Ceph Based Storage Systems at the RACF

Alexandr Zaytsev
alezayt@bnl.gov

BNL, USA
RHIC & ATLAS Computing Facility

Outline

- Ceph Evolution
  - Recent developments
- Ceph Installations in RACF
  - Current Configuration
  - Main users and associated challenges
  - Recent production experience
  - Performance optimizations with ATLAS ES
- Future Plans (2016Q4-2017Q2)
- Summary & Conclusion
- Q & A

Reference to Previous Presentations

- It's a 4th dedicated report on our Ceph related developments to the HEPiX community, so please refer to the following materials in case you are interested in the pre-history of the subject
- HEPiX presentations:
  - HEPiX 2014 Fall: https://indico.cern.ch/event/320819/contributions/742951/
  - HEPiX 2015 Spring: https://indico.cern.ch/event/346931/contributions/817793/
  - HEPiX 2015 Fall: https://indico.cern.ch/event/384358/contributions/909232/
- CHEP presentations
  - CHEP 2015 (talk): https://indico.cern.ch/event/304944/contributions/1672364/
  - CHEP 2016 (poster): https://indico.cern.ch/event/505613/contributions/2230970/

Recent Developments with Ceph

- *Jewel* production releases since Mar 2016 (v10.a.b)
  - CephFS declared stable by the developers
  - CephFS related recovery/repaire tools are declared feature-complete
  - RedHat includes CephFS for the first time in their Ceph Storage 2 platform (Jewel release based) as a technology preview (since Aug 2016)
  - AWS v4 Client Signatures support added
  - Fixes for supporting multipath devices and NVMe partitions
  - Significant improvements in MDS
  - Significant fixes for XIO (Ceph over RDMA features)
    - Based on the 3rd party Accelio high-performance asynchronous reliable messaging and RPC library
    - Adds RDMA/Infiniband transport to Ceph, extending Ceph's Messenger and integrates the new Messenger with Mon, OSD, MDS, librados (RadosClient), rados, and libcephfs (client)
  - *We are currently on v9.a.b for production deployment with our "new" Ceph cluster (bound by SL6 OS on the head nodes)*

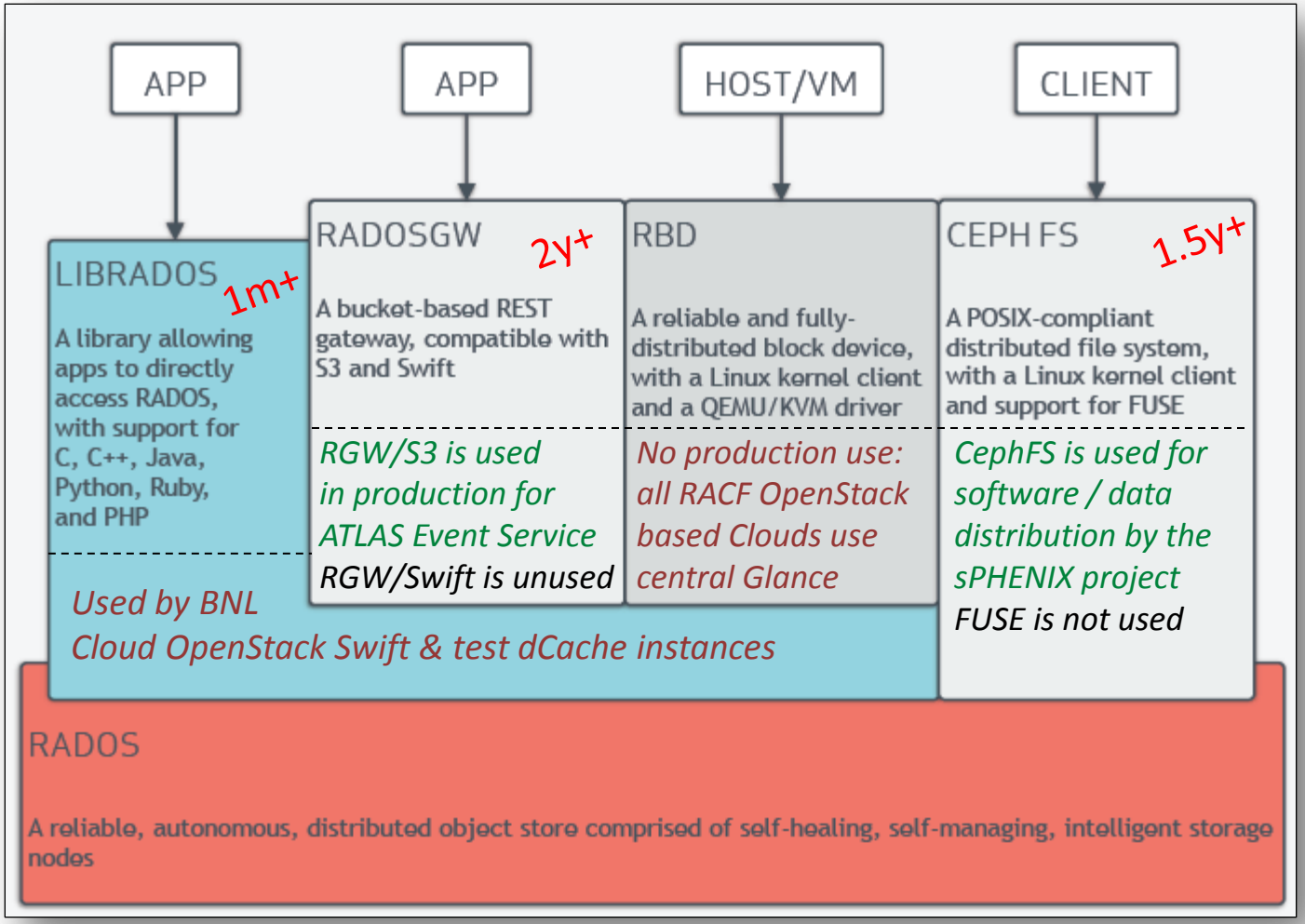
Recent Developments with Ceph

- *Kraken* release (v11.a.b)
  - Sync modules introduced to to RadosGW
  - Metadata (elastic) search functionality based on sync modules introduced to RadosGW
  - Add static website support for Swift API
  - Main development focus is now on the new OSD backend for Ceph called bluestore (key-value store), that is expected to become the new default for Ceph eventually
  - Asynchronous features for RBD image creation/replication added (yet they are of limited interest to us as we don't have RBD used in production)
  - *The expectation is to jump on v11.a.b branch with the next underlying storage upgrade and revisit the Ceph RDMA features stability analysis in the process*

RACF: Main Use Cases for Ceph in 2014-2016

- Production
  - Object Store for ATLAS Event Service (ES): RadosGW/S3 over civetweb, other options are under investigation (Swift REST API), **100M objects stored, up to 1 GB/s WAN transfers demonstrated, up to 24k simultaneous client connections supported** – *not yet a pledged resource for ATLAS; still in the "best possible effort" support mode*
  - GridFTP/CephFS software repository for PHENIX/sPHENIX collaborations needed for supporting large scale production on the opportunistic OSG resources, **0.5M objects stored**
  - Swift-Ceph for the BNL Cloud Installation (direct access via Rados API with some of the OpenStack Swift (Kilo release) components running on the side of the Ceph cluster gateways), **200 TB usable space allocated, up to 1.7 GB/s demonstrated (network limited)**
- Testing/Evaluation
  - CephFS directly mounted on the clients on the scale of more than 50 nodes, **up to 8.7 GB/s demonstrated (network limited)** – *awaiting for the RACF farms to upgrade on RHEL7 derived distributives in order to reach the scale of hundreds of nodes; using FUSE is not considered an option*
  - dCache-Ceph for BNL ATLAS dCache (direct access via Rados API, highly experimental at the moment; many features are yet to be added)

RACF: Current Use of Ceph Components

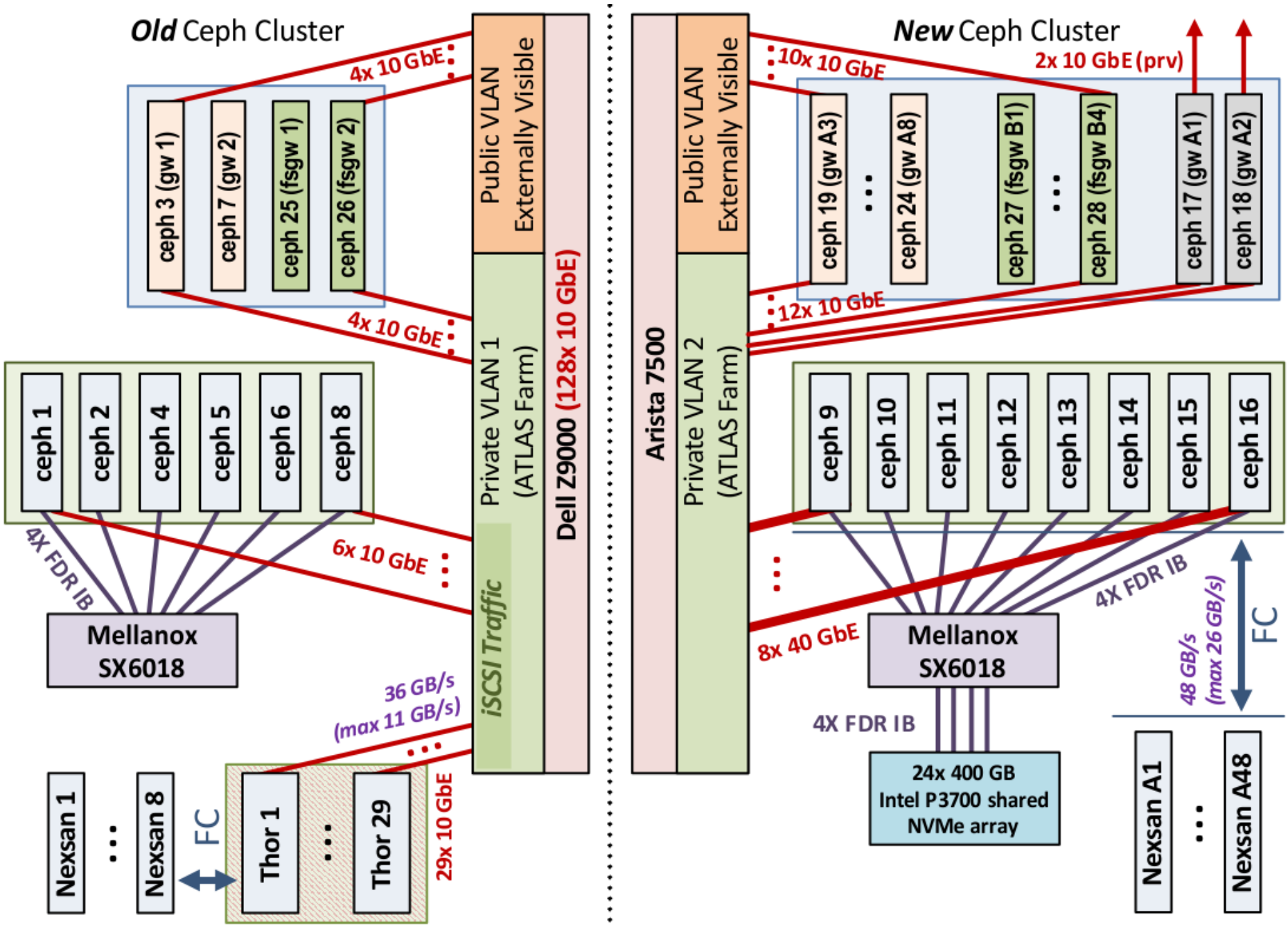


RACF Ceph Clusters: Changes Since 2015Q4

- Infrastructure
  - Both clusters are now on redundant power (normal power + Flywheel UPS)
  - All the single PSU nodes in the new cluster are now getting power from dual attached APC ATS devices installed locally in the racks (not the case for the old cluster yet)
  - No more “recovery stress tests” by normal (“LIPA”) power cuts for us
- Networking
  - No major changes other than some transparent interventions on the external uplinks of both clusters
  - One scheduled Dell Z9000 switch reboot in order to fix issue with the management interface on the switch (short scheduled downtime for the old cluster only)
  - Preparing the Ceph clusters for the major ATLAS network reorganization as a part of migration to the BNL Scientific Network Fabric (former HPC-Core) network (Ceph clusters are expected to remain on external NTP synchro-nization as provided by the ATLAS Farm network infrastructure)

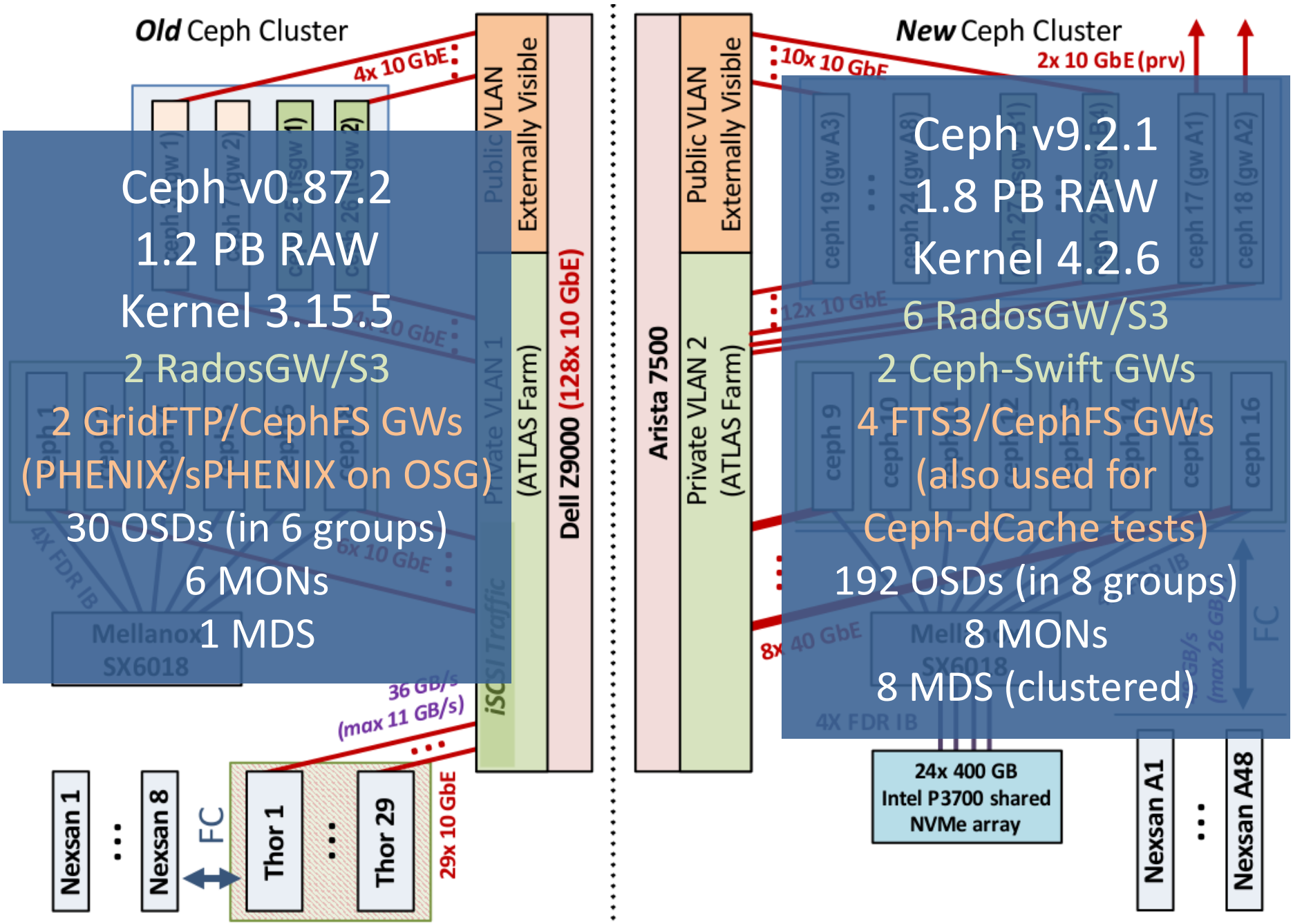
Two Ceph clusters deployed in RACF as of 2016Q4

0.6 PB + 0.4 PB usable capacity split



Two Ceph clusters deployed in RACF as of 2016Q4

0.6 PB + 0.4 PB usable capacity split



Ceph v0.87.2
1.2 PB RAW
Kernel 3.15.5
2 RadosGW/S3
2 GridFTP/CephFS GWs
(PHENIX/sPHENIX on OSG)
30 OSDs (in 6 groups)
6 MONs
1 MDS

Ceph v9.2.1
1.8 PB RAW
Kernel 4.2.6
6 RadosGW/S3
2 Ceph-Swift GWs
4 FTS3/CephFS GWs
(also used for
Ceph-dCache tests)
192 OSDs (in 8 groups)
8 MONs
8 MDS (clustered)

RACF Ceph Clusters: Changes Since 2015Q4 – cont.

- Headnodes & gateways:
  - Intel P3700 800 GB NVMe PCIe device is added on each head node of the new cluster (one PCIe slot had to be vacated on each head node for NVMe deviced by replacing 2 low density 4 Gbps FC cards with one high density 8 Gbps FC card)
  - Four new FTS3/CephFS gateways added to the new Ceph cluster in Oct'16
  - OS/Ceph components reconfiguration in order to better deal with new scale of ATLAS Event Service operations (June 2016)
  - Ceph version upgrades on both clusters (up to v0.87.2 and v9.2.1)
- Storage backend
  - No hardware changes other than the regular fixes of broken RAID controllers and HDDs in the Nexsan arrays
  - All RAID controllers in the Nexsan arrays in the new cluster are reconfigured for better handling the random I/O in order to match the ATLAS Event Service access pattern (June 2016)
  - Shared NVMe array (Supermicro 2U server with 24x 400 GB Intel P3700 devices and 4x FDR IB ports) is added to the IB fabric of the new Ceph cluster

RACF Ceph Clusters: Current Building Blocks

1st gen. head nodes, 1st and 2nd gen. gateways

Dell PowerEdge R420 (1U) servers

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*

2nd gen. head nodes

Dell PowerEdge R720xd (2U) servers

x8

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

1st gen. storage backend components (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

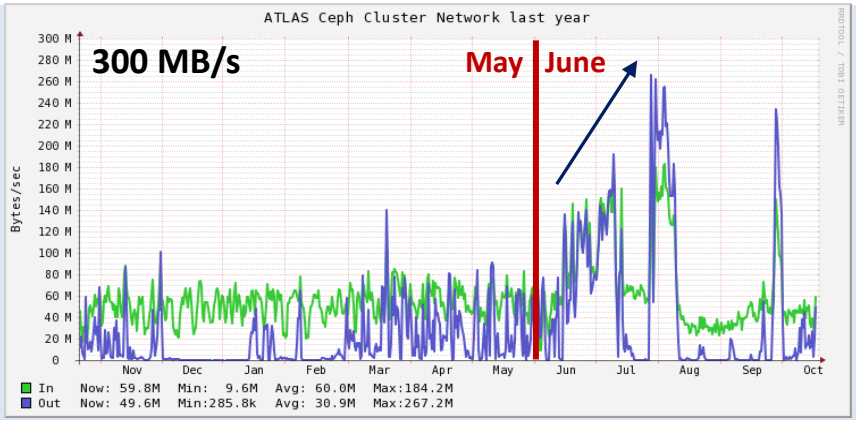
x56

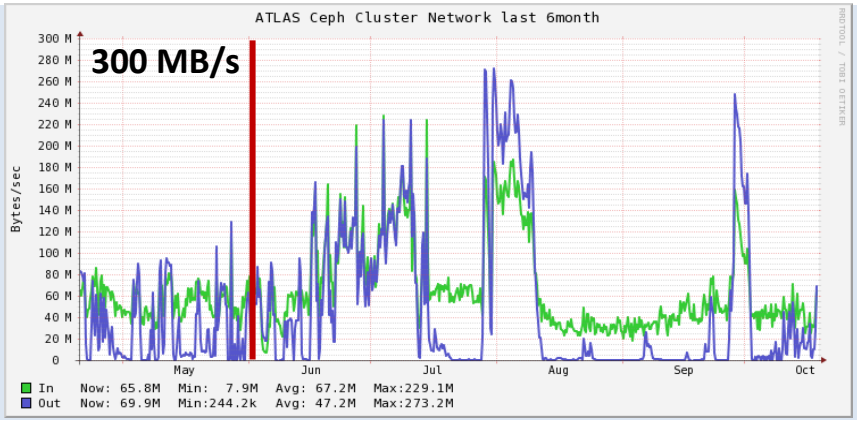
Ceph Workload Evolution Since 2015Q4

Last 12 months

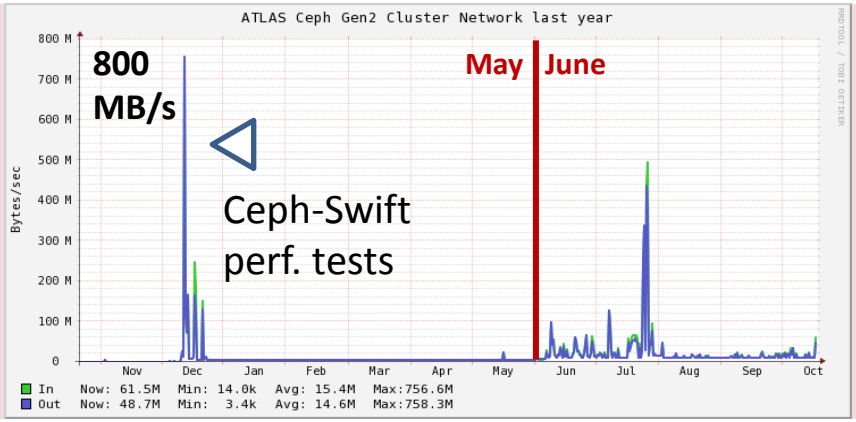
Last 6 months

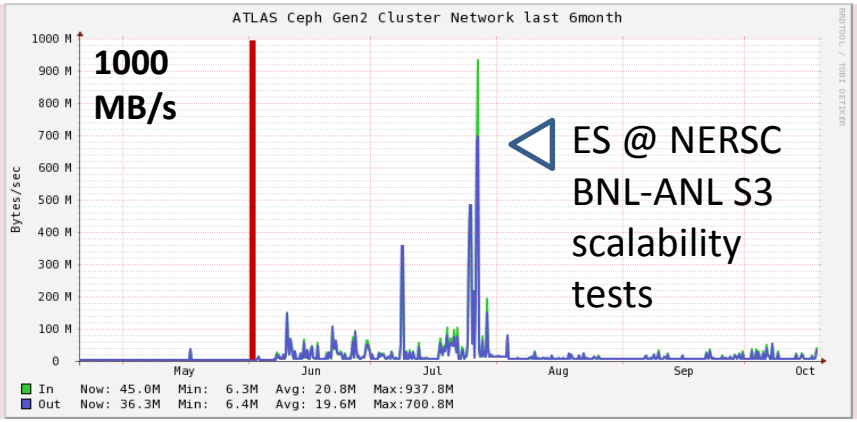
Old cluster





New cluster





May/June 2016 transition: ATLAS ES begins to ramp up the number of ES jobs in PROD queues and performs multiple test runs at NERSC (Edison & Cori machines, **RTT = 71 ms** from BNL) with up to 700 nodes x 24 = 16.8k concurrent jobs (Edison); 100k events produced and saved to BNL Object Store in each run over the WAN directly from compute nodes.

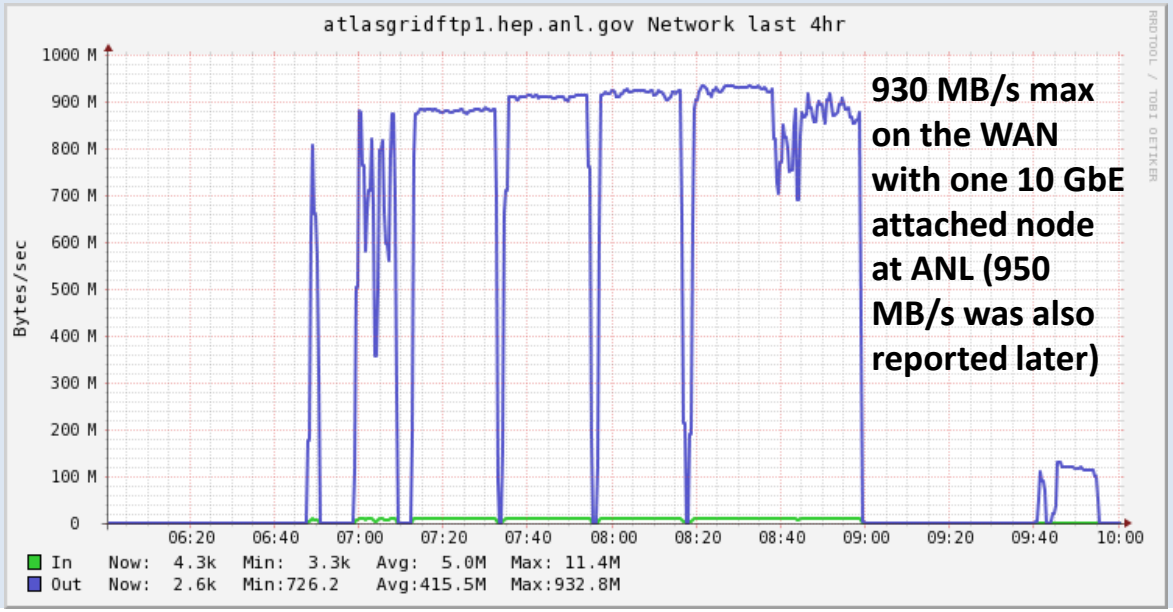
Configuration Changes Needed on Both Clusters

- NERSC workloads are steered away from the old cluster ('cephgw' alias) and placed exclusively on the new cluster ('cephgw-test' alias) on the ATLAS side
- RAID controller cache settings change from streaming I/O optimized to random I/O optimized on all Nexsan arrays of the new Ceph cluster in order to cope better with highly concurrent workloads
- Ceph version upgrade up to 0.87.2 on the old cluster and its RadosGW/S3 subsystem migration from Apache to civetweb
- Ceph version upgrade up to 9.2.1 upgrade on the new cluster and its RadosGW/S3 subsystem to civetweb
- **All the RadosGW/S3 nodes:** performance tuning (OS & Ceph RGW level in order to scale up 0.25k to 4k concurrent connections per server with the hardware available)
- **All the head nodes:** OS & Ceph level performance tuning in order to cope with increased number of internal TCP sessions within the cluster caused by allowing up to 24k external client connections (OSD settings)
- More performance monitoring metrics added to Ganglia for the new cluster to monitor the number of TCP connections in various states (both external and internal)

BNL-ANL Ceph Workload Evolution Since 2015Q4

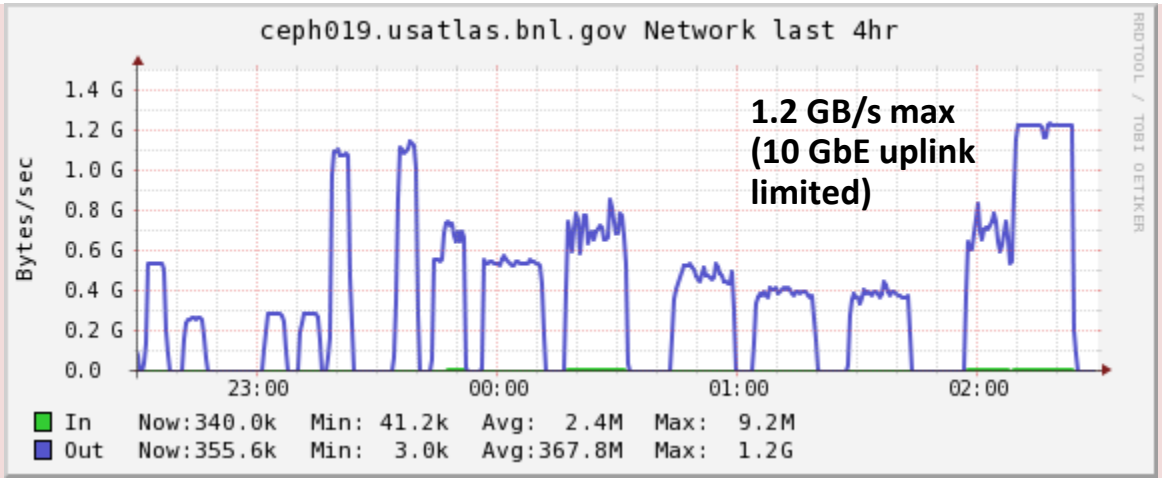
Reaching up to 1 GB/s with RadosGW/S3 with up to 24k concurrent client sessions

WAN tests (ANL-BNL)



Blocks of 10-40 typical ATLAS events of 8.4-34 MB in size were pushed from ANL in 320 concurrent threads to 6 gateways nodes of the RACF new Ceph cluster using the Boto library based test suite developed by Doug Benjamin **(RTT = 28 ms)**

Local tests (BNL)

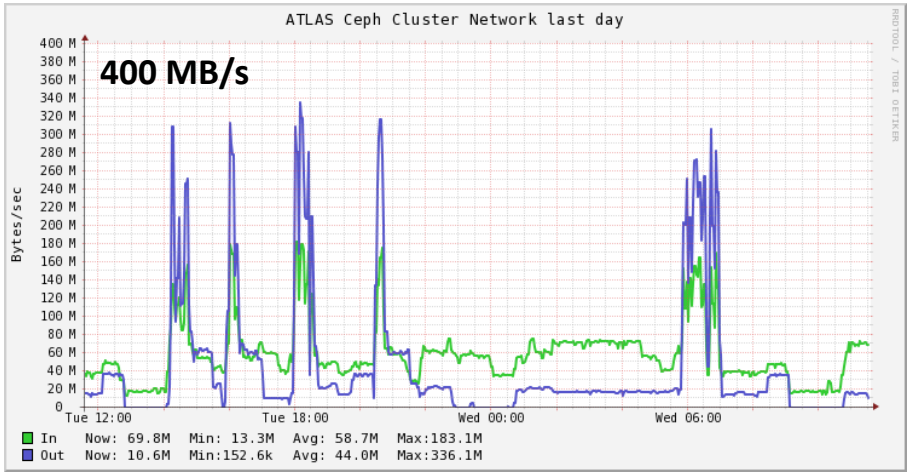


Local tests at BNL performed with the same test suite are capable of saturating 10 GbE links on RadosGW/S3 nodes of the new Ceph cluster for payloads > 8 MB in size

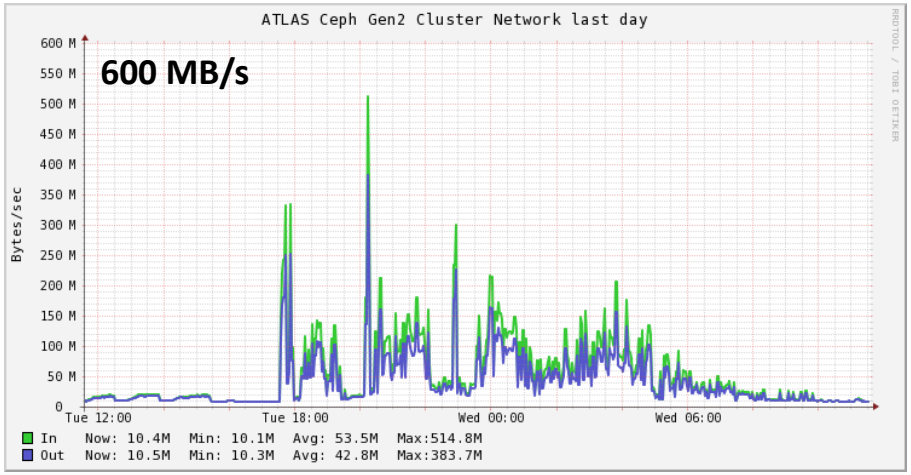
BNL-ANL Ceph Workload Evolution Since 2015Q4

Typical workloads we see now (last 24 hours)

Old cluster (2k connections limit, 2 GWs)



New cluster (24k connections limit, 6 GWs)



Even though we are no longer pushing the limits of what's possible with our current setup, the regular load generated by the ATLAS ES jobs is now an order of magnitude higher than what we saw a year ago

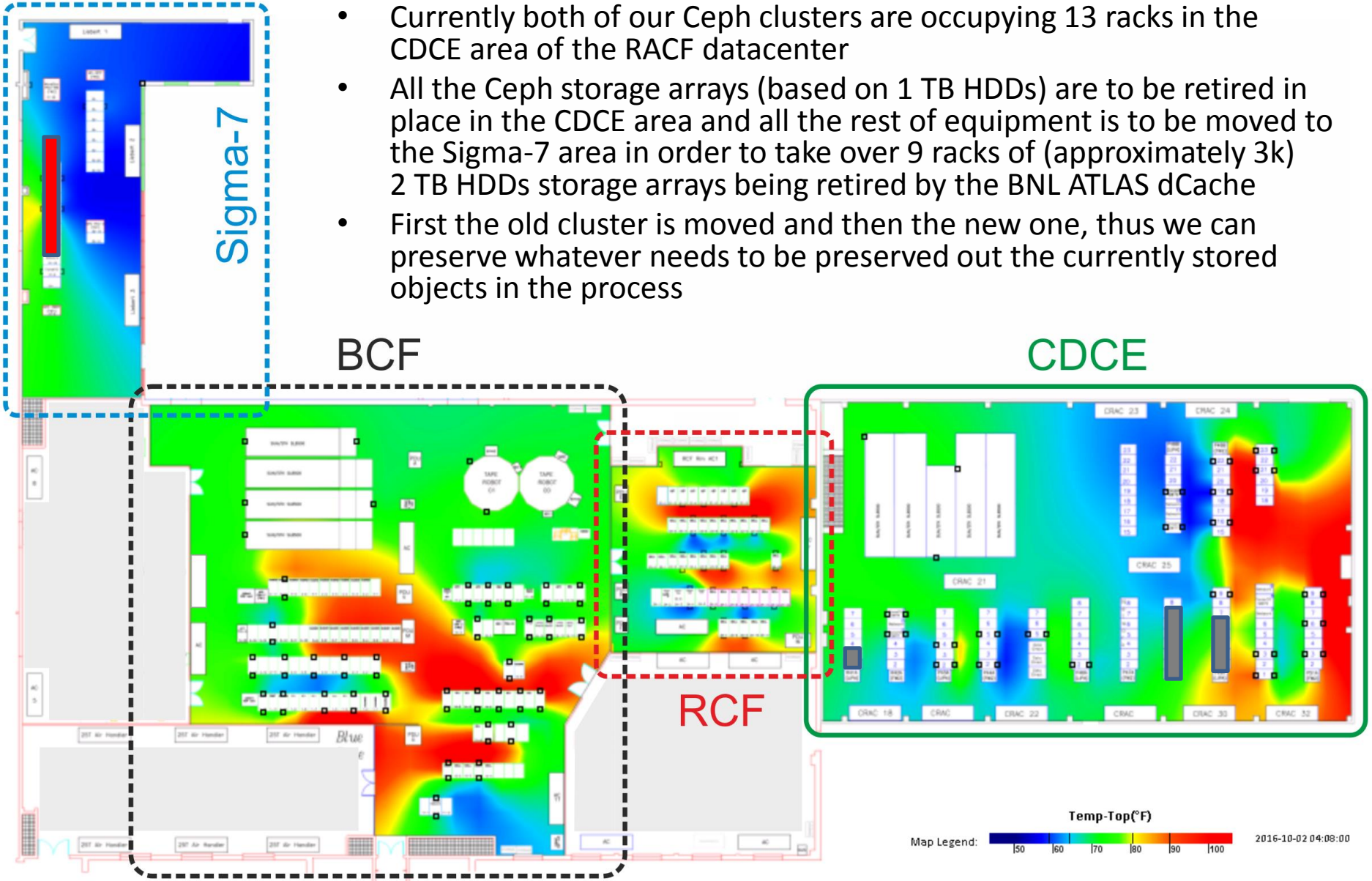
The combined ES workload at NERSC, BNL ATLAS Tier-1, US Tier-2s and opportunistic Cloud/OSG resources is expected to reach the level of 80k concurrent jobs running, which likely means one more order of magnitude increase of the regular workloads for the BNL Ceph clusters

Preparing for Handling 80k Concurrent ATLAS ES jobs

- On the ATLAS Event Service (software) side
  - Consider output events buffering on the machine level for large production runs at NERSC
  - Consider using more sophisticated data transfer mechanisms used in WLCG with queuing capabilities (such as FTS)
  - Consider merging the output events on the level of single compute node in order to achieve better overall transfer efficiency without tuning TCP settings at the source and further understanding of the network path being used
- On the BNL Ceph clusters (hardware) side
  - Consider reducing the internal RadosGW/S3 and OSD latency by introducing both distributed (on each head node) and centralized (array attached to the OSD-to-OSD IB fabric) NVMe PCIe SSD capacity using the 20 μs read/write Intel P3700 devices to be used for handling OSD journals and adding 10 TB cache tier on top of the OSDs deployed on the spinning drives in HW RAIDs
  - Migrating to the newer storage hardware for both Ceph cluster backends

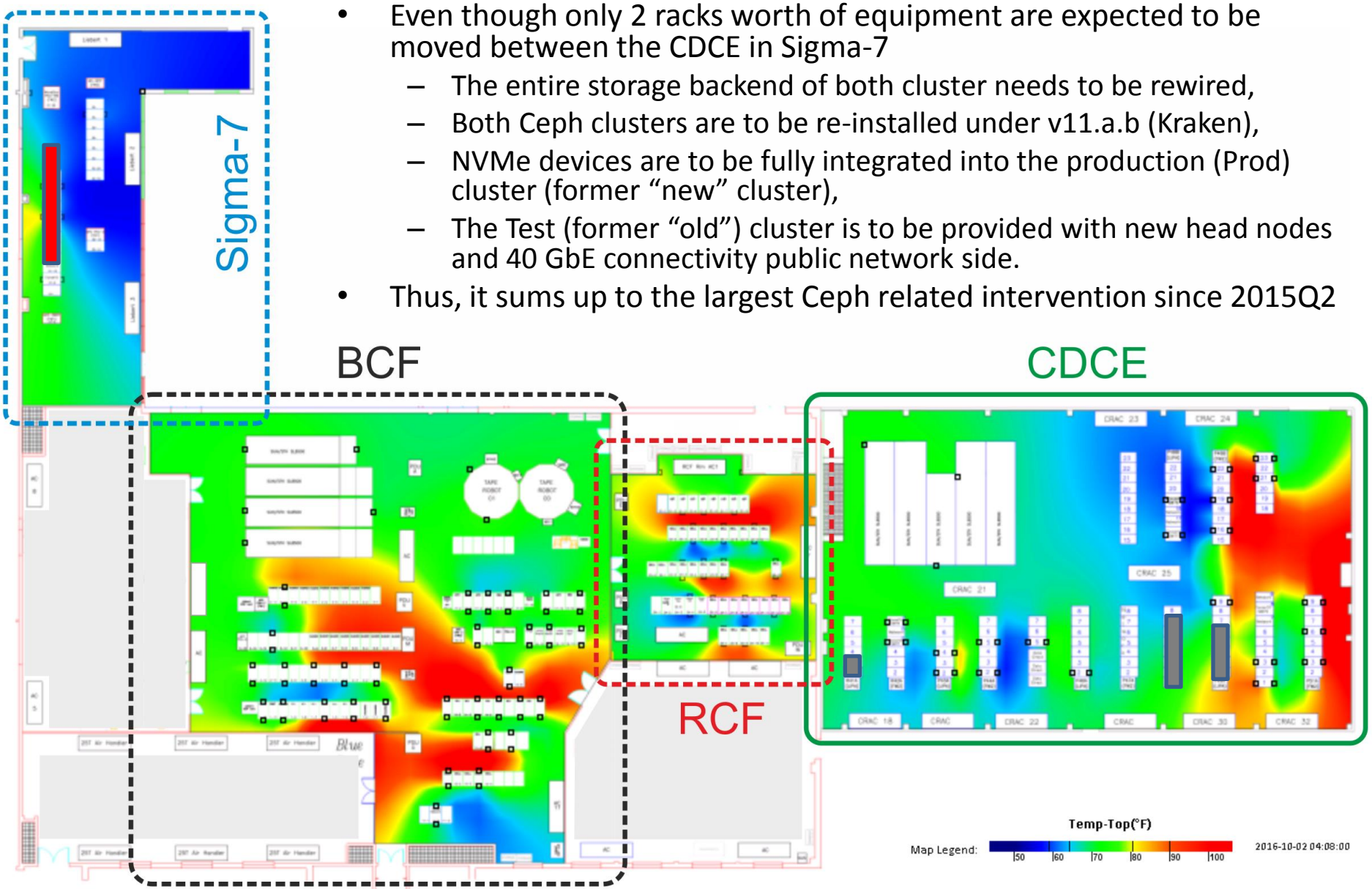
Next Upgrade Plans (2016Q4-2017Q1)

- Currently both of our Ceph clusters are occupying 13 racks in the CDCE area of the RACF datacenter
- All the Ceph storage arrays (based on 1 TB HDDs) are to be retired in place in the CDCE area and all the rest of equipment is to be moved to the Sigma-7 area in order to take over 9 racks of (approximately 3k) 2 TB HDDs storage arrays being retired by the BNL ATLAS dCache
- First the old cluster is moved and then the new one, thus we can preserve whatever needs to be preserved out the currently stored objects in the process



Next Upgrade Plans (2016Q4-2017Q1) – cont.

- Even though only 2 racks worth of equipment are expected to be moved between the CDCE in Sigma-7
  - The entire storage backend of both cluster needs to be rewired,
  - Both Ceph clusters are to be re-installed under v11.a.b (Kraken),
  - NVMe devices are to be fully integrated into the production (Prod) cluster (former “new” cluster),
  - The Test (former “old”) cluster is to be provided with new head nodes and 40 GbE connectivity public network side.
- Thus, it sums up to the largest Ceph related intervention since 2015Q2



RACF Ceph Clusters: Next Upgrade Building Blocks

1st gen. head nodes, 1st and 2nd gen. gateways

Dell PowerEdge R420 (1U)

x22

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

2nd 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 + 800 GB NVMe
1x 40 GbE + 1x IPoIB/4X FDR IB (56 Gbps) +
FC layout is yet to be defined

2nd gen. storage backend (retired ATLAS dCache HW RAID disk arrays)

The next upgrade planned for 2016Q4-2017Q1 implies the complete replacement of the storage backend

FC attached storage arrays with SAS extension chassis

Nexsan E60 series arrays

Approx. 3k x 2 TB HDDs in total for approx. 5.5 PB of raw capacity
8 Gbps FC ext. connectivity

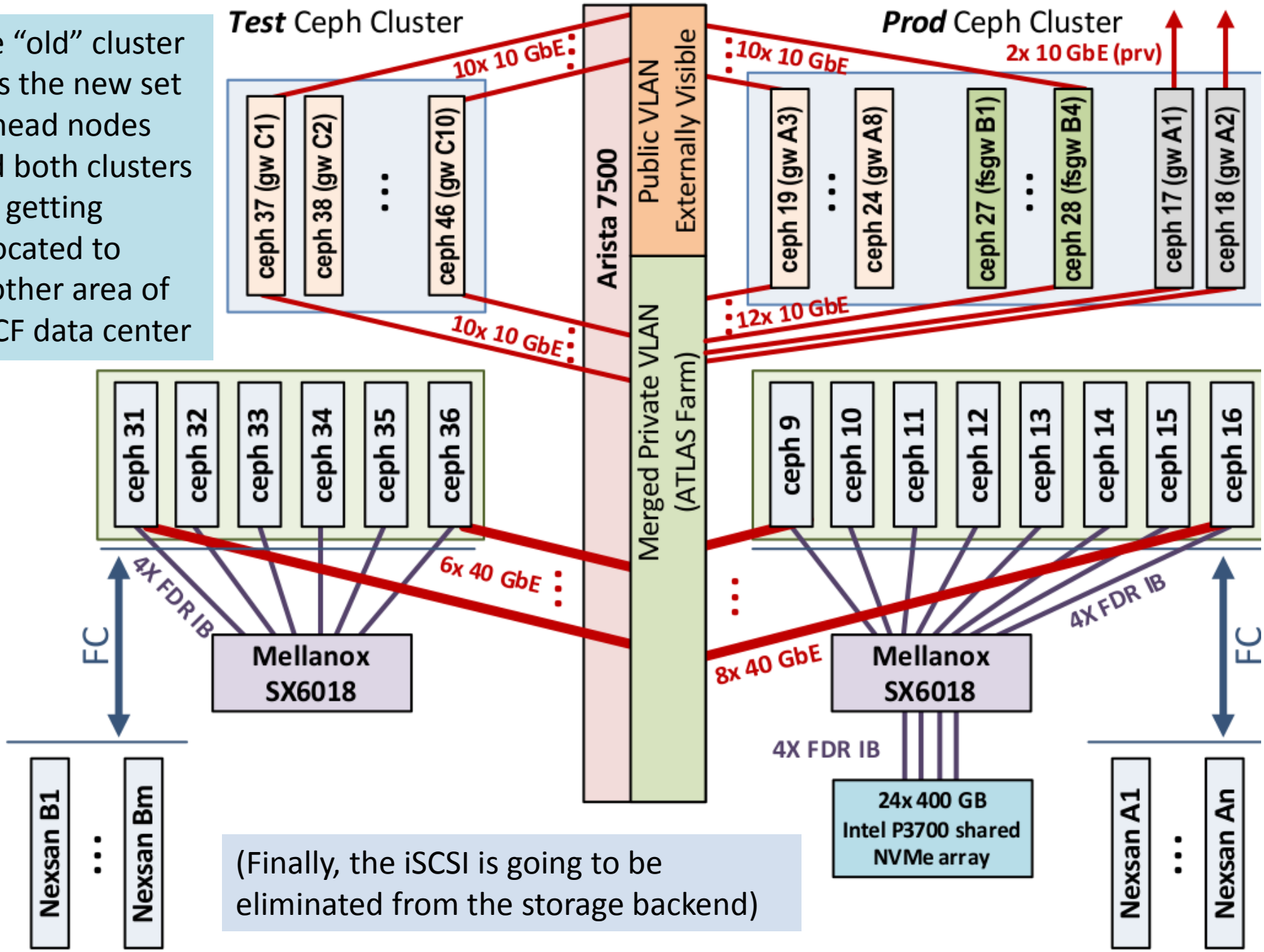
Exact quantities are yet to be defined

The 3rd gen. head node specs are yet to be defined: the retired BNL ATLAS dCache servers are the most likely choice

We intend to preserve dual cluster layout in the future

~ 5.5 PB of total RAW capacity

The "old" cluster gets the new set of head nodes and both clusters are getting relocated to another area of RACF data center



(Finally, the iSCSI is going to be eliminated from the storage backend)

Summary & Conclusion

- 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 and operated ever since.
- Originally, these installations were only supporting CephFS and RadosGW/S3 clients, but other gateway systems such as GridFTP/CephFS, FTS3/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 Event Service, PHENIX/sPHENIX collaboration production on the OSG opportunistic resources) and internal (BNL Cloud) clients. In the process of doing so the following performance characteristics were demonstrated so far:
  - Up to 8.7 GB/s of aggregated throughput with CephFS (client network uplink limited),
  - Up to 1.7 GB/s of throughput via OpenStack Swift gateways (client network uplink limited),
  - Up to 1 GB/s of I/O capability demonstrated with RadosGW/S3 gateways subsystem with ANL to BNL object store tests (up to 24k simultaneous client connections permitted).
- We plan to increase the raw capacity up to approximately 5.5 PB in early 2017 and further increase the I/O performance by using the cache tiering mechanism and low latency (20 μs for both read/write) NVMe PCIe SSD devices from Intel P3700 series.

Q & A