Status Report on Ceph Based Storage Systems at the RACF

Alexandr Zaytsev (alezayt@bnl.gov)
Hironori Ito, *presented by Ofer Rind*

BNL, USA
RHIC & ATLAS Computing Facility

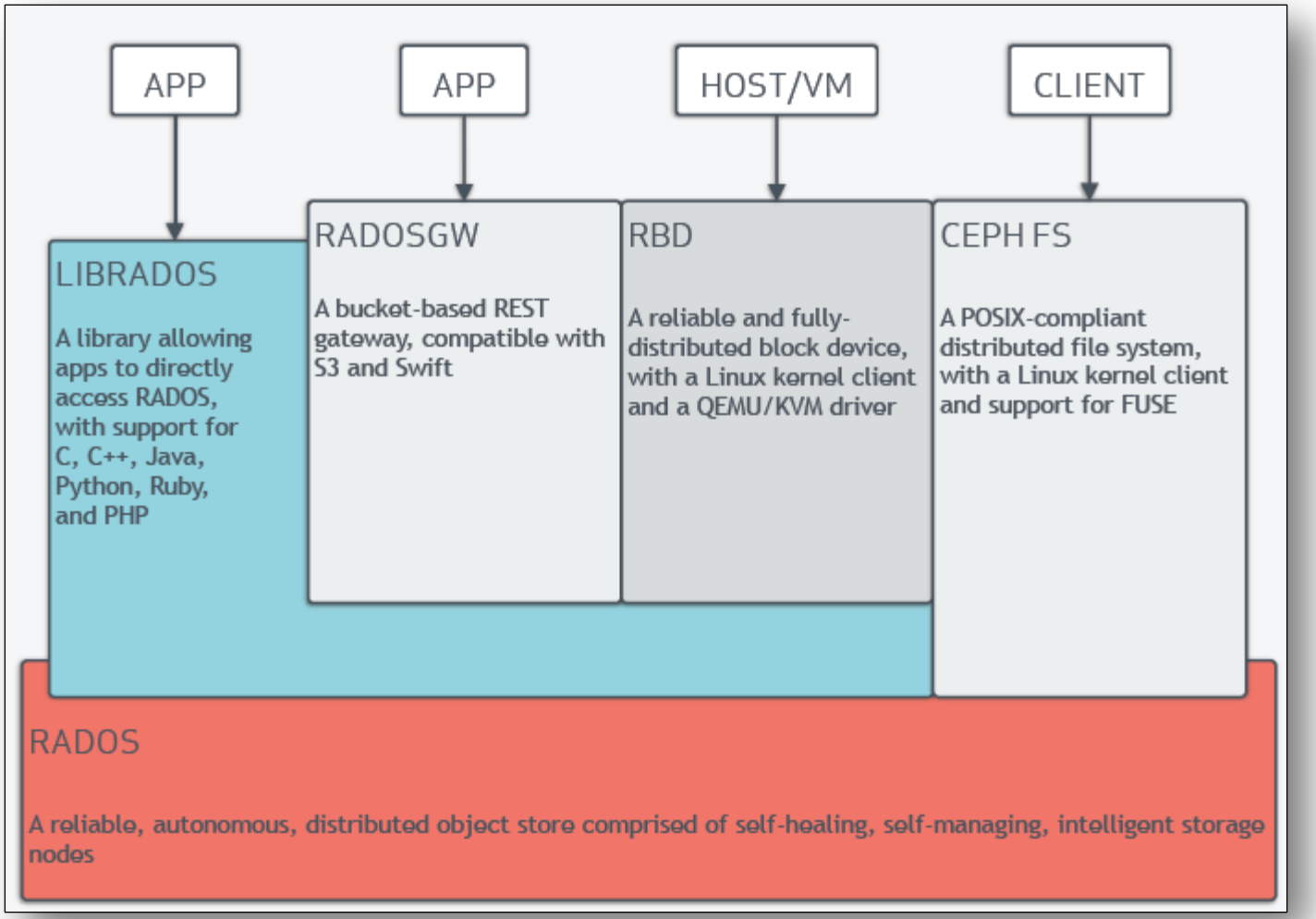
Outline

- Ceph Project Overview
- Early Steps for Ceph in RACF (up to Oct 2014)
  - Early tests (2012Q4-2013Q3)
  - Pre-production test systems (2013Q4-2014Q2)
  - First production systems (2014Q3)
- Recent Developments (since last HEPiX report)
  - Major Ceph cluster(s) upgrade (Feb-Mar 2015)
  - Recent I/O performance measurements
  - Current production load and future plans
- Summary & Conclusions
- Q & A

Ceph Project: General Overview & Status

- Ceph provides all three layers of clustered storage one could reasonably want in an HTC and/or HPC environment:
  - Object storage layer
    - Native API storage cluster APIs (librados)
    - Amazon S3 / Openstack Swift compatible APIs (Rados Gateway)
    - Object storage layer does not require special kernel modules, just needs TCP/IP connectivity between the components of the Ceph cluster
  - Block storage layer (Rados Block Device, RBD)
    - Mainly designed to be used by the hypervisors (QEMU/KVM , XEN, LXC, etc.)
    - It can also be used as a raw device via 'rbd' kernel module first introduced in the kernel **2.6.34** (RHEL 6.x is still on 2.6.32; RHEL 7 derived OS distributions need to get though to the majority of HEP/NP resources in order to get Ceph RBD supported everywhere)
    - Current recommended kernel is 3.16.3 or later (works fine with 3.19.1)
    - Libvirt interfaces with RBD layer directly, so the kernel version restrictions do not apply, though kernel 2.6.25 or newer in order to utilize the paravirtualized I/O drivers for KVM
  - (Nearly) POSIX-compliant distributed file system layer (CephFS)
    - Just recently becoming ready for the large scale production
    - Requires kernel modules ('ceph' / 'libceph') also first introduced in kernel **2.6.34**
    - Current recommended kernel is 3.16.3 or later (works fine with 3.19.1)

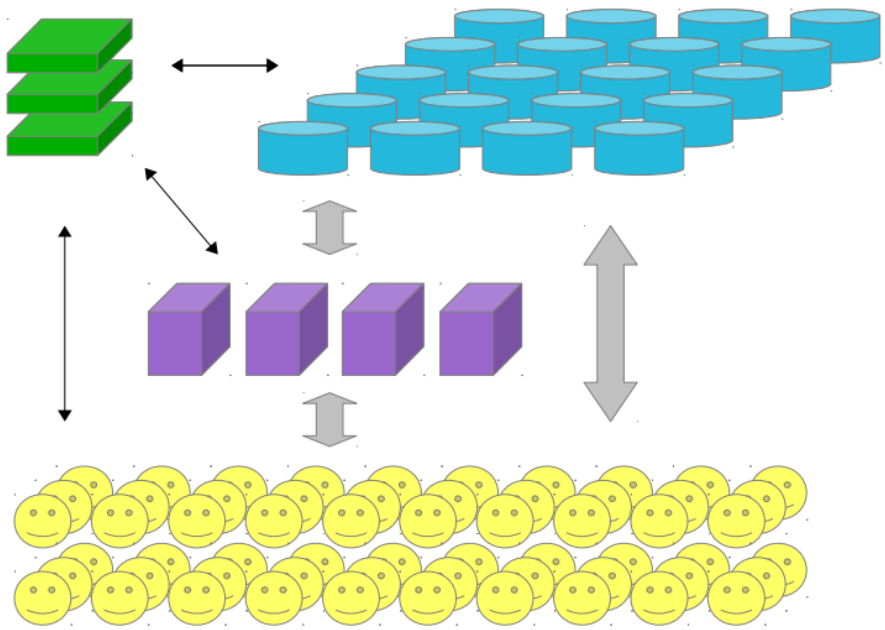
Ceph Architecture: Protocol Stack



Ceph Architecture: Components / Scalability

http://storageconference.us/2012/Presentations/M05.Weil.pdf

- monitors (ceph-mon)
  - 1s-10s, paxos — Achieving consensus in a network of unreliable processors
  - lightweight process
  - authentication, cluster membership, critical cluster state
- object storage daemons (ceph-osd)
  - 1s-10,000s
  - smart, coordinate with peers
- clients (librados, librbd)
  - zillions
  - authenticate with monitors, talk directly to ceph-osds
- metadata servers (ceph-mds)
  - 1s-10s
  - build POSIX file system on top of objects



More detailed architectural overview was given in our previous report at HEPiX Fall 2014: https://indico.cern.ch/event/320819/session/6/contribution/39/material/slides/1.pdf

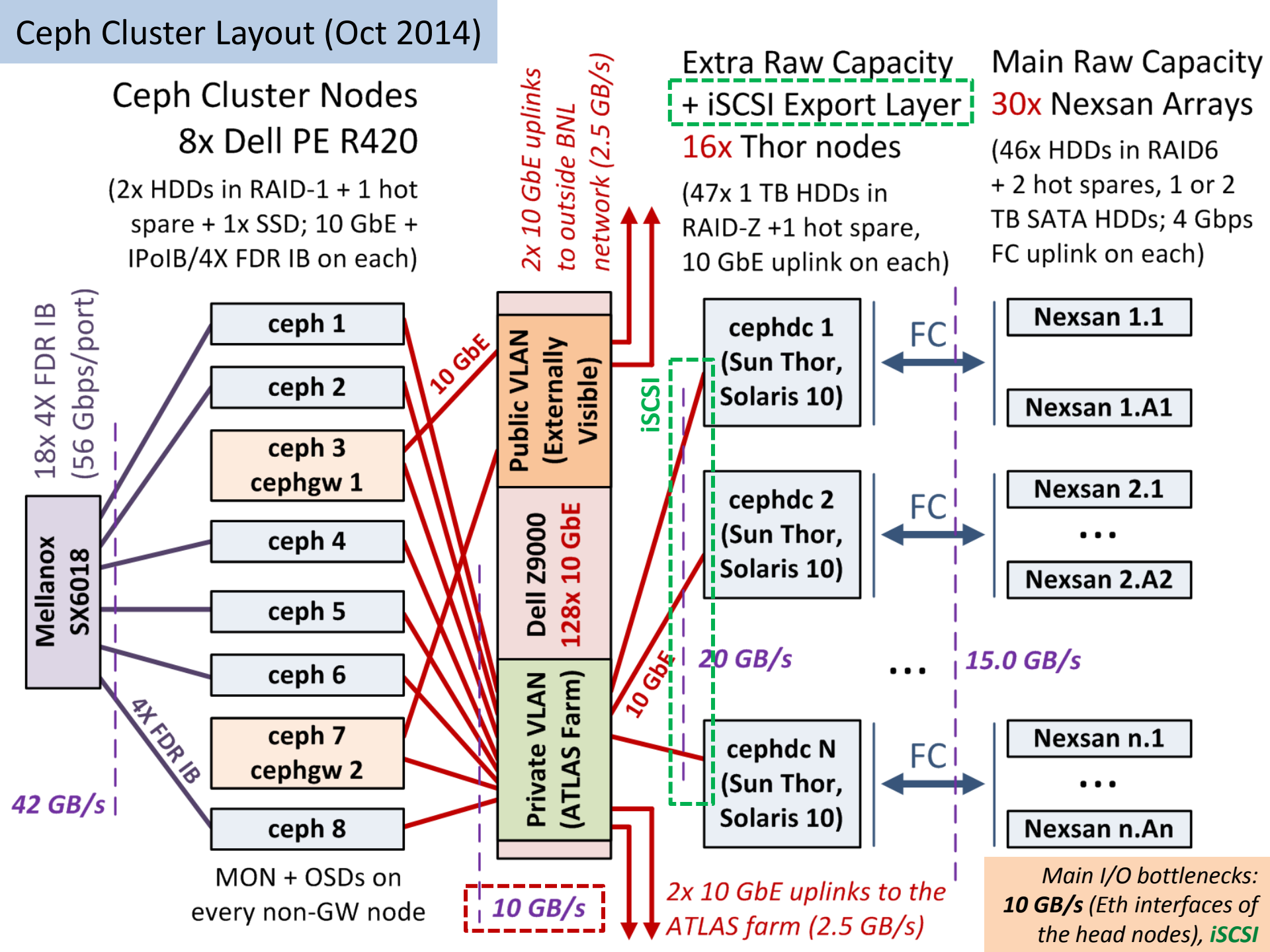
Availability and Production Readiness of Some New Game Changing Features of Ceph

- *Giant* release (v0.87.x)
  - CephFS becomes stable enough for production (6+ months survivability of the mount points observed in RACF with kernels 3.14.x and ≥3.16.x)
    - Should be in a usable state for any RHEL 7.x derived OS environment now
  - Erasure (forward error correction) codes supported by CRUSH algorithm since Ceph v0.80 Firefly release become production ready as well
    - Albeit this feature normally requires more CPU power behind the Ceph cluster components
- Future *Hammer* release (targeted v0.94.x or higher)
  - Long-awaited Ceph over RDMA features (RDMA “xio” messenger support) are to appear in this release
    - Implementation based on Accelio high-performance asynchronous reliable messaging and RPC library
    - Enables the full potential of the low latency interconnects such as Infiniband for Ceph components (particularly for the OSD cluster network interconnects) via eliminating additional IPoIB / EoIB layers and dropping the latency of the cluster interconnect down to the microsecond range

Ceph @ RACF: From Early Tests to the First Production System (2012Q4-2014Q3)

- First Tests in the virtualized environment (2012Q4, Ceph v0.48)
  - Single virtualized server (4x OSD, 3x MON: **64 GB** raw capacity**, 2 GB/s** peak)
- Ceph/RBD testbed using refurbished dCache hardware (2013Q1, Ceph v0.61)
  - 5 merged head & storage nodes (5x OSD, 5x MON, 5x MDS: **0.2 PB, 0.5 GB/s**)
- Ceph in the PHENIX Infiniband Testbed (2013Q4, Ceph v0.72)
  - 26 merged head & storage nodes (26x OSD, 26x MON: **0.25 PB, 11 GB/s**)
- Pre-production Test Systems with Hitachi HUS 130 Storage System (2014Q2, Ceph v0.72.2)
  - 6 head nodes + 3 RAID array groups (36x OSD, 6x MON: **2.2 PB, 2.7 GB/s**)
- CephFS evaluation with HP Moonshot test system (2014Q2, Ceph v0.72.2)
  - 30 merged head nodes (16x OSD, 8x MON: **16 TB,** *only stability tests*)
- First Production System (2014Q3, Ceph v0.80.1)
  - 8 head and gateway nodes + 45 RAID arrays (45x OSD, 6x MON, 2x Rados GW: **1.8 PB, 4.2 GB/s**, 6+ GB RAM/OSD)

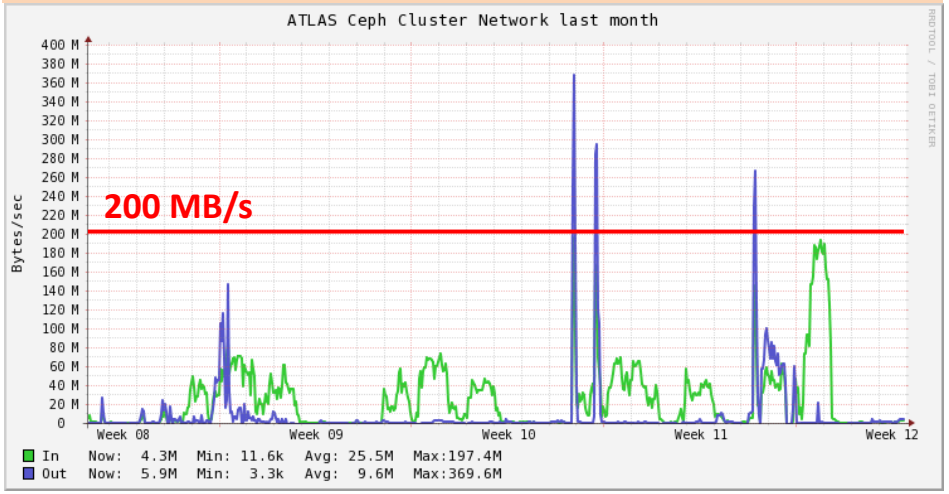
Ceph Cluster Layout (Oct 2014)



Ceph @ RACF: Current Production Load

- ATLAS Event Service (S3):
  - Logs (since 2014Q4)
  - Events (since 2015Q1)
  - 7M objects are currently stored in the PGMAP
- Testing Interfaces to Other Storage Systems:
  - ATLAS dCache
  - XRootD
  - Amazon S3 interoperation

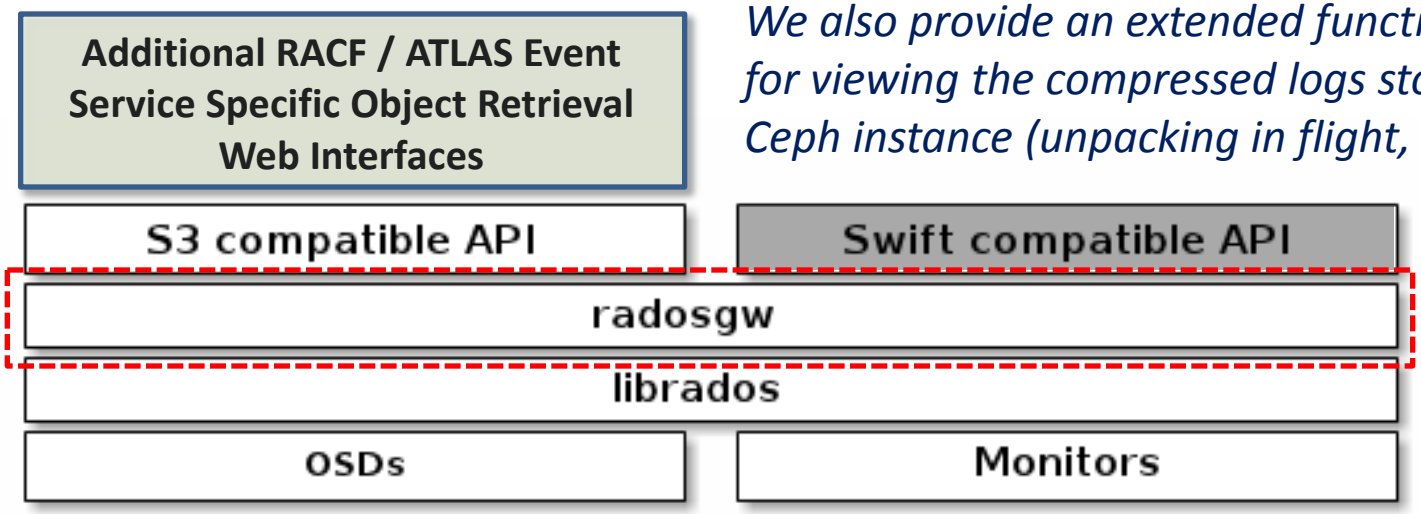
Ceph cluster network activity over the last month



- BNL, MWT2 and AGLT2 are also discussing the possibility of deploying the Federated Ceph storage
  - A distributed group of Ceph clusters each provided with a group of Rados gateways and the shared Region / Zone aware data pool configuration
  - The delays with deployment of the new Ceph cluster hardware that we suffered in Dec 2014 – Jan 2015 caused shifting this activity forward in time; now we expect to be ready for this in Apr 2015

Ceph @ RACF: Amazon S3 / Rados GW Interfaces

- We provide a redundant pair of Ceph Rados Gateways over the httpd/FastCGI wrapper around the RGW CLI that facing both into BNL intranet and to the outside world
  - Allows one to use Amazon S3 compliant tools to access the content of the RACF Ceph object storage system (externally mostly used by CERN for initial ATLAS Event Service tests)
  - We do not support an Amazon style bucket name resolution via subdomains such as http://<bucket_id>.cephgw01/<object_id>, but have an easier custom implemented schema instead that supports URLs like http://cephgw02/<bucket_id>/<object_id>



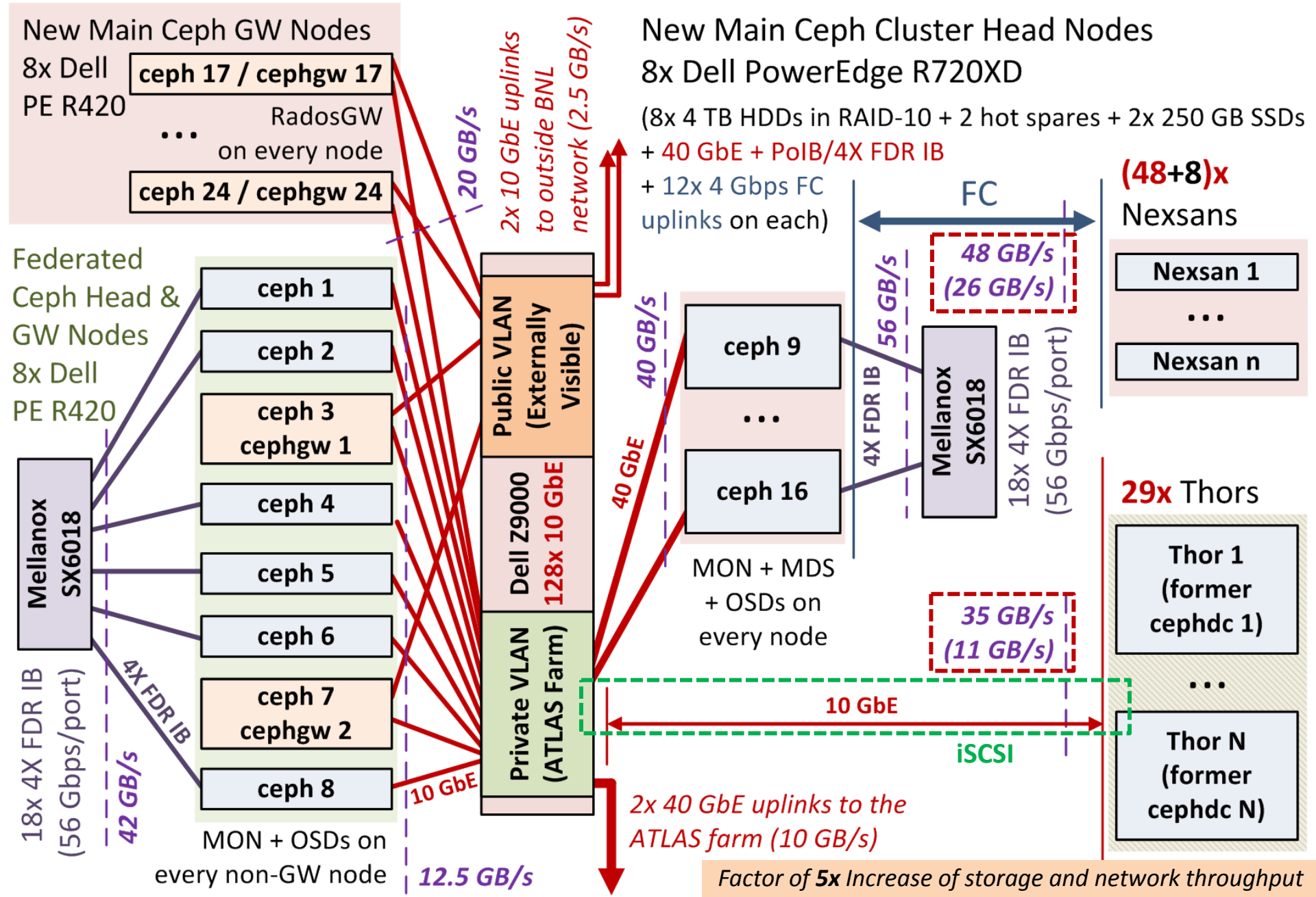
We also provide an extended functionality for viewing the compressed logs stored in our Ceph instance (unpacking in flight, etc.)

Ceph @ RACF: Recent Developments (2014Q4-2015Q1)

- The storage backend of the Ceph cluster is extended up to 85 disk arrays consisting of 3.7k HDDs (most of which are still 1 TB drives) with total usable capacity of 1 PB (taking into account data replication factor of 3x)
- Major upgrade of the Ceph cluster head nodes and Rados Gateways was performed in Feb-Mar 2015 increasing the internal network bandwidth limitation of the system from 10 GB/s up to 50 GB/s
  - **Adding 16 new head and gateway nodes**
  - **102x OSDs, 8x MONs, 8x MDS, 8x Rados GW, 6+ GB RAM per OSD**
- We are now running two physically separate Ceph clusters with the usable capacity split as 0.6 PB (new main production cluster for the ATLAS event service) + 0.4 PB (Federated Ceph cluster)
  - The newly installed Ceph cluster components are based on Linux kernel 3.19.1 and Ceph v0.87.1 (*Giant*)
  - This new configuration is expected to be finalized in Mar-Apr 2015
  - Hoping to switch the new main Ceph cluster to the Hammer release before going into production with the new system (in Apr 2015)
  - Once the new main cluster goes into production the former production Ceph cluster is going to be re-deployed as a new Federated Ceph cluster (still provided with the separate S3 interfaces)
- Scalability and performance tests of CephFS and mounted RBD volumes were performed on the scale of 74x 1 GbE attached physical client hosts (latest batch of the BNL ATLAS Farm compute nodes) for the first time

Expected Post-upgrade Layout (Apr 2015)

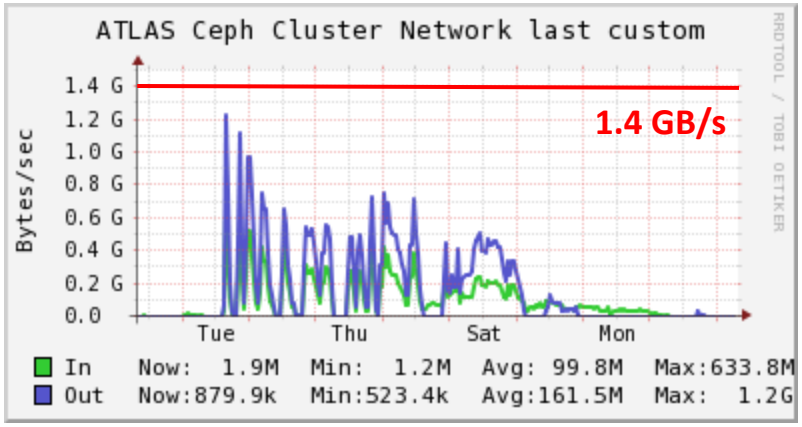
0.6 PB + 0.4 PB Usable Capacity Split



*Factor of **5x** Increase of storage and network throughput compared previous configuration of the Ceph cluster*

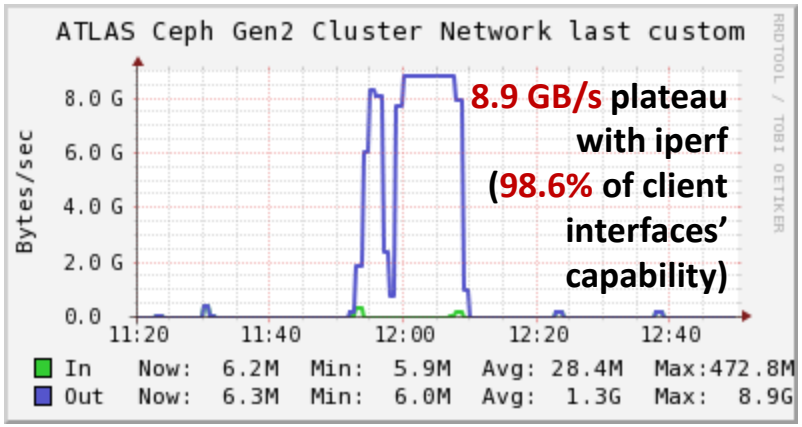
Ceph @ RACF: Some Notes / Observations

- The “old” cluster:
  - Excellent recoverability observed over the last 6 months with Ceph v0.80.1
  - Transparent phasing out of 30x Nexsans disk arrays from under the 3x replicated storage pool with re-balancing (going from 45x OSDs down to 15 remaining OSD in 4 days) that was performed as a part of the cluster upgrade worked as expected

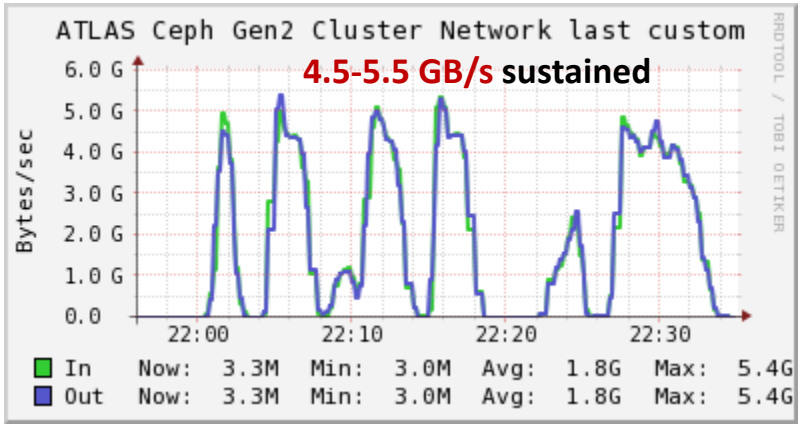


- The new cluster:
  - Extracting maximum performance out of the old Nexsan RAID arrays and the new Dell PE R720XD based head nodes (≈550 MB/s per RAID array)
    - Splitting the capacity of each RAID array 4 ways: 2 volumes per each P2P FC uplink
    - IRQ affinity optimizations (1x Mellanox + 1x Emulex + 4x Qlogic cards in each head node)
    - Custom build automation layer for affinity and I/O scheduler optimizations, spinning disk arrays and OSD journal SSDs discovery/mapping, and the OSD/OSD journal (re-)deployment
    - Choosing between IPoIB / EoIB layers over 4X FDR IB for internal Ceph cluster interconnect deployment (IPoIB over the Infiniband interface in connected mode is performing better mainly because of much larger 64k MTU)
    - 8x PCI-E v3.0 (8 GT/s) data throughput limitation of maximum throughput of about 63 Gbps = 7.9 GB/s is actually limiting the network performance in our current installation
  - Dealing with the RAM requirements of the OSDs
    - Splitting the OSDs further, e.g. with factor of 2x (up to 200+ OSDs in total) is not possible right now because of OSD RAM requirements in the recovery mode (it was experimentally verified that 3.5+ GB RAM per OSD is insufficient for stable operations)

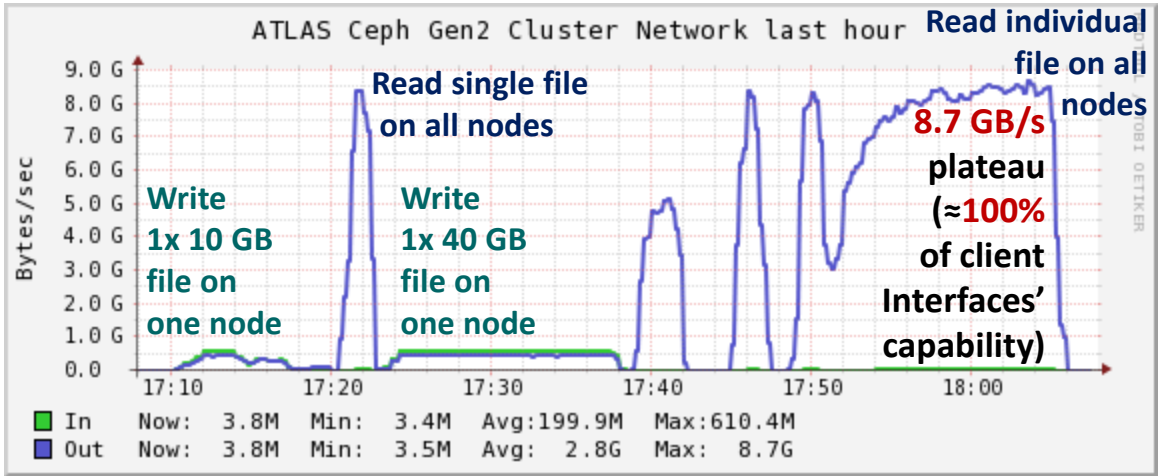
Recent Performance Tests (74x 1 GbE Attached Client Nodes): CephFS



[STEP1] Measure the network bandwidth available between the clients and the Ceph cluster nodes with iperf: (a) one stream per client node, (b) two streams per client node. (80 Gbps limitation is also in place on the level of internal uplinks to the ATLAS farm, but it is not restrictive in this case.)

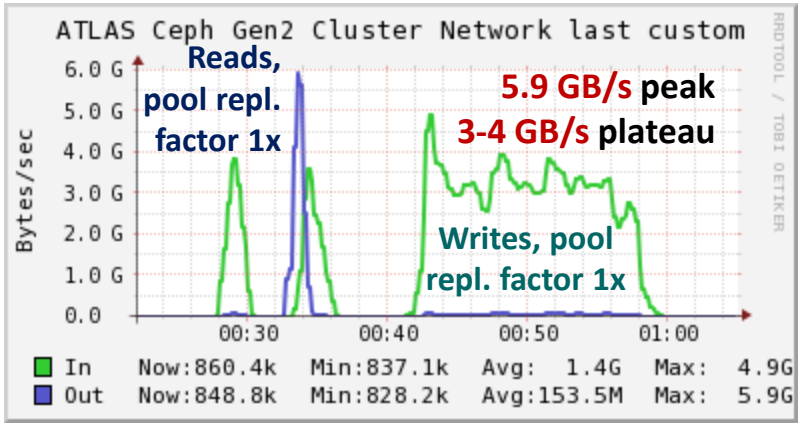


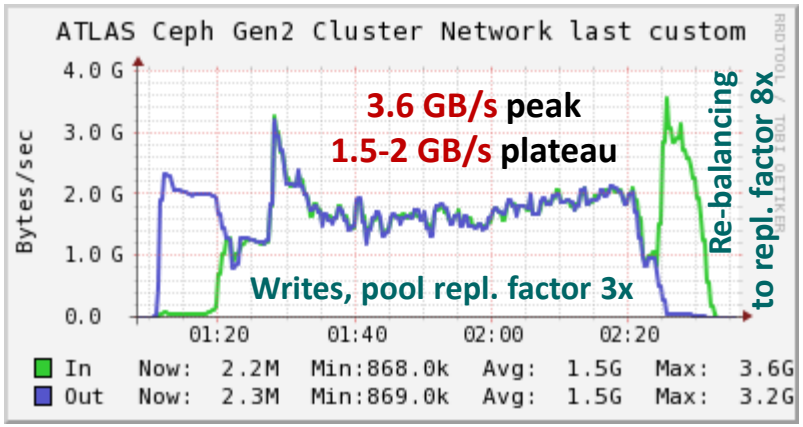
[STEP2] (a) Write data from one client to one file in CephFS (1 GB, 10 GB and 40 GB files were used; not all the cases are shown) and (b) write data to individual file to CephFS on all the clients (while both data and metadata pool replication factors are kept at 3x)



[STEP3] Read tests: retrieving (a) one file from CephFS on one node, (b) one file from all the nodes, (c) individual file on each node in parallel. (Dropping pool replication factor to 1x doesn't change the situation much.)

Recent Performance Tests (74x 1 GbE Attached Client Nodes): RBD





10 GB RBD image is created for every client (with the replication factor 3x on the level of the RBD pool), then mapped individually as a block device to each of the client node, formatted under XFS and mounted locally.

The I/O tests were similar to those of CephFS, though the maximum test file size was limited to 4 GB.

The aggregated I/O rate is approximately factor of 2x smaller compared to CephFS while the CPU load on the OSD daemons is twice higher. Dropping the pool replication factor down to one helps with write performance, but doesn't change much for reads (as expected).

Perhaps, further optimizations are needed in order to bring the mounted RBD performance close to the one of CephFS.

- The CephFS aggregated performance is only limited by the client NIC throughput with some particular load patterns (particularly in a "file shared among many hosts" scenario)
- Larger scale tests (such as those with 8x 10 GbE + 4x 40 GbE clients scheduled for the near future) are needed to probe the actual I/O limits of our current Ceph installation
- CephFS now looks like a better option for exporting raw capacity of the Ceph cluster to the clients not willing / unable to use librados and/or Rados GW interface layers (at least on the systems based on the 3.x kernels)

Summary & Conclusions

- Ceph is a rapidly developing open source clustered storage solution that provides object storage, block storage and distributed file system layers
  - Ceph is spreading fast among the HTC community in general and HEP/NP community in particular
  - Major increase in the number of deployments for HEP/NP over the last 12 months
- A long sequence of Ceph functionality and performance tests was carried out in RACF during the period of 2012Q4-2014Q2 that resulted in design and deployment of a large scale RACF Ceph object storage system that entered production in 2014Q3 and operated successfully ever since
  - **Scaling from 1x 1U server to 13 racks full of equipment in less than 2 years**
  - 4X FDR Infiniband technology was successfully used in all our production Ceph installations (used as internal cluster interconnect in combination with IPoIB). Still waiting for the production support for RDMA in Ceph for exploring its full potential.
  - We are continuing to evaluate Ceph/RBD and CephFS for possible applications within the RACF facility, e.g. as a dCache storage backend and a distributed files system solution alternative to IBM GPFS (with both simple replication and erasure codes)
  - The newly deployed part of the RACF Ceph cluster is provided with **26 GB/s** backend storage bandwidth and completely non-blocking network interconnect capable of reaching the level of **40 GB/s** of aggregated network traffic coming through it
  - The ability to generate up to **8.7 GB/s** of useful client traffic was demonstrated with CephFS scalability tests using 74x compute nodes of the BNL ATLAS farm
  - The new cluster layout is to be finalized in Apr 2015
- The next status update is to be given at CHEP2015 conference in Okinawa

Q & A

http://apod.nasa.gov/apod/ap150321.html