Caltech

Ceph Update

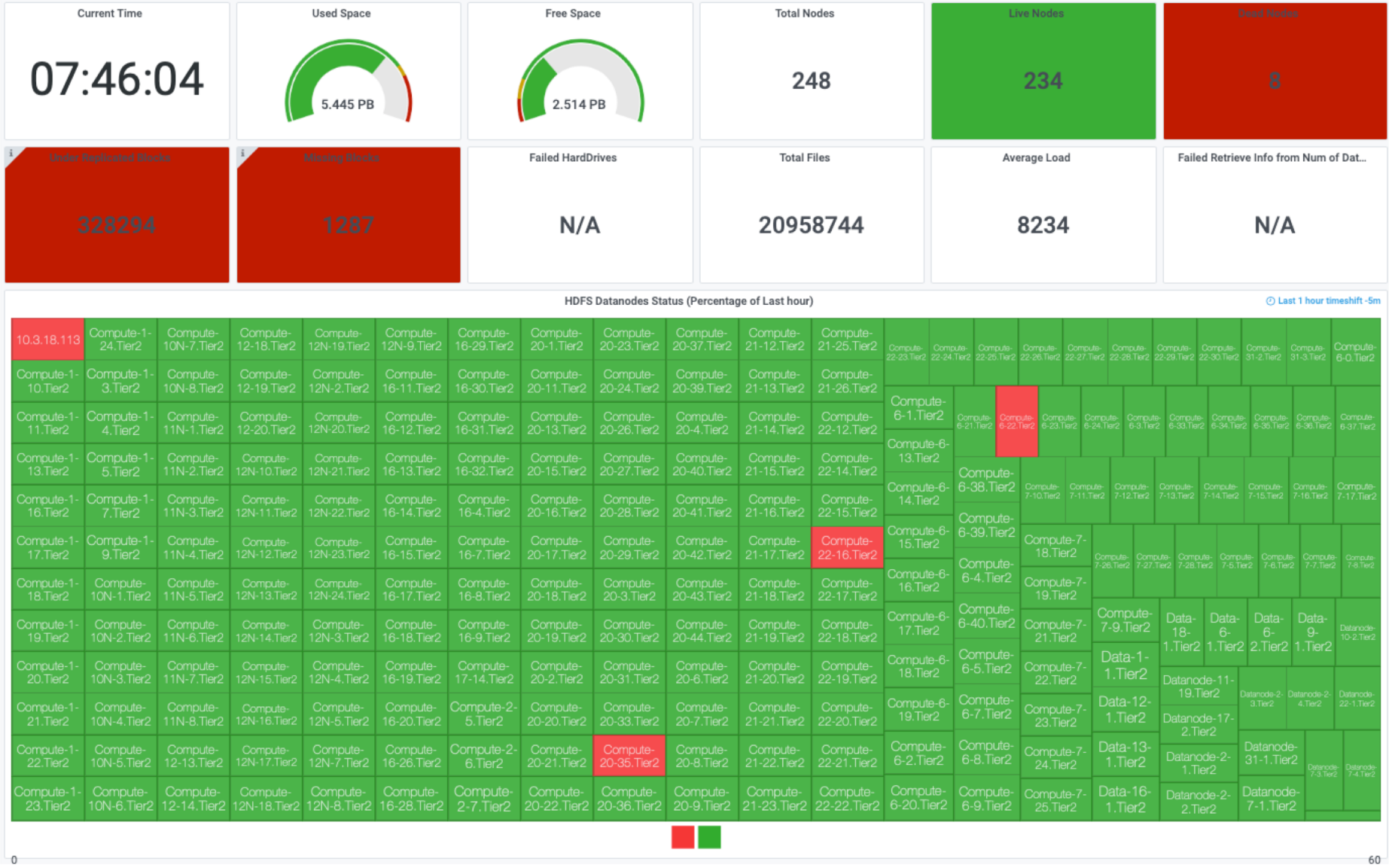
J. Balcas

Continuation update from OSG AHM 2020: https://indico.fnal.gov/event/22127/contributions/194938/

Hadoop Cluster (in Prod since Feb. 2009)

Caltech

- Moved from dCache in 2009.
- 1 Metadata Server, 234 Datanatodes (moving to big JBODs).
- **Running 2.6 Release (last latest available from OSG).**
- **8PB RAW (Usable 4PB). EC not available till 3.0 Release. EC Not good for small files. See link[1]**
- **Network is mixed of 1G, 10G, 40Gb.**
- **Used to store all CMS Data (anything what fits under /store).**
- **Using Tiered Storage with SSDs. Manually enforcing speed up for some users.**



[1] https://blog.cloudera.com/hdfs-erasure-coding-in-production/

Why do we want to move from HDFS?

Caltech

- **Open Source:**
  - HDFS/OSG Issue: Cloudera used to distribute source RPMs for CDH 5, which contains Hadoop 2.6.0. However, with CDH 6 containing Hadoop 3.0.0, the only distribute binary RPMs and require installation of Cloudera Manager for installation and license management. (Of course we tried to go tarball way – but this does not look production ready type installation). CEPH so far is fully opensource (clone, modify, propose, redeploy, build)
- **Reliable and scalable:**
  - HDFS Does not have Rolling upgrades till 3.0 release. Metadata single point of failure (or need to run standby/active via NFS

★ **Erasure coding (Rep 2 || 3 is expensive)- gain space on same hardware; We could go to HDFS 3.x – still does not work for small files. See** https://blog.cloudera.com/hdfs-erasure-coding-in-production/

★ **POSIX Compliant for user home dirs, analysis facilities. (jupyter notebooks, gpu access)**

★ **One FS for many purposes: Block device (for containers, VMs,), Single FS**

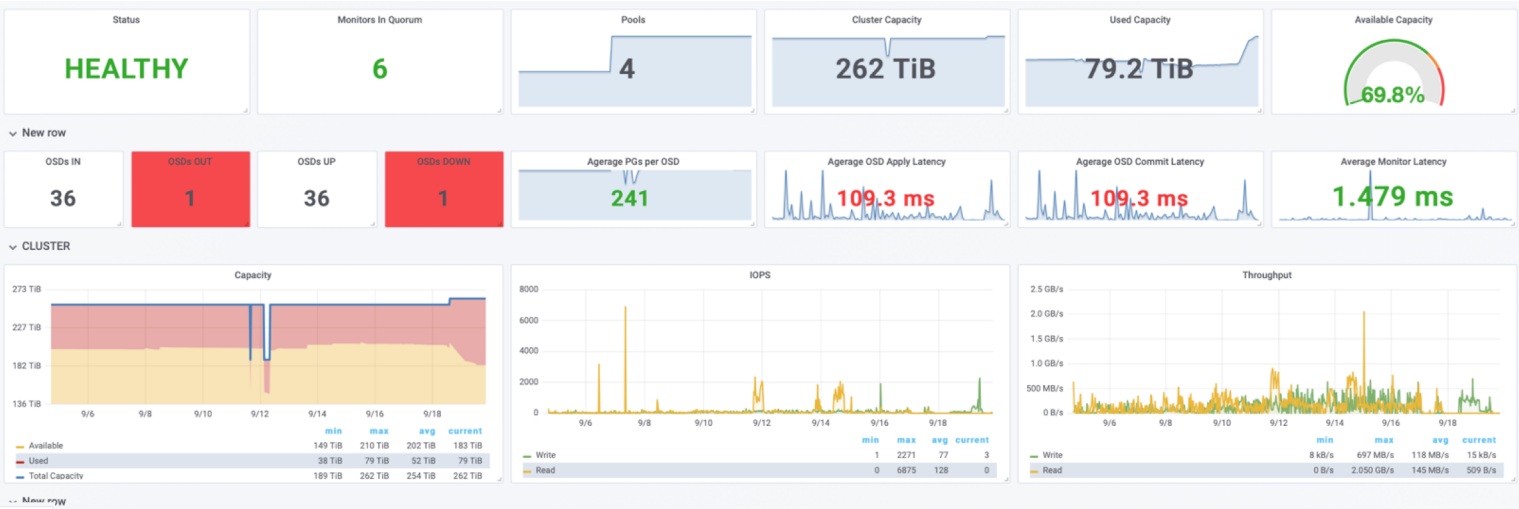
- **Latest C libhdfs does not have xattrs**

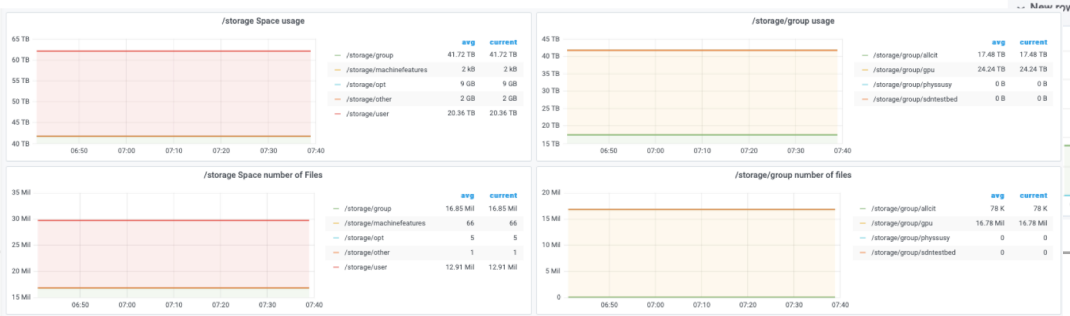
★ **Running 3 diff FS is a burden on admins**

Small CEPH Cluster (in Prod since 2018)

Caltech

- We have **6 monitor hosts, 3 storage hosts, and 2 metadata hosts (1 in use)**
- **262TB total of HGST 12Gb SAS3 disks** in the 3 storage hosts (Rep 2, No Erasure coding) + 6 PCI NVMe P4600 for Metadata
- **Network is 40Gb to the 3 storage hosts**
- Upgraded from Jewel -> Luminous -> Mimic (13.2.4) -> Mimic (13.2.6) -> (Octopus since December 15.2.9)
- **Used as HOME DIR and as Analysis Facility Storage**
- **Mounted on all Tier2, already used by CRAB, Local Condor Jobs, interactive access, GPU Workflows, Jupyter**
- **Block device for containers (Docker, OpenVZ)**





Movement plans for CMS Storage

Caltech

- **Storage situation:**
  - 8 PB RAW in HDFS
  - 1.4 PB RAW in very old hardware; Will not be moved to Ceph
  - 1 PB in a new hardware JBOD
  - We are left with 7.5 PB (~7PB RAW due to mix OS disk/data)
- **2019 Oct – 2020 Dec:**
  - Different Erasure coding and failure tests
  - Memory Usage decrease tests
  - New Storage Element in CMS and testing with CMS Transfer tools (Rucio, FTS)
- **2021 Jan – 2021 Mar:**
  - Infrastructure preparation: Clean OpenVZ, UPS, Puppet code changes to support multiple separate FS
  - XRootD and GridFTP tests and hacks (more details later)
  - Drain 2PB from HDFS, Move to Ceph
  - Redirect all /store/{unmerged,temp} to CEPH
- **2021 Apr – 2021 Jun:**
  - ***<u>Pacific release! And multiple FS is not experimental anymore</u>***.
  - Drain HDFS, Decrease CMS Space, Move datanodes to Ceph.
  - Move users 1 by 1 (~500TB)
- **2021 Summer – Only Ceph, Bye Hadoop**

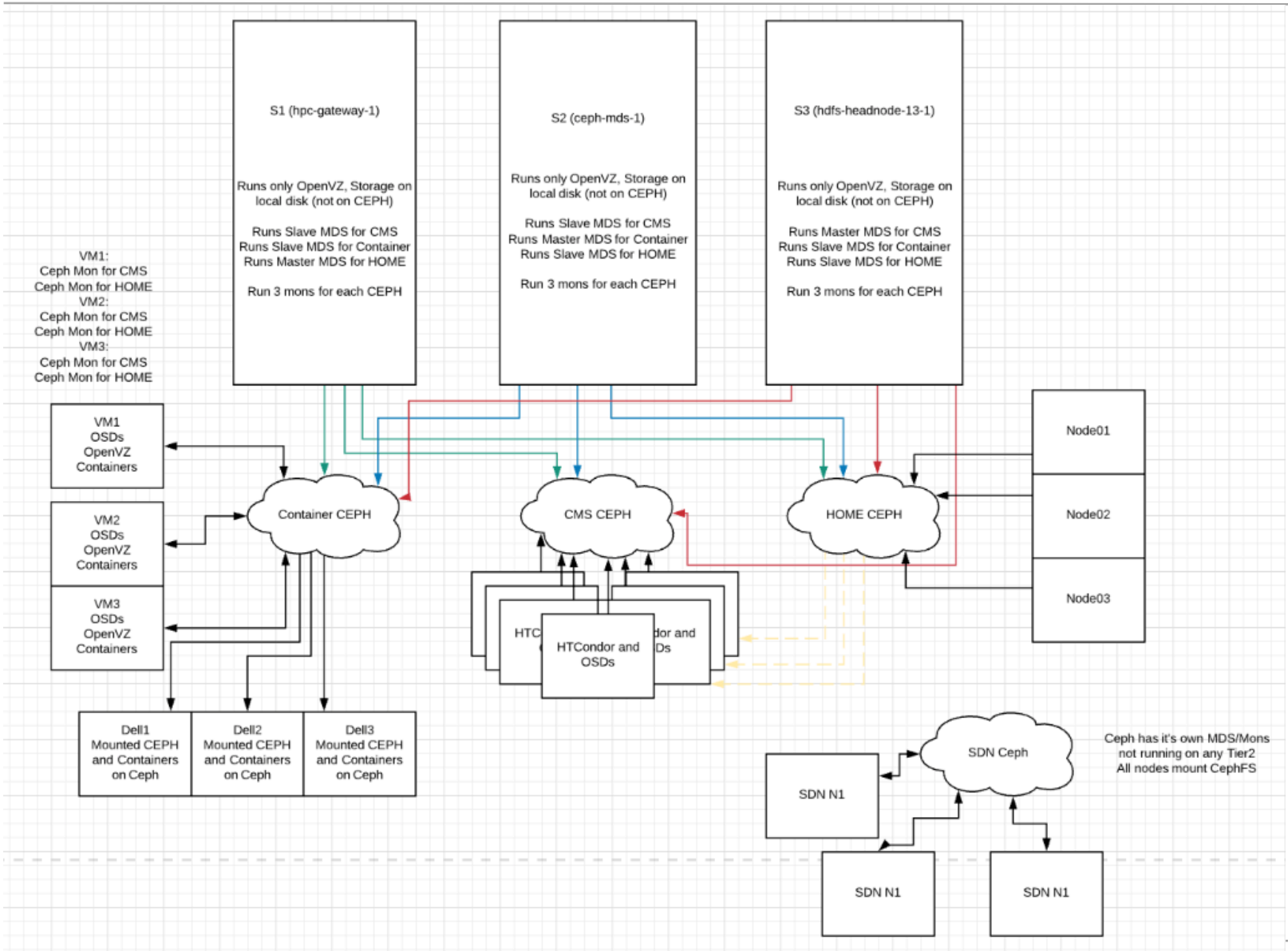
Ceph Cloud Storage Erasure Coding Calculations

Num of OSDs	880	Raw Physical Space (TB)	7,040
*~Storage Overhead Ratio	0.0907	*~Raw Storage Overhead (TB)	638
Full Ratio	0.95	Raw Full Ratio Space (TB)	7,008
OSD Size (TB)	8	Raw Storage Avail (TB)	6,370

Usable TB Matrix	M					
K	1	2	3	4	5	6
1	3,041	2,027	1,520	1,216	1,014	869
2	4,054	3,041	2,433	2,027	1,738	1,520
3	4,561	3,649	3,041	2,606	2,281	2,027
4	4,865	4,054	3,475	3,041	2,703	2,433
5	5,068	4,344	3,801	3,379	3,041	2,764
6	5,213	4,561	4,054	3,649	3,317	3,041
7	5,321	4,730	4,257	3,870	3,547	3,275
8	5,406	4,865	4,423	4,054	3,742	3,475
9	5,473	4,976	4,561	4,210	3,909	3,649
10	5,529	5,068	4,678	4,344	4,054	3,801
11	5,575	5,146	4,778	4,460	4,181	3,935
12	5,614	5,213	4,865	4,561	4,293	4,054
13	5,647	5,271	4,941	4,650	4,392	4,161
14	5,676	5,321	5,008	4,730	4,481	4,257
15	5,701	5,366	5,068	4,801	4,561	4,344

Tier2 Plans (multiple FS while on Octopus)

Caltech



3 Dedicated machines (S1,S2,S3):
2x E5-2640 v4 @ 2.40GHz
8x 16GB DDR4 2133 (up to 32x)
6x SAMSUNG MZ7KM480 (up to 10x)
1x Each node has dedicated UPS

Each FS:
Container/CMS/Home Runs:
3 MDS (1active, 2 slave)
5 Mons (all active)

Even 2 machines(S1/S2/S3) down, it will not affect cluster.
Up to 2 mons down supported;

- CephFS and XRootD/GridFTP
- Ceph Object Storage and XRootD/GridFTP
- Memory consumption
- Lessons learned hard way (Kernel mount)

CephFS (XRootD/GridFTP)

Caltech

- GridFTP (EOL 2021/2022 [1]):
  - Use GridFPT POSIX DSI (Data Storage Interface)
  - Checksum calculation: export GRIDFTP_CKSUM_EXT_ADLER32="/usr/bin/xrdadler32”
- XRootD:
  - Use XRootD Multiuser plugin:
    - Checksum issue:
      https://github.com/opensciencegrid/xrootd-multiuser/issues/14
      https://github.com/xrootd/xrootd/issues/1294
    - We run custom XRootD (taken from stable release 4.12.x): Once XrdSysFAttr is called – it writes checksum metadata under /tmp/xrootd/cksums
    - Along with it – another python script reads it and publish as xattr in CephFS;

[1] See Diego’s talk: https://indico.fnal.gov/event/47040/contributions/208459/

Ceph Object Storage (XRootD/GridFTP)

Caltech

- GridFTP (EOL 2021/2022):
  - Use Ceph DSI Plugin from RAL: https://github.com/stfc/gridFTPCephPlugin
- Xrootd:
  - Use RAL Plugin: https://github.com/stfc/xrootd-ceph
- Issues so far:
  - OSG is not building XRootD/GridFTP Ceph plugins. Doable – but…
    - XRootD plugin depends on a specific Ceph Release (14.2.15 Nautilus), while our clusters already 15.2.9 (Octopus) and we look a lot at 16.0 (Pacific)
    - XRootD developers stated that it would be a disaster to maintain separate RPMs for each Ceph Release (not only complication for maintainer, but also for an end user).
    - What are OSG plans for storage support now and in HL-LHC? HDFS Already dropped? (**Question after in Ajit's talk**)
  - Vector Reads are in development. More details: https://github.com/xrootd/xrootd/issues/1259
  - XRootD Ceph TPC Issues. James Walder working on it. Many of the issues found have been handled by the core XrootD team however there are some that fall on us such as how XrdCeph handles the mkdir call and the command to calculate checksums on large files correctly
  - We already moved to 15.2.8 and XRootD plugins do not build nicely
    - From the performance point – in our tests CephFS was performing better than any of these plugins

Memory requirements/consumption

Caltech

Process	Criteria	Mimic/Nautilus	Octopus (latest stable) Pacific (So far)	Is it an issue?
ceph-osd	Processor	1x 64-bit AMD-64 1x 32-bit ARM dual-core or better	1 core minimum; 1 core per 200-500 MB/s 1 core per 1000-3000 IOPS Results are before replication. Results may vary with different CPU models and Ceph features. (erasure coding, compression, etc)	**Huge memory decrease. 4GB/disk sustainable (most nodes 2.5GB/core, 2GB for condor) This year we decommission all <2gb/core nodes and all <24 core nodes Most nodes are minimum dual 10G/25G/40G Also for osd memory (scrub,recovery) – we limit it to 1(min)/node. If things break (lost osd, node) – we want to make decision: Restore hardware if possible or increase backup.**
	RAM	~1GB for 1TB of storage per daemon	4GB+ per daemon (more is better) 2-4GB often functions (may be slow) Less than 2GB not recommended	
	Volume Storage	1x storage drive per daemon	1x storage drive per daemon	
	Journal (DB/WAL)	1x SSD partition per daemon (optional)	1x SSD partition per daemon (optional)	
	Network	2x 1GB Ethernet NICs	1x 1GbE+ NICs (10GbE+ recommended)	
ceph-mon	Processor	1x 64-bit AMD-64 1x 32-bit ARM dual-core or better	2 cores minimum	**Mons run in containers on 3 dedicated nodes and on 3 shared with other containers (simple websites) More memory – Not an issue. 24 memory slots are empty and can be upgraded. Disk Space – 1.3TB in RAID10 in each server; Network – dual 10G (2x25G soon)**
	RAM	1 GB per daemon	24GB+ per daemon	
	Disk Space	10 GB per daemon	60 GB per daemon	
	Network	2x 1GB Ethernet NICs	1x 1GbE+ NICs	
ceph-mds	Processor	1x 64-bit AMD-64 quad-core 1x 32-bit ARM quad-core	2 cores minimum	**MDS run in containers on 3 dedicated nodes.**
	RAM	1 GB minimum per daemon	2GB+ per daemon	
	Disk	1 MB per daemon	1 MB per daemon	
	Network	2x 1GB Ethernet NICs	1x 1GbE+ NICs	

Release dates: Mimic – 06/2018, Nautilus – 03/2019, Octopus – 03/2020, Pacific – N/A

Lesson learned the hard way

Caltech

- We set-up dual MDS (active+backup)
- During the stop of active MDS – firewall was not open on backup MDS, all clients staled connection
  - P.S. There was no Ceph MDS/Mons complains about closed firewall – it showed all OK
- All client's failed to communicate with MDS (Same issue can happen with network down for a long time), crash and fail with:

libceph: mds0 10.5.0.1:6800 socket closed (con state NEGOTIATING)

 # ls -la /storage

ls: cannot access /storage: Permission denied

- Bringing MDS back (or network up) – Clients will not restore mount point
- Ceph recommends 4.19.z/4.14.z/5.x – and there are fixes already in place for reconnect, but:
  - All Tier2 is under 3.10.0-1160.15.2.el7 - way too low of what is needed
  - Easiest fix: reboot machine (**not doable for ~300 machines at once**), harder way:
    - Ensure there are no locks (*lsof*) – all condor jobs (singularity) keep lock on CephFS; Stop Condor
    - Unload ceph and libceph kernel modules; Load them again (*rmmod, modprobe*)
    - Unmount mount point; Mount it back;
    - Start condor.

We are still unsure if we will add kernel mount point on all WN's and or redirect all jobs to read via local redirector. Each option has it's own drawbacks.