Recovery vs. Backfill

Zhiqiang Wang @XSKY

2016.11.25

X·SKY
www.xsky.com

PG info, PG log和Peering

- PG info记录一个PG的元数据，包括PG上数据最后更新的时间版本，最后一次peering完成的时间，最后一次recovery完成的时间，等等
- PG Log记录一个PG中所有对象最近的更新，跟数据库和某些文件系统中的日志是同样的概念
- Peering指的是将一个PG中所有对象的数据和元数据的状态在该PG所处的所有OSD上达成一致的过程
- Peering过程依赖于PG info和PG log，会找出一个权威的OSD上有最新的PG info和PG log，并分发给这个Pg所处的其他OSD成员
- 当OSDMap发生改变时，OSD上的所有的PGs就需要做peering

Recovery和Backfill

- 根据peering过程中选出的权威OSD上的PG log，将其与该PG的其他OSD成员上的PG log比较
- 如果二者有重叠，则可以做log based recovery，即增量恢复。根据log算出哪些对象缺失，只需恢复缺失的对象
- 否则的话，需要做backfill recovery，即全量恢复。根据hash排序扫描PG中所有对象，依次进行恢复
- OSD短时间下线 - recovery
- OSD长时间下线 - backfill

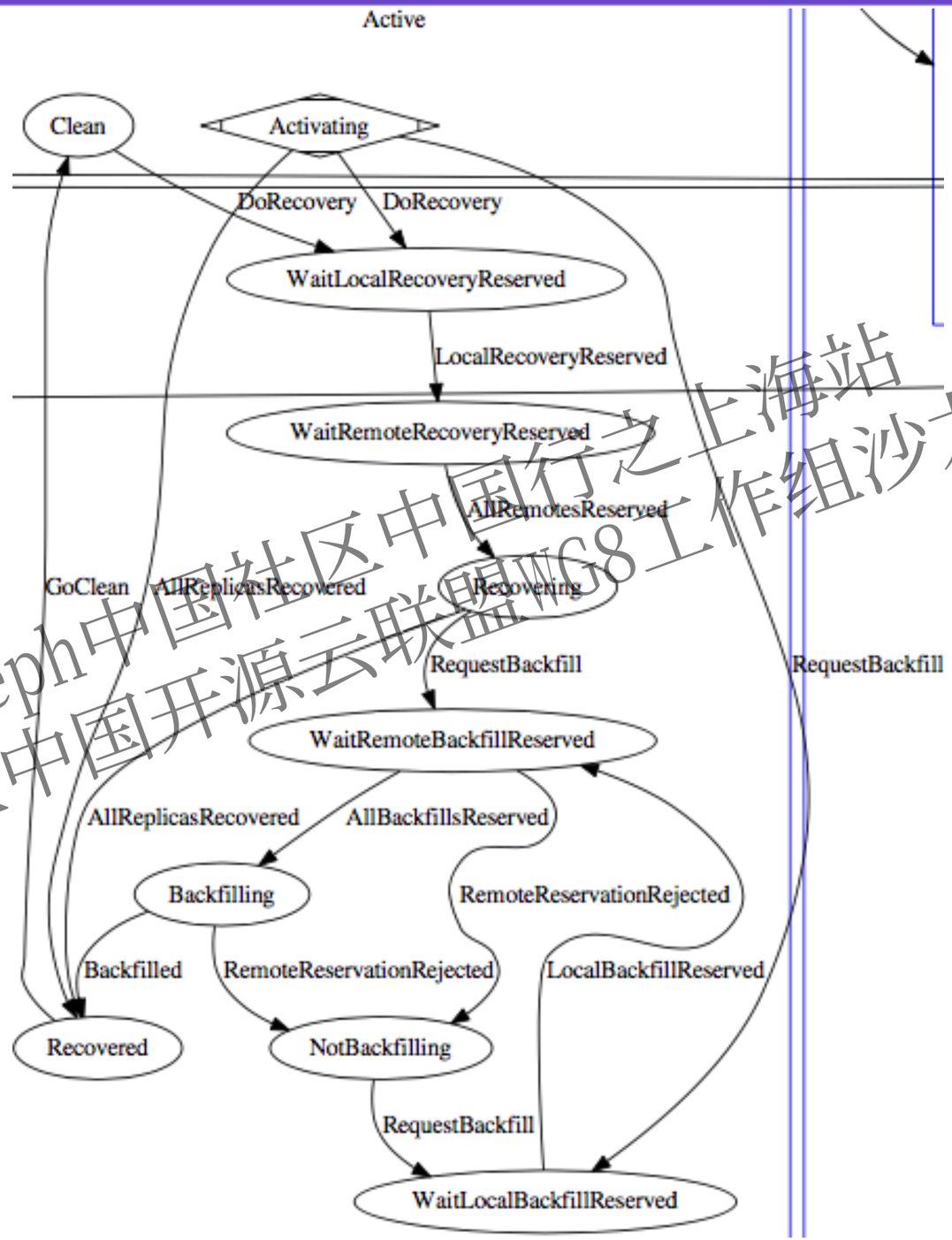
Backfill和PG Temp

- 如果在完成peering后，PG的某些副本需要做backfill，需要生成PG temp
- PG temp将一个PG临时映射到一组其他的OSD上，生成新的OSDMap
- PG根据其以前OSDMap的那些OSD成员来选择临时OSD组的成员
- Epoch 1000: [0, 1, 2]
- Epoch 1100: [3, 4, 5]
- Epoch 1110: [0, 1, 2]
- Epoch 1190: [3, 4, 5]

Recovery/Backfill优先级

- 主副本OSD发起recovery/backfill操作时，需先从所有副本OSD“预约”
- 先“预约”主副本，然后再“预约”从副本
- Log based recovery，主副本对象缺失 – 最高优先级
- Log based recovery，从副本对象缺失
- Backfill，PG多副本降级
- Backfill，PG单副本降级
- Backfill，PG没有降级 – 最低优先级

Recovery/Backfill状态转换



Recovery对IO影响

- 主动recovery
- 被动recovery

Backfill对IO影响

- Epoch 1000: [0, 1, 2]
- Epoch 1100: [3, 4, 5]
- Epoch 1110: [0, 1, 2], backfill targets – [3, 4, 5]
- Epoch 1190: [3, 4, 5]

Backfill对IO影响

- Epoch 1000: [0, 1, 2]
- Epoch 1100: [3, 4, 5]
- Epoch 1110: [0, 1, 2], backfill targets – [3, 4, 5]
- Epoch 1190: [3, 4, 5]

Recovery/Backfill速度控制参数

- osd_recovery_max_active
- osd_recovery_max_single_start
- osd_recovery_max_chunk
- osd_recovery_op_priority
- osd_recovery_threads
- osd_max_backfills

But..

Async recovery

- 介于recovery和backfill中间的一种数据恢复方式
- 也是一种log based recovery
- 不阻塞IO
- https://github.com/ceph/ceph/pull/11918
- 测试结果显示在集群需要做数据恢复操作时，对IO的影响降低到几秒钟，然后恢复正常性能

THANK YOU

2016.11.26

X·SKY
www.xsky.com