EOS at the Fermilab LHC Physics Center (LPC)

Dan Szkola - Fermi National Accelerator Laboratory

EOS Workshop 2020

01 Mar 2021

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 At Fermilab LPC - Early Days

- 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

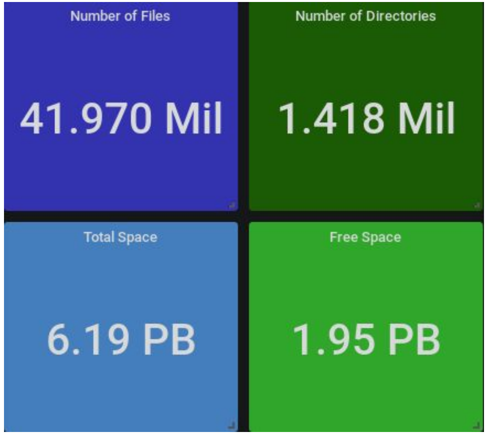
By May 2013, more than 600 TB was in use with EOS still not being officially in production

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 11 PB of replicated space

EOS Space Allocation, Usage, and Growth

2018



Number of Files

41.970 Mil

Number of Directories

1.418 Mil

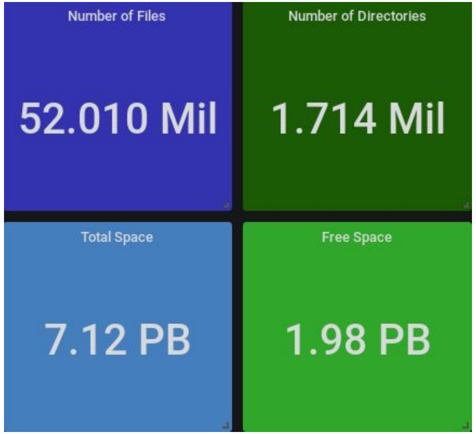
Total Space

6.19 PB

Free Space

1.95 PB

2019



Number of Files

52.010 Mil

Number of Directories

1.714 Mil

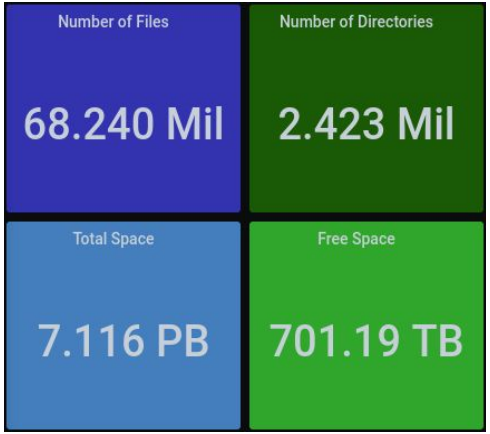
Total Space

7.12 PB

Free Space

1.98 PB

2020



Number of Files

68.240 Mil

Number of Directories

2.423 Mil

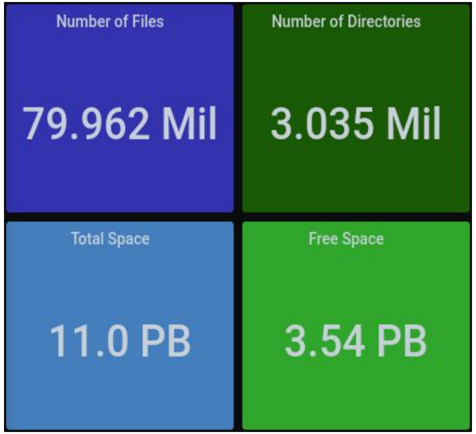
Total Space

7.116 PB

Free Space

701.19 TB

2021



Number of Files

79.962 Mil

Number of Directories

3.035 Mil

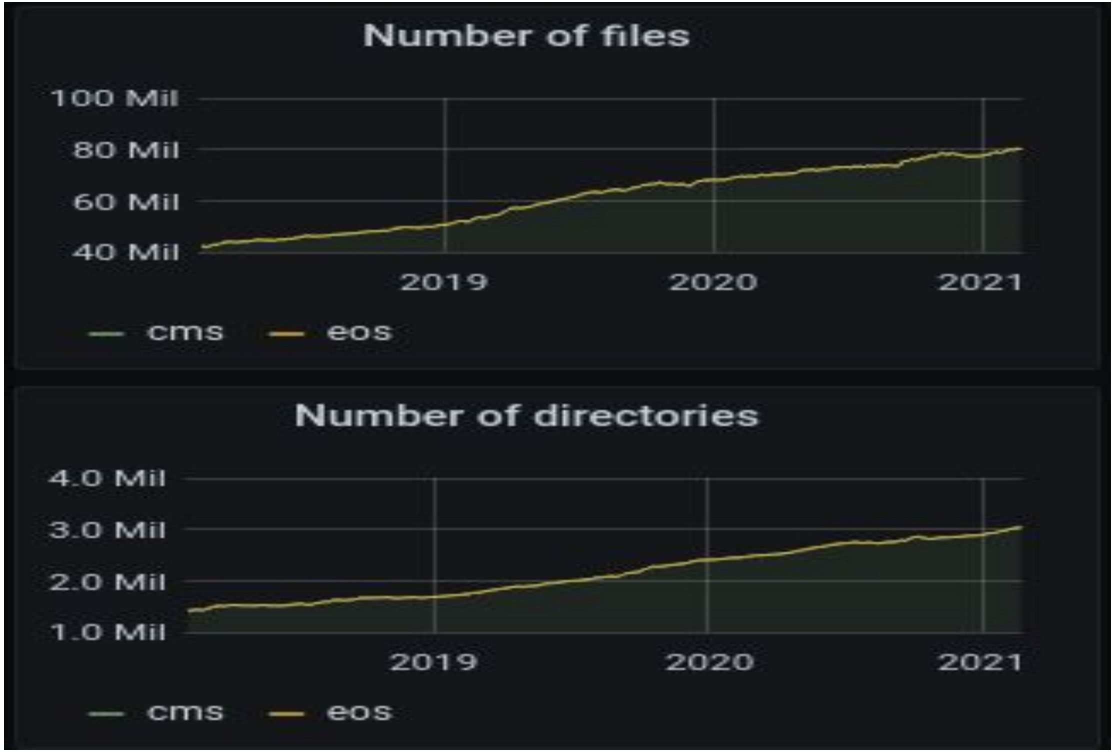
Total Space

11.0 PB

Free Space

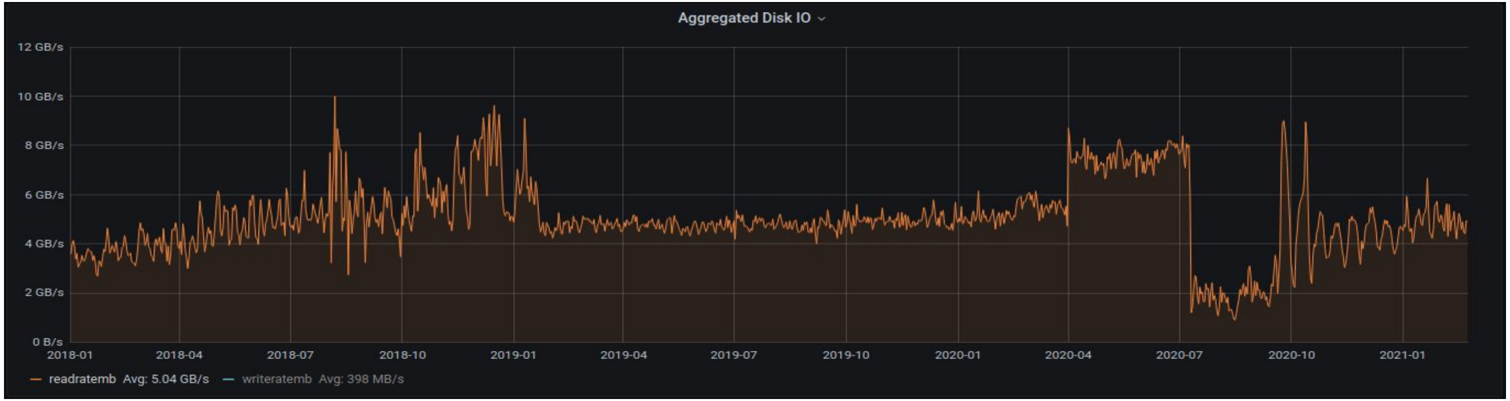
3.54 PB

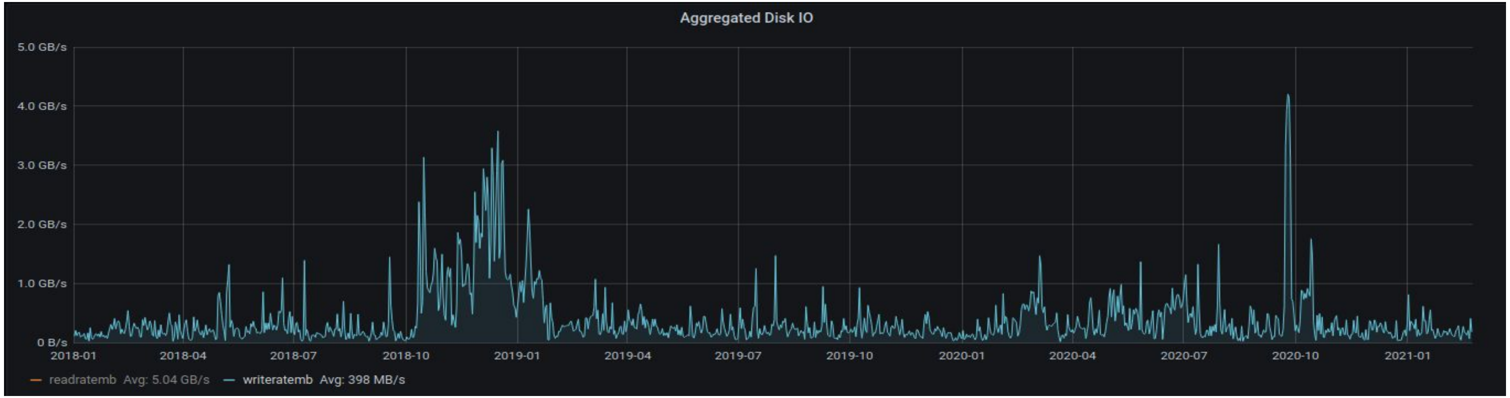
EOS Growth



Number of files
100 Mil
80 Mil
60 Mil
40 Mil
2019
2020
2021
— cms
— eos
Number of directories
4.0 Mil
3.0 Mil
2.0 Mil
1.0 Mil
2019
2020
2021
— cms
— eos

Aggregate I/O





EOS Hardware and Layout

- 2 MGM nodes and a third similarly configured running only QuarkDB
- 77 FST nodes
- 186 filesystems
- 4 groups

```
┌──────────┬────────────────┬────────────┬──────┬────────────┬────────────┬────────────┬──────────┬──────────┬──────────┐
│type      │            name│      status│ N(fs)│ dev(filled)│ avg(filled)│ sig(filled)│ balancing│   bal-shd│ drain-shd│
└──────────┴────────────────┴────────────┴──────┴────────────┴────────────┴────────────┴──────────┴──────────┴──────────┘
 groupview        default.0           on     47         4.49        66.51         1.09  balancing          0          0
 groupview        default.1           on     47        13.57        66.43         2.09  balancing          0          0
 groupview        default.2           on     46         8.67        68.33         1.35  balancing          0          0
 groupview        default.3           on     46         4.26        67.74         0.99  balancing          0          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 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, 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
- /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
```

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. There are FUSE mounts on all LPC interactive nodes. On CMS worker (job) nodes, users use XROOTD to access EOS files.

Upgrading from 4.5.14 to 4.6.8

- The upgrade was done on 2021-03-31
- This upgrade included moving the namespace to QuarkDB
- This upgrade also included running IPv4/IPv6 dual stack on all nodes
  - Added a third ‘MGM’ node (not running MGM services) for QuarkDB
  - FST nodes were upgraded while the namespace conversion was being done
  - All MGM metadata was copied off (just in case)
  - MGM nodes were upgraded to 4.6.8 before the namespace conversion
  - As big a change as this was, it was very straightforward and mostly uneventful
    - It was this upgrade where we started to have issues with some older hardware (see later slides)

Upgrading from 4.6.8 to 4.7.16

- The upgrade was done on 2021-06-10
  - Minor upgrade
  - We were experiencing intermittent FST shutdowns in 4.6.8, this upgrade was done to mitigate that issue
  - This is the version we are currently running, but we’ll be upgrading to 4.8.39 (or higher?) soon

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 SSE2 instructions
- It was discovered that a few releases prior to 4.6.8, optimizations that included SSE2 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 SSE2 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 SSE2 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
- All 14 of our nodes used for test instances also have Opteron 6128 CPUs

EOS In The Near Future At Fermilab LPC

- Production instance will move to 4.8.39 (or later)
- Still using FUSE, we will look at FUSEx, but we’re not sure the added complexity is worth any gains
- Single replica space provided for some groups will be changed to standard two replica space
- Test RAIN on one of the test environments

EOS - What Could Be Improved

- There still exists no ideal method (at least not a documented one) for dealing with a floating instance IP when an MGM is promoted
  - When an MGM becomes primary, the IP address pointing to the instance does not move with it. I’ve tried some automated (cron checks, etc) methods, but these tend to be crude and not always work
- Documentation could 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

EOS - What Could Be Improved

- Community support
  - The EOS Community site is great
    - Good level of participation
    - CERN devs and admins often answer questions
- Release notes
  - I know they exist for some releases, but sometimes important information is easy to miss
    - ‘Nominal Capacity’ was a good example of this

Users EOS Complaints and Requests for Enhancements

- *eos rm* prints extra text:

```
# pre-configuring default route to /eos/user/p/pedrok/
# -use $EOSHOME variable to override
```

- (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

Users EOS Complaints and Requests for Enhancements (cont)

- *eos cp -r* should work when source dir is on EOS
- *eos ls* should have a ‘-t’ option to sort by time
- 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
- Users are unclear as to why there is a separate *eoscp* command and how it differs from *eos cp*
- Have EOS commands properly handle wildcards (this is a long-standing complaint)

Contributors

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

- David Mason
- Marguerite Tonjes
- Kevin Pedro