Evolution of the Ceph Based Storage Systems at the RACF (Poster Contribution)

Alexandr Zaytsev
alezayt@bnl.gov

BNL, USA
RHIC & ATLAS Computing Facility

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

Summary / Highlights

- 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 Event Service, PHENIX detector 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 1.1 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 double the capacity (up to approximately 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 (Intel P3700 series).