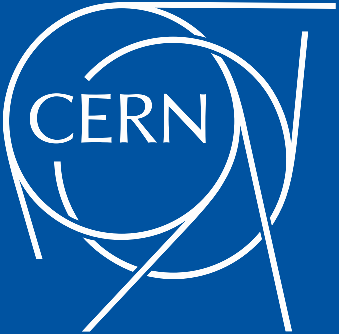

CERN

Providing large-scale disk storage

Herve Rousseau | on behalf of CERN IT Storage group

Table of Contents

EOS News

Optimizing resource usage

miscellaneous

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/

CentOS7

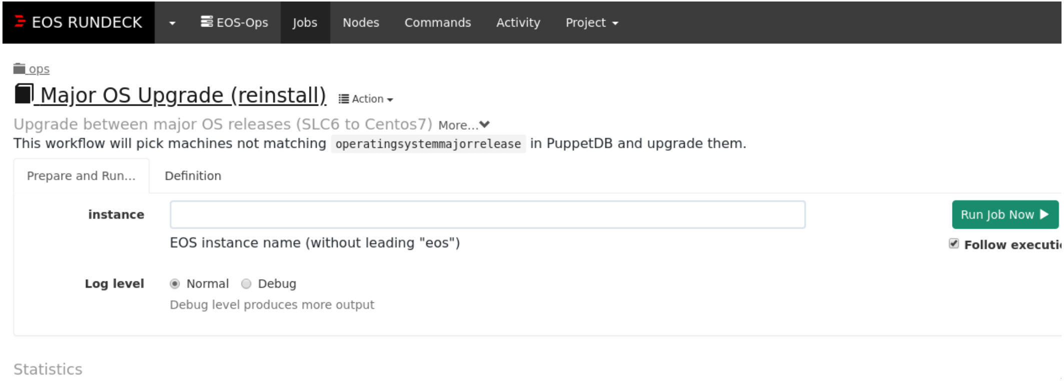
How to upgrade ~1300 machines with minimal disruption?

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



Major OS Upgrade (reinstall)

Upgrade between major OS releases (SLC6 to Centos7)

This workflow will pick machines not matching `operatingsystemmajorrelease` in PuppetDB and upgrade them.

instance

EOS instance name (without leading "eos")

Log level ● Normal ○ Debug

Debug level produces more output

WLCG Accounting

EOS now supports CRIC[2] compatible reporting

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

[2] Computing Resource Information Catalogue

Table of Contents

EOS News

Optimizing resource usage

miscellaneous

BEER (Batch on EOS Extra Resources)

See D. Smith's talk[3]

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

“Monster” machines

Storage node

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

Storage array (8x)

- Dummy SAS array
- 24x 12TB drives

“Monster” machines

Lower the server overhead

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

Fault-detection

EOS data transfers

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

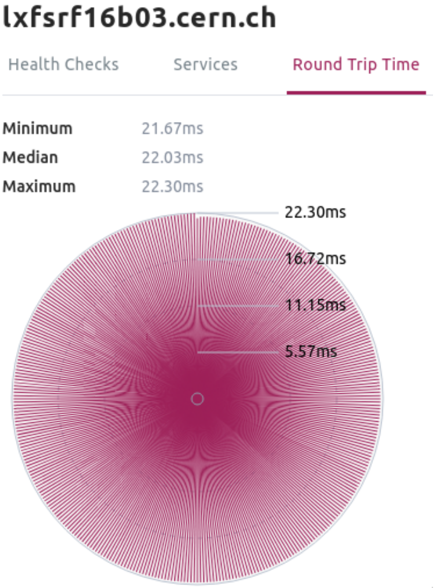
“It’s always the network !”

Fault-detection

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

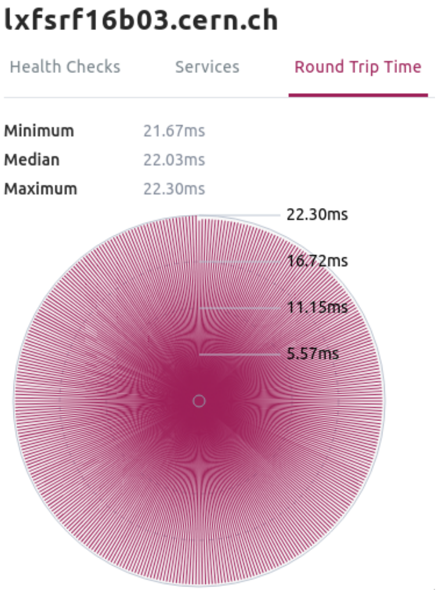
- Was meant for some internal experiment
- Nodes monitor each other[a]
- Ended up identifying possible network problems

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



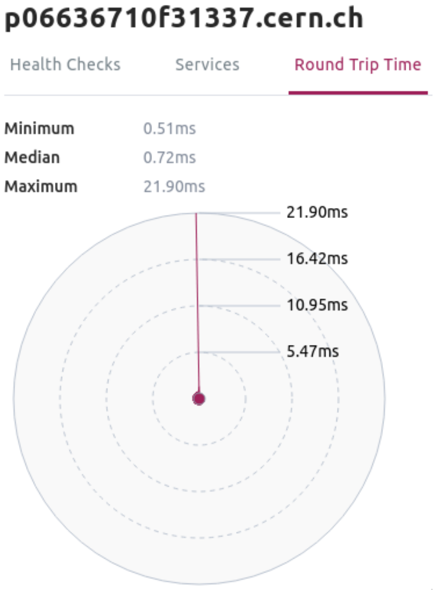
lxfsrf16b03.cern.ch

Minimum	21.67ms
Median	22.03ms
Maximum	22.30ms



lxfsrf16b03.cern.ch

Minimum	21.67ms
Median	22.03ms
Maximum	22.30ms



p06636710f31337.cern.ch

Minimum	0.51ms
Median	0.72ms
Maximum	21.90ms

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
```

CERNBox

See H. Gonzalez Labrador's talk
https://indico.cern.ch/event/587955/contributions/2936817/

Table of Contents

EOS News

Optimizing resource usage

miscellaneous

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
- IPv6-only internal traffic in the cluster

[a] Mainly for disaster recovery use cases

NFS

Virtual NFS filer service

Currently

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

Evolving to Manila-based self-service using CephFS

HPC

CephFS for HPC:
https://indico.cern.ch/event/587955/contributions/2936868/

Thank you !

www.cern.ch