# EOS at the Fermilab LHC Physics Center (LPC)

Dan Szkola - Fermi National Accelerator Laboratory

EOS Workshop 2020

04 Feb 2020

# The LHC Physics Center At Fermilab

The LHC Physics Center (LPC) at Fermilab is a regional center of the **Compact Muon Solenoid Collaboration**. The LPC serves as a resource and physics analysis hub primarily for the seven hundred US physicists in the CMS collaboration. The LPC offers a vibrant community of CMS scientists from the US and overseas who play leading roles in analysis of data, in the definition and refinement of physics objects, in detector commissioning, and in the design and development of the detector upgrade. There is close and frequent collaboration with the Fermilab theory community. The LPC provides outstanding computing resources and software support personnel. The proximity of the Tier-1 and the Remote Operations Center allow critical real time connections to the experiment. The LPC offers educational workshops in data analysis, and organizes conferences and seminar series.

LHC Physics Center At Fermilab - https://lpc.fnal.gov/index.shtml

CMS Experiment - https://cms.cern

# EOS Back Then At Fermilab LPC

- Needed POSIX compliant online area for LPC analysis data
- EOS testbed built at Fermilab around June 2012
- Initially 1 MGM and 3 FST nodes
- Access was by FUSE mount and XROOTD

By May 2013, more than 600 TB was in use with EOS still not being officially in production

# EOS Today At Fermilab LPC

- LPC cluster is a 4500 core user analysis cluster
- LPC cluster supports several hundred users annually from CMS users across the U.S.
- EOS is used for LPC user data which tends to be small files with very random access
- EOS storage is approximately 7.12 PB

# EOS Space Allocation, Usage, and Growth


2 years ago

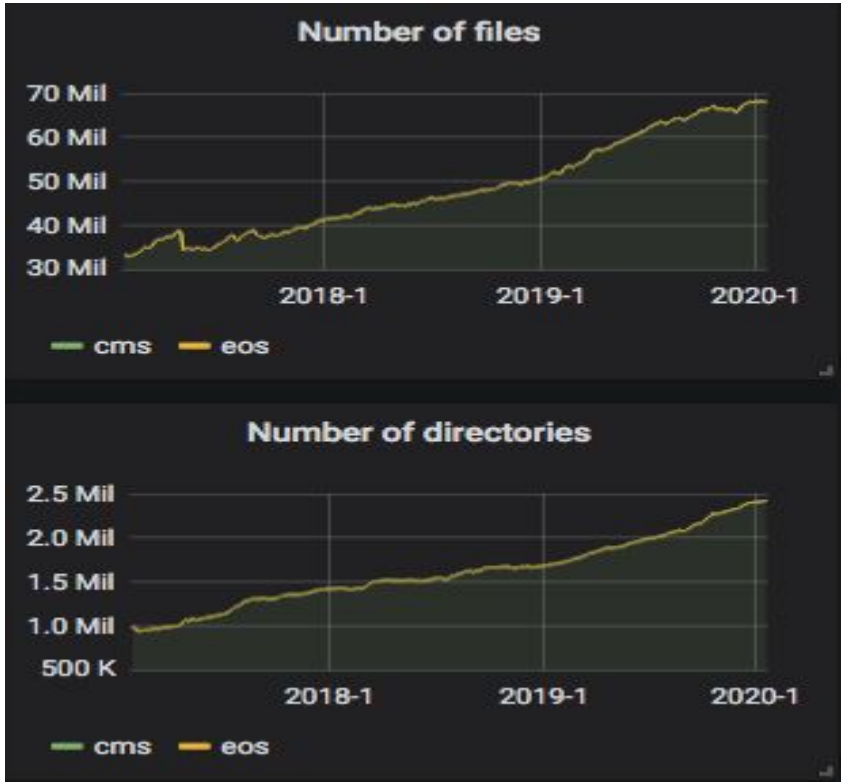

1 year ago


01/2020



Fermilab

# EOS Growth






Fermilab

# EOS Hardware and Layout

- 2 MGM nodes (a third is standing by for the QuarkDB namespace)
- 44 FST nodes
- 96 filesystems
- 4 groups

| type | name | status | N(fs) | dev(filled) | avg(filled) | sig(filled) | balancing | bal-shd | drain-shd |
|---|---|---|---|---|---|---|---|---|---|
| groupview | default.0 | on | 24 | 8.25 | 89.75 | 1.96 | balancing | 0 | 0 |
| groupview | default.1 | on | 25 | 10.24 | 86.76 | 2.40 | balancing | 0 | 0 |
| groupview | default.2 | on | 24 | 8.08 | 89.92 | 2.48 | balancing | 0 | 0 |
| groupview | default.3 | on | 23 | 6.17 | 92.83 | 1.81 | balancing | 0 | 0 |



Fermilab

# MGM Hardware

MGM servers each have an individual IP address and hostname. An ‘instance’ IP address and hostname is defined and a virtual NIC is brought up on the MGM currently defined as the master using this instance name and IP address.

- Dual Intel Xeon CPUs @ 2.10 GHz
- 256 GB RAM
- 1 TB system disk
- 2 TB SSD (for /var/eos)
- 10 Gb Ethernet

| | |
|---|---|
| eth0 | cmseosmgm01.fnal.gov |
| eth0:0 | cmseos.fnal.gov |

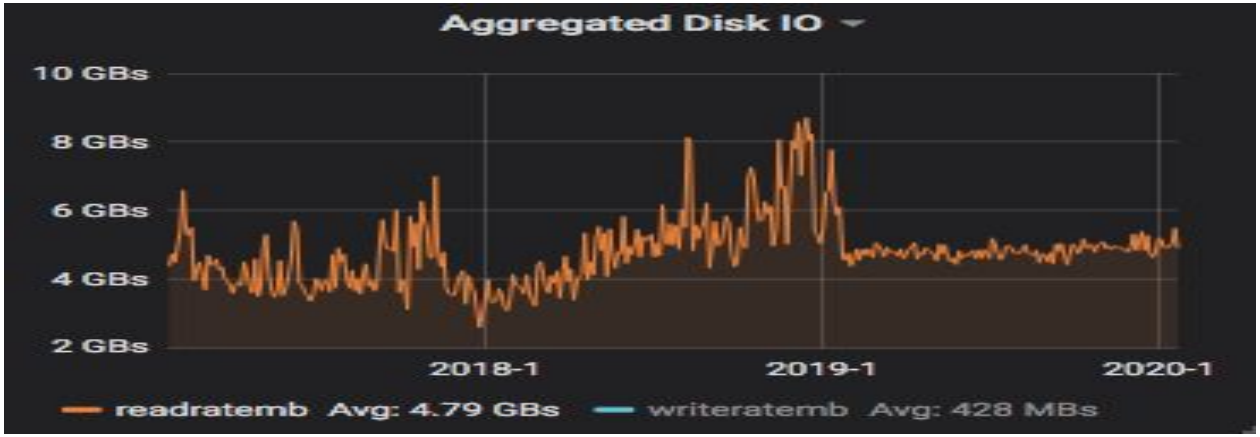
| | |
|---|---|
| eth0 | cmseosmgm02.fnal.gov |
| eth0:0 | cmseos.fnal.gov |

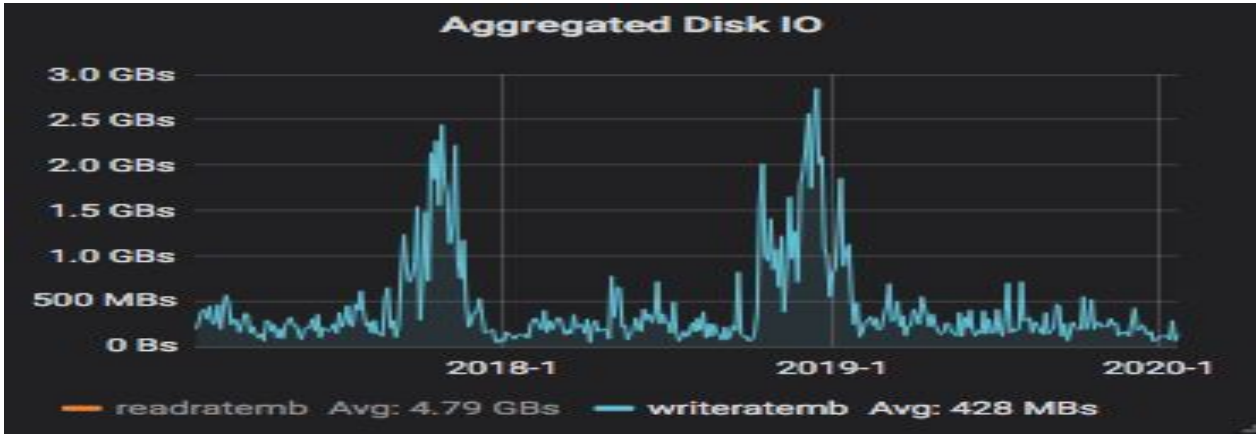
# FST Hardware

FST hardware varies as FST nodes have been added and removed over time. Typically they will have:

- Dual or Quad CPU (usually AMD Opteron)
- 64 GB RAM
- 1 - 2 TB system disk
- 10 Gb Ethernet
- 2 or 3 Nexsan volumes, most are ~71TiB a few are ~50TiB, all formatted as xfs

Older FST hardware usually had 3 volumes assigned for EOS use. Newer hardware is assigned 2 volumes.


Aggregated Disk IO
10 GBs
8 GBs
6 GBs
4 GBs
2 GBs
2018-1
2019-1
2020-1
readratemb Avg: 4.79 GBs
writeratemb Avg: 428 MBs


Aggregated Disk IO
3.0 GBs
2.5 GBs
2.0 GBs
1.5 GBs
1.0 GBs
500 MBs
0 Bs
2018-1
2019-1
2020-1
readratemb Avg: 4.79 GBs
writeratemb Avg: 428 MBs



Fermilab

# How Is EOS Space Allocated?

- Most users get a 2 TB logical (4 TB physical) area enforced by quota
- For groups (usually associated with experiments or projects), a user account is created and a quota is set based on their need for space.
- Some of the EOS space is used to hold rotated EOS logs
- /lustre/unmerged - This area is used to hold job output files that are later merged into bigger (4 - 5 GB) files.
- A temp area is defined to hold initial output of user analysis jobs.

# LRU Usage

A directory hierarchy exists in EOS for the initial output of CRAB (CMS Remote Analysis Builder, a CMS grid job tool) jobs. This user analysis data is later picked up by a separate process and moved to a user-defined area. At some point in 2019, the LRU was left on to run continuously and we have not experienced any LRU-related crashes as we had in prior releases.

- LRU rules are defined to clean up this job data after a week or so.

```
attributes.sys.lru.expire.match=*:86400
attributes.sys.lru.expire.match=*:1w
```




Fermilab

# Access To Files In EOS

- XROOTD (xrdcp, etc.)
- GridFTP
- FUSE mount (heavy use is discouraged for performance reasons)

The gridFTP service runs on all FST nodes. An F5 load balancer front-ends the gridFTP service. BeSTMan2 was eliminated with the citrine upgrade, as it has not been updated in years and no longer works properly with SL7. There are FUSE mounts on all LPC interactive nodes. On CMS worker (job) nodes, users use XROOTD to access EOS files.

# Upgrading from 0.3.268 on SL6 to 4.4.10 on SL7 (01/2019)

- First things first
  - All FST nodes were upgraded to SL7 and 0.3.268 is reinstalled on them.
    - We drained each node before upgrading it, which lengthened this step to about 3 months. Late in the process we added 6 new FSTs which gave us a bit of extra headroom and sped things up slightly.
    - No surprises here

# Upgrading from 0.3.268 on SL6 to 4.4.10 on SL7 (cont.)

CHANGE PLAN

- Update our local EOS RPM repo with the 4.4.10 RPMs and dependencies
- Shut down FSTs and MGMs
- Compact the namespace with the *eos-log-compact* command
- Copy /var/eos hierarchy to another node, just in case
- Copy scripts, cron-jobs, etc to another node
- Turn over MGM nodes to be updated to SL7
- Make any puppet changes necessary for SL7 and EOS Citrine

# Upgrading from 0.3.268 on SL6 to 4.4.10 on SL7 (cont.)

CHANGE PLAN (cont.)

- MGM nodes should now be at SL7 and have EOS 4.4.10 RPMs installed
- Update FST nodes to EOS 4.4.10

# So How Did It Actually Go?

- I had allotted 12 hours (8am - 8pm) for the whole upgrade
- MGM upgrade went smoothly
  - I missed a config item or two in puppet, but this was easily remedied
- Most of the upgrade work was finished by about 12:30pm
- The FST upgrades went a bit less smoothly
  - The nodes came up fine, but did not seem to recognize that a full metadata sync was required. Instead an empty LevelDB was created.
  - This was easily fixed by booting the filesystems by hand with the --syncmgm option.

# Upgrading from 4.4.10 to 4.5.14 (11/2019)

- Local EOS RPM repo updated to 4.5.14
  - Automatic upgrade is prevented with excludes in repo config
- A few FST nodes at a time are upgraded to 4.5.14 until all are done
- Compact the namespace
- Backup EOS metadata as a precaution
- Stop and upgrade the secondary MGM
- Stop and upgrade the primary MGM

# Upgrading from 4.4.10 to 4.5.14 (cont.)

- Upgrade went very smoothly
- 4.5.14 has several bugs that aren't show-stoppers, but are quite annoying
  - fsck report text output is broken (newlines missing)
  - FST by-node network activity is not reported
  - 4.5.14 uses XROOTD 4.10.1
    - XROOTD 4.10.1 has a nasty bug that causes problems with empty directories
      - https://github.com/xrootd/xrootd/issues/1038 [XrdCl] Client fails to parse dirlist response if the stat was requested and the directory is empty
    - EOS now has an included XROOTD, so we backed the system XROOTD down to 4.8.6 on all FST nodes.

# Users EOS Complaints and Requests for Enhancements

- (Still) Annoyed that *path=* is prepended to EOS 'find' command output
  - *path=* also added with --xurl option, makes output worse/useless

    e.g. `root://cmseos.fnal.gov/path=/eos/uscms/store/user/dszkola/`
- Would like a 'du' command, currently using a user-written *eosdu* script which usually has to be upgraded when EOS is upgraded
  - https://github.com/FNALLPC/lpc-scripts/blob/master/eosdu
- Better doc and examples for EOS commands
- Recent fuse usage during DAS highlighted poor fuse performance (ver 4.5.14)
- *eos cp -r* should work when source dir is on EOS

# Users EOS Complaints and Requests for Enhancements (cont)

- Users don't like being forced to add the / at the end of directories), i.e. fix:

```
Remark:
    If you deal with directories always add a '/' in the end of source or target paths
    e.g. if the target should be a directory and not a file put a '/' in the end. To
    copy a directory hierarchy use '-r' and source and target directories terminated
    with '/' !
```

- *eos mv -h* is confusing because it returns help for the 'file' command
- Add a '-t' option to sort files by time
- Have EOS commands properly handle wildcards (this is a long-standing complaint)

# EOS In The Near Future At Fermilab LPC

- Already have a test stand running QuarkDB, will move production to QuarkDB when it is feature complete
- Still using FUSE, we will look at FUSEx
- We have an eye towards IPv6 (RSN - real soon now)
- Eliminate or limit the amount of single replica space
- Some users have expressed interest in a CERNBox type implementation, but unlikely we will do this anytime soon
- EOS growth is expected to be approximately 1 PB per year if funding permits

# EOS - What Could Be Improved

- Still many ‘mysterious’ MGM crashes
  - This was determined to be caused by us updating our password/group DB without an NSCD or similar cache in place. Still no explanation as to why this causes a crash though.
- Documentation could be more complete and is sometimes out of date
- Output of some of the commands is cryptic and is not explained anywhere
- No complete list of config statements for xrd.cf.* files or /etc/sysconfig files and no explanation for some options without digging into the source

# EOS - What Could Be Improved

- Community support
  - The EOS Community site is a big improvement
    - Good level of participation
    - CERN devs and admins often answer questions
  - Could maybe pin some useful information at the top such as…
    - Officially supported or recommended procedures for upgrades (to citrine, to QuarkDB, etc.)
    - Changelogs, release notes for latest supported or recommended release
    - Any useful tidbits, scripts, configuration suggestions admins might find useful
    - As previously mentioned, description of all available config directives

# Contributors

Thanks to the following people at Fermilab who provided information for this Presentation:

- David Mason
- Marguerite Tonjes