

CRAY

The Supercomputer Company

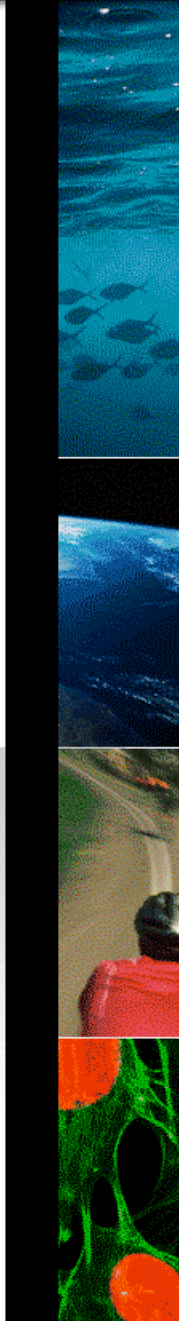
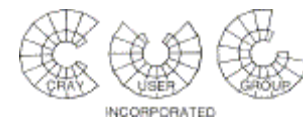
Cray's Advanced Storage Architecture CASA

Cray Users Group Meeting
May 2006

CUG 2006



This Presentation May Contain Some Preliminary Information, Subject To Change



Introduction

- Cray making progress in storage
- An important issue: our customers spend 90% or more on computing
 - Limits investments Cray and other HPC vendors can reasonably make
- In contrast, commercial customers now spend more than 50% of data center budget on storage
 - Large investments possible by storage vendors that serve this market
 - Fortunately, Cray can leverage the economics of these large markets to create scalable storage technologies for HPC

HPC Customer Storage Problems

- High-speed local storage for HPC systems
 - Very high speed, mostly used for scratch
- Site-wide data storage
 - High speed, permanent files, shared by multiple systems
- Data management for permanent files
 - Backup, HSM, archiving
- Configuration and management of storage solutions
- Interoperability with current environments

High-Speed Local Storage

- Largely used for scratch with big HPC system
- Extremely high performance for scalable applications is critical
 - Need to drive data in and out of large parallel machines
 - Used for intermediate results from large calculations
 - Also application restart dumps
- High capacities – related to total memory sizes

Site-Wide Data Storage

- Permanent files
- Data you cannot easily or cheaply regenerate
- Data needs to be accessible by multiple systems (having different operating systems)
- Good, but not extreme, performance is important
- Capacity requirements vary by application

Cluster/Shared File Systems

- Question: What role should this technology play in high performance computing storage?
- Have been marketed to HPC community as a potential single solution for most storage problems
- But there is still major work to be done:
 - Interoperability
 - Management tools
 - Integration with hardware environments
 - Stringent demands on data center operations

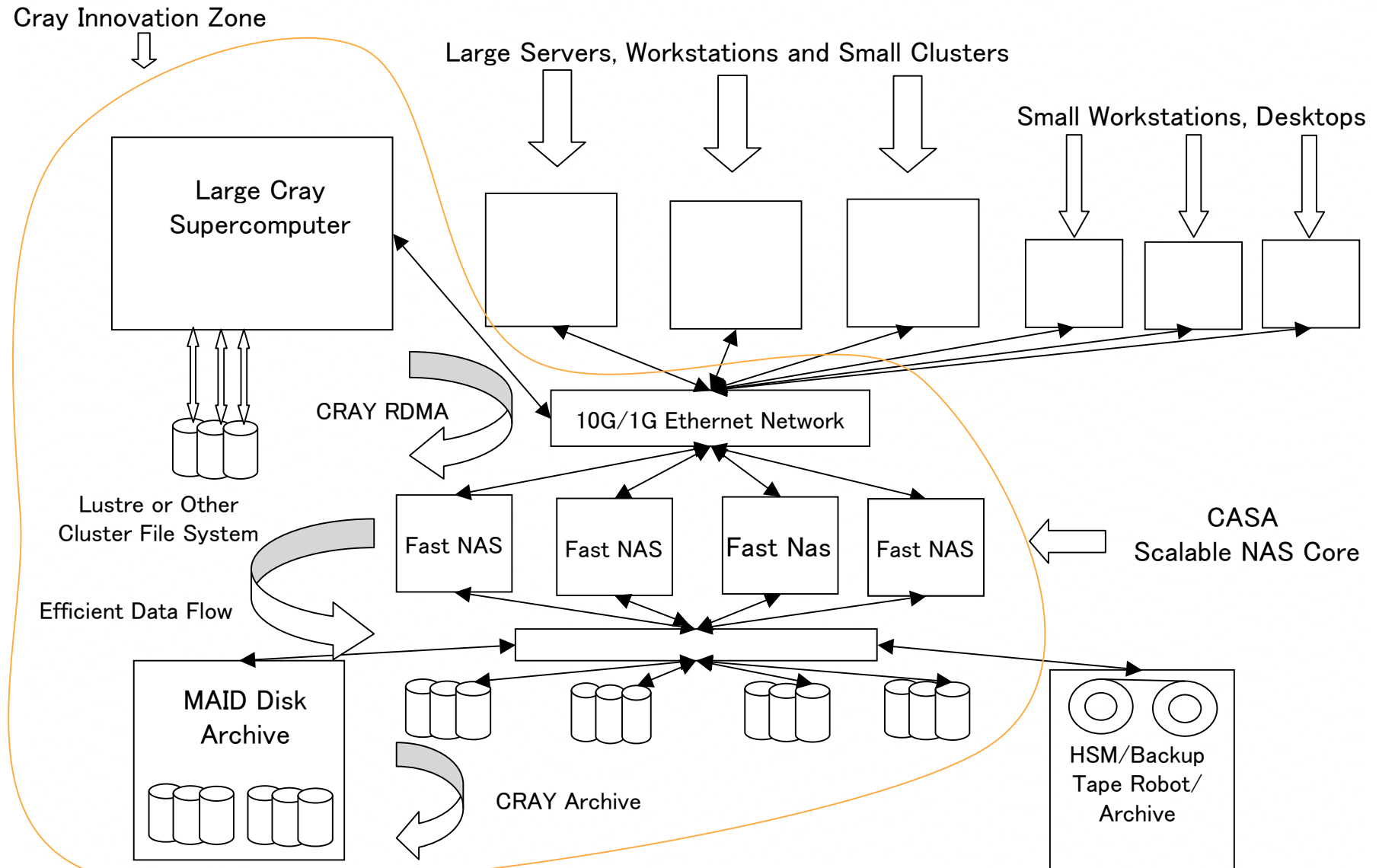
Cluster/Shared File Systems

- Need to meet a very wide range of product requirements:
 - High performance – both large and small I/Os
 - Highly scalable – numbers of client systems
 - Highly scalable – storage capacities, numbers of files
 - Heterogeneous clients – many OSs and systems
 - Full suite of data management tools
 - Excellent system admin tools
 - Open – non-proprietary software
 - Open – no hardware lock-in
 - Inexpensive – license, install, manage, service
 - Full and complete support and service
- A tough problem to solve

Cray Storage Strategy

- Instead of a search for the single solution that can solve all customers' storage problems
- Pull together a set of tools to solve these problems
 - Find products excellent at meeting key requirements
 - Focused R&D with partners to strengthen products for HPC and with Cray systems

CRAY Advanced Storage Architecture (CASA)



Why NFS?

- NFS is the basis of the NAS storage market
 - 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
 - Leverages Cray core competency: building fast switching networks

User Model

- Cray user sees multiple storage environments
- /scratch – used for highest performance for applications (cluster file system)
- /home – used for permanent files (scalable NAS; backed up)
- /archive – permanent files that will be archived or migrated
- Job scripts and applications can access any files
 - Will often want to stage files between /scratch and other environments

Administrator Model

- Sees two main environments – local scratch and shared storage
- Data management focuses on shared storage
- Options to use multiple technologies
 - FC disk
 - SATA disk
 - MAID arrays
 - Tape libraries and drives

Fast Network Attached Storage

- Partnering with Blue Arc: NAS performance leader
- **First stage:** connect Blue Arc to Cray IO nodes using standard NFS and TCP/IP connections
- **Second stage:** Accelerate TCP/IP on Cray IO nodes via TOE
- **Third stage:** Investigating transfer of NFS RPC's directly across Cray networks from compute nodes to Blue Arc IO blades (avoids TCP/IP overhead)
- Investigating disk-to-disk storage migration from Blue Arc to MAID platform

MAID (Massive Arrays of Idle Disks)

- Basic idea: power-up only a fraction of disk drives in an array
 - Increases packing density while lowering power density and related cooling requirements
 - Dramatically increases disk lifetimes (5 to 20x the standard lifetime of ATA drives)
- Virtual tape and file archive interfaces
- Other interfaces available: Cray working with Copan to exploit these

Cluster File System

- Cray has partnered with Cluster File Systems Inc. for 3 years
- Deploying Lustre across all platforms
- Working out the kinks: focusing on productization issues rather than features
 - Documentation, training
 - Coordinated joint testing on Cray hardware
 - Error logging and reporting
 - File system monitoring

Current CASA Status

- First level of integration with Cray partners
- Cray CASA testbed installed in Chippewa Falls: pilot project to benchmark, test ideas, perform storage R&D
- Deploying systems in the field with a variety of storage partners
 - Cray learning quickly and leveraging strong commercial products for HPC needs
- Investigating potential for greater integration and, where possible, leveraging product attributes to meet HPC requirements

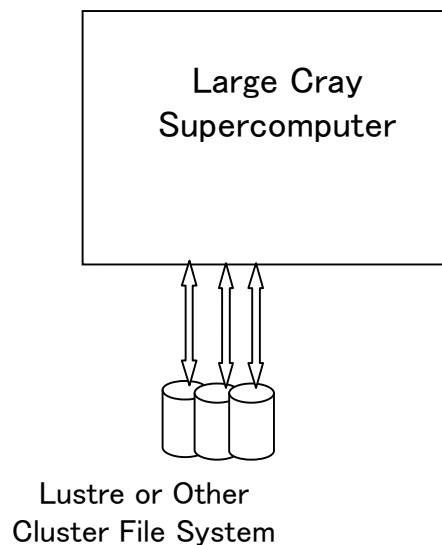
Summary

- Cray recently been putting a lot of effort into I/O recently
- Lustre is a powerful product, but is best suited for high-performance temporary storage
- Goal: offer a more comprehensive solution that uses Lustre with other products to meet a wider range of customer requirements
- CASA: A toolkit of offerings that have been checked out and integrated by Cray

Questions? Comments?

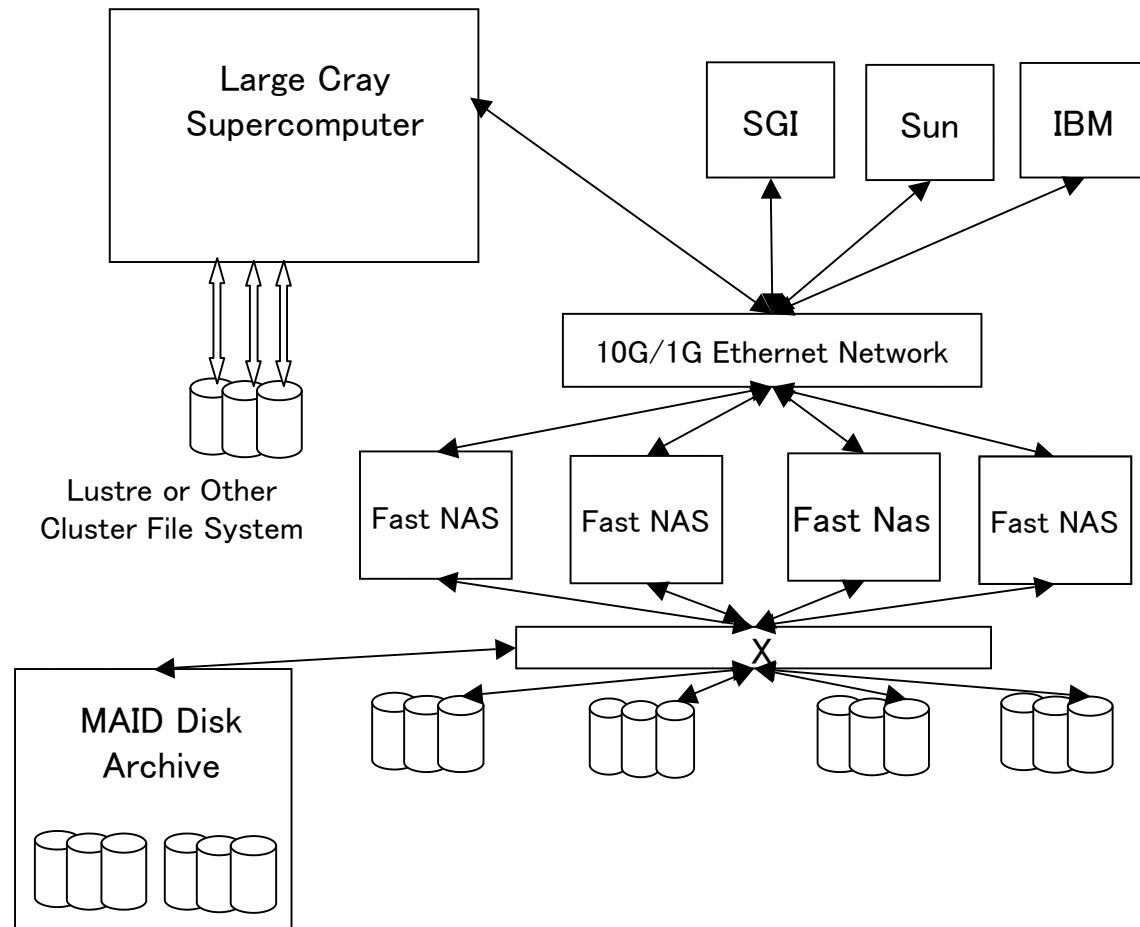


Phase 0: Cluster File System Only



- 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 protocols, for sharing