

ORNL EOS Environment

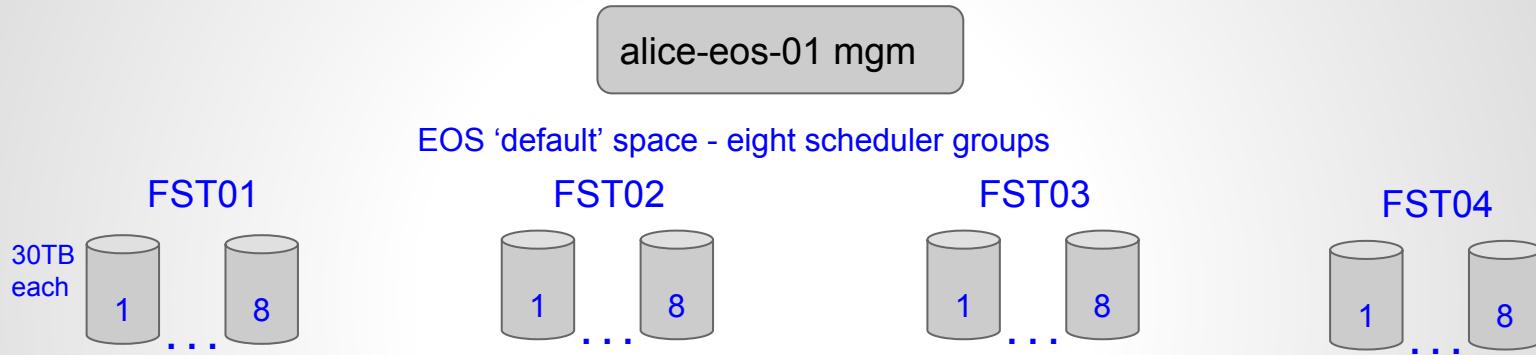


ALICE T2 Summit, Strasbourg 2017
April 2017

EOS - Hardware Lifecycle Mgmt

- New models of JBODs and drives
 - Dense storage (600TB per JBOD)
 - Differing storage pool layouts, redundancy, etc.
- How to compartmentalize same gen hardware for lifecycle management?
- How to ensure ability to drain / rebalance
- Can't store hardware descriptor in schedule group metadata
- **Solution:** Talk with EOS experts on how to best address lifecycle management goals (thanks Andreas!)

EOS Original 1PB Layout



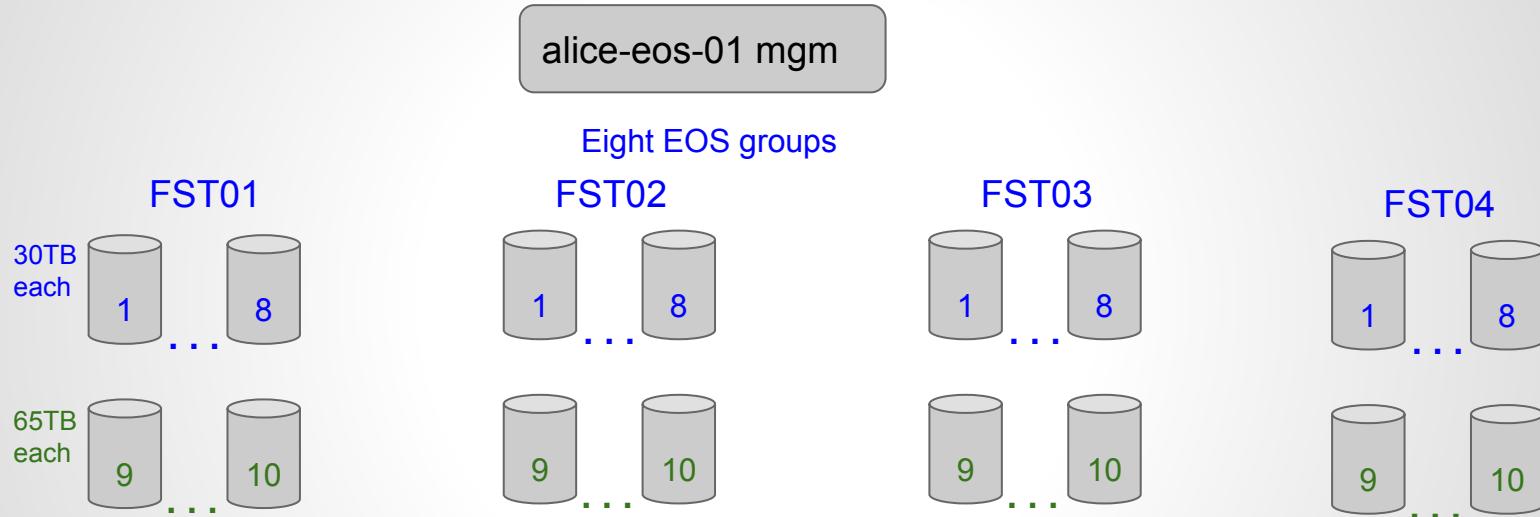
- Four FSTs
 - Each mounts 8 filesystems
- One eos space: 'default'
 - Eight eos scheduler groups:
 - default.0 - default7
 - Schedule groups equal size
 - Each group consists of four file systems (one from each FST)
- Easy!

EOS - Add another .5PB

- Creating a new ‘eos space’ not too helpful
 - Spaces are branches in eos namespace
 - Could designate all new files be stored in new space - meh.
 - Creating a ‘spare’ space was helpful for adding / configuring new storage
- **Solution:** Use eos scheduling groups, carve up new storage accordingly, use logical naming to map resources to hardware

- ★ A word about scheduling groups:
 - Should be kept same size
 - Data placement is round-robin
 - Draining/Balancing works within a group
 - If new filesystems are added to existing groups, possible to use the simple 'drain' mechanism to move all data from old filesystems to the new ones and retire them.
 - If the new hardware is in separate groups or spaces, you have to convert all files from a given old group to live in the new space.
 - There is a high-level group balancer to balance between groups, but *there is no high-level draining between groups*

EOS Storage Expansion



Carved up the new 544TB in eight equally sized 65T (quoted) file systems, to distribute evenly among existing eos groups. (Each eos group gain one new 65TB fs.)

EOS: Eight Groups

Cloud icon: Drain/Balance within groups

19:28:19 # eos group ls

| # | type # | name # | status | #nofs | #dev(filled) | #avg(filled) | #sig(filled) | #balancing | # bal-shd |
|---------------------|--------|--------|--------|-------|--------------|--------------|--------------|------------|-----------|
| #drain-shd | | | | | | | | | |
| # | | | | | | | | | |
| groupview default.0 | | | | | | | | | |
| | on | 5 | 70.17 | 82.45 | 35.09 | idle | 0 | 0 | |
| groupview default.1 | | | | | | | | | |
| | on | 5 | 71.84 | 82.03 | 35.92 | idle | 0 | 0 | |
| groupview default.2 | | | | | | | | | |
| | on | 5 | 61.85 | 84.53 | 30.93 | idle | 0 | 0 | |
| groupview default.3 | | | | | | | | | |
| | on | 5 | 65.60 | 83.59 | 32.80 | idle | 0 | 0 | |
| groupview default.4 | | | | | | | | | |
| | on | 5 | 61.16 | 84.70 | 30.58 | idle | 0 | 0 | |
| groupview default.5 | | | | | | | | | |
| | on | 5 | 60.74 | 84.80 | 30.37 | idle | 0 | 0 | |
| groupview default.6 | | | | | | | | | |
| | on | 5 | 63.02 | 84.23 | 31.51 | idle | 0 | 0 | |
| groupview default.7 | | | | | | | | | |
| | on | 5 | 53.65 | 73.75 | 32.63 | idle | 0 | 0 | |



EOS ‘spare’ pool and fs config

```
(eos)space define spare 4 2
```

Registering new fs with eos:

```
(fst)eosfstregister alice-eos-01.ornl.gov /warpfs/cern-09 spare:1
```

Unresgistering:

```
(eos) fs config NN configstatus=empty  
fs rm NN
```

Config new fs, moving to production:

```
fs ls -m spare
```

View properties, configure as desired.

```
fs config NN headroom=50G
```

```
fs config NN configstatus=ro (or rw when ready)
```

```
fs mv NN default.7 (groups should be kept same size)
```

```
fs boot NN
```

Mapping groups to hardware

OS perspective:

Use naming convention in volume/pool names to designate hardware type/generation:

```
df -h
```

```
volume-n01-v09/cern-09-hgst
    65T 8.1T 57T 13% /warpfs/cern-09
volume-n01-v09/cern-10-hgst <= Hardware type/gen
    65T 6.7T 59T 11% /warpfs/cern-10
```

EOS perspective:

```
20:31:07 # eos group ls default.1 -l
```

```
#
#-----#
#   type #      name #  status #nofs
#-----#
groupview      default.1      on      5
#
#-----#
#           host #port #  id #
#                           uuid #
#                           path #
#                           schedgroup #
#-----#
...ornl-cern-02.ornl.gov 1095  26 cb6ccf1a-7ef3-4a05-86ee-b9fd17af9514 /warpfs/cern-02      default.1
...ornl-cern-03.ornl.gov 1095  33 d1dc1706-fd5a-4635-90e3-d135add31808 /warpfs/cern-02      default.1
...ornl-cern-04.ornl.gov 1095  40 02d50c75-9f01-411c-a911-7d6c2b957adb /warpfs/cern-02      default.1
...ornl-cern-01.ornl.gov 1095  47 2bdc9c9d-5a76-4328-b8d3-d018408663c3 /warpfs/cern-02      default.1
...ornl-cern-01.ornl.gov 1095  58 0aea6f97-cea5-408e-8cfb-613547af5d32 /warpfs/cern-10      default.1
```

Quotas and EOS Headspace

Blah blah, what do you know... it crashes if
you hit 100%

ZFS Compression 7% Return

Enabled on new 65T filesystems

- 100TB populated thus far
- Consistently seeing 6-7% compression ratio
 - = 112TB “for free” in our 1.6P EOS
- No measurable performance impact

```
#zfs get refcompressratio /warpfs/cern-09  
volume-n02-v09/cern-09-hgst refcompressratio 1.06x -
```

| Filesystem | Size | Used | Avail | Use% | Mounted on |
|-----------------------------|------|------|-------|------|-----------------|
| volume-n02-v09/cern-09-hgst | 65T | 15T | 51T | 24% | /warpfs/cern-09 |