BROOKHAVEN NATIONAL LABORATORY

Collider-Accelerator Department

Search for a Reliable Storage Architecture For RHIC

S. Binello, R.A. Katz, J.T. Morris

Brookhaven National Laboratory, Upton, New York, USA

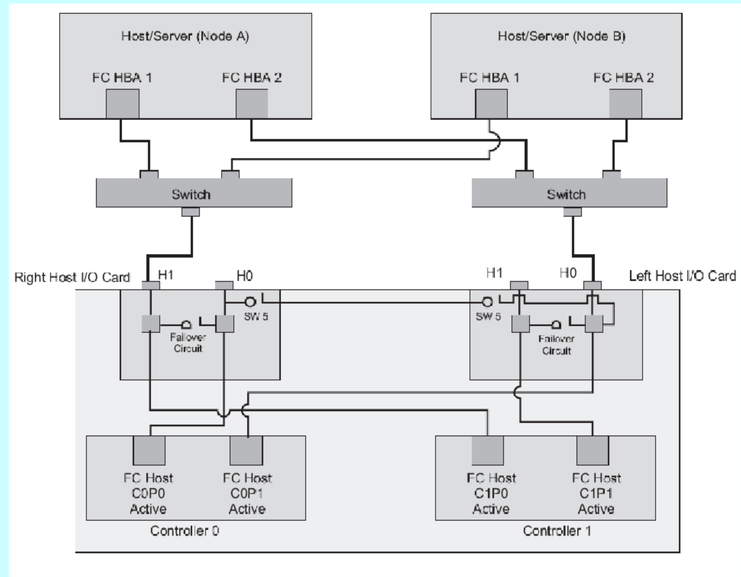
ABSTRACT:

Software used to operate the Relativistic Heavy Ion Collider (RHIC) resides on one operational RAID storage system. This storage system is also used to store data that reflects the status and recent history of accelerator operations. Failure of this system interrupts the operation of the accelerator until backup systems are brought online. In order to increase the reliability of this critical control system component, the storage system architecture has been upgraded to use Storage Area Network (SAN) technology and to introduce redundant components and redundant storage paths. This paper describes the evolution of the storage system, the contributions to reliability that each additional feature has provided, further improvements that are being considered, and real-life experience with the current system.

SYSTEM :

Storage:

- All operational components are redundant, power supplies, fans, cards.
- Fibre drives for increased reliability.
- Fibre drives connected to dual loops, i.e. can be accessed through 2 independent paths.
- RAID 5 provides protection from a drive failure.
- Spare disk available to rebuild after drive failure.
- Active/Active dual controllers provide protection in case of controller failure, and provide increased throughput.



Connections:

- Switches allow multiple host to connect to storage.
- Dual switches provide redundancy in case of a switch failure.
- Dual Host Bus Adapters (HBAs) that provide independent paths to storage.
- Failover driver that detects a link failure (hba, cable, switch) and automatically reroutes I/O to surviving link.

Software:

- Web based software to configure and monitor RAID system; Email sent when error detected.
- Host based software to configure and monitor links; Email sent when error detected.

EXPERIENCE:

- Experienced problems with RAID array from nStor/Xyratex.
  - Controller would hang but would not be failed over.
  - Ext3 file system was susceptible to corruption during these controller failures.
- Imperative to have a hot backup system and procedures in place in the event of RAID failure, or extensive file system corruption.

PLANS:

- Replace RAID array, while keeping same SAN architecture in place. Products from Dell/EMC under consideration.
- Cluster hosts to provide automatic failover at host level.
- Provide capability to mirror data in real-time.