异构存储感知的Ceph存储系统数据放置方法

刘飞，蒋德钧，熊劲
中国科学院 计算技术研究所

背景

Ceph分布式存储系统正成为广泛使用的开源云环境存储解决方案，异构存储如果应用有效的数据管理策略能够在保持低成本的同时提供大容量和高性能存储，Ceph的Crush算法不能将SSD与HDD区分对待，使得在Ceph中使用异构存储设备不能有效发挥异构存储设备的性能，由于数据多个副本可以存放到不同的存储介质，不同的副本组合的性能和成本都不一样。

现有工作

- SSD作为本地Cache使用：
  - E.g.: FlashCache, Azor, SmartSaver
- SSD作为本地持久化存储：
  - E.g.: Hybrid Store, Hybrid Aggregates
- 这些工作都没有针对公布式系统

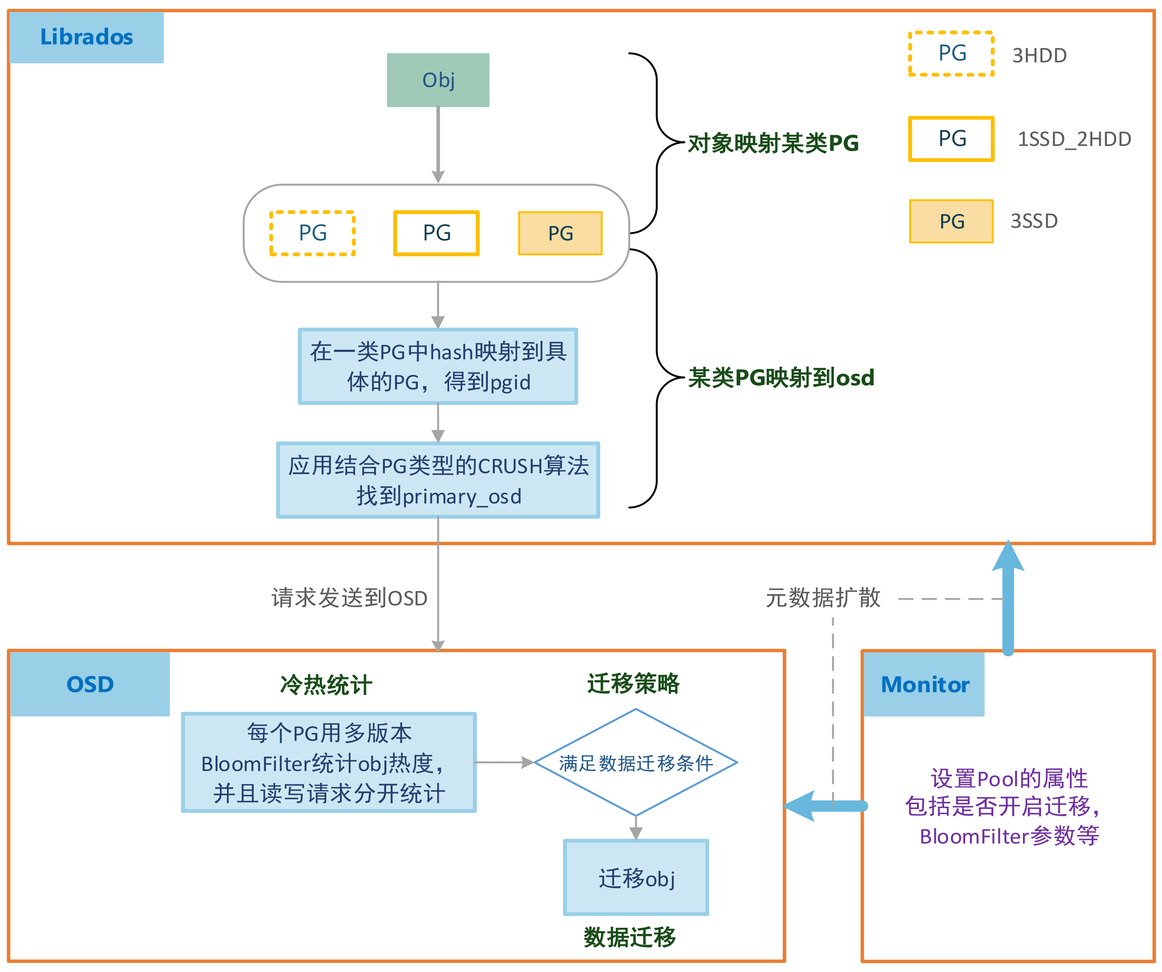
分布式异构存储系统关键问题

- 数据映射：怎样确定请求数据位置？
- 数据热点识别：如何区分热点数据？
- 数据迁移：数据该存储在何种介质中，何时迁移

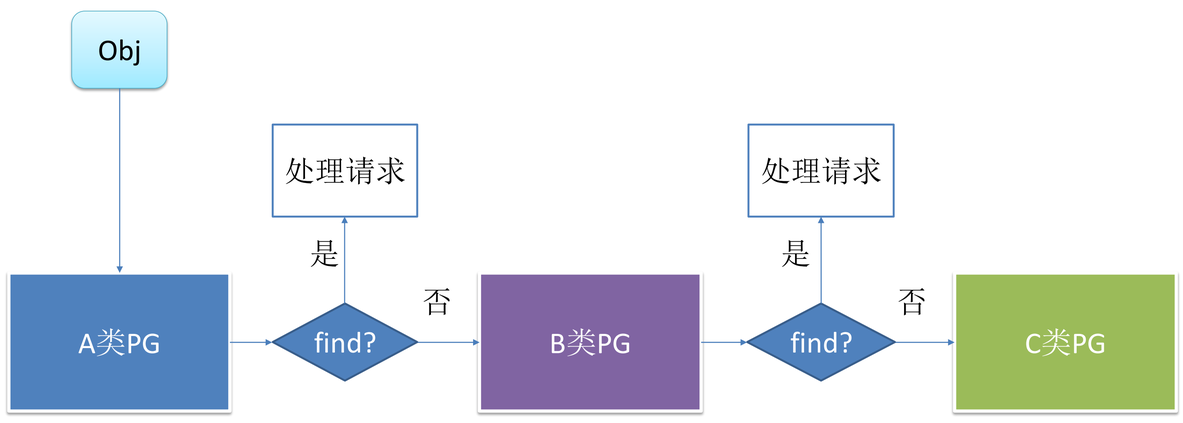
系统设计

- 把副本组合分为多种group：
  - 3HDD用于存储冷数据
  - 3SSD用于存放读写频繁的热数据
  - 1SSD+2HDD用于存放读多写少的数据
- 收集对象的读写热度信息：用Bloom Filter实现
- 定期根据热度让对象在group之间迁移

系统架构



对象查找



评测结果

TPC-W性能测试

Browsing Mix访问模式SSD使用量

Block I/O traces性能测试