
CERN

Providing large-scale disk storage

Herve Rousseau | on behalf of CERN IT Storage group

Table of Contents

CentOS7

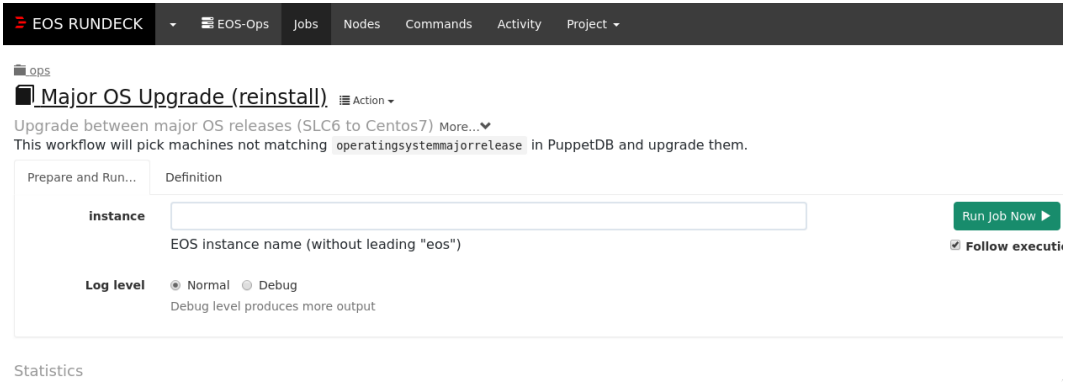
How to upgrade ∼1300 machines with minimal disruption?

Automation is key

- `Rundeck`: IT Operations management platform
- Leveraged components of CERN's "Agile" infrastructure
- Only raise attention when stuck

∼30 machines per day ⇒ 2 months

CentOS7


EOS RUNDECK
EOS-Ops
Jobs
Nodes
Commands
Activity
Project
ops
Major OS Upgrade (reinstall)
Action
Upgrade between major OS releases (SLC6 to Centos7) More...
This workflow will pick machines not matching operatingsystemmajorrelease in PuppetDB and upgrade them.
Prepare and Run...
Definition
instance
EOS instance name (without leading "eos")
Run Job Now
Follow executi
Log level
Normal
Debug
Debug level produces more output
Statistics

Namespace

Service grows faster than available hardware

Scale-up limitations

- Routine maintenance becomes a burden
- Boot time skyrockets

QuarkDB

"A highly available datastore with a Redis-like interface"

Namespace

See A. Manzi's talk right after[1]

[1] https://indico.cern.ch/event/587955/contributions/2936873/

WLCG Accounting

EOS now generates SRR[2] compatible JSON

```
{
    "numberoffiles" : 35551,
    "path" : [ "/eos/opstest/fts/tbtest/" ],
    "timestamp" : 1530540012,
    "totalsize" : 3000000000000,
    "usedsize" : 2928224959894,
    "vos" : [ "dteam" ]
}
```

[2] Storage Resource Reporting:
https://indico.cern.ch/event/587955/contributions/2936951/

Table of Contents

BEER (Batch on EOS Extra Resources)

See D. Smith's talk[3]

[3] https://indico.cern.ch/event/587955/contributions/2937728/

“Monster” machines

Goal: Lower the server overhead

- EOS also has a lifecycle/workflow engine
- EOS has Erasure Coding support
- Target is cold-er data

“Monster” machines

Storage node

- Compute node
- 10 (or 40) Gbit/s network interface
- 4x SAS expander

Storage array (8x)

- Dummy SAS array
- 24× 12TB drives

Fault-detection

EOS data transfers

- Diskserver to diskserver traffic
- Users see strange errors on `write()` or `close()`

"It's always the network!"

Fault-detection

Consul: distributed key-value store (and service catalog)

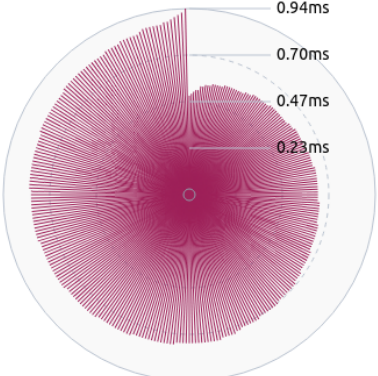
- Was meant for some internal experiment
- Nodes monitor each other[a]
- Ended up identifying possible network problems

[a] SWIM `http://www.cs.cornell.edu/info/projects/spinglass/public_pdfs/swim.pdf`

p05151207145936.cern.ch

Health Checks | Services | Round Trip Time

Minimum	0.50ms
Median	0.75ms
Maximum	0.94ms



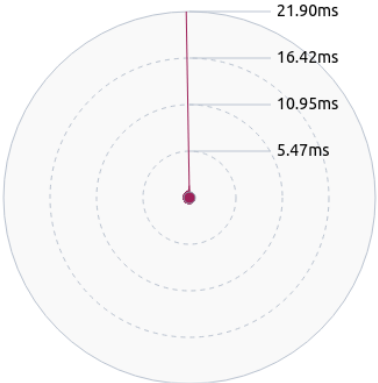
Fault-detection

```
2018/07/02 14:41:57 [WARN] memberlist: Was able to connect to lxfsrf16b03.cern.ch
↪  but other probes failed, network may be misconfigured
2018/07/02 15:06:32 [WARN] memberlist: Was able to connect to lxfsrf16b03.cern.ch
↪  but other probes failed, network may be misconfigured
2018/07/02 15:25:35 [WARN] memberlist: Was able to connect to lxfsrf16b03.cern.ch
↪  but other probes failed, network may be misconfigured
2018/07/02 15:43:41 [WARN] memberlist: Was able to connect to lxfsrf16b03.cern.ch
↪  but other probes failed, network may be misconfigured
2018/07/02 16:03:21 [WARN] memberlist: Was able to connect to lxfsrf16b03.cern.ch
↪  but other probes failed, network may be misconfigured
```

p06636710f31337.cern.ch

Health Checks | Services | Round Trip Time

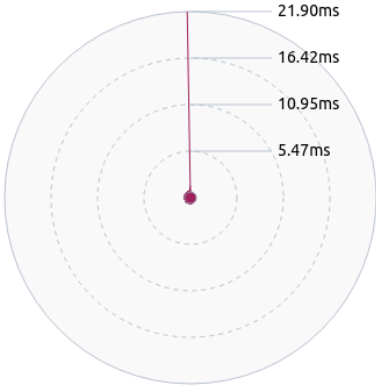
Minimum	0.51ms
Median	0.72ms
Maximum	21.90ms



p06636710f31337.cern.ch

Health Checks | Services | Round Trip Time

Minimum	0.51ms
Median	0.72ms
Maximum	21.90ms



lxfsrf16b03.cern.ch

Health Checks | Services | Round Trip Time

Minimum	21.67ms
Median	22.03ms
Maximum	22.30ms

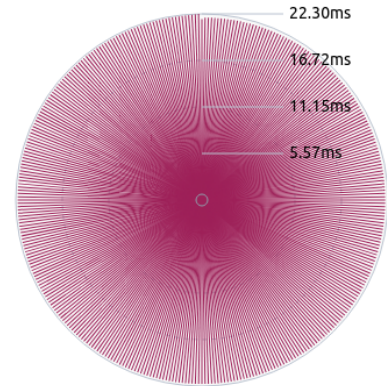
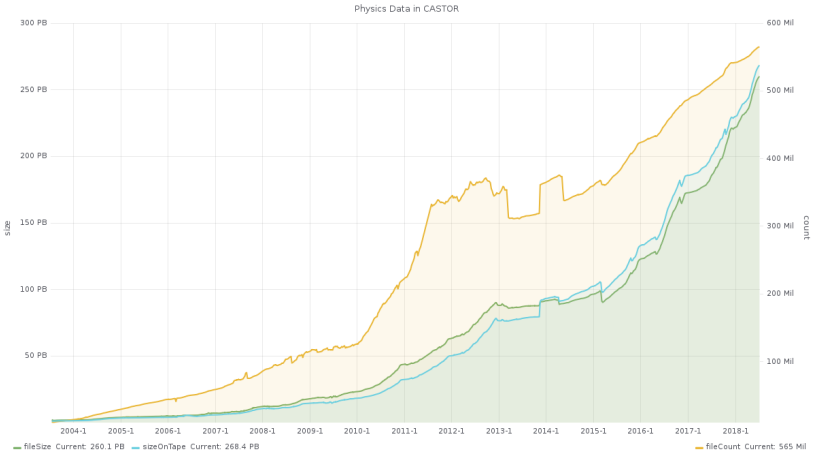


Table of Contents

CASTOR: Still going strong



S3: Simple Storage Service

HTTP-based object store (AWS S3-like) based on Ceph

- Became an official service this year[a]
- Pre-signed URLs, lifecycle policies, static websites
- ∼1 PB using Erasure Coding

[a] Mainly for disaster recovery use cases

NFS

Virtual NFS filer service

Currently

- Labour-intensive creation of new filers
- Performance doesn't scale horizontally

Evolving to Openstack-based self-service using CephFS

HPC

See “CephFS for HPC” talk in 45 minutes[4]

[4] https://indico.cern.ch/event/587955/contributions/2936868/

Thank you!

www.cern.ch