - 4 FC Channels, SGI TP9500 Disk, ~12 TB

McData FC Topology

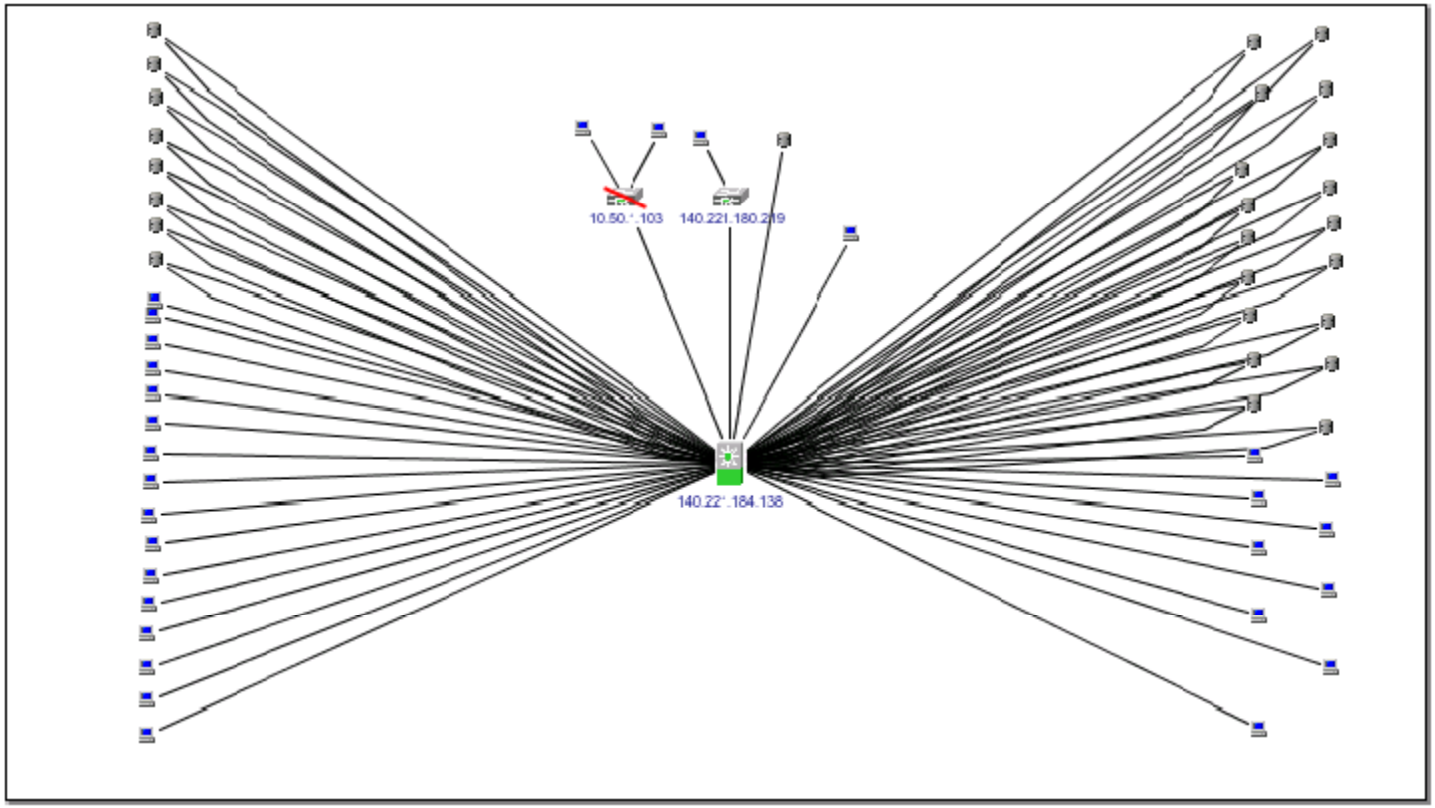




Brocade FC Topology



Cisco FC Topology

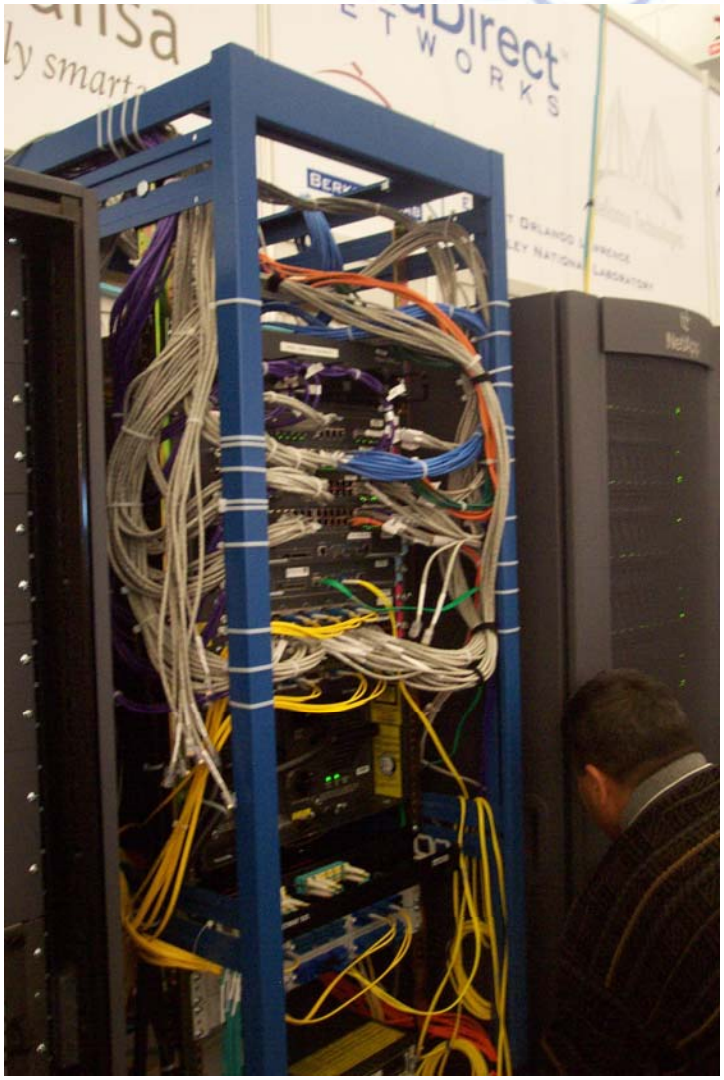


IP Network(s)

- Public Networks – Provide Conference & World Access Into StorCloud
  - Regular SCInet Network
    - Set Up and Maintained Via SCInet (Routes & Addresses Through SCInet)
  - Mostly NFS and ISCSI Access To StorCloud Resources (No Logins Supported)
  - StorCloud Resources Preallocated to SCInet Addresses
    - Booths Notified of StorCloud Allocations & Addresses
  - 5 10GbE Channels Into StorCloud From SCInet
- Requested - Requested Networks by Specific Booths For Specific StorCloud Resources
  - Medium Specifically Requested & Installed By StorCloud
  - Maintained by StorCloud For Applications Use
  - Routed Only Within StorCloud (No World/Conference Access)
  - Statically Assigned Addresses
  - 2 10GbE Channels Into StorCloud From SNL
- Private – Non-Routed Management Network
  - For Equipment Setup, Configuration and Monitoring
  - Maintained by StorCloud
  - Only 10/100/1000 Ethernet Drops Supported
  - Statically Assigned Addresses for Equipment and Drops
  - Available Within StorCloud Booth & to Vendor Booth on IP Pair Basis

IP Switch Resources

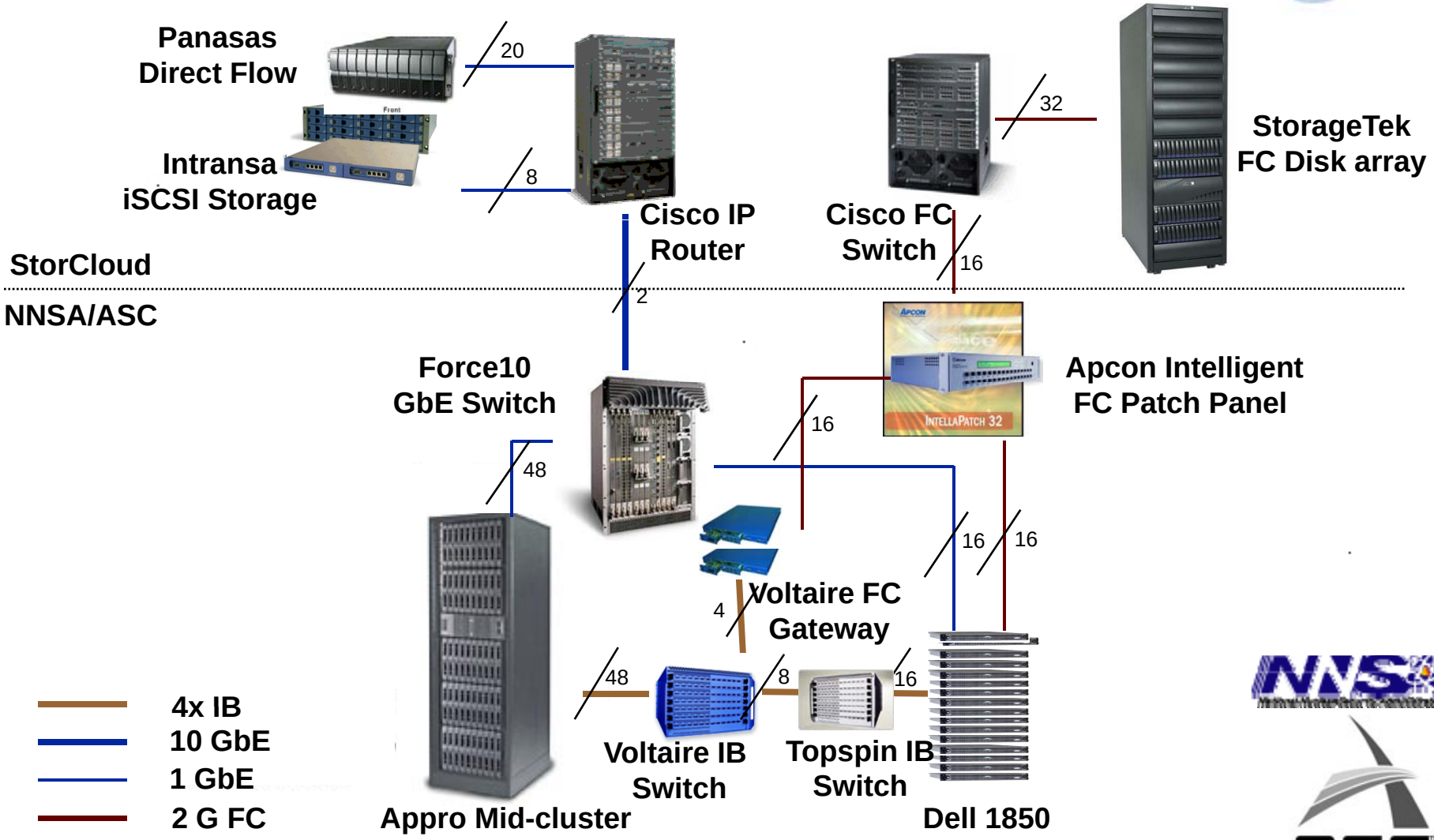
- All IP Access Via Cisco 6500 Switch
  - 7 10Gb Ports
    - 5 SCInet Controlled (Public Access)
    - 2 StorCloud Controlled (Private)
  - ~156 Ports 1Gb
    - 42 Private
    - 64 Internal Host Access
    - 8 Mgmt Network
    - 42 General Purposed NAS Ports
  - 7 VLANS
    - Public
    - Mgmt (3)
    - PNNL
    - SNL
    - MARCONI



IP Storage Resources

- NAS
  - 52-1GbE Ports
  - 62 TB Disk
  - Network Appliance
    - 8-1GbE FAS980 (General Purpose), 12 TB
    - 8-1GbE FAS980 (Dedicated DTSB), 12 TB
    - 8-1GbE SpinServer 4100 (FAS270, Dedicated NRL/Marconi), 16 TB
  - Panasas
    - 20-1GbE Interfaces (Dedicated)
    - 12 TB, 5 Trays SATA Disk
  - Intransa
    - 8-1GbE Interfaces (Dedicated)
    - 9 TB, 4 Trays FC Disk

IP/FC/IB Booth Demo Configuration



IP Storage Resources

- Pacific Northwest National Lab (PNNL) Lustre Cluster
  - 5 48U Racks
  - 41 Western Scientific FUSION-SA (Server Appliance)
  - Dual Opteron Node w ~ 10 TB SATA Disks
  - ~400 TB (400 GB SATA Disk)
  - Lustre File System Distributed Across 40 SA Nodes
  - 80-1GbE Internal (Private) Interfaces
  - 16-1GbE Public IP Interfaces
  - Integrated Distributed Application
    - Nodes Running Application and File System

IP Storage Access

- NetApp
  - NFS & CIFS (None)
  - ISCSI (FAS270, Private)
- Panasas
  - ISCSI (Private Network)
- Intransa
  - ISCSI (Private Network)
- IBRIX
  - NFS is Default Access Mode on 25 Shared Nodes
  - ISCSI Upon Request (Will Provide SW Upon Request)
- Lustre
  - Portals Library
- StorNext
  - NFS
- NFS
  - Public
    - 58 Research Booths w SCInet Connections
    - StorCloud Resources Pre-Allocated
  - Private
    - Dedicated Connections for Requesting Applications (DTSP, SpinServer)

Storage Management

- Veritas Command Central Storage Resource Manager (SRM)
  - Used Primarily to View and Monitor SAN
  - 2-Dedicated Hosts Provided Graphics and Traffic View
  - Storage Resources Managed By Switch SW
- Proprietary SRM (Zoning & Monitoring)
  - Brocade
  - McData
  - Cisco
  - 3Par

Internal StorCloud Projects

- Interoperability Fabric
  - Designed & Implemented by Vendors to Showoff Capability
  - Minimal Production Traffic Involved
  - Layer 2 Flat Fabric
    - Fabric consisting of Qlogic, Cisco and McData switches meshed together
  - Cisco IVR Routed Fabric
    - Fabric using Cisco's Inter VSAN Routing capability
  - McData IP Routed Fabric
    - Fabric using McDATA IP router appliance
- Performance Measurement/Display
  - SNMP Polling and Data Ingest/Storage (MySQL)
  - Graphical Presentation of MIB Data
  - Able to Acquire Large Amounts of Traffic for Post Analysis

Agenda

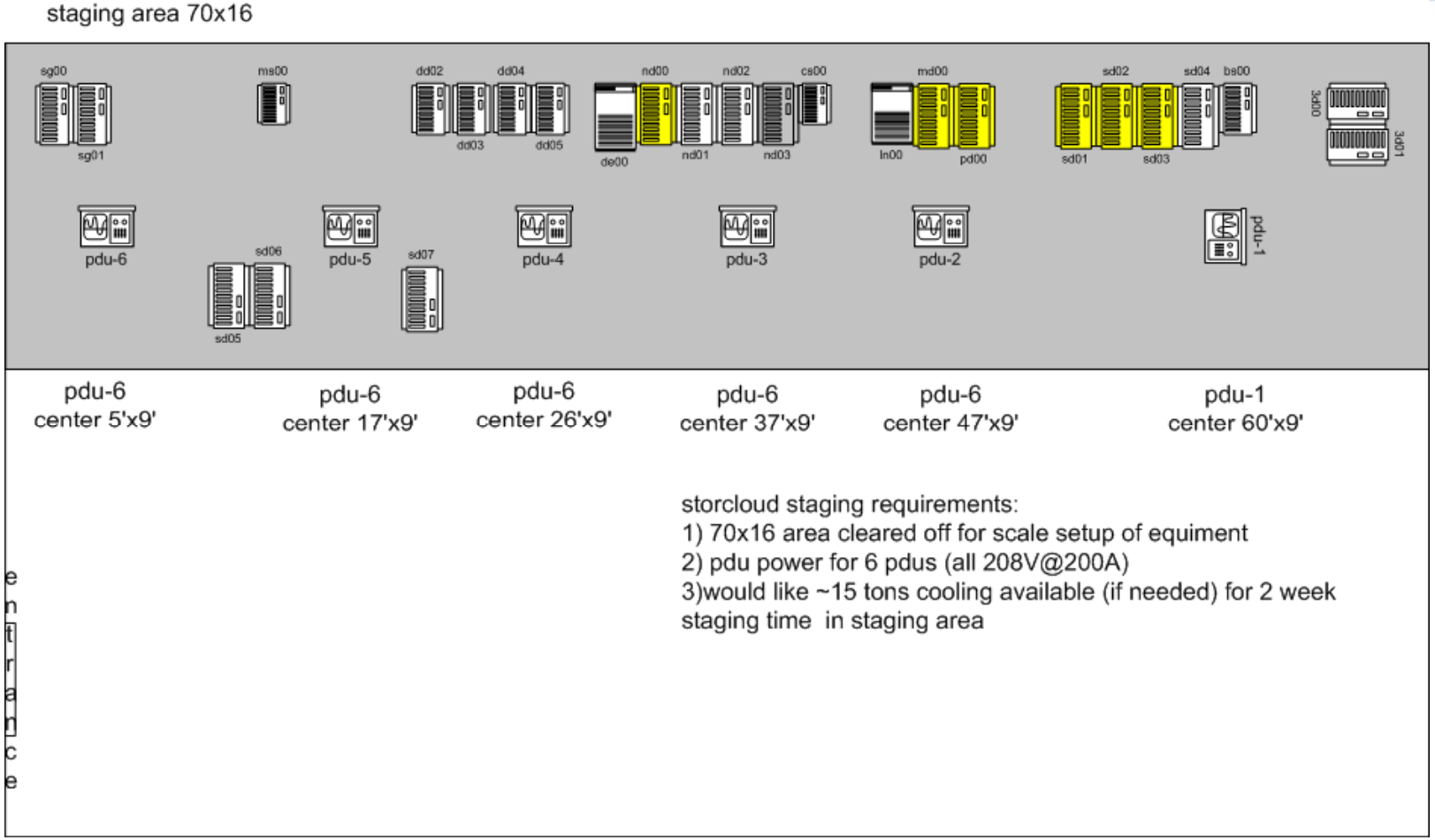
- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Logistics

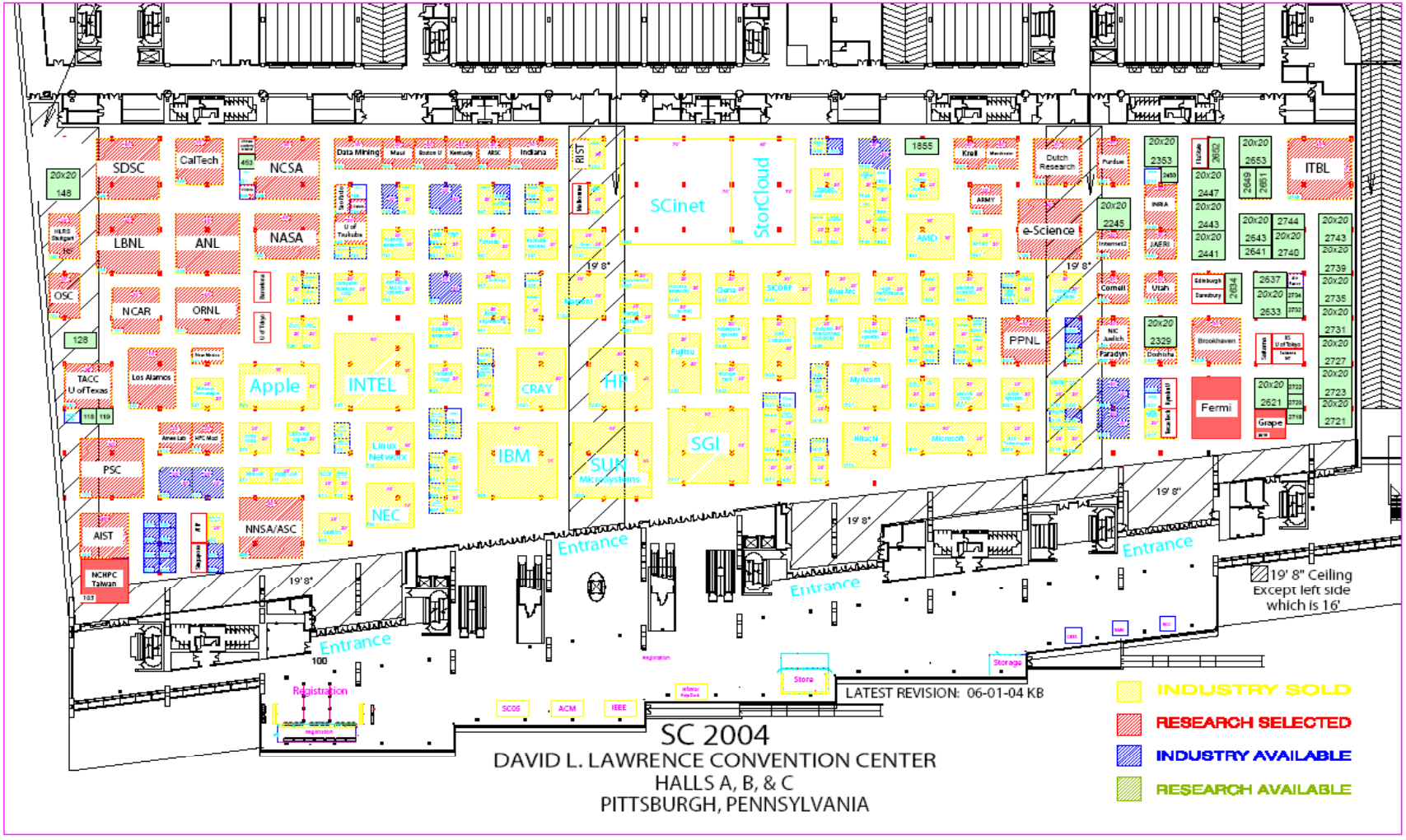
- Pre-Staging
  - Vendor/Applications Working Together Prior to Show at Application Site
  - Allow Applications to Tune Show Configuration
- Staging
  - Build & Configure As Much As Possible
  - Validate PDU and Booth Structure Configuration
  - Fibre Plant Configurations (Bundle and Install Multi-Fibre Cable)
  - Located in Pittsburgh DLLCC (Hall E)
  - Receive/Inventory/Track Equipment Prior to Show
- Booth Build
  - Coordinate with Show Manager (Booth, Labor, Shipping), DLLCC (Power, Cooling, Security), Vendors and Applications
  - Receive/Inventory/Track Pre-Staged Equipment Arriving on Trucks
  - Deal with Packing Materials
  - Move Staged Equipment to Show Floor from Staging Area
- TearDown
  - Coordinate with Show Manager (Booth, Labor, Shipping), DLLCC (Power, Cooling, Security), Vendors and Applications
  - Deal with Packing Materials
  - Help Vendors Pack and Ship Equipment
  - Pack and Ship StorCloud Resources

Pre-staging at Sandia

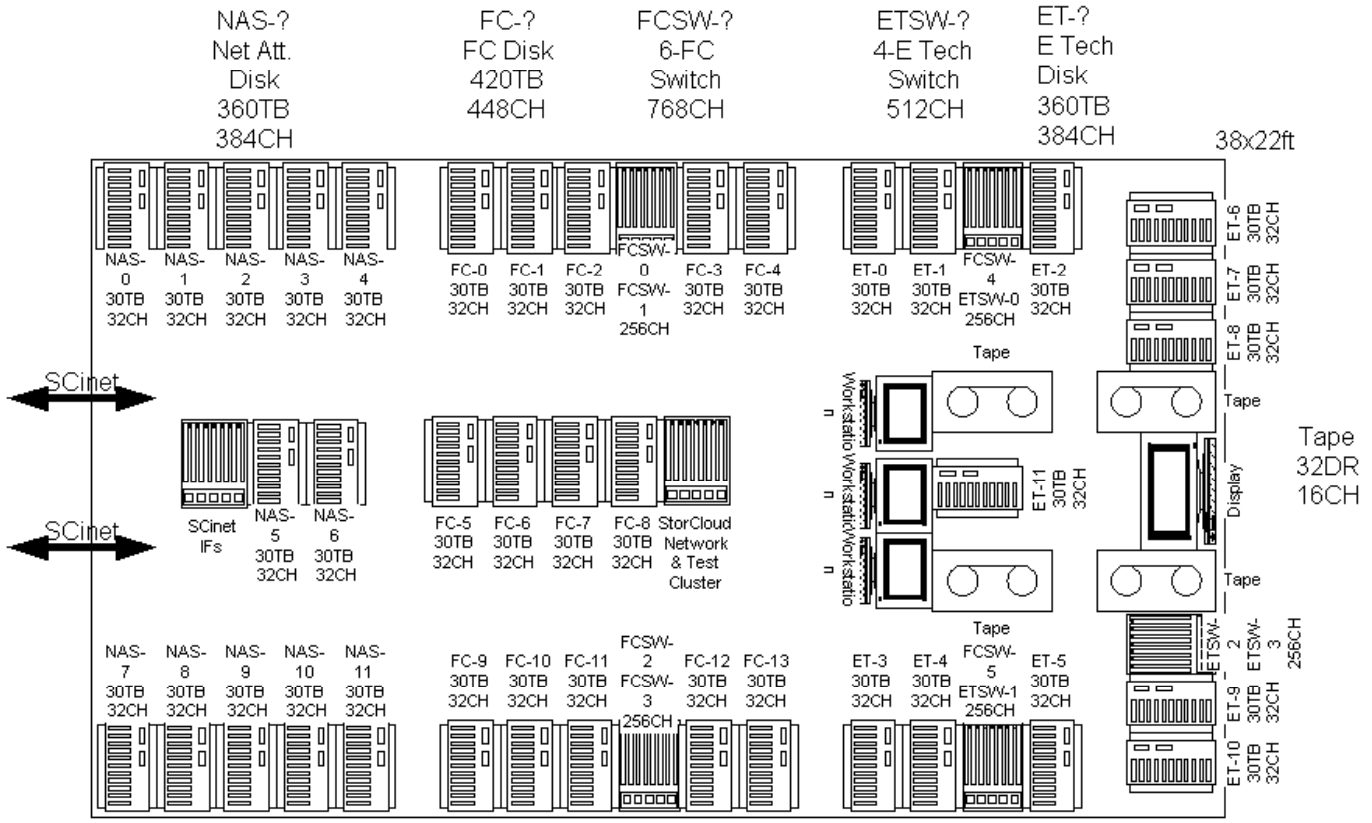
Staging Physical Layout



SC2004
Exhibits Floor Plan

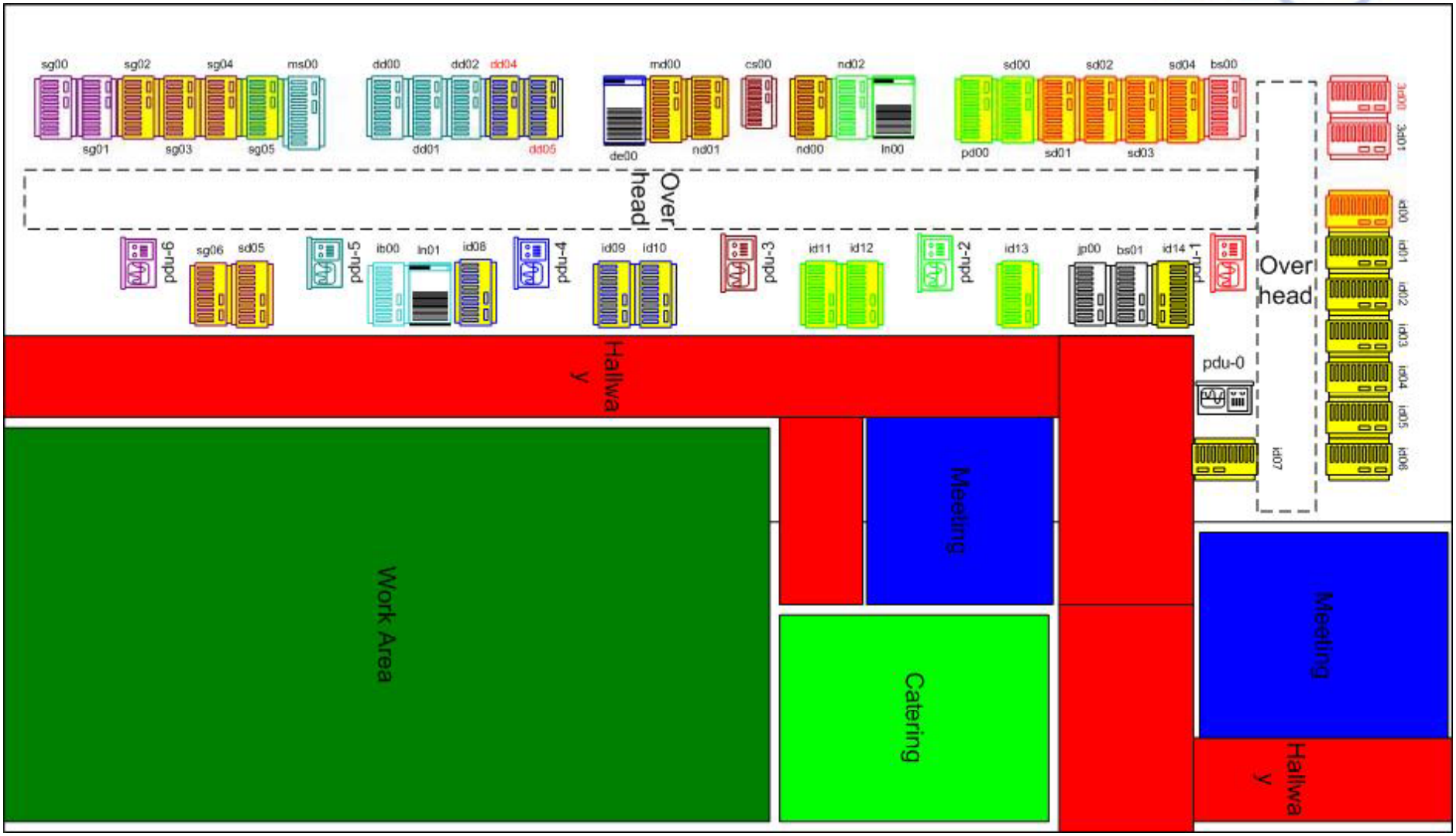


Physical Layout
6/5/03 (Iteration 3)



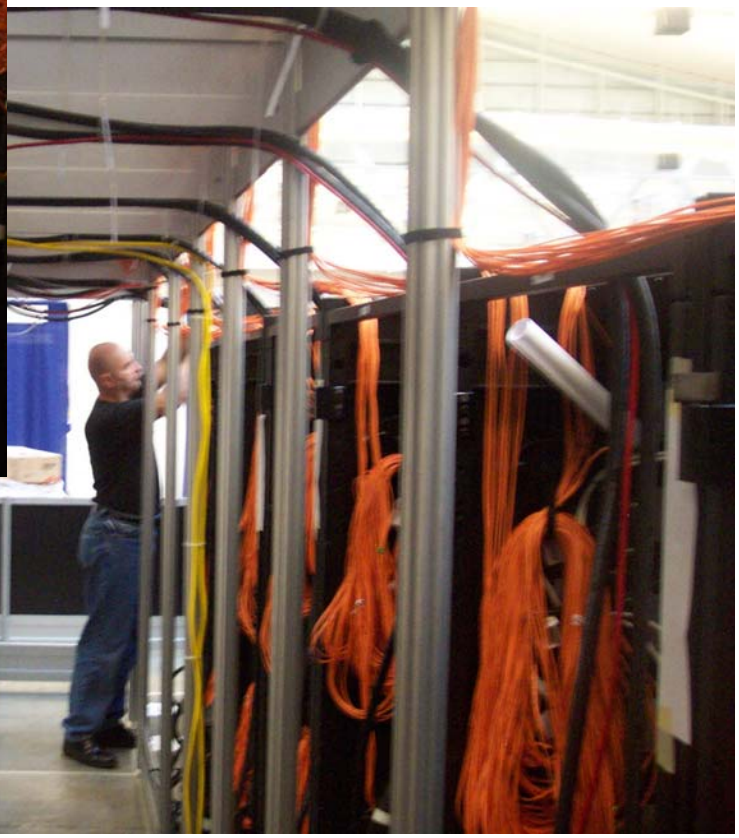
Totals:
Inbound CH: 974
Disk TB:1140
SQ FT:836
Racks:48

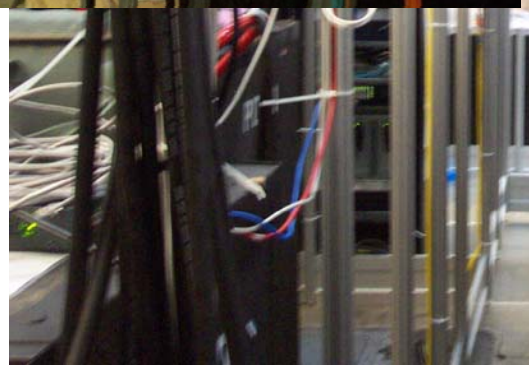
Physical Layout 10/29/04 (Iteration 997)



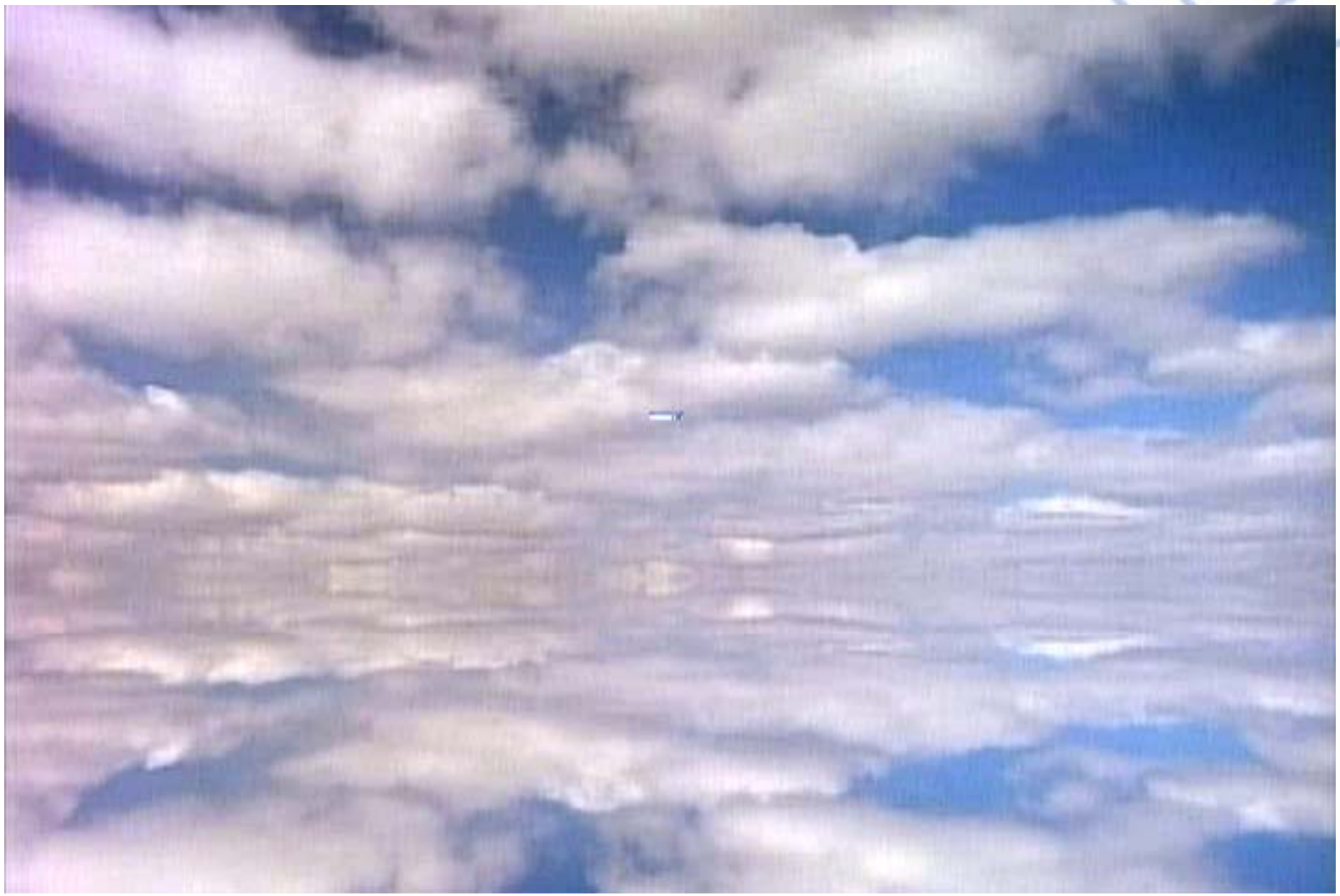
Booth Build - Overhead

Booth Build – In Deep





Booth Setup at Pittsburgh



Security

- Physical
  - Staging
    - Coordinate ($) with DLLCC Security (7x24 Access)
    - Access List Coordinated w All Vendors & Applications People
    - Dead Week (LOCKUP)
  - Booth
    - Shared Security with SCInet for DLLCC Security (7x24 Access)
    - Access List Coordinated w All Vendors & Applications People
    - Tours and Guests Allowed in Limited Numbers
- Electronic
  - Heavy Public Network Filtering To Privileged Management Network
  - No Public Network Login Accounts to File Servers
  - SCInet Tools and Monitoring

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Environmentals

- 49 cabinets
- ~850 Terabytes Disk Capacity
- 2.3 Terabits/sec Raw Bandwidth
- ~30 Tons of Equipment
- 45 Miles of Backbone Fiber
- 13 Miles of Jumper Fiber
  - ~600 2Gb FC Channels
  - ~188 1Gb Ethernet
  - ~7 10Gb Ethernet
- Consuming ~300 KW
- Generating ~800k BTU/hr

Power

- 7 Custom Built PDU(s)
  - Each 200A 220V Single Phase Feed From Building
  - 220V & 120V Circuits
  - Load Balanced Across All PDUs
  - 5 PDUs Required for Pittsburgh Staging
  - 2 Minor Modifications Prior To Show
- 133 Custom Built Power Drops
  - 4 Drops Rewired Prior To Show

StorCloud SC04 Power Distribution Unit End State

PDU	L6-30	L6-20	L5-30	L5-20	5-20	L14-30	Notes
0	16	4					
1	16			4			
2	8	6	5		2		
3	2	8		10			
4	8				10	2	
5	4	2			4	4	
6	6	8			4		(3 total circuits)
Totals:	**60**	**28**	**5**	**14**	**20**	**6**	**133**

Cooling

- 40 Commodity Box Fans
  - Mounted Overhead with Upward Trajectory
  - ~2000 CFM Each
  - ~70000 CFM Total
  - Rack Delivered, Installed & Customized by Show Manager
- Agreed with DLLCC to Keep Ambient ~70'F
  - Using 4 1500 Ton Chillers (One Overhead)
  - Generated ~70 Tons Heat
  - Chilled Us Well at Night
- Monitoring
  - 5 Standard Outdoor Thermometers
  - Student Volunteers Check Every Hour
  - 87'F Highest Recorded Temp
    - Then DLLCC Froze Us!

DLLCC
Building Characteristics

- Building Characteristics
  - ~290K Sq Ft (357'x830')
  - ~23M Cubic Ft
  - 34-125' Ceilings
  - 4 1500 Ton Chillers w Overhead Distribution
  - 480V 3-Phase Power
  - Windows & Curtains Used to Control Solar Heat
  - Fabric Roof (only one leak)
  - Great Staff

Cloudy on the outside

Hot and Humid Inside

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Application Guidelines

- In this first year our goal was to have a relatively small set of applications that exercise the broadest range possible of StorCloud's many capabilities
  - Fiber Channel, iSCSI, Lustre OSTs, NFS connections
  - Utilize storage from many contributing vendors
  - Local and WAN access
  - Multiple storage and switch vendors
- Real world applications from a variety of scientific disciplines
  - National labs
  - Universities
  - Industry
- StorCloud committee is sponsoring an application to test the interoperability of HPC storage and switch equipment from different vendors
- Four applications will compete in the StorCloud challenge

Challenge Categories

- **Most innovative use of storage**
  - Use of storage grids - the application that best demonstrates the use of storage in a grid environment
  - Most innovative use of storage - The application or infrastructure that shows the most ingenious, unique, novel way to use storage networks
  - The scientific merit or real world applicability of the application
- **Best use of storage bandwidth**
  - Best use of storage bandwidth - the application that best takes advantage of the increased bandwidth available in modern storage systems
  - Maximum use of Bandwidth - the application that is able to generate the fastest sustained transfer rate over a given period of time.
  - The scientific merit or real world applicability of the application
- **Best use of IOPs**
  - Maximum use of IOPs - the application that is able to generate the most I/O operations per second during a specified period of time.
  - The scientific merit or real world applicability of the application

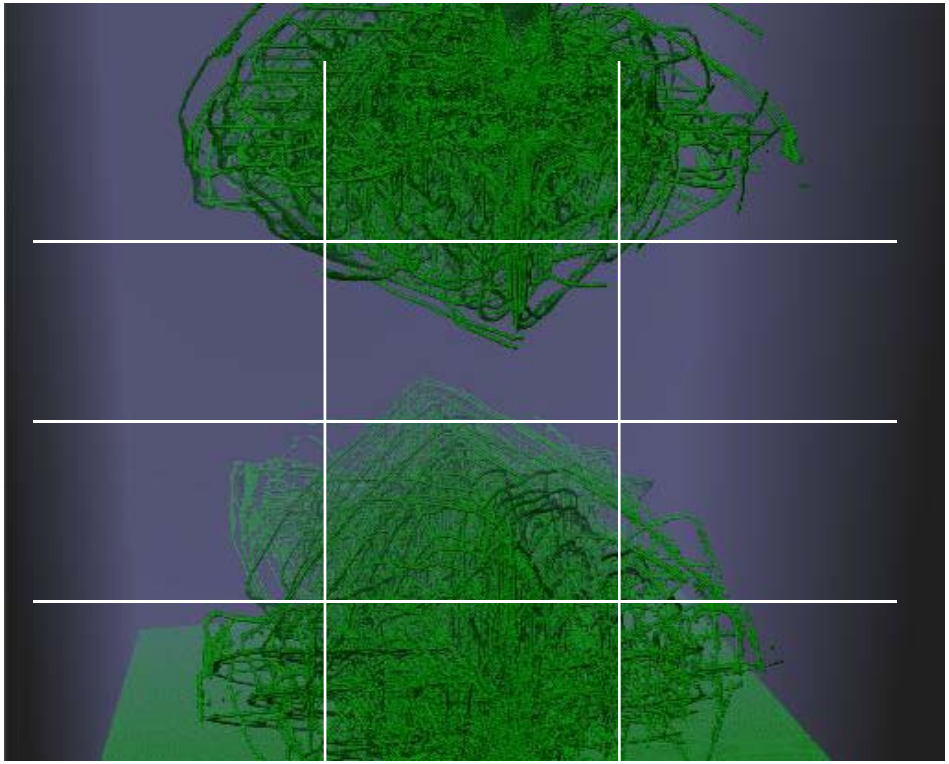
The Challenge Applications stretch StorCloud to its limits

Application	Booth (Booth #)	Storage Required (Vendor Preference)	Storage Bandwidth (does not include general IP traffic)	Connection	Switch
LLNL Blockbuster Holger Jones, Dave Bremer	NNSA/ASC (booth 403)	9 TB (Intransa) And 8 TB (STK – shared w/ SNL)	1 Gbps/node x 16 nodes total = 16 Gbps	2 x 10-GbE 16 x 2Gb FC	Force10 -10GbE (at NNSA booth) Cisco – FC, 10 GbE (StorCloud)
SNL MPQC Multi-pass (ABS) MP2-R12 Curtis Janssen, Joe Kenny	NNSA/ASC (booth 403)	8 TB (STK – shared with LLNL)	64 Gbps via I/O nodes and 32 Gbps via Voltaire gateway	2 x 10GbE 16 x 2Gb FC	Force10 -10GbE (at NNSA booth) Cisco – FC, 10 GbE (StorCloud)
PNNL PNNL Active Storage David Cowley, Kevin Regimbal	PNNL (booth 2027)	40 storage bricks built into a 400TB Lustre filesystem	2 Gb/s External 80 Gb/s Internal	2 x 1 GbE over SCInet 80 x 1 GbE Private	SCInet to StorCloud Cisco Switch
SDSC ENZO Phil Andrews	SDSC (booth 149)	90+ TB (IBM)	240 Gb/sec	120 x 2 Gb FC 3 x 10GbE	Brocade

The Blockbuster Demonstration

David Bremer and Holger Jones, LLNL

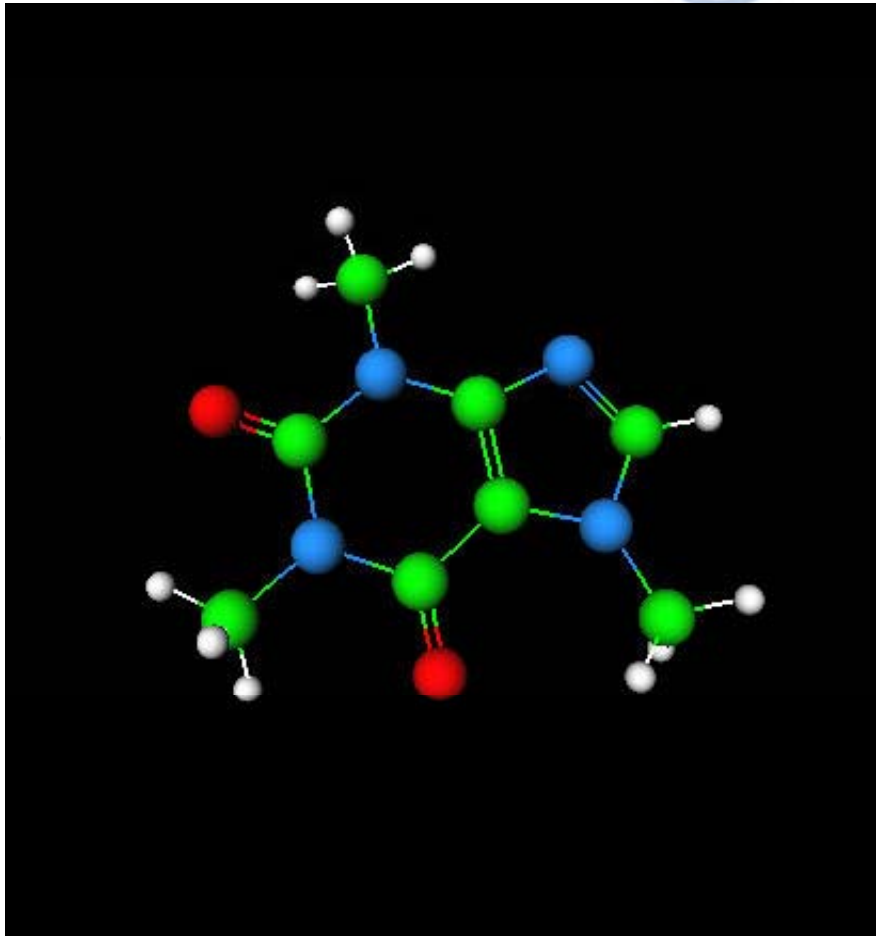
- Movie playing software developed by LLNL
- Scales to handle high resolution, high frame count movies
- Multiple nodes of a cluster read a portion of the tiles and render to tiled display in parallel
- Can be used by high-resolution capable movie player to leverage power walls



Massively Parallel Quantum Chemistry (MPQC)

Joseph P. Kenny and Curtis L. Janssen, SNL

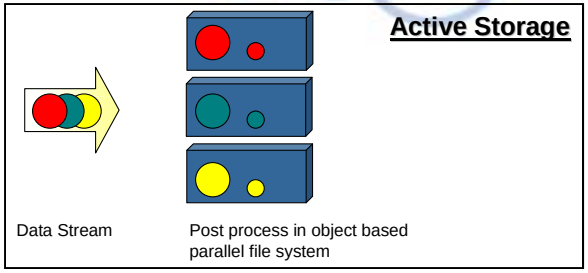
- Prototypical systems for study of molecular interactions
- High-performance parallel storage systems allow larger biological systems to be studied
- LAMPI and TerraGRID storage offer parallel I/O via InfiniBand Interconnect to TerraGRID I/O servers using StorageTek FC and Intransa iSCSI storage arrays



Applying Active Storage to Data-Intensive Proteomics Research

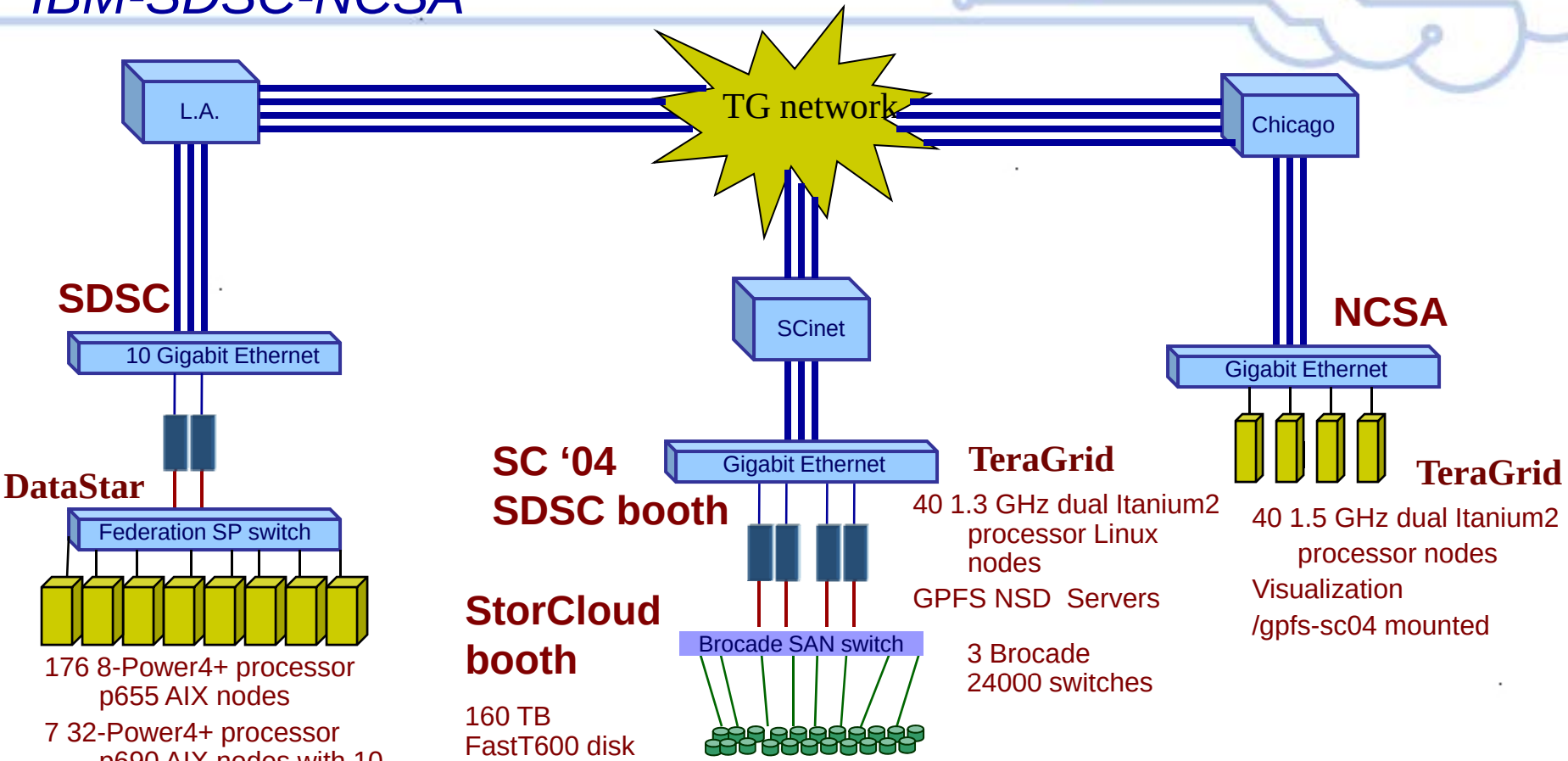
Kevin Regimbal, PNNL

- **Active Storage**
  - PNNL technology to exploit the otherwise under-utilized computational resources of Lustre OST servers
  - Lustre system with 40 OST servers each containing 24 400GB disks totaling 384 TB



- **Proteomics Scientific Application**
  - Goal is to build a reference table that could quickly identify a set of possible protein sequences based on experimental data
  - The application uses Active Storage technology to calculate and store all possible amino acid sequences
- **StorCloud Challenge**
  - a master application running on a workstation in the PNNL booth is writing files to the Lustre file system that contain a target mass and tolerance
  - The Active Storage module reads this information and calculates and stores the resulting protein mass and sequence data to a 2nd file within the file system.
  - Active storage enables very little data to cross the network.
  - Results monitored from PNNLs booth using a 3-D cluster visualization application

ENZO Computation
IBM-SDSC-NCSA



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

Challenges Awarded

- Best Random Access Use of StorCloud
  - Blockbuster
- Best Use of StorCloud to Advance a Scientific Application
  - MPQC
- Most Innovative Use of StorCloud
  - PNNL Active Storage
- Highest Achieved StorCloud Bandwidth
  - ENZO (15GB/s Achieved)
  - Also, entered in SCInet Bandwidth Challenge

Non-challenge Apps exercise other dimensions of Storage

Application	Booth (Booth #)	Storage Required (Vendor Preference)	Storage Bandwidth (does not include general IP traffic)	Connection	Switch / Cluster
DSTP Web Service Michal Sabala	National Center for Data Mining (Booth # 653)	12 TB (NetApp)	400 MB/s	2 x 10GbE (Public) to 4 x 1GbE	Cisco
NRL/Marconi/NetApp High Performance NFS Global Name Space Eric Melvin, Mark Buczynski, John Rutemiller	Marconi (booth 1132) NetApp (booth 1854)	10 TB at NRL (NetApp) 10TB StorCloud (NetApp)	OC192 WAN 4 x GbE	4 x GbE	SCInet provides direct drop to NRL (for WAN) Cisco (local)
Interoperability App Cary Whitney	StorCloud Booth	1 TB (STK tape maybe) MultiVendor (1 shelf per)	Negligible	N/A	Cisco, Qlogic, McData

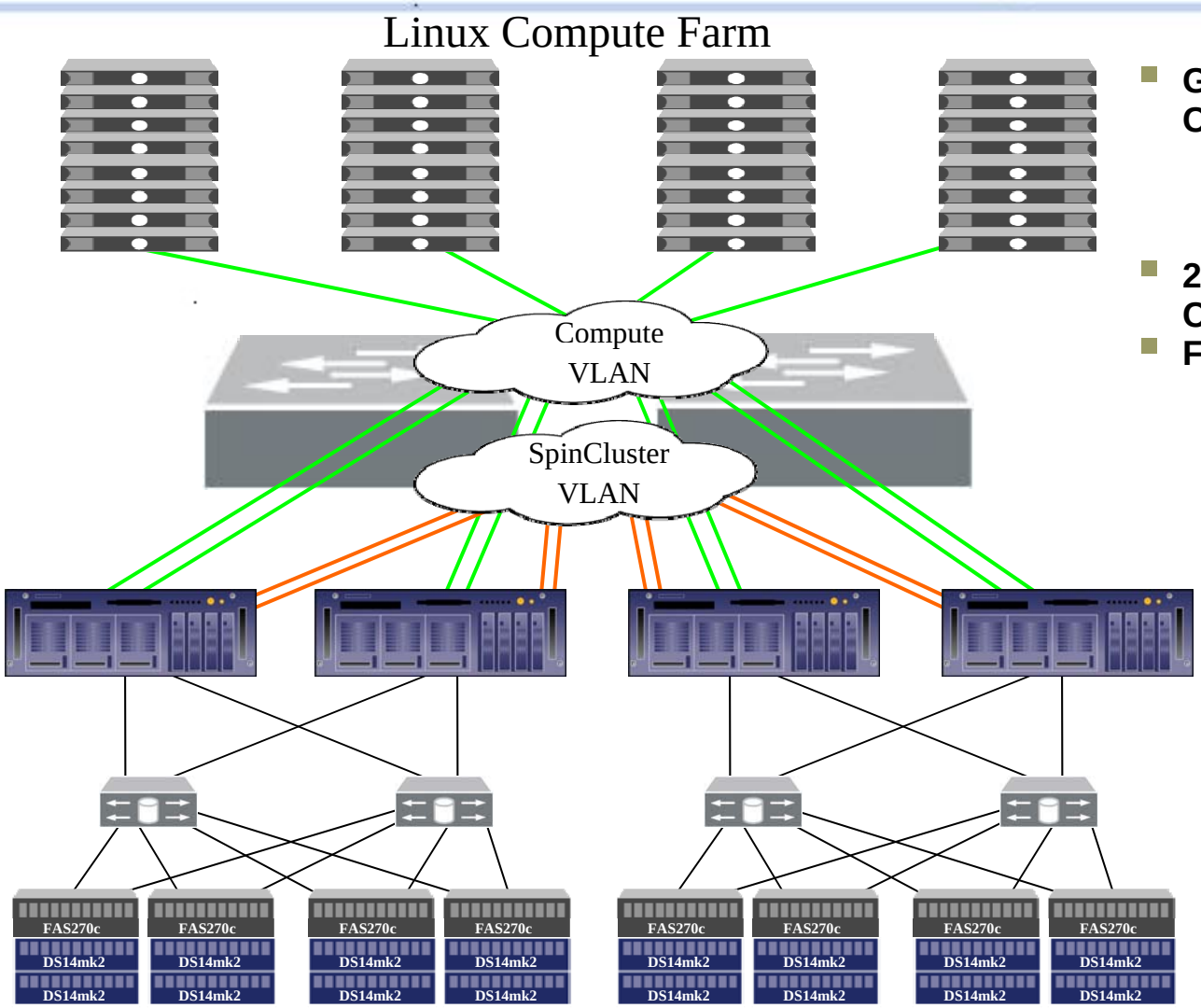
Non-challenge Apps exercise other dimensions of Storage

Application	Booth (Booth #)	Storage Required (Vendor Preference)	Storage Bandwidth (does not include general IP traffic)	Connection	Switch / Cluster
TeraShake: An Earthquake Storage Intensive Simulation Marcio Faerman	SDSC – Southern California Earthquake Center (booth 149)	60 TB (STK)	~500 MB/s	1 GbE over SCInet	McData
Image Analysis and Visualization of Very Large Biomedical and Engineering Datasets Leslie Southern, Pete Wyckoff	OSC booth	40 TB (DDN)	16 Gb/s	1 GbE over SCInet (Compute Nodes in StorCloud Booth)	McData
S3D – Direct Numerical Simulation (DNS) of Combustion Phenomena Philipe PeBay	NNSA/ASC (booth 403)	12 TB (Panasas)	20 1GbE to Panasas (internal to StorCloud)	2 x 10GbE	Force10 - 10GbE

DSTP (Michal Sabala)

- DataSpace is an infrastructure for creating a web of data instead of documents. DataSpace is designed to provide an infrastructure for the analysis and mining of business, e-business, scientific, engineering, and health care data.
- Data is accessed in DataSpace using a protocol called the dataspace transfer protocol (DSTP). DSTP web service requires fast sequential reads and, will use high performance NFS service from the StorCloud. This web service will request an I/O bandwidth of 400 MB/s using four Gigabit Ethernet connections to the StorCloud booth.
- Demonstrate on StorCloud proper operation of DSTP with datasets of multi terabyte size.

Typical SpinServer Deployment

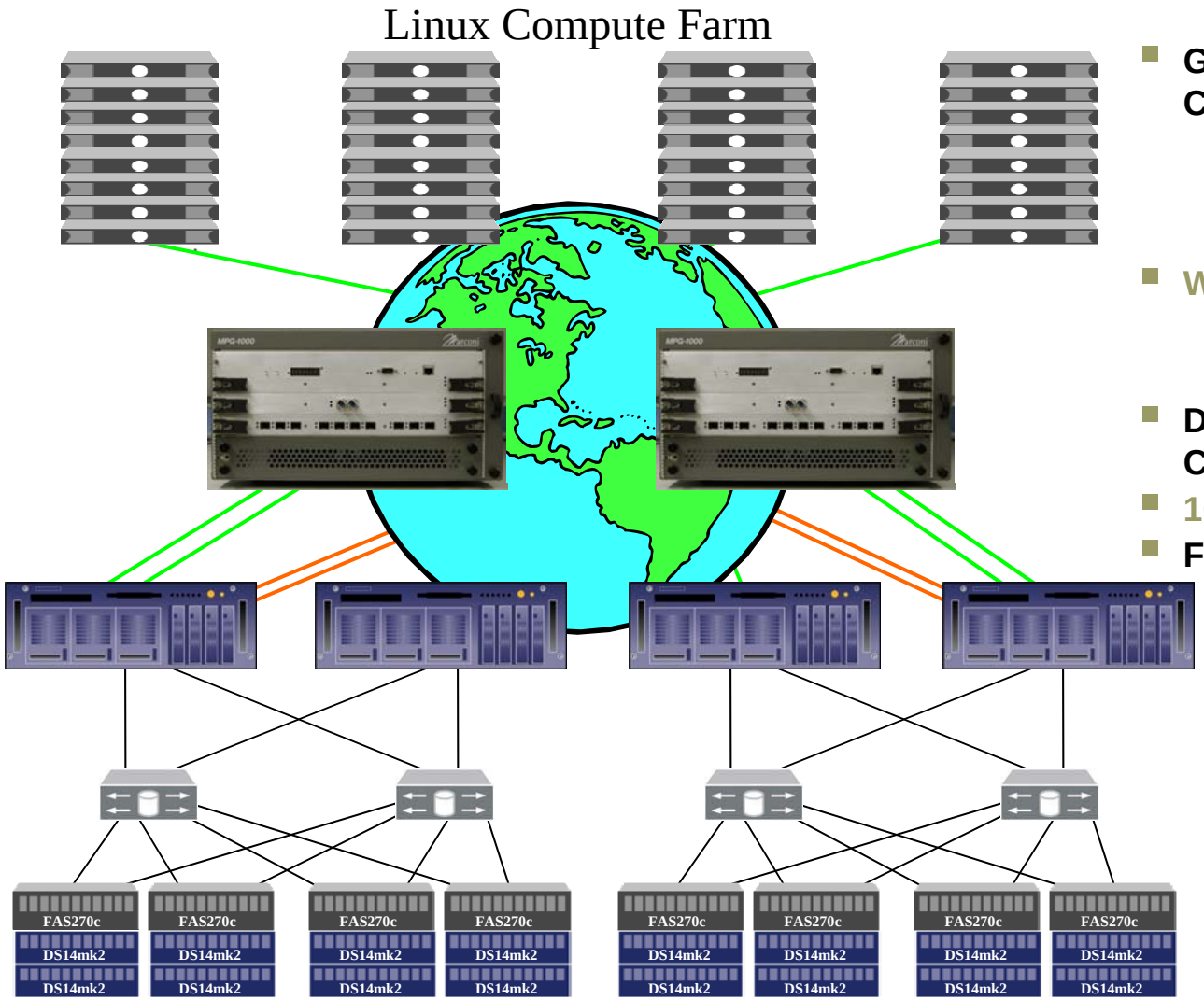


- **Global Namespace Storage Clustering**
  - **Up to 512 SpinServers per Cluster**
  - **Up to 12TB per SpinServer**
- **2x Gigabit Ethernet Connections**
- **Fibre Channel Interconnects**
  - **2x FAS270c**
  - **4x DS14mk2**

Marconi

NetApp

WAN SpinServer Deployments with the Marconi MPG-1000



- **Global Namespace Storage Clustering**
  - **Up to 512 SpinServers per Cluster**
  - **Up to 12TB per SpinServer**
- **Wide Area Cluster Extension**
  - **Up to 10Gbps**
  - **Type 1 Encryptable**
- **Dual Gigabit Ethernet Connections**
- **10 Gigabit WAN Interconnect**
- **FibreChannel Interconnects**
  - **2x FAS270c**
  - **4x DS14mk2**

NetApp®

Marconi™

TeraShake (M. Faerman)

- Magnitude 7.7 earthquake on southern San Andreas
- Mesh of 1.8 Billion cubes, 200 m
- 0.011 sec time step, 20,000 time steps: 3 minute simulation
- Kinematic source from Cajon Creek to Bombay Beach
  - 75 sec source duration
  - 18,886 point sources
- 240 processors on SDSC DataStar
- ~20,000 CPU hours, ~5 days wall clock
- ~47M Mbytes (TB, 30M floppies) generated
- StorCloud Demo
  - Seamless integration of StorCloud resources into the TeraGrid Storage Resource Broker Data Grid
  - Replication of data onto StorCloud
  - Demonstrate Remote Access of StorCloud files
  - Real Time Visualization

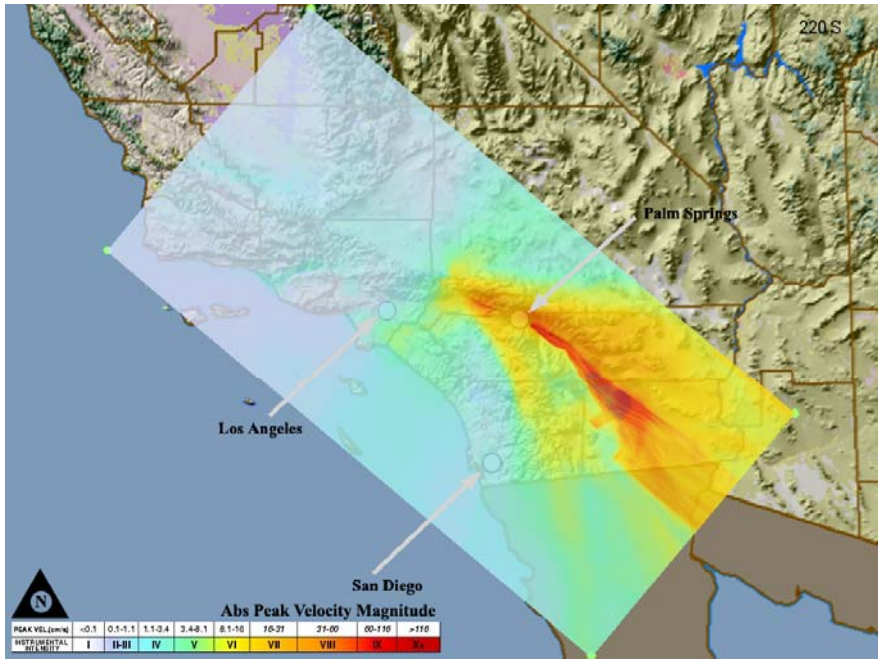


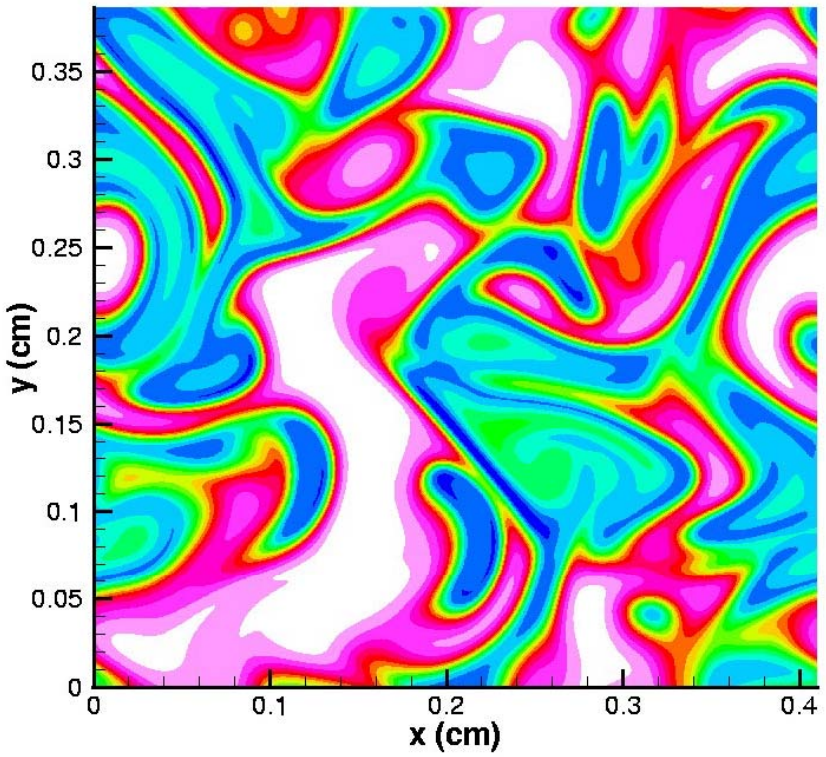
Image Analysis and Visualization of Very Large Biomedical and Engineering Datasets (P. Wycoff)

- Vast volumes of image data are generated from advanced imaging systems (e.g., high-resolution digitizing microscopes and MRI). Effective use of rich image information in biomedical research requires support for distributed storage and image processing.
- Showcase a suite of middleware tools that are designed to support storage, retrieval, and processing data for image analysis and visualization of very large multi-dimensional datasets.
- Demonstrates execution of spatio-temporal queries into large-scale, on-line databases of images and time-dependent 3D volumes on disk-based storage clusters and execution of image and visualization workflows on distributed collections of datasets.

S3D, A Direct Numerical Simulation (DNS)

Philippe P´ebay, Jackie Chen, Helen Chen, and E. R. Hawkes, SNL

- S3D is a compressible Navier-Stokes solver developed at Sandia
- We demonstrate the acceleration of high fidelity, HPPC simulation using the Panasas Direct Flow parallel I/O
  - Avoid potential loss of critical data
  - Simplify post data analysis
  - Allows real-time code-steering



Sandia National Laboratories

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Timeline 2004

- **January 25-26**: Executive Committee Meeting
- **March 9-10**: Meeting with vendors, define architecture, Baltimore, MD
- **April 20-23**: Executive Committee, San Francisco, CA
- **May 16-17**: All Hands Meeting, San Diego, CA
- **June 1-3 2004**: SCinet/StorCloud integration architecture meeting at Argonne National Labs, Chicago
- **August 3-4 2004**: With Executive Committee and SCInet, Seattle, WA
- **October 18-22**: SCinet/StorCloud Staging, Pittsburgh Convention Center
- **November 1-7**: Conference Setup
- **November 7-11**: SC2004
- **November 11-13**: Teardown/Ship
- **November 14-25**: Sleep

Timeline Staging

DATE	TASKS	WHO
Oct 15	- Layout Booth w/ Tape in Staging Area -Equipment Delivery -Bundle & Label Fibre	- SC - SC - SC
Oct 18	- Layout Booth w/ Tape in Staging Area - Setup Ladder Rack in Staging Area - Equipment Delivery - Move PDUs into place & Power On	- SC - Freeman - Freeman - SC, PCC
Oct 19	- Verify Power Connections - Set up management Network - Power on Racks - Set up Minimal FC Network	- Vendors - SC - Vendors - SC
Oct 20	- Create LUNS on Disk Arrays - Zone Switches - Verify/Rebuild Server OS - Ensure Restricted Access on Servers - Interoperability Install - SC DINNER (1930:Mark's Bar & Grill (next to Courtyard), plz sign up)	- Vendors - Vendors - SC - SC - Vendors -All
Oct 21	- Install File Systems on Servers - Test File Systems w Minimal FC access - Install/Acquire SRM Configuration - Ensure SRM/Performance Display - Interoperability Testing	- Vendors - Vendors, SC - Vendors - SC, Vendors - Vendors, SC
Oct 22	- Dismantle Mgmt Network and Fibre Runs - Power off equipment - Take Down Ladder Rack	- SC - Vendors - Freeman

Timeline Setup

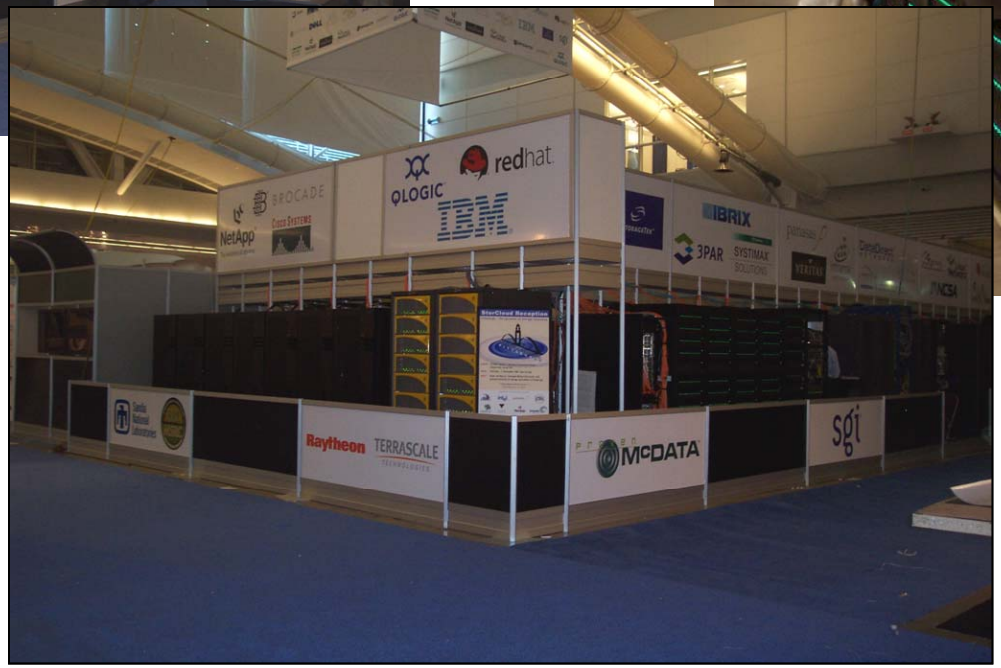
DATE	TASKS	WHO
Nov 1 Monday	- Booth & Overhead Assembly - Power To Booth -Prepare Staged Equipment to Movement	- Freeman - PCC - SC/Vendors
Nov 2	- Equipment Moved From Staging Area - Equipment Delivered From Warehouse - Power Up	- Freeman/Teamsters - Freeman/Teamsters - PCC/SC/Vendors
Nov 3	- FC & Network Setup - Equipment Configuration - Server Configuration	- SC - Vendors - SC/Vendors
Nov 4	- FC & Network Setup - Equipment Configuration - Server Configuration - Application & Storage Integration	- SC - Vendors - SC - SC, Apps, Vendors
Nov 5	- Equipment Configuration - SCInet Hookup - Server Configuration - Application & Storage Integration - Trouble Shooting	- Vendors - SC, SCInet - SC - SC, Apps, Vendors - SC, Apps, Vendors
Nov 6	- Application & Storage Integration - Trouble Shooting	- SC, Apps, Vendors - SC, Apps, Vendors
Nov 7	- Application & Storage Integration - Trouble Shooting	- SC, Apps, Vendors - SC, Apps, Vendors
Nov 8 Monday)	- Application & Storage Integration - Trouble Shooting - SHOWTIME (1800-2100)	- SC, Apps, Vendors - SC, Apps, Vendors - Conference

The Calm before the Storm

Timeline SC2004

DATE	TASKS	WHO
Nov 8 (Monday)	- SHOWTIME (1800-2100) Gala Opening	- Conference
Nov 9	- SHOWTIME (1000-2000) - Application Tuning - Problem Resolution	- Conference - SC, Apps, Vendors - SC, Vendors
Nov 10	- SHOWTIME (1000-2000) - Application Tuning - Bandwidth Challenge (1600, 1700, 1800, 1900) - Problem Resolution	- Conference - SC, Apps, Vendors - SC, Apps, Vendors - SC, Vendors
Nov 11	- SHOWTIME (1000-1600) - MaxBandwidth (1200) - ENDOFSHOW (1600) - Coil Backbone Fibre (1600) - Equipment Teardown and Ship	- Conference - Conference - Conference - Freeman - Freeman, Vendors
Nov 12	- Equipment Teardown and Ship (by 1200) - Power Down - FC & Network Recovery - SC05 StorCloud Meeting (0900-1100) - Conference Awards Ceremony	- Freeman, Vendors - PCC - SC -All - All
November 13	- Finish Packing/Shipping - Go Home	- All

Opening Night



Communications/Public Relations

- Public Website
  - http://www.sc-conference.org/sc2004/storcloud.html
  - http://sc05.supercomputing.org/initiatives/storcloud.php
- Show recognition
  - Signage on contributed hardware and StorCloud booth
  - Vendor recognition on SC website

Press

- HPCWire
- GridToday
- Supercomputing Online
- Access Online
- Business 2.0
- CNET.com
- InfoWorld
- Associated Press
- Business Week
- Computer World
- New York Times
- PC World
- Reuters
- Wall Street Journal
- Post News Week
- Computerworld.com
- Bio-itworld.com
- Forbes
- Linux Magazine

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Budget

	SC2004 Budget	Notes:	SC2004 Projection
Meals & Refreshments	10,040	3/04,04/04Staging,8/04,10/04 CtteMtgs	2,000
Convention Center Catering	16,000	Per Elvira2/4/04-20 ctte brkfst&lunch@12days-11/17/03Staging&Conf	6,000
Committee Support	1,500	Shirts, gifts etc for 50 volunteers	1,669
Facilities & Equipment Rental-Computer	0	Sandia,ARL,others donating equipmnt	3,200
Facilities & Equipment Rental	22,500	Capital SANS items(switches, wiring &hdwe)and Furniture rental	0
Facilities & Equipment Rental-Fiber	45,000	Capital NetwrkItems(Fiber,FiberSupplies)	0
Facilities & Equipment Rental-Networking			1,000
Electric Power	18,400	Fcst:Chrg for circuits&service	28,762
CC Telecommunications Support			
Network Drops			
Equipment Purchases		Fcst:Need fans,cable trays,fiber,PDUs can be reused	111,000
Telephone & Office Supplies	1,200		0
Postage & Shipping	1,000	Vendors to return own stuff	5,722
Printing & Copying	1,000	do own printing	0
StorCloud Challenge		Awards, brochures & publicity	2,000
Service Contractor Fees			7,000
Other (Misc)	5,000	Contingency	5,000
Grants (from new vendors)	0	$25K fr DOE&$25K fr DARPA	
Grants (from sponsors)	0	$2K Awards Sponsor+$61.2K New Booth Sales	0
	121,640	Note:LostInc(50x30ExhSpace-$51K,+10VendorRegcomps-$5K)	**172,353**

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

Lessons Learned - Staging

- Too much time was spent tracking missing or incomplete shipments.
  - Need complete set of vendor POCs for shipment tracking prior to departing for staging.
  - Need tracking information for ALL items prior to departing for staging, so that we don't need to contact those in (1) above.
  - Need vendors to commit to shipping the items for staging.
- Require vendor provide repair POCs for all equipment shipped for StorCloud team integration (i.e. Dell, LNXi)
- Purchase a full set of tools for setup and teardown. Spent too much time hunting and borrowing tools
- Ensure we have extra electrical parts/pieces and rack space to handle the inevitable extra gear that shows up
- Have available extra FC HBAs
- Have available extra 120 VAC building power for unplanned piece parts that arrive.
- Make sure all vendors/participants understand that they are responsible for packing and shipping their gear back

Lessons Learned – Setup

- Pre-Build Booth: Booth itself was built while we were building StorCloud
- Engage Research Booths Earlier To Ensure Adequate Resource Utilization
- Require Privileged Access to All Equipment for Committee Member(s)
- Require Primary & Backup Individuals on All Equipment to Provide 7x24 (Better Scheduling)
- Provide Dedicated Nodes for Data Capture and Display
- Too Many IP Networks to Deal With, Let SCInet Provide ALL External IP Access
- Build More Flexible Architecture Because Some Equipment Will Not Arrive

Lessons Learned - Applications

- Write a vendor “Terms for Participation” to clearly outline the expectations from the vendor. All participating vendors will be required to acknowledge they understand the agreement.
  - Participation in Staging (Pre-staging) by all vendors is absolutely necessary.
  - Shipping to/from convention center is vendor’s responsibility
- Develop/find a means to storage and transfer large datasets to & from the users site and convention center
- Finalizing the StorCloud architecture must be accomplished by early July.
- Find more large scale StorCloud applications to experiment with and showcase the storage resources

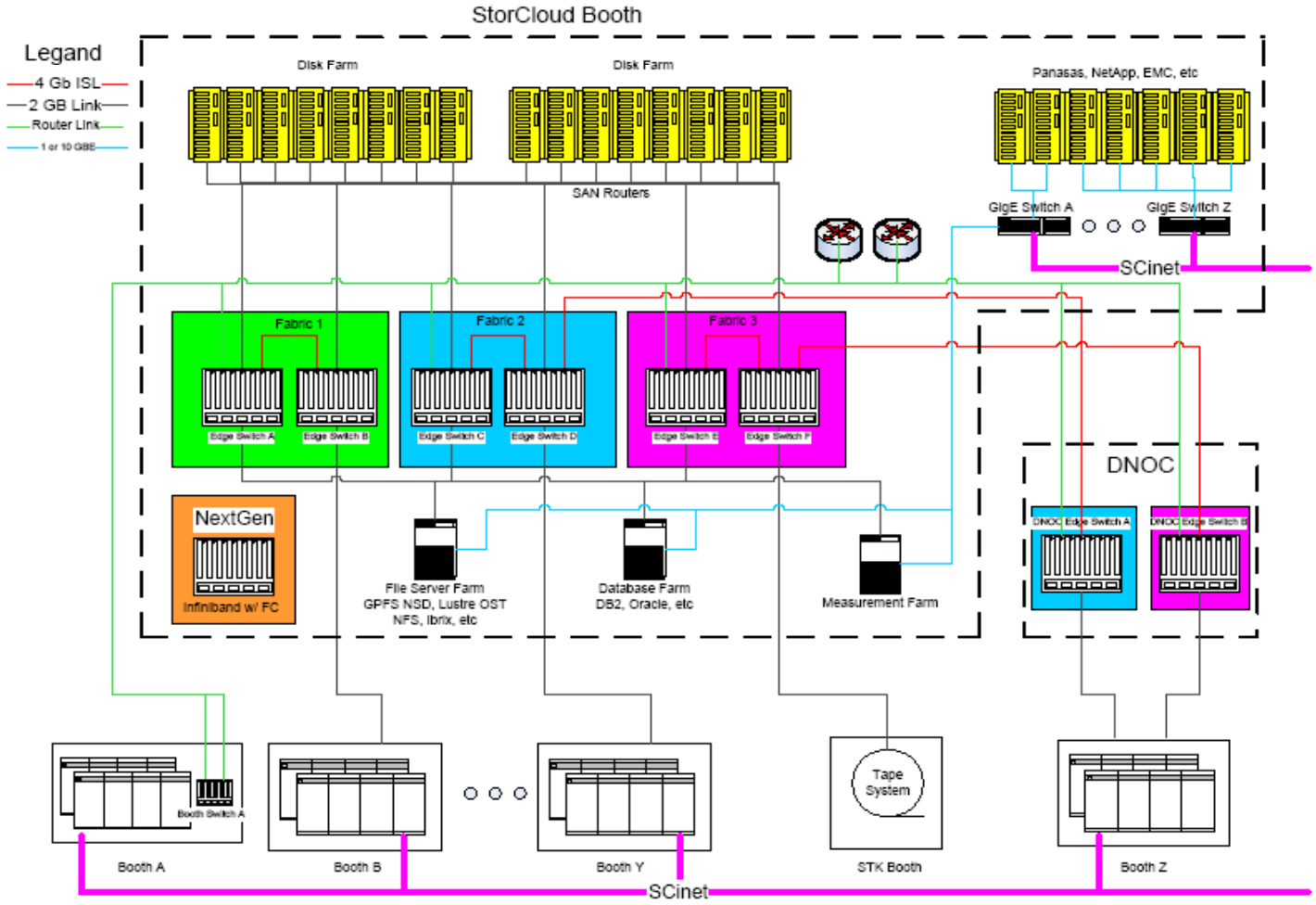
Communicate, communicate, communicate….

- Continue to support a SCInet Liaison
- Assign a primary POC to facilitate user requirements with their assigned vendors
- Develop and implement database to track vendor information such as 1) POC information; 2) hardware or software contributions; 3) graphics/logos
- Develop on-line submission process for applications
- Need complete set of information for vendor shipment (POCs, equipment list, tracking info, delivery dates, # crates, etc)
- Assign a primary system administrator with assistants
- Order extra shirts!!

Agenda

- Background
- Storage Architecture
- Physical Attributes/Logistics/Security
- Environmental Attributes
- Application Profiles
- Schedule/Communication
- Budget
- Lessons Learned
- StorCloud 05?
- Questions

SC05 Logical Diagram (Iteration 3)



SC05 Objectives

- The goals of “production” StorCloud are to:
  - Provide 1 PetaByte multi-vendor rotating storage
  - Provide A 1 PB File System (Name Space)
  - Approach 1 TeraByte/sec bandwidth to/from storage
  - Provide 1 GigaByte/sec backup capability
- Manage and allocate resources to SC2005 participants
- Raise the storage intensive applications performance bar

- New to SC05, there will be a NextGen StorCloud that will focus on one emerging technology: IB

SC'05 Timeline

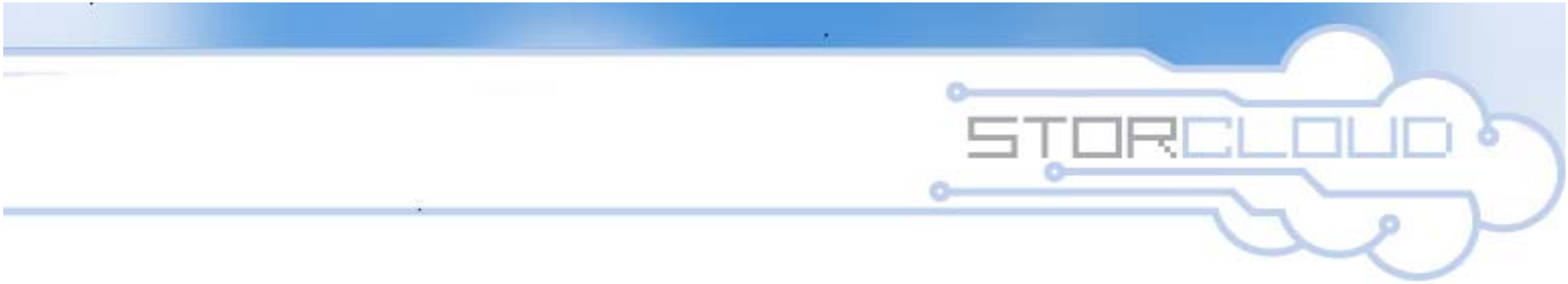
- May 23 2005: Full abstracts for applications are due.
- May 23 2005: Notice of intent to participate in StorCloud Challenge due
- June 6-10 (1 day), Seattle WA: All hands meeting, finalize architecture. Assign applications to HW.
- June 10 2005: Install tape library at ?? (SDSC, PNNL)
- July 1 2005: start loading application data into tape library
- July 11 2005: Order StorCloud Fiber and other equip.
  - Two month lead time for 50 micron bundles
- August 1 2005: Design StorCloud booth
- August 1 2005: Final exact equipment list and commitment
- August 1, 2 2005: SDSC All hands meeting
- September 15 2005: Insurance submission
- October 27- Nov 5, 2005, Seattle, WA: StorCloud Staging
  - Ship to arrive no later than October 24th
- November 7 - 19, 2005, Seattle, WA: Conference Setup and Show

Contact Information

- http://sc05.supercomputing.org/initiatives/storcloud.php
  - White Paper, Flyers, Presentations, Documents
- E-mail lists:
  - StorCloud team: StorCloud@sc05.org
  - Email archive is available for viewing past discussion:

Thanks To:

- IEEE & ACM
- SC'04 Executive Committee
- HW/SW Vendors & Families
- Volunteers & Families
- Volunteer Organizations
- Press
- SC'04 Attendees



Questions?

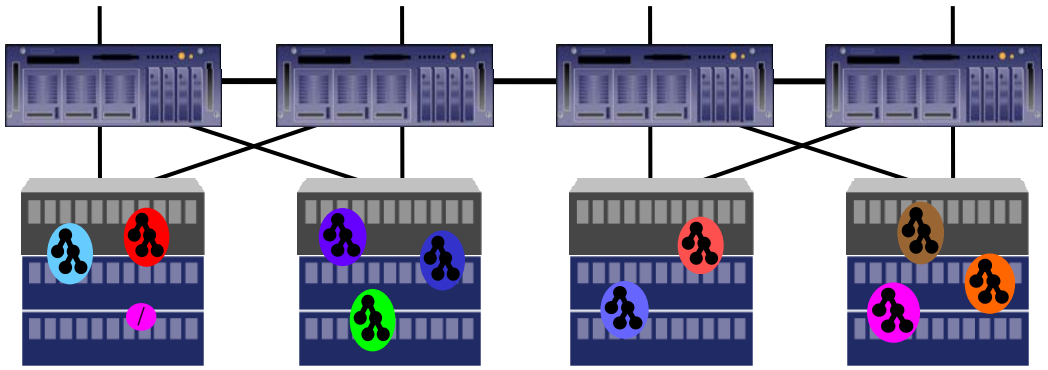
Non-challenge Apps will exercise other dimensions of StorCloud

- DSTP Web Services (National Center for Data Mining)
  - fast sequential reads that will use a high performance NFS service from the StorCloud
  - I/O bandwidth of 500 MB/s using four Gigabit Ethernet connections to the StorCloud booth
- TeraShake: an Earthquake Storage Intensive Simulation (Southern California Earthquake Center)
  - models a large earthquake occurring on the southern San Andreas Fault in California
- Image Analysis and Visualization of Very Large Biomedical and Engineering Datasets (OSC)

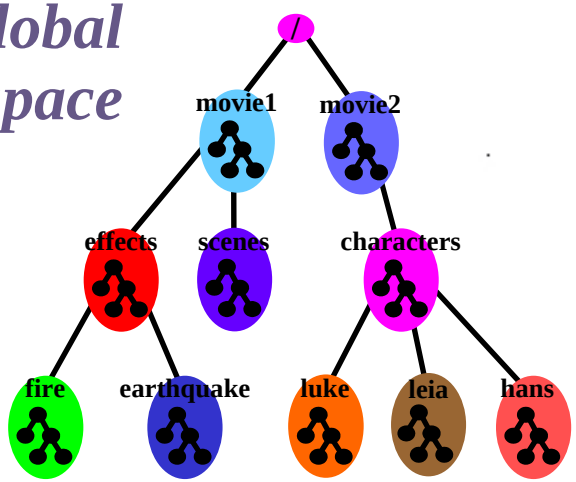
High performance NFS Global Namespace (SpinFS)

- Better manageability for large, growing, and changing data sets
  - Data spread across multiple SpinServers to achieve required performance and capacity
  - Data organized in a single logical file system for manageability
- All Linux clients in compute farm can see all the data
  - No need to reprogram Linux clients as SpinServers and data added
- Single NFS exported file system provided
  - Made up of multiple VFSs (virtual file systems)
  - File systems to multiple petabyte built by linking VFSs at mount points

SpinServer Cluster

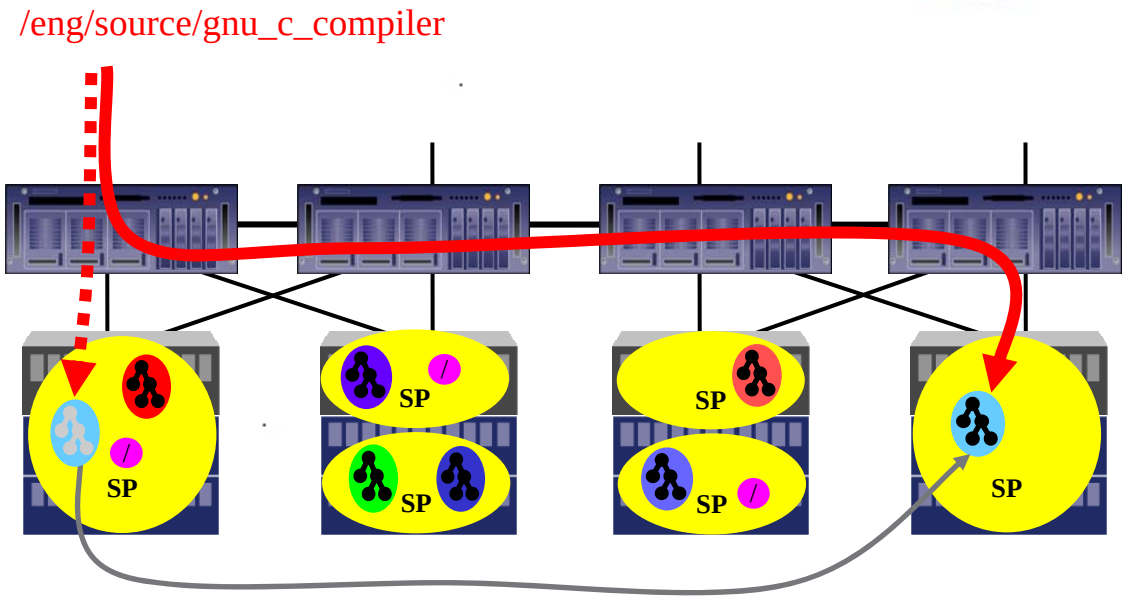


Global Namespace



SpinMove – Transparent Data Migration

- Non-disruptive data movement
- Rebalance data across existing or newly added storage and servers
- File movement during file access
- File location in namespace is unchanged
- Storage admin controlled
- Transparent to clients



Non-disruptive VFS move

Miles and Miles…..