Ceph: A Distributed File System

Audience/Presented to LC Staff Meeting

M. Tran, M. Wan, L. Zhang

August 10, 2017

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Introduction

- What is a Distributed File System (DFS)?
  - A file system that permits various hosts on separate machines to access and share files through a computer network.
  - Data may be distributed across many nodes, but users can access their files as though they were stored on one server.

- Why use distributed file systems?
  - High availability
  - Redundancy
  - Location-independent access
  - Scalability

- Why Ceph?
  - Provides block and file storage
  - Can handle large-scale file systems
  - Reduces traffic to metadata clusters using CRUSH algorithm
  - POSIX compliant

Source: http://3.bp.blogspot.com/-B_UA0D0I6xI/T2ycHjkPdvI/AAAAAAAAAIs/nIm3cjymTwk/s1600/dfs.jpg

How does Ceph work?

- Components of the Ceph Storage Cluster
  - Monitors
  - Managers
  - Object Storage Daemons
  - Metadata Server (for use with Ceph File System)

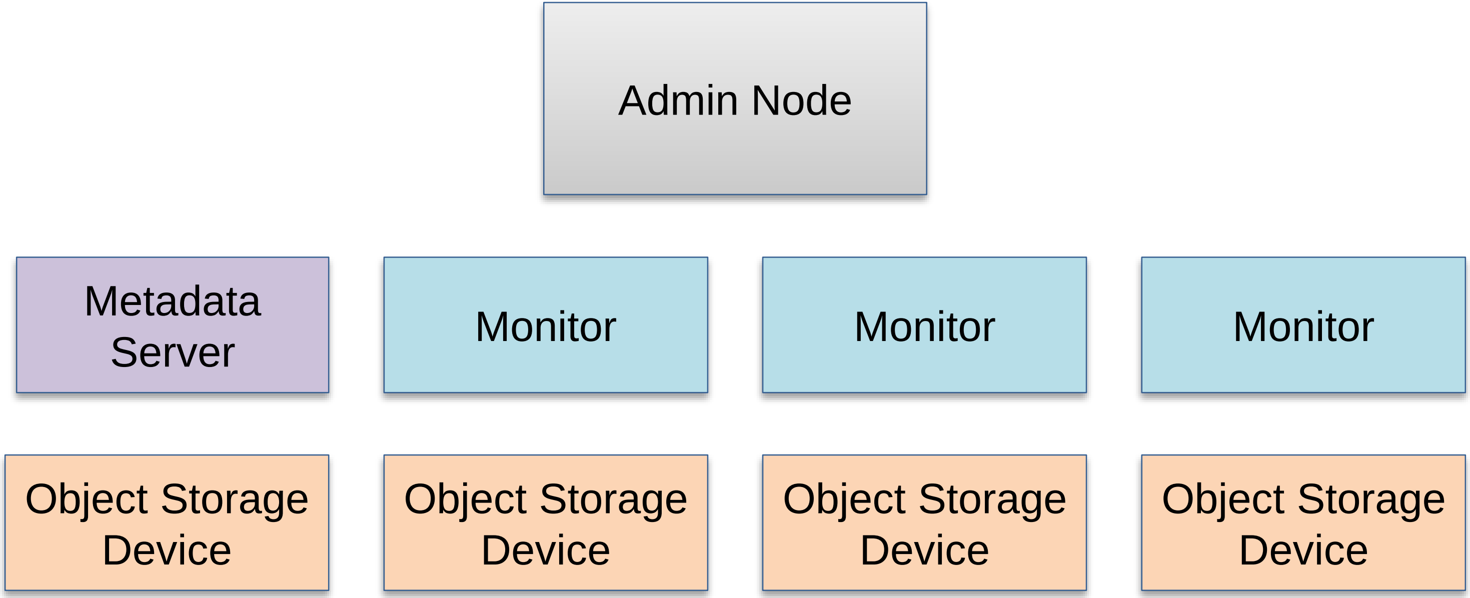
- Stores data as objects within logical storage pools

- CRUSH algorithm
  - Controlled Replication Under Scalable Hashing
  - Determines which OSD stores the placement groups
  - Enables scaling, rebalancing, and recovery dynamically

- Ceph File System
  - POSIX-compliant interface
  - Files are mapped to objects and stored in the Ceph Storage Cluster.
  - Metadata Server prevents filesystem operations from consuming resources excessively

Source: http://ceph.com/wp-content/uploads/2016/07/Ceph_Logo_Stacked_RGB_120411_fa.png

Ceph Storage Cluster



Admin Node

Metadata Server

Monitor

Monitor

Monitor

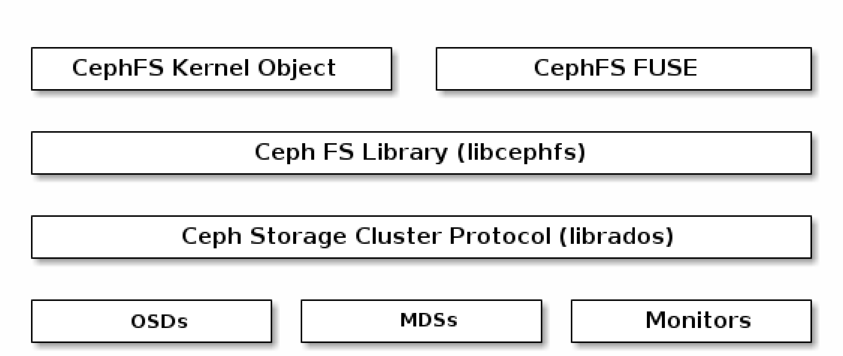
Object Storage Device

Object Storage Device

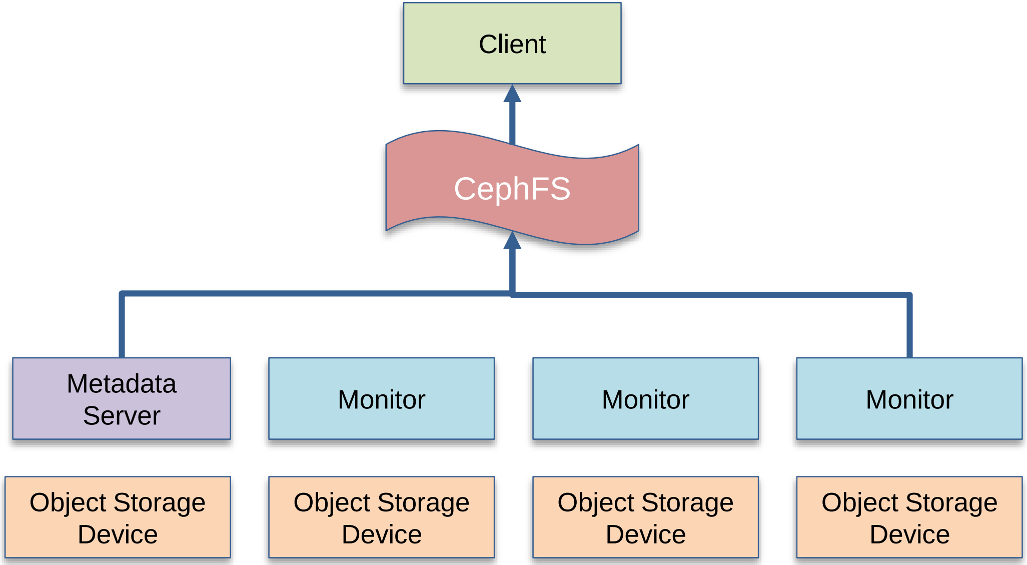
Object Storage Device

Object Storage Device

Ceph File System


CephFS Kernel Object
CephFS FUSE
Ceph FS Library (libcephfs)
Ceph Storage Cluster Protocol (librados)
OSDs
MDSs
Monitors

Source: http://docs.ceph.com/docs/master/_images/ditaa-1cae553f9d207d72257429d572673632afbd108c.png



Client

CephFS

Metadata Server

Monitor

Monitor

Monitor

Object Storage Device

Object Storage Device

Object Storage Device

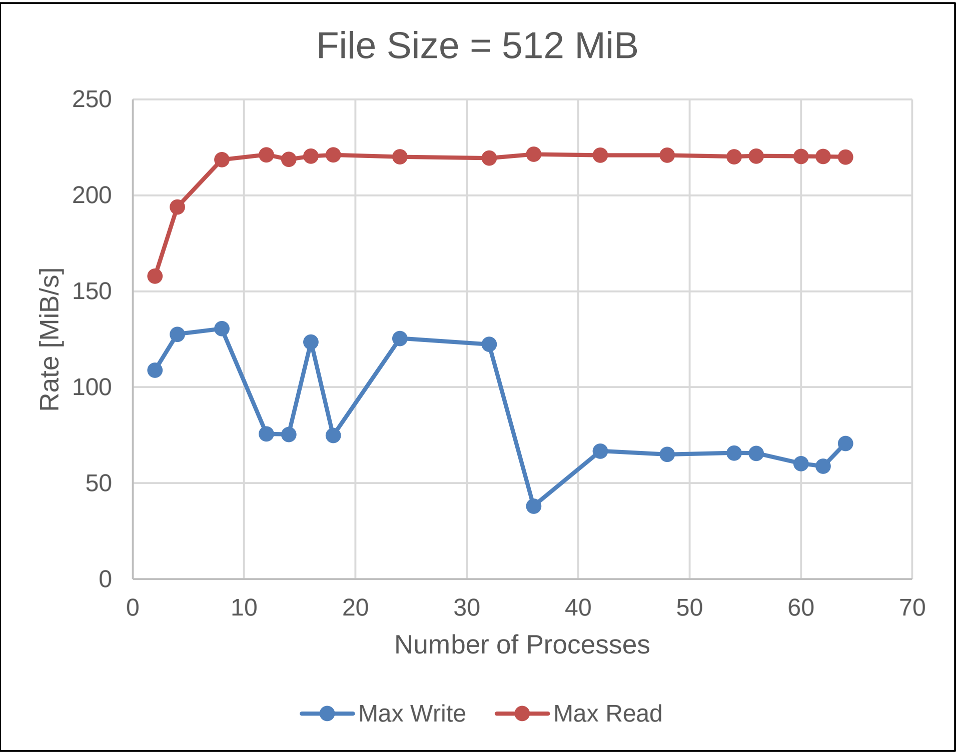
Object Storage Device

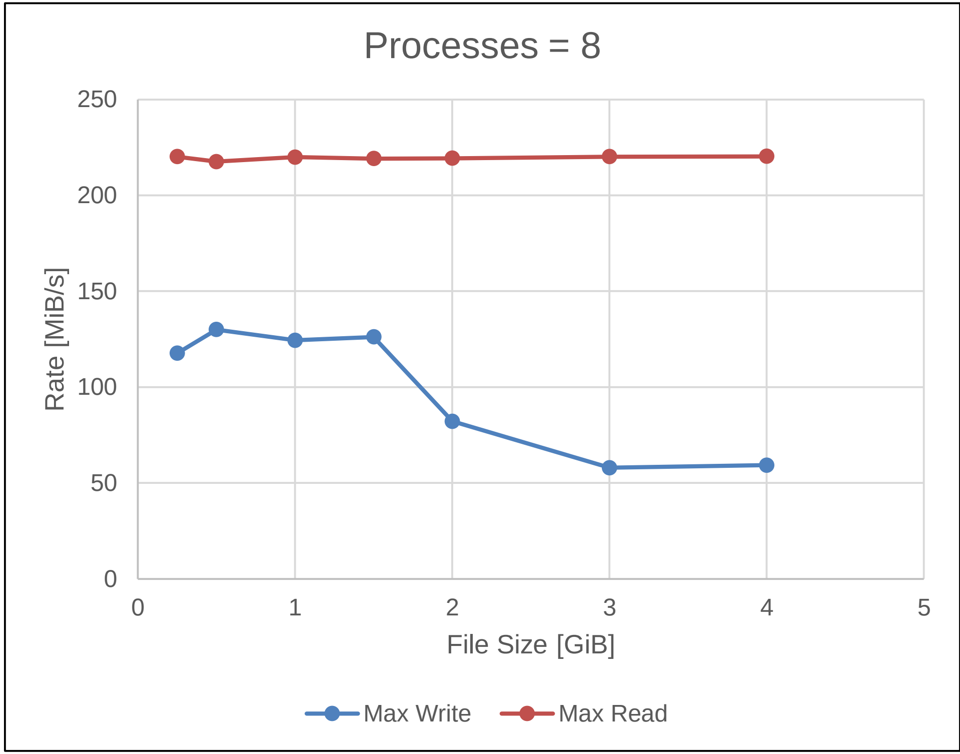
Benchmarking & Results

- Tested POSIX compliance using the POSIX Filesystem Test Suite
  - Passed 1951/1957 tests; failed 6/1957
  - Most UNIX systems aren’t 100% POSIX compliant
- Tested read/write speeds using IOR
- Tested file creation/deletion speeds with mdtest

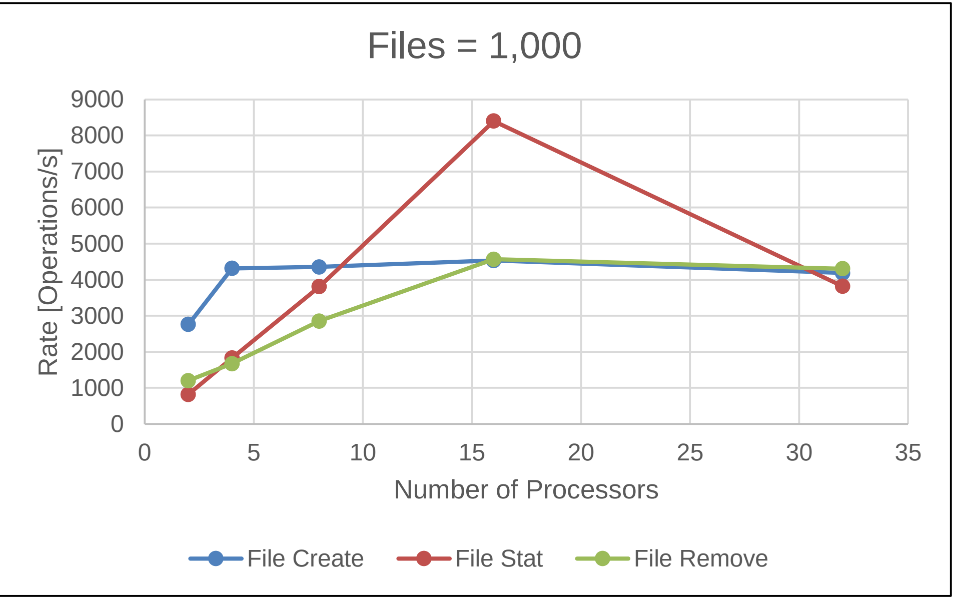
https://www.iag.biz/wp-content/uploads/2016/08/223-Executive-Summary-Business-Analysis-Benchmark.jpg

IOR Performance Testing

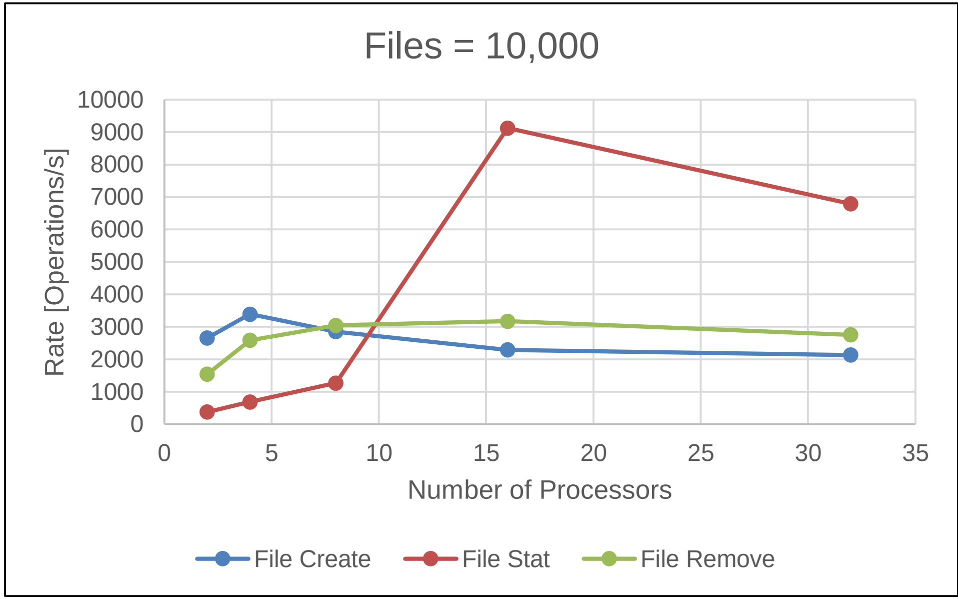




mdtest

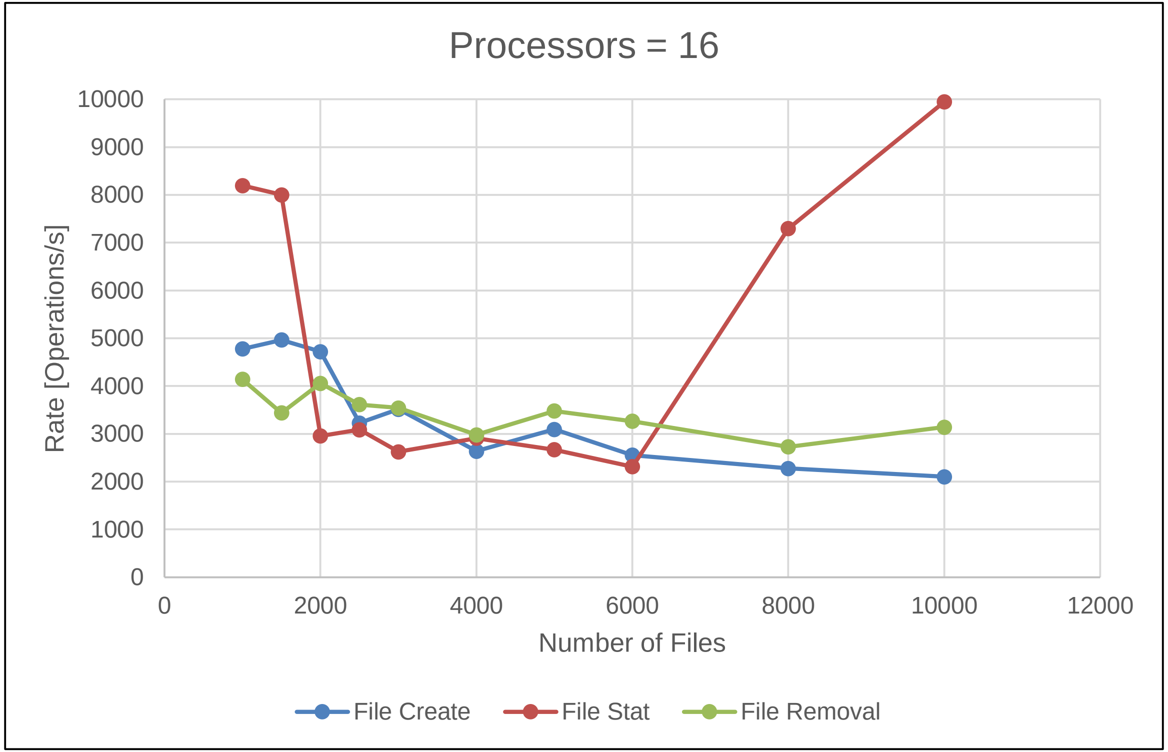


Files = 1,000
Rate [Operations/s]
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
35
Number of Processors
File Create
File Stat
File Remove



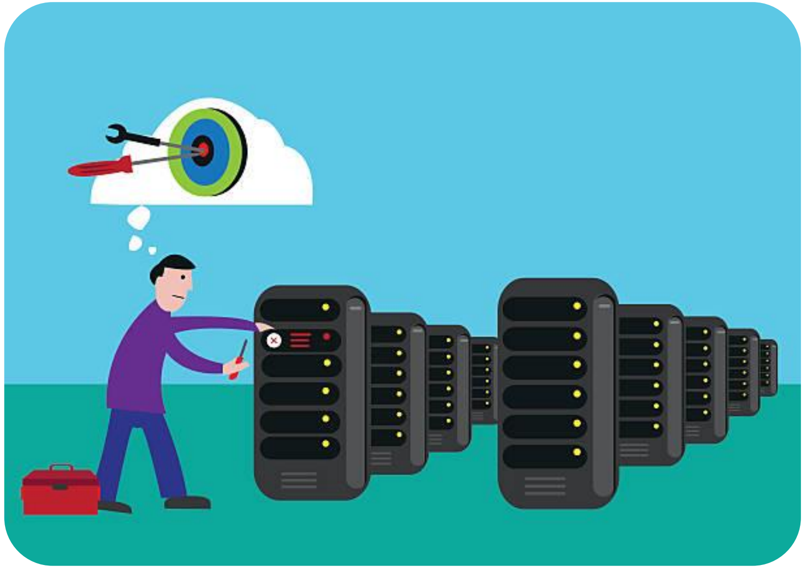
Files = 10,000
Rate [Operations/s]
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
35
Number of Processors
File Create
File Stat
File Remove

mdtest


Processors = 16
Rate [Operations/s]
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000
12000
Number of Files
File Create
File Stat
File Removal

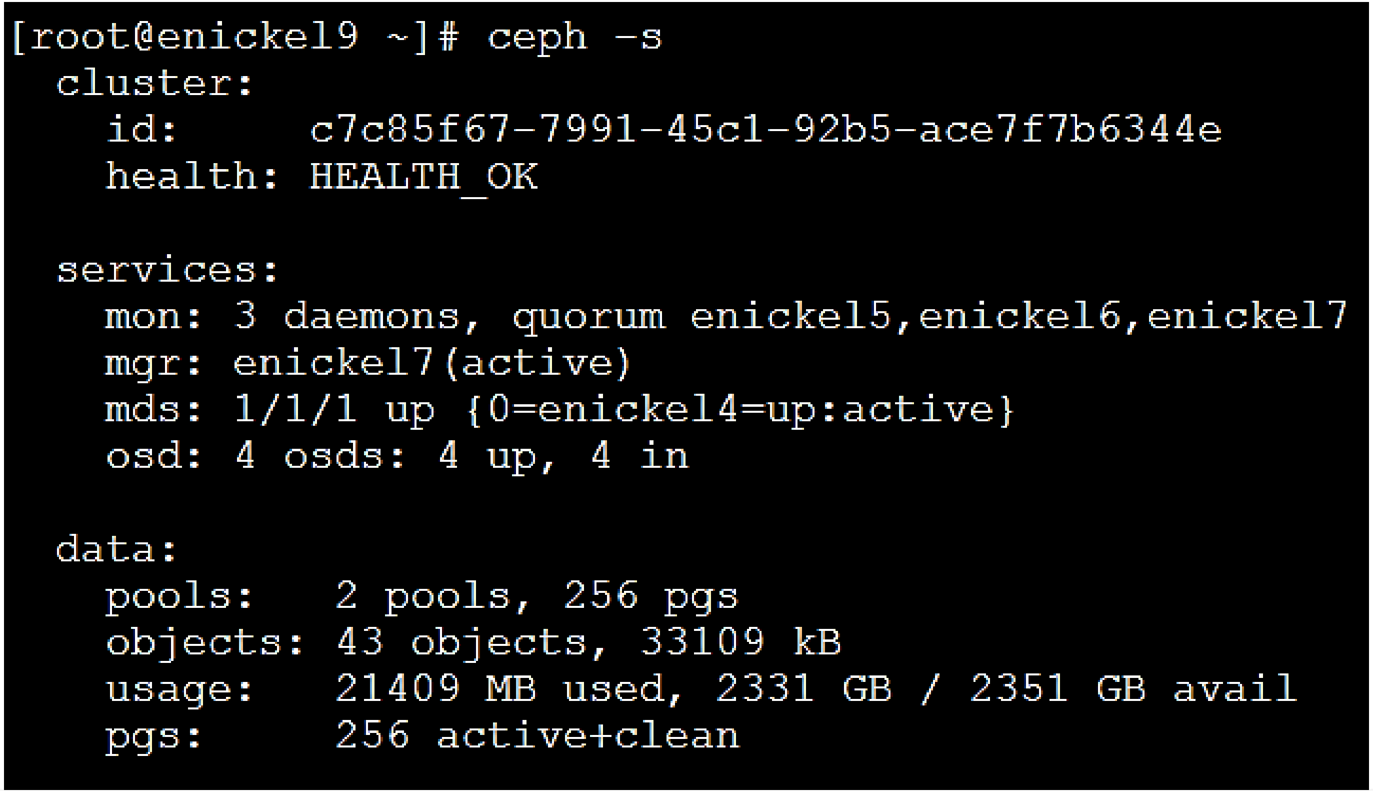
Failover Testing

- Demonstrate that our implementation of CephFS can survive a failure of up to ¼ of the total system.
- Explore how the system responds when an OSD, a monitor, or a manager fails.



Source: http://www.istockphoto.com/illustrations/failover?excludenudity=true&sort=mostpopular&mediatype=illustration&phrase=failover

Initial State: Healthy

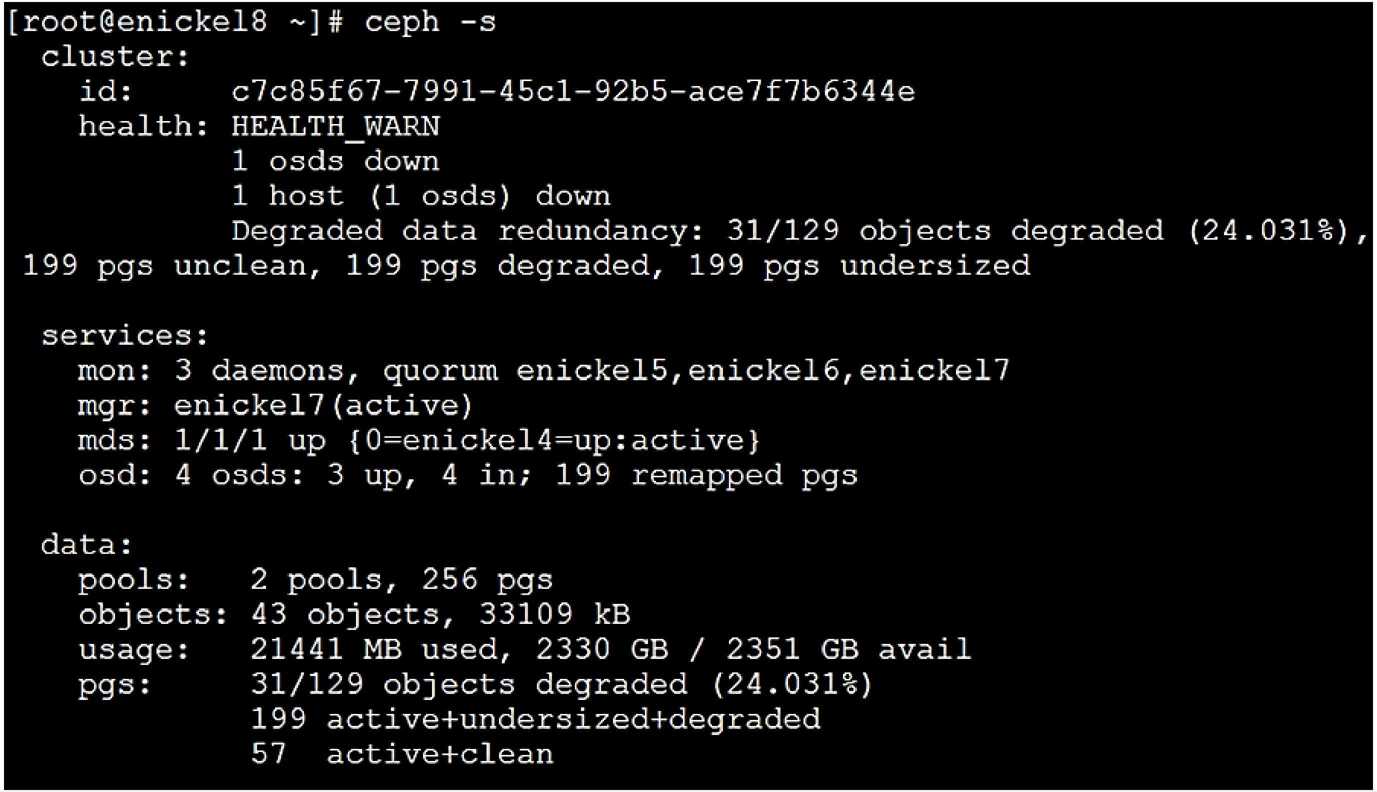


```
[root@enickel9 ~]# ceph -s
  cluster:
    id:     c7c85f67-7991-45c1-92b5-ace7f7b6344e
    health: HEALTH_OK

  services:
    mon: 3 daemons, quorum enickel5,enickel6,enickel7
    mgr: enickel7(active)
    mds: 1/1/1 up {0=enickel4=up:active}
    osd: 4 osds: 4 up, 4 in

  data:
    pools:   2 pools, 256 pgs
    objects: 43 objects, 33109 kB
    usage:   21409 MB used, 2331 GB / 2351 GB avail
    pgs:     256 active+clean
```

After Taking Down an OSD

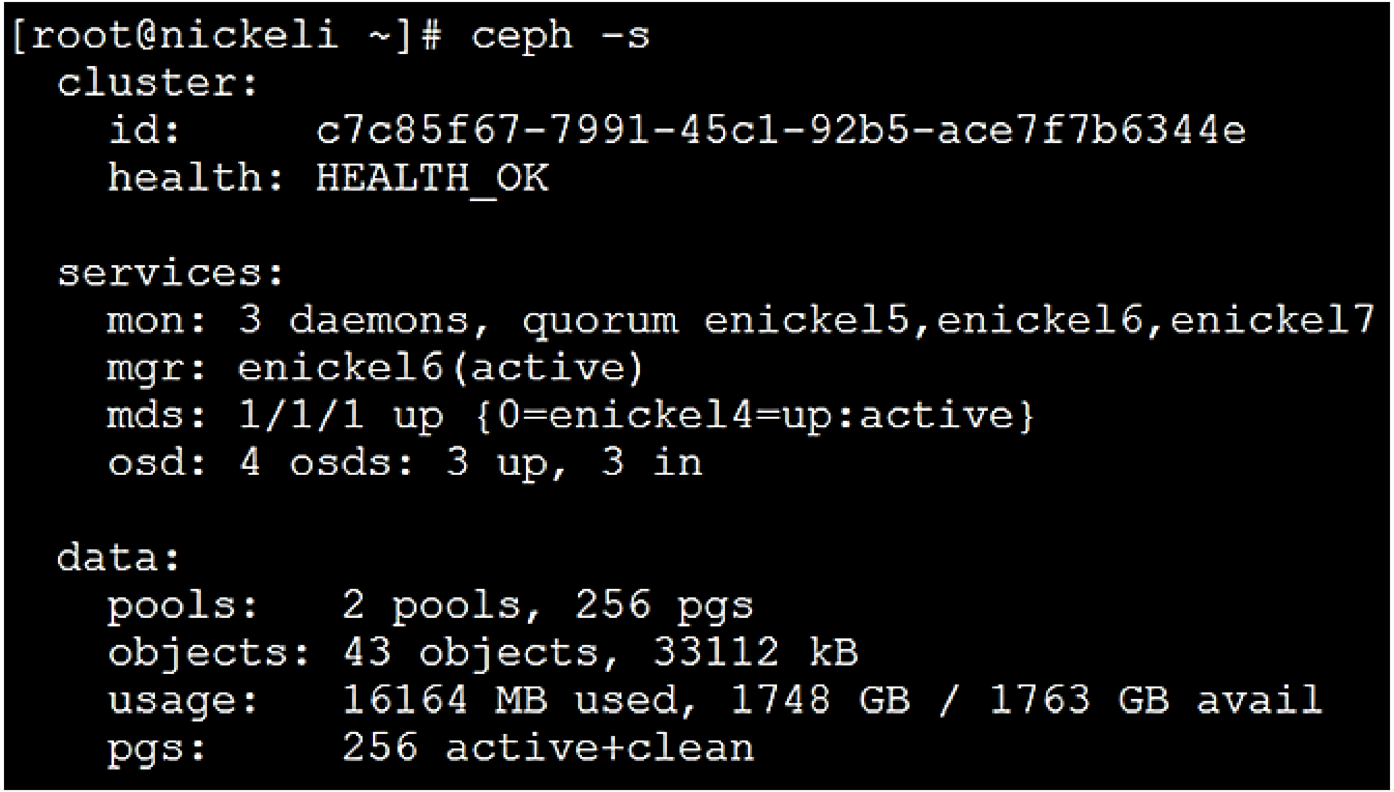


```
[root@enickel8 ~]# ceph -s
  cluster:
    id:     c7c85f67-7991-45c1-92b5-ace7f7b6344e
    health: HEALTH_WARN
            1 osds down
            1 host (1 osds) down
            Degraded data redundancy: 31/129 objects degraded (24.031%),
 199 pgs unclean, 199 pgs degraded, 199 pgs undersized

  services:
    mon: 3 daemons, quorum enickel5,enickel6,enickel7
    mgr: enickel7(active)
    mds: 1/1/1 up {0=enickel4=up:active}
    osd: 4 osds: 3 up, 4 in; 199 remapped pgs

  data:
    pools:   2 pools, 256 pgs
    objects: 43 objects, 33109 kB
    usage:   21441 MB used, 2330 GB / 2351 GB avail
    pgs:     31/129 objects degraded (24.031%)
             199 active+undersized+degraded
             57  active+clean
```

Recovered State

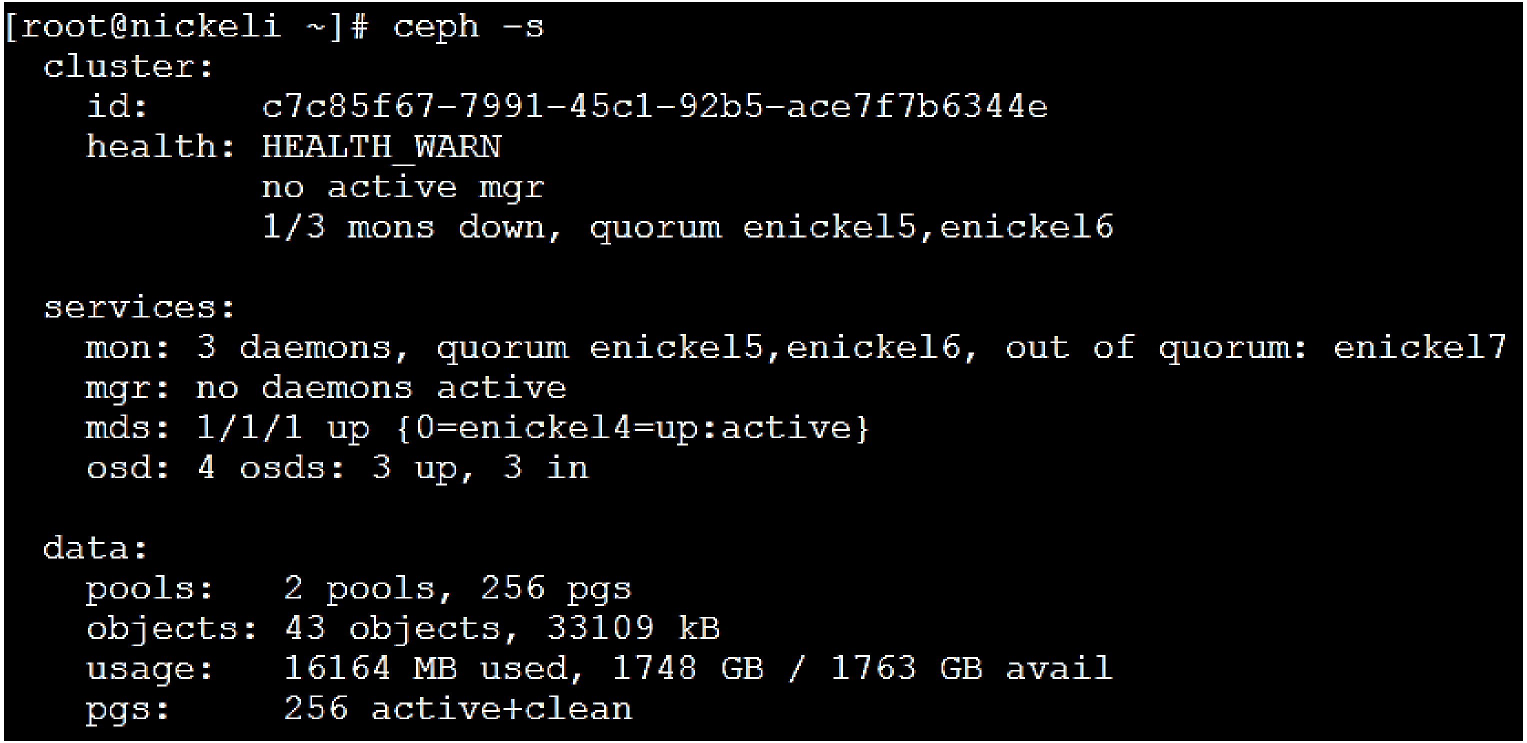


```
[root@nickeli ~]# ceph -s
  cluster:
    id:     c7c85f67-7991-45c1-92b5-ace7f7b6344e
    health: HEALTH_OK

  services:
    mon: 3 daemons, quorum enickel5,enickel6,enickel7
    mgr: enickel6(active)
    mds: 1/1/1 up {0=enickel4=up:active}
    osd: 4 osds: 3 up, 3 in

  data:
    pools:   2 pools, 256 pgs
    objects: 43 objects, 33112 kB
    usage:   16164 MB used, 1748 GB / 1763 GB avail
    pgs:     256 active+clean
```

After Taking down a Monitor

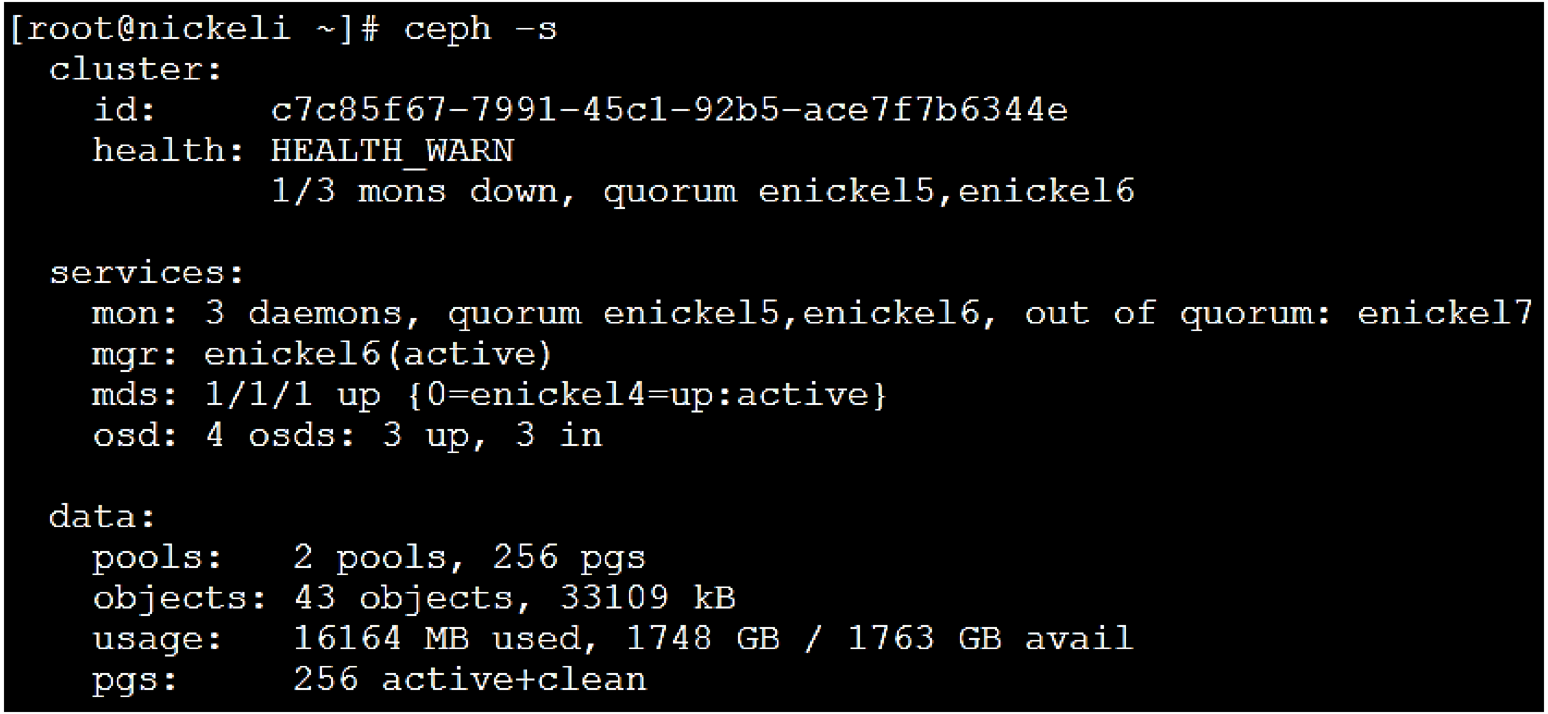


```
[root@nickeli ~]# ceph -s
  cluster:
    id:     c7c85f67-7991-45c1-92b5-ace7f7b6344e
    health: HEALTH_WARN
            no active mgr
            1/3 mons down, quorum enickel5,enickel6

  services:
    mon: 3 daemons, quorum enickel5,enickel6, out of quorum: enickel7
    mgr: no daemons active
    mds: 1/1/1 up {0=enickel4=up:active}
    osd: 4 osds: 3 up, 3 in

  data:
    pools:   2 pools, 256 pgs
    objects: 43 objects, 33109 kB
    usage:   16164 MB used, 1748 GB / 1763 GB avail
    pgs:     256 active+clean
```

Partially Recovered



```
[root@nickeli ~]# ceph -s
  cluster:
    id:     c7c85f67-7991-45c1-92b5-ace7f7b6344e
    health: HEALTH_WARN
            1/3 mons down, quorum enickel5,enickel6

  services:
    mon: 3 daemons, quorum enickel5,enickel6, out of quorum: enickel7
    mgr: enickel6(active)
    mds: 1/1/1 up {0=enickel4=up:active}
    osd: 4 osds: 3 up, 3 in

  data:
    pools:   2 pools, 256 pgs
    objects: 43 objects, 33109 kB
    usage:   16164 MB used, 1748 GB / 1763 GB avail
    pgs:     256 active+clean
```

Discussion

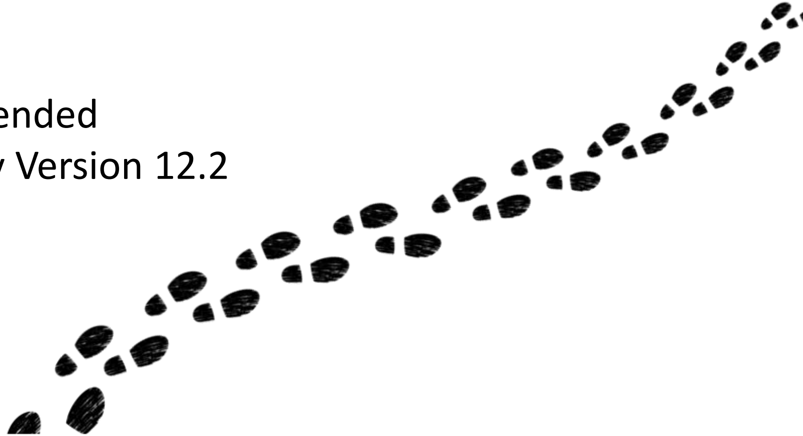
Challenges	Solutions
LVMs already installed	Installed OSDs manually
Benchmarking: Tests read from cache	Re-ran tests using 2 clients
New version released halfway through	Updated all of our nodes to Version 12.1.2
Not enough troubleshooting documentation	Trial and error; reinstalling Ceph

Next Steps

- Integrate Ceph with NFS
  - We would like to mount CephFS on clients that don’t have Ceph installed.
  - Currently, we do this by having one node of the cluster act as a NFS server.
  - This methods is flawed: if the NFS server goes down, clients lose access to the file system.

- Improve performance, particularly write speeds

- Incorporate additional metadata servers
  - Multiple metadata servers is not currently recommended
  - Ceph plans to support multiple metadata servers by Version 12.2



Thank you for your help and support!

Thomas Bennett

Elsa Gonsiorowski

Dave Fox

Geoff Cleary

Bryan Dixon

Pam Hamilton

Go Team Cephalopod!

Source: https://www.pinterest.com/pin/445504588117025745

Lawrence Livermore
National Laboratory