SciOps + ENP November 2022

Status Updates

Bryan Hess

Thursday, November 3, 2022

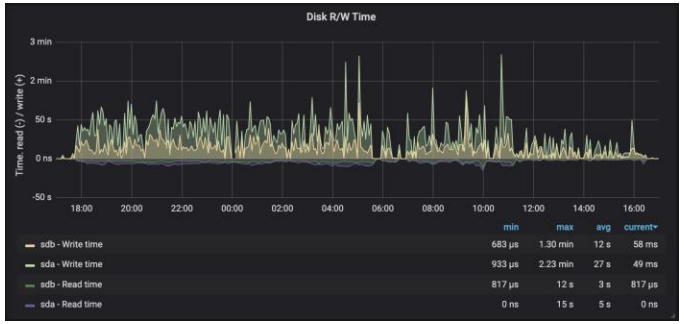
Jefferson Lab

U.S. DEPARTMENT OF ENERGY | Office of Science

JSA

Farm Swapping and the Fix

- Prior to 2020
  - farm jobs that could run away and consume all system memory
  - The Linux System OOM killer can be arbitrary
  - OOM caused system and Lustre stability issues



- As a fix, we implemented the Linux CGROUP OOM for Slurm
  - The immediate goal was to keep systems up and stable
  - because we had never enforced memory limits we were conservative, and set the hard limit to 150% of the slurm memory request
  - This solved the problem and we moved on with minimal disruption to job submission
- 2022: Swapping observed on some nodes, slowing all jobs
  - This means jobs are exceeding their slurm memory request and running the nodes short of memory
  - As part of November Maintenance, we will limit jobs to their actual memory request
  - This will cause some OOM killed jobs at first
  - This will prevent nodes from swapping and make job wall times more even

Cache Disk Management

- Hall A data rates have increased, which affects disk use accordingly
  - During a sample 2week period in October, Hall A processed 2.2PB using SWIF
  - We are monitoring cache disk pressure (file lifetime, throughput) to avoid chokepoints
  - In the coming months the impact will become more clear
  - No big changes in the short term
- Longer term
  - This will inform disk purchase and allocation
  - We are planning for a storage expansion in this fiscal year

Farm Hardware Changes

- Firmware upgraded on farm19 chassis
  - A power management bug would down-clock some CPUs
  - Nodes have been reserved and re-flashed, four at a time.
- Farm 13 node phase out
  - The farm 13 nodes will be decommissioned in the coming weeks
  - They will be repurposed as a testbed for farm development
  - They represent <1% of the current farm compute capacity
- Farm 23 nodes
  - Awaiting delivery. Supply chain continues to be problematic.
  - Farm23 "Milan" nodes benchmarked at just under 2x the performance per core of the farm19 era "Rome" nodes.
  - New nodes include more local scratch (>10TB) and faster ethernet and IB (10GigE, EDR)

Farm Software Changes

- The next Farm OS it targeted for Rocky 8
  - 1.5 years ofCentOS 7 support remain
  - We will begin that project soon. Nothing user-visible before the new year.
  - Slow roll out with test queues first
  - Rocky 9 may be too far to go in one step; Key software may not be mature on 9 soon enough.
  - Rocky 8 keeps us close to RHEL8 where many deskstops are
- Containers
  - We strongly encourage running jobs in Apptainer (Singularity) containers
  - Containers decouple the farm OS from the application OS
  - Containers are the norm for OSG, NERSC, and most remote processing
- The farm is now routinely patched, including kernel patches
  - Slurm feature is used to reserve/patch/reboot nodes
  - Can be done opportunistically or more aggressively
  - Patches support bug fixes and security
  - For tight configuration control, run in containers

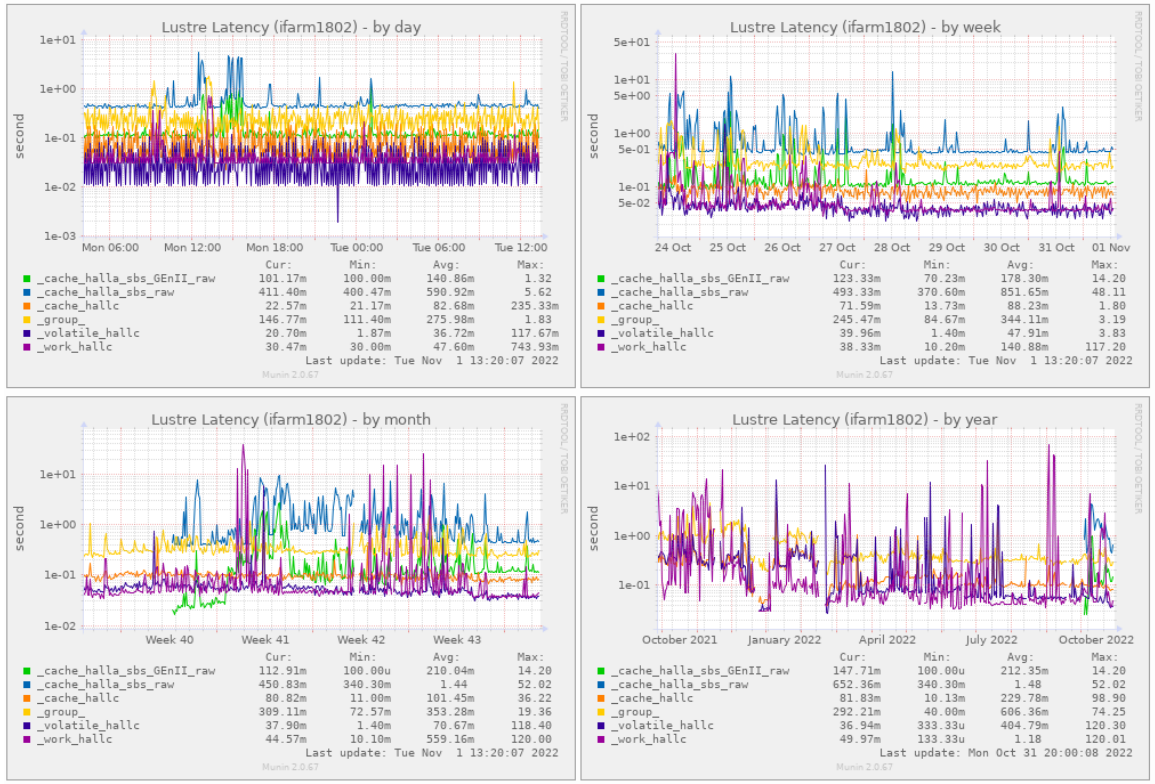
Wide Area Network Upgrade: 100Gbit for JLab

- ESNet is the DOE Office of Science provider for the National Labs
  - Specialized focus: High-throughput, lossless networks for science
  - Tuned for large flows
  - Advanced capability for building virtual circuits/ overlays (e.g. LHCONE L3VPN)
- We have prepared for 100Gbit service, and the final step is in progress now.
- Dark fiber from ESNet to JLab will be in place in the next 2-3 months.
- We will have two diverse 100Gbit paths from the lab to ESNet
- The scientific computing networks, including the routers, Science DMZ and Data Transfer Nodes (DTNs) are already at 100Gbit, so this is the missing link.
- By using dark fiber instead of circuits from Telcos, we can work with ESNet to swap out equipment and upgrade to 400Gbit or do wavelength multiplexing.
- This positions us well for data intensive projects
- We already have the need. GlueX processing at NERSC routinely caps out the existing 2x10Gbit links.

Storage Issue Tracking

- Oct 19 – 24 Work file server incident
  - Four work file servers crashed overnight each night
  - One was LQCD, so not farm related
  - Case open with RedHat. Initially points to a udev bug. Could be IB stack too.
  - Resolution was a kernel patch, but exact trigger not known yet

- Lustre high latency metadata operations
  - Ongoing issue; example is `ls -l` on a large directory
  - History at right from Brad
  - Getting file size requires going to storage server, scatter/gather. Not just MDS.

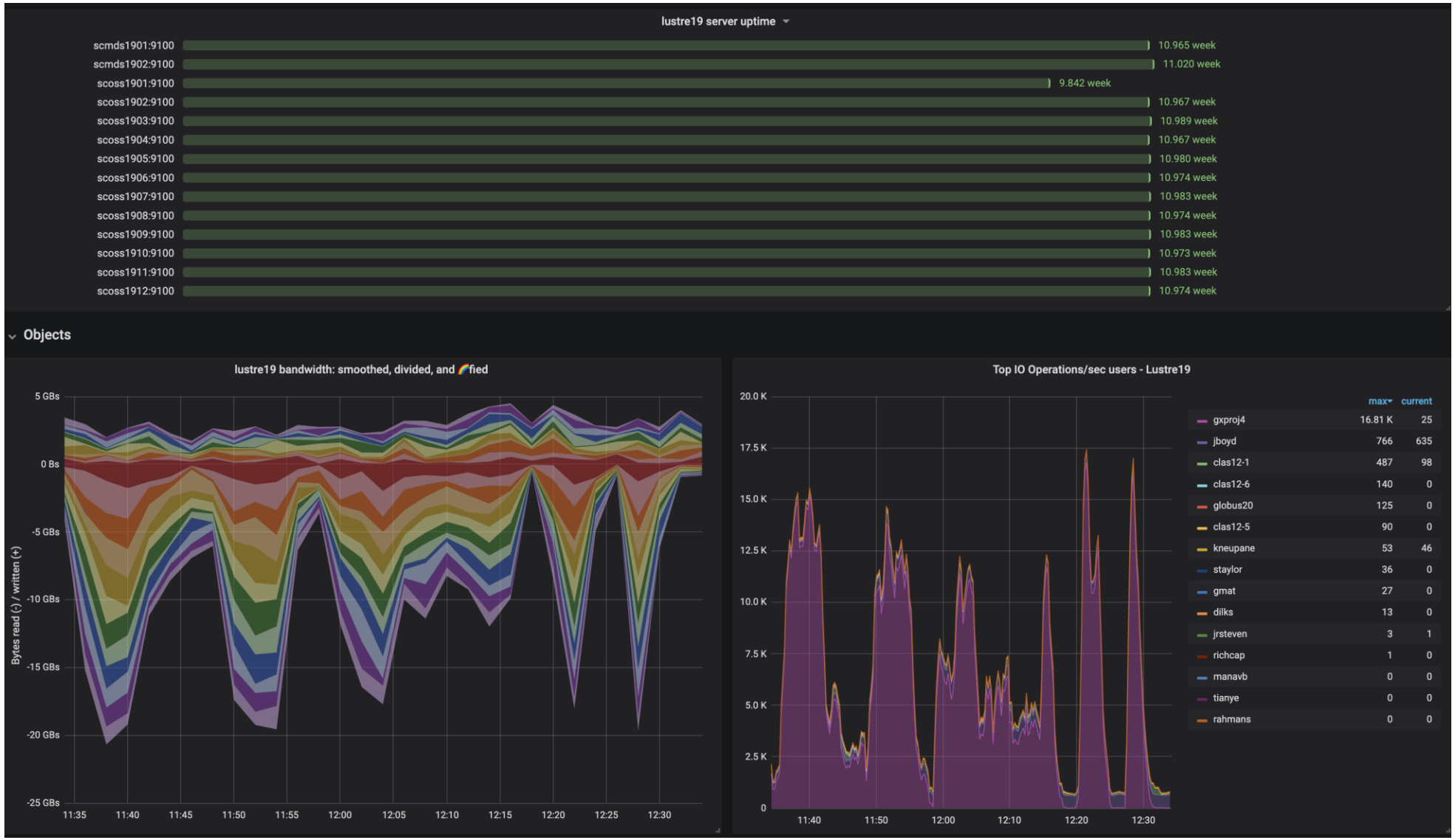


Field	Internal name	Type	Warn	Crit	Info
_cache_halla_sbs_GEnII_raw	_cache_halla_sbs_GEnII_raw	gauge			Time to 'ls -la /cache/halla/sbs/GEnII/raw'
_cache_halla_sbs_raw	_cache_halla_sbs_raw	gauge			Time to 'ls -la /cache/halla/sbs/raw'
_cache_hallc	_cache_hallc	gauge			Time to 'ls -la /cache/hallc'
group	_group_	gauge			Time to 'ls -la /group/'
_volatile_hallc	_volatile_hallc	gauge			Time to 'ls -la /volatile/hallc'
_work_hallc	_work_hallc	gauge			Time to 'ls -la /work/hallc'

NODE_FAIL return from Slurm

- We had an examination of some recent job failures.
- One recent case related to scheduled reboots (e.g. for kernel patches)
  - There is a race condition in Slurm for node booting that can cause jobs to start before the node has Lustre mounted.
  - We will reboot nodes to a "down" state to work around this for now.
  - May not be caught in all cases (e.g. unplanned reboot), but should reduce the number of NODE_FAIL cases
- NODE_FAIL also caused by name lookup failures (NIS, for example) in sssd
  - Ongoing issue with sssd being tracked
  - Specific to NIS

Lustre Performance Example: Uptime, Throughput, and IOPS



Maintenance Day Planning -- November 15, 2022

- OOM Fix for Slurm to avoid swapping
- Routine Kernel and ZFS patching for file servers
- Routine security and OSG patches for internet-facing services

Discussion & Questions