The Cray Advanced Storage Architecture

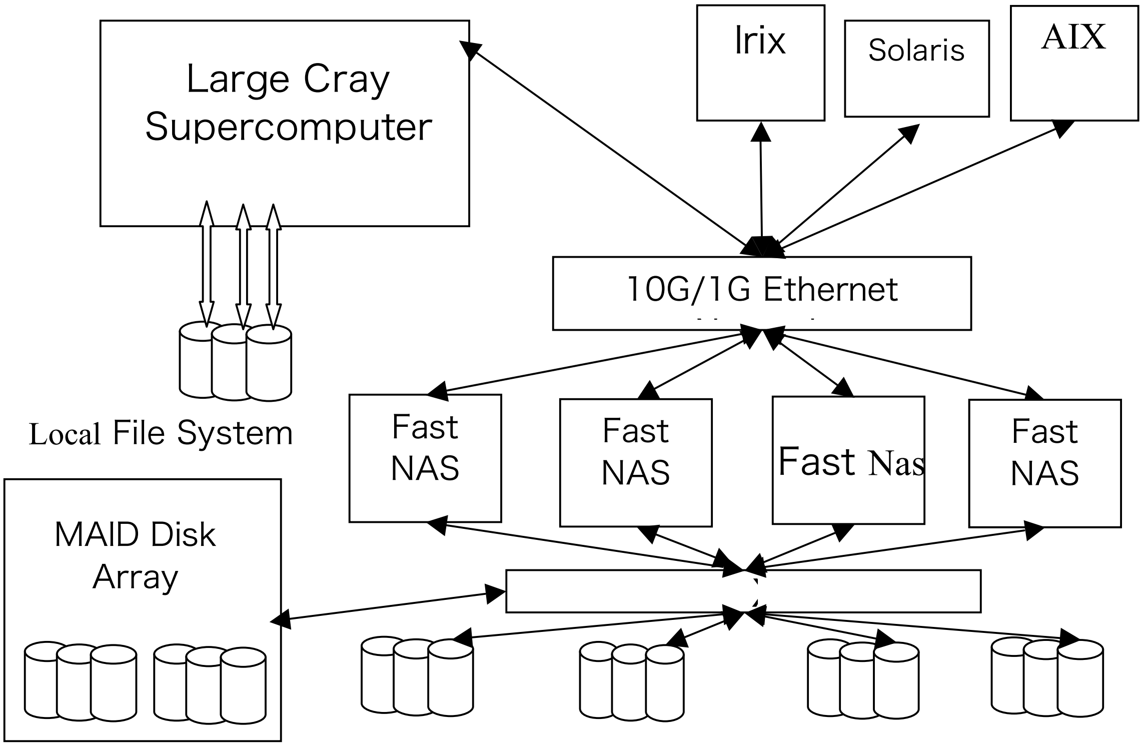
Matthew O'Keefe, *Cray Inc*.

ABSTRACT: *In this paper we outline the Cray Advanced Storage Architecture (CASA). CASA is a framework for deploying storage to meet Cray customer requirements for high speed IO transfers and efficient data management.*

KEYWORDS: Network-Attached Storage (NAS), Storage Area Network (SAN), Massive Arrays of Idle Disks (MAID).

Cray has recently focused on improving its ability to deliver fast, robust, and scalable storage systems. This strategy is based on the Cray Advanced Storage Architecture (CASA).

CASA is an integrated storage system that leverages scalable cluster file system technology for fast IO transfer speeds within our compute hardware, combined with fast, interoperable, highly-manageable network-attached storage outside our machine. The following figure shows a typical Cray CASA configuration.



CASA includes industry-leading technologies based on open standards in five key areas: disk arrays, object-based cluster file systems, network-attached storage (NAS), backup software, and disk-based storage for backup and archive. Cray's goal is to use the best technologies for each storage task:

Cray's storage architecture provides fast, scalable local scratch space and easily-managed, interoperable network attached storage. Local storage space will be provided by a cluster file system. This cluster file system runs on each Cray compute node, providing efficient parallel access to high-speed local storage. CASA balances IO requirements at the Cray compute nodes with sufficient storage capacity and bandwidth at Cray SIO blades. The SIO blades are also a conduit for transfers between the cluster file system scratch storage used for data files and checkpoint restart files, and the fast NAS storage pool used for home directories and persistent data. This NAS pool of storage can be easily managed and scaled through the use of the following features:

- The NFS protocol implemented in FPGA hardware, with multiple fast data paths to provide good performance
- A bladed, modular architecture for increased scalability and simplified upgrades
- Tiered storage to effectively deliver hybrid fast disk and capacity disk configurations that match customer requirements with the lowest-cost storage
- Advanced high-availability features for maximum system uptime
- Scalable horizontally to 8 or more clustered NAS servers exporting the same volume, with industry-leading SPECmarks performance (100K SPEC ops, 6 GigE ports — 600 Megabytes/second, up to 512 Petabytes per NAS server)
- Virtualization via virtual NFS/CIFS servers and virtual volumes
    - Over-provisioning to reduce the need for volume resizing and reconfiguration
    - Dynamic volume expansion or contraction
    - Parallel striping across FC or SATA disk arrays
    - Policy-based management for data migration between volumes (from high-speed storage to less expensive capacity-oriented storage)
- Multi-protocol support (NFS/CIFS/FTP/iSCSI/NDMPv2/v3/v4)
- File system snapshots for simplified, efficient backup
- Synchronous and asynchronous remote replication
- Disk-to-disk (including MAID) and disk-to-tape backup (via NDMP)
- Checkpoints ensure file system consistency and recovery
- NVRAM cache for fast writes without data loss; this cache can be mirrored to a partner server
- Centralized management combined with sophisticated statistics gathering and performance monitoring

Cray is creating fast paths between its compute nodes and this external NAS storage. In addition, Cray is utilizing MAID (Massive Arrays of Idle Disks) technology to provide a fast backup capability from the shared NAS storage pool. MAID storage consists of large, fractionally-powered, high-density disk arrays that can emulate tape devices, but with much higher (and consistent speeds) and without the complexity and error-prone mechanical behavior of tapes. Tapes are still used, of course, to provide off-site vaulting capability for the MAID storage.

CASA includes fast, scalable network-attached storage (NAS) to meet HPCS data management and data sharing requirements. NAS technology (based on the NFS protocol) offers sophisticated backup and management capabilities. Cray is partnering with a leading NAS provider with a roadmap to achieve very high bandwidth and IO operations per second using multiple NAS filers (from 2 to 8 or more) in a cluster.

MAID (Massive Arrays of Idle Disks) technology has now been commercialized and offers intriguing opportunities for large, low-latency secondary storage tiers. Traditionally, large HPC data archives have been kept on tape technology due to its volumetric efficiency, low power needs, transportability, and extended shelf life. MAID allows disk technology to match these tape characteristics, without the high (and highly variable) latency of tape access and complicated management requirements of tape.

Cray plans to work with a block storage array vendor to improve the bandwidth, capacity, and manageability of large block-interface-based storage systems. HPC customers require high array port counts (8 or more) and drive densities (hundreds) in a single array. In addition, large supercomputers require many storage arrays to provide sufficient IO bandwidth into and out of the system. Its imperative that integrated reliability, availability, and serviceability features be provided so that all these arrays can be managed, upgraded, and repaired in a consistent and efficient manner. Without this capability, a single array mis-configuration or failure can result in large amounts of downtime.

Acknowledgments

A variety of people at Cray have helped develop the storage architecture defined in this paper including Peter Rigsbee, who heads Cray's IO product team, and Curt Fawcett, who leads Cray's IO product integration team. John Carrier helped develop the IO networking strategy for CASA.

About the Author

Matthew O'Keefe is a storage architect at Cray Inc. From 1990 to May 2000, Matthew taught and performed research in storage systems and parallel simulation software as a professor of electrical and computer engineering at the University of Minnesota. He founded Sistina Software in May of 2000 to develop storage infrastructure software for Linux. His product teams at Sistina won numerous industry awards, including the first Linux Enterprise Achievement Award (OSDL), Best Linux Backup Product (Linux Journal), and two Best Server Software Product Awards (Linux World Expo). Sistina was acquired by Red Hat in December 2003. Matthew developed Red Hat's storage product strategy around Sistina's technology.