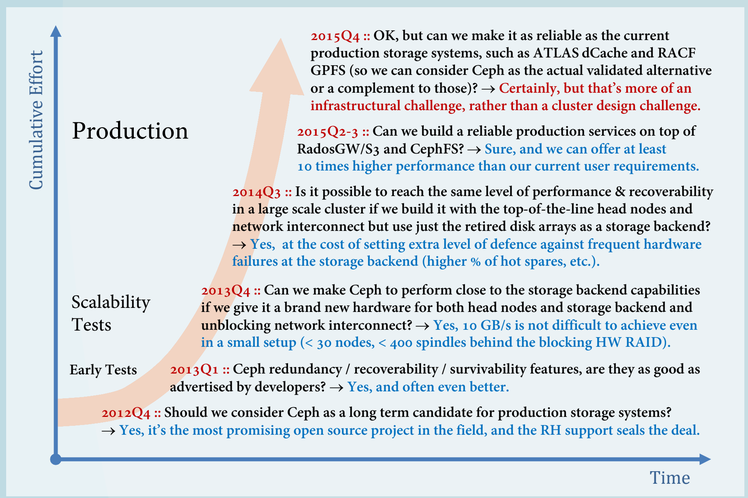
Evolution of the Ceph Based Storage Systems at the RACF

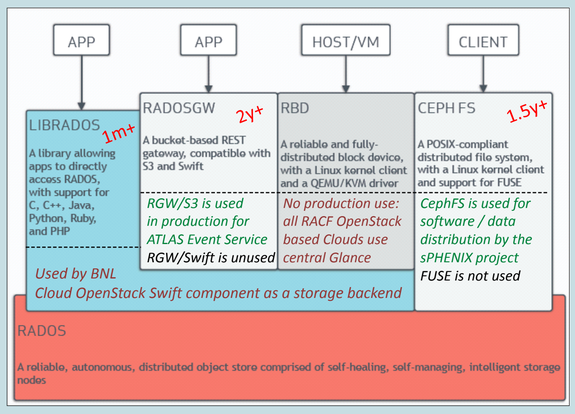
Hironori Ito[1], Tony Wong[1], Tejas Rao[1], Alexandr Zaytsev[1*], Xin Zhao[1]
[1]RHIC & ATLAS Computing Facility (RACF), Brookhaven National Laboratory (BNL), Upton, NY, USA

Computing in High Energy and Nuclear Physics, CHEP 2016: San Francisco, CA, USA (Oct 8-14, 2016)

Evolution of the RACF Approach to Ceph

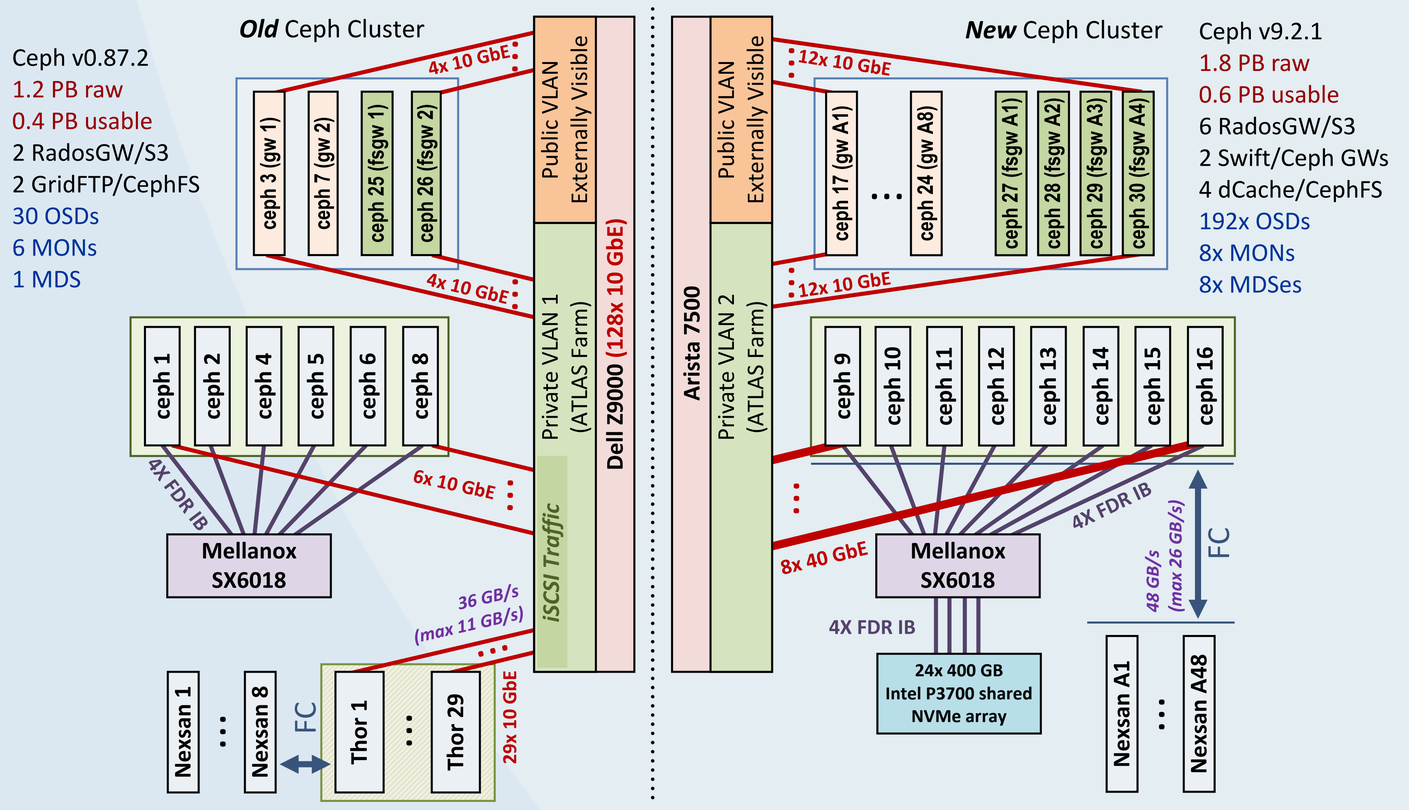


Current Use of Ceph Components in RACF



The BNL Cloud instance is the first user of our Ceph installations that utilizes the low level object store API of Ceph directly, and thus benefiting from the lowest API-driven overhead possible (such as swift-ceph-backend and dCache deployed directly on top of Ceph pools).

Current Network & Storage Layout of the RACF Ceph Clusters



RACF Ceph Cluster Building Blocks

Head nodes and gateways

First gen. head nodes, first and second gen. gateways

Dell PowerEdge R420 (1U) x22

2x 1 TB HDDs in RAID-1 + 1 hot spare
50 GB RAM + 1x 250 GB SSD (up to 10 OSDs)
1x 40 GbE + 1x IPoIB/4X FDR IB (56 Gbps) – *Head nodes*
2x 10 GbE – *Gateways*

Second gen. head nodes

Dell PowerEdge R720XD (2U) x8

8x 4 TB HDDs in RAID-10 + 2 hot spares
128 GB RAM + 2x 250 GB SSDs (up to 24 OSDs)
800 GB PCIe SSDs // 1x 40 GbE + 1x IPoIB/4X FDR IB (56 Gbps) + 12x 4 Gbps FC ports

Storage backend (retired ATLAS dCache HW RAID disk arrays)

iSCSI export nodes

SUN Thor servers (Thors)

48x 1 TB HDDs under ZFS
8 GB RAM
1x 10 GbE
4x 4 Gbps FC (no longer used)
x29

FC attached storage arrays

Nexsan SATABeast arrays (Nexsans)

40x 1 TB HDDs in HW RAID-6 + 2 hot spares
2x 4 Gbps FC
x56

Summary & Future Plans

After nearly two years of building proof of concept installations in 2012–2014, two permanent Ceph cluster installations with total 3 PB raw (1 PB usable) capacity were established in RACF in 2014-2015.

Originally, these installations were only supporting CephFS and RadosGW/S3 clients, but other gateway systems such as GridFTP/CephFS, OpenStack Swift/Ceph and (experimental) dCache/Ceph gateways were added shortly after.

Since mid-2015 our main focus stayed on performance optimization of our Ceph clusters and providing the uninterrupted service to our biggest external (ATLAS ES, PHENIX production on the OSG) and internal (BNL Cloud) clients.

We plan to double the capacity (up to ~ 2 PB usable) early in 2017 and further increase the I/O performance by using the cache tiering mechanism and low latency NVMe PCIe SSD devices.

Performance tuning, Ceph upgrades, adding New Access Capabilities (2015Q2–2016Q3)

2015Q2: Up to 8.7 GB/s of aggregated throughput with CephFS (client network uplink limited)
2015Q4: Up to 1.7 GB/s of throughput via OpenStack Swift gateways (client network uplink limited)
2016Q2: upgrades of Ceph on both clusters to v.87.2 ("old" cluster) and v9.2.1 (new cluster); up to 1.1 GB/s of I/O capability demonstrated with RadosGW/S3 gateways subsystem with ANL to BNL object store tests
2016Q3: adding Intel P3700 series NVMe PCIe SSDs to the "new" Ceph cluster (6.3 TB of distributed plus 9.4 TB of centralized capacity in total) — to be used for the OSD journals and cache tiering on top of high latency OSDs for better handling of high spikes of I/O (with up to 24k concurrent sessions).

First large scale permanent deployment (2014Q3–2015Q2, Ceph v0.80.1–v0.94.3)

CHEP'16

Based on the refurbished storage hardware retired from the BNL ATLAS dCache. Main purpose is to serve as a temporary object storage in front of BNL ATLAS dCache — to be used by the ATLAS Event Service):
Two Ceph clusters with 1.2 PB plus 1.8 PB of raw capacity
1 PB of usable space in total assuming replication factor 3
13 racks of equipment, mixed set of 3.7k HDDs in total

Hitachi HUS 130 Storage System & HP Moonshot test system (2014Q2, Ceph v0.72.2)

3x storage head nodes (each holding a redundant 1+1 RAID controller serving 240x 4 TB HDDs each, organized into 8x RAID-6 sets) plus 6x server nodes (32x HT CPU cores, 64 GB RAM; 1x MON, 6x OSDs on each, OSDs deployed over XFS):
Bonded 2x 10 GbE plus IPoIB/4X FDR IB network uplinks on every node (IPOIB is used for internal cluster network)
2x SSDs on every node for the OSD journals (up to 1 GB/s of write I/O capability)
2.2 PB of raw capacity / 720x HDDs in total
2.4 GB/s (sustained) / 3.7 GB/s (peak) internal Ceph cluster throughput was reached in the replica re-balancing & RBD tests (CPU saturation on the RAID controller up to 95%)

CHEP'15

Ceph in the PHENIX Infiniband Testbed (2013Q4, Ceph v0.72)

26x brand new compute nodes from the RACF PHENIX farm (32x HT CPU cores, 64 GB RAM, 12x 2 TB 7.2krpm SATA HDDs in HW RAID5 on each):
Provided with IPoIB/4X FDR IB interconnect solution based on Mellanox hardware (3.5x oversubscribed tree topology)
Symmetrical configuration of the cluster: 1x MON+OSD on every node (no MDSes since CephFS is not used)
0.25 PB of raw capacity over 312 HDDs (OSDs deployed on top of XFS, 9.4 TB per OSD)
18 GB/s of internal cluster throughput is demonstrated with object replica re-balancing operations
11 GB/s of throughput is demonstrated with read tests over the rbd-fuse mountpoints

CHEP'13

Ceph/RBD testbed using refurbished dCache hardware (2013Q1, v0.61)

Small real hardware testbed built out of 5x Sun Thor (4540) storage servers retired from BNL ATLAS dCache:
"Symmetrical" cluster configuration: MON + MDS + OSD on every node; the same nodes used as clients; OSDs deployed on top of Ext4; OSD journals are on the same devices
235 HDDs / 20 CPU cores in total
Thors are re-installed from Solaris to Fedora 18
Single 10 GbE Myricom NICs on every node

First tests in the virtualized environment (2012Q4, v0.48)

A single hypervisor based virtual testbed (64 GB RAM):
4x OSD SL 6.x VMs (data disks in RAM, OSDs deployed with XFS underneath)
3x MON SL 6.x VMs
12x Fedora 17 client VMs (mapping RBD volumes & mounting CephFS)

CHEP'12

2017 Q1 Q2 Q3 Q4
2016 Q1 Q2 Q3 Q4
2015 Q1 Q2 Q3 Q4
2014 Q1 Q2 Q3 Q4
2013 Q1 Q2 Q3 Q4
2012 Q1 Q2 Q3 Q4

BROOKHAVEN NATIONAL LABORATORY

Designed by Alexandr S. Zaytsev (RACF, BNL), Oct 2016, E-mail: alezayt@bnl.gov