Incomplete PGs & Atlas data loss

Tom Byrne

Timeline (18th-25th August)

- OSD removed due to disk errors
  - This OSD was primary of a backfilling PG with a newly introduced OSD (which was seeing thousands on read errors
- The first 3 OSDs in the set started crashing, flapping and finally died
  - Removing the problem OSD(s) didn't help
- Finally, the PG was manually removed and recreated from all OSDs in the set, causing data loss of ~23K files
- In the following week, we have had several other PGs become incomplete due to this issue
  - Fortunately, these have all been recoverable after removing OSD with problem disk

Bug

- In Kraken, if a read error is encountered during a backfill operation on an EC pool, the primary OSD will crash.
  - This will cascade unless the problem OSD is identified and removed
- When adding large amounts of new OSDs (with an expected amount of problem disks), this bug suddenly became very apparent.

http://tracker.ceph.com/issues/18162

Bug

OSD with bad disk:

```
2017-08-31 16:26:58.490330 7f268f2f6700 -1 log_channel(cluster) log [ERR] : handle_sub_read: Error -5 reading
1:6a822d17:::datadisk%2frucio%2fdata16_13TeV%2fb2%2fe7%2fAOD.11192195._009893.pool.root.1.0000000000000005:head
```

Primary OSD:

```
2017-08-31 16:26:58.490760 7f5459ed1700  0 osd.17 pg_epoch: 74771 pg[1.156s0( v 74771'735383 (74646'732237,74771'735383] local-les=74692 n=23152 ec=764 les/c/f 74692/74659/0
74691/74691/73684) [17,1279,640,505,1307,1045,1718,683,508,263,848]/[17,1279,640,505,1307,1045,1713,683,508,263,848] r=0 lpr=74691 pi=74647-74690/1 luod=74771'735382 rops=1
bft=1718(6) crt=74771'735382 lcod 74771'735381 mlcod 74771'735381 active+remapped+backfilling] failed_push
1:6a822d17:::datadisk%2frucio%2fdata16_13TeV%2fb2%2fe7%2fAOD.11192195._009893.pool.root.1.0000000000000005:head from shard 1713(6), reps on  unfound? 0
2017-08-31 16:26:58.527089 7f54576cc700 -1 osd.17 pg_epoch: 74771 pg[1.156s0( v 74771'735383 (74646'732237,74771'735383] local-les=74692 n=23152 ec=764 les/c/f 74692/74659/0
74691/74691/73684) [17,1279,640,505,1307,1045,1718,683,508,263,848]/[17,1279,640,505,1307,1045,1713,683,508,263,848] r=0 lpr=74691 pi=74647-74690/1 bft=1718(6) crt=74771'735382
lcod 74771'735382 mlcod 74771'735382 active+remapped+backfilling] recover_replicas: object added to missing set for backfill, but is not in recovering, error!
2017-08-31 16:26:58.753517 7f54576cc700 -1 *** Caught signal (Aborted) **
 in thread 7f54576cc700 thread_name:tp_osd_tp


 ceph version 11.2.1 (e0354f9d3b1eea1d75a7dd487ba8098311be38a7)
 1: (()+0x9323ca) [0x7f5478dcf3ca]
 2: (()+0xf370) [0x7f5475eeb370]
 3: (gsignal()+0x37) [0x7f5474d091d7]
 4: (abort()+0x148) [0x7f5474d0a8c8]
 5: (PrimaryLogPG::recover_replicas(unsigned long, ThreadPool::TPHandle&)+0x6cd) [0x7f5478a8cdbd]
 6: (PrimaryLogPG::start_recovery_ops(unsigned long, ThreadPool::TPHandle&, unsigned long*)+0x568) [0x7f5478a949f8]
 7: (OSD::do_recovery(PG*, unsigned int, unsigned long, ThreadPool::TPHandle&)+0x40c) [0x7f547890f1ac]
 8: (OSD::ShardedOpWQ::_process(unsigned int, ceph::heartbeat_handle_d*)+0x68b) [0x7f5478915d1b]
 9: (ShardedThreadPool::shardedthreadpool_worker(unsigned int)+0x945) [0x7f5478f76ca5]
 10: (ShardedThreadPool::WorkThreadSharded::entry()+0x10) [0x7f5478f78e00]
 11: (()+0x7dc5) [0x7f5475ee3dc5]
 12: (clone()+0x6d) [0x7f5474dcb76d]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.
```