Hall D Storage Management

Elliott Wolin and David Lawrence

16-Dec-2011

GlueX Doc #1893 covering 12GEV line 1532055

Requirements for the Hall D Computing Cluster (HDCC) online storage management (RAID) hardware system include:

1. Accept data from DAQ system at 300 MBytes/sec average rate continuously.
2. Simultaneously transfer data to JLab tape silos at 300 MBytes/sec average rate continuously.
3. Handle multiple input and output data streams.
4. Able to store minimum 72 hours of data (75 TB) if silo unavailable.
5. Extremely high reliability, implement RAID 6.
6. Minimal down time for repair operations, hot-swappable repair desirable.
7. Compatible with JLab silo and networking systems.
8. Compatible with JLab maintenance and repair contracts.

Requirements for online storage management software include:

1. Accept data from DAQ system and write to storage management hardware.
2. Automatically transfer data to JLab silo system.
3. Automatically free up space as needed after data is safely transferred to silo system.
4. Monitor and log silo input/output rates, remaining capacity, performance, etc.
5. Page system managers if problem detected.
6. Post alarm to operators if problem persists.

Hardware

We expect to accumulate about 1 TB/hour during production running, or about 25 TB/day. The storage system needs to accept this data from the DAQ system while simultaneously transferring data to the JLab silo system. Note that there may be more than one process writing to the RAID system at a time, and data may be transferred to the silo system via multiple streams

If the connection to the silo fails we need to buffer a minimum of three day's worth of data (weekend plus one business day repair time). If the silos fail for a significant length of time we may simply transfer

data to the large central file servers until they are repaired. We are considering installing multiple smaller RAID systems in the HDCC so that if one fails we can still operate at reduced capacity.

The HDCC will employ two networking systems, and either one might be used to write to the RAID system. Ethernet will be used to transfer data from the front-end ROC's to the event building system, and Infiniband will be used to transfer data between event builder nodes. The event builders will send data via either Ethernet or Infiniband to online farm nodes, which then will transfer data to an event recorder via Ethernet or Infiniband, and the choice might be different for the two cases. The event recorder will then transfer the data to files on the RAID system.

Which network will be used by the event recorder to write to the RAID system will depend on hardware availability and performance in FY14, when we purchase the RAID system, although Infiniband seems to be a good choice for all cases noted above.

Note that transfer from the RAID system to the tape silos over a long fiber must be via 10 GBit Ethernet.

There may be more than one event recorder accepting data from the farm nodes and writing to the RAID system, but no more than three or four, with some being low-rate streams (e.g. scalers and special events). We may require more than one stream transferring data to the tape silos, depending on details of the RAID system and JLab tape silo system in FY14.

Extremely high reliability and RAID6 redundancy is required since if the RAID system is down the experiment cannot take data, and lost data due to hardware failures is very costly. Repair times must be kept to an absolute minimum. It is desirable for the HDCC RAID system to be similar to those used by the Computer Center so we can share maintenance contracts and spare parts inventories. Important considerations when purchasing the RAID system are the performance degradation expected if a disk fails and recovery/rebuild time when replacing the bad disk.

Software

Data will be written to the RAID system by the event recorder, a component of the CODA DAQ system developed by the JLab DAQ group. Standard Unix system calls will be used. Depending on the nature of the RAID hardware and details of the DAQ system there may be more than one event recorder running, but this is not expected to be a problem.

Details of the software needed to transfer data from the RAID system to the tape silos and free up space on the RAID system depends on the actual hardware purchased and the JLab tape silo hardware and software installed at the time. This software will be written by the Hall D computer manager, a position we will fill in mid-2012. The software will be written in close cooperation with JLab Computer Center personnel.

The HPCC computer manager will further need to develop a system to:

- monitor RAID performance and log RAID performance data
- generate early warnings for incipient problems
- send alarm to the Hall D operator alarm system if problem occurs

Manpower Estimates

The 12GeV schedule allocates 11.3 man-weeks for planning and 11.5 man-weeks for writing the storage management system software. The amount allocated for planning seems excessive, less than half would be more than enough, although the allocation for writing seems adequate. Separate lines in the schedule cover the actual specification, purchase and installation of the RAID system. The schedule further includes 12 man-weeks for check-out of the storage management system, which also seems adequate.

Finally, this report completes the planning stage for Hall D Storage Management.