CephFS in 2017 (*Kraken*→*Luminous* update)

John Spray

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

CephFS at Red Hat

- POSIX compatible scale-out distributed filesystem built on Ceph object store (RADOS).
- Tech preview feature in Red Hat Ceph Storage 2.0, to be supported fully in 2017 in Red Hat Ceph Storage 3.0
- Red Hat development team focussed on stabilisation, testing, supportability.
- Integration with OpenStack via Manila driver

The *Kraken* release (January 2017)

- Mainly **bug fixes** (80+ in src/mds, src/client)
- **Directory fragmentation** improvements
- **Mantle** (Lua plugins for multi-mds balancer)
- **libcephfs API** changes (enable proper uid/gid enforcement in samba/nfs-ganesha bindings)
- **statx** support (in userspace library for now, in kernel client later once kernel finalizes statx interface)
- New “cephfs-data-scan **pg_files**” command for identifying files damaged with bad PGs.

Meanwhile in RADOS...

- **Bluestore** (experimental)
  - Disk format stabilized in Kraken
- **Erasure coding** with object overwrites
  - Enable direct consumption of EC pools without a cache tier
  - Precursor to use of EC pools as CephFS data pools
- **ceph-mgr**
  - Precursor to better usability for CephFS commands (e.g. central view of clients, eviction, fsck)

The *Luminous* release (Summer 2017)

- Stabilise **MDS scale-out** and directory fragmentation
- Feature to **pin** directories to MDS ranks for better control in multi-MDS clusters
- More scalable **deletion** (files waiting to be purged no longer occupy space in metadata cache)
- More robust **client eviction** (integrate with OSD blacklisting)
- Better **fsck usability** (check completion of ongoing scrub)
- **Kernel client** updates to match userspace client features (ENOSPC handling, multi-fs, namespace layouts)

Relation to Red Hat products

- Luminous Ceph code: **RHCS 3.0**
- Updated kernel client: **RHEL 7.4**
- CephFS Manila driver: **RHEL OSP 12**

Feature flags

	Kraken	Luminous
Multi-MDS	No (known unstable)	Yes
Snapshots	No (known unstable)	No (known unstable)
Inline data	No (limited testing)	No (limited testing)
Multi-FS	No (limited testing)	No (limited testing)
Directory fragmentation	No (limited testing)	Yes

Default settings of the feature flags by upstream Ceph release