EOS at the Fermilab LHC Physics Center (LPC)

Dan Szkola - Fermi National Accelerator Laboratory

EOS Workshop 2024

March 14-15, 2024

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.

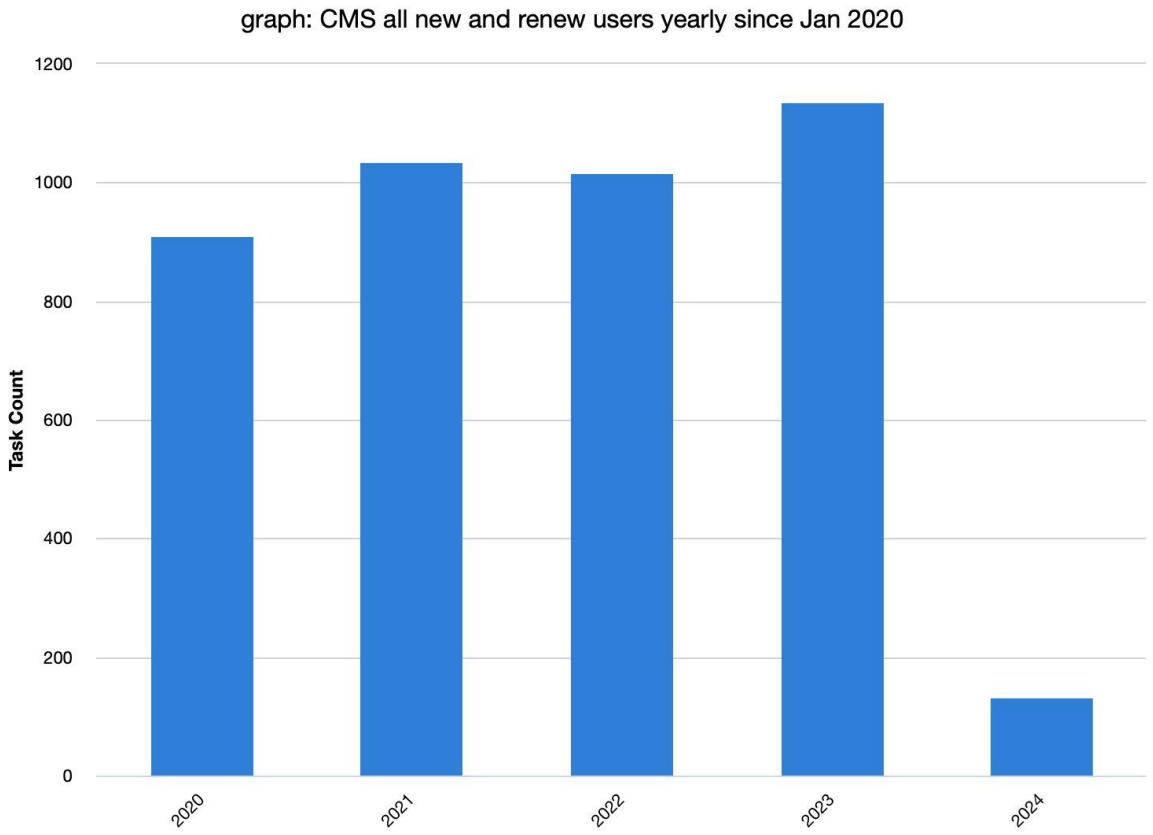
LHC Physics Center At Fermilab - https://lpc.fnal.gov/index.shtml

CMS Experiment - https://cms.cern

EOS Today At Fermilab LPC

- First EOS instance built in 2012
- LPC cluster is a 4000 core user analysis cluster
- LPC cluster supports over 1100 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 with most space having 2 replica layout

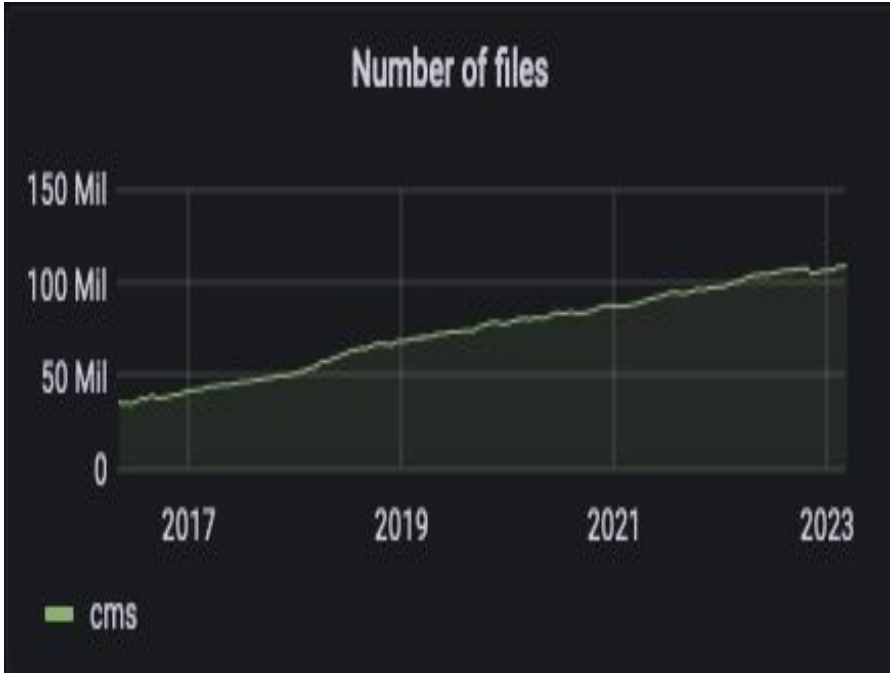
LPC Cluster User Count

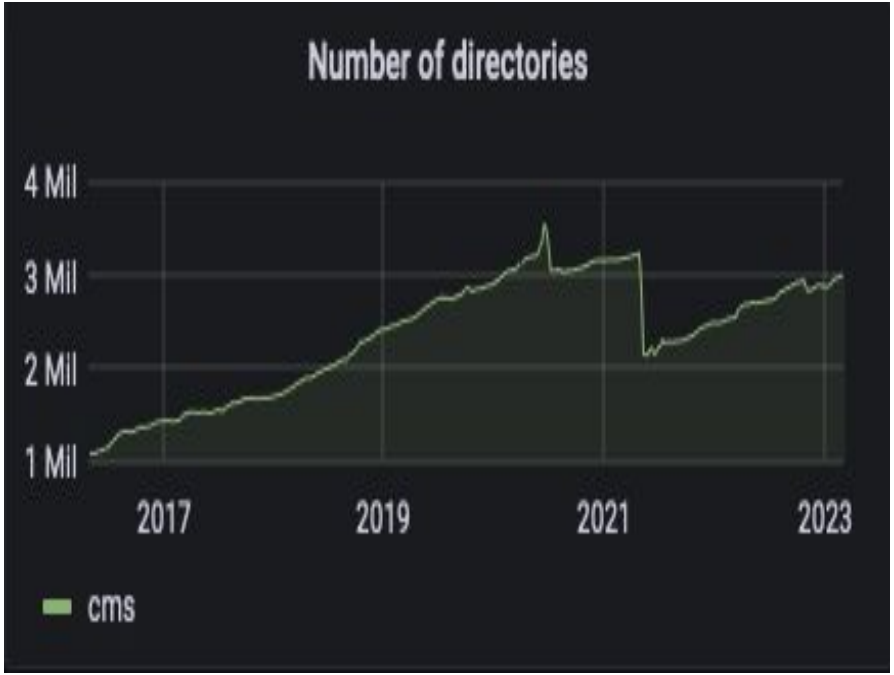


EOS Space Allocation, Usage and Growth

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
2024	12.9 PB	1.39 PB

EOS Space Allocation, Files and Directories





Files: 4/2017 - 37.4 million
3/2024 - 108.60 million

Directories: 4/2017 - 1 million
3/2024 - 2.99 million

Current EOS Hardware and Layout

- 3 nodes with similar hardware, all are QuarkDB nodes, two are MGM nodes, an active and a standby
- 83 FST nodes
- 170 filesystems
- 4 groups

type	name	status	N(fs)	dev(filled)	avg(filled)	sig(filled)	balancing	bal-shd
Groupview	default.0	on	42	6.87	92.13	2.40	idle	0
Groupview	default.1	on	43	8.94	90.08	1.69	idle	0
Groupview	default.2	on	43	9.48	87.36	1.93	idle	0
Groupview	default.3	on	42	9.53	87.20	1.68	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
- FUSEx mount (heavy use is discouraged for performance reasons)
- HTTP(S) with x509, macaroons, and scitokens

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

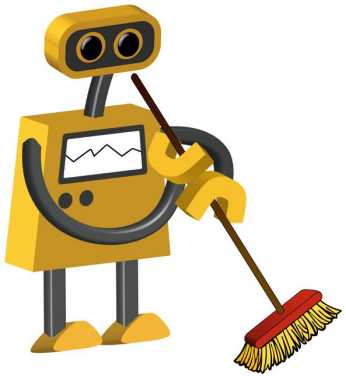
EOS Features used in FNAL EOS Instance

LRU

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
```



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 and later 5.1.23 for many months

Recent EOS Instance changes

- 11/2022 - FST node FUSE mounts moved to FUSEx without issues
- 05/2023 - LPC interactive nodes moved from FUSE to FUSEx
- 10/2023 - FST nodes with Opteron 6128 CPUs (no SSE4_1 and SSE4_2 instructions) were replaced with newer hardware, eliminating the need for the NOSSE EOS build (some test nodes still have this CPU, replacements are coming)

EOS In The Near Future At Fermilab LPC

- EOS v5 production upgrade to 5.1.x scheduled for 2024-Mar-20
  - Node OS upgrades and upgrade to EOS 5.2.x will follow
- SL7 is EOL soon, we will be switching to AlmaLinux 8 or 9, which version will depend on support for our hardware
- Would like to eventually test erasure coding in one of the test environments

EOS - What We Would Like To See (RFE)

- New (better) 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 is much improved, but some additions could be made
  - What is represented in the output of some commands (‘eos group ls’ is a good example, e.g. what do dev(filled) avg(filled) sig(filled) represent?)
- A complete, up-to-date list of config statements for xrd.cf.* files and /etc/sysconfig files with a short comment on each, what effect the option has, case when it should be used or not used, etc. would be appreciated

EOS User Observations and Requests for Enhancement

- 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'?
- The quality of the help text and examples for EOS commands could 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