Luminous: pg upmap (dev)

Dan van der Ster, CERN IT Storage Group

2 October 2017 | Clermont-Ferrand, France

ceph osd pg-upmap

```
ceph osd pg-upmap <pgid> <osdname (id|osd.id)> [<osdname (id|osd.id)>...]
ceph osd pg-upmap-items <pgid> <osdname (id|osd.id)> [<osdname (id|osd.id)>...]
```

- Upmap allows us to map the *up* set for a PG to a different set of OSDs.
- Ex: suppose we have PG 1.7 with `up=[0,2,1]` (osd.0, osd.2, osd.1)
  - We do `ceph osd pg-upmap-items 1.7 0 4`
  - For pg 1.7, use osd.4 in place of osd.0.
  - PG 1.7 after upmap has `up=[4,2,1]`

osdmaptool --upmap

- Start with an unbalanced cluster

```
# ceph osd getmap -o om
got osdmap epoch 3331

# osdmaptool om --mark-up-in --test-map-pgs --pool 4 | egrep 'avg|min|max'
osdmaptool: osdmap file 'om'
 avg 31 stddev 6.606 (0.213097x) (expected 5.59854 0.180598x))
 min osd.71 19
 max osd.45 52
```

osdmaptool --upmap

- Call the OSDMap::calc_pg_upmaps routine

```
# osdmaptool om --mark-up-in --upmap-max 500 --upmap-pool test --upmap c
--upmap-save c
osdmaptool: osdmap file 'om'
marking all OSDs up and in
writing upmap command output to: c
checking for upmap cleanups
upmap, max-count 500, max deviation 0.01
 limiting to pools test (4)
osdmaptool: writing epoch 3333 to om
```

Bug in 12.2.0 related to cyclic upmaps
Luminous upmap does not handle reweighted OSDs. (PR#17944)

osdmaptool --upmap

- Inspect the output of calc_pg_upmaps

```
# wc -l c
145 c

# head -n4 c
ceph osd pg-upmap-items 4.0 60 72
ceph osd pg-upmap-items 4.1 37 28 79 71
ceph osd pg-upmap-items 4.2 69 42 51 83
ceph osd pg-upmap-items 4.3 15 19
```

osdmaptool --upmap

- Check the new distribution

```
# osdmaptool om --mark-up-in --test-map-pgs --pool 4 | egrep
'avg|min|max'
osdmaptool: osdmap file 'om'
 avg 31 stddev 0.989637 (0.0319238x) (expected 5.59854 0.180598x))
 min osd.0 31
 max osd.13 33
```

ceph-mgr balancer

- PR#16272 introduced mgr/balancer

- [x] back off when degraded, unknown, inactive
- [x] throttle against misplaced ratio
- [x] upmap (luminous+)
- [ ] crush-legacy (compat with pre-luminous)
- [ ] crush (luminous+)
- [ ] osd_weights (legacy osd weight-based approach) (probably not worth doing this!)
- [ ] phase out balance optimizations from other modes (e.g., phase out osd_weight if we are optimizing crush weights)

https://github.com/ceph/ceph/pull/16272
Merged to master, not yet in Luminous.

References

- OSDMap::calc_pg_upmaps:
  - https://github.com/ceph/ceph/blob/master/src/osd/OSDMap.cc#L3792
- ceph-mgr balancer module:
  - https://github.com/ceph/ceph/blob/master/src/pybind/mgr/balancer/module.py