SciOps + ENP October 2022

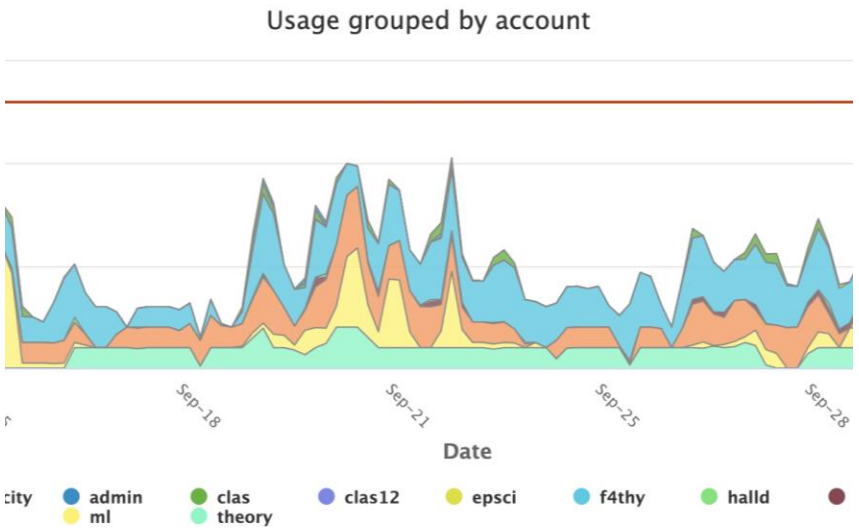
Agenda

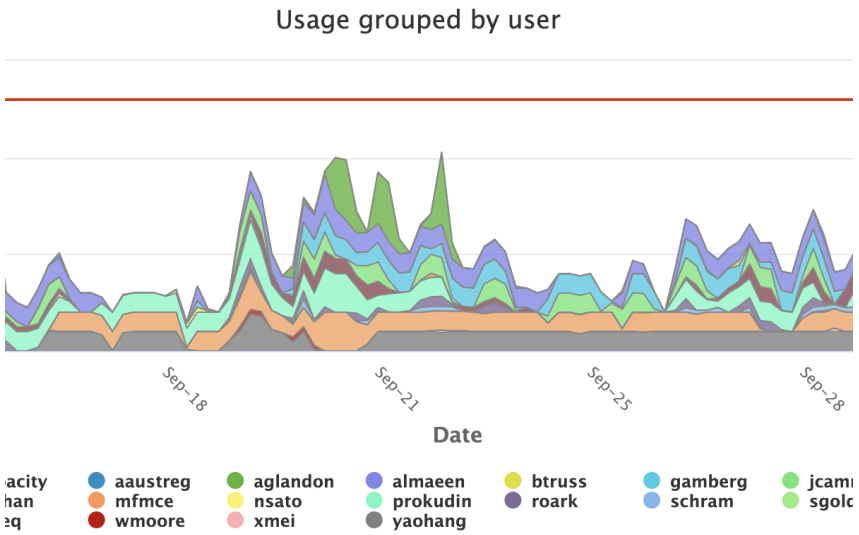
- Data Center
- GPU use
- Maintenance Plans
- Tape Storage Update

Data Center Operations

- The repair to the Data Center Chilled Water line is in progress, FM&L has the contract in flace
- Data center cooling will continue to operate on the rental chiller in the CEBAF Center parking lot for several more weeks
- We are working with our partners in Facilites Management on other maintenance and repair issues, including UPS battery replacement for the 800KVA UPS that services the farm (Oct 13)
- We do not anticipate an interruption in services

GPU Accounting in Slurm





- GPU usage is now tracked in slurm and summarized on the web portal in the usage section
- https://scicomp.jlab.org/scicomp/farmUsage/gpu

- Two A100 GPUs are now in production. More will be added in coming months.

```
bhess@ifarm1801 ~ $ sinfo -O "NodeHost:15,Gres:60" -p gpu
HOSTNAMES       GRES
sciml1901       disk:750G,gpu:TitanRTX:2,gpu:A100:2
sciml2101       disk:750G,gpu:T4:8
sciml2102       disk:750G,gpu:T4:16
sciml1902       disk:750G,gpu:TitanRTX:4
sciml1903       disk:750G,gpu:TitanRTX:4
sciml2103       disk:750G,gpu:T4:16
```

October Maintenance Planning

- Reminder that scheduled maintenance is the third Tuesday of every month.
  - Tuesday, October 18th is the next scheduled maintenance window
  - Scientific Computing maintenance prior to 5pm, CNI maintenance after 5pm.
- When there are there is beam changes are more cautious in scope
- Work to be done this month
  - Patching and rebooting of internet-facing servers (OSG, etc)
  - Patching and rebooting of ifarm machines (time to be determined)
  - A100 GPUs will be out of service to add intifity fabric bridges
  - RPM package updates to farm nodes for OSG and SciToken support

Small File Performance improvements coming for Jasmine Tape Storage

- Long Standing problem: Small files severely degrade tape drive performance
- A new Strategy
  - Small files represent a large number of files, but not so much space
  - Store all small files on a Jasmine internal disk cache
  - Write small files to tape (still) but use tape as a last resort for retrieval (cold copy)
  - Read small file requests from internal disk cache
  - Use XRootD storage for internal cache. Good failure semantics, no kernel linkage.
  - This is in testing now.
- Impact
  - Need for fewer tape drives because utilization is more predictable.
  - Go to tape file files less often as Jasmine internal cache grows
  - On a tape with 653K small files, it was able to achieve 99% utilization and ~350 MB/sec. A HUGE improvement from 1MB/sec and 99% seek.
  - Can lift the restriction on small files in /cache not being written to tape, which mitigates a data loss risk.
  - A caution: small files on Lustre are still problematic in large quantities because of high metadata ops.

Jasmine Upcoming Tape System Changes (2)

- Functional Differences
  - Small files will go to jasmine internal disk cache (tape look-aside, distinct from lustre user visible/writeable cache)
  - Auto cache hall files are stashed in jasmine cache
  - Raw duplicate stashes also kept in jasmine cache
  - All reads will traverse the cache: requested files will be retrieved from tape to cache, then sent out to requesting client.
- These changes are part of an overall plan for tape I/O
  - Good bounds on tape drive performance through scheduling and small file handling improvements
  - Fewer Round trips to tape for files; Improve disk caching at the system level
  - More caution about data protection; do not delete files from internal cache until verified on read-back.
  - Utilize community supported software where possible, Integrate. (example: XRootD cluster for cache)

NFS to Lustre (/cache) gateway hangs

- This has been an irregular, recurring bug
- Appears to be a kernel-level interaction betweek nfsd and Lustre
- Have tried a few work-arounds, time for something new
- Option 1: We are exploring User-Space NFS server option, which may work around the problem entirely. This will be vetted on scigw20b shortly. More details to follow
- Option 2: Looking ahead-- would a read-only XRootD gateway to Lustre with redundant servers be an acceptable alternative?
  - POSIX semantics (if needed) with the LD_PRELOAD option
  - Streaming with xrdcp, metadata ops with xrdfs
  - Can operationalize this if it looks viable.
  - Working Test Examples:

```
xrdfs xrdmgr1.jlab.org ls /cache/halla/sbs/raw
xrdcp -f xroot://xrdmgr1.jlab.org//cache/halla/sbs/raw/sbsvme29_6.evio.0 /tmp/
```

Farm Next Steps

- Hardware
  - Farm Node Procurement has been awarded (Thanks to lots of hard work by Amitoj)
  - Awaiting a delivery date
  - We will begin decommissioning farm13 nodes and making plans for the new node installation
  - Ethernet improvements coming to the farm in support of CVMFS, XRootD, OSG.
- Software
  - We have a Rocky 8 build for Linux ready
    - Good, Long lifetime
    - Support for essential software and kernel modules
  - A farm upgrade will be a gradual approach.
  - Plans for this are just starting
  - We will establish some test queues
  - One discussion topic will be what CUE dependencies can be deprecated (e.g. /site) in a next software revision of the farm.

Update on XRootD storage with Federated Identities (SciToken based)

- We have deployed the production token issuer with CILogon
- We have enrolled the first test users
- Documentation for two cases
  - OSG/Batch User
  - Interactive Access
- Need to identify early/beta test users for GlueX and EIC and a good contact
  - Identify mapping to /work areas for storage
  - Understand the sticking points