Ceph Development Update (HEPiX 2015, Oxford)

John Spray

john.spray@redhat.com
jcsp on #ceph-devel

Agenda

- What's new in Ceph 0.94 *Hammer*?
- CephFS work in Firefly->Hammer period
- Ongoing CephFS work

Ceph 0.94 *Hammer*

Emperor

Firefly

Giant

Hammer

Infernalis

Jewel

RADOS

- Performance:
  - more IOPs
  - exploit flash backends
  - exploit many-cored machines
- CRUSH straw2 algorithm:
  - reduced data migration on changes
- Cache tiering:
  - read performance, reduce unnecessary promotions

RBD

- Object maps:
  - per-image metadata, identifies which extents are allocated.
  - optimisation for clone/export/delete.
- Mandatory locking:
  - prevent multiple clients writing to same image
- Copy-on-read:
  - improve performance for some workloads

RGW

- S3 object versioning API
  - when enabled, all objects maintain history
  - GET ID+version to see an old version
- Bucket sharding
  - spread bucket index across multiple RADOS objects
  - avoid oversized OMAPs
  - avoid hotspots

Agenda

- What's new in Ceph 0.94 *Hammer*?
- **CephFS work in Firefly->Hammer period**
- Ongoing CephFS work

CephFS Stats (Firefly->Hammer)

- Code:
  - src/mds: 366 commits, 19417 lines added or removed
  - src/client: 131 commits, 4289 lines
  - src/tools/cephfs: 41 commits, 4179 lines
  - ceph-qa-suite: 4842 added lines of FS-related python
- Issues:
  - 108 FS bug tickets resolved since Firefly (of which 97 created since firefly)
  - 83 bugs currently open for filesystem, of which 35 created since firefly
  - 31 feature tickets resolved

Diagnostics

- `session ls` & client metadata
- `dump_ops_in_flight` (OpTracker)
- MDS Health warnings: soft checks for misbehaving clients

Testing

- New ceph-qa-suite code in tasks/cephfs
- Beyond typical filesystem tests:
  - Verify RADOS updates are as expected
  - Test specific damage/corruption scenarios
  - Test multi-client lock interaction
  - Test recovery behaviour, client eviction

Metadata scrub

- `flush_path` admin socket command
  - traverse tree
  - validate on-disk objects vs. in-memory state
  - validate recursive statistics
  - building block for continuous background scrub
- `mds_verify_backtrace` check + fix dirfrag backtraces on load

Journal diagnostic & repair

- `cephfs-journal-tool`
  - characterise damaged journals
  - extract inodes/dentries from surviving regions
  - reset journal
- `ceph fs reset`
  - restore a single-MDS map after offline disaster recovery

Better behaviour

- Ability to more reliably fence clients
- Redesigned ENOSPC handling
- `mds_max_file_recover` – avoid overloading RADOS cluster
- Functional client cache trimming
- Write error handling (readonly mode)

Agenda

- What's new in Ceph 0.94 *Hammer*?
- CephFS work in Firefly->Hammer period
- **Ongoing CephFS work**

CephFS: Ongoing work

- Full online scrub
- Repair tools for handling errors in scrub
- Debugging snapshot code
- More refined behaviour (error handling, op throttling)
- Client authorization (root squash & more)

Discussion