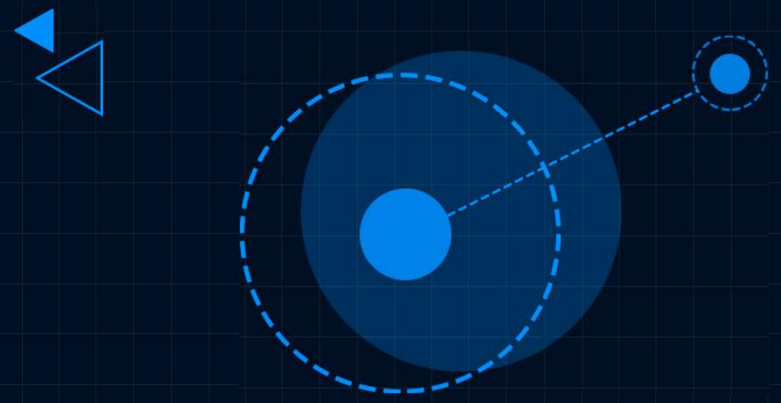
ceph纠删码在网易对象存储中的实践

俞乐勤

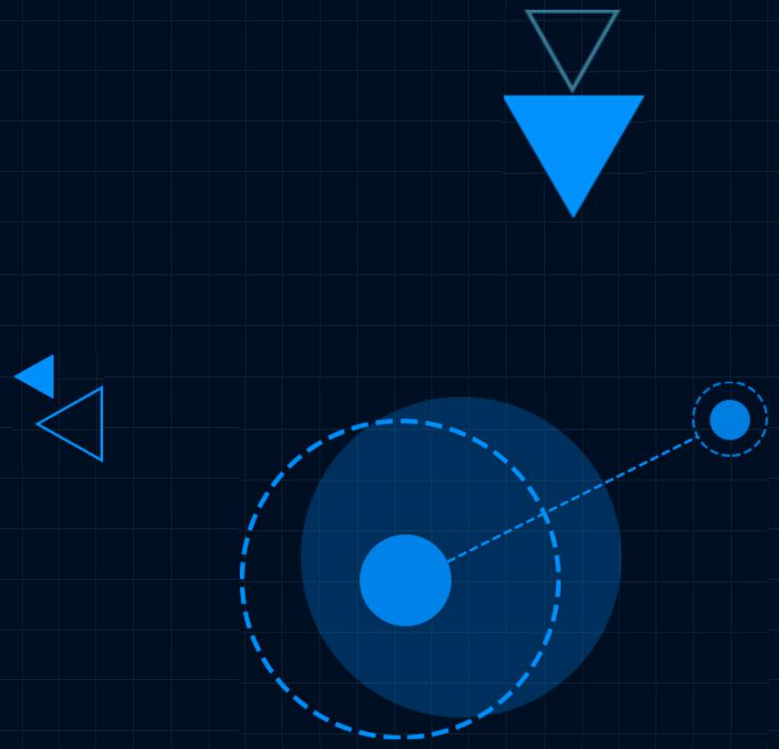
网易杭研数帆存储部门



网易对象存储

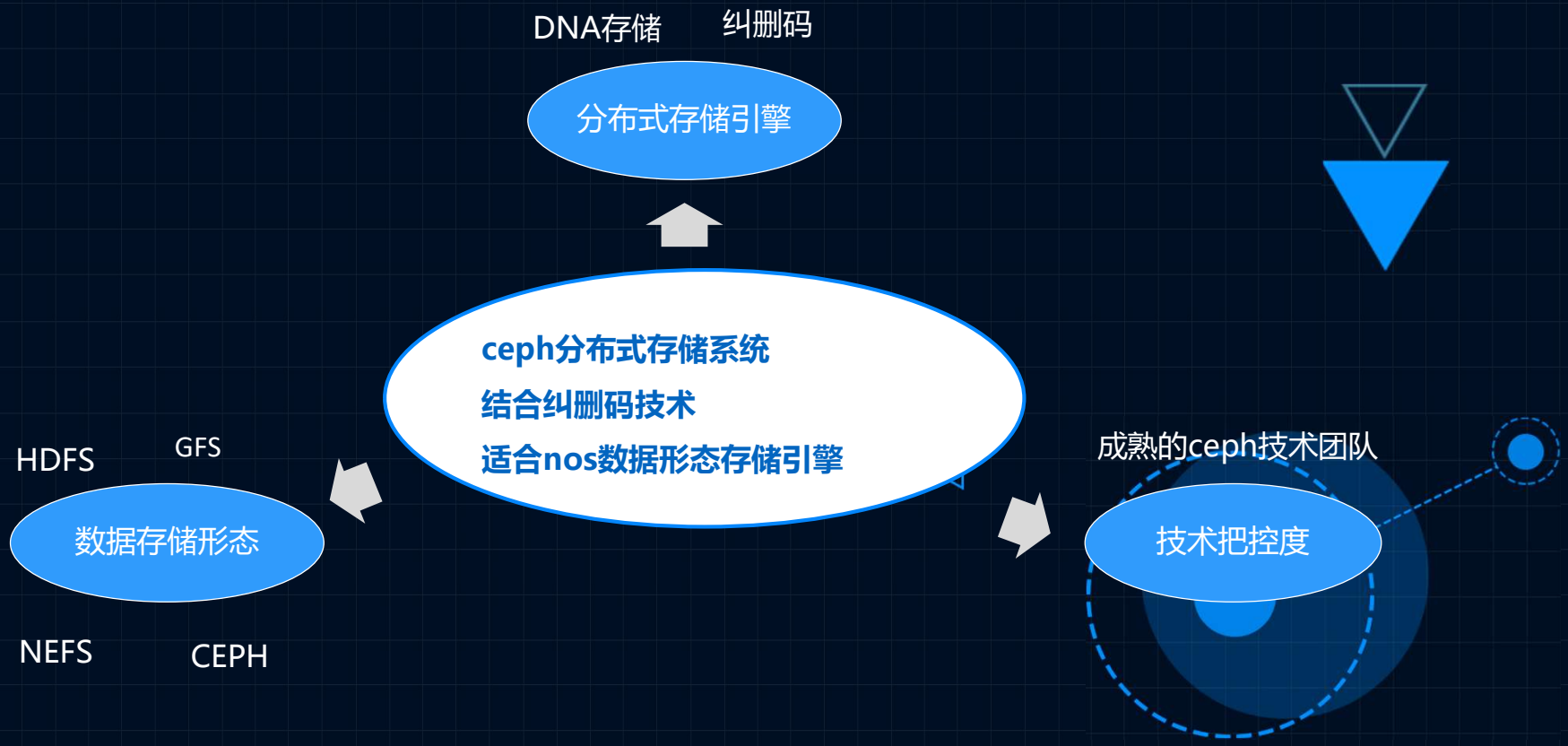
网易对象存储，Netease Object Storage，简称NOS，主要提供了以下服务：

- 数据存储服务
- 全球直传加速
- 图片/音视频服务
- CDN服务



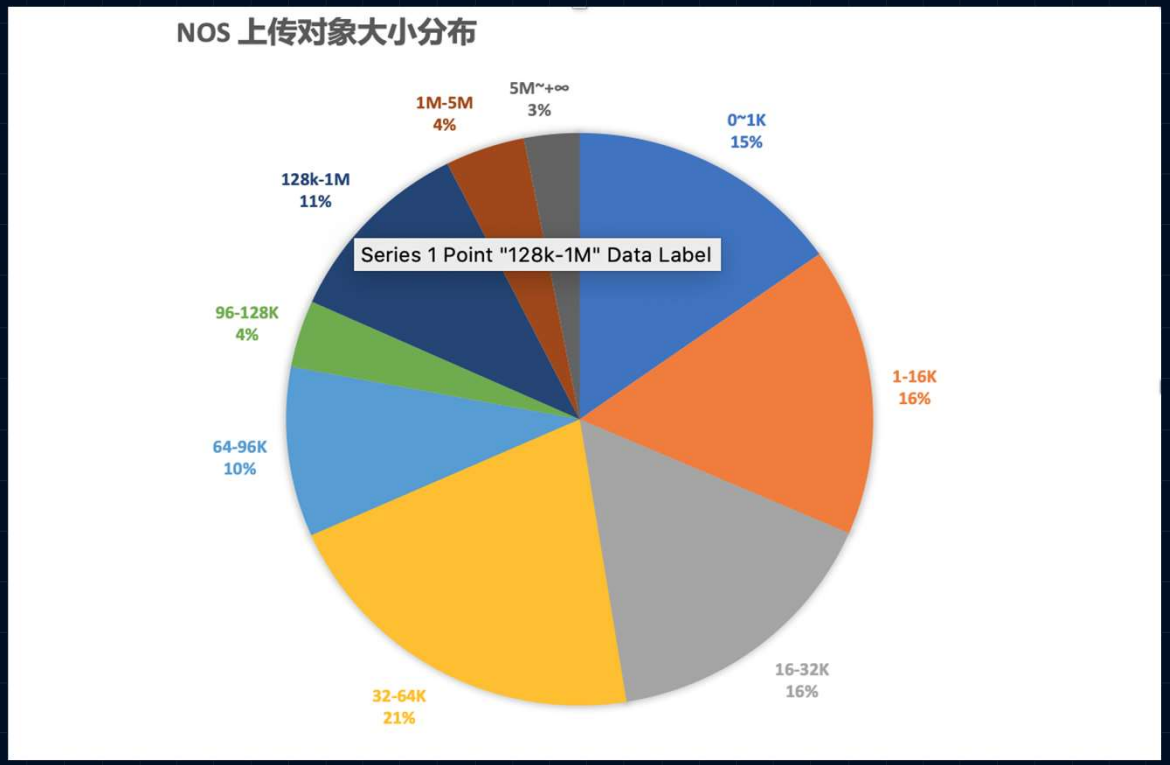
NOS低成本存储引擎探索

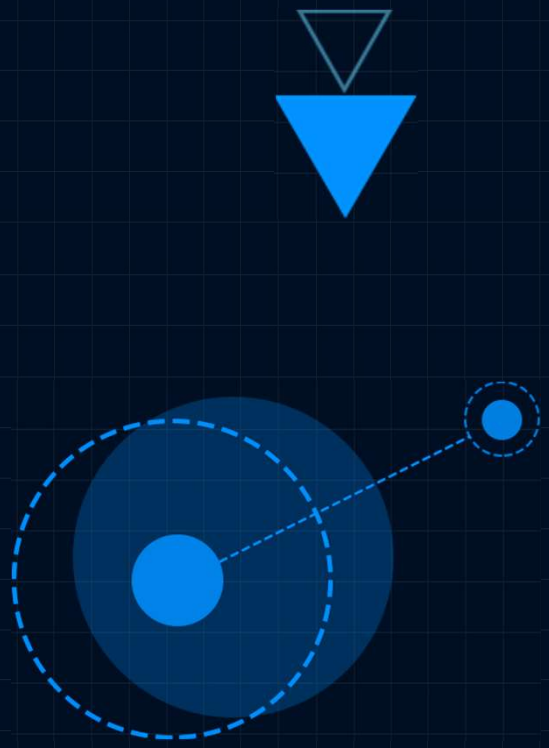
不管何种业务形态，随着数据量的爆炸式增长，数据存储的成本敏感度越高，NOS在不断探索一种低成本/高可靠性/快速迭代的存储模型。



存储空间管理之数据分布形态

从下图中可以总体观测一下NOS的文件大小分布情况，其中小文件的占比高达80%（0-128k），而ceph纠删码必须要条带对齐，大量的小文件写入会引发很多ceph空间存储和性能问题。





存储空间管理之小文件合并技术

- ceph的写入缺陷

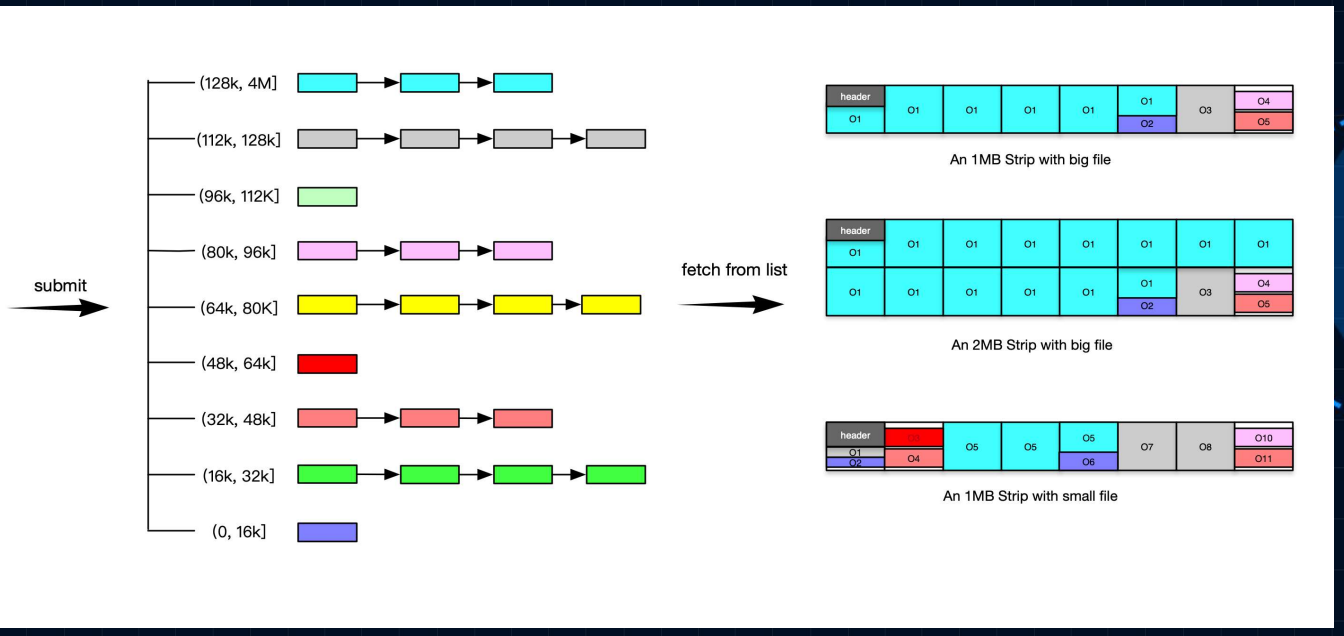
适合于大文件/流式文件的读写

小文件，虽有ceph append技术，但其写入性能极差

随着系统的小文件数量增多，ceph的元数据管理和检索时间成指数倍增加

- 小文件合并技术

nos文件列表以一定数据管理策略，进行条带合并，分析其合并率，最终写入底层ceph对象



存储空间管理之回收机制

- ceph对象空洞率

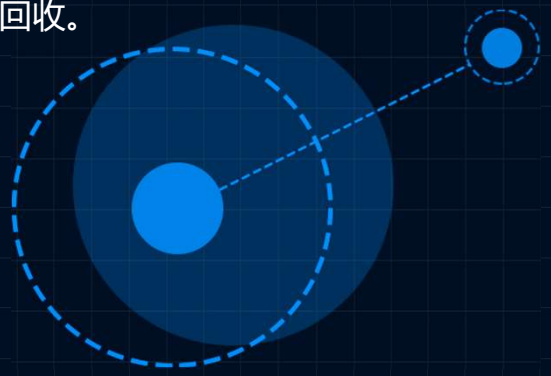
NOS小文件的合并策略导致删除的时候，ceph对象空洞率很高

- 条带写入

纠删码的条带写入，导致nos文件的有效空间和ceph对象的数据写入量存在一定的差距

- 空间回收机制

数据合并写入,nos的数据信息合并到一个header结构体中存储到ceph对象，后期不断地分析nos文件的数据有效性，进行多个ceph对象的有效空间汇聚，进行空间整合和有效回收。



系统IO性能之读性能优化

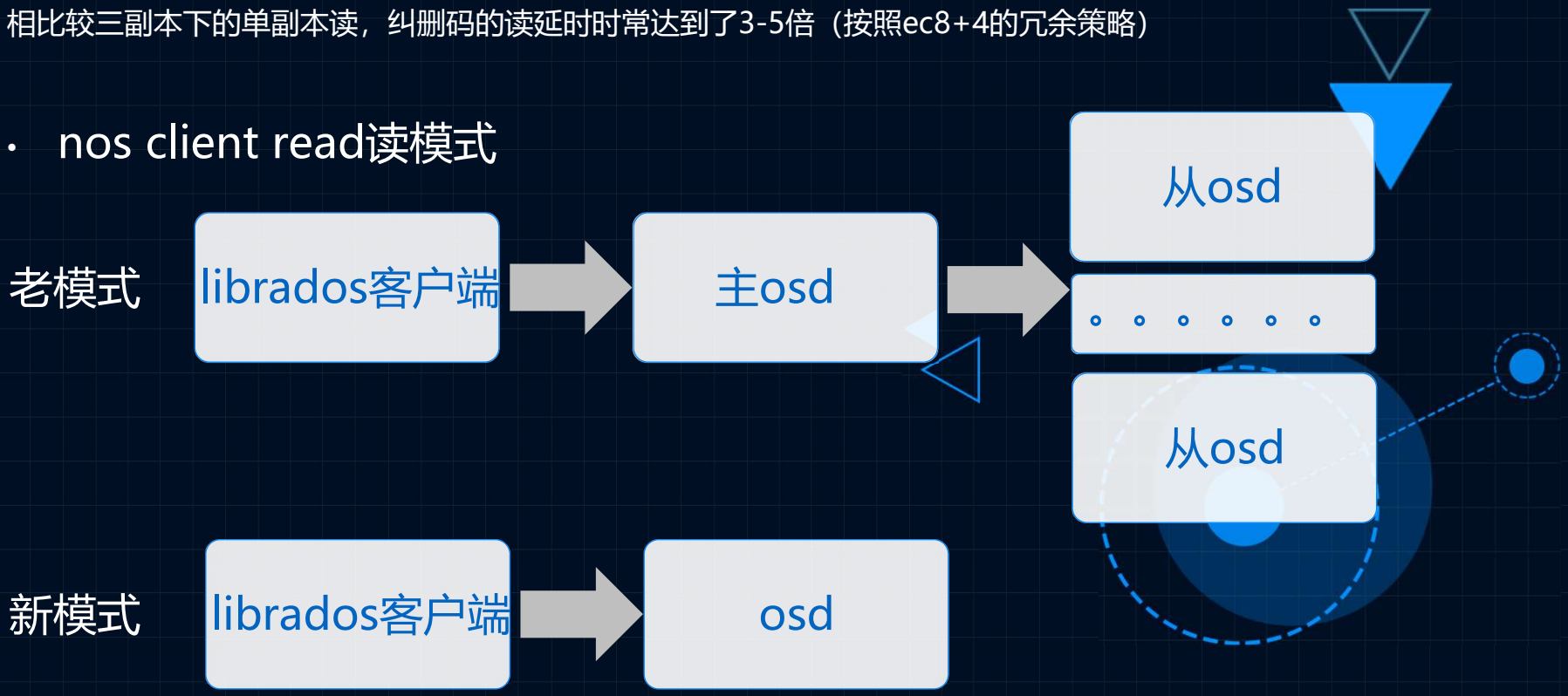
- ceph+纠删码的性能缺陷

必须满条带读，即使fast read也必须基于k+m的k份数据进行读

纠删码本身的数据离散度特别高，当k+m中的k和m值越大，osd长尾效应加剧读延时

相比较三副本下的单副本读，纠删码的读延时时常达到了3-5倍（按照ec8+4的冗余策略）

- nos client read读模式



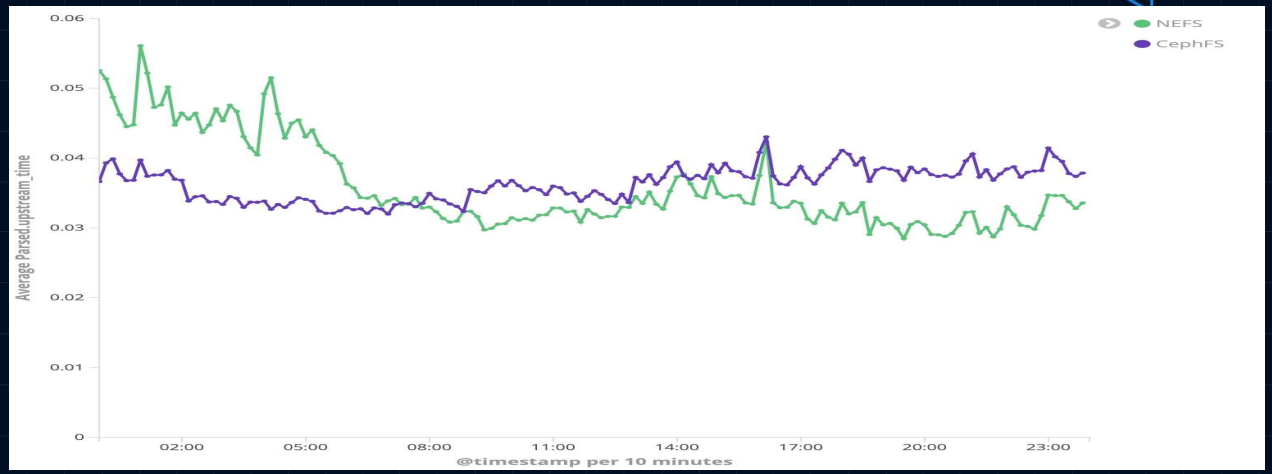
系统IO性能之读性能优化

改善后的小文件性能比ceph读提升3倍以上（8+4，条带为1M）

读类型	ceph	NOS-client-read
64k读延时	144ms	39ms
1M读延时	160ms	156ms

小文件性能和三副本基本上持平，在40ms左右

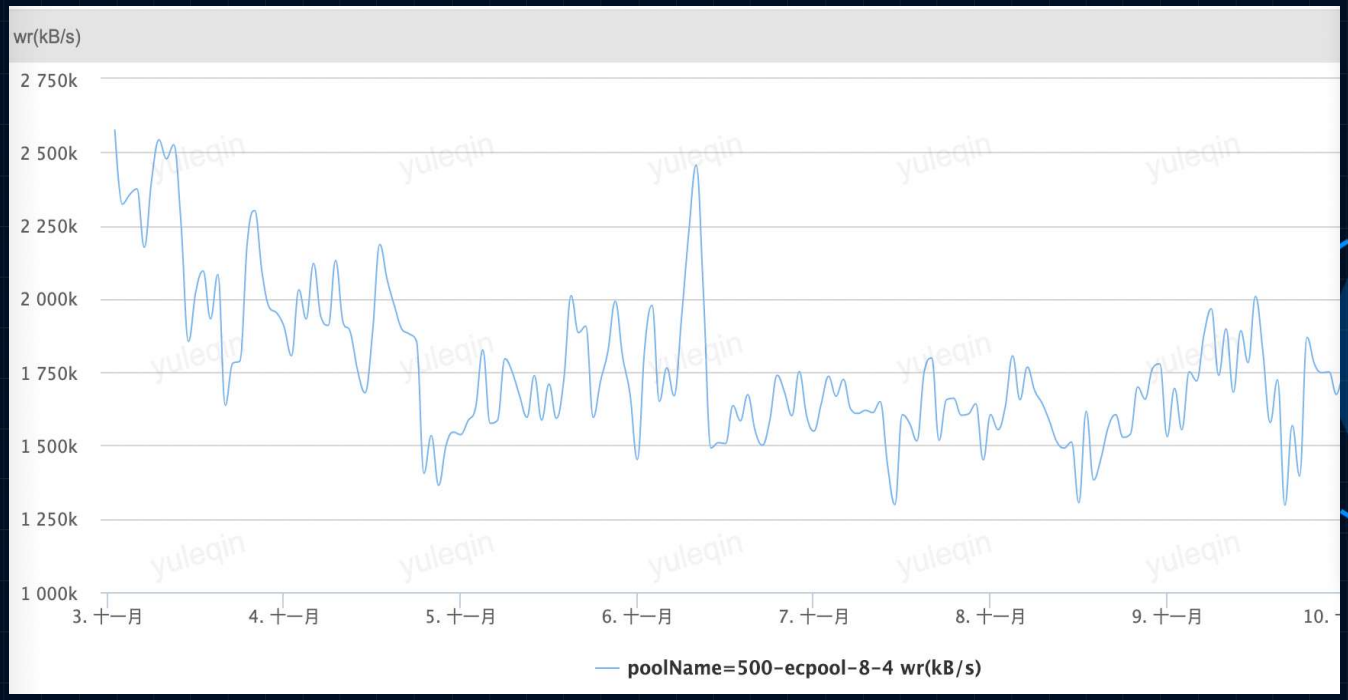
（NEFS为网易三副本存储引擎/cephfs为改进的网易纠删码存储引擎）



系统IO性能之写性能优化

小文件的合并写入不仅仅提升了物理空间利用率，从另一方面也降低了ceph系统的qps，提升了nos对象存储的系统写带宽。

- 机器信息：Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz 40 核心 / 128G RAM /西数16T 7200RPM *36
- ceph集群：432 osds
- 写带宽：正常情况下，为了平衡读写速率，可以达到2GB/s，其**极限性能在6GB/s**



数据可靠性之原地恢复

- osd故障导致集群数据的大量迁移

数据迁移：osd发生故障时，有可能导致集群的pg重新映射，使得大部分的osd数据进行迁移

集群上osd越多迁移的数据量越大，集群上的存储水位越高迁移的数据量越大

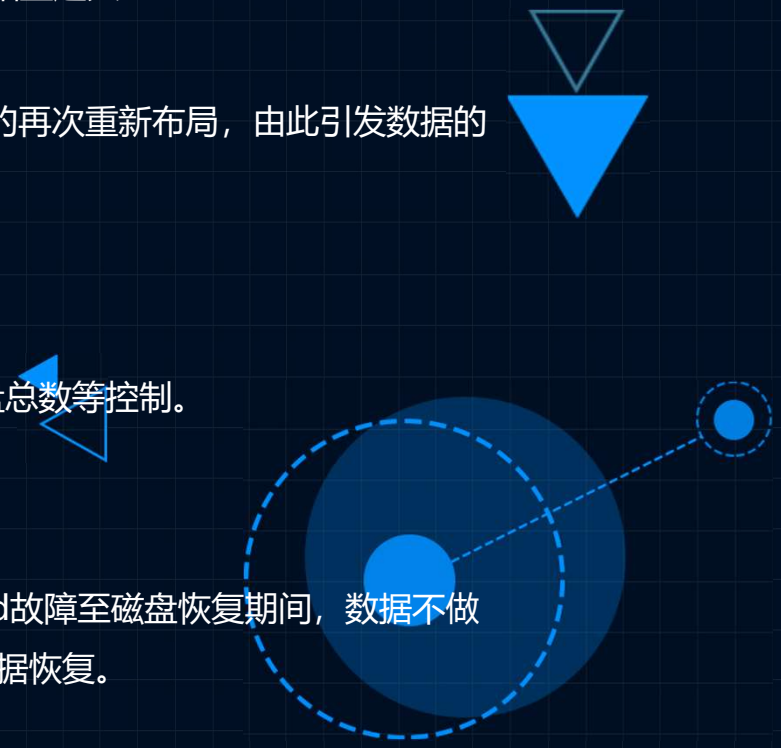
数据重迁移：故障或者换盘后的osd重新入集群，可能会导致pg的再次重新布局，由此引发数据的再次重新迁移。

- 迁移速度影响数据可靠性

数据可靠性由磁盘的故障率，磁盘的恢复数据恢复速度，集群的磁盘总数等控制。

- 原地恢复方案

nos根据数据迁移量，磁盘恢复速度，最终确定原地恢复方案。在osd故障至磁盘恢复期间，数据不做任何处理，一旦出现故障恢复或者磁盘更换完成后，进行该磁盘的数据恢复。



系统容灾/故障

- 磁盘故障时的数据完备性

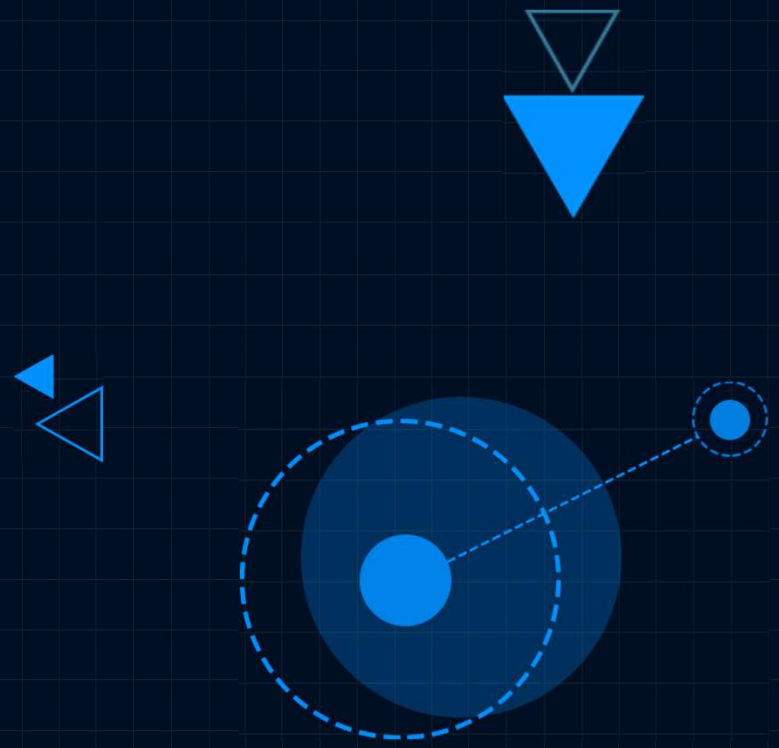
提前感知各种参与故障的pg，保证数据的写入全在状态ok的pg

ceph集群做反向通知机制来同步客户端集群的一些信息

- 系统升级时的数据可靠性

整体上做好ceph集群的升级组件过程控制

nos提前屏蔽升级节点的数据写入



整体收益

- 成本收益

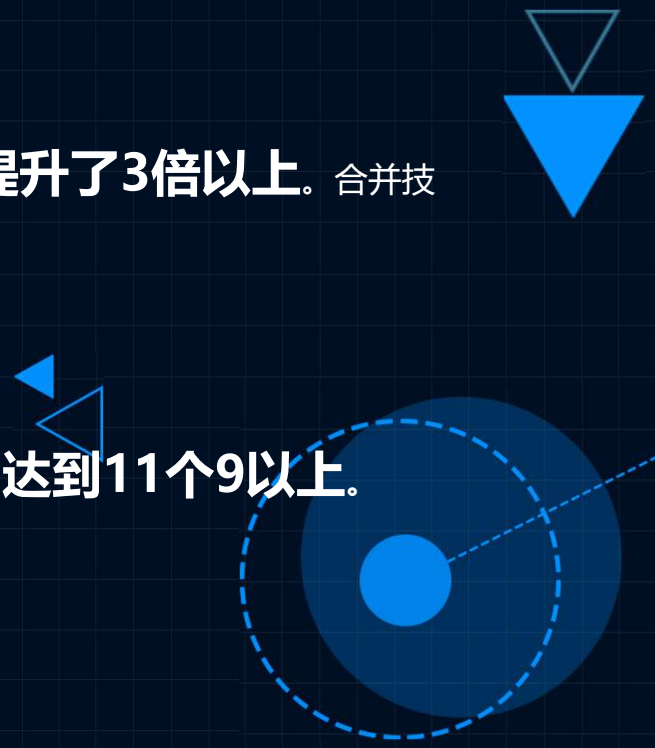
ceph纠删码的引入，给nos的对象存储带来了成本上的降低。结合我们自研的小文件合并方案/数据回收策略等，更好地提升了nos场景下的ceph物理空间利用率。

- 读写性能

nos从根本上改善了读io延时，特别是小文件的io延时，其**读性能提升了3倍以上**。合并技术提升ceph系统的带宽，其带宽接近磁盘的裸盘性能。

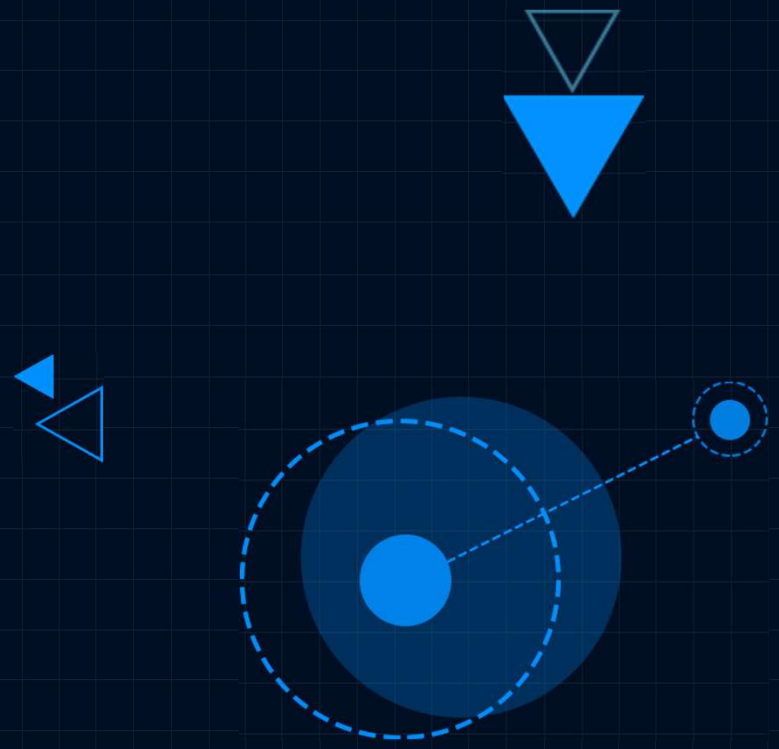
- 数据可靠性

故障磁盘原地恢复，写时数据完备性都从本源上提升了数据的可靠性，**达到11个9以上。**



未来探索

- 更高效的数据合并算法
- 更低成本的存储引擎
- 更高的数据可靠性
- 集群物理资源管控/隔离



Thanks