# 课题列表

## Ceph下的EC性能和可靠性研究

* **小IO更新优化**：对于数据块的更新，尤其是更新数据量（少于64KB）较少时的性能不理想
* 故障后的恢复效率与代价，主要考虑恢复时引起的数据传输总量
* 底层存储设备需要基于HDD，并且需要支持掉电保护
* 理想的解决办法：新的编码+新的架构+强一致性

## Ceph文件系统中多MDS的可靠性研究

* **元数据重建**：在海量小文件场景中，海量元数据的可靠性保证当MDS集群中出现某台MDS宕机等故障时，MDS集群可以自适应重建整个命名空间，并且对在线性能影响较小
* **热点文件的优化**：针对热点文件，现有的技术多是基于热点多备份的方式，将热点文件的元数据复制到多台MDS中以提高访问的并发性，但是现有的CephFS是基于动态子树对文件系统的命名空间进行划分，导致迁移的开销很大，性能呈断崖式下降
* **文件系统快照技术**：海量文件、快照的数量比较大、在制作快照时保证在线性能的稳定性

## Ceph下的存储接口优化

* **小对象存取性能优化**：对几个字节大小的对象存取性能较差（对接S3接口），现有的Ceph对象存储RADOS和对象网关RGW难以提供良好的接口对接性能
* **查询优化**：针对数据查询操作（对象-PG-OSD多层映射），需要在多个进程和线程间切换（编程问题），由于线程池的切换开销较大，从而导致查询操作的软件开销较大 （需要再沟通）

## 多副本存储的智能分布和选源方式

* **副本分布优化**：针对特定的多副本分布式存储系统中，当数据写入时如何确定副本的物理存储位置（从全局角度考虑）？其中在分布式需要考虑到整个存储系统中副本和数据分布的负载均衡性，同时分布策略的动态调整
* **副本选源优化**：当处理用户读请求时，如何选择具体的副本响应读请求从而保证良好的存取性能？在进行副本选源时需要考虑cache命中率、选源方式的动态转换
* **【参考】**optimizing hash-based distributed storage using client choices @ ApSys’16

## 磁盘、固态硬盘存储系统的故障预测

根据用户的访问模式、请求延迟、垃圾回收、寿命等多维度信息分析设备的存取特征，设计故障预测模型。现在深信服还在收集数据，可协助调研。

* 【**参考**】HPE infosight，Pure 1，DiskProphet

## 集群数据去重

在分布式存储系统中，进行全局去重，考虑以下几点问题：

* 对重复数据的快速识别，从而提高重删效率
* 查询元数据以确定数据块位置和大小时，需要进行跨结点请求，加重网络流量
* 全局去重时需要跨结点比较指纹，增加网络传输量
* 对存在多副本和EC的存储系统中，如何保证重删后数据的可靠性？
* 【**参考**】华为dorado V3

## 分层存储系统的智能预取

典型的分层存储系统使用固态盘作为性能层，机械盘作为容量层，通过将热点数据放在性能层，为应用提供远远超过容量层的性能。分层存储系统的研究重点包括淘汰和预取，经典的淘汰算法包括RU、ARC、LIRS、LARC等以及理论最优的Belady算法。能否设计一种智能预取算法（例如人工智能），预测用户未来访问的数据并提前预取到性能层，突破Belady算法的理论极限，接近全闪存的性能。

可考虑利用access pattern进行预取、数据冷热识别方案。

## open-channel SSD

以往在SSD上的实验大多是采用Disksim或Flashsim等模拟器完成，为提高实验的真实性，open-channel SSD向OS敞开了大门，用户可自行管理和设计SSD内部组间的执行方式，现open-channel SSD正逐步取代模拟器成为主流的实验平台。

## 其他问题

SDS软件定义存储、Ceph架构下的热升级方案和数据重删、MDS的扩容和缩容、异构存储管理方法（包含云端存储对接与管理）、存储系统的性能预测