Performance, Reliability, and Operational Issues for High Performance NAS Storage on Cray Platforms

Cray User Group Meeting

June 2007

CRAY
THE SUPERCOMPUTER COMPANY

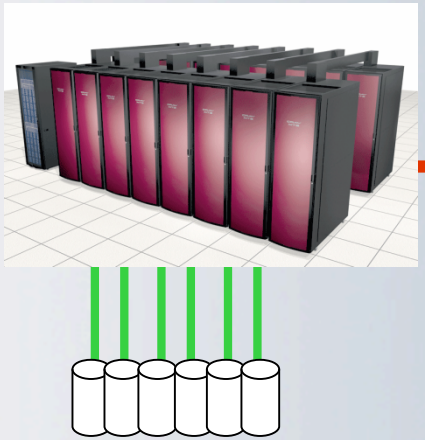
Cray's Storage Strategy

- Background
  - Broad range of HPC requirements – big file I/O, small file I/O, scalability across multiple dimensions, data management, heterogeneous access…
  - Rate of improvement in I/O performance lags significantly behind Moore's Law

- Direction
  - Move away from "one solution fits all" approach
  - Use cluster file system for supercomputer scratch space and focus on high performance
  - Use scalable NAS, combined with data management tools and new hardware technologies for shared and managed storage

Cray Advanced Storage Architecture (CASA)

High Performance Scratch Files:
- Shared file system
- RAID disk



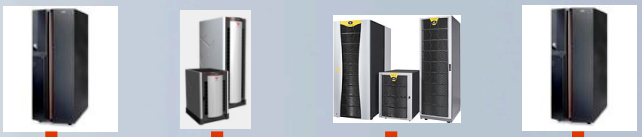
CASA Partners

High Performance Scratch Files:
- Shared file system
- RAID disk

CFS – Lustre

DDN, Engenio – RAID disk

Servers, Clusters, and Workstations



10G/1G Ethernet Network

Fast NAS | Fast NAS | Fast NAS | Fast NAS

FC fabric

HSM/ Backup Server | HSM/ Backup Server

ADIC – HSM

TBD – backup

Virtual Tape Archive

COPAN – VTL (MAID)

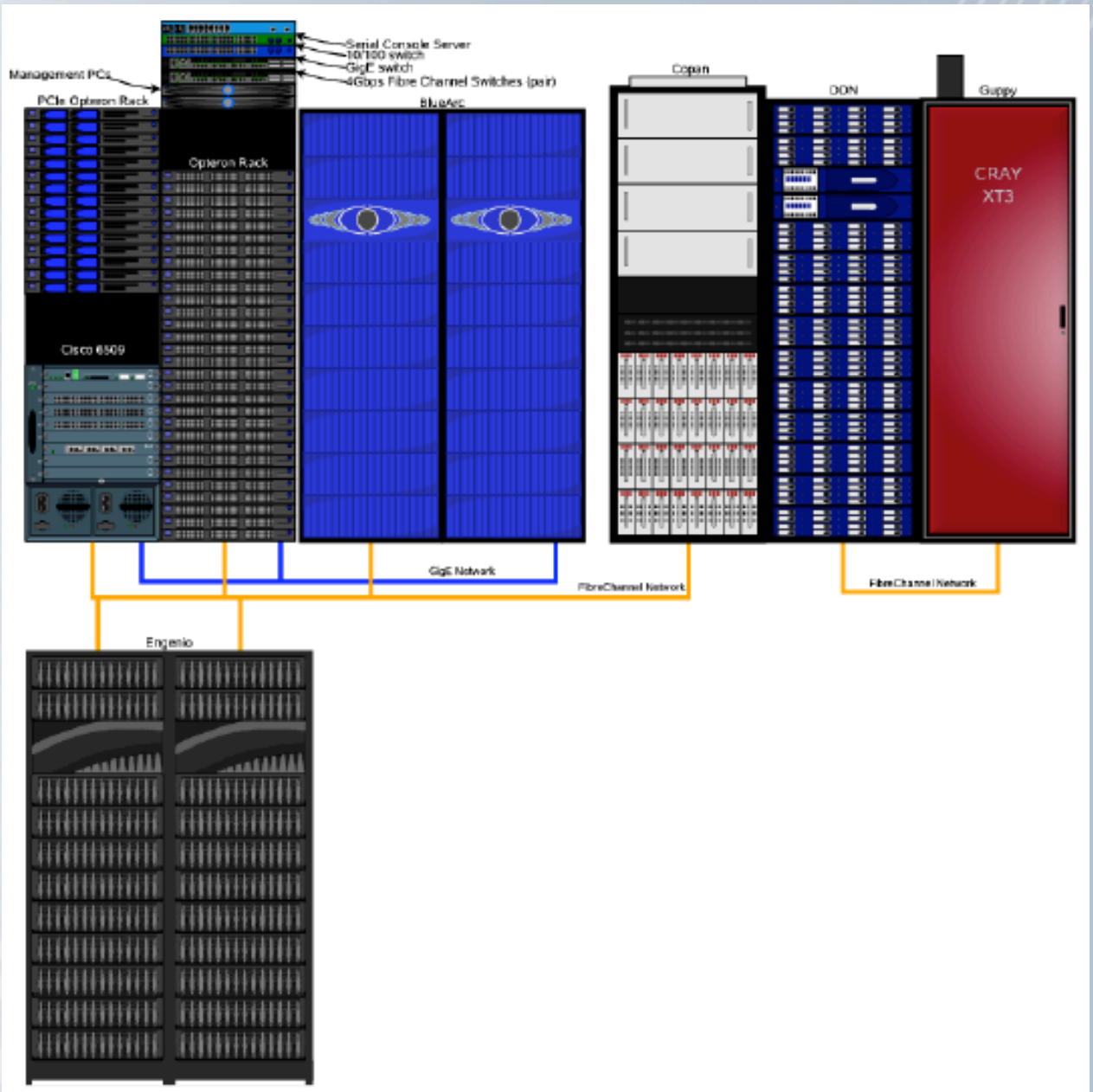
BlueArc – scalable NAS

HSM/Backup Tape Robot/ Archive

ADIC – tapes

CASA Lab Chippewa Falls

- Opteron Cluster
- Cisco 6509 switch
- Engenio Storage
- COPAN MAID



CASA Lab

- CASA Lab
- Opteron Cluster
- Related Storage

CASA Lab

- Blue Arc Titan Servers
- Engenio Storage

Blue Arc Titan-2 Dual Heads

CASA Lab

- COPAN Revolution System

How will this help?

- Use a cluster file system for big file (bandwidth) I/O for scalable systems
  - Focus on performance for applications
- Use commercial NAS products to provide solid storage for home directories and shared files
  - Vendors looking at NFS performance, scalability
- Use new technologies – nearline disk, virtual tape – in addition to or instead of physical tape for backup and data migration
  - Higher reliability and performance

Major HPC Storage Issue

- Too many HPC RFPs (esp for supercomputers) treat storage as secondary consideration
  - Storage “requirements” are incomplete or ill-defined
    - Only performance requirement and/or benchmark is maximum aggregate bandwidth
      - No small files, no IOPS, metadata ops
    - Requires “HSM” or “backup” with insufficient details
    - No real reliability requirements
  - Selection criteria don’t give credit for a better storage solution
    - Vendor judged on whether storage requirements are met or not

- Result: vendor proposes the minimum cost solution that meets the storage requirements
  - Gets rewarded for putting the rest of the budget towards TFLOPS

Why NFS?

- NFS is the basis of the NAS storage market (but CIFS important as well)
  - Highly successful, adopted by all storage vendors
  - Full ecosystem of data management and administration tools proven in commercial markets
  - Value propositions – ease of install and use, interoperability
- NAS vendors are now focusing on scaling NAS
  - Various technical approaches for increasing client and storage scalability
- Major weakness – performance
  - Some NAS vendors have been focusing on this
  - We see opportunities for improving this

CASA Lab Benchmarking

- CASA Lab in Chippewa Falls provides testbed to benchmark, configure and test CASA components
  - Opteron cluster (30 nodes) running Suse Linux
  - Cisco 6509 switch
  - BlueArc Titan — dual-heads, 6x1 Gigabit Ethernet on each head
  - Dual-fabric Brocade SAN with 4 FC controllers and 1 SATA controller
  - Small Cray XT3

Test and Benchmarking Methodology

- Used Bringsel tool (J. Kaitschuck — see CUG paper)
  - Measure reliability, uniformity, scalability and performance
  - Creates large, symmetric directory trees, varying file sizes, access patterns, block sizes
  - Allows testing of the operational behavior of a storage system: behavior under load, reliability, uniformity of performance
- Executed nearly 30 separate tests
  - Increasing complexity of access patterns and file distributions
  - Goal was to observe system performance across varying workloads

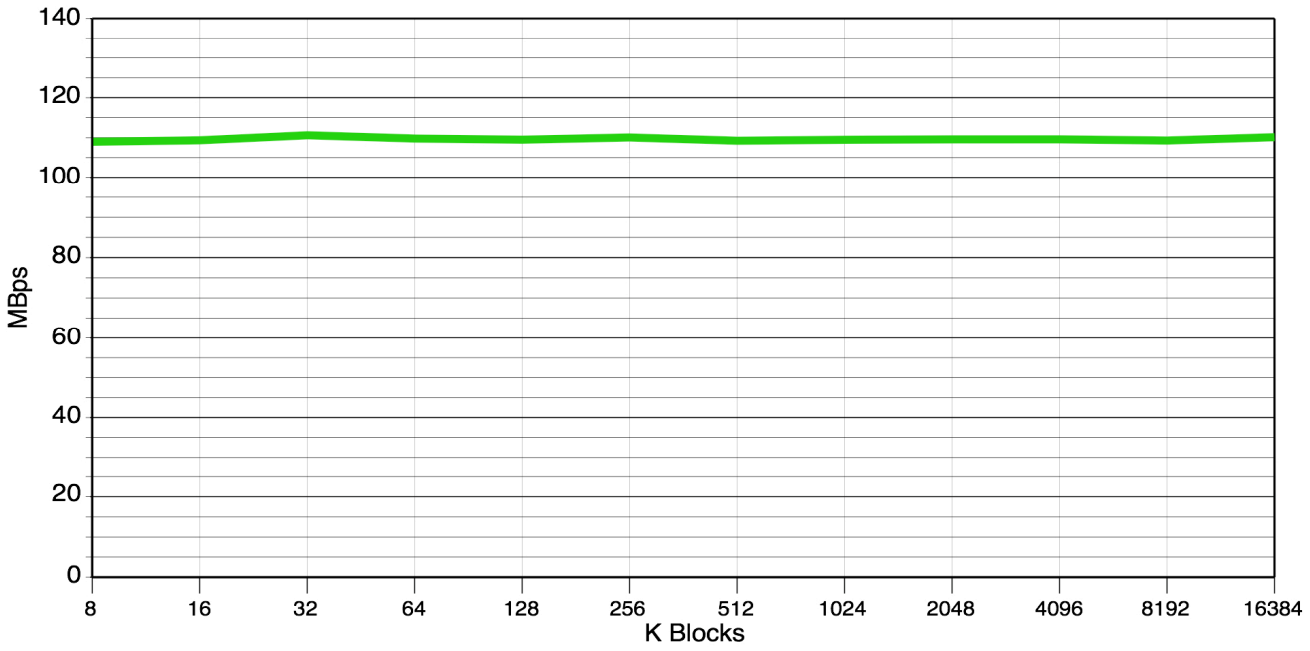
Quick Summary of Benchmarking Results

- A total of over 400 TB of data has been written without data corruption or access failures
- There have been no major hardware failures since testing began in August 2006
  - predictable and relatively uniform.
  - with some exceptions, the BlueArc aggregate performance generally scales with the number of clients
  - Recovery from injected faults was fast and relatively transparent to clients
  - 32 test cases have been prepared, about 28 of varying length have been run, all file checksums to date have been valid
- Early SLES9 NFS client problems under load, detected and corrected via kernel patch; this led to the use of this patch at Cray's AWE customer site, who experienced the same problem

Sequential Write Performance: Varying Block Size

Test Results TC01.07

[1 : 1 : 1 : 1,0 : 0,0] POS:CR:VAR:125G - SLES10 2.6.16.21-0.8 with 6x Dedicated @ 0% Spatial Utilization

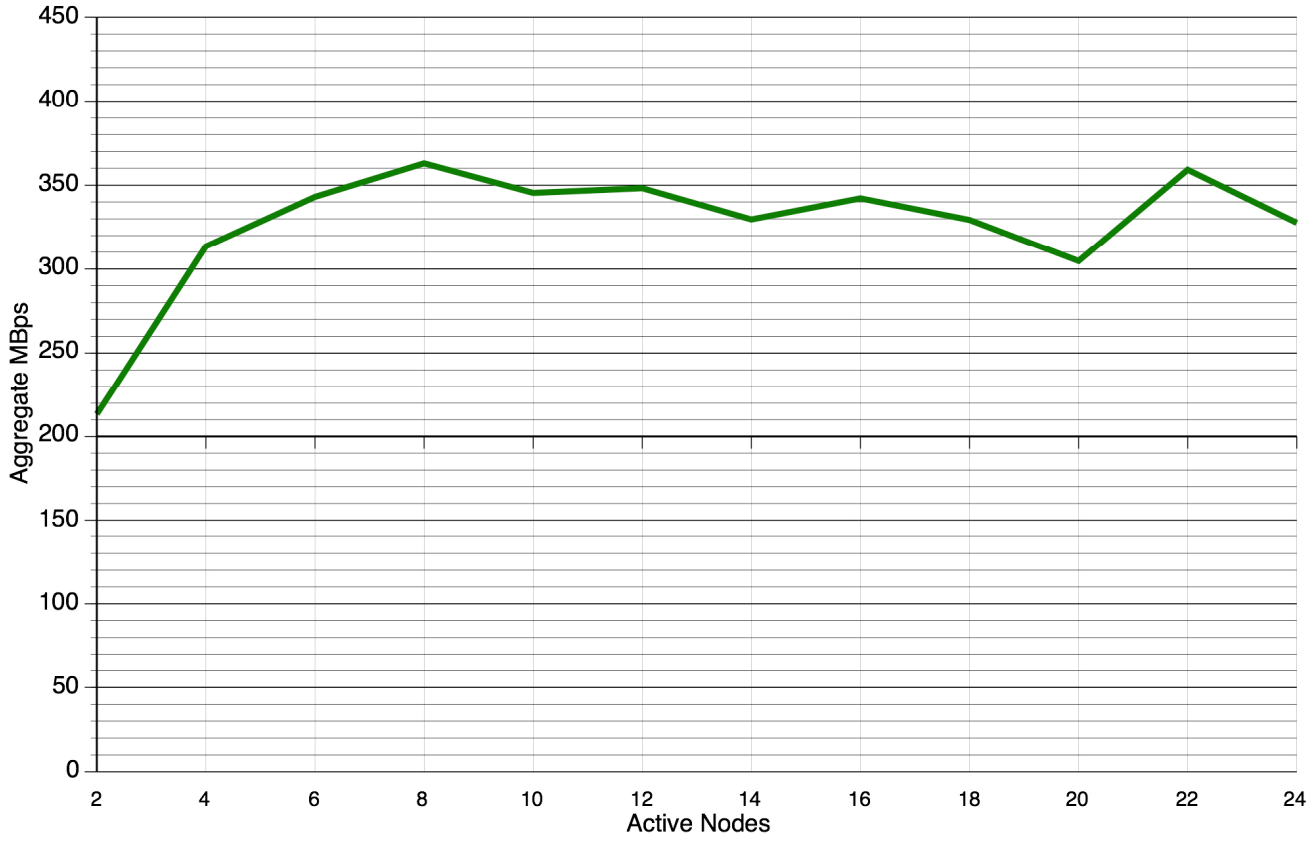


Cray Proprietary

Large File Writes: 8K Blocks

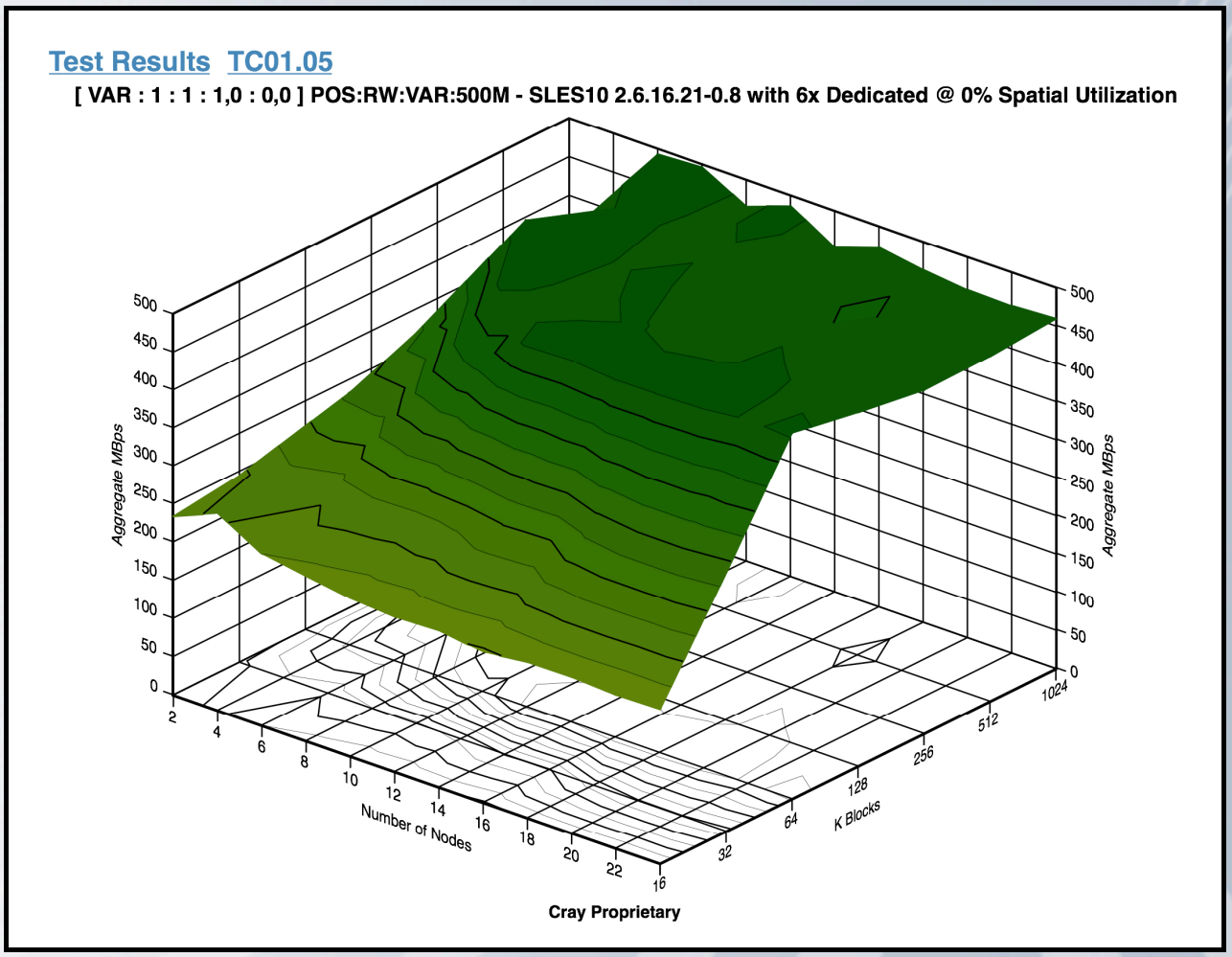
Test Results TC01.01

[VAR : 1 : 1 : 1,0 : 0,0] POS:CR:8K:500M - SLES10 2.6.16.21-0.8 with 6x Dedicated @ 0% Spatial Utilization

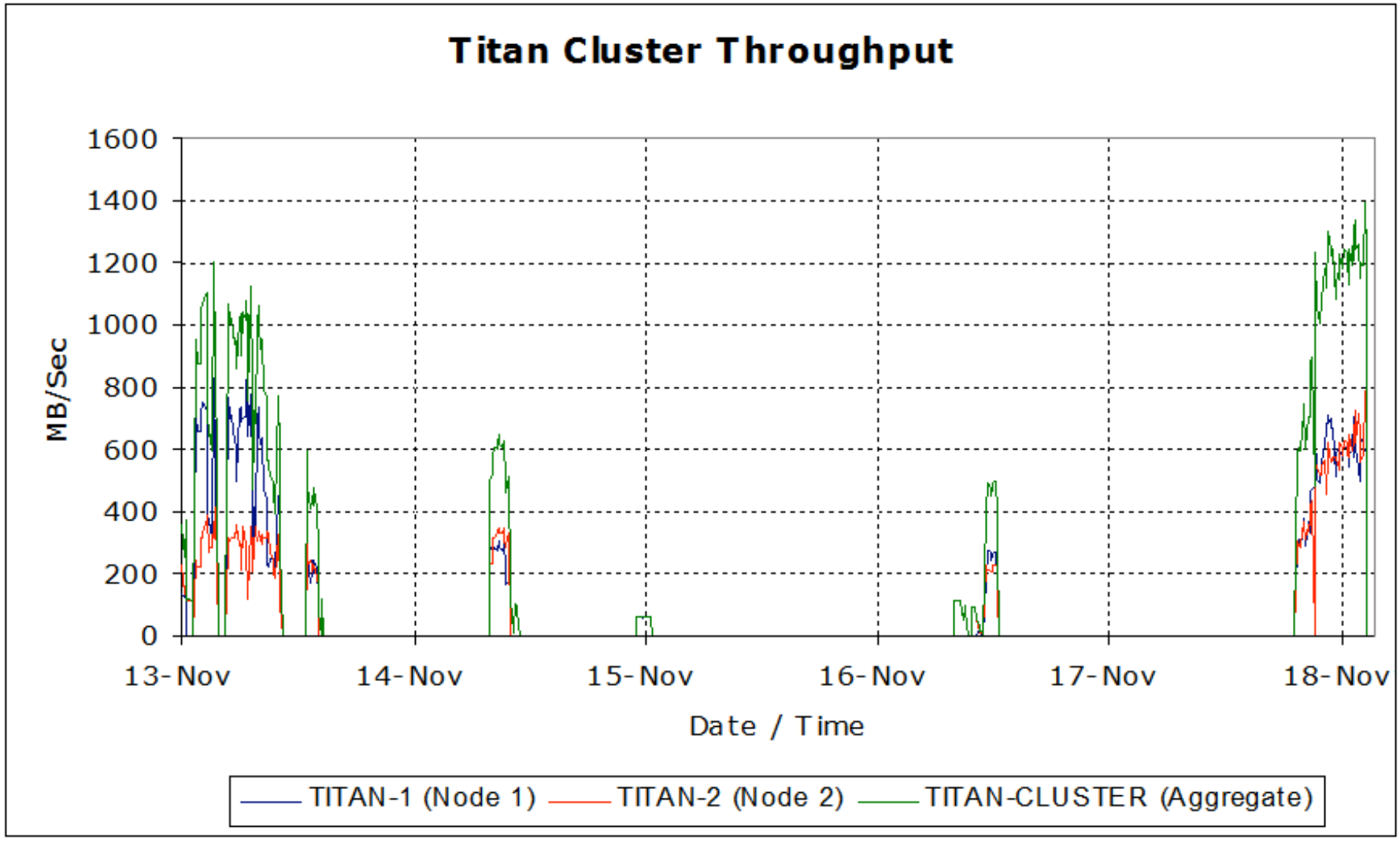


Cray Proprietary

Large File, Random Writes, Variable Sized Blocks: Performance Approaches 500 MB/second for Single Head



Titan Performance at SC06



Summary of Results

- Performance generally uniform for given load
- Very small block size combined with random access performed poorly with SLES9 client
  - Much improved performance with SLES10 client
- Like cluster file systems, NFS performance sensitive to client behavior
  - SLES9 Linux NFS client failed under Bringsel load
  - Tests completed with SLES10 client
- Cisco link aggregation reduces performance by 30% at low node counts
  - Static assignment of nodes to Ethernet links increases performance
  - This effect goes away for 100s of NFS clients

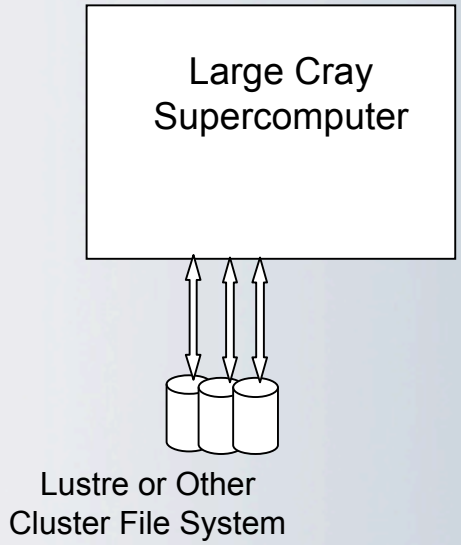
Summary of Results

- BlueArc SAN backend provides performance baseline
- The Titan NAS heads cannot deliver more performance than these storage arrays make available
  - Need sufficient storage (spindles, array controllers) to meet IOPS and bandwidth goals
  - Stripe storage for each Titan head across multiple controllers to achieve best performance
- Test your NFS client with your anticipated workload against your NFS server infrastructure to set baseline performance

Summary

- BlueArc NAS storage meets Cray goals for CASA
- Performance tuning is a continual effort
- Next big push: efficient protocols and transfers between NFS server tier and Cray platforms
- iSCSI deployments for providing network disks for login, network, and SIO nodes
- Export SAN infrastructure from BlueArc to rest of data center
- Storage Tiers: fast FC block storage, BlueArc FC and SATA, MAID

Phase 0: Cluster File System Only



- All data lands and stays

In the cluster file system

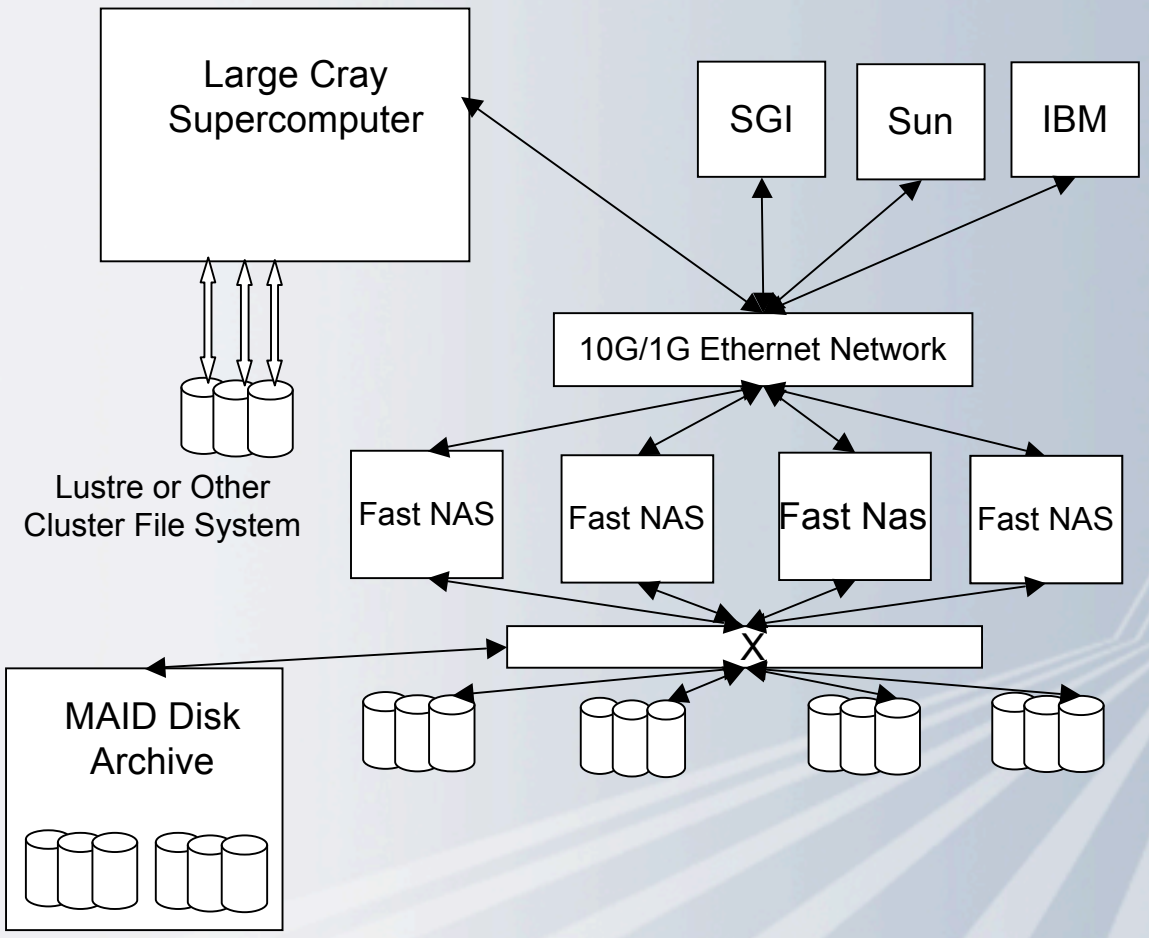
- Backup, HSM, other data

management tasks all handled

here

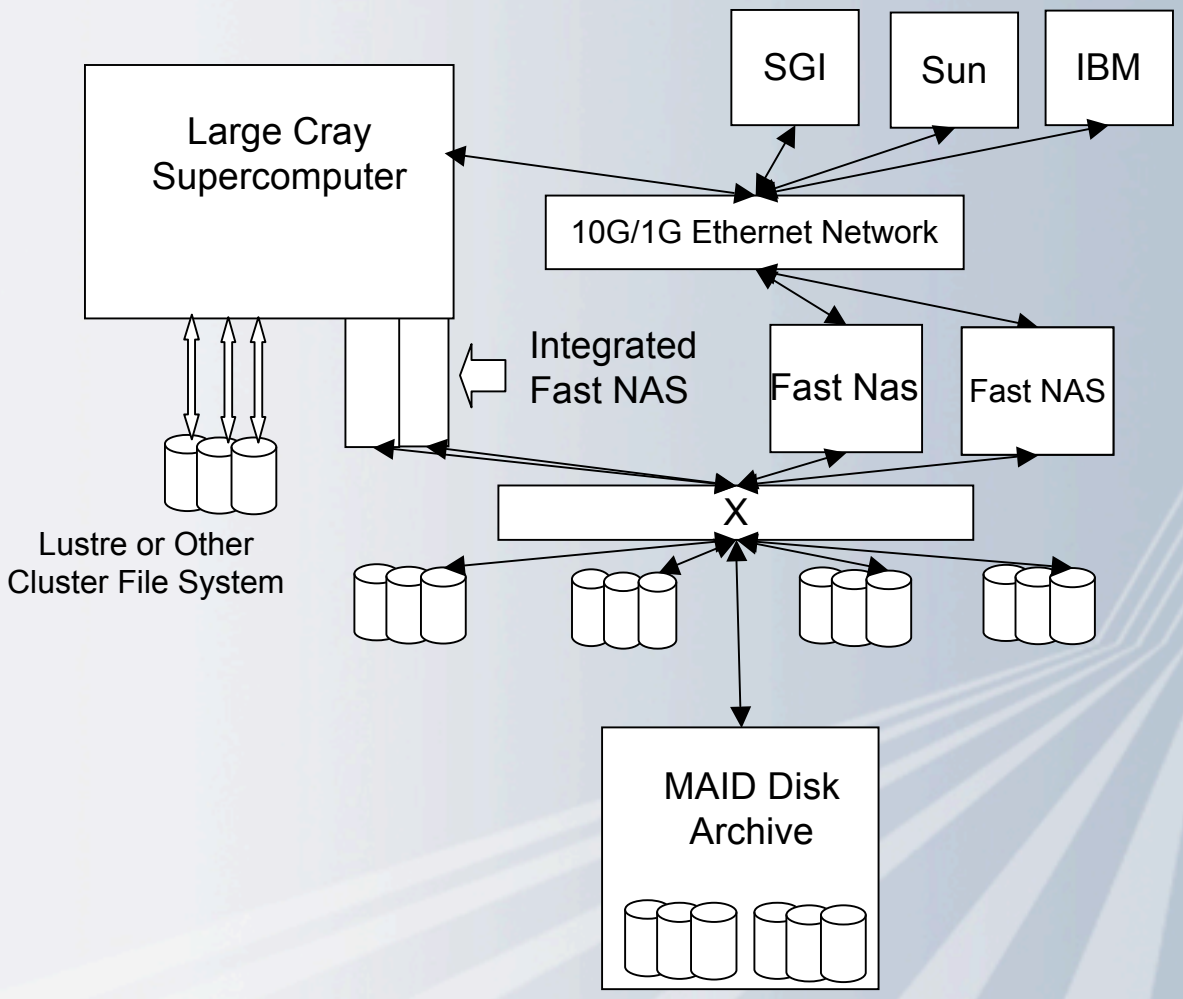
- Data sharing via file transfers

Phase 1: Cluster File System and Shared NAS



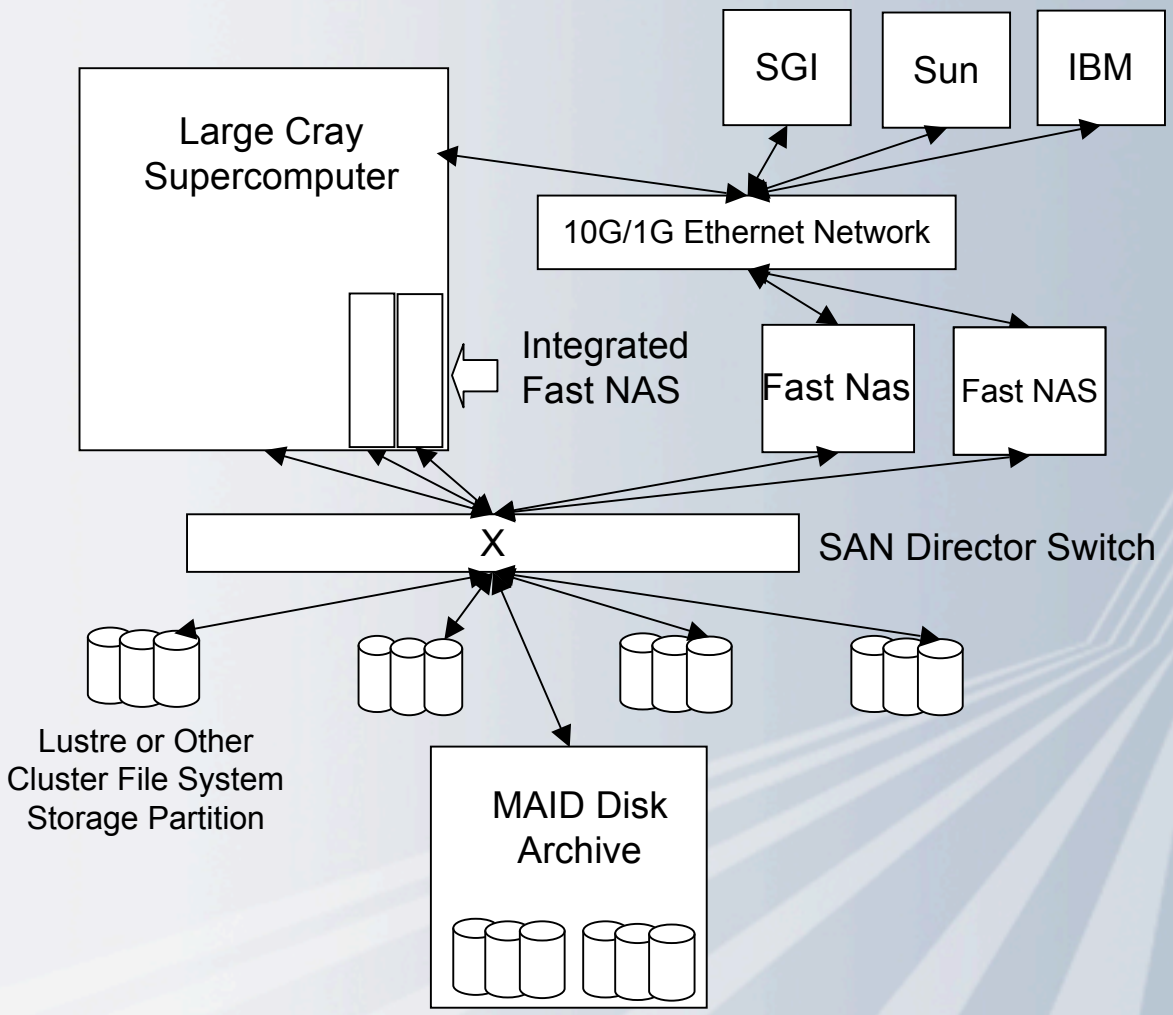
- Add NAS storage for data sharing between Cray and other machines
- NAS backup and archive support
- Long-term, managed data
- MAID for backup
- Separate storage networks for NAS and CFS stores
- GridFTP, other software, for sharing and data migration

Phase 2: Integrate NAS with Cray Platform



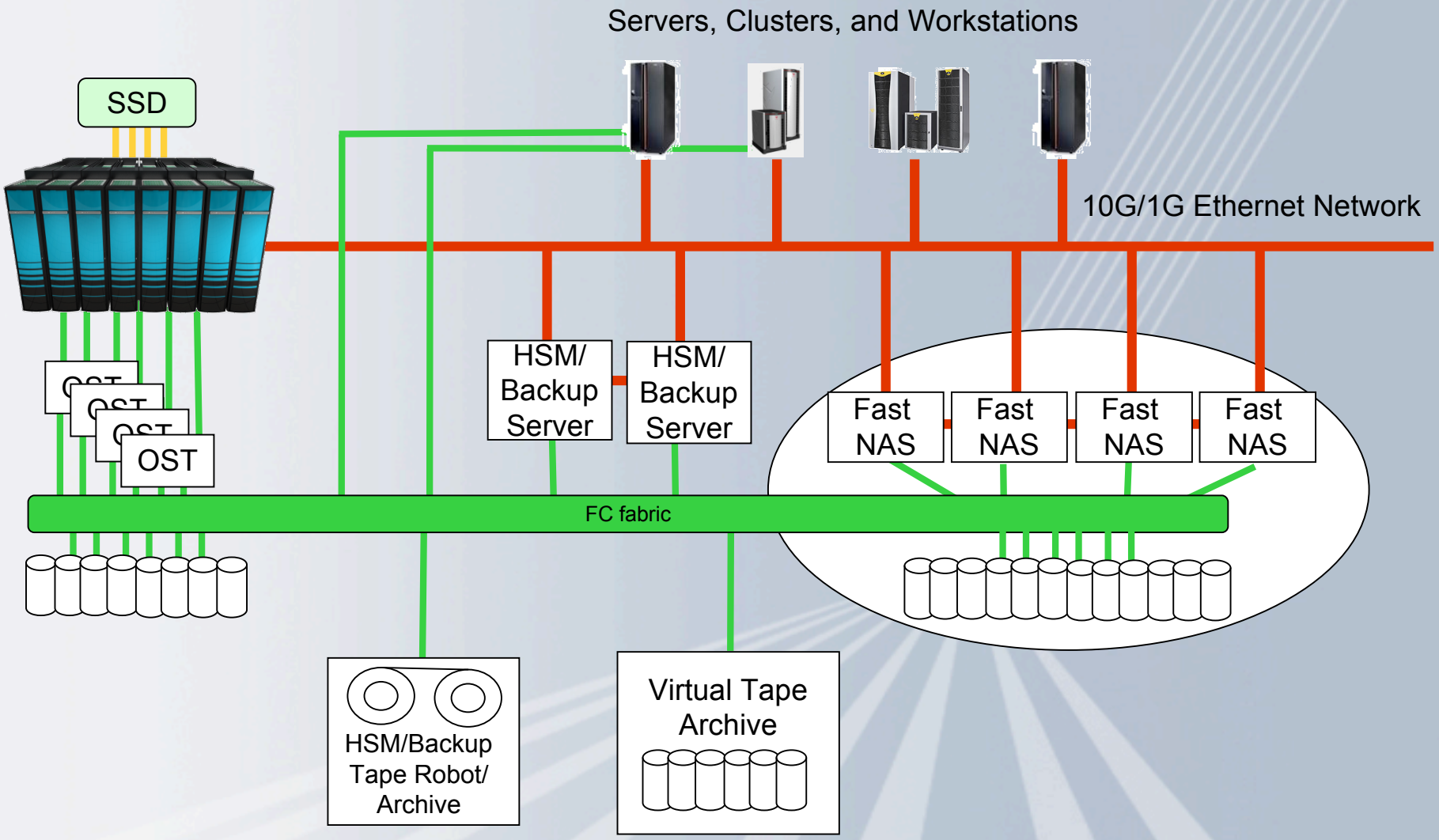
- Integrate fast NAS with Cray network: reduced NFS overhead, compute node access to shared NFS store
- Single file system name space: all NFS blades share same name space — internal and external
- MAID for backup and storage tier underneath NAS: FC versus ATA
- Separate storage networks for NAS and CFS

Phase 3: Integrated SANs



- Single, integrated Storage Area Network for improved efficiency and RAS
- Volume mirroring, snapshots, LUN management
- Partition storage freely between shared NAS store and the cluster file System
- Further integration of MAID storage tier into shared storage hierarchy

CASA 2.0 Hardware (Potential)



Questions? Comments?