使用Ceph构建高效的分布式存储平台

王 刚

平安科技

目录

企业存储应用领域遇到的难题

海量数据存储管理难题--种类繁多

EMC²

NetApp™

DELL

DataDirect NETWORKS

IBM

HITACHI

hp

HUAWEI

浪潮 inspur

FUJITSU

海量数据存储管理难题--种类繁多（续）

- 架构复杂
- 过保迁移困难
- 支持响应慢
- 个性化需求难实现
- 厂商放弃维护
- 专有硬件

块存储-应用现状

单机块存储

硬盘是一个块设备，内核检测到硬盘然后在/dev/下会看到/dev/sda/。

LVM & Device-mapper

LVM是一种逻辑卷管理器

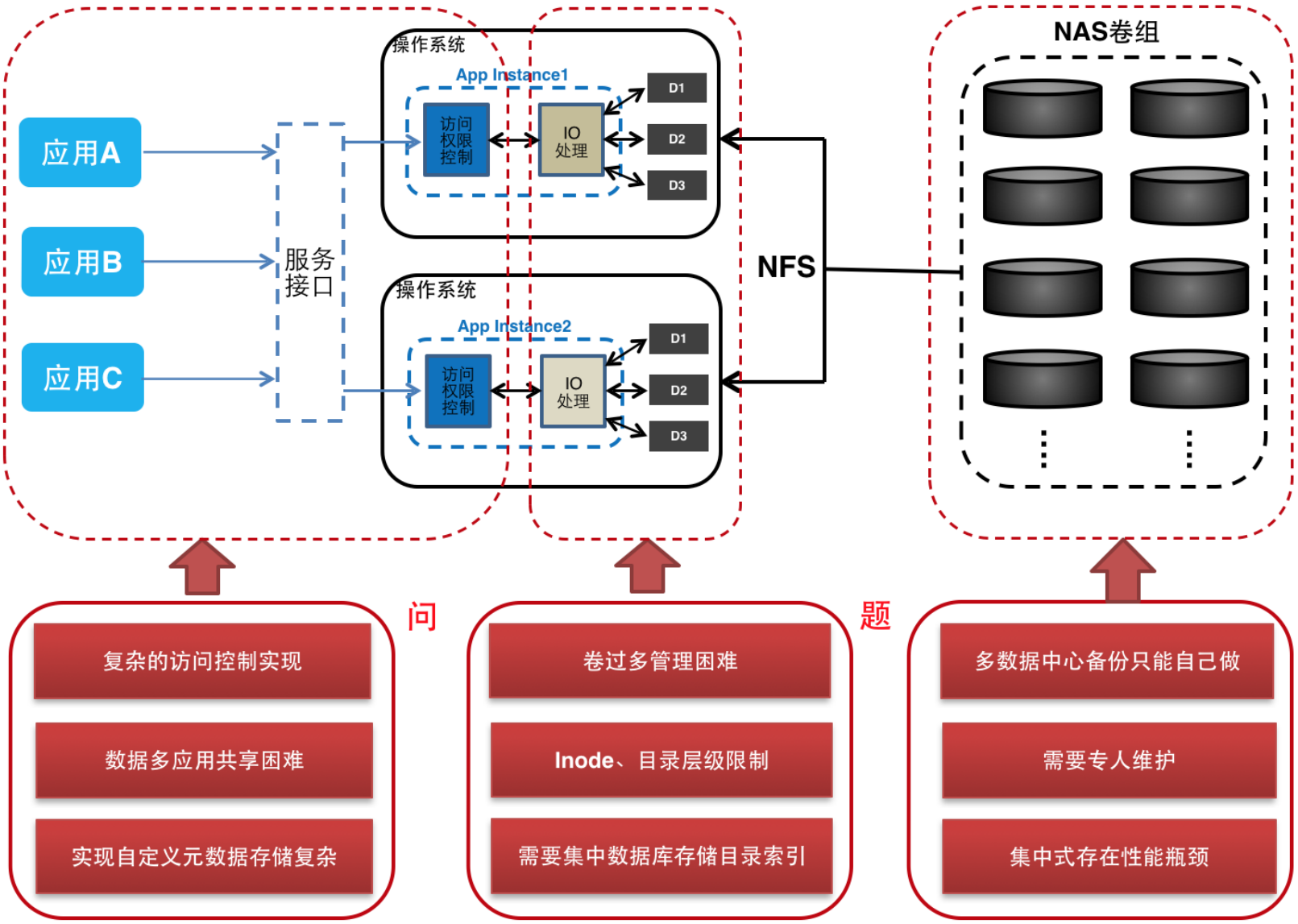
Device-mapper是一种支持逻辑卷管理的通用设备映射机制

SAN & iSCSI

大部分SAN使用SCSI协议在服务器和存储设备之间传输和沟通。

常见的有iSCSI，FC等。

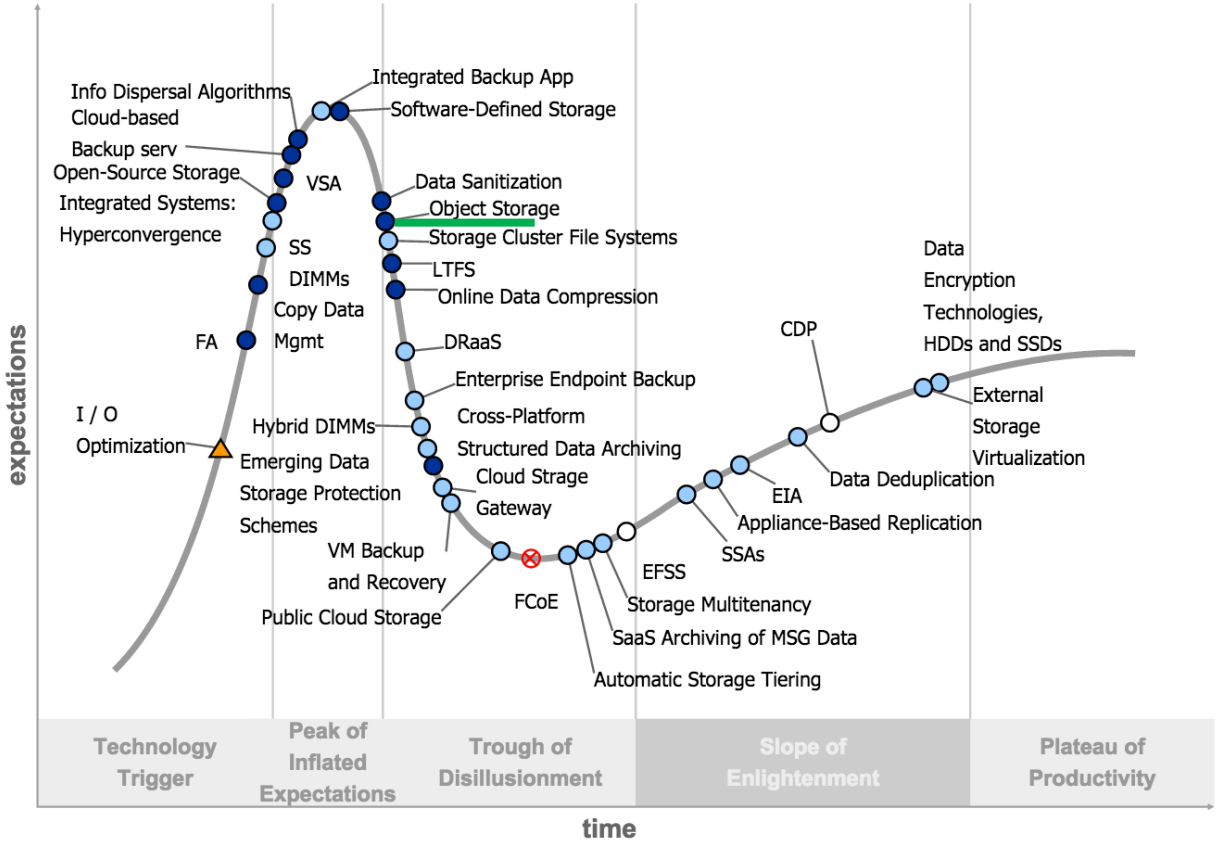
文件系统-应用难题（NAS举例）



对象存储-发展趋势

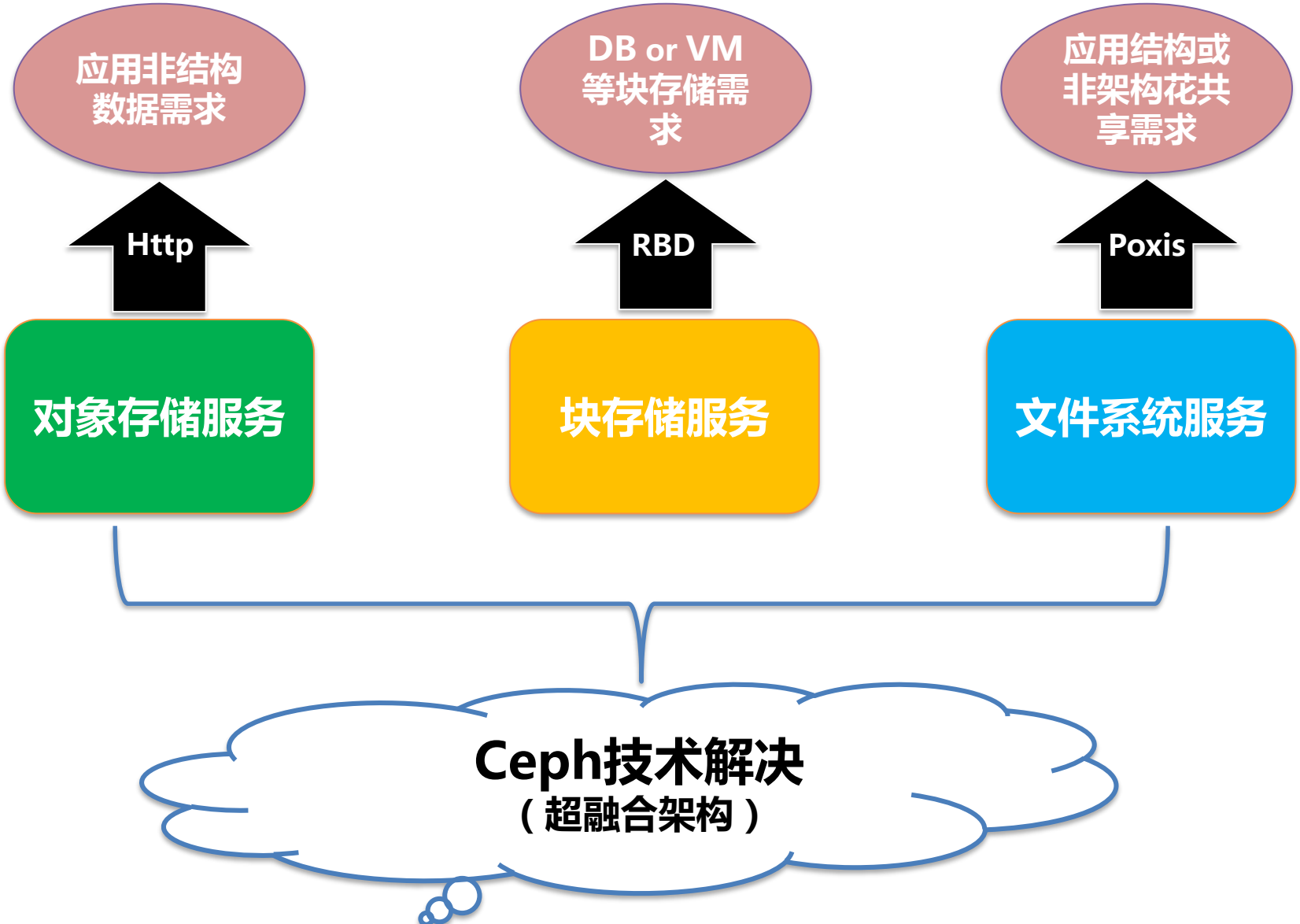
- 非结构化数据年平均增长率(60%~80%)带来的成本和管理复杂性问题；
- 共享存储资源要求多租户和以对象为粒度的权限控制，在传统存储中难于满足；
- 以程序直接访问存储和基于对象的权限控制、描述信息提供简化应用程序开发，为快速应用开发和存储自动管理、自我修复提供基础；
- 对象存储的主要应用方向，为其他新兴存储技术和解决方案提供后台非结构化存储方案;

Gartner 《Hype Cycle for Storage Technologies, 2015》

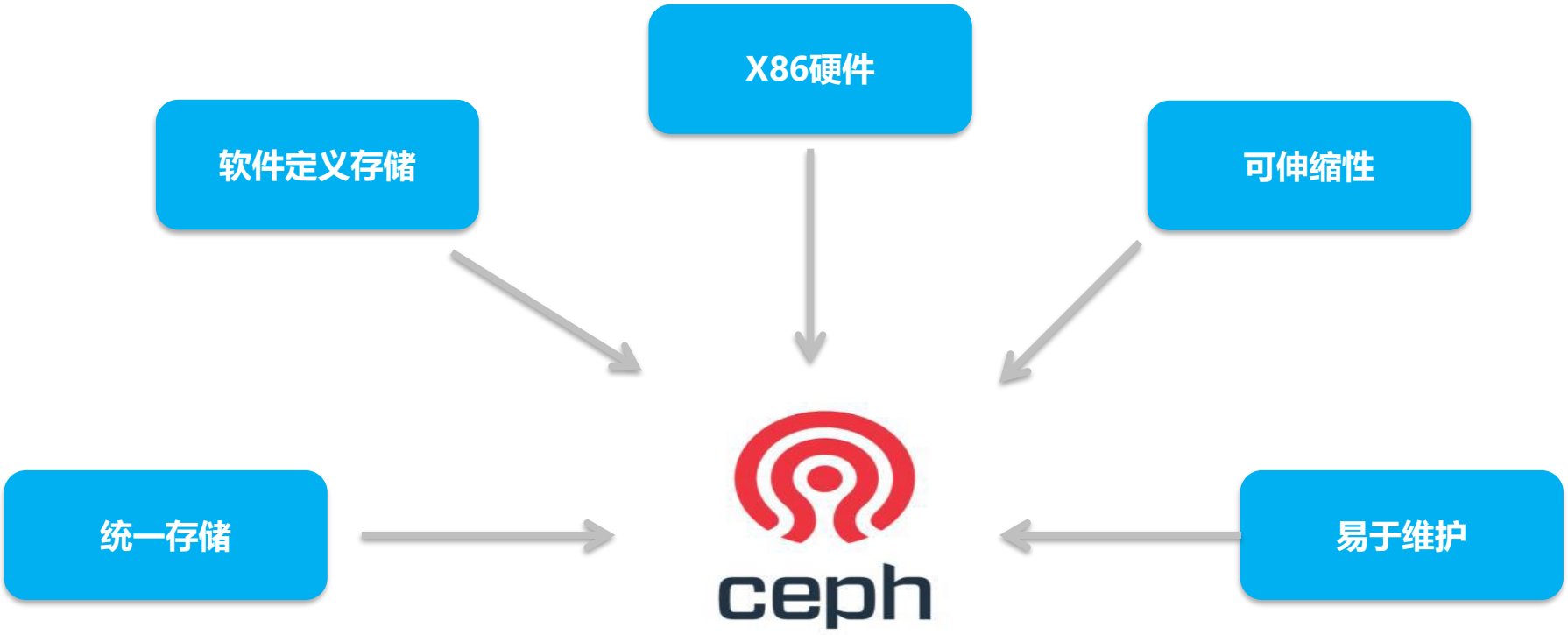


为什么选择Ceph做为存储核心架构

超融合存储架构


应用非结构数据需求
DB or VM等块存储需求
应用结构或非架构花共享需求
Http
RBD
Poxis
对象存储服务
块存储服务
文件系统服务
Ceph技术解决
（超融合架构）

为什么是Ceph



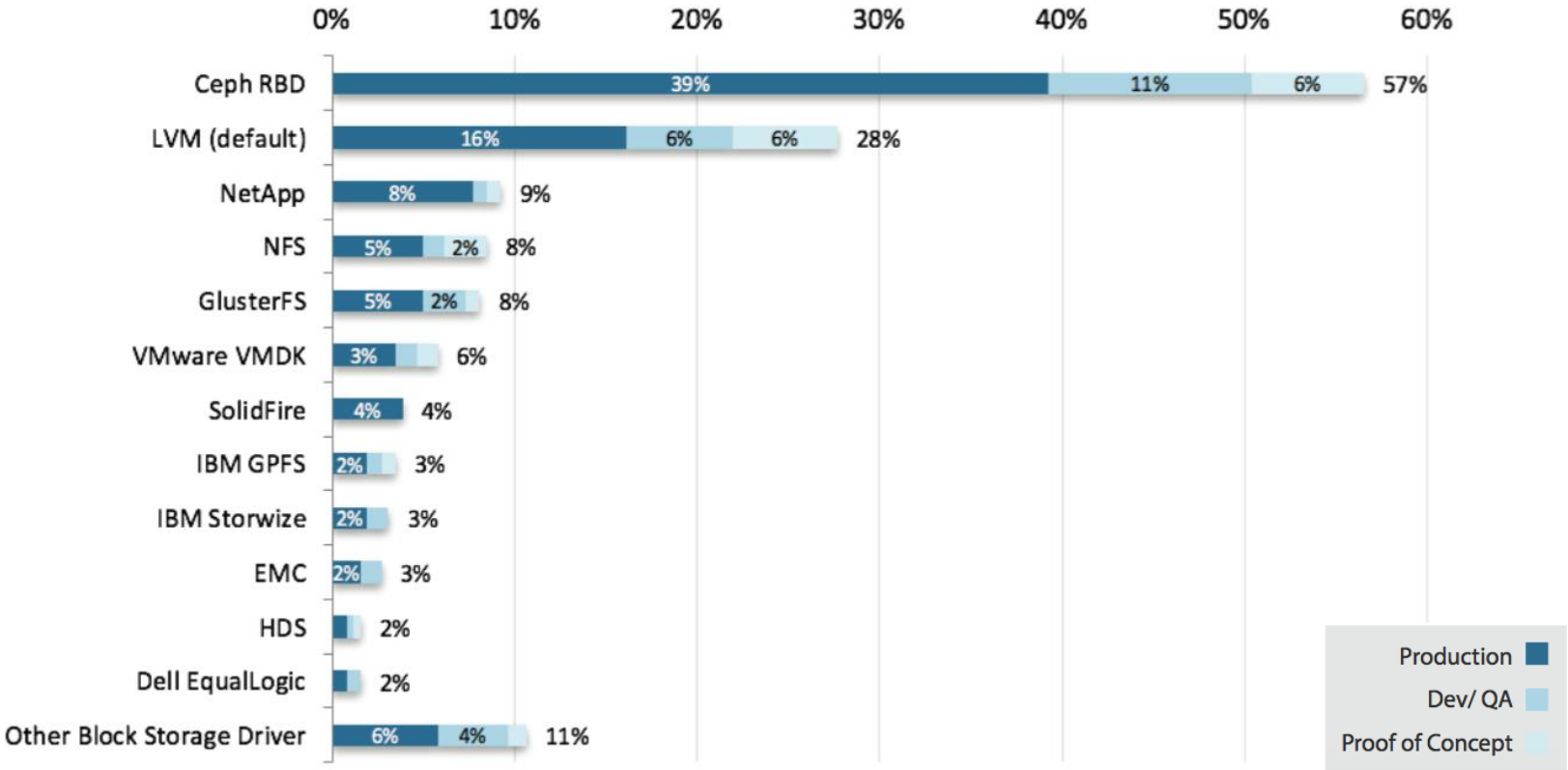
块存储领域-OpenStack2016技术白皮书

Which OpenStack Block Storage (Cinder) drivers are in use?

Ceph RBD continues to dominate Cinder drivers, though its share declined 5 points while second-place LVM (default) increased 6 points.

NetApp lost 3 points, EMC and NFS lost 2, and Gluster FS and Dell EqualLogic were down 1.

The portion of users indicating other storage drivers rose markedly from 7% to 11%, with users writing in DRDB, Dell Storage Center, ZFS, Fujitsu Ethernus, HPE MSA, and Quobyte.



对象存储领域-应用情况

苏州研发中心-对象存储开发服务项目_比选采购_变更公告_(2)

本采购项目为苏州研发中心-对象存储开发服务项目，比选编号:CMEETC-167XL313KK55已立项，采购人为中移（苏州）软件技术有限公司，采购代理机构为中国机电工程招标有限公司。项目具备采购条件，现进行公开比选，具有服务能力的供应商均可前来报名。

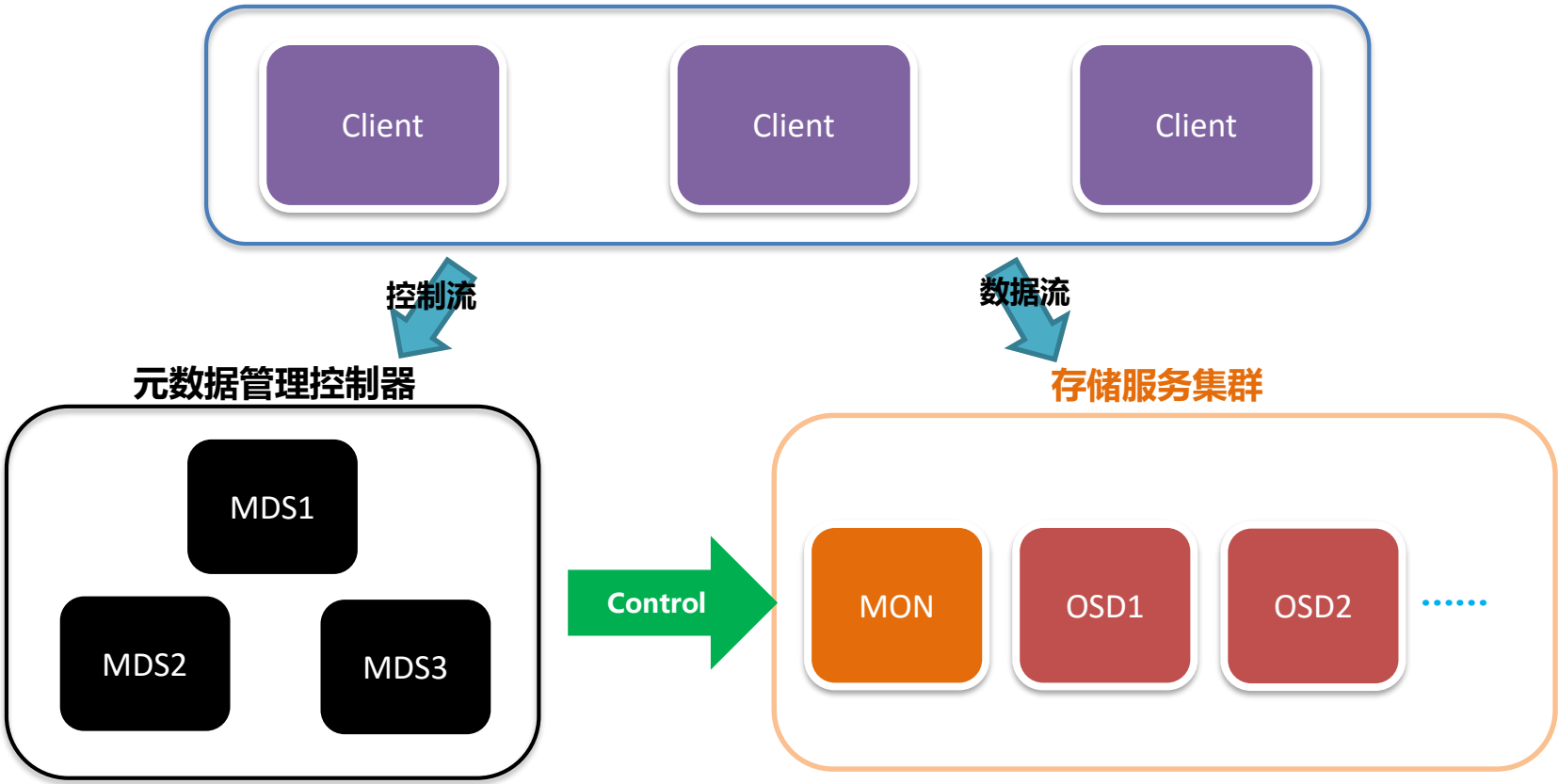
一、项目概况与采购范围

1.1项目名称：苏州研发中心-对象存储开发服务项目。

1.2项目内容： 对象存储系统架构调整为CEPH架构，此次采购为合作开发新架构对象存储系统、共同进行系统部署实施以及技术支持服务。

合作周期：合同签订之日起两年。预估金额：1000万元。

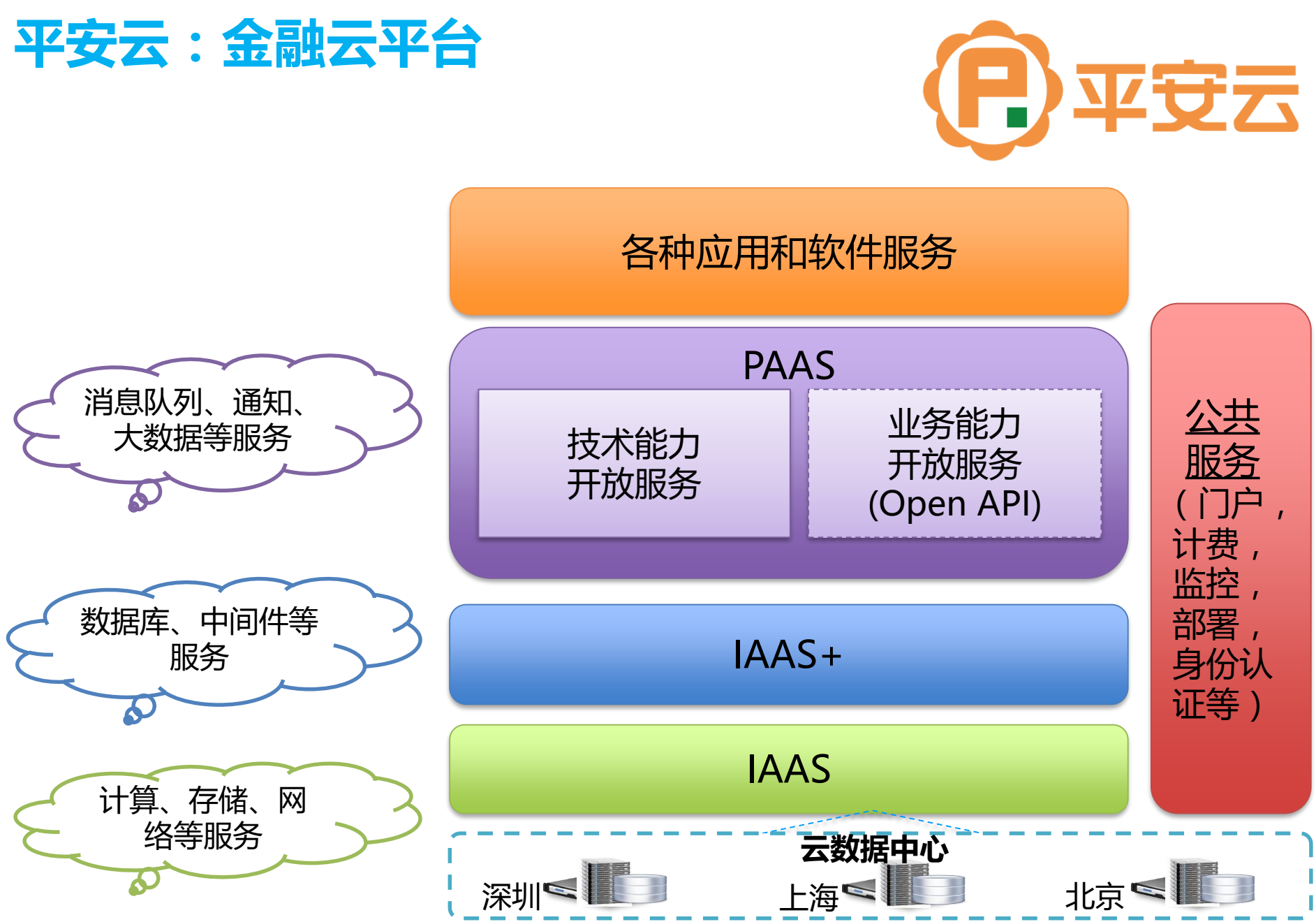
文件系统领域-技术分析


Client
Client
Client
控制流
数据流
元数据管理控制器
存储服务集群
MDS1
MDS2
MDS3
Control
MON
OSD1
OSD2
.....


多个MDS的数据一致性
元数据修复辅助工具
只适合大文件存储
稳定性验证

Ceph大规模实践运行情况报告

平安云：金融云平台


各种应用和软件服务
PAAS
技术能力
开放服务
业务能力
开放服务
(Open API)
IAAS+
IAAS
公共
服务
（门户，
计费，
监控，
部署，
身份认
证等）
消息队列、通知、
大数据等服务
数据库、中间件等
服务
计算、存储、网
络等服务
云数据中心
深圳
上海
北京

问题现状-非结构化数据挑战

每天业务量

上传 下载 流量 增量




（1000万+） （2000万+） （100MB+/S） （1TB+）

业务量

深圳数据中心

NAS卷 1000+

PB级数据容量

DATA DATA

容量

单证 图片 文档


容量

上海数据中心

NAS卷 1000+

PB级数据容量

DATA DATA

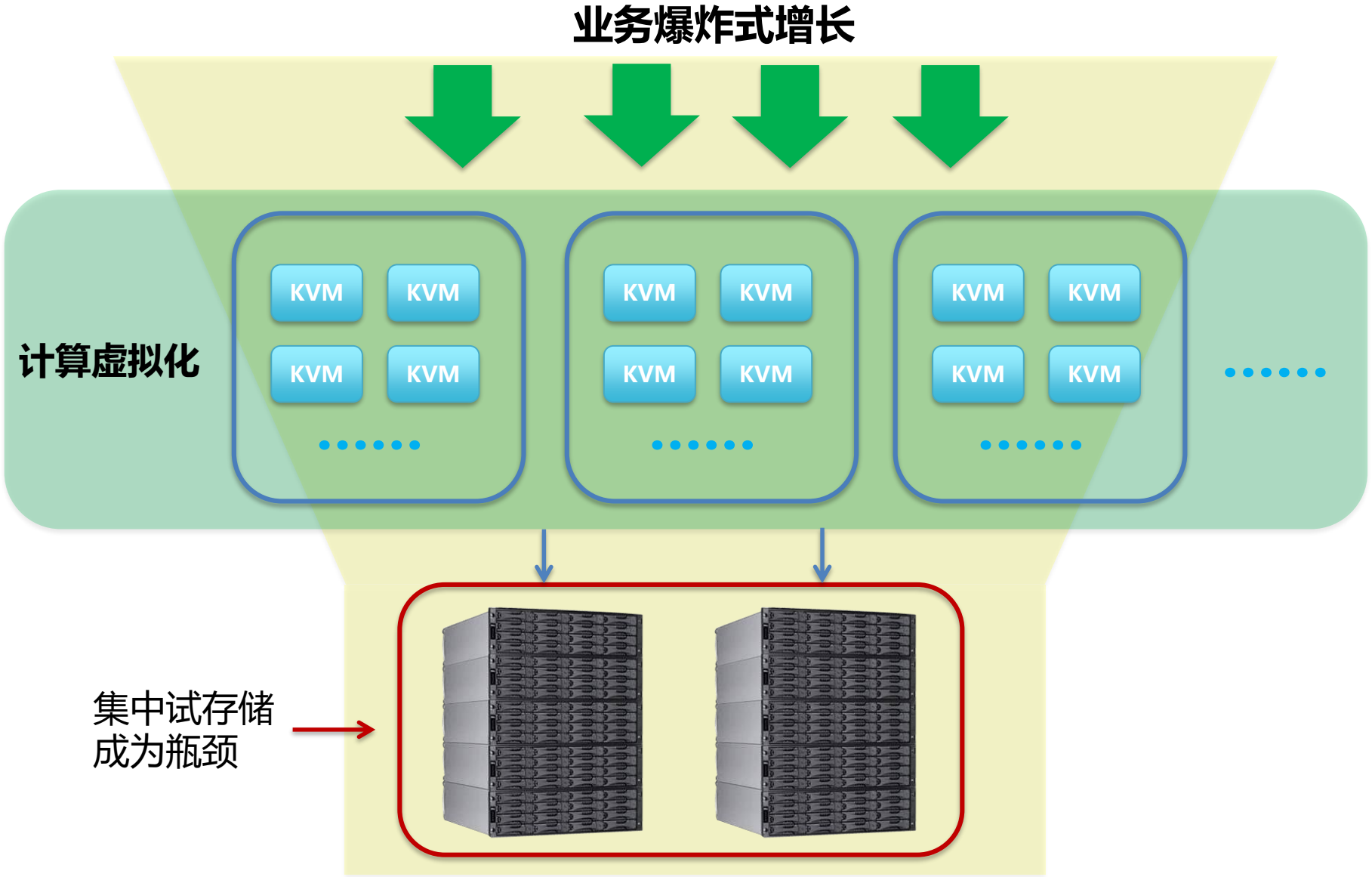
性能瓶颈

卷管理复杂性

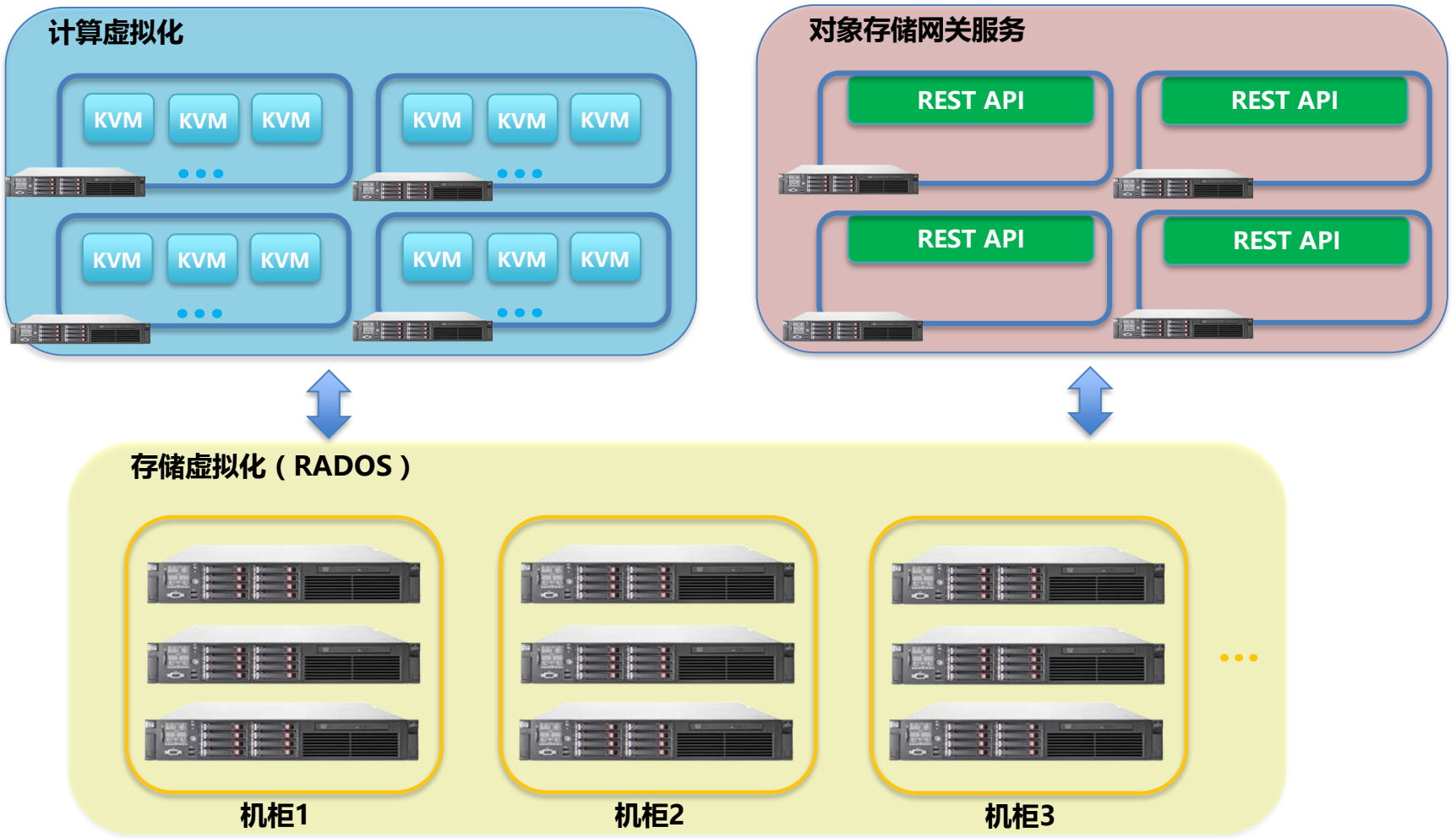
过保替换

高昂的运维成本

问题现状-存储虚拟化的紧迫需求



平安存储集群物理架构示意图



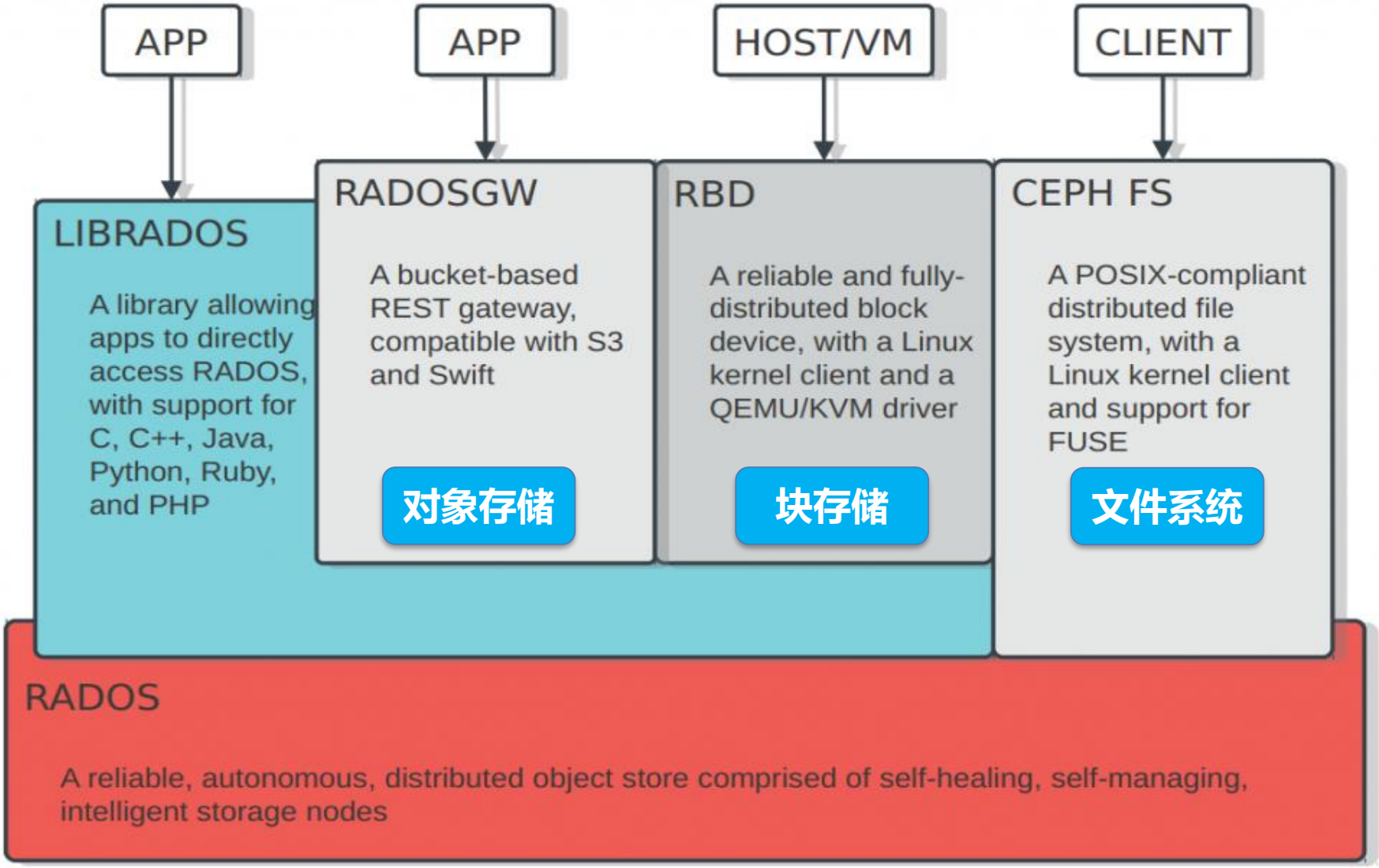
Ceph技术起源

Ceph孕育过程

2004年	Sage Weil在加州大学Santa Cruz分校攻读博士期间的研究课题
2006年	在OSDI学术会议上，Sage发表了介绍Ceph的博士论文
2007年	大学毕业后继续，继续进行此系统的开发
2010年	Linux Kernel 2.6.34 加入Ceph RBD的特性支持。
2011年	创立了Inktank公司以主导Ceph的开发和社区维护.（LGPL）
2014年	RedHat 以 1.75亿美元收购Inktank，加速了社区发展。
2016年	越来越多的资源投入到Ceph社区中……

Ceph技术架构

技术架构

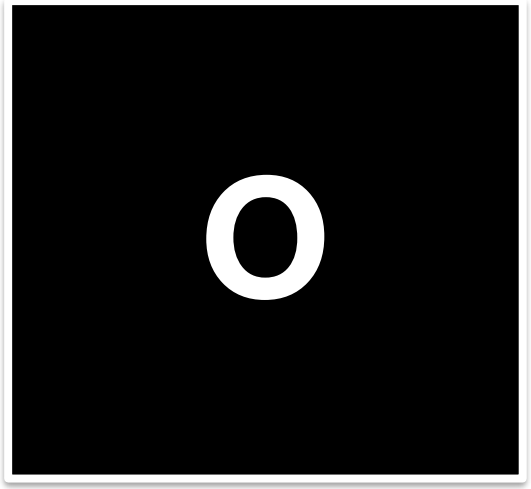

APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
对象存储
RBD
A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver
块存储
CEPH FS
A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE
文件系统
RADOS
A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

RADOS-MONITOR

Monitors：

- **维护集群物理地图及状态。**
- **提供分布式决策。**
- **非常小。**
- **不会存放Data。**

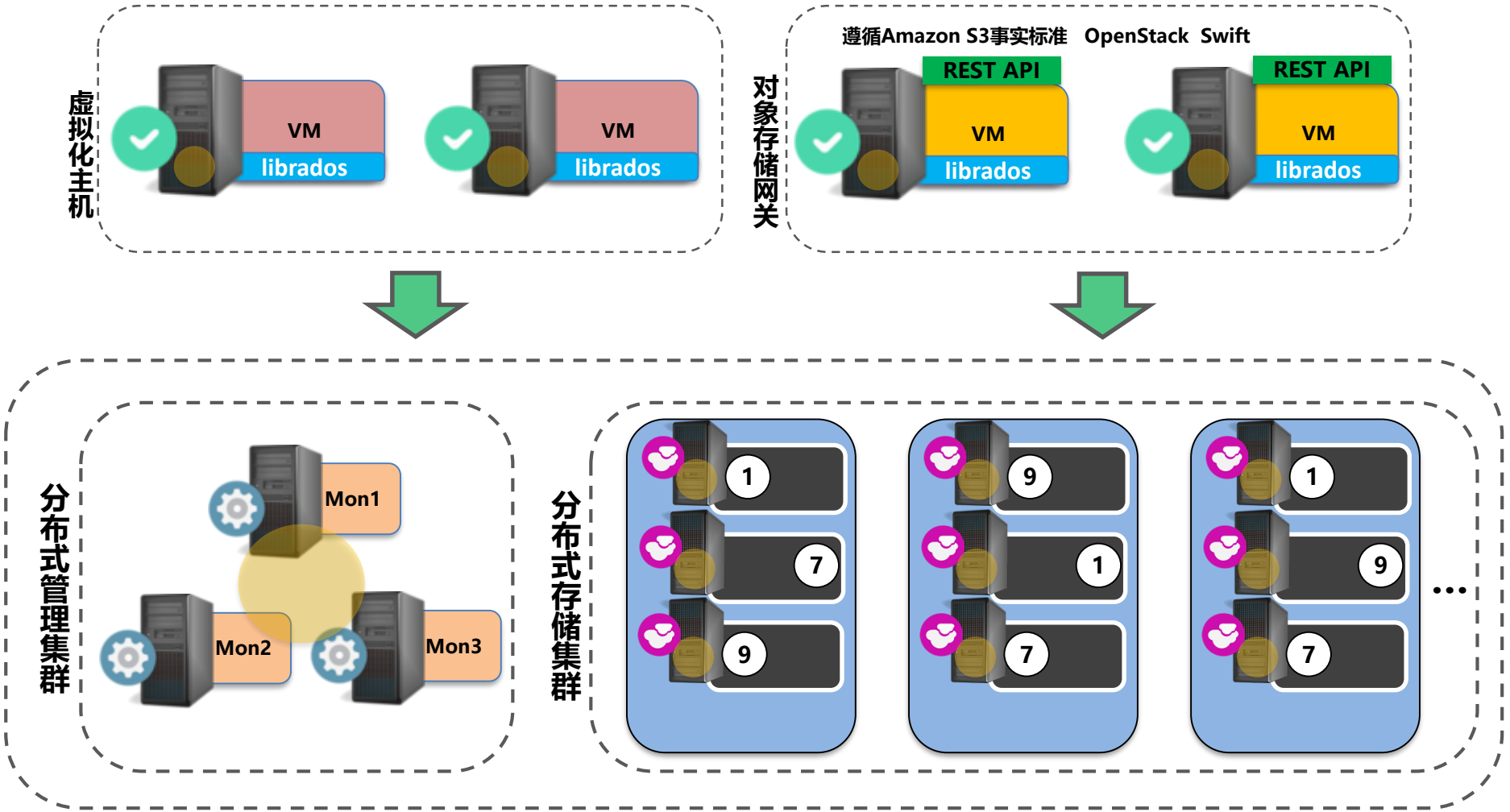
RADOS-OSD



OSD(Object Storage Daemon)：

- 提供计算和存储服务能力
- HDD OR SSD
- 存放Data

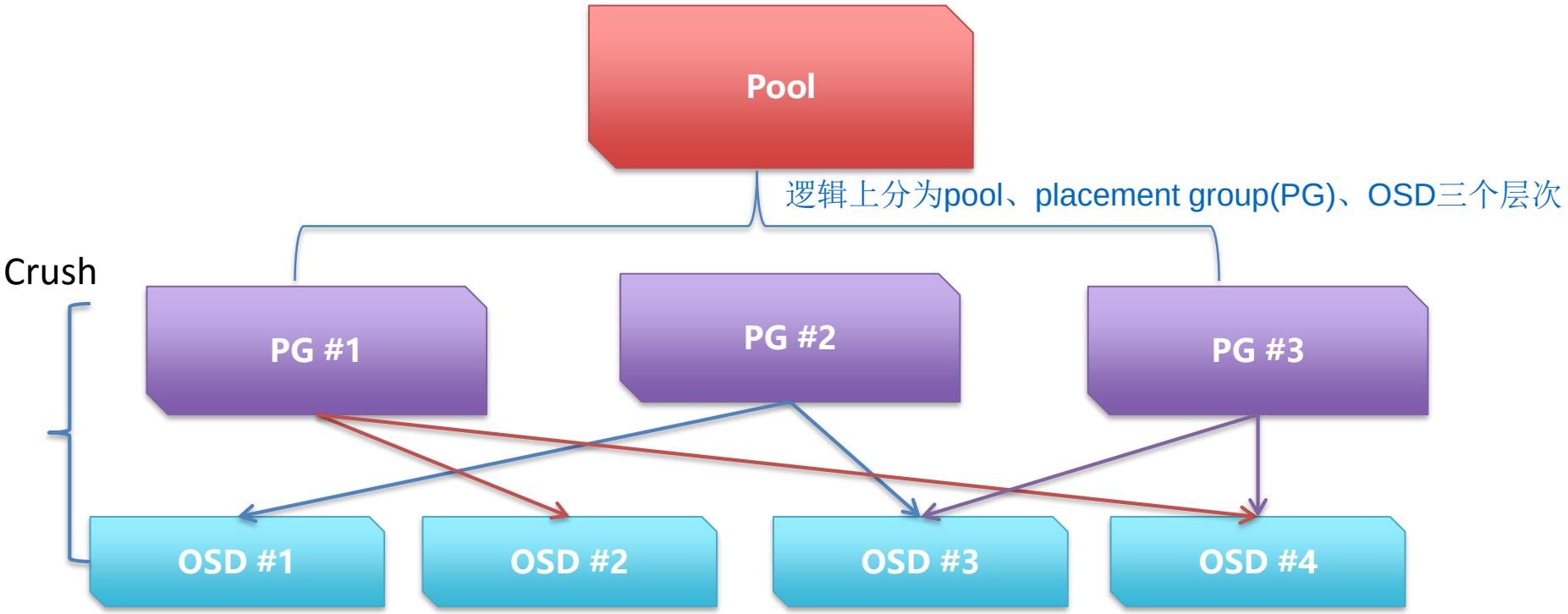
架构部署逻辑示意图



说明：◯ Object 存储多份副本分布示例

数据定位过程

逻辑数据结构



此两层映射关系由OSD Map、CRUSH Map和PG Map + CRUSH算法共同确定

新增PG或者OSD都将导致映射关系动态改变， OSD Map和PG Map也将随之更新不变的是数据所在的pool

集群PG总数 = 集群OSD总数 * 100 / 集群大部分pool的副本数

每个pool的PG个数 = 集群OSD总数 * 100 / 集群大部分pool的副本数 / 预估集群总pool个数

数据保护

- **很好的可靠性，耐用性，性能好**

多副本 Multi-Copies

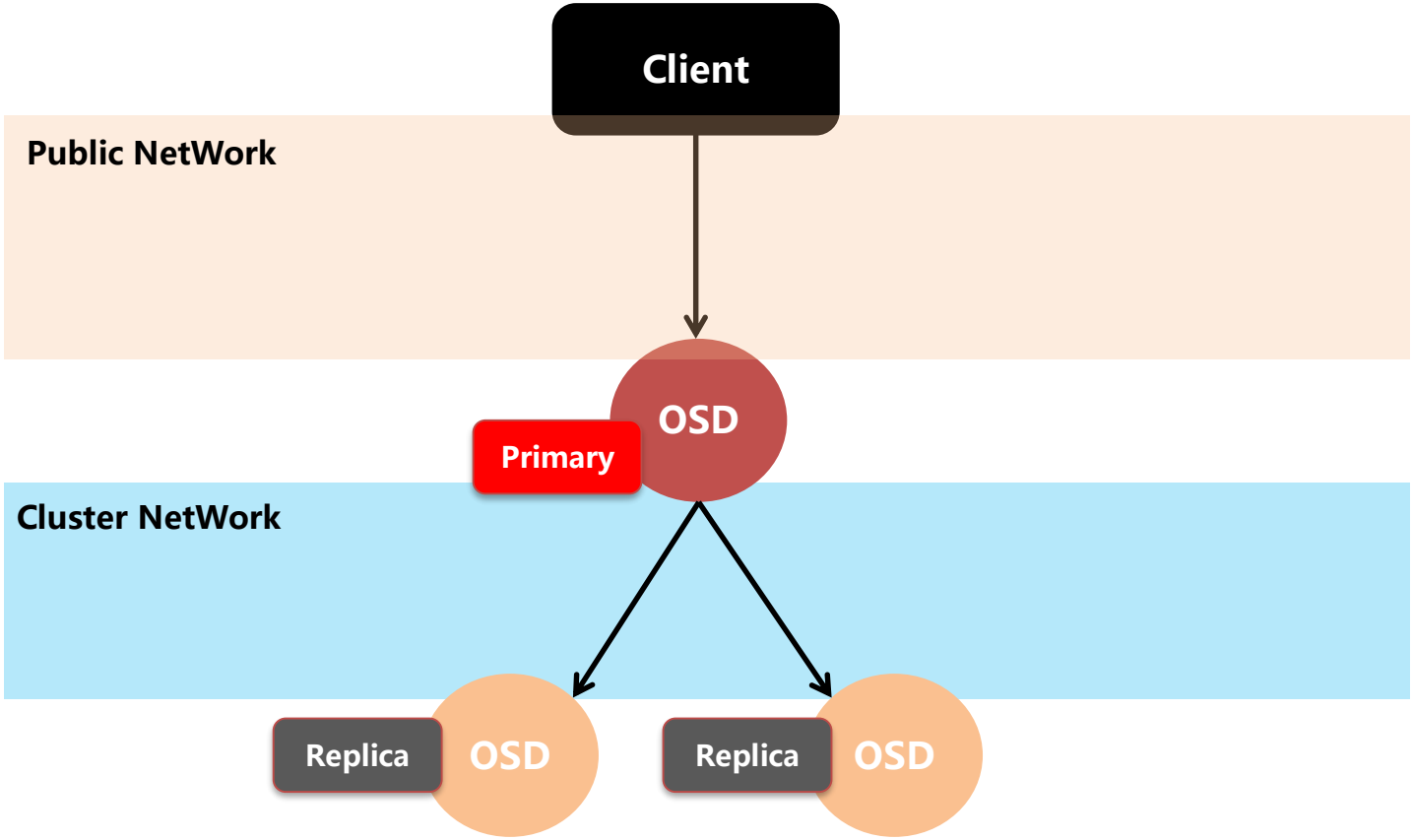
直接将文件对象存储多份，尽量分布在不同数据中心、不同的机房、不同的机柜、不同主机、不同对象存储服务上。来保证数据的可靠性。

- **很好的可靠性，耐用性，成本低**

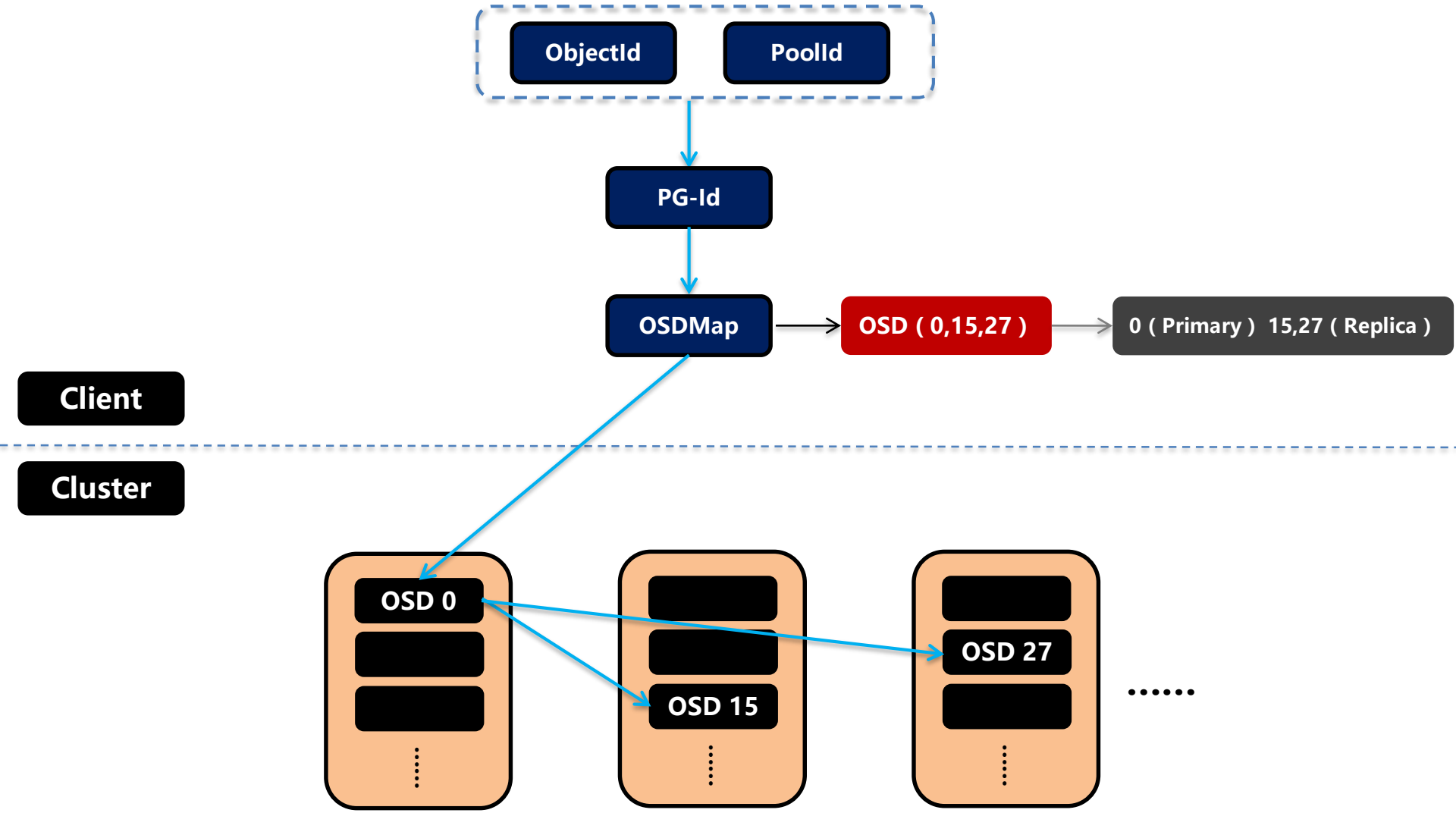
EC 模式 Erasure Code

指定数据块（K）和校验块（M）的个数，然后将文件拆分成K+M总数的数据对象进行存储。当且仅当丢了超过M个数据对象，数据才会丢失。

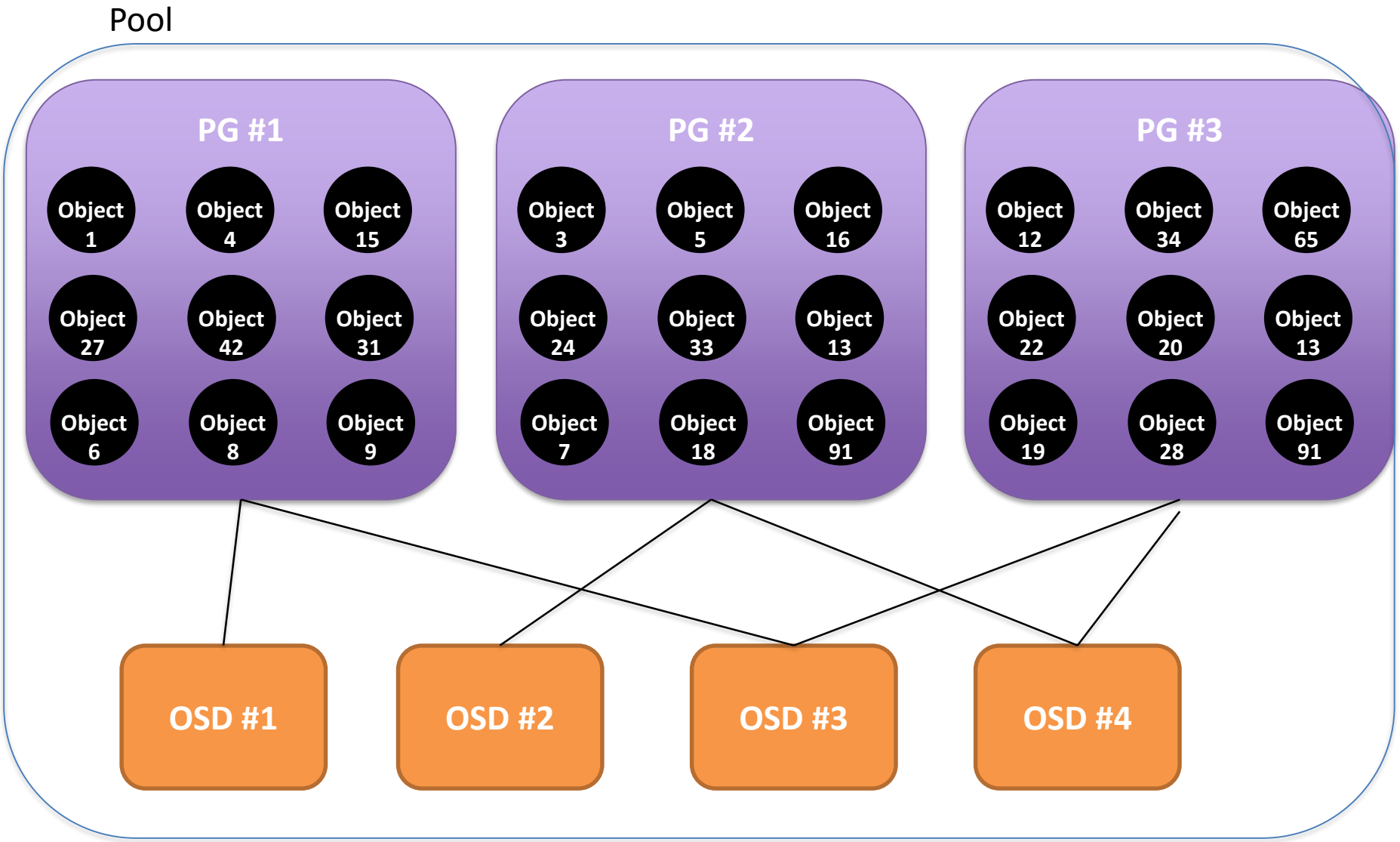
OSD角色区分


Client
Public NetWork
OSD
Primary
Cluster NetWork
Replica
OSD
Replica
OSD

读写流程

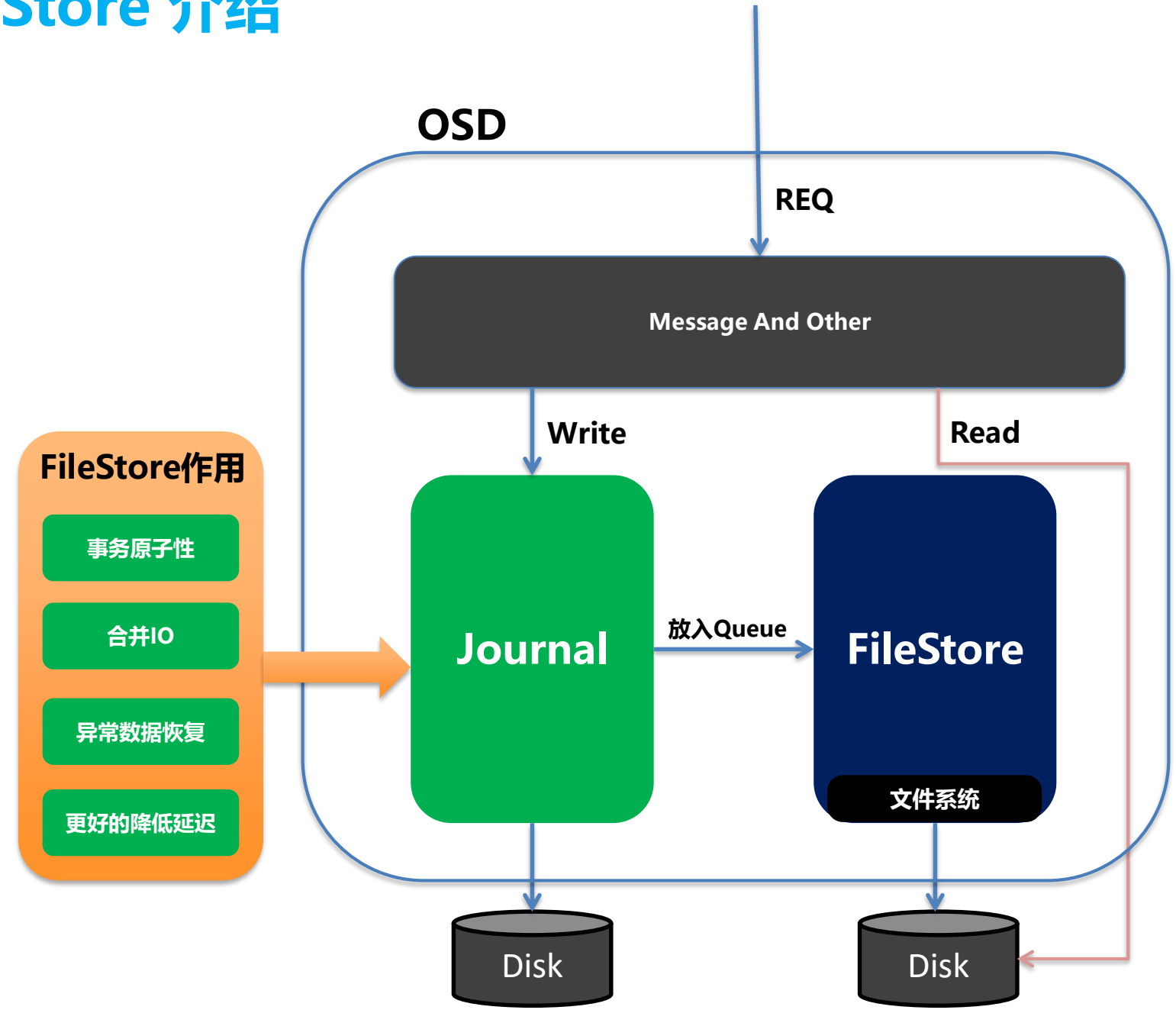

ObjectId
PoolId
PG-Id
OSDMap
OSD (0,15,27)
0 (Primary) 15,27 (Replica)
Client
Cluster
OSD 0
OSD 15
OSD 27
.....

数据分布

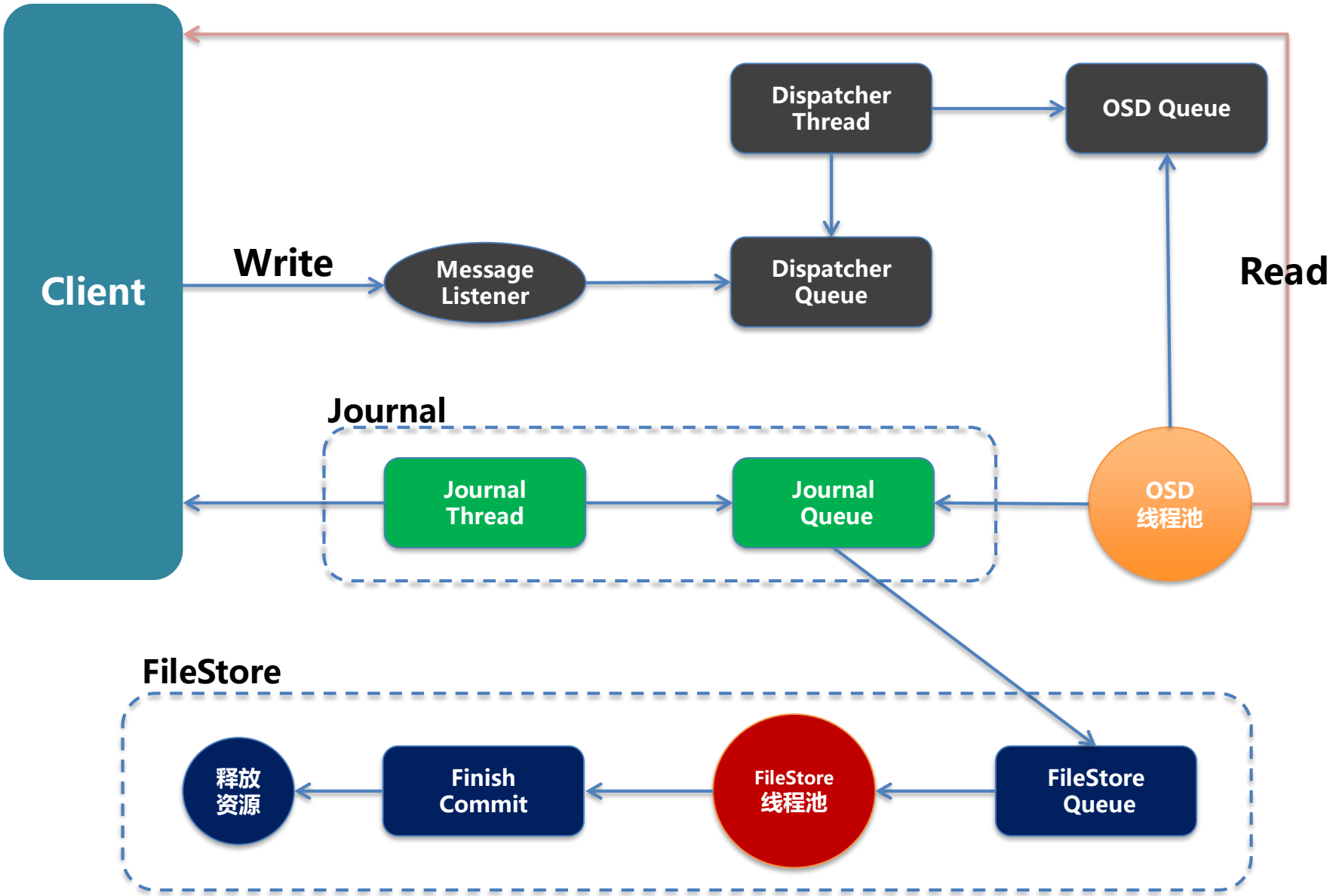

Pool
PG #1
Object 1
Object 4
Object 15
Object 27
Object 42
Object 31
Object 6
Object 8
Object 9
PG #2
Object 3
Object 5
Object 16
Object 24
Object 33
Object 13
Object 7
Object 18
Object 91
PG #3
Object 12
Object 34
Object 65
Object 22
Object 20
Object 13
Object 19
Object 28
Object 91
OSD #1
OSD #2
OSD #3
OSD #4

FileStore & BlueStore

FileStore 介绍



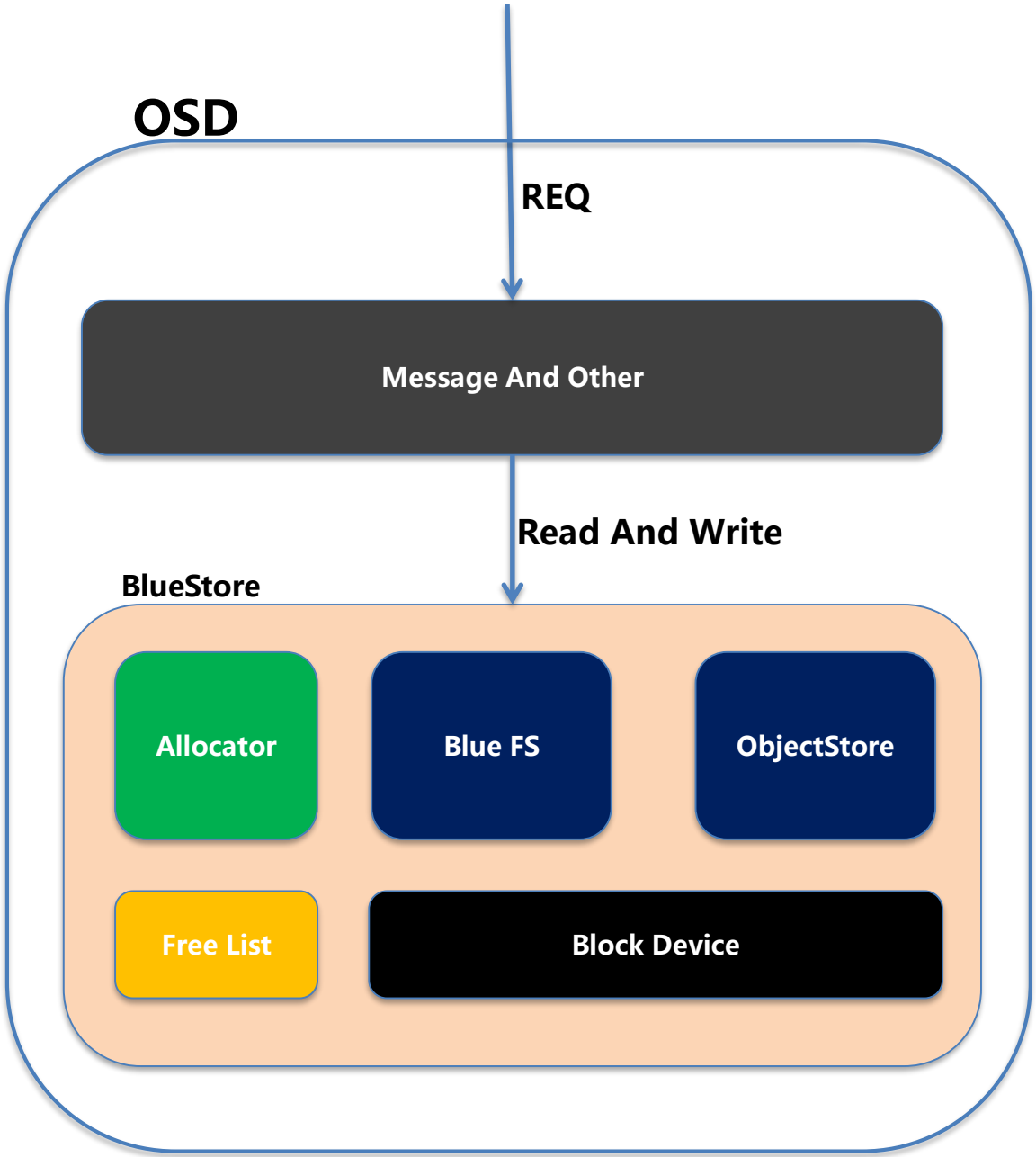
FileStore 读写流程



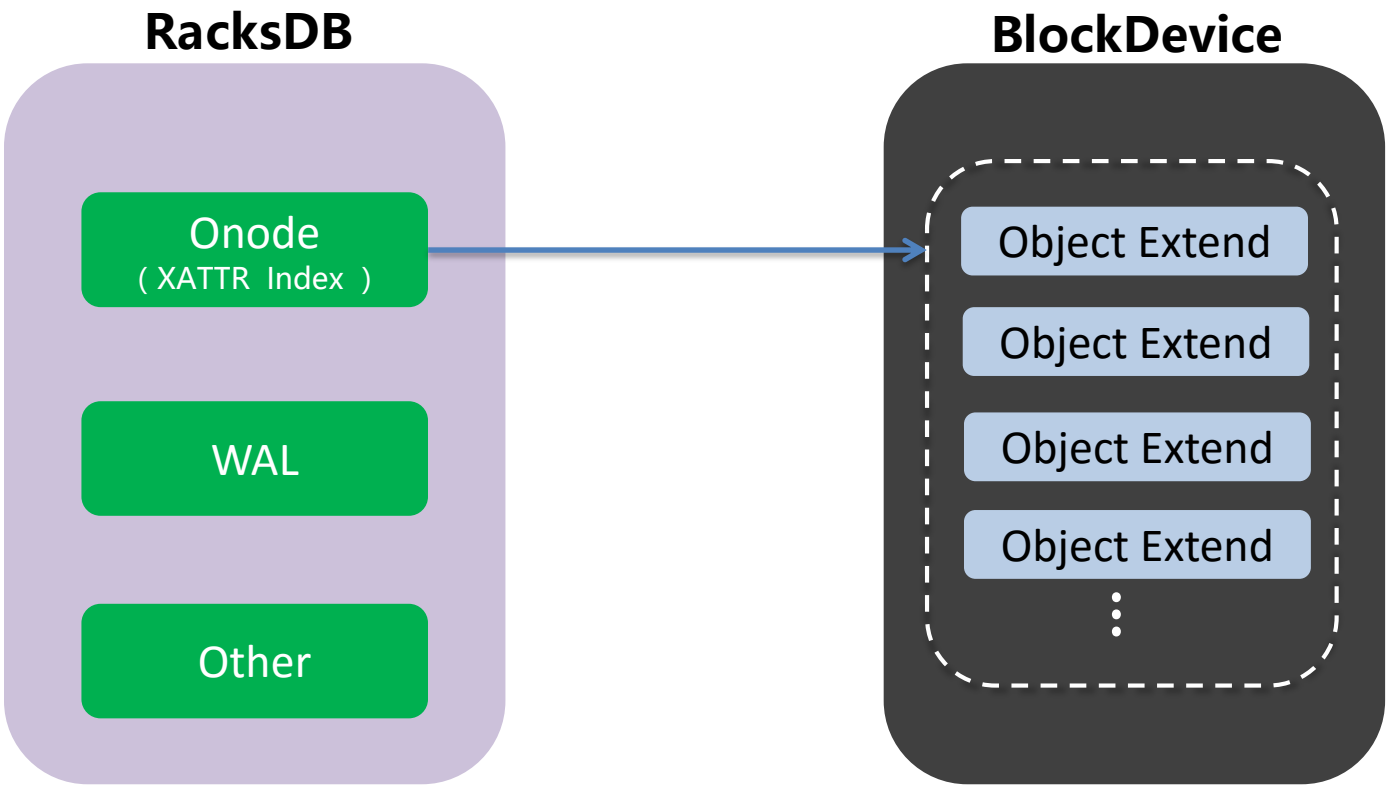
FileStore 参数调整

```
[root@host01 ~]# ceph daemon osd.0 config show | grep filestore_wb
    "filestore_wbthrottle_enable": "true",
    "filestore_wbthrottle_btrfs_bytes_start_flusher": "41943040",
    "filestore_wbthrottle_btrfs_bytes_hard_limit": "419430400",
    "filestore_wbthrottle_btrfs_ios_start_flusher": "500",
    "filestore_wbthrottle_btrfs_ios_hard_limit": "5000",
    "filestore_wbthrottle_btrfs_inodes_start_flusher": "500",
    "filestore_wbthrottle_xfs_bytes_start_flusher": "41943040",
    "filestore_wbthrottle_xfs_bytes_hard_limit": "419430400",
    "filestore_wbthrottle_xfs_ios_start_flusher": "500",
    "filestore_wbthrottle_xfs_ios_hard_limit": "5000",
    "filestore_wbthrottle_xfs_inodes_start_flusher": "500",
    "filestore_wbthrottle_btrfs_inodes_hard_limit": "5000",
    "filestore_wbthrottle_xfs_inodes_hard_limit": "5000",
```

BlueStore (1)


OSD
REQ
Message And Other
Read And Write
BlueStore
Allocator
Blue FS
ObjectStore
Free List
Block Device

BlueStore (2)


RacksDB
Onode
(XATTR Index)
WAL
Other
BlockDevice
Object Extend
Object Extend
Object Extend
Object Extend

存储海量小文件的难题-小结

文件没有进行合并处理

受限文件系统

BlueStore解决

写放大过于严重

BlueStore解决

如何让Ceph跑的更快

硬件调优（1）

X86 PC

磁盘密度高

热插拔

低耗电

硬件调优（2）

CPU

2 HDD = 1 CORE

内存

1 TB HDD = 1 GB Mem

硬盘

1 SSD = 5 HDD

趟坑经验分享

分享-Cannot Create Thread

2015-11-27 21:47:22.543027 7f9a8acb0700 0 osd.99 11830 do_command r=0

2015-11-27 21:47:57.113237 7f9a9b0ca700 -1 common/Thread.cc<http://thread.cc>;: In function 'void Thread::create(size_t)' thread 7f9a9b0ca700 time 2015-11-27 21:47:57.103392

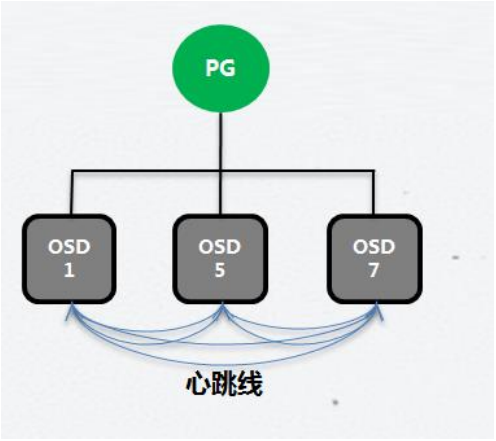
common/Thread.cc<http://thread.cc>;: 129: FAILED assert(ret == 0)

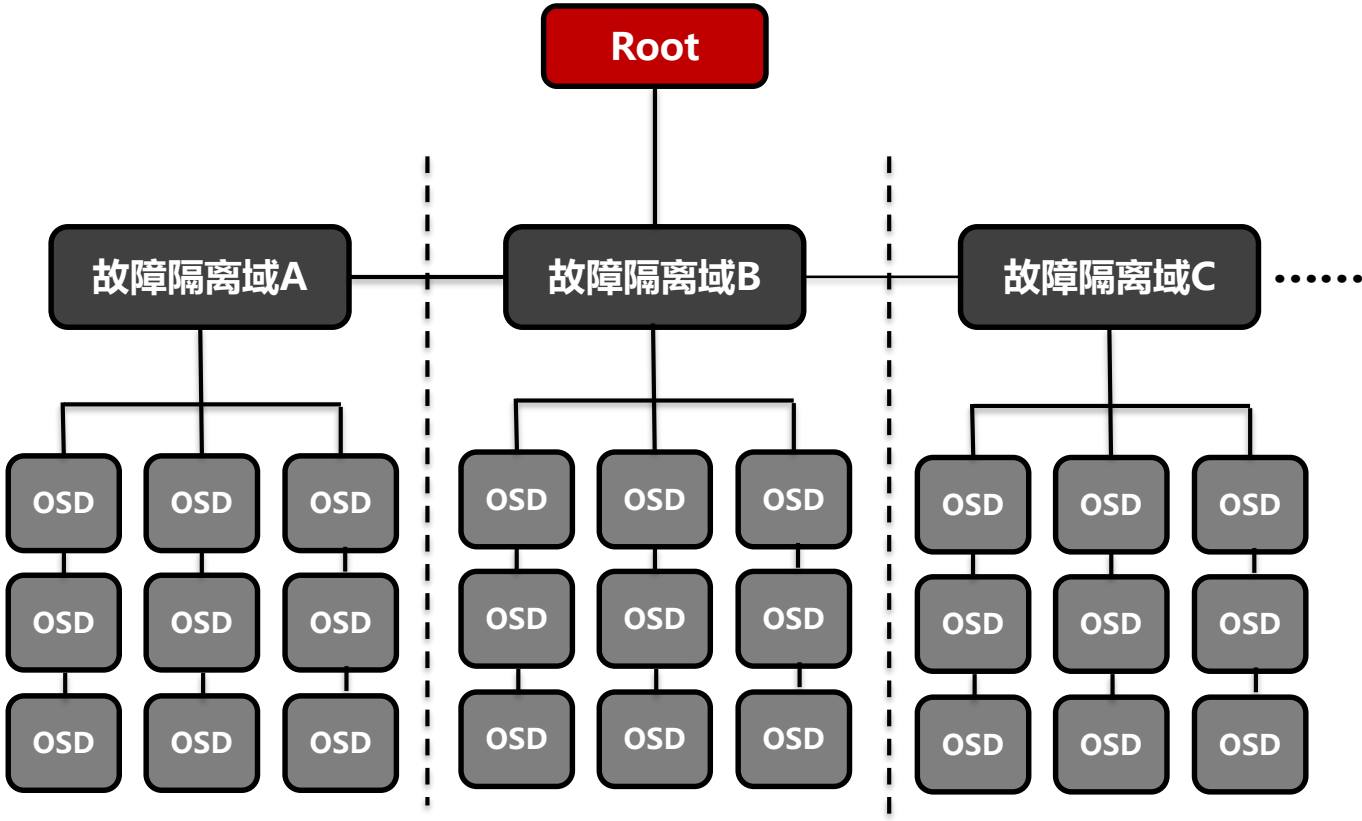
- [ceph.conf] max_open_files=200000（默认32768）
- [Kernel] threads-max=1028467
- [Kernel] pid_max=65535(甚至更大)

```
        # Increase max_open_files, if the configuration calls for it.
        get_conf max_open_files "32768" "max open files"

        # build final command
    wrap=""
    runmode=""
    runarg=""

    [ -z "$docrun" ] && get_conf_bool docrun "0" "restart on core dump"
    [ "$docrun" -eq 1 ] && wrap="$BINDIR/ceph-run"

    [ -z "$dovalgrind" ] && get_conf_bool valgrind "" "valgrind"
    [ -n "$valgrind" ] && wrap="$wrap valgrind $valgrind"

    [ -n "$wrap" ] && runmode="-f &" && runarg="-f"
    [ -n "$max_open_files" ] && files="ulimit -n $max_open_files;"
```

分享-控制故障隔离域

合理的规划CrushMap，控制一个故障域中OSD的数目。



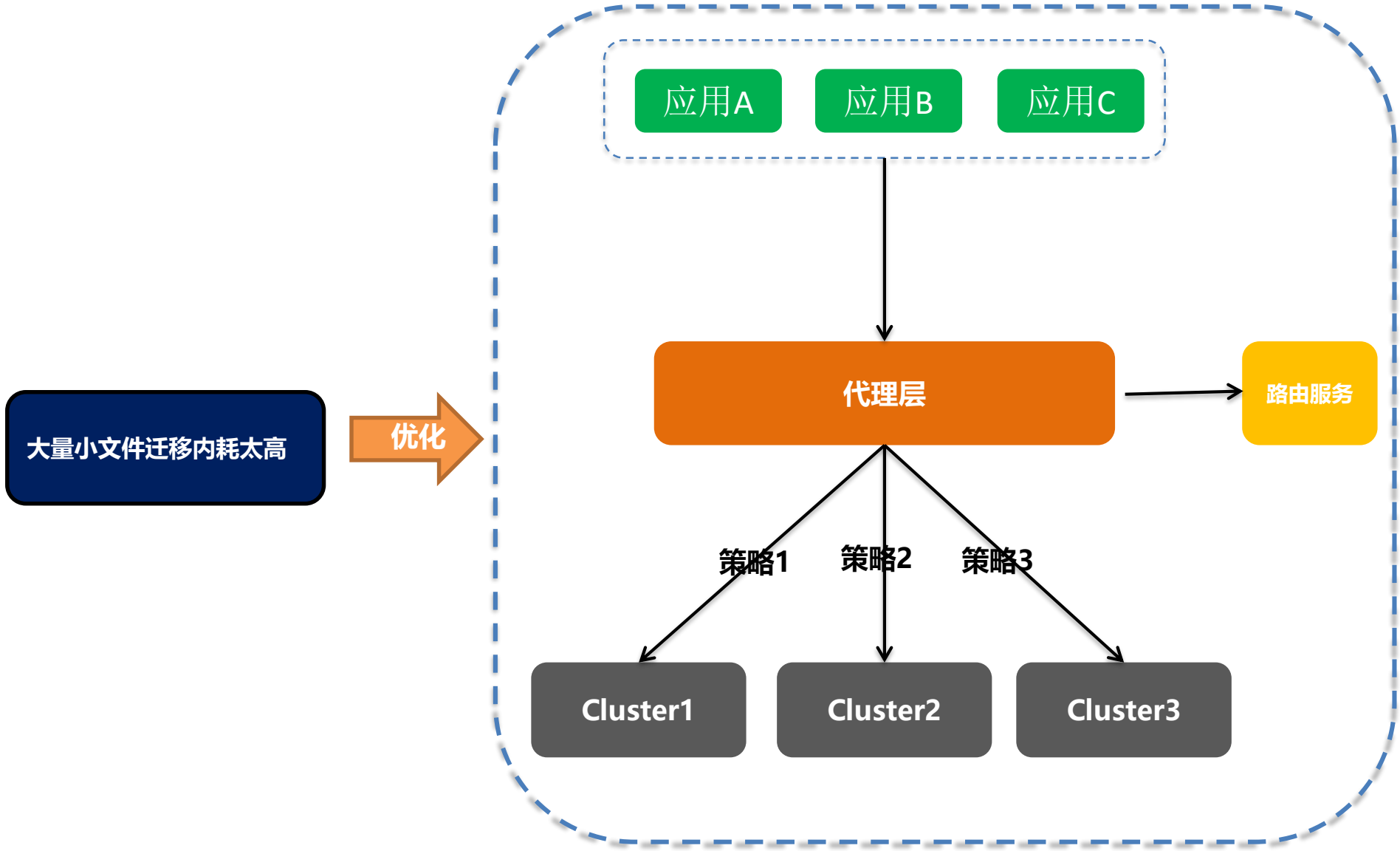


Placement Rule：

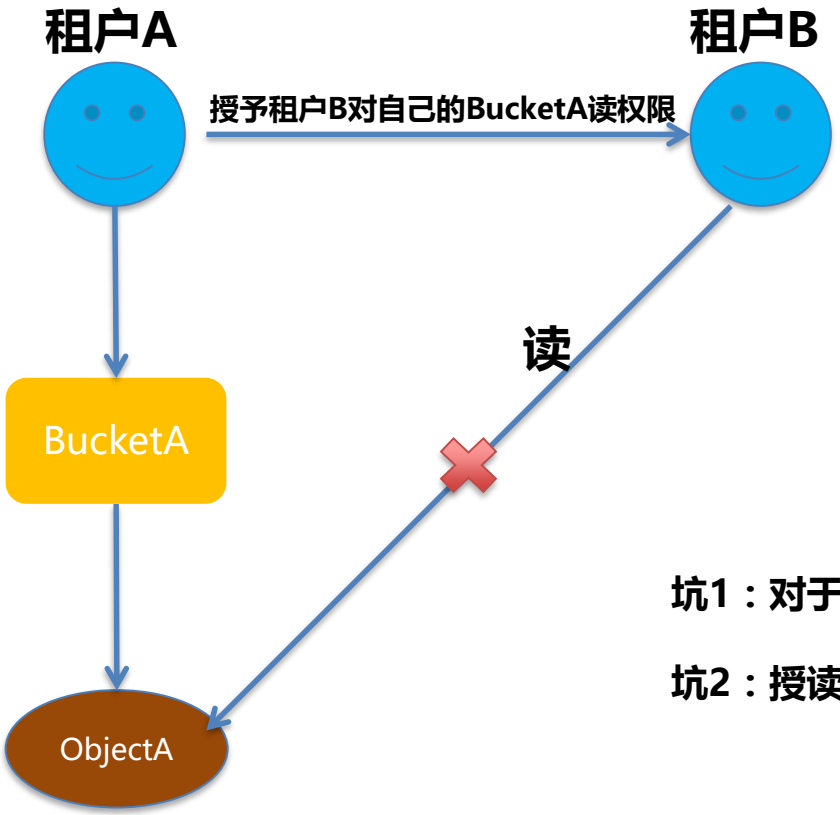
```
[ruleA]
    ruleSet   2
    take from 故障隔离域
    take chooseleaf 0 type host
    take emit
```

Note：Jewel中已经加入Async messager，但预估要在K、L中可以稳定应用。

分享-分集群来控制稳定的性能



分享-S3协议中的大坑



坑1：对于Bucket授权读写是一个权限。

坑2：授读不能读。

分享-多数据中心自动灾备