
Architect
SACC
2022 中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2022
激发架构性能 点亮业务活力
云上会议 网络直播 | 2022年10月27-29日
ITPUB

同程旅行对象存储实践

研发中心-架构师-周祝群

CONTENTS

01. 背景介绍

02. 存储选型与落地

03. 基于S3的其它落地场景

04. 未来展望

Part1:背景介绍

背景介绍

文件

图片

Css/Js等静态资源

视频

容器仓库

网盘

背景介绍

- 图片 20亿+
- Js/Css 千万+
- 视频 100+TB
- 容器仓库 200+TB
- 静态资源 100+TB

ceph

对象存储 COS
Cloud Object Storage

MINIO

阿里云OSS

FastDFS+Redis+Hbase

Part2:系统设计与落地

系统设计与落地-设计目标

- **可扩展**: 至少需要支持到30亿+的对象数,并且需要有水平扩展的能力
- **高可用**: 要做到高可用,至少要有隔离,多租户,限流,灾备/双活等能力,最核心业务甚至可以做到不同存储产品的灾备。
- **高性能**: 需要足够快,类似ceph rados,后续需要可以作为分布式文件存储的存储底座
- **低成本**: 可以用远低于ceph的成本支撑所有的业务
- **接入简单**: 能够支持主流对象存储S3协议的接入
- **无缝升级**: 可以在业务无感知的情况下,稳定的、无缝将业务从ceph s3,公有云 s3迁到新oss

系统设计与落地-设计目标拆解

- 一个强大的对象存储oss服务

- 如何在架构层面解决业务无缝切换,存储无缝升级从而保障业务的稳定性

系统设计与落地-oss服务选型-minio优点

- 友好的UI
- 部署比较简单,很容易上手
- 支持文件级别的自愈,在节点故障时无需人工干预
- 全EC存储,成本相对比较低
- 中大文件性能比较好
- 基于文件系统设计,无需额外的存储来存储元数据

系统设计与落地-oss服务选型-minio劣势

- 仅支持EC
- 扩容不太友好
- 支持的文件数量有限

系统设计与落地-oss服务选型-SeaweedFS优点

- 性能比较强大:
- 架构设计比较灵活:
- 功能齐全: 存储比较关心的冷热分离,EC存储,TTL等功能
- 部署简单: 部署非常简单,很容易上手

系统设计与落地-oss服务选型-SeaweedFS劣势

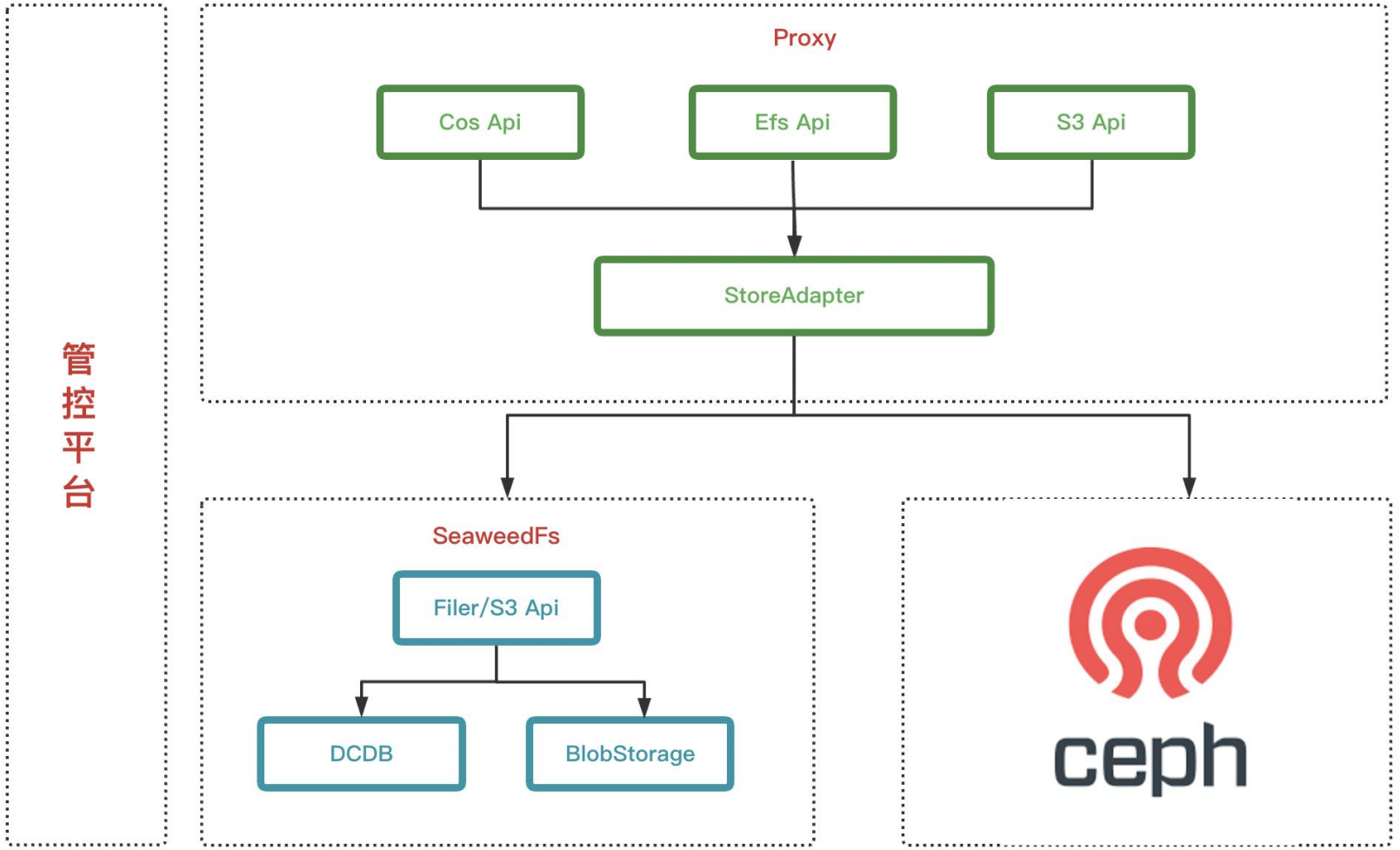
- **S3的适配不完全: 如object acl等**

系统设计与落地-oss服务选型

新OSS底层基于SeaweedFS来做共建

系统设计与落地-架构设计

- Proxy层
- 存储层
- 管控平台

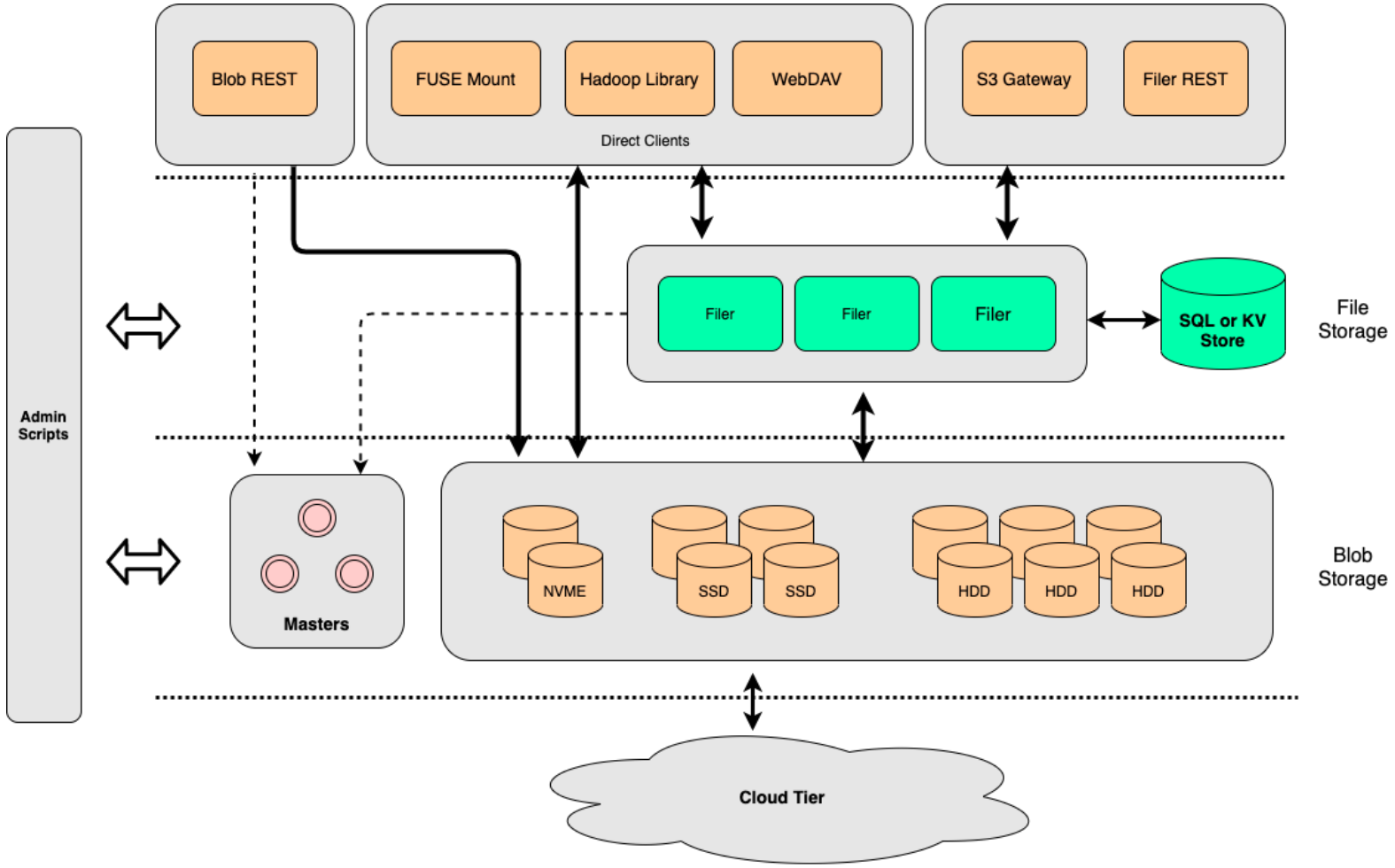


系统设计与落地-可扩展性

Proxy :无状态支持水平扩展

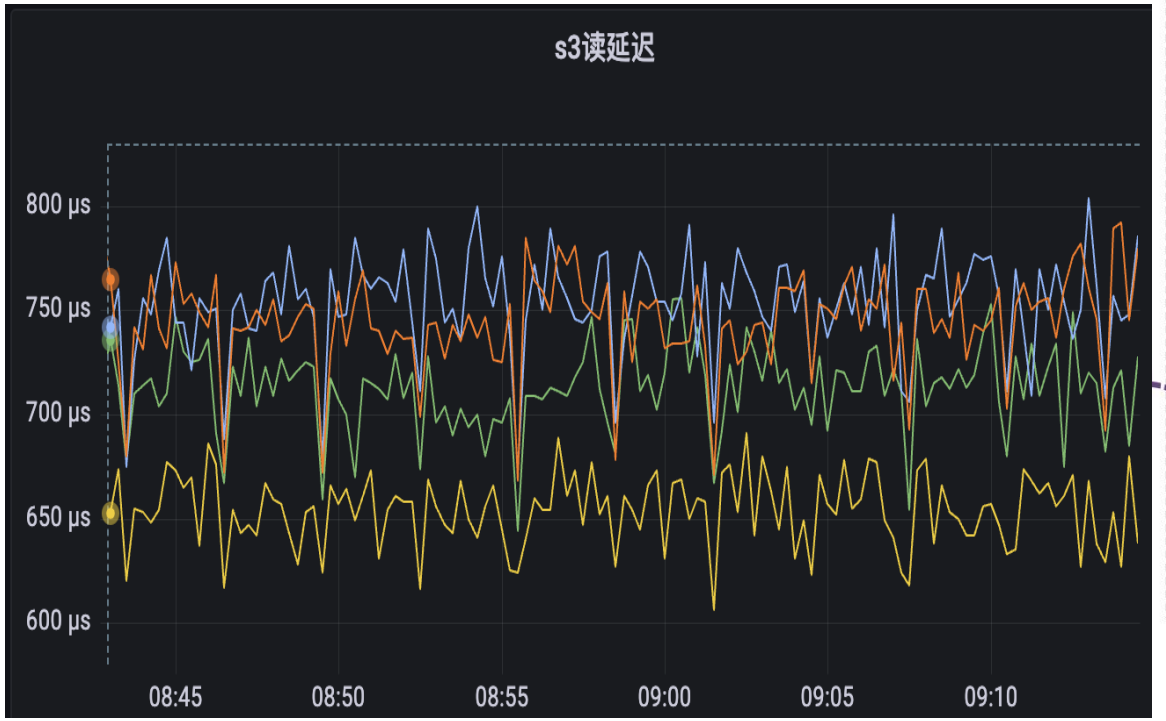
File Storage：用来做metadata元信息的存储以及做api的适配

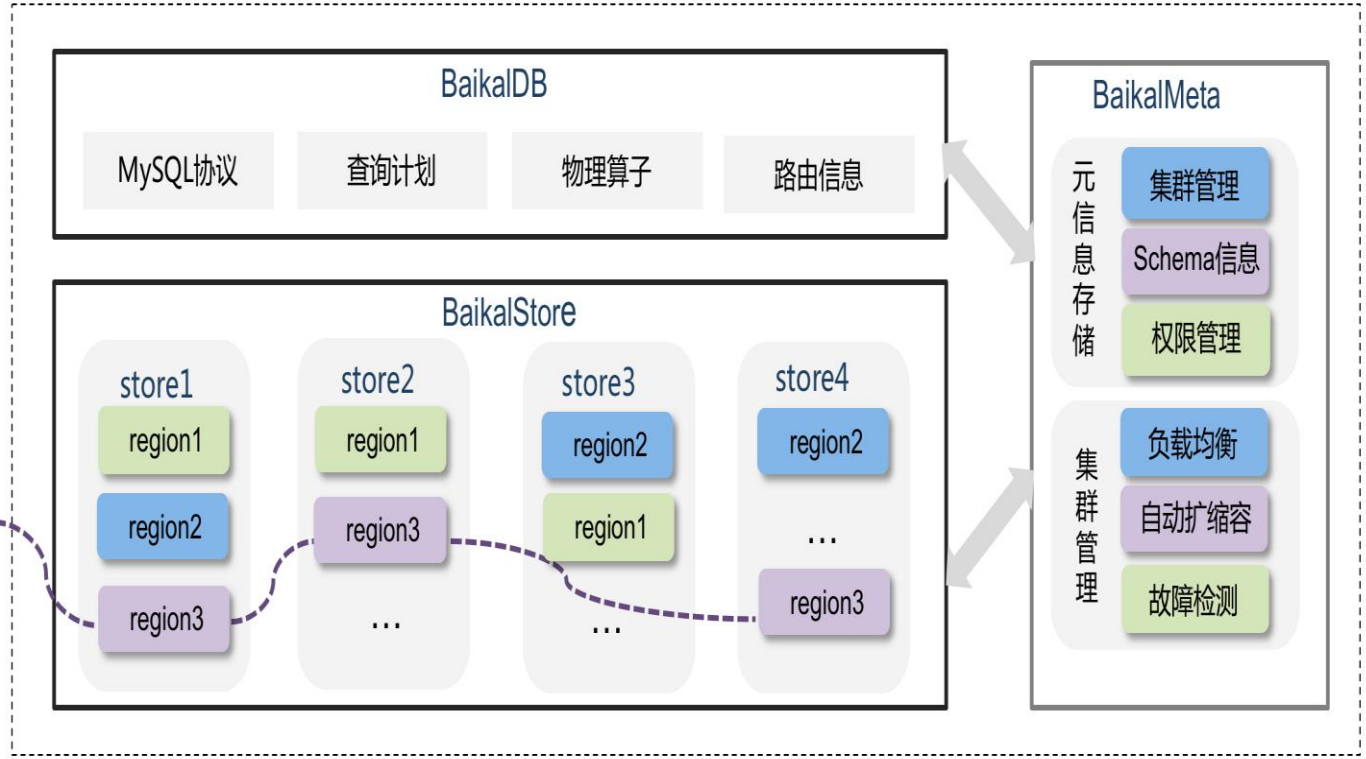
Blob Storage：对象存储的底座



系统设计与落地-可扩展性

MetaData Storage:BaikalDB



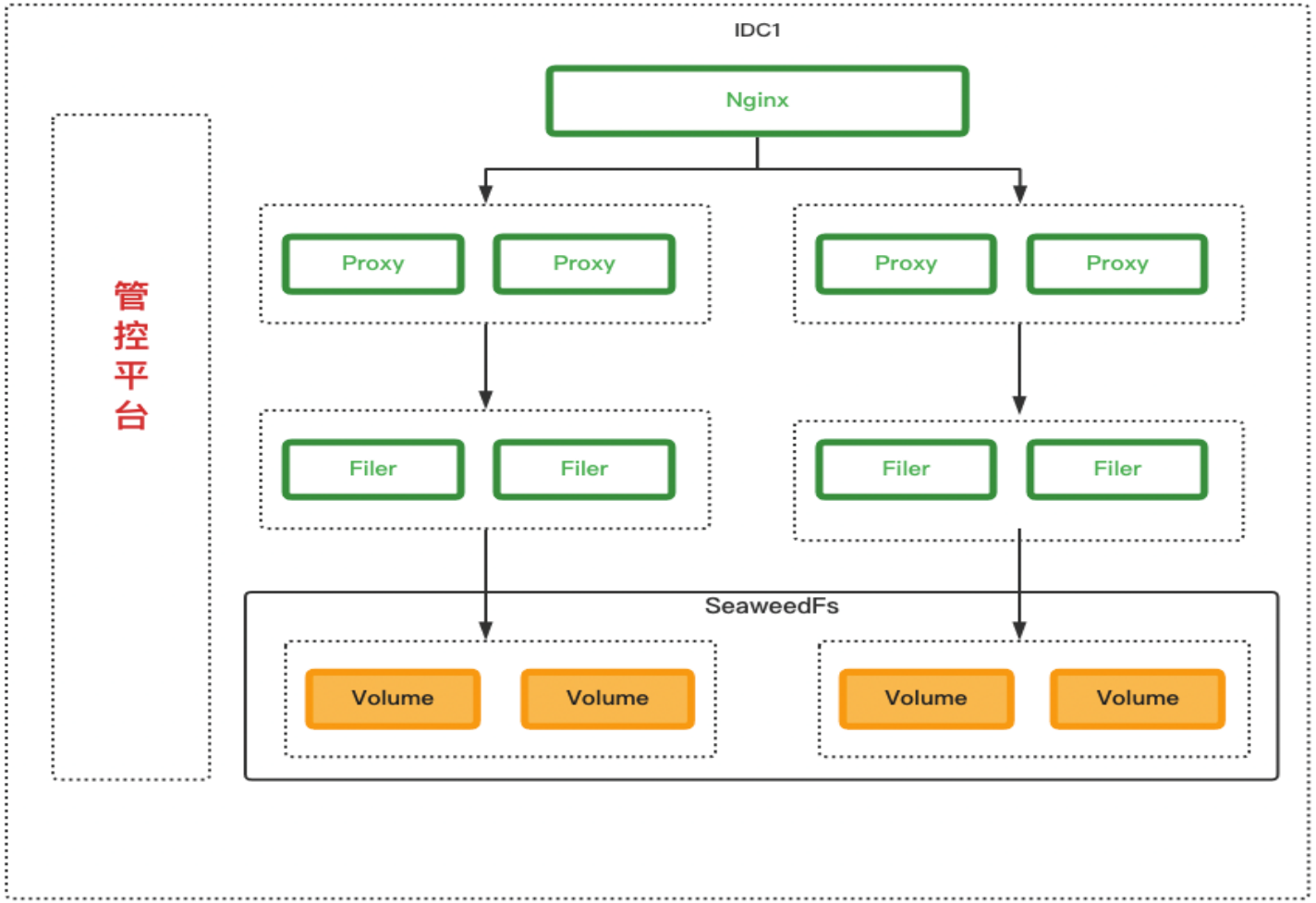


系统设计与落地-高可用

隔离

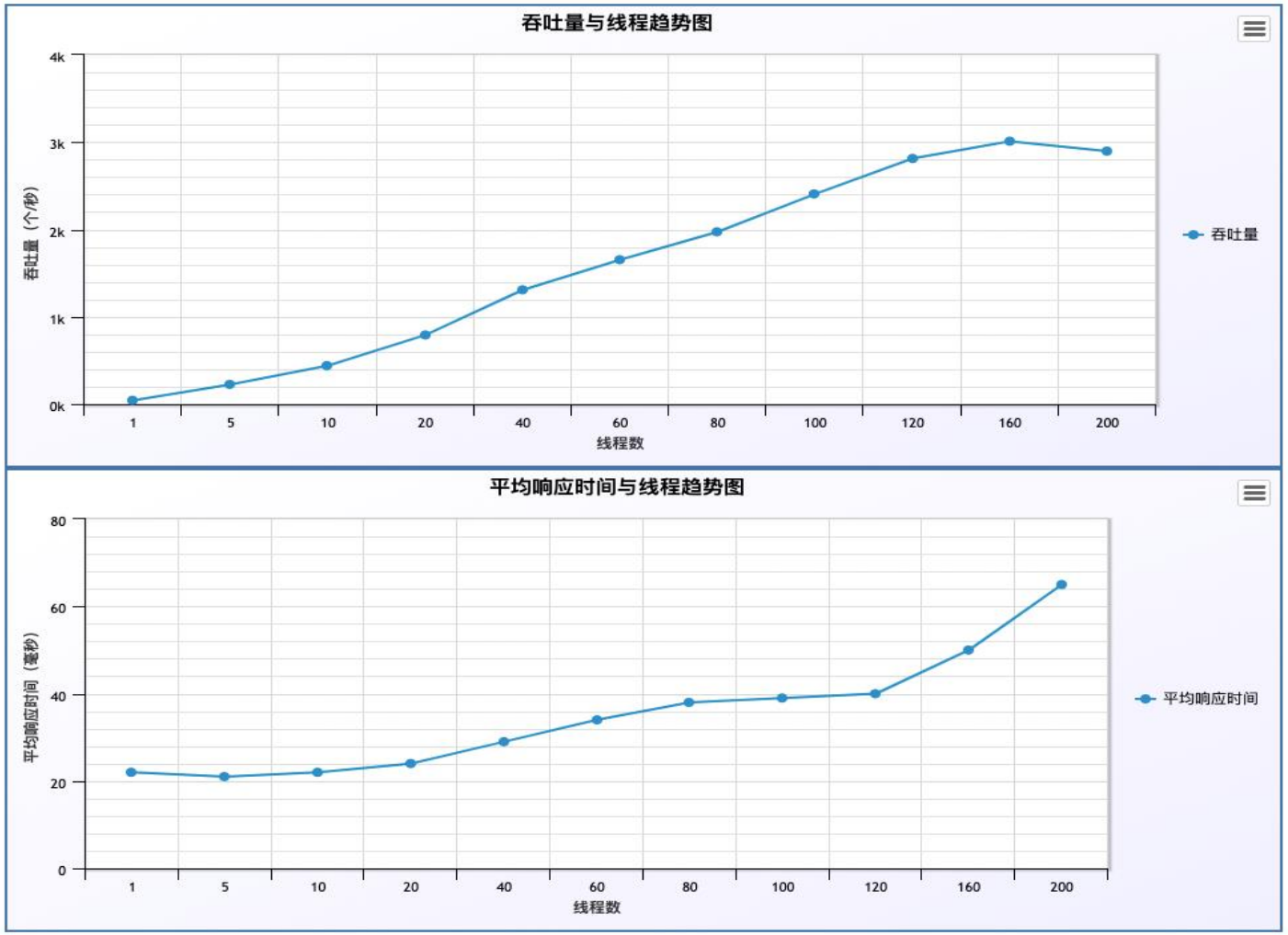
多租户与限流

全组件高可用



系统设计与落地-高性能

主页 / 历史项目 / 20220613S3Proxy / SuiteTest1-S3Dss-1MB



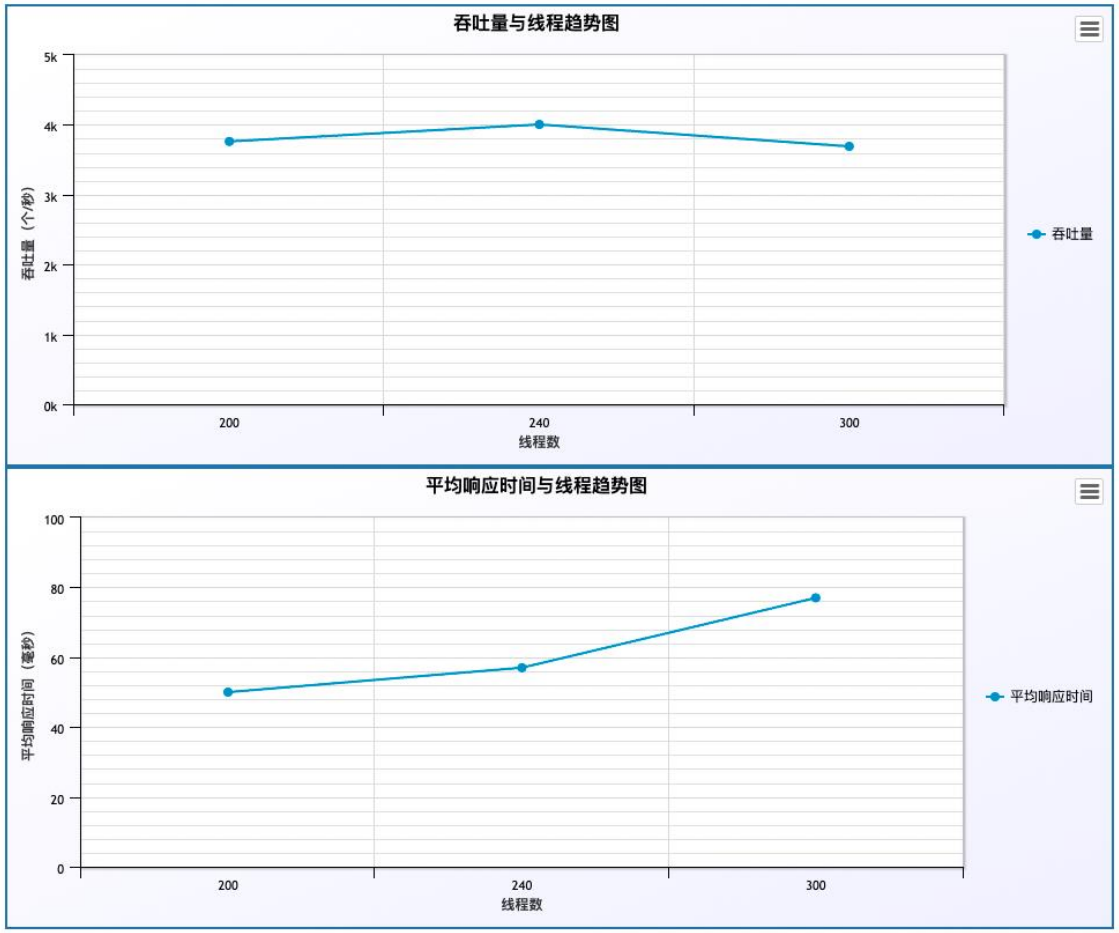
6台256G,12*8TB SATA盘,2副本
压测的是1M的写

IO 优先到了瓶颈

系统设计与落地-高性能

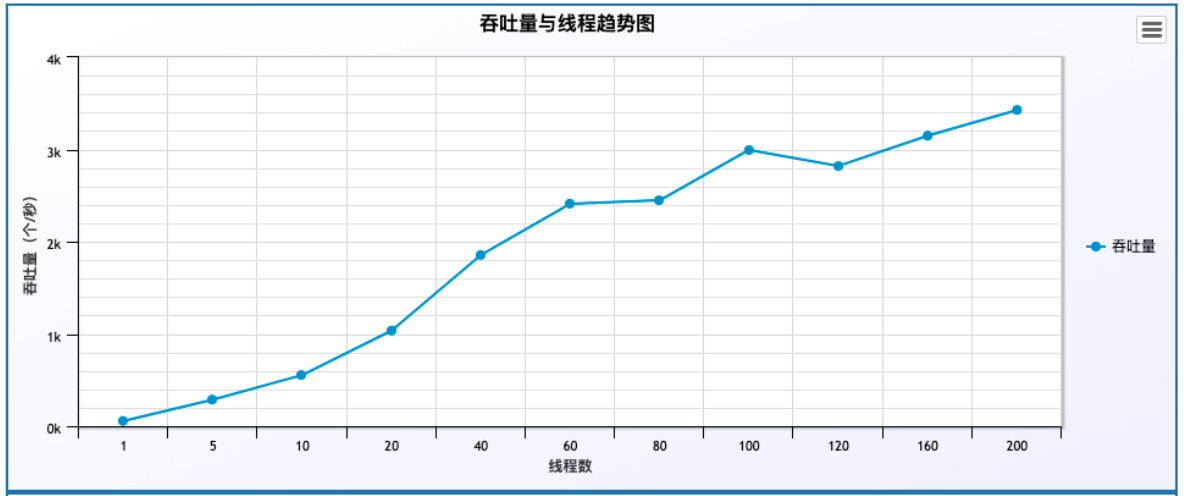
6台256G,10TB SSD盘,2副本,压测的是1M的写

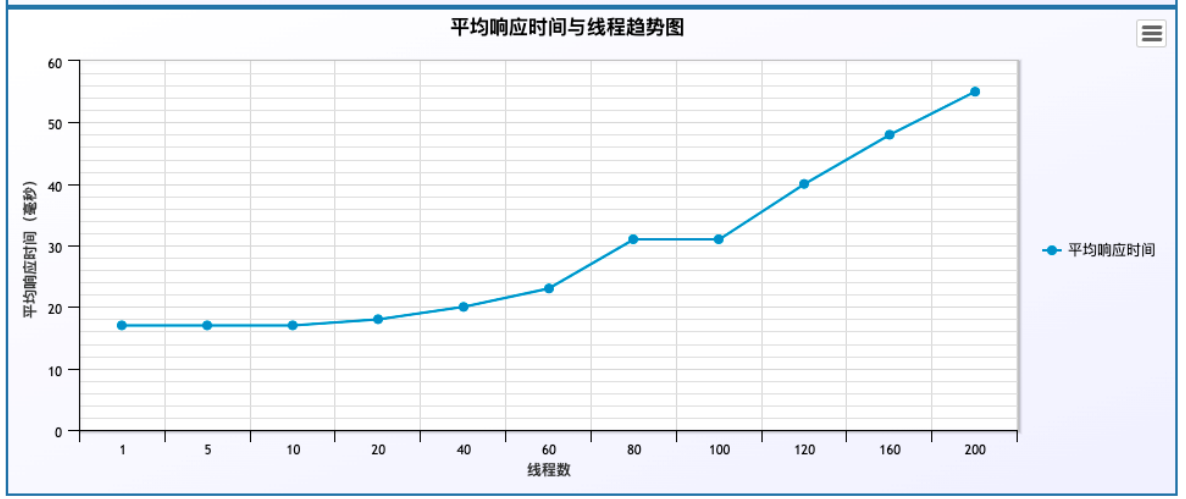
出于稳定性考虑,没有继续压测,各项指标远没到瓶



系统设计与落地-高性能

主页 / 历史项目 / 20220613S3Proxy / SuiteTest2-S3Dss-1MB





6台256G,SATA盘,2副本

写入,查询,删除混合场景(6:3:1)压测下,1M的文件

系统设计与落地-低成本

存储类型

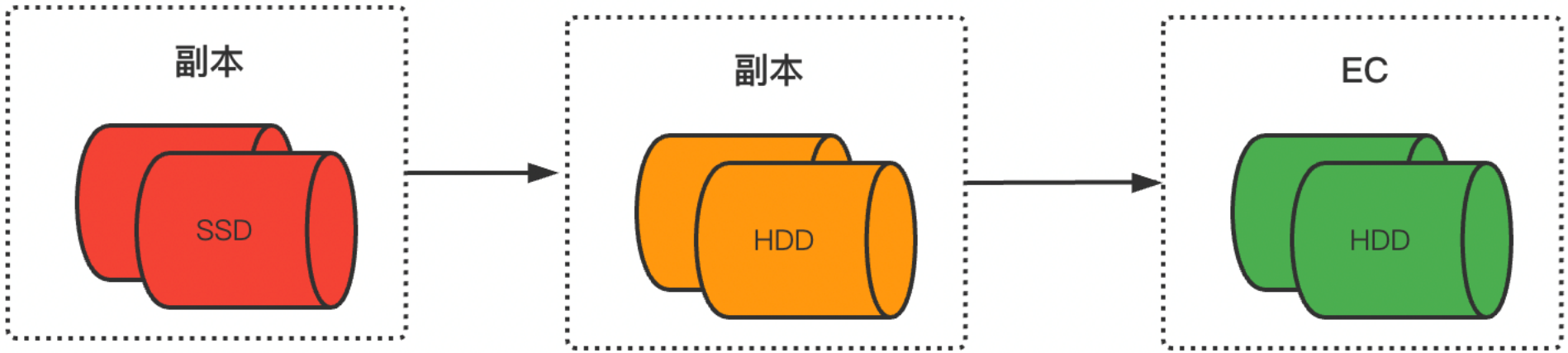
	标准类型	低频访问类型	归档类型	冷归档类型
适用场景	单文件每月访问大于1次	单文件月访问不到1次	单文件90天访问不到1次	单文件年访问不到1次
对象最小计量大小	按照对象实际大小计算	64KB，即小于64KB的文件 按64KB计费	64KB，即小于64KB的文件 按64KB计费	64KB，即小于64KB的文件 按64KB计费
最少存储时间要求	无	30天	60天	180天
数据访问特点	实时访问	实时访问	解冻后才能读取 解冻时间1分钟	解冻后才能读取 解冻时间1~12小时可选
图片处理	支持	支持	支持，但需要先解冻	支持，但需要先解冻
数据取回费用	无	按实际获取的数据量收取 单位GB	按实际解冻的数据量收取 单位GB	按实际解冻的数据量以及选择的 数据解冻时间收取，单位GB

系统设计与落地-低成本

副本 VS EC

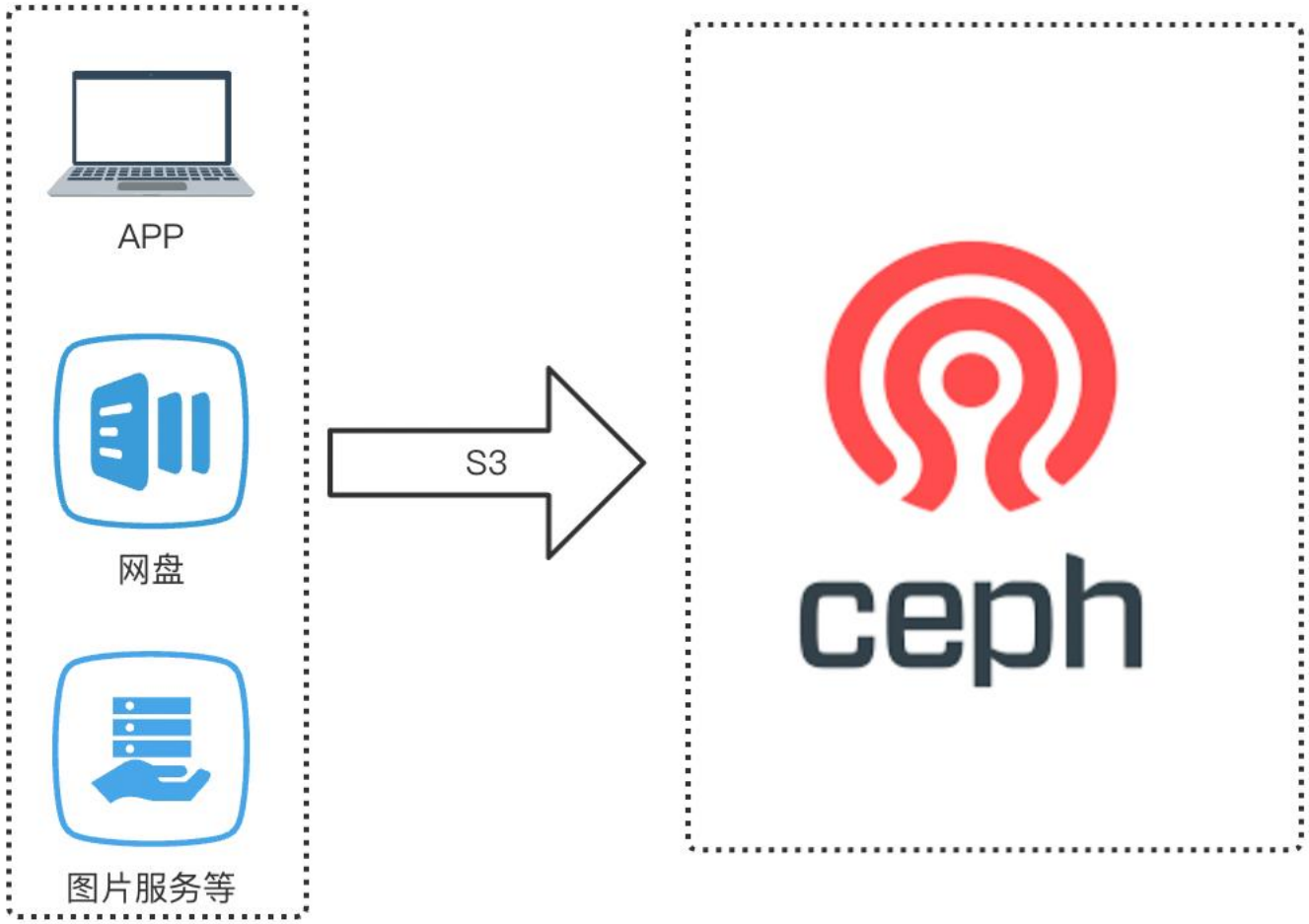
比较项	多副本(N)	EC(M+N)
可用容量	1/N,较低	M/(M+N),较高
读写性能	较高	较低,小块io更明显
重构性能	无校验计算,较快	有校验计算,较慢
容忍节点故障数量	N-1	N
适用场景	小文件场景	大文件场景

系统设计与落地-低成本



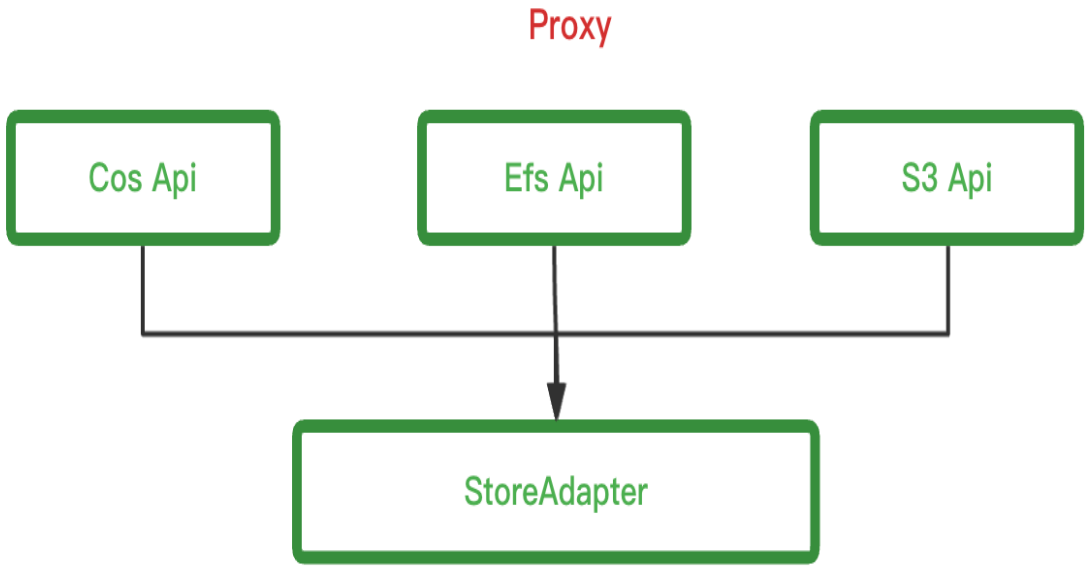
现在我们可选的存储介质包括NVME SSD,SATA SSD,HDD。可选的数据存储方式有多副本,EC

系统设计与落地-无缝升级-当前架构



系统设计与落地-无缝升级-流程适配

全api的适配

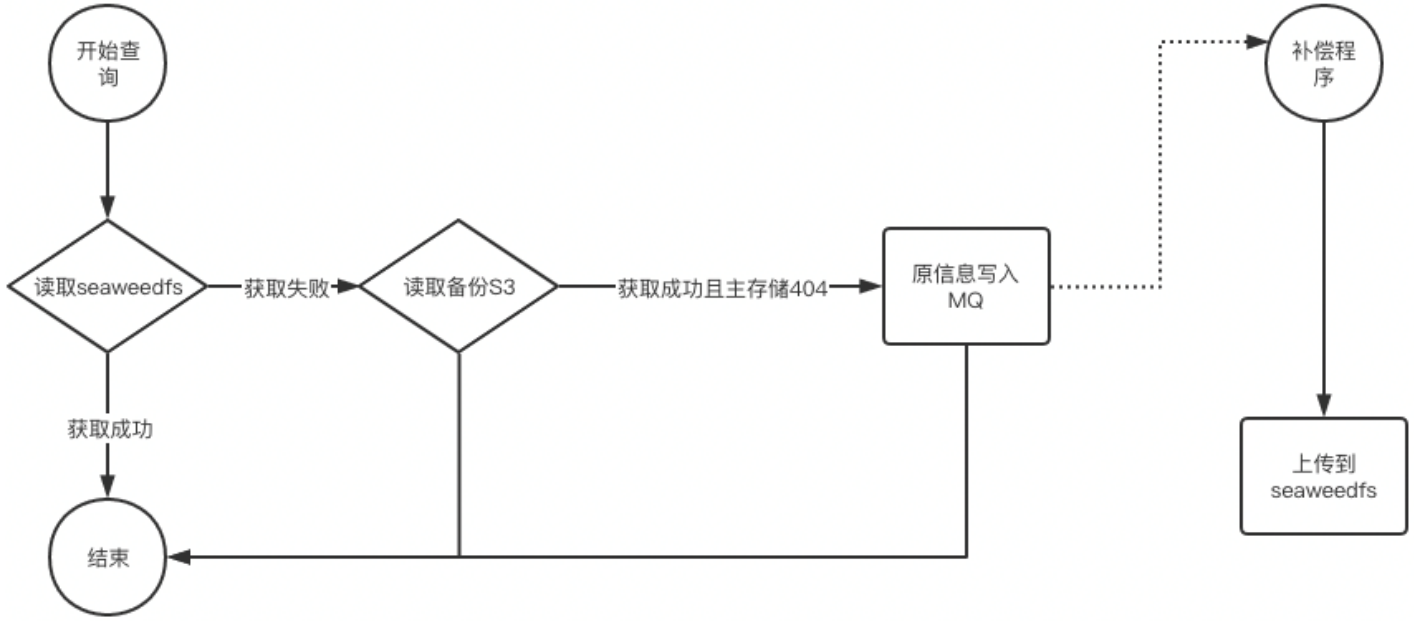


系统设计与落地-无缝升级-读流程适配

无需数据全部导入

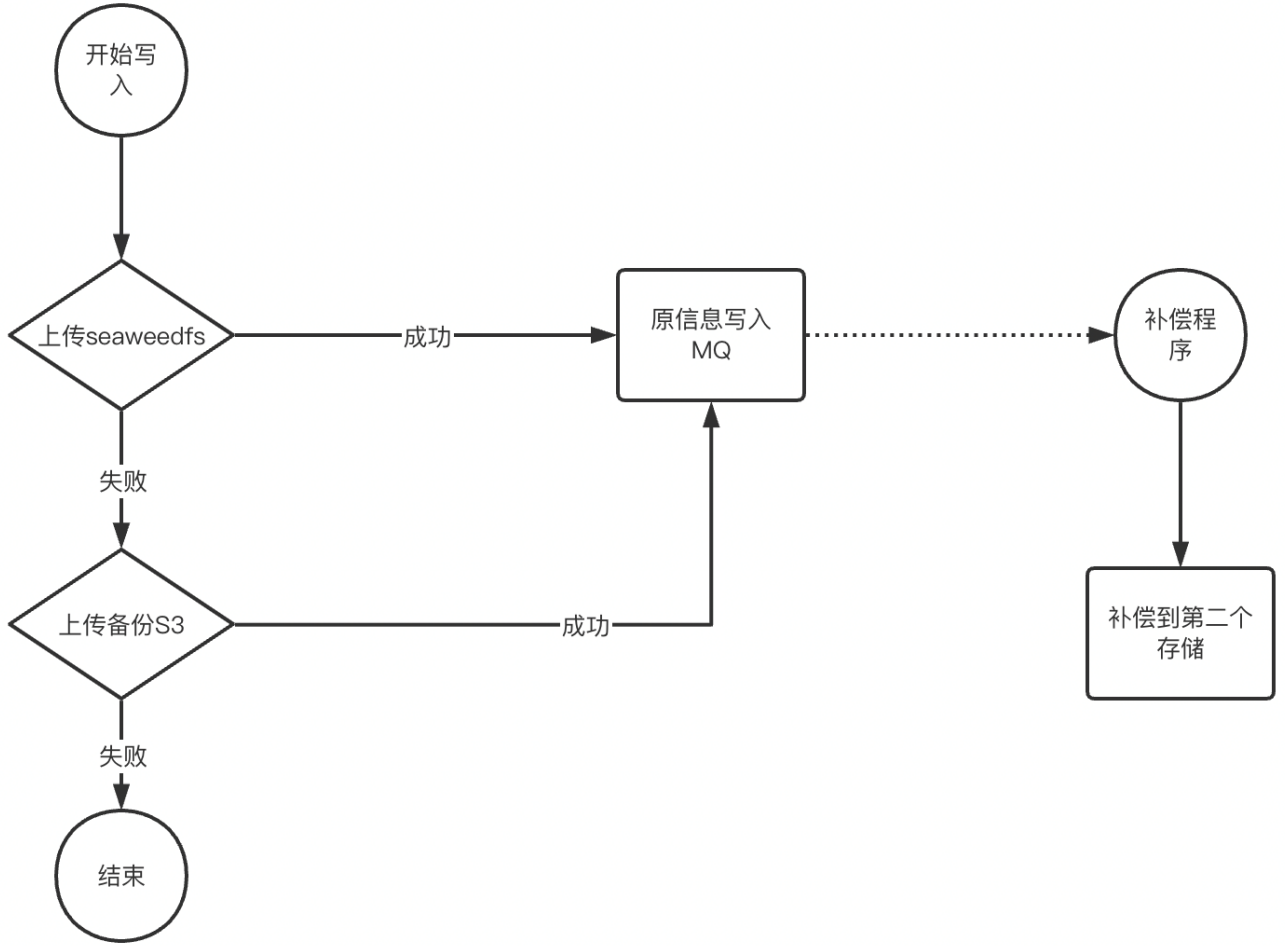
存储级别的高可用

自动数据补齐



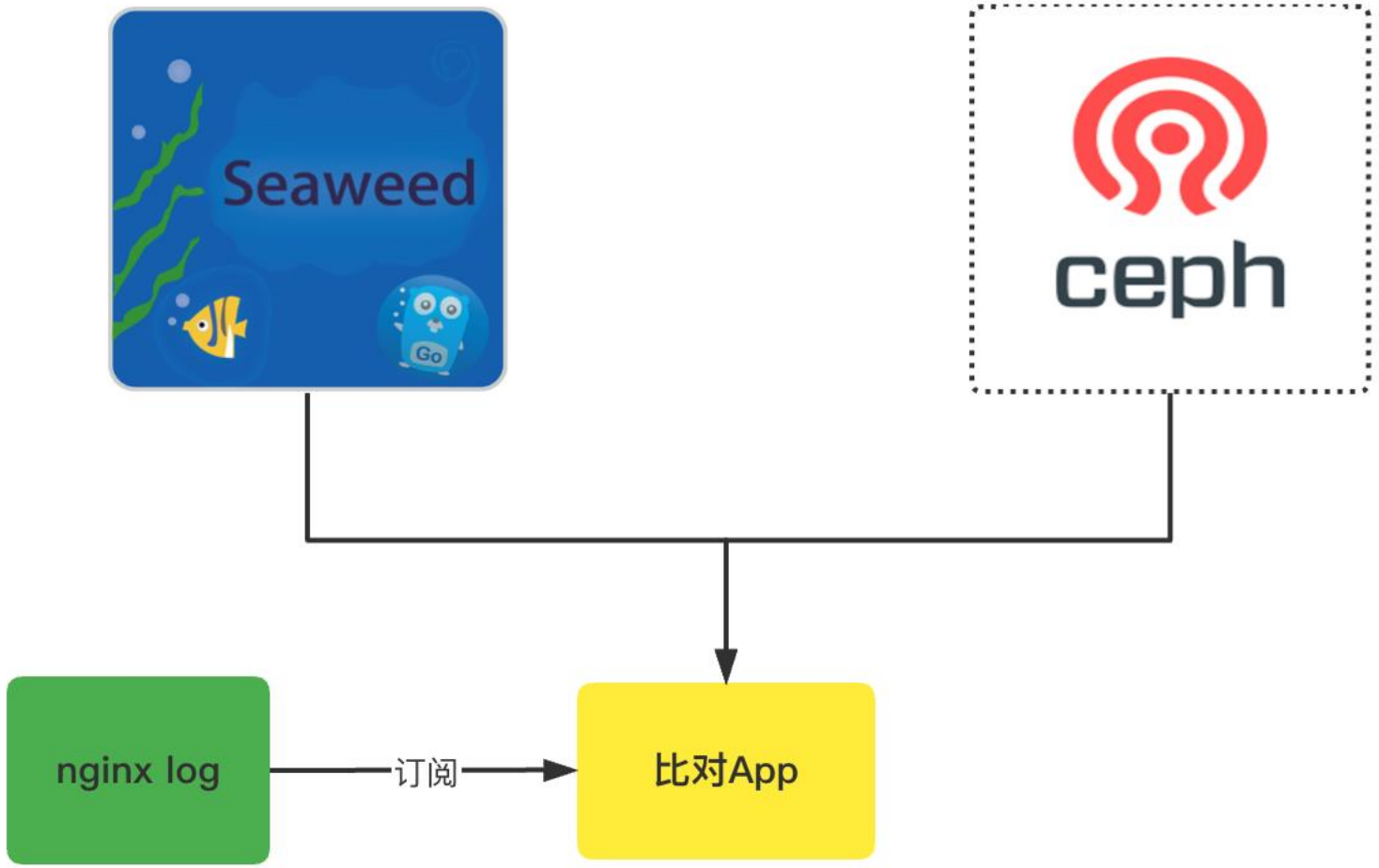
系统设计与落地-无缝升级-写流程适配

写入一个就写入成功, 提升性能



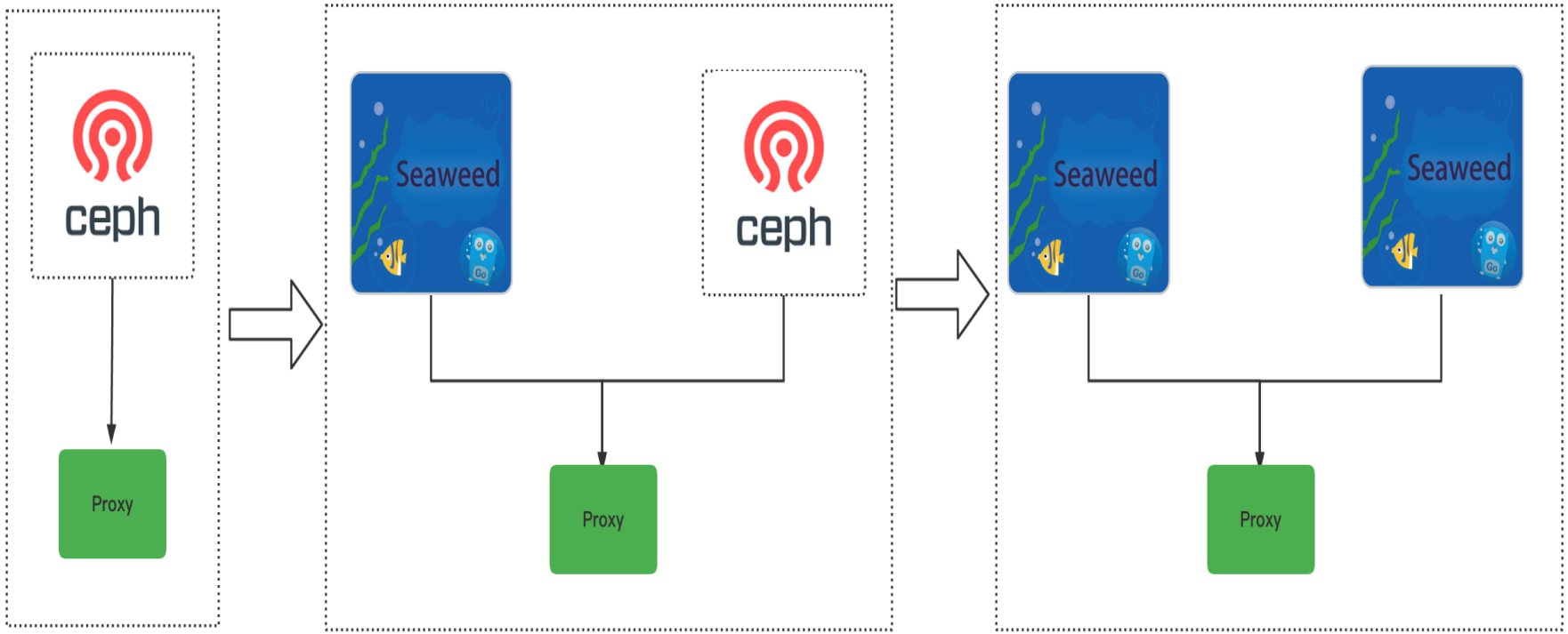
系统设计与落地-无缝升级-升级步骤

数据比对



系统设计与落地-无缝升级-升级步骤

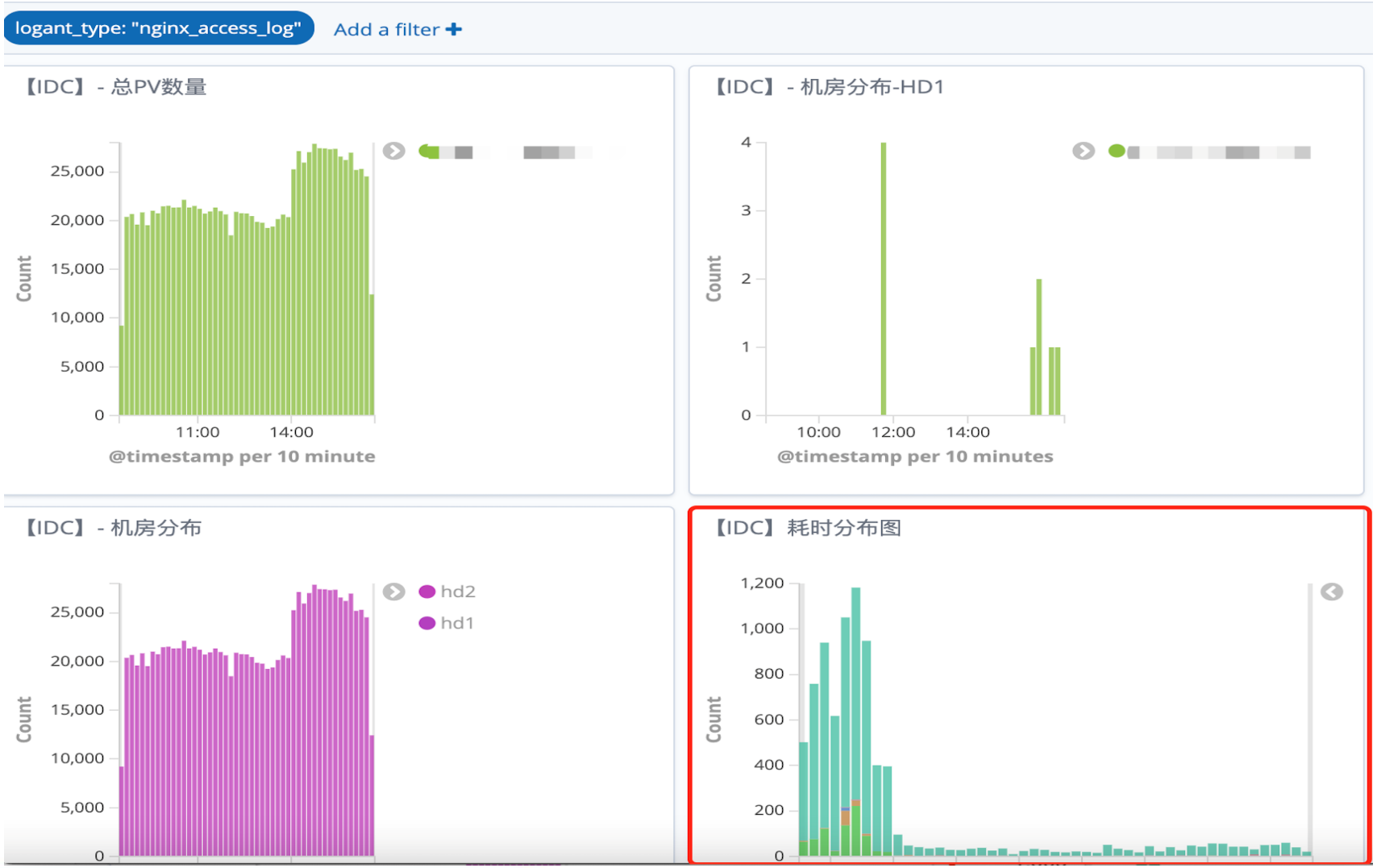
过程



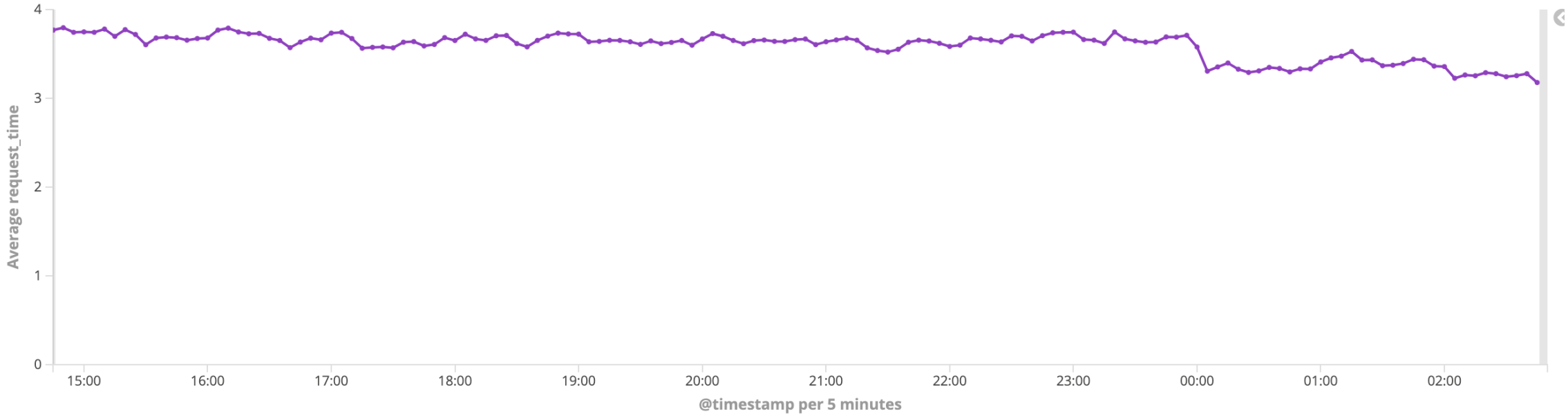
系统设计与落地-无缝升级-额外功能

基于proxy的无缝去重

系统设计与落地-收益


logant_type: "nginx_access_log"
Add a filter +
【IDC】 - 总PV数量
Count
25,000
20,000
15,000
10,000
5,000
0
11:00
14:00
@timestamp per 10 minute
【IDC】 - 机房分布-HD1
Count
4
3
2
1
0
10:00
12:00
14:00
@timestamp per 10 minutes
【IDC】 - 机房分布
hd2
hd1
Count
25,000
20,000
15,000
10,000
5,000
0
【IDC】 耗时分布图
Count
1,200
1,000
800
600
400
200
0

系统设计与落地-收益


4
3
2
1
0
Average request_time
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
00:00
01:00
02:00
@timestamp per 5 minutes

系统设计与落地-使用tips

volumeGrowthCount

fs.configure

filer.sync

Part3:基于S3的其它存储落地

落地场景-clickhouse

TTL

落地场景-prometheus

TTL

Part4:未来展望

未来展望

分布式文件存储

QA