The GUPFS Project at NERSC

Greg Butler, Rei Lee, Michael Welcome

NERSC

Lawrence Berkeley National Laboratory

One Cyclotron Road

Berkeley CA USA

NASA/IEEE MSST 2004

12th NASA Goddard/21st IEEE Conference on Mass Storage Systems & Technologies

The Inn and Conference Center

University of Maryland University College

Adelphi MD USA

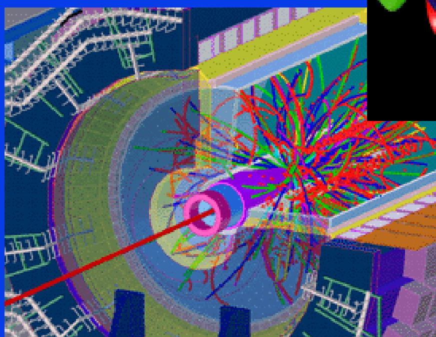
April 13-16, 2004

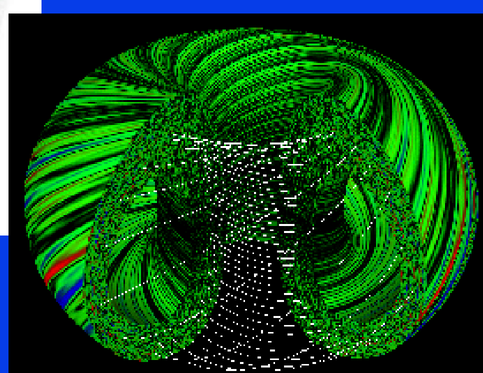
The GUPFS Project at NERSC

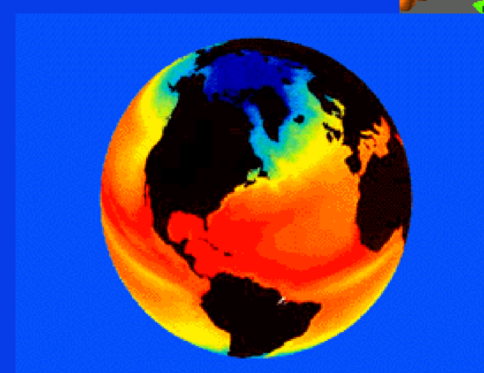
This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

National Energy Research Scientific Computing Center

- Serves all disciplines of the DOE Office of Science
- ~2000 Users in ~400 projects
- Focus on large-scale capability computing







Office of Science
U.S. DEPARTMENT OF ENERGY

LBNL & Department of Energy

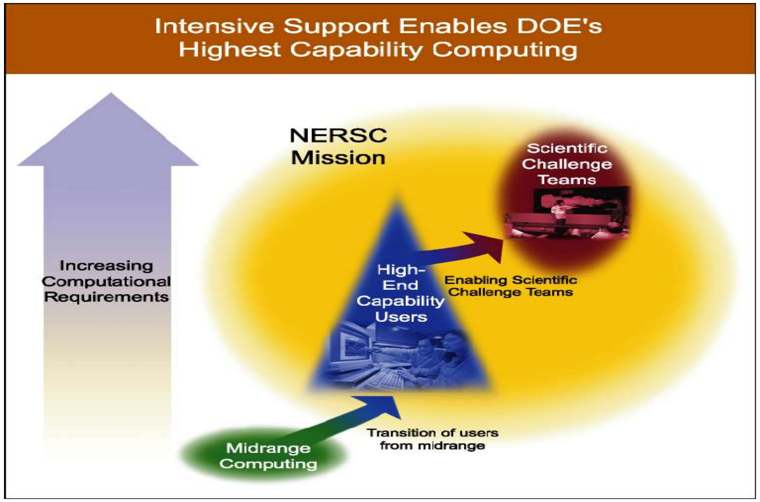
- NERSC is an DOE National Facility located at Lawrence Berkeley National Laboratory (LBNL)
- LBNL is operated by the University of California for the Department of Energy
- For most large scale computing, DOE has two major parts:
  - Office of Science
    - Fundamental research in high-energy physics, nuclear physics, fusion energy, energy sciences, biological and environmental sciences, computational science, materials science, chemical science, climate change, geophysics, genomics, life sciences.
    - Manages Berkeley Lab, Oak Ridge, Brookhaven, Argonne, Fermilab, SLAC, and others.
  - National Nuclear Security Administration (NNSA)
    - Nuclear weapons and defense
    - Manages Los Alamos, Sandia, Lawrence Livermore
    - ASC (previously ASCI) is associated with NNSA

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC Center Overview

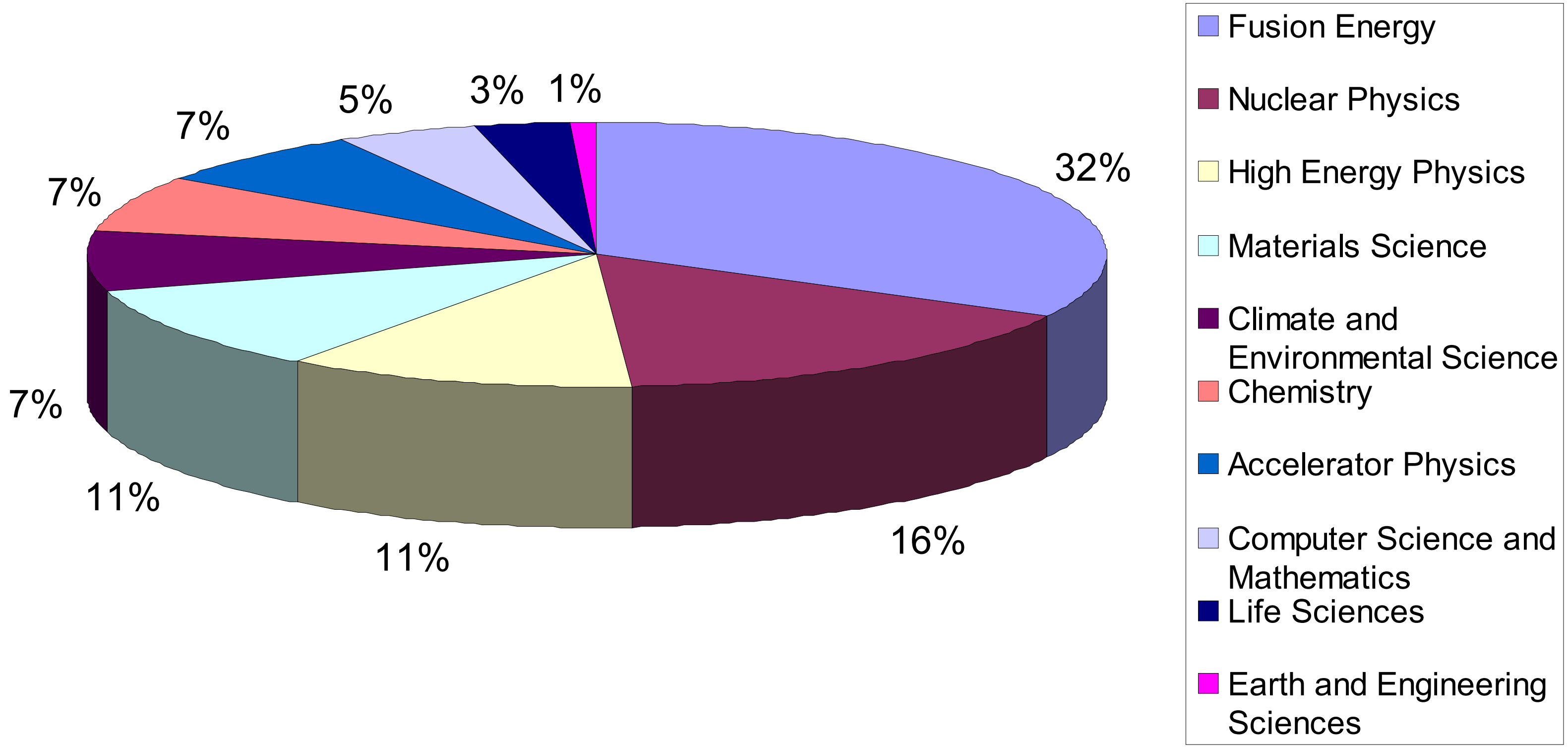
- Funded by DOE Office of Science, annual budget $28M, about 65 staff
- Supports open, unclassified, basic research, open to all researchers regardless of organization or funding agency
- Located at LBNL in the hills next to U of California, Berkeley campus
- Focus on large scale science that cannot be done elsewhere
  - Computational and Data Intensive Application Areas
  - Capability vs. Capacity

NERSC's Mission



NERSC is a world leader in <u>accelerating scientific discovery through computation</u>. Our vision is to provide high-performance computing resources and expertise to <u>tackle science's biggest and most challenging problems</u>, and to play a major role in advancing large-scale computational science and computer science.

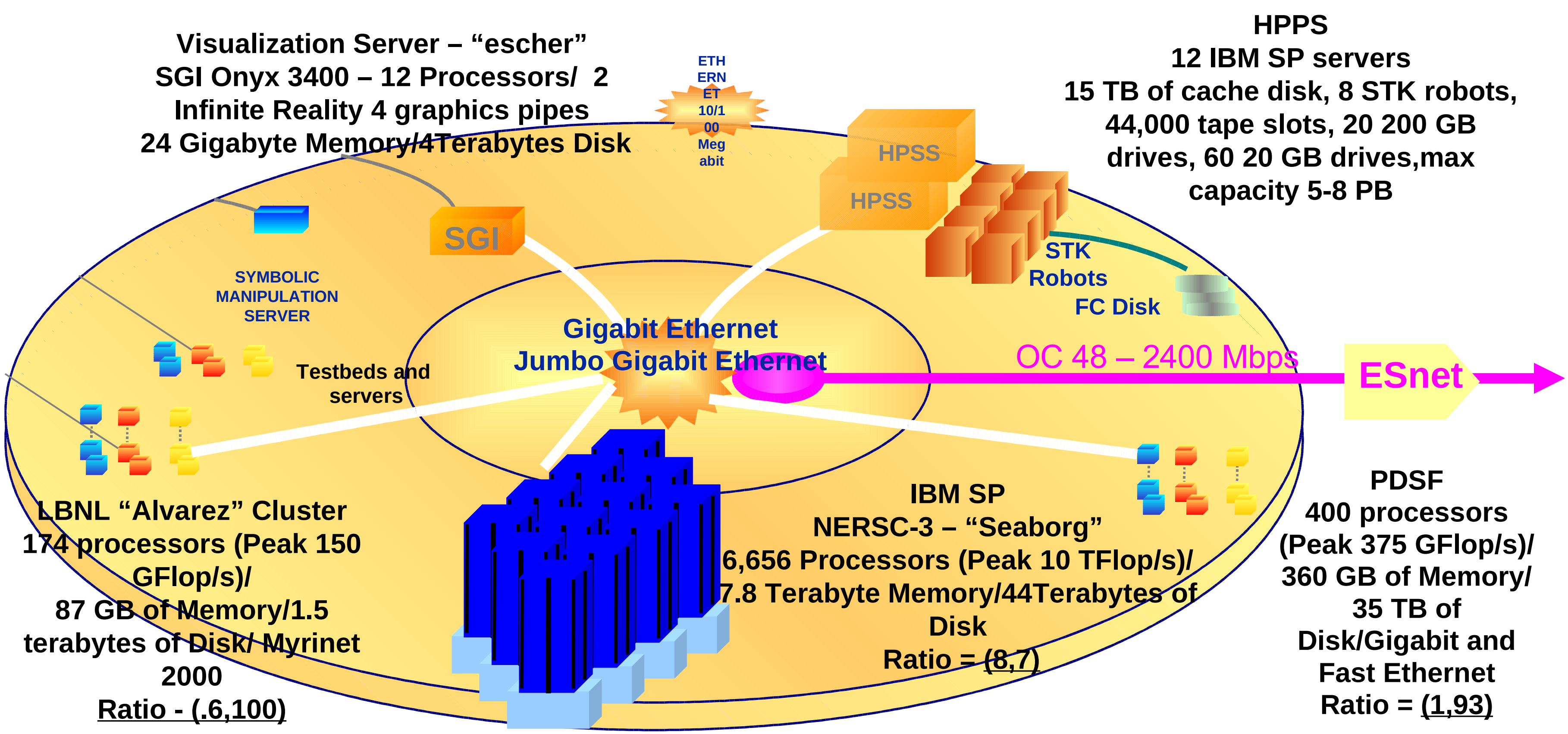
FY 03 Usage by Scientific Discipline



NERSC Resources

- “Seaborg” – 6656 processor IBM SP
  - 10 TFlop/s peak; 1-4 Tflop/s sustained
  - 7.8 TB memory; 50 TB online storage
- “PDSF” – 400+ processor Linux cluster
- “Alvarez” – 174-processor Linux cluster
- HPSS Storage system
  - 8STK robots; 20000 tapes; 1.2 PB data; 9 PB capacity
- “Escher” visualization server – SGI Onyx 3400
- Gigabit ethernet infrastructure; OC48 (2.4Gb/s) connection to ESNet – DOE network also managed by Berkeley Lab. Moving to 10 Gb/s in FY2004.

NERSC System Architecture



Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

National Energy Research Scientific Computing Center (NERSC)

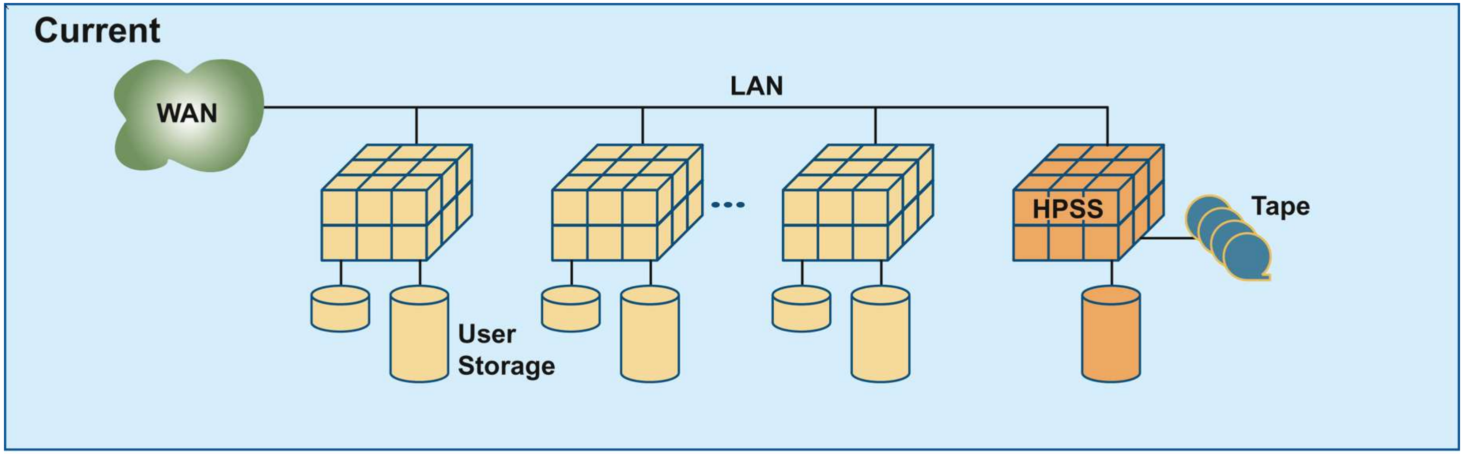
GUPFS Project Overview

- Five year project to deploy a center-wide shared file system at NERSC
- Purpose to make advanced scientific research using NERSC systems more efficient and productive
- Simplify end user data management by providing a shared disk file system in NERSC production environment
- An evaluation, selection, and deployment project
  - May conduct or support development activities to accelerate functionality or supply missing functionality

Global Unified Parallel File System (GUPFS)

- Global/Unified
  - A file system shared by major NERSC production systems
  - Using consolidated storage and providing unified name space
  - Automatically sharing user files between systems without replication
  - Integration with HPSS and Grid is highly desired
- Parallel
  - File system providing performance that is scalable as the number of clients and storage devices increase

Current NERSC Storage Configuration

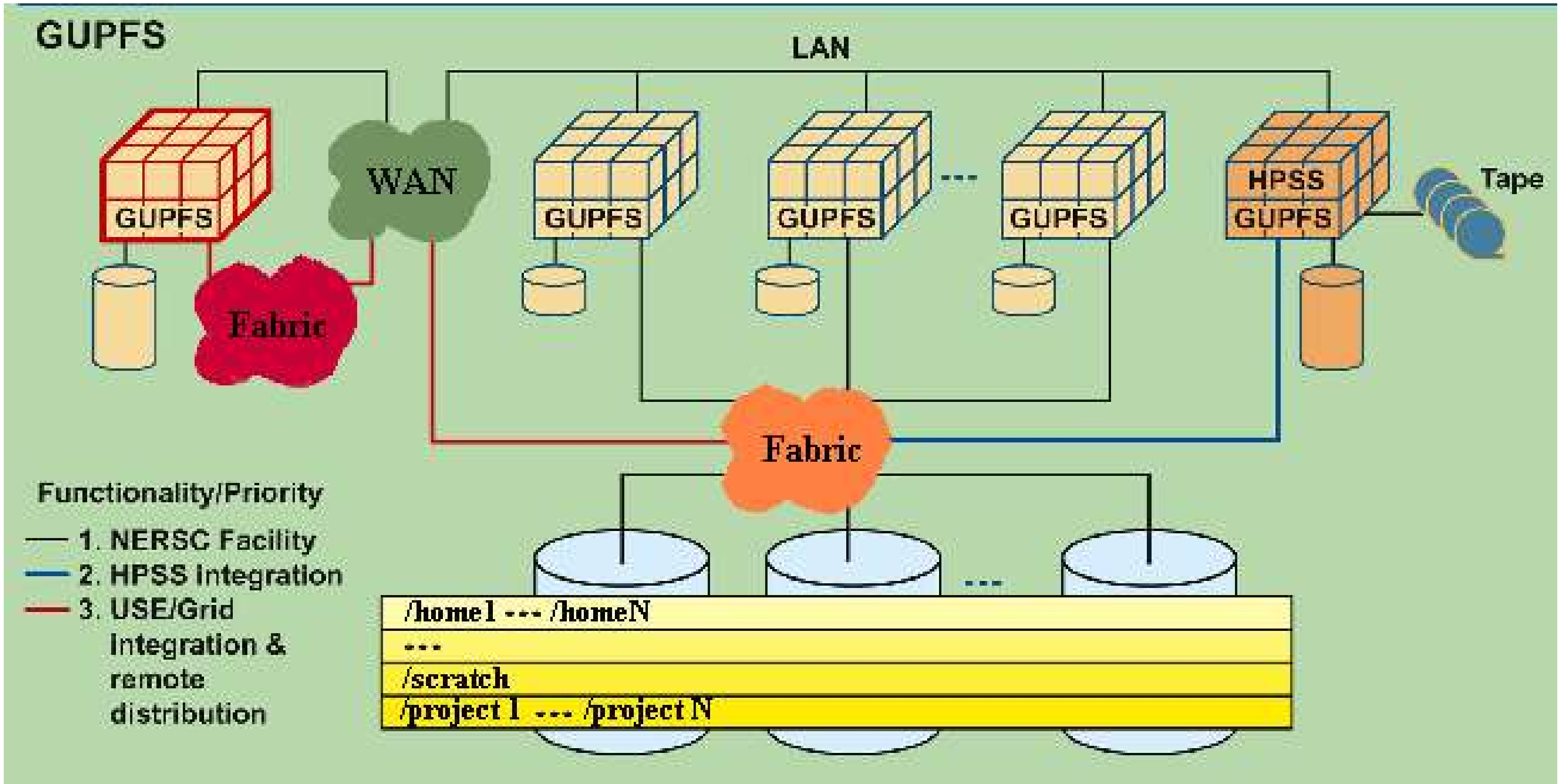


- Each system has its own separate direct-attached storage
- Each system has its own separate user file system and name space
- Data transfer between systems is over the LAN
- Includes large computational systems, small systems, and support systems

NERSC Storage Vision

- Single storage pool, decoupled from NERSC computational systems
  - Diverse file access - supporting both home file systems and large scratch file system
  - Flexible management of storage resource
  - All systems have access to all storage – require different fabric
  - Buy new storage (faster and cheaper) only as we need it
- High performance large capacity storage
  - Users see same file from all systems
  - No need for replication
  - Visualization server has access to data as soon as it is created
- Integration with mass storage
  - Provide direct HSM and backups through HPSS without impacting computational systems
- Potential geographical distribution

Envisioned NERSC Storage Configuration (GUPFS)



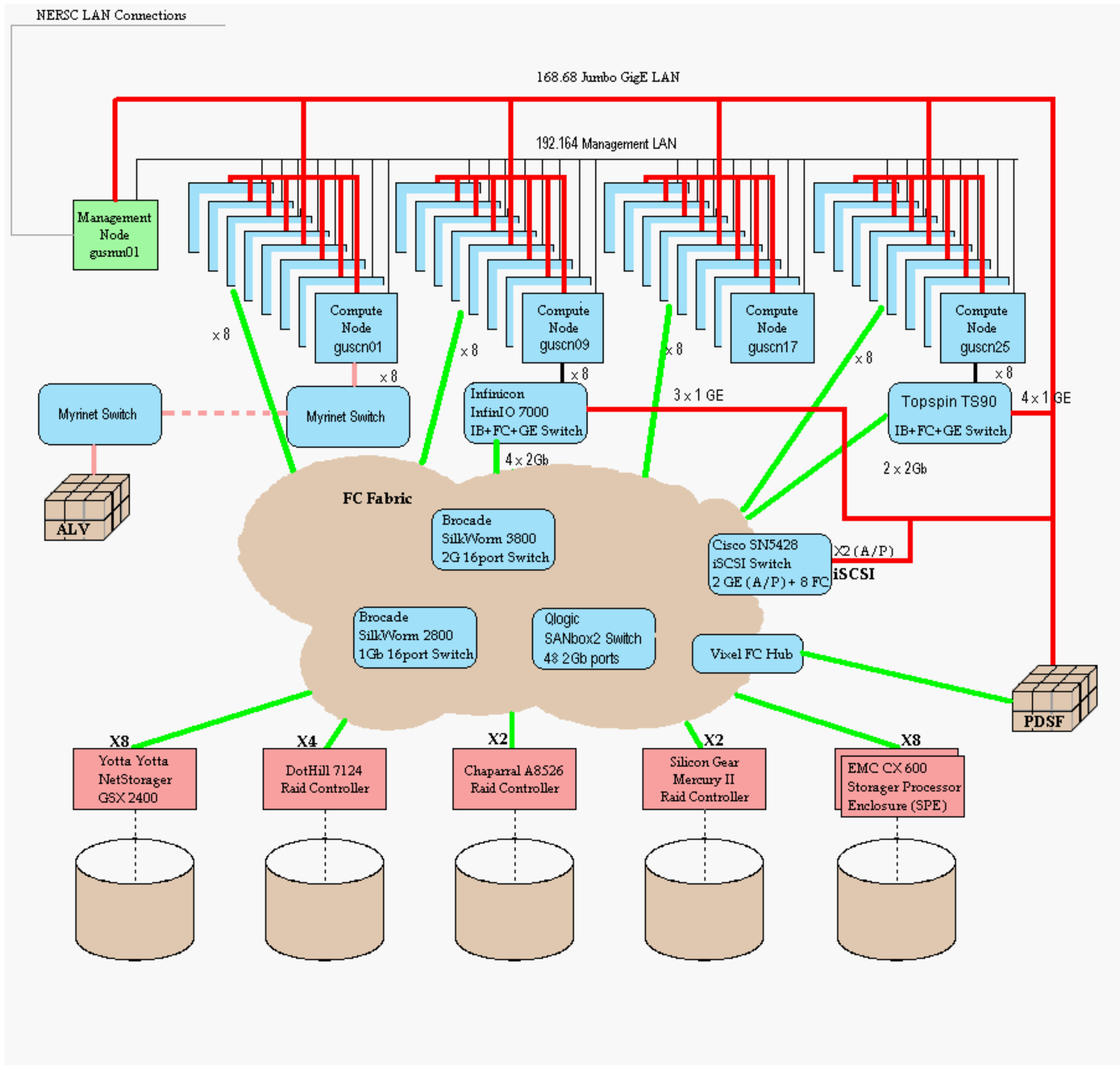
Where We Are Now

- Middle of the 3rd year of the 5-year project
  - Transition from component evaluation to deployment planning
- Evaluation of technology components needed for GUPFS (Shared File System, Network/SAN Fabric, and Storage)
  - Complex testbed simulating envisioned NERSC environment
  - Testing methodologies for evaluation
  - Collaborating with vendors: Emphasis on HPC I/O issues
- Focus now shifting to solution evaluation and deployment planning
  - Evaluation of solutions/systems rather than components.
  - Deployment planning: towards RFI, RFP, acquisition, integration.

GUPFS Testbed (FY2004)

- 32*P4 Compute
- 4*P4 Special
- 2*P3 Management
- 1*P3 Interactive
- 3*P3 Development
- GigE Interconnect
- Fibre Channel
- InfiniBand
- Myrinet
- Fiber Patch Panel
- 7.2 TB Disk Space
- 5 GB/s aggregate I/O

- Alvarez: 87*P3 nodes w/ Myrinet 2000
- PDSF: 414*P3/P4 nodes w/ 100bT & GigE



Office of Science
U.S. DEPARTMENT OF ENERGY

Technologies Evaluated

- File Systems
  - Sistina GFS 4.2, 5.0, 5.1, and 5.2 Beta
  - ADIC StorNext File System 2.0 and 2.2
  - Lustre 0.6 (1.0 Beta 1), 0.9.2, 1.0, 1.0.{1,2,3,4}
  - IBM GPFS for Linux, 1.3 and 2.2
  - Panasas
- Fabric
  - FC (1Gb/s and 2Gb/s): Brocade SilkWorm, Qlogic SANbox2, Cisco MDS 9509, SANDial Shadow 14000
  - Ethernet (iSCSI): Cisco SN 5428, Intel & Adaptec iSCSI HBA, Adaptec TOE, Cisco MDS 9509
  - Infiniband (1x and 4x): InfiniCon and Topspin IB to GE/FC bridges (SRP over IB, iSCSI over IB),
  - Inter-connect: Myrinnet 2000 (Rev D)
- Storage
  - Traditional Storage: Dot Hill, Silicon Gear, Chaparral
  - New Storage: Yotta Yotta GSX 2400, EMC CX 600, 3PAR, DDN S2A 8500

GUPFS Benchmark Methodology

- First determine baseline storage device raw performance
- Next connect storage over selected fabric
- Then install file system on storage and run benchmarks to determine performance impact of file system component
  - Parallel and metadata benchmarks used
- Impact of fragmentation as file system ages of interest but not explored yet

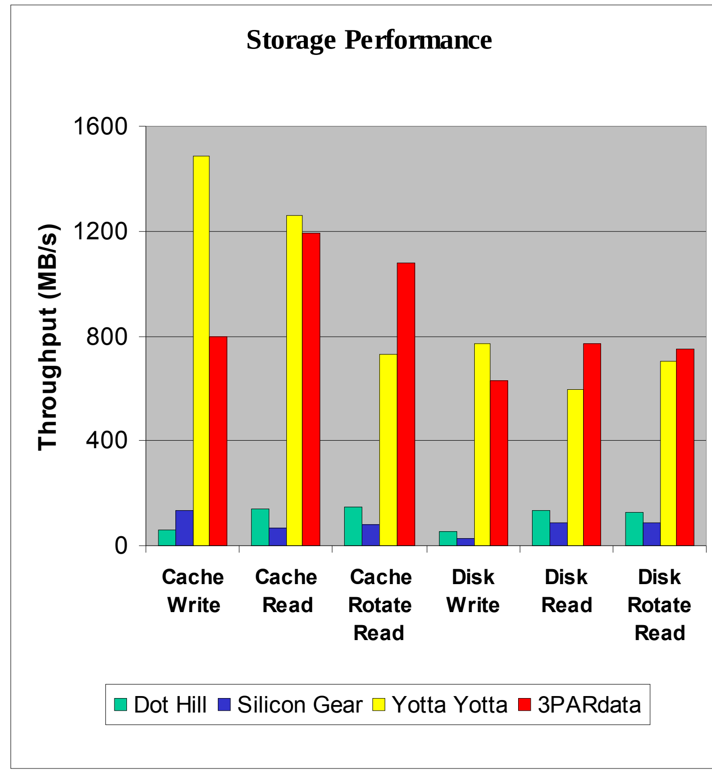
GUPFS Benchmarks

- Pioraw – parallel I/O to raw storage and file system
- Mptio – MPI based parallel I/O to file system
  - Cache Write
  - Cache Read
  - Cache Rotate Read
  - Disk Write
  - Disk Read
  - Disk Rotate Read
- Metabench – meta-data performance benchmarks
- Ioverify – file system integrity check

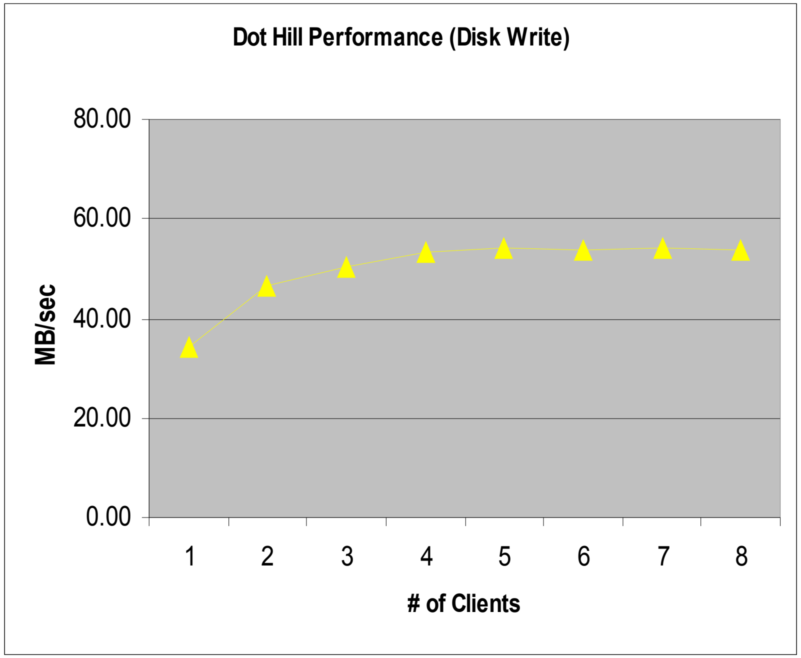
Storage Performance

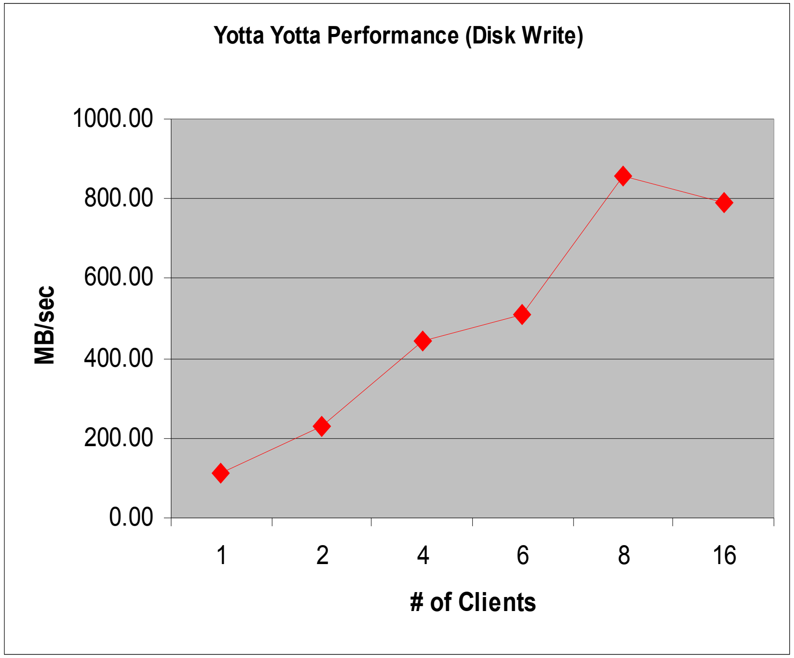
Max. Aggregate Performance (Raw I/O w/ 8 clients)				
	Dot Hill	Silicon Gear	Yotta Yotta	3PARdata
Cache Write	60	136	1487	799
Cache Read	143	66	1256	1193
Cache Rotate Read	144	83	733	1075
Disk Write	54	29	768	627
Disk Read	131	85	593	770
Disk Rotate Read	127	87	706	749

Storage Details				
	cache size	frontend ports	backend bus/ports	backend disks
Silicon Gear	256MB	FC: 2 X 1Gb	CSI: split bus	2 X 5+1 R5
Dot Hill	1GB	FC: 2 X 1Gb	C: 2 X 1Gbs	2 X 6-way R0
YottaYotta	4 X 4GE	FC: 8 X 2Gb	C: 8 X 2Gbs	28-way R0
3PARdata	4 X 8GE	FC: 16 X 2G	C: 8 X 2Gbs	R5 (160 disks)



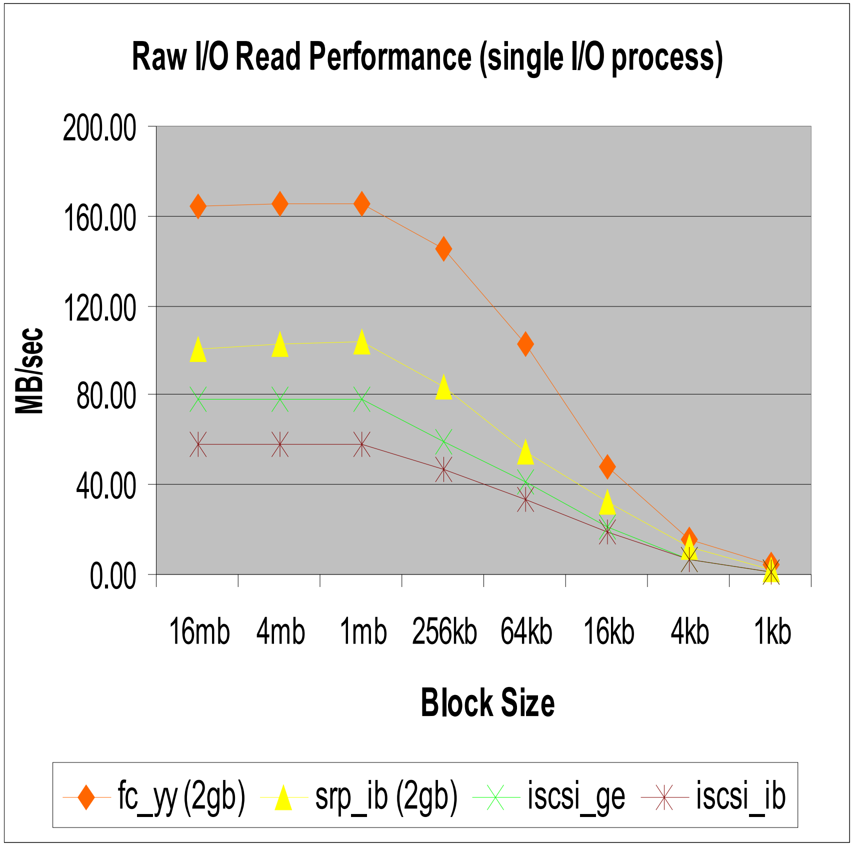
Storage Scalability

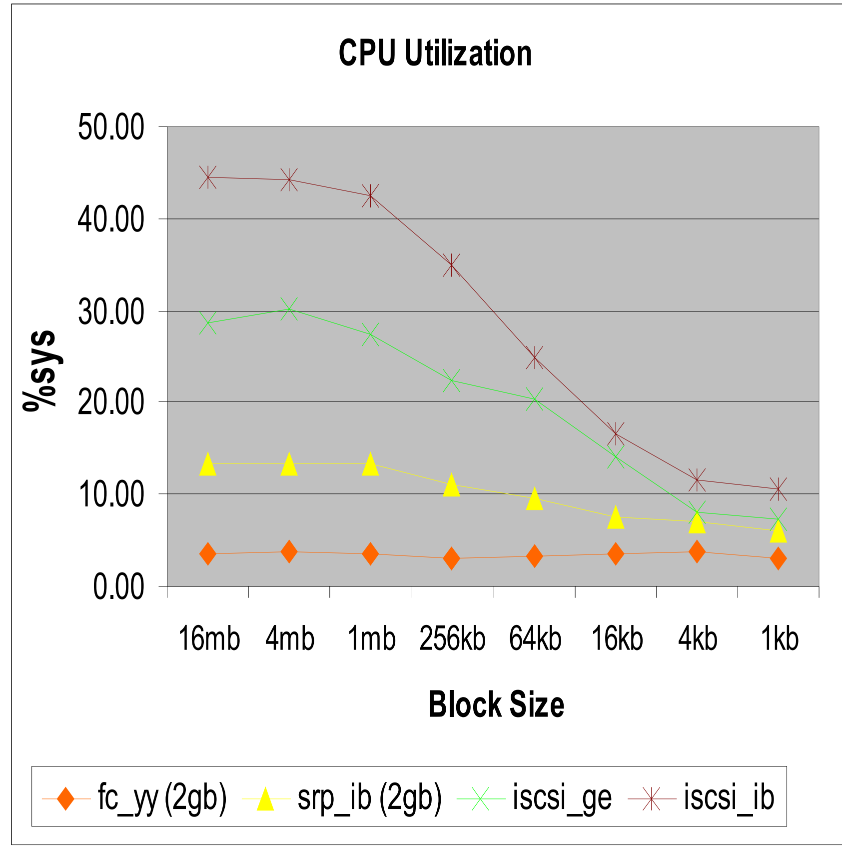




Traditional Storage: DotHill, 1 1Gb/s port, 6-way R0
New Storage: YottaYotta. 4 blades, 2 FC ports per blade, 32-way R0
Client: Dual P4 Xeon, 1GB Mem, Linux 2.4.18-10smp
HBA: Qlogic QLA2340 2Gb HBA

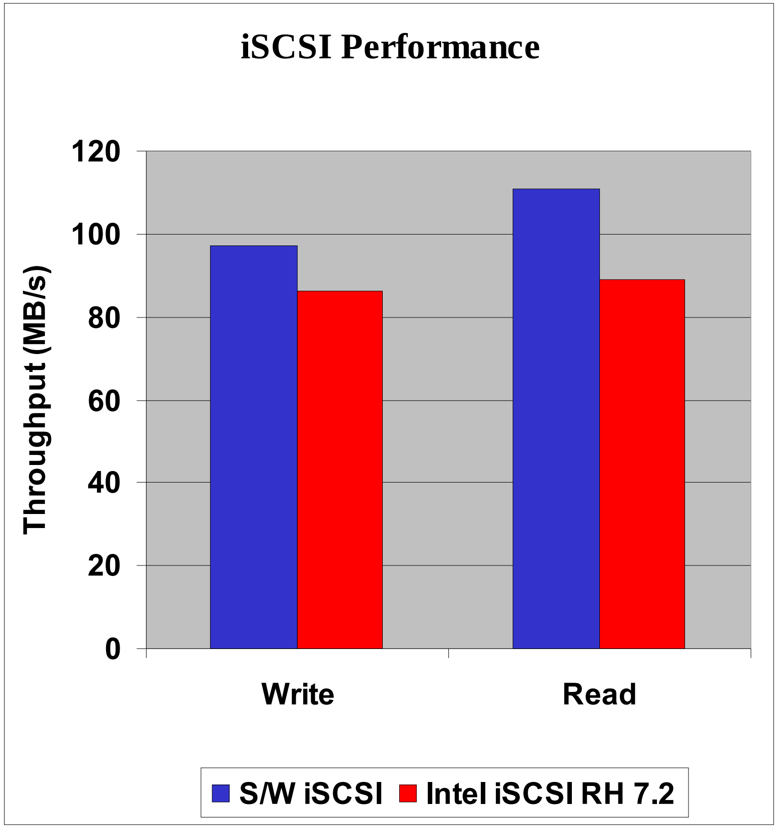
Comparison of Fabric Technologies

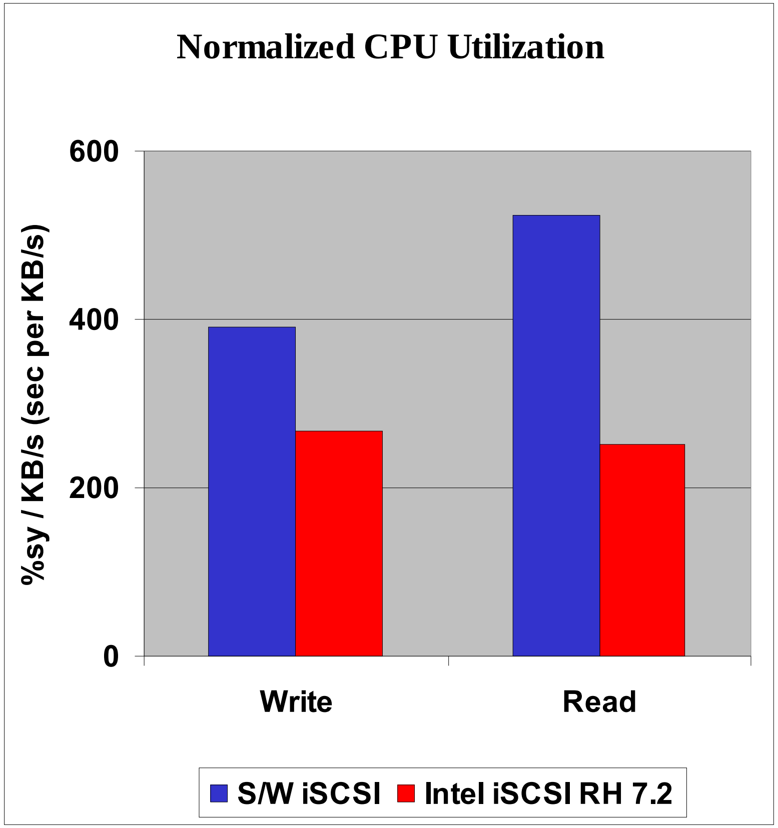




Storage: YottaYotta (1 2Gb/s FC Raid-0 LUN)
GE: Intel E1000
FC: Qlogic QLA2340 2Gb/s HBA
IB: InfiniCon Infin7000 Gen 1 switch & 4x HCA

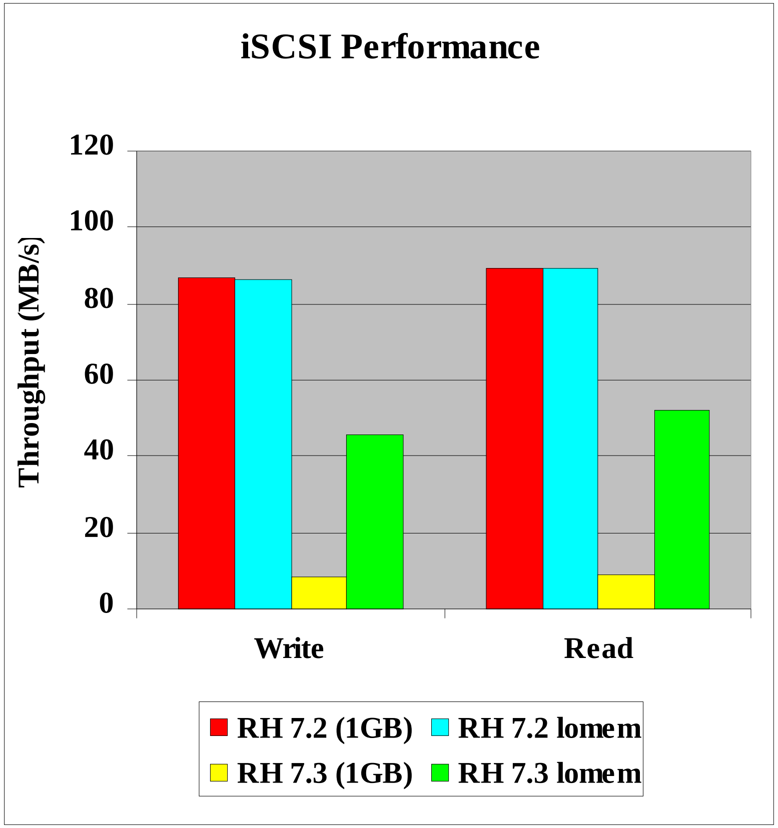
Software iSCSI vs. iSCSI HBA

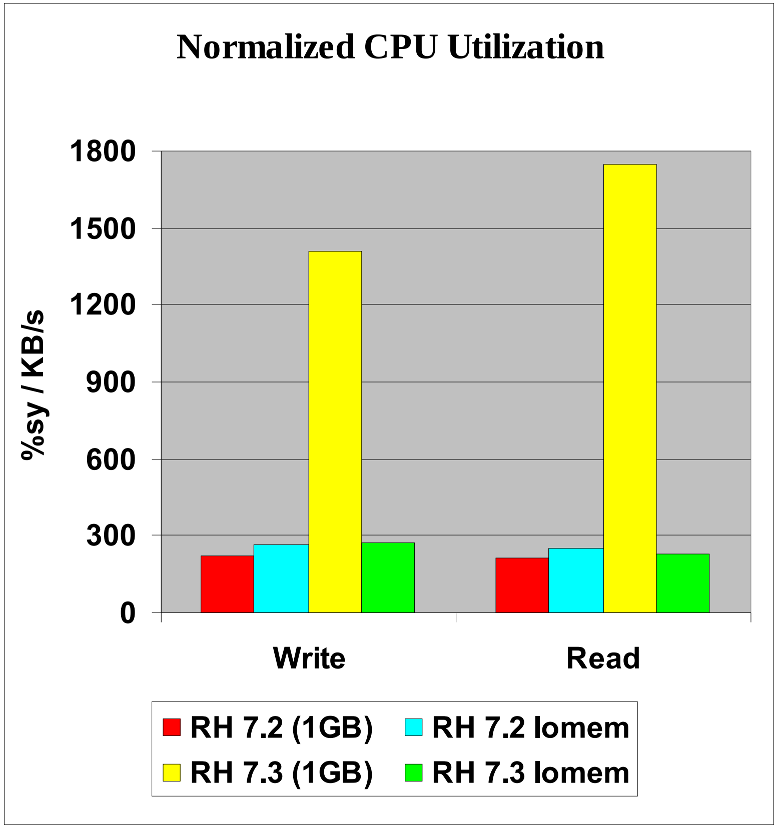




Storage: Chaparral (1 2Gb/s FC Raid-0 LUN)
iSCSI: Intel PRO/1000 IP Storage Adapter
Switch: Cisco SN5428 iSCSI Switch (1 1Gb/s iSCSI port)
Normalized CPU Utilization=%SYS/Throughput (sec per KB/s)
Date: 07/2003

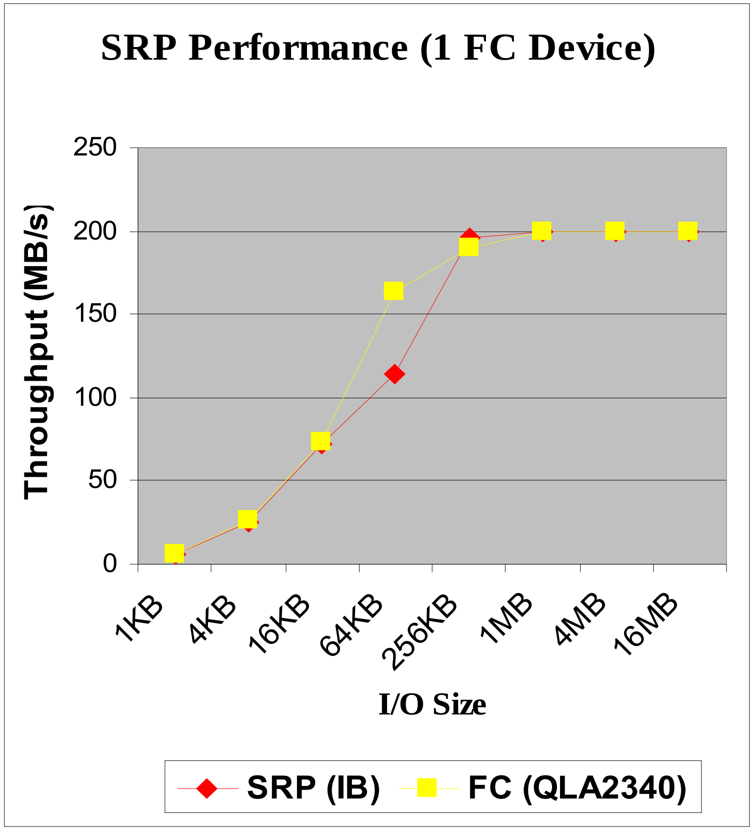
Performance Impact of Kernel Changes

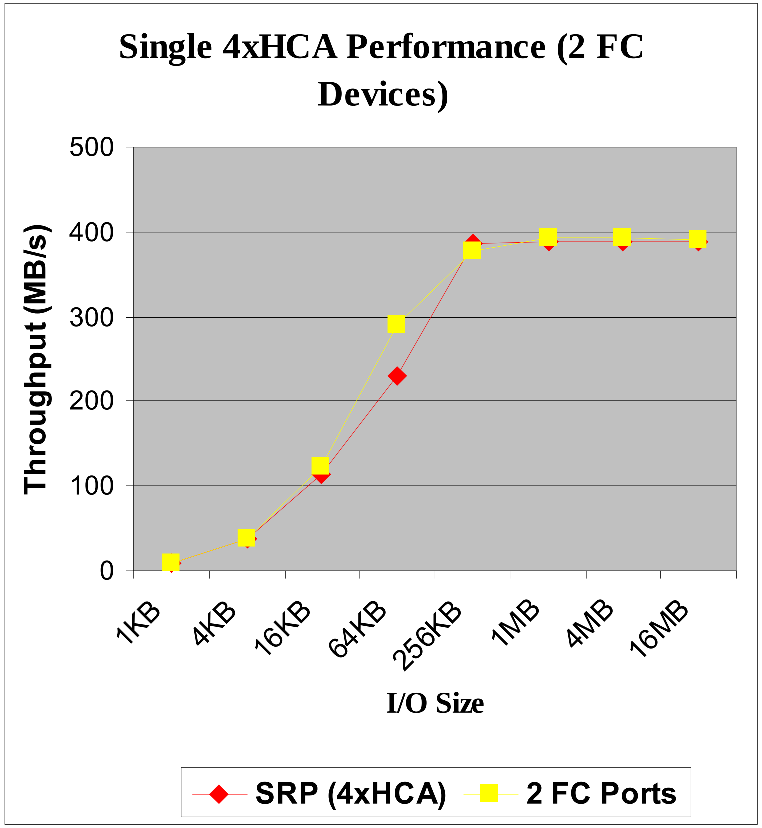




Storage: Chaparral (1 2Gb/s FC Raid-0 LUN)
iSCSI HBA: Intel PRO/1000 IP Storage Adapter
Switch: Cisco SN5428 iSCSI Switch (1 1Gb/s iSCSI port)
Date: 07/2003

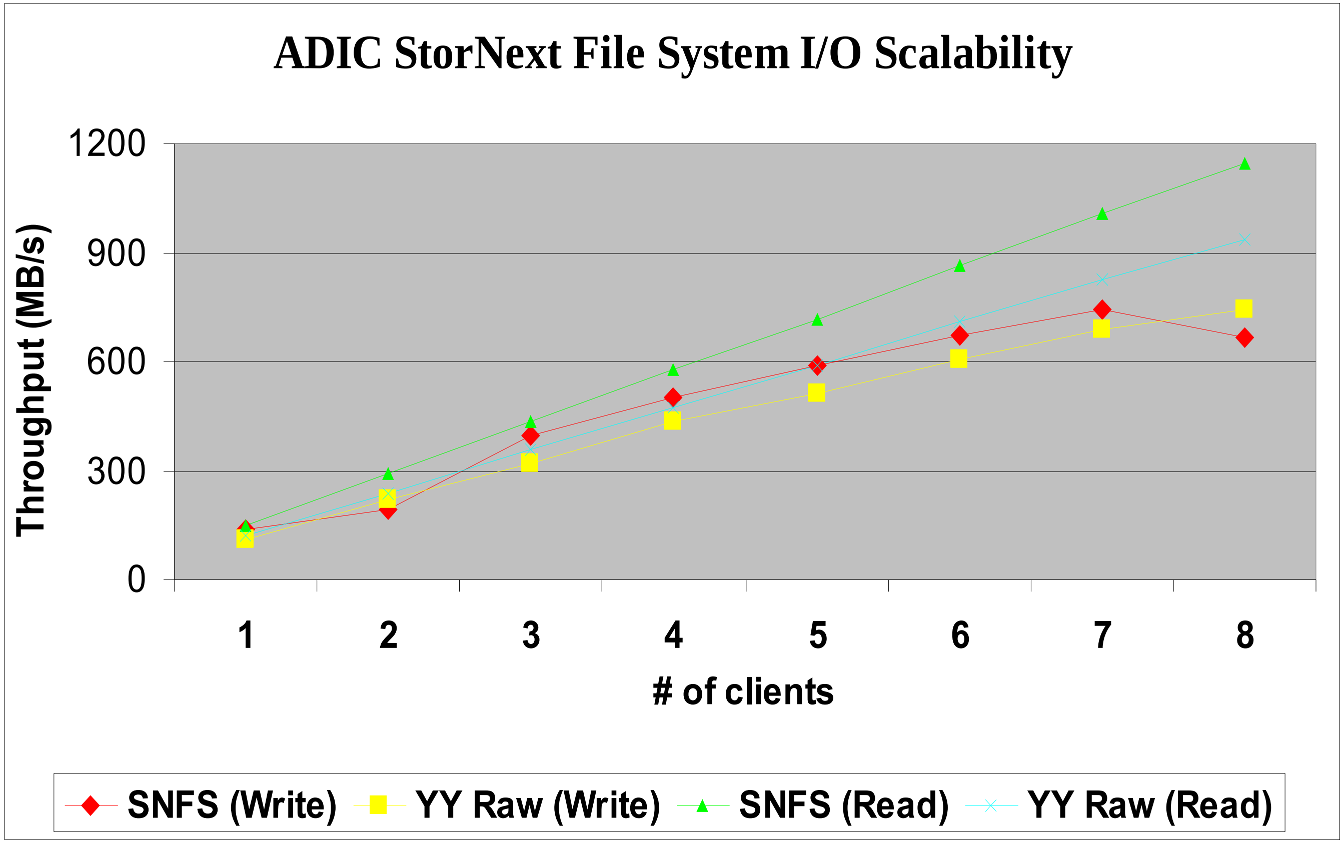
InfiniBand SRP Performance





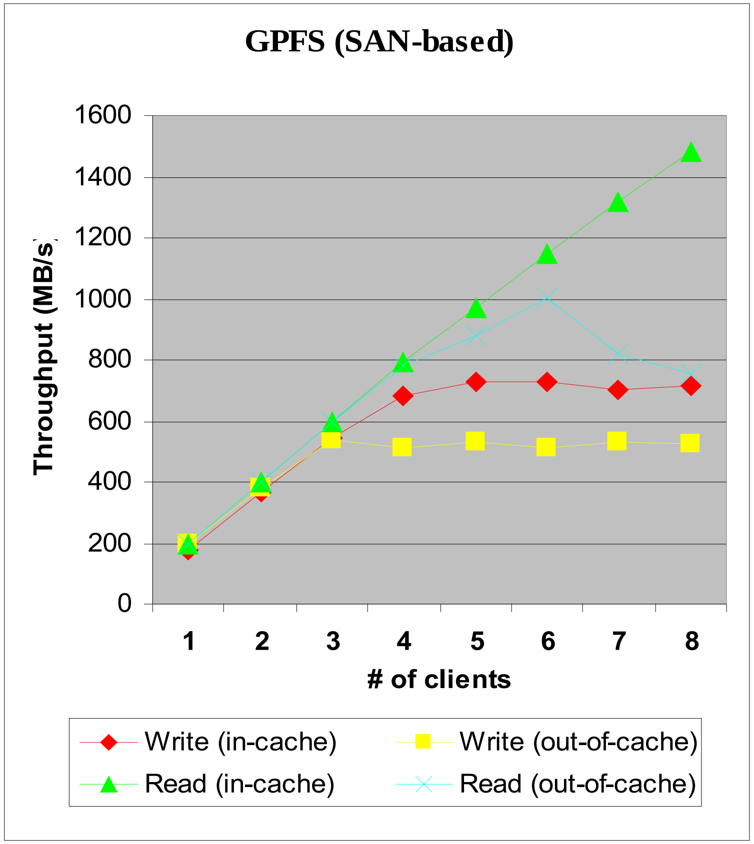
Storage: 3PARdata & EMC 600 (2Gb/s FC Raid-5 LUNs)
FC: Qlogic QLA2340 2Gb/s HBA
IB: Topspin 90 switch & 4x HCA
Date: 08/2003

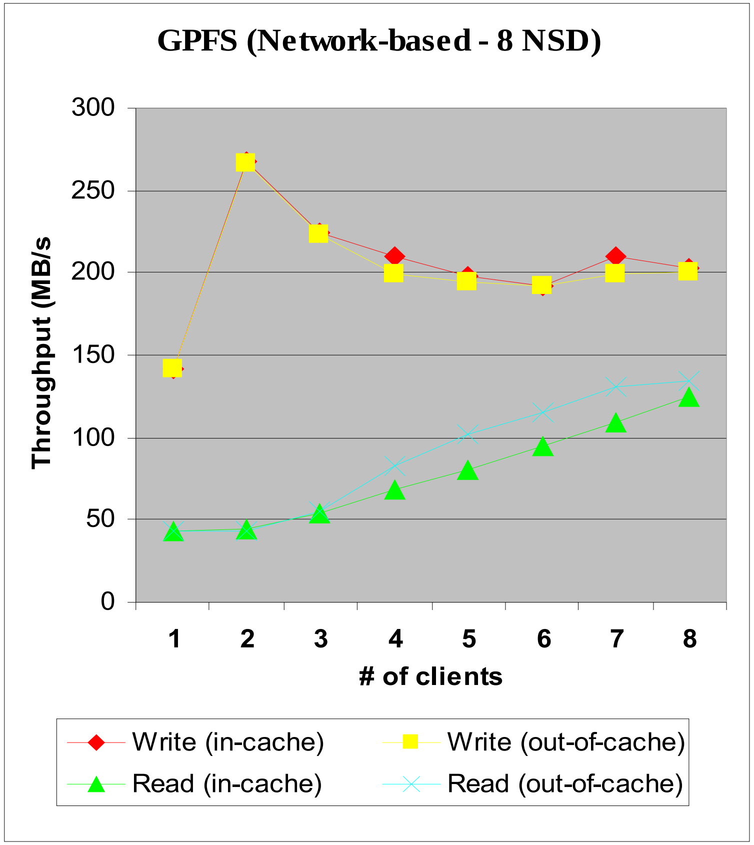
ADIC StorNext File System 2.0



Storage: YottaYotta, Single 28-way RAID-0 LUN (exported on 8 2Gb/s ports)
Clients: Dual P4 Xeon, 1GB Mem, Linux 2.4.8-10smp
FC: Qlogic QLA 2340

GPFS/Linux 1.3 Performance



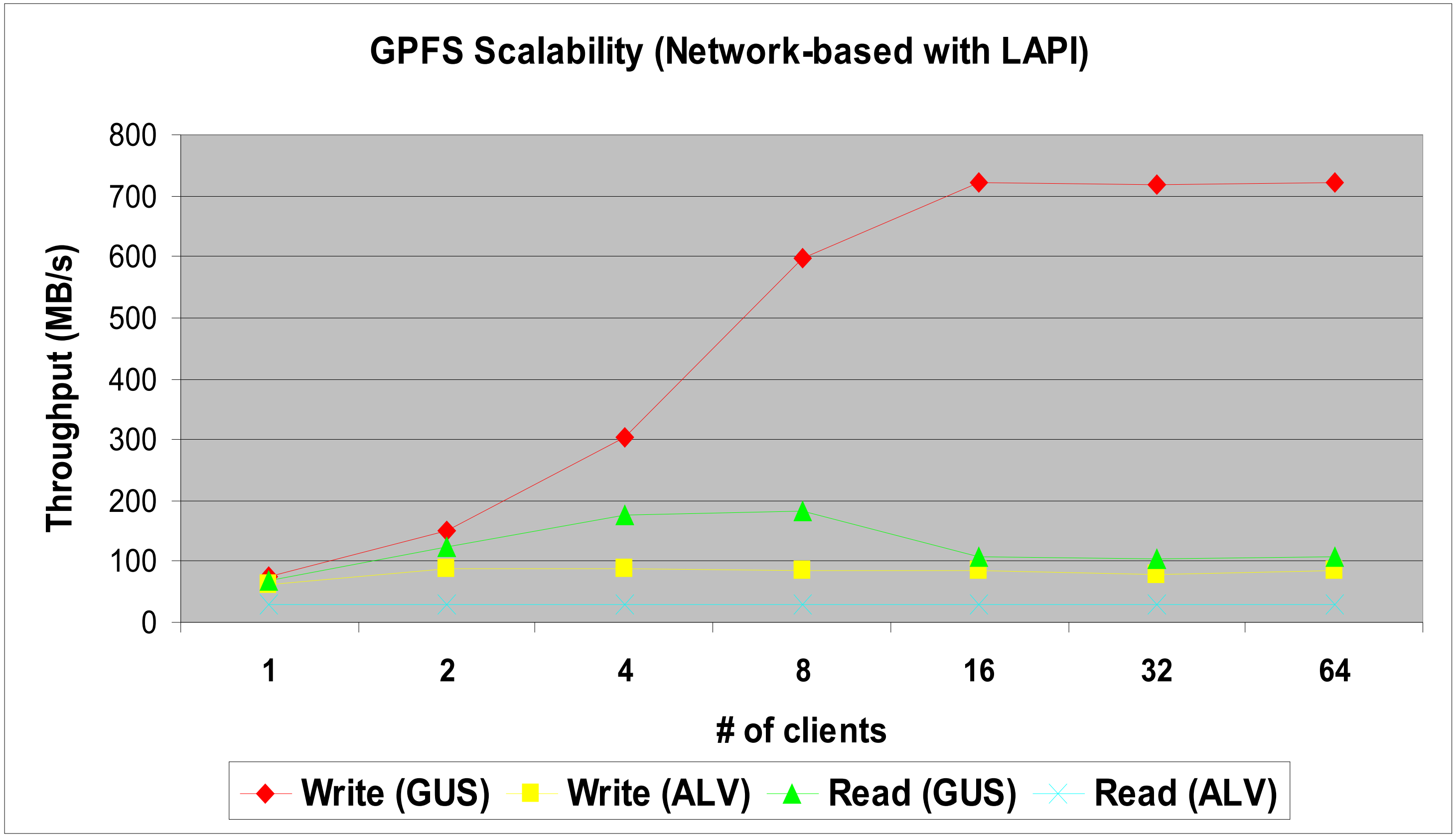


SAN: 3PARdata (8 2Gb/s FC ports over 1 Raid-5 LUN)
Clients: Dual P4 Xeon, 2GB Mem, Linux 2.4.18-10smp
FC: Qlogic QLA2340
Interconnect: Myrinet (Rev D) LAPI

NSD: 8 YottaYotta 4-way RAID-0 LUNs (8 2Gb/s ports)
Clients: Dual P4 Xeon, 2GB Mem, Linux 2.4.8-10smp
FC: Qlogic QLA 2340
Interconnect: GigE

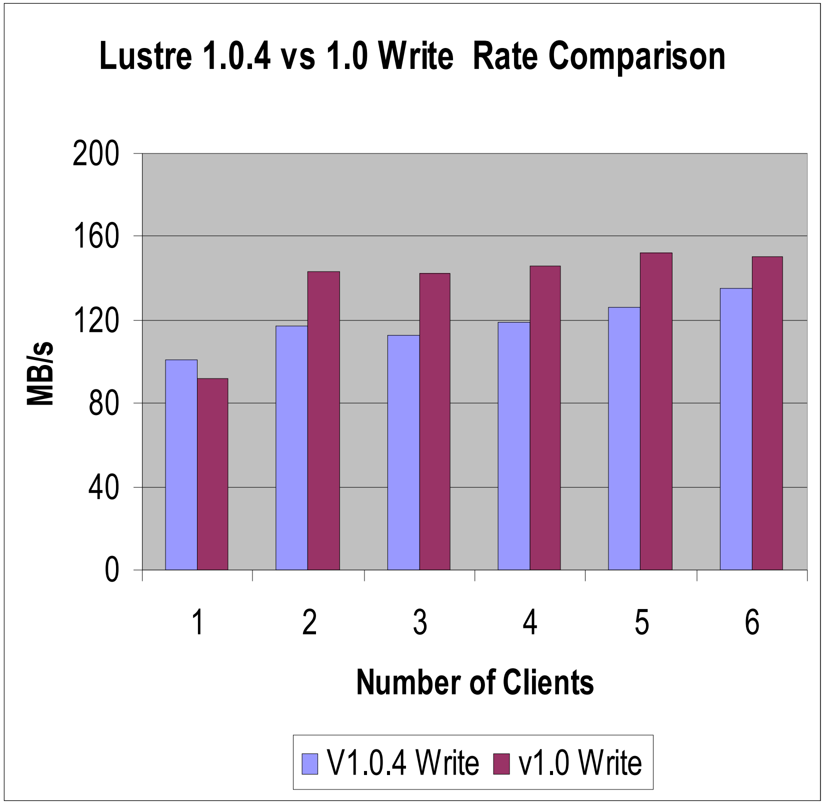
Office of Science
U.S. DEPARTMENT OF ENERGY

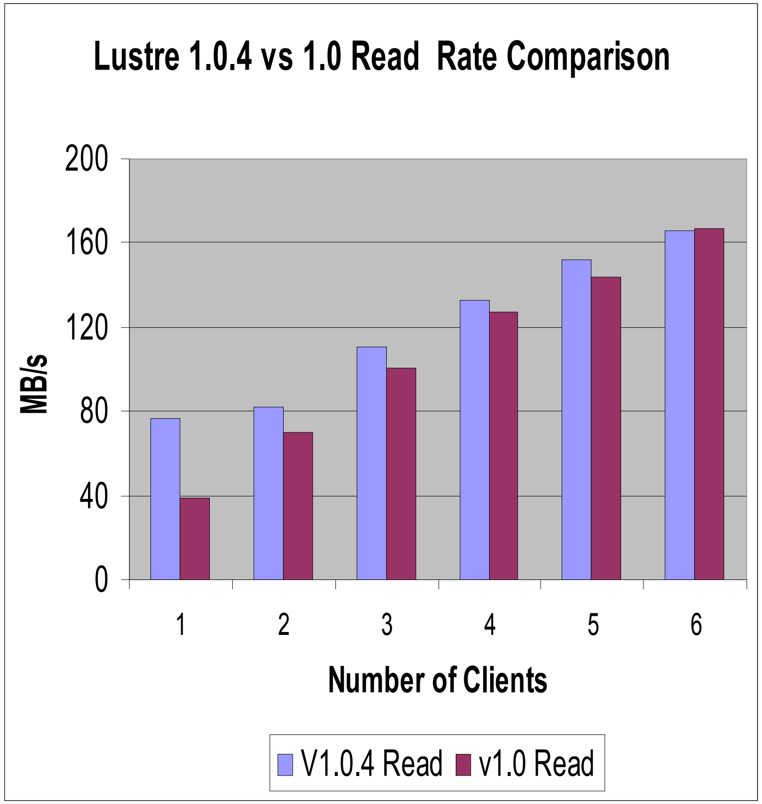
GPFS/Linux 1.3 I/O Scalability



Linux Cluster: NERSC Alvarez (87 nodes)
Storage (ALV): 2 storage nodes (IBM Raid-5 LUNs, 100MB/s max)
Storage (GUS): YottaYotta 8 Raid-0 LUNs, 8 x 2Gb FC ports
Interconnect: Myrinet 2000 (Rev C)
Date: 08/2003 – 10/2003

Lustre PIORAW Results

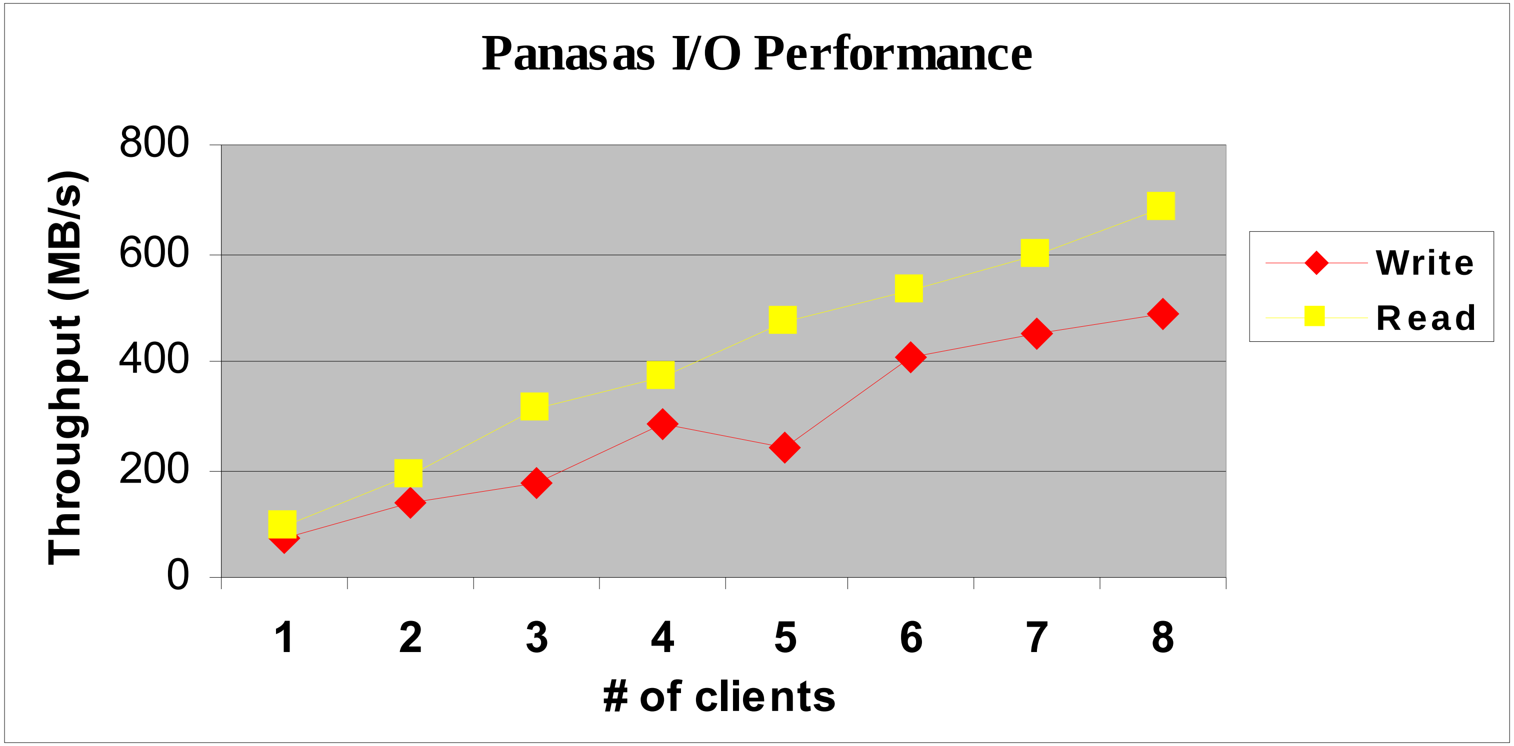




Host: RH 9.0
Lustre: 6 clients, 2 OSSs
Storage: EMC CX 600 2 LUNs
Interconnect: GigE (Dell GigE switch)
Date: Jan 2004 - March 2004

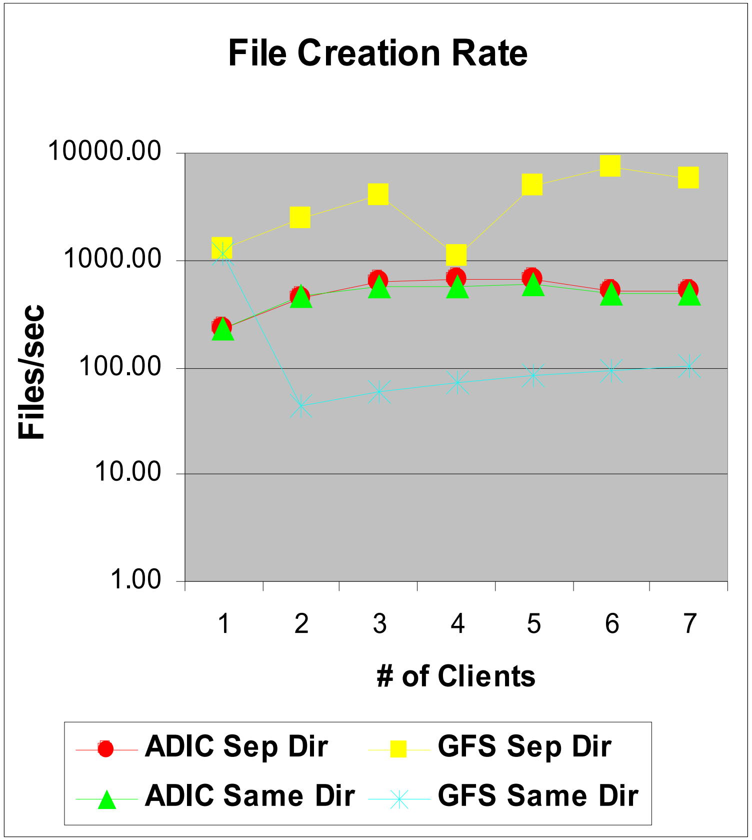
Office of Science
U.S. DEPARTMENT OF ENERGY

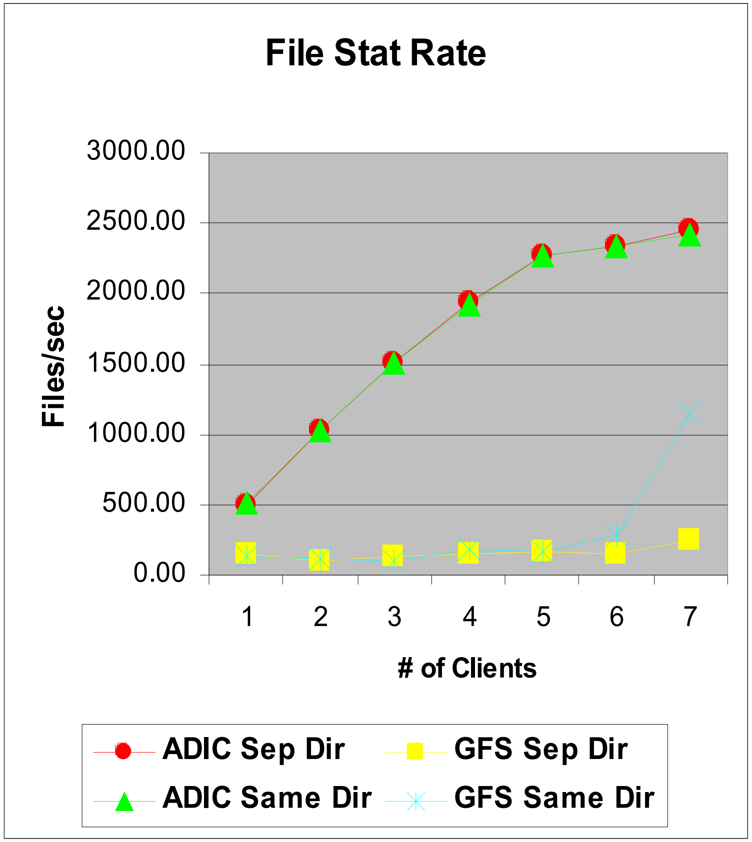
Panasas File System 2.0.7



Host: Linux 2.4.21
Panasas: 2 Shelves, 8 x 1 GigE with trunking, 2 DirectorBlade, 20 StorageBlade, with Direct Flow (Release 2.0.7)
Interconnect: GigE
File Size per I/O process/client: 1GB
Date: Mar 2004

File System Metadata Performance





Storage: YottaYotta (Raid-0 LUNs, 8 2Gb/s FC ports)
FC: Qlogic QLA2340 2Gb/s HBA
GFS: 5.1
ADIC: 2.0
Date: 03/2003

Office of Science
U.S. DEPARTMENT OF ENERGY

Near Term Activities

- Conduct preliminary selection and deployment planning
  – Preparing for RFI, RFP
- Narrow focus to candidate file systems
- Possible future evaluations (fewer and more focused):
  - IBM SAN File System (StorageTank)
  - Myrinet Gigabit Ethernet blade (fabric bridge)
  - 10Gb Ethernet
  - 4Gb and 10Gb Fibre Channel

Summary

- GUPFS Project has evaluated components necessary for center-wide file system
- Developed institutional knowledge base needed to select and deploy
- Working with vendors to incorporate HPC requirements into their products
- Fabric and storage have been advancing and are not major concerns
- File systems, improving but still need work for parallel production environments

Observations

General

- Many vendors are involved and have various products available. But are more focused on commercial market.
- Need standardized, centralized monitoring and management for fabric and storage.

Filesystem

- Progress but still need more work, remains the component with the highest risk
  - Stability
  - Parallel I/O performance and functionality
  - Scalability
- Not enough filesystems can or plan to support many platforms/OS's.
  - ADIC SNFS greatest variety now
  - IBM Storage Tank has many now and more planned
  - Lustre plans to support OS-X
  - IBM GPFS supports AIX and Linux

Observations (cont'd)

Filesystem (cont'd)

- File system vendors should open source their client software to assist wide scale adoption
  - Storage Tank Linux client and Lustre already open source
  - Open sourcing under consideration by others

Fabric

- Better and more extensive bridging capabilities needed
- Need better inter-switch link and aggregate bandwidth
- Need policy driven quality of service (QoS) capabilities for all fabrics
- Applaud open sourcing of IB drivers

Storage

- Multi-port storage and multiple interconnect storage desired
- Need devices with higher aggregate and individual port bandwidth
- Still need work on supporting very large number of initiators

Thank You

- GUPFS Project Web Site
  - http://www.nersc.gov/projects/gupfs
- Contact Info:
  - Project Lead: Greg Butler (gbutler@nersc.gov)

GUPFS Team

Greg Butler (GButler@nersc.gov, Project Lead)
Will Baird (WPBaird@lbl.gov)
Rob Farber (RMFarber@lbl.gov)
Rei Lee (RCLee@lbl.gov)
Craig Tull (CETull@lbl.gov)
Michael Welcome (MLWelcome@lbl.gov)
Cary Whitney (CLWhitney@lbl.gov)