Scientific Computing Operations for ENP

June 2024 Updates

Bryan Hess

Wednesday, June 5, 2024

Jefferson Lab

U.S. DEPARTMENT OF ENERGY | Office of Science

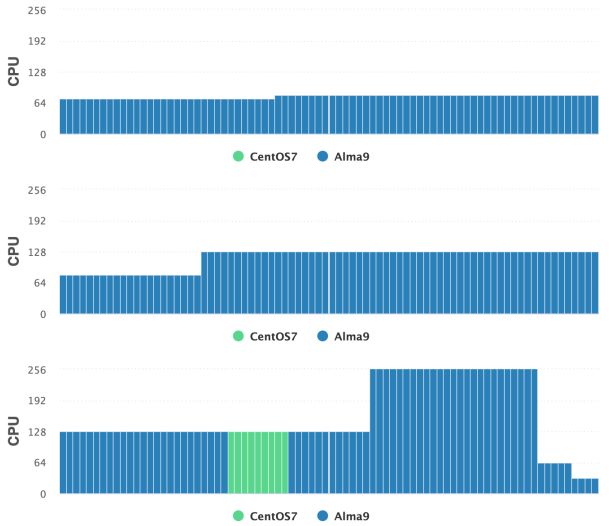
JSA

Lustre Outage June 5, 2024

- Around 11:45am, Lustre (/cache and /volatile) went offline when four servers (two fail over pairs) began a reboot loop.
- Other servers were unaffected, but at least one server in each pair is required for the filesystem to be up and stable.
- The affected servers had not seen any configuration change or rpm updates in months since they are end of life.
- The stack trace of the kernel thread hangs pointed to the InfiniBand Lustre Networking (LNET) layer.
- The bug appeared to be caused by client traffic, but the source of the “poison pill” was not clear. Eventually, we shutdown the entire farm and were able to stabilize the system.
- Today (June 6) We are bringing up farm nodes in batches of common hardware and software configurations to see if a particular event will trigger the bug.
- This is an ongoing debugging process.

Farm Transition to EL9 (AlmaLinux 9)

- Transition to Alma9 is 96% done. All GPU nodes are running Alma9.
- Nine farm19 nodes remain on EL7.
  - Users running on El7 are being contacted individually..
- New ifarm machines have arrived on site and will be online in the coming week.
- On Maintenance day, June 18th, the DNS alias ifarm.jlab.org will connect to EL9 machines.
  - Ifarm9 alias goes away (on July Maintenance day)
  - Ifarm.jlab.org points to the two new servers
  - Keep the same ssh keys for the hosts
  - Will leave old ifarm nodes for another month
- OSG servers at JLab will be completely on EL9 in June.



Lustre Transition plans for this summer (already started)

- Reminder: Old Lustre19 is 4.7PB. New Lustre24 is 11.2PB, soon to be ~14PB.
- First, we will transition /cache areas:
  - All current cache reads from tape are being copied to both /lustre19 and /lustre24
  - This will fully populate the /cache area in coming weeks.
  - Some /cache areas that have slower churn will be read back from tape manually
  - /cache files that have not been committed to tape will be synced manually.
  - Once the two /cache areas are largely in sync, we will schedule a maintenance window (per hall) to do a final sync pass and switch /cache areas to point to the new filesystem.
- Second, we will transition /volatile areas:
  - These files are not on tape, so they must be synced manually
  - Once we have the bulk copied we will schedule a maintenance window (per hall) and make the cutover.
- Once the /volatile cutover is complete, we will make the old areas read-only
- Lustre19 will be dismounted and shut off on the following maintenance day.

Proposed Adjustments to Maintain Fairshare

- Last month it was noted that one Slurm accounting group was far exceeding its fair share despite having the lowest priority in the production queue.
- How did this happen?
  - The low-priority jobs did the following:
    - Jobs were submitted to the production partition from multiple users
    - This group of users almost always had jobs queued
    - Jobs were submitted with a 4 day wall clock time (the maximum allowed) and they ran for most of that time
  - The higher priority / larger Fair Share jobs had this pattern:
    - Jobs ran for a shorter time
    - Jobs were submitted in batches, but there were pauses between submissions
  - The pauses between completion and the next submission provided a chance for the lower priority/long running jobs to rise to the top of the queue and run
- Proposed Maintenance Day Change to fix this (for discussion)
  - 24 hour limit on production partition
  - Create a QOS policy to allow longer run times with preemption.
    - Consequence: checkpointing will be important

Proposal: A return to the Read-Only Cache

- We have operated a write-through cache (/cache) since ~2018.
- It has proven to be problematic in several ways
  - Has many opportunities for tape and disk to get out of sync (file names re-written, permissions mismatches, accidental re-writes of the same files that are not found until they cannot be written to tape)
  - Collects “small files” that cannot go to tape and are never deleted
- The former read-only cache had advantages
  - Simpler, more understandable quota enforcement
  - Easier to grow/shrink as needed
  - One-to-one mapping between /cache and tape
  - Can be integrated with other systems (e.g. Rucio) because it is system managed
  - Makes file storage to tape more explicit
- Proposed Change for read-only /cache (for Discussion – target late summer/Autumn)
  - Announce the change, document workflow examples for read-only /cache
  - On a maintenance day, change /cache to Read-Only.
  - Allow writes to /cache only by the tape system
  - Adjust SWIF to ensure that files destined for tape land on /cache immediately.
  - Need to understand non-SWIF workflows (e.g. – write to /volatile, then jput)

Code.jlab.org (gitlab) and OpenShift (Kubernetes)

- Steps before full production
  - OpenShift being installed (needed for CI/CD)
  - Gitlab user on-boarding hook development nearly done
  - Backup/Restore testing to be completed (Cohesity/Ceph integration)

- Next Steps
  - CI/CD usage documentation, templates

July/August Maintenance Look Ahead

- All login gateways are now equivalent (login.jlab.org, scilogin, hallgw). Proposed Timeline for merging them all into login1-4:
  - JUNE MD:
    - 4 new RHEL9 VMs for login1-4.
    - CNAMEs for hallgw and scilogin for compatibility/migration time.
    - Announce Change
  - JULY/AUG MD:
    - Remove CNAMEs for hallgw and scilogin.
- Slurm Upgrade for the Farm Likely July or August, but TBD