From: **Chip Watson** watson@jlab.org
Subject: Fwd: GlueX DAQ performance (file server and beyond) -- how to reach 1.6 GB/s
Date: August 9, 2016 at 10:10 PM
To: Curtis A. Meyer cmeyer@cmu.edu

WC

Curtis,

Meant to cc you on this...

Chip

-------- Forwarded Message --------
Subject: GlueX DAQ performance (file server and beyond) -- how to reach 1.6 GB/s
Date: Tue, 9 Aug 2016 21:06:18 -0400
From: Chip Watson <watson@jlab.org>
To: Mark Ito <marki@jlab.org>, David Lawrence <davidl@jlab.org>, Sandy Philpott <philpott@jlab.org>, Christopher Larrieu <larrieu@jlab.org>, Graham Heyes <heyes@jlab.org>, Sandy Philpott <philpott@jlab.org>

Mark, David and Colin,

For this Fall, you will have 2 changes related to writing GlueX data to tape compared with last Spring: (1) You will have access to dual 10g Ethernet links, and (2) you will have a new higher performance file server. This email describes how to get the most from these resources.

Feel free to post the following information to a wider audience. I'm available for additional interaction on these points as you find useful.

regards,

Chip

Scicomp operates (currently) 2 DAQ gateways through which you can NFS mount the Lustre filesystem (which is how you push data to CEBAF Center). Bryan Hess has put into place a routing entry such that one of these is visible on the first 10g link, and the other is visible on the second 10g link. Each of the gateways can accept >800 MB/s, so that you should be able to acquire data at 1.6 GB/s, perhaps more. We will bring up a third gateway in September so that you are 2+1 fault tolerant.

The DAQ gateway lets you write data to the Lustre filesystem, which in practice means it is writing the data to a single RAID stripe (RAID z2 8+2, similar to RAID-6, tolerating up to 2 failed disks). Our more recent file servers can do read or write at 800 MB/s to a single RAID set . In observing the 800 MB/s performance of the transfer last Spring, it means that you were fortunate to land on one of the higher performing nodes. Other nodes could be as low as 30% of this figure (roughly random placement). Also, if the node and RAID set (stripe) you land on is busy with another large transfer, you would share this bandwidth and so get only 120 MB/s to 400 MB/s.

Lustre supports higher single file throughput through striping a file across multiple RAID sets (round robin). I recommend a stripe size of 2, so that you would see bandwidth of 500 - 1600 MB/s on a pair of idle file servers/stripes. We are slowly decommissioning the slower file servers, so this will improve to 900-1600. Some fraction of the time (i.e. landing on a busy stripe), you would see that fall to 450 - 800. Increasing the stripe size to 4 (for example) doubles the likelihood of hitting a busy server, so is of limited utility, however we can later experiment.

One DAQ gateway can transfer multiple files at the same time, so that even if one file is running at only 400 MB/s, two files might manage 800 MB/s and thus better fill the 10g link.

Two DAQ gateways each transferring 2 files at the same time thus has a high likelihood of providing >= 1.6 GB / s aggregate bandwidth.

Takeaway #1: to achieve high bandwidth, you need to have a system which will transfer up to 4 files to 2 gateways at the same time. This means that the current cron job launched transfer script will need to be modified -- we'll do that some time in September; earlier if time allows.

The next challenge is to make sure you have enough bandwidth in the counting house. Your new file server has 2 RAID controllers and a 40 disk array, so 4 stripes of 8+2. The file server's controllers each have 2 cables of 4 lanes of 12 Gbps SAS-3, thus at the controller level, you have lots of SAS3 cable bandwidth (the disks have FIFO buffers), Each controller will control 2 8+2 stripes. Each stripe can handle ~ 800 MB/s, single file, with no stripe competition. So to have a throughput of 1.6 GB/s, you actually need to have 2 files going in and 2 files going out at the same time in the counting house. You should even be OK with 2 files going in and up to 4 going out (to deal with the fact that Lustre in CEBAF Center has many competing file clients). In other words, 2 stripes are receiving data from DAQ, and 2 stripes are serving data to Lustre.

It might be possible to stripe this server's disks (to double stripe size and bandwidth) so that one double stripe is receiving data, and one double stripe is serving out data. We'll need to play with this. Alternatively, if Graham et al get the multi-event-builder scenario running in production, this striping would not be necessary (2 files created simultaneously, each with half of the events).

The 8 TB disks in your system can sustain 200 MB/s/disk for the outermost disk cylinders, and maybe 100 MB/s for the innermost cylinder. At the outer edge, an 8+2 stripe of such disks theoretically run at 1600 MB/s, but there are overheads which degrade this by more than a third. It will probably be necessary to avoid writing more than half of the disk to maximize the disk performance in the presence of various overheads.

Takeaway #2: don't use this disk server to hold as much data as possible for local analysis; delete data early and periodically erase the stripe to remove fragmentation

Takeaway #3: don't use this disk server as a source of high rate online analysis (do that from memory before writing the files)

(p.s. we will help you get your file server configured if you like once our new LQCD resources are in production)

With 4*8 data drives (removing RAID) each 8 TB, you have 256 TB, which formats down to ~230 TB in the file system. Using the outer half of these disks means you have 115 TB for high throughput mode. At 1.6 GB/s, this is a 20 hour buffer. If you intentionally get data transferred to CEBAF Center within 1 hour, then this server gives you a 40 hour buffer at this rate, since if you are not reading the data off the disk, you can write into the inner cylinders at this rate. To meet the "contractual" requirements of a 72 hour buffer, you will need a second file server to run at this rate.

Takeaway #4: you can only run steady state at 1.6 GB/s while the transfers to Lustre are working most of each day

Takeaway #5: if you DON'T transfer data out, you can run at something of order 3 GB/s for up to ~20 hours, and even higher for a few hours starting from an empty disk; this could be useful this Fall to SAMPLE higher luminosity running in GlueX

Takeaway #6: keep the ability to fall back to the older RAID servers to achieve the mandated 72 hours at some reasonably high data rates

For a few days of high rate running, Lustre could be a buffer. If you want to run in this mode for weeks at a time, additional provisioning will be necessary.

In the absence of highly reliable / mature level 3 trigger software, it would be possible to sustain writing 1.6 GB/s to LTO-6 tape ONLY if IT buys additional LTO-4 tape drives. Each tape drive can only sustain writing 140 MB/s of incompressible data, and costs (with a computer) about $8K. Adding an extra 8 drives to make up for this very high rate (1.2 GB/s above the originally planned 400 MB/s) is feasible, but not free. IT currently plans to add 2 drives, so adding 6 more means an extra $50K more or less.

Lustre could also be used as a rate buffer, where we could intend to sustain 800 MB/s to tape, buffering the rest on disk, averaged over many days. When beam is off or of low quality, the drives could catch up with acquired data if the duty factor were 50%. Physics could buy one extra file server for the farm which could be used for this purpose when needed ($30K).

Running at these high rates to tape costs $12K / PByte ($1.6K / day) per copy; however, if the data could be filtered offline and only 20% of the data later kept, then the tapes could be immediately reused and this cost vanishes in the same fiscal year (2 copies at 20% would thus cost $640/day, $20K/month -- not a serious cost)

Takeaway #7: running for weeks at 1.6 GB/s with no level 3 trigger will require a modest amount of I/O infrastructure growth

Takeaway #8: tape cost is not a driver for storing before level 3 filtering if the filtering is done within a few months, although it does require more tape drives than filtering first would require

Unless we hear otherwise from you, we will assume that you are not expecting IT to provision to steady state 1.6 GB/s running. 800 MB/s is fine with no additional hardware. Please make sure you update your computing requirements as you adopt changes in your running model. Thanks.