EOS at the Fermilab LHC Physics Center (LPC)

Dan Szkola - Fermi National Accelerator Laboratory

EOS Workshop 2023

24–27 Apr 2023

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

History of EOS 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

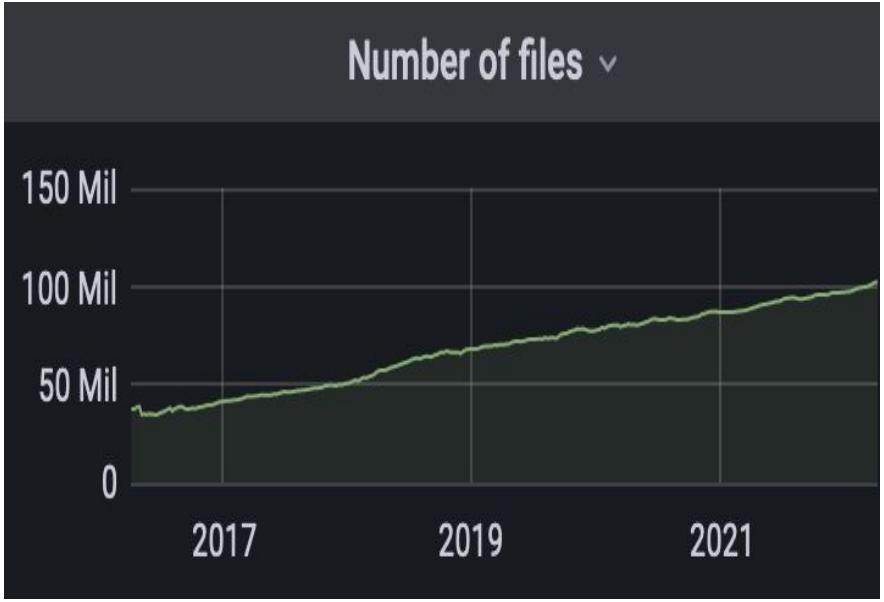
EOS Today At Fermilab LPC

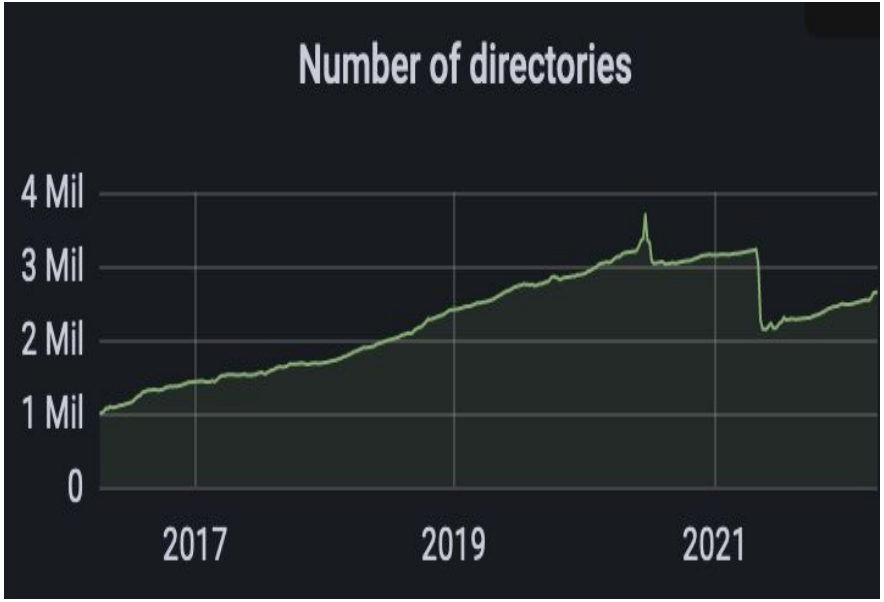
- LPC cluster is a 4500 core user analysis cluster
- LPC cluster supports over eight hundred users annually who ask for new accounts or renew their existing accounts
- EOS is used for LPC user data which tends to be small files with very random access
- EOS storage is approximately 13 PB of replicated space

EOS Space Allocation, Usage and Growth

Year	EOS Total Space	EOS Free Space
2017	4.75 PB	552.03 TB
2018	6.19 PB	1.62 PB
2019	7.12 PB	1.69 PB
2020	7.64 PB	1.13 PB
2021	11.0 PB	3.31 PB
2022	13.2 PB	3.21 PB
2023	13.2 PB	2.90 PB

EOS Space Allocation, Files and Directories

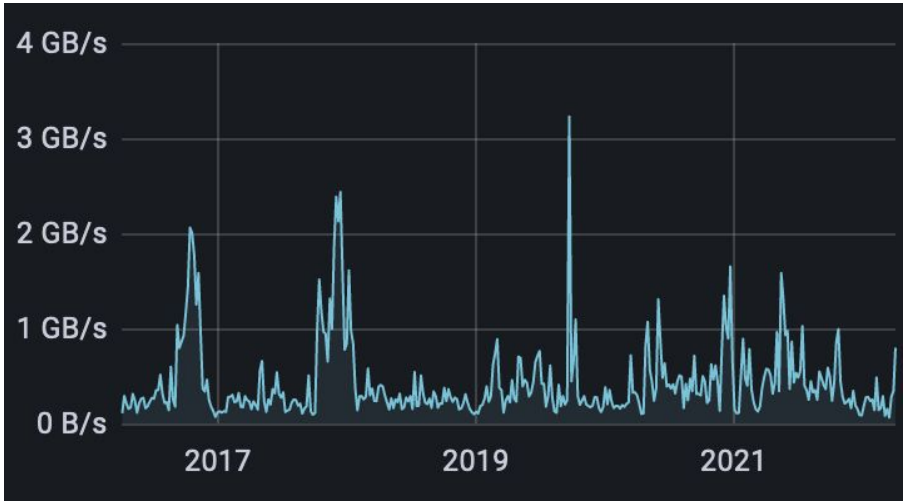




Files: 4/2017 - 37.4 million
4/2023 - 102 million

Directories: 4/2017 - 1 million
4/2023 - 2.64 million

EOS Aggregate I/O



EOS Hardware and Layout

- 3 MGM nodes configured similarly, only 2 run as MGMs, the other is a third QuarkDB node
- 91 FST nodes
- 215 filesystems
- 4 groups

type	name	status	N(fs)	dev(filled)	avg(filled)	sig(filled)	balancing	bal-shd
groupview	default.0	on	54	9.88	78.30	2.52	idle	0
groupview	default.1	on	54	3.89	79.01	0.89	idle	0
groupview	default.2	on	54	8.65	77.79	1.41	idle	0
groupview	default.3	on	53	9.86	77.76	2.30	idle	0

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 E5-2620 v4 CPUs @ 2.10 GHz
- 256 GB RAM
- 1 TB system disk
- Mirrored 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, the sizes of these volumes vary from 36TB to 77TB and are formatted as XFS volumes

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
- An area is designated 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.

File Access In EOS

- XROOTD (xrdcp, etc.)
- GridFTP
- FUSE mount (heavy use is discouraged for performance reasons)
- HTTP(S) with x509, macaroons, and scitokens - this was enabled last year and seems to be working well

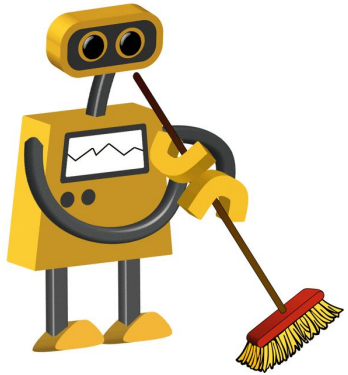
The gridFTP service runs on some of the FST nodes. An F5 load balancer front-ends the gridFTP service. There are FUSE mounts on all LPC interactive nodes. On CMS worker (job) nodes, users use XROOTD to access EOS files.

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. The LRU runs continuously with an interval of 86400s (1 day).

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

```
attributes.sys.lru.expire.empty="1mo"
attributes.sys.lru.expire.match=*:1w
```



EOS and older hardware

- At the time of the upgrade to EOS 4.6.8, we had 3 nodes with Opteron 6128 CPUs
  - These CPUs do not support SSE4_1 and SSE4_2 instructions
- It was discovered that a few releases prior to 4.6.8, optimizations that included SSE4 instructions were being done at build time
  - This caused EOS to crash with ‘illegal instruction’ on the FSTs with Opteron 6128 CPUs
  - The FSTs with these CPUs were held back to a previous version
  - When the upgrade to 4.7.16 was done, a special build was done that did not include SSE4 instructions and all nodes were able to run 4.7.16 code.
  - At that time, it was hopeful that we would only ever have these three nodes and once they were retired we could avoid using a special build

EOS and older hardware (cont.)

- 26 new nodes were retired from use in dCache and added to EOS as FST nodes
  - All 26 nodes have Opteron 6128 CPUs
- As of EOS 4.8.39, devs have added a build path for binaries that do not contain SSE4 instructions
  - This is obviously not ideal, but we have to run with the hardware that is available
  - The extra space was required for an upcoming elimination of single replica space used in some group areas
    - This was provided for some groups due to quota concerns, but has caused too many issues and is now being eliminated
- Most of our nodes used for test instances also have Opteron 6128 CPUs

Current EOS version at FNAL: 4.8.66

- The upgrade to 4.8.66 was done on 2022-02-01
  - Minor upgrade
  - No unexpected issues
  - We are still running this version with an upgrade to Diopside (EOS 5) expected soon
    - A test environment has been running 5.1.8 for a few months

Recent EOS changes at FNAL

- HTTP transfer (x509, sci-tokens, and macaroons) is in place, gridFTP is still being used for now
- FUSEx is on a few LPC nodes and all FSTs are running FUSEx

EOS In The Near Future At Fermilab LPC

- EOS v5 production upgrade will take place very soon
- SL7 is EOL soon, we will be switching to AlmaLinux 8 or 9
- Most LPC nodes are still using FUSE, we need to move to FUSEx
- Would like to eventually test erasure coding in one of the test environments
- Some nodes will soon be retired from dCache and moved to EOS, hopefully enough to eliminate all Opteron 6128 nodes to eliminate the need for a NOSSE build

EOS - What Could Be Improved and What Has Improved

- We are awaiting the new HA implementation in EOS5 as we still rely on a VIP on the primary MGM node that must be moved manually to failover properly
- Documentation has gotten better but could still be more complete and is sometimes out of date
- Output of some of the commands is cryptic and is not explained anywhere
  - ‘eos group ls’ is a good example
- 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. This is especially important when upgrading and new config statements are required.

EOS - What Could Be Improved and What Has Improved

- Release notes
  - These should be part of the documentation and new required configuration options should be listed as part of the release notes. Perhaps recommended default configurations could be provided for such releases.
- Community support
  - The EOS Community site is a good resource
    - Good level of participation
    - CERN devs and admins often answer questions

EOS User Observations and Requests for Enhancement

- The 'eos find' command prepends "path=" to output (including with '--xurl' option)
- no 'eos du' command: user-developed script is currently used, but usually needs to be updated when the EOS version changes. 'eos ls -lh' showing full directory sizes is not adequate because:
  - to know the size of a specific directory, have to call 'eos ls -l' on the *parent* directory
  - hard to list the size of all subdirectories ('eos ls' shows all files and directories; can't restrict to just directories)
  - no way to list the number of files rather than the total size (since eos has quotas for both, this is important; user tool has this feature)
- 'eos ls' should have a '-t' option to sort by time (following standard Linux ls)
- 'eos cp -r' (recursive) should work when source directory is on EOS

EOS User Observations and Requests for Enhancement (cont)

- fix this behavior (from 'eos cp --help') so / doesn't need to be added if EOS already knows that the path is a directory:

```
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 '/' !
```

- Why is there a separate 'eoscp' command distinct from 'eos cp'?
- Overall, the quality of the documentation and examples for EOS commands should be improved.
  - eos mv -h' and 'eos ln -h' return the help message for 'eos file ...', which does not include the terms 'mv' or 'ln' anywhere and in general is not written clearly.
- The syntax for 'eos ln' is backwards vs. the Linux system ln command (for making symlinks) and the help message is written especially unclearly

Users EOS Complaints and Requests for Enhancements (cont)

- Wildcard support in EOS commands is inconsistent: some commands now support wildcards, but others don't (e.g. 'eos info'), and they do not always work correctly. Examples:
  - *eos root://cmseos.fnal.gov ls /store/user/pedrok/r** lists any file or directory in my area that contains the character 'r' *anywhere*, which is incorrect; should only list files *starting with* r
  - *eos root://cmseos.fnal.gov ls /store/user/pedrok/*z* correctly lists only files ending in z (e.g. .tar.gz files).

Contributors

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

- David Mason
- Marguerite Tonjes
- Kevin Pedro