WHAT’S NEW IN LUMINOUS AND BEYOND

Douglas Fuller – Red Hat
2017.11.14
UPSTREAM RELEASES

- **New release cadence**
  - Named release every 9 months
  - Backports for 2 releases
  - Upgrade up to 2 releases at a time
    - (e.g., Luminous → Nautilus)
LUMINOUS
BLUESTORE: STABLE AND DEFAULT

- New OSD backend
  - consumes raw block device(s) – no more XFS
  - embeds rocksdb for metadata
- Fast on both HDDs (~2x) and SSDs (~1.5x)
  - Similar to FileStore on NVMe, where the device is not the bottleneck
- Smaller journals
  - happily uses fast SSD partition(s) for internal metadata, or NVRAM for journal
- Full data checksums (crc32c, xxhash, etc.)
- Inline compression (zlib, snappy)
  - policy driven by global or per-pool config, and/or client hints
- Stable and default
HDD: RANDOM WRITE
HDD: MIXED READ/WRITE
RGW ON HDD+NVME, EC 4+2

4+2 Erasure Coding RadosGW Write Tests

32MB Objects, 24 HDD/NVMe OSDs on 4 Servers, 4 Clients
RBD OVER ERASURE CODED POOLS

- aka erasure code overwrites
- requires BlueStore to perform reasonably
- significant improvement in efficiency over 3x replication
  - \(2+2 \rightarrow 2x\) \(4+2 \rightarrow 1.5x\)
- small writes \textit{slower} than replication
  - early testing showed 4+2 is about half as fast as 3x replication
- large writes \textit{faster} than replication
  - less IO to device
- implementation still does the “simple” thing
  - all writes update a full stripe
CEPH-MGR

- ceph-mgr
  - new management daemon to supplement ceph-mon (monitor)
  - easier integration point for python management logic
  - integrated metrics
- ceph-mon scaling
  - offload pg stats from mon to mgr
  - validated 10K OSD deployment (“Big Bang III” @ CERN)
- restful: new REST API
- prometheus, influx, zabbix
- dashboard: built-in web dashboard
  - webby equivalent of 'ceph -s'
## CEPH-METRICS

### At a Glance

<table>
<thead>
<tr>
<th>Health</th>
<th>Pool Status</th>
<th>OSD Status</th>
<th>OSD Hosts</th>
<th>Capacity Utilization</th>
<th>7d Growth Rate</th>
<th>Weeks Remaining</th>
<th>Placement Group Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>WARN</td>
<td>Normal</td>
<td>Normal</td>
<td>12/15</td>
<td>53%</td>
<td>-2.0 TB</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Scrub</td>
<td>ACTIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client IOPS</td>
<td>15.9 MB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client Throughput</td>
<td>3.7 MB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBDs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Client Read/Write Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSD Hosts CPU Busy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk IOPS</td>
<td>28%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk Throughput</td>
<td>2503</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nearly Full Disks</td>
<td>189.1 MB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAM Util.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk Latency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk Utilization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ASYNCMESSENGER

- new network Messenger implementation
  - event driven
  - fixed-size thread pool
- RDMA backend (ibverbs)
  - built by default
  - limited testing, but seems stable!
- DPDK backend
  - prototype!
PERFECTLY BALANCED OSDS (FINALLY!)

- CRUSH choose_args
  - alternate weight sets for individual rules
  - fixes two problems
    - **imbalance** – run numeric optimization to adjust weights to balance PG distribution for a pool (or cluster)
    - **multipick anomaly** – adjust weights per position to correct for low-weighted devices (e.g., mostly empty rack)
      - backward compatible ‘compat weight-set’ for imbalance only
- pg up-map
  - explicitly map individual PGs to specific devices in OSDMap
  - requires luminous+ clients
- Automagic, courtesy of ceph-mgr ‘balancer’ module
  - ceph balancer mode crush-compat
  - ceph balancer on
RADOS MISC

- CRUSH device classes
  - mark OSDs with class (hdd, ssd, etc)
  - out-of-box rules to map to specific class of devices within the same hierarchy
- streamlined disk replacement
- require_min_compat_client – simpler, safer configuration
- annotated/documentated config options
- client backoff on stuck PGs or objects
- better EIO handling
- peering and recovery speedups
- fast OSD failure detection
S3
Swift
Multisite federation
Compression
Erasure coding
Multisite replication
NFS
Tiering
Deduplication
Encryption
RADOSGW
RGW METADATA SEARCH

ZONE A
RADOSGW
LIBRADOS

ZONE B
RADOSGW
LIBRADOS

ZONE C
RADOSGW
LIBRADOS

CLUSTER

REST

REST

REST

elasticsearch

redhat
RGW MISC

- NFS gateway
  - NFSv4 and v3
  - full object access (not general purpose!)
- dynamic bucket index sharding
  - automatic (finally!)
- inline compression
- encryption
  - follows S3 encryption APIs
- S3 and Swift API odds and ends
RBD

- RBD over erasure coded pool
  - `rbd create --data-pool <ecpoolname> ...`
- RBD mirroring improvements
  - cooperative HA daemons
  - improved Cinder integration
- iSCSI
  - LIO tcmu-runner, librbd (full feature set)
- Kernel RBD improvements
  - exclusive locking, object map
CEPHFS

- multiple active MDS daemons (finally!)
- subtree pinning to specific daemon
- directory fragmentation on by default
  - (snapshots still off by default)
- so many tests
- so many bugs fixed
- kernel client improvements
CONTAINERS CONTAINERS

- ceph-container ([https://github.com/ceph/ceph-container](https://github.com/ceph/ceph-container))
  - multi-purpose container image
  - plan to publish upstream releases as containers on download.ceph.com
- ceph-helm ([https://github.com/ceph/ceph-helm](https://github.com/ceph/ceph-helm))
  - Helm charts to deploy Ceph in Kubernetes
- openshift-ansible
  - deployment of Ceph in OpenShift
- mgr-based dashboard/GUI
  - management and metrics (based on ceph-metrics, [https://github.com/ceph/ceph-metrics](https://github.com/ceph/ceph-metrics))
- leverage Kubernetes for daemon scheduling
  - MDS, RGW, mgr, NFS, iSCSI, CIFS, rbd-mirror, ...
- streamlined provisioning and management
  - adding nodes, replacing failed disks, ...
The future is:
- NVMe
- DPDK/SPDK
- futures-based programming model for OSD
- painful but necessary

BlueStore and rocksdb optimization:
- rocksdb level0 compaction
- alternative KV store?

RedStore?
- considering NVMe focused backend beyond BlueStore

Erasure coding plugin API improvements:
- new codes with less IO for single-OSD failures
EASE OF USE ODDS AND ENDS

- centralized config management
  - all config on mons – no more ceph.conf
  - ceph config …
- PG autoscaling
  - less reliance on operator to pick right values
- progress bars
  - recovery, rebalancing, etc.
- async recovery
  - fewer high-latency outliers during recovery
QUALITY OF SERVICE

- Ongoing background development
  - dmclock distributed QoS queuing
  - minimum reservations and priority weighting
- Range of policies
  - IO type (background, client)
  - pool-based
  - client-based
- Theory is complex
- Prototype is promising, despite simplicity
- Basics will land in Mimic
IOPS1 (R=50 ,W=1) IOPS2 (R=100 ,W=1) IOPS3 (R=0 ,W=10) IOdepth = 32

- IOPS1 (Avg = 49.8083333333)
- IOPS2 (Avg = 99.725)
- IOPS3 (Avg = 115.925)
- new RADOS ‘redirect’ primitive
  - basically a symlink, transparent to librados
  - replace “sparse” cache tier with base pool “index”
Generalize redirect to a “manifest”
  - map of offsets to object “fragments” (vs a full object copy)

Break objects into chunks
  - fixed size, or content fingerprint

Store chunks in content-addressable pool
  - name object by sha256(content)
  - reference count chunks

TBD
  - inline or post?
  - policy (when to promote inline, etc.)
  - agent home (embed in OSD, …)
CephFS

- Integrated NFS gateway management
  - HA nfs-ganesha service
  - API configurable, integrated with Manila
- snapshots!
- quota for kernel client
CLIENT CACHES!

- **RGW**
  - persistent read-only cache on NVMe
  - fully consistent (only caches immutable “tail” rados objects)
  - Mass Open Cloud
- **RBD**
  - persistent read-only cache of immutable clone parent images
  - writeback cache for improving write latency
    - cluster image remains crash-consistent if client cache is lost
- **CephFS**
  - kernel client already uses kernel fscache facility
GROWING DEVELOPER COMMUNITY

Authors (May 2016 - Jan 2012)

Commits (Feb 2012 - Feb 2016)
GROWING DEVELOPER COMMUNITY

- Red Hat
- Mirantis
- SUSE
- ZTE
- China Mobile
- XSky
- Digiware
- Intel
- Kylin Cloud
- Easystack
- Istuary Innovation Group
- Quantum
- Mellanox
- H3C
- Quantum
- UnitedStack
- Deutsche Telekom
- Reliance Jio Infocomm
- OVH
- Alibaba
- DreamHost
- CERN
OPENSTACK VENDORS

- Red Hat
- Mirantis
- SUSE
- ZTE
- China Mobile
- XSky
- Digiware
- Intel
- Kylin Cloud
- Easystack
- Istuary Innovation Group
- Quantum
- Mellanox
- H3C
- Quantum
- UnitedStack
- Deutsche Telekom
- Reliance Jio Infocomm
- OVH
- Alibaba
- DreamHost
- CERN
CLOUD OPERATORS

- Red Hat
- Mirantis
- SUSE
- ZTE
- China Mobile
- XSky
- Digiware
- Intel
- Kylin Cloud
- Easystack
- Istuary Innovation Group

- Quantum
- Mellanox
- H3C
- Quantum
- UnitedStack
- Deutsche Telekom
- Reliance Jio Infocomm
- OVH
- Alibaba
- DreamHost
- CERN
HARDWARE AND SOLUTION VENDORS

- Red Hat
- Mirantis
- SUSE
- ZTE
- China Mobile
- XSky
- Digiware
- Intel
- Kylin Cloud
- Easystack
- Istuary Innovation Group
- Quantum
- Mellanox
- H3C
- Quantum
- UnitedStack
- Deutsche Telekom
- Reliance Jio Infocomm
- OVH
- Alibaba
- DreamHost
- CERN
<table>
<thead>
<tr>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat</td>
</tr>
<tr>
<td>Mirantis</td>
</tr>
<tr>
<td>SUSE</td>
</tr>
<tr>
<td>ZTE</td>
</tr>
<tr>
<td>China Mobile</td>
</tr>
<tr>
<td>XSky</td>
</tr>
<tr>
<td>Digiware</td>
</tr>
<tr>
<td>Intel</td>
</tr>
<tr>
<td>Kylin Cloud</td>
</tr>
<tr>
<td>Easystack</td>
</tr>
<tr>
<td>Istuary Innovation Group</td>
</tr>
<tr>
<td>Quantum</td>
</tr>
<tr>
<td>Mellanox</td>
</tr>
<tr>
<td>H3C</td>
</tr>
<tr>
<td>Quantum</td>
</tr>
<tr>
<td>UnitedStack</td>
</tr>
<tr>
<td>Deutsche Telekom</td>
</tr>
<tr>
<td>Reliance Jio Infocomm</td>
</tr>
<tr>
<td>OVH</td>
</tr>
<tr>
<td>Alibaba</td>
</tr>
<tr>
<td>DreamHost</td>
</tr>
<tr>
<td>CERN</td>
</tr>
</tbody>
</table>
GET INVOLVED

- Mailing list and IRC  
  - [http://ceph.com/IRC](http://ceph.com/IRC)

- Ceph Developer Monthly  
  - first Weds of every month  
  - video conference (Bluejeans)  
  - alternating APAC- and EMEA-friendly times

- Github  
  - [https://github.com/ceph/](https://github.com/ceph/)

- Ceph Days  

- Meetups  
  - [http://ceph.com/meetups](http://ceph.com/meetups)

- Ceph Tech Talks  

- ‘Ceph’ Youtube channel  
  - (google it)

- Twitter  
  - @ceph
CEPHALOCON APAC

2018.03.22 and 23
BEIJING, CHINA
THANK YOU!

Douglas Fuller
Senior Engineer

dfuller@redhat.com
@djfuller0