Ceph

数据一致性浅析

可靠性和一致性的矛盾

Ceph简介

Ceph处理一致性的机制

Ceph数据异常处理机制

Ceph的scrub机制

CRUSH与一致性Hash

Ceph的应用场景

例 高性能架构中Ceph的应用

可靠、高效与一致性的矛盾

动态 是现阶段云计算平台的共同特征与挑战

计算资源动态

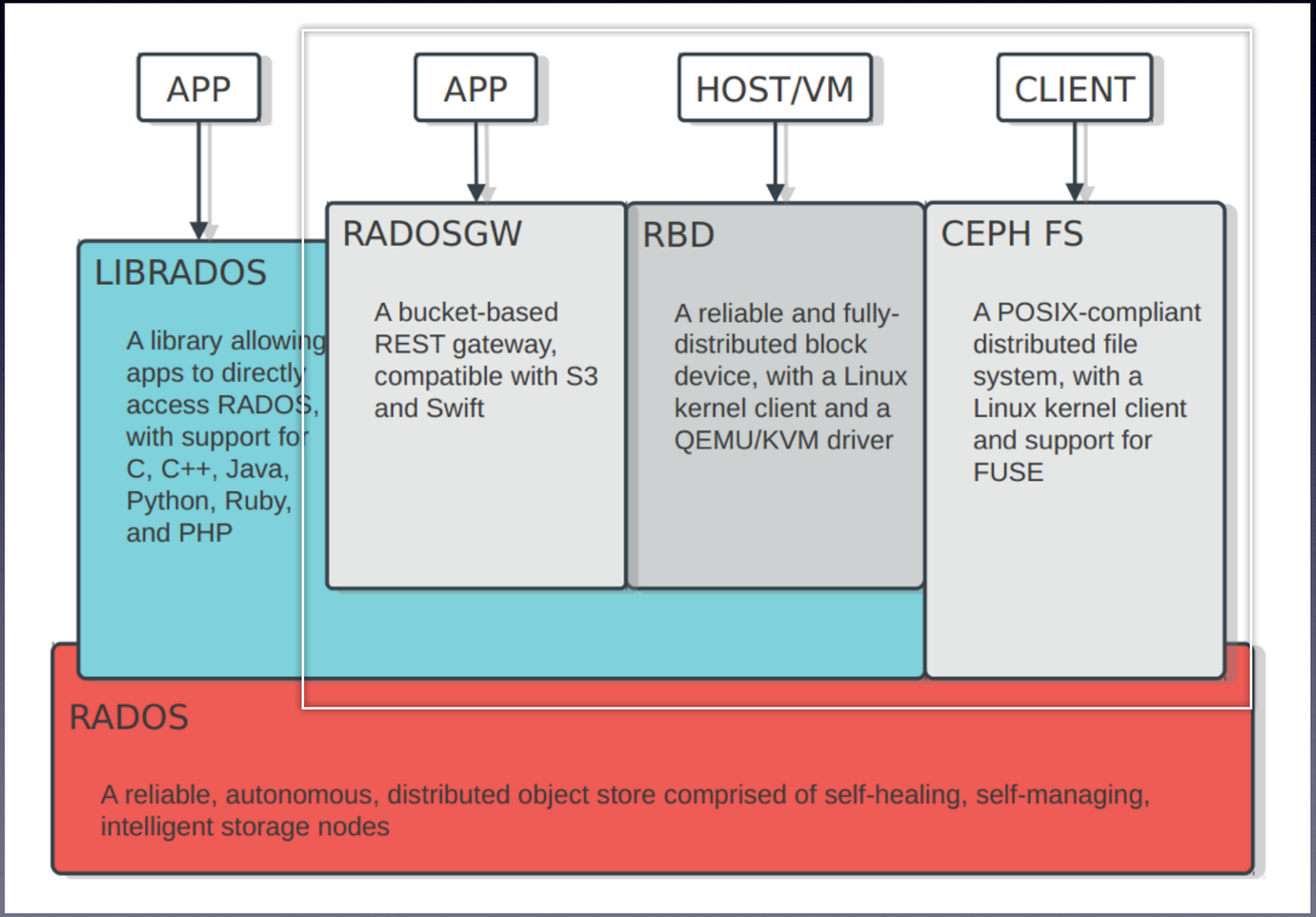
存储资源动态

管理信息动态

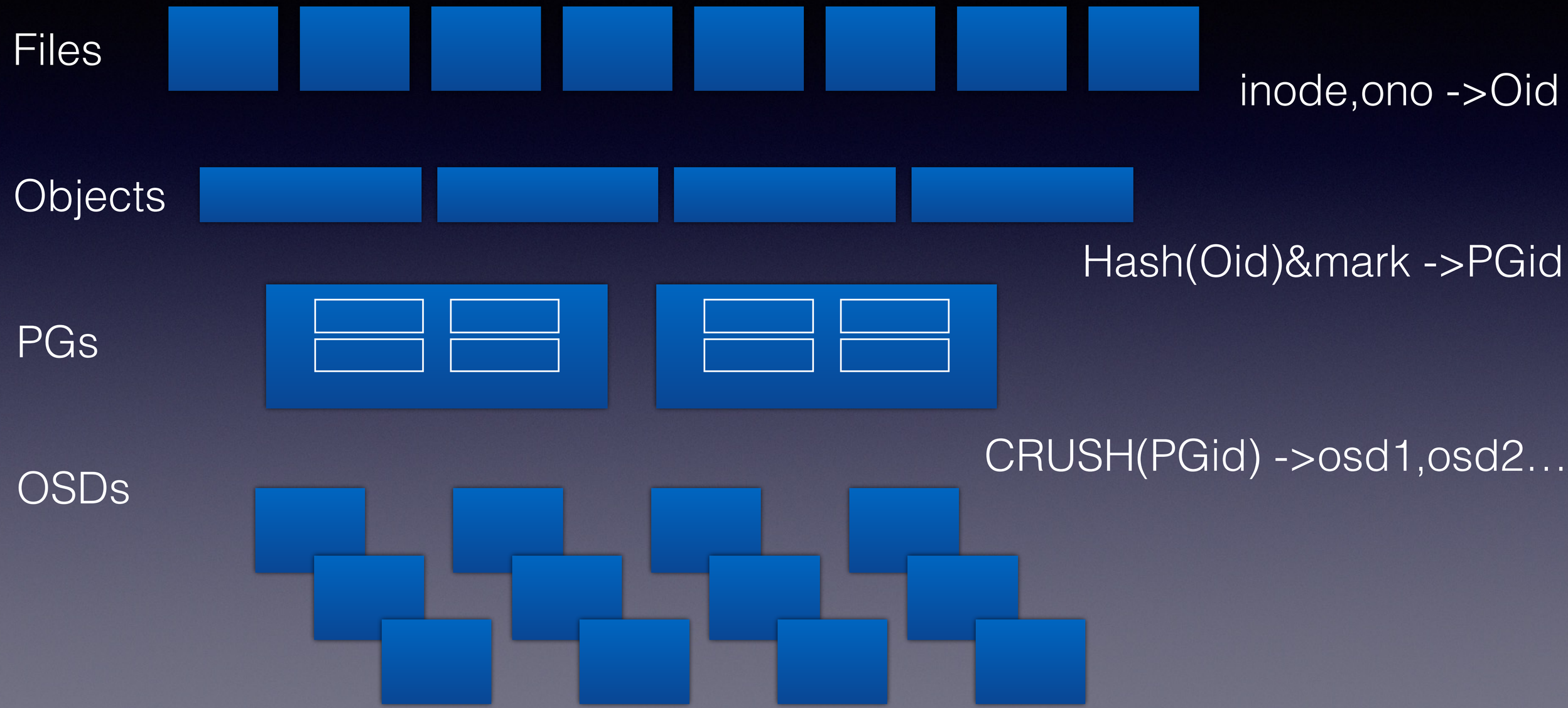
动态的需求造成对云的可靠性、高效、一致性都有较高的期望
这些要求是相互制约和矛盾的

Ceph简介

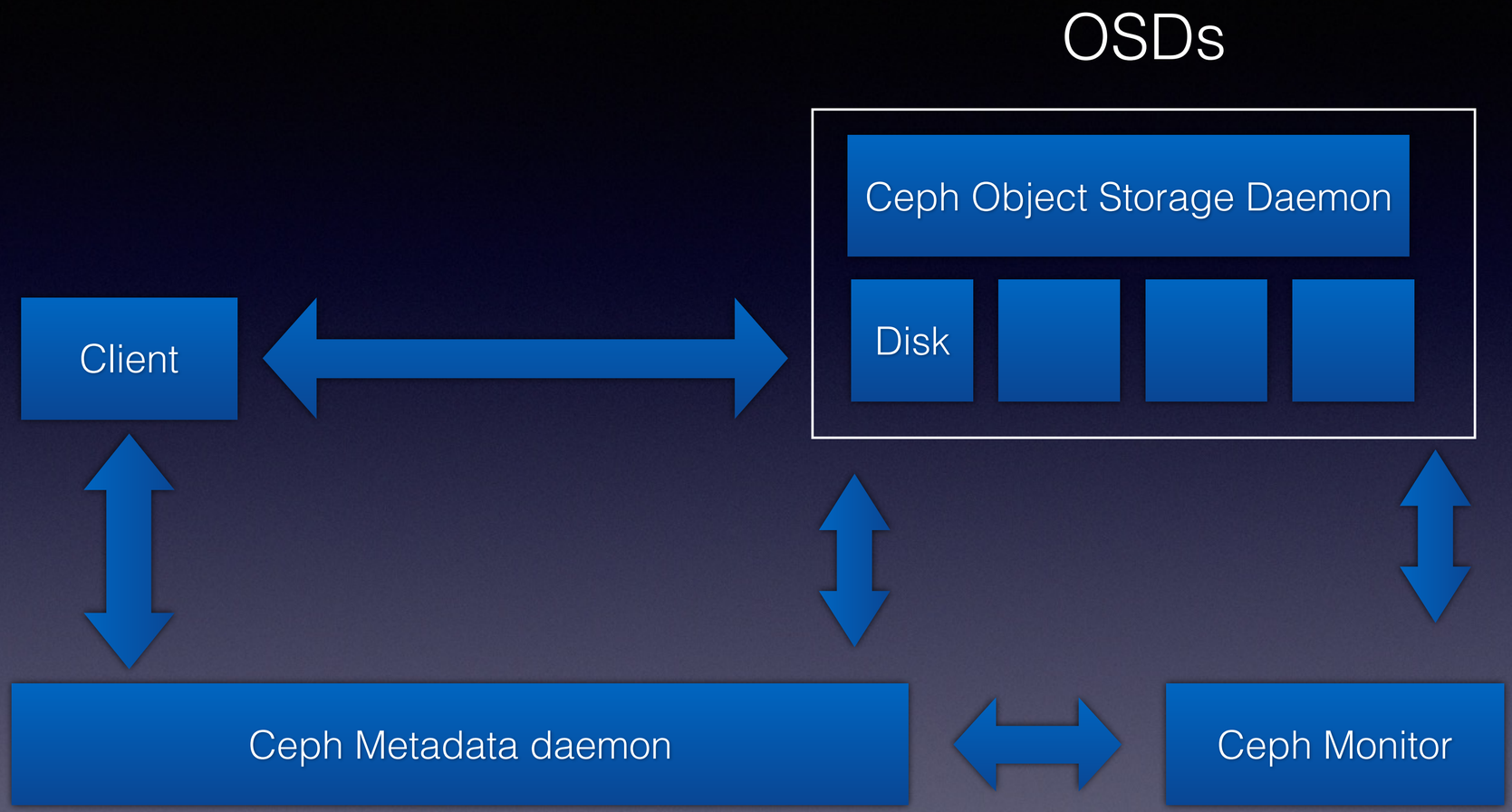
一个提供了对象、块以及兼容POSIX文件访问的分布式统一存储系统



Ceph简介

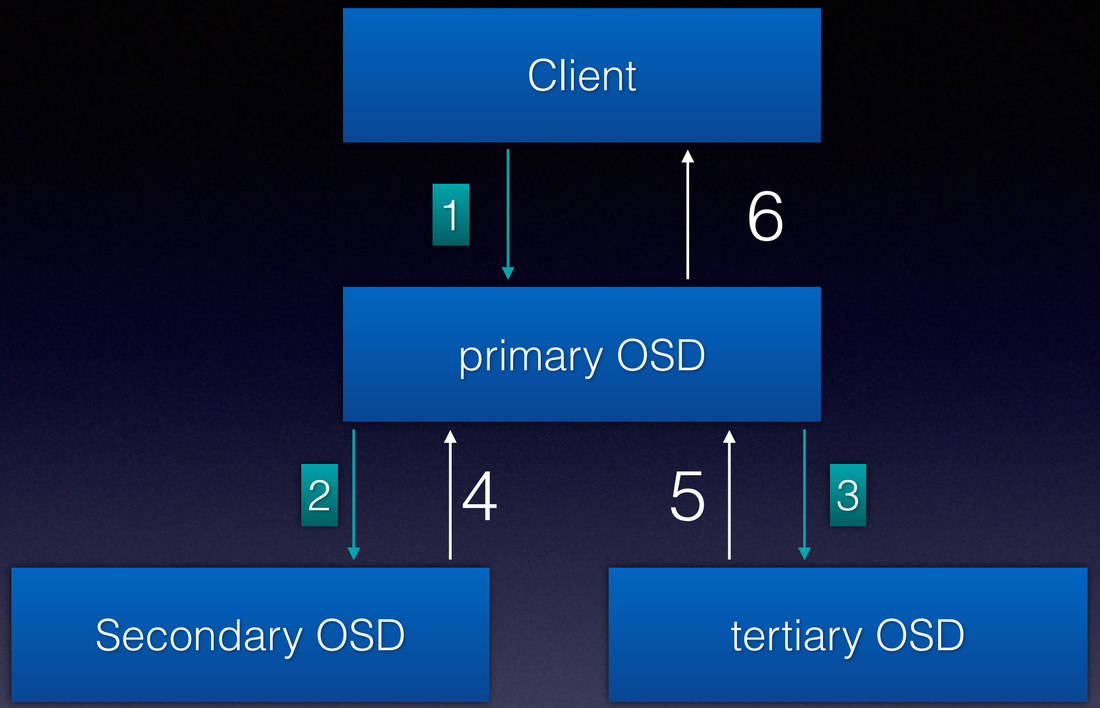

Files
inode,ono ->Oid
Objects
Hash(Oid)&mark ->PGid
PGs
CRUSH(PGid) ->osd1,osd2…
OSDs

Ceph简介



CRUSH Map　　OSD Map　　CRUSH Ruleset

Ceph处理数据一致性的机制



- Ceph的读写操作采用主-副模型，Client只向Object所对应OSD 集合的主OSD发起读写请求，这保证了数据的强一致性。
- 由于每个Object都只有一个主OSD，因此对Object的更新都是顺序的，不存在同步问题。
- 当主OSD收到写请求时，它负责把数据发送给其他副本，只有这个数据被保存在所有的OSD上时，主OSD才确认完成写请求，这保证了副本的一致性。

Ceph的数据异常处理机制

系统断电、重启、网络故障

PGLog

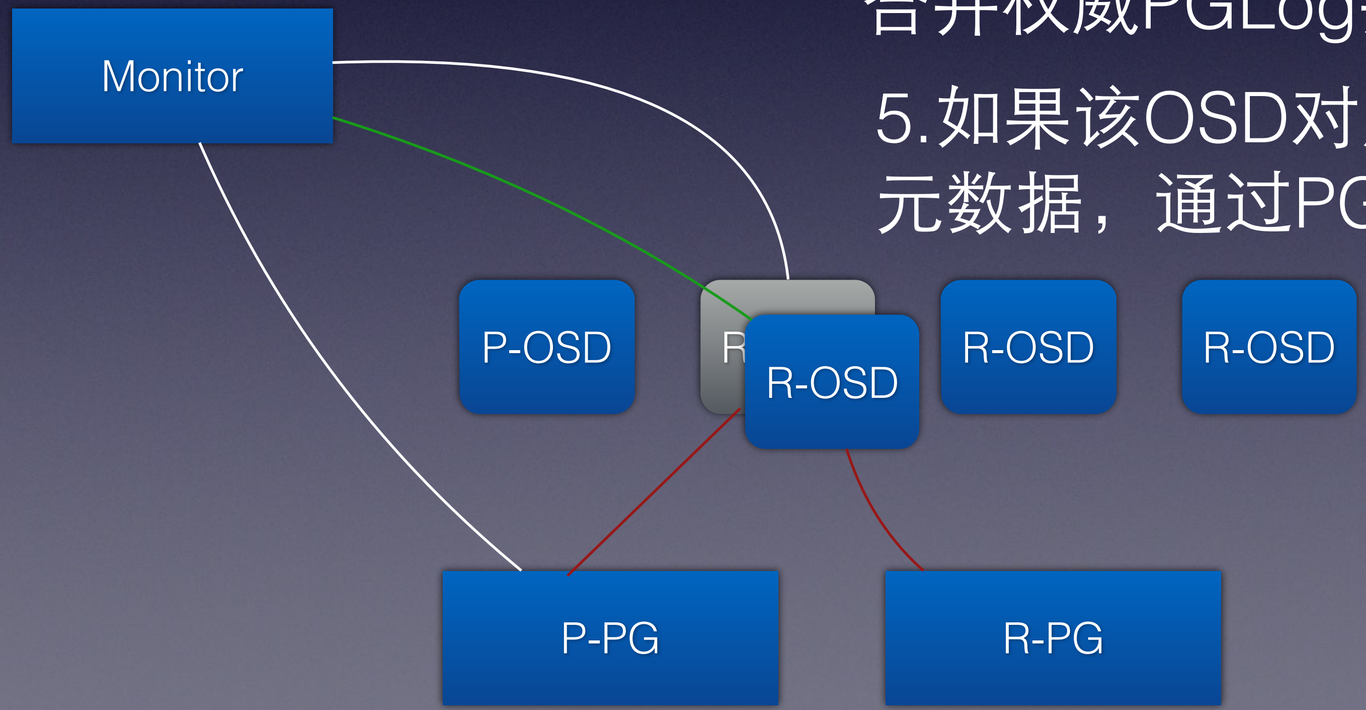
1.异常发生，Monitor发现并通知对应的PrimaryPG

2.PG标示为Degraded状态，并增加PGLog记录

3.OSD重新上线，先在Monitor注册，读取PGLog

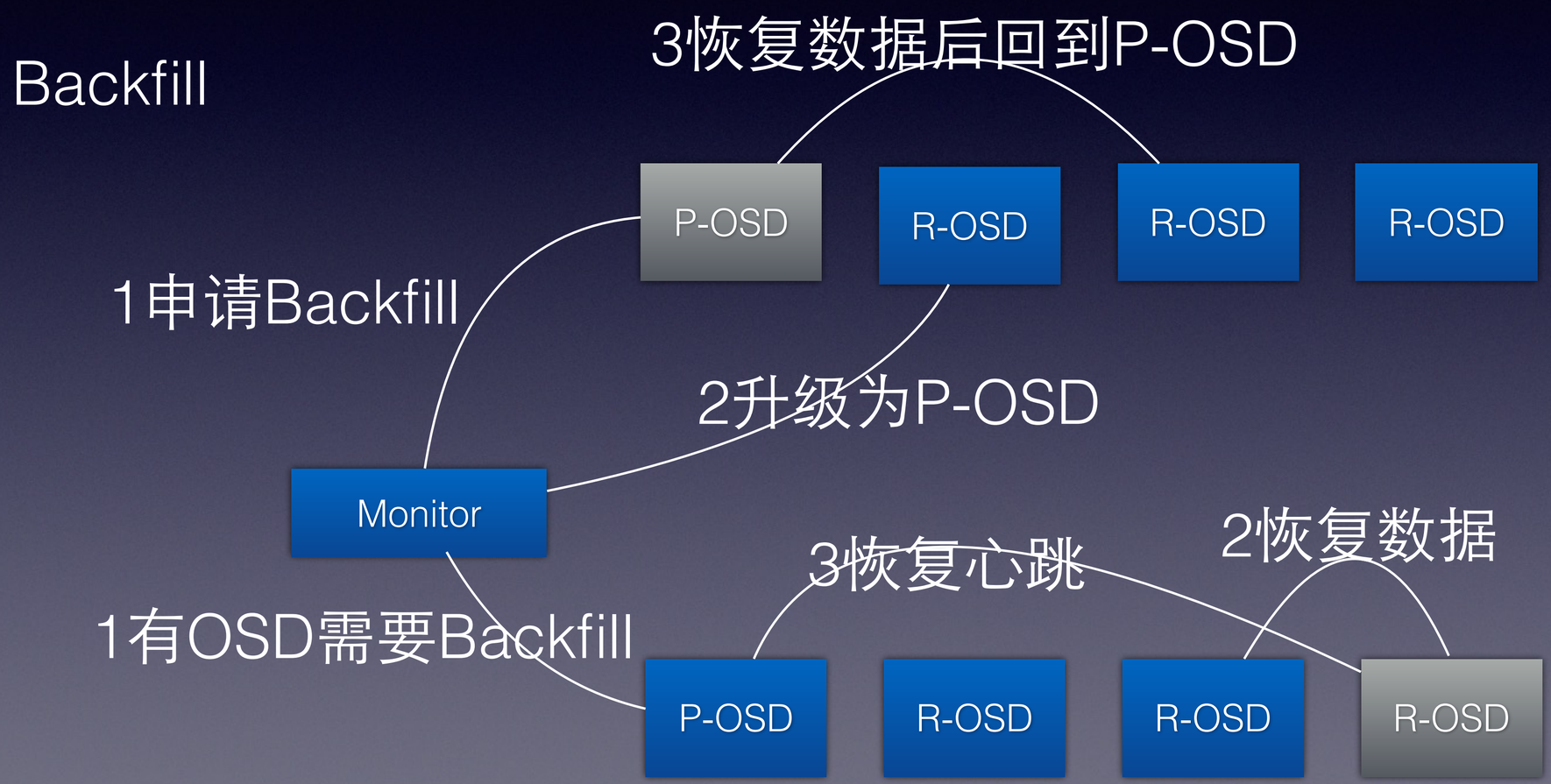
4.如果该OSD对应的是PrimaryPG，需要发起元数据查询，故障期接替的PG记录了权威的PGLog，该OSD合并权威PGLog并更新其落后状态

5.如果该OSD对应的是ReplicatePG，上线会被查询元数据，通过PGLog的Missing表格更新元数据



Ceph的数据异常处理机制

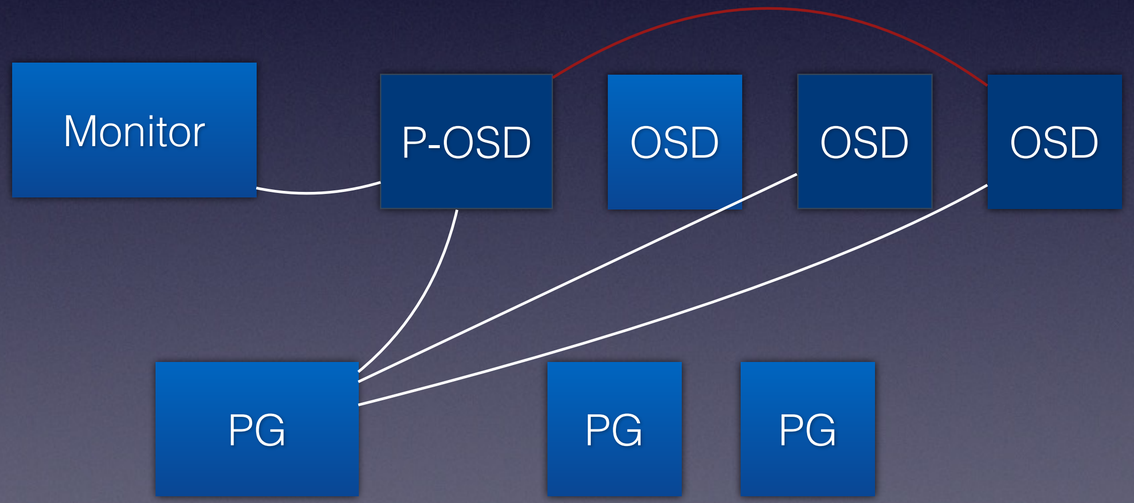
OSD失效，损坏



Ceph的scrub机制

Read verify方式定时扫描部分对象，副本间的对比发现非一致数据

Ceph会使用每一个对象名哈希值的部分作为依据，每次启动scrub时，对一部分不会受到修改的对象进行校验。



PG对应的P-OSD发起

P-OSD收集对象集信息

计算校验信息ScrubMap并对比

不一致对象信息会发给Monitor

启动PG Repair

CRUSH与一致性Hash

	CRUSH	一致性Hash
存储池内负载平衡	Hash(x)%PGs PG是抽象的存储节点，在PG层面，数据是均匀分布的	Hash(x)%N 基本算法 在N的层面数据平均分布
存储节点变化应对	由于PG是抽象的存储节点，不会随着物理节点变化而变化 分区变化时，在PG的管理范围内进行计算与数据迁移	一致性Hash通过将数据和存储节点映射到同个Hash空间减少节点变化的数据迁移 分区变化时，已写入数据需要重新计算Hash
副本分布风险控制	PG划分了固定分区 副本可以存储在不同的故障隔离区，确保数据安全	引入虚拟节点、固定分区等方法对数据分布做更优化处理 Dynamo等使用了CRUSH类似的思想来改进一致性Hash

Ceph的应用场景

存储需求多样化

- 统一存储简化开发复杂度，提供多种存储接口
- 为VM提供快照、克隆的高性能块存储

高可扩展

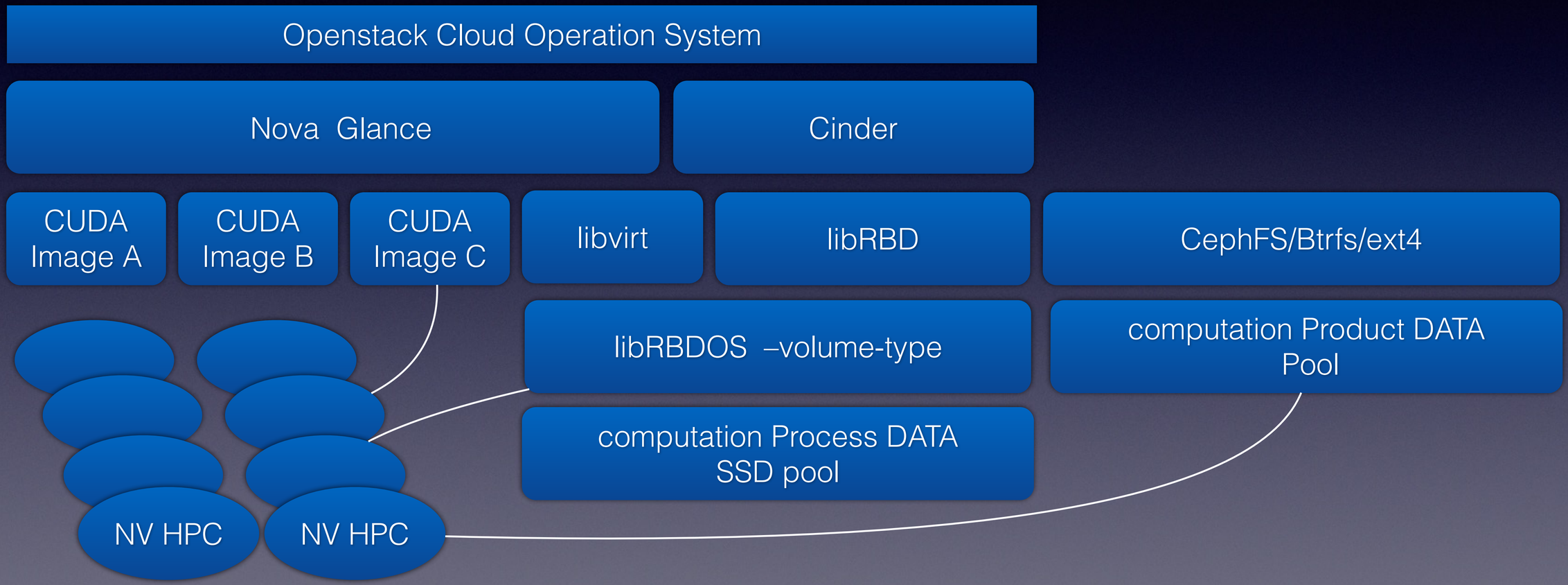
- 支持不同层次的硬件
- 动态可靠性

可用性

- 增量扩展操作性强，维护管理方便
- 无缝迁移，数据可靠性高

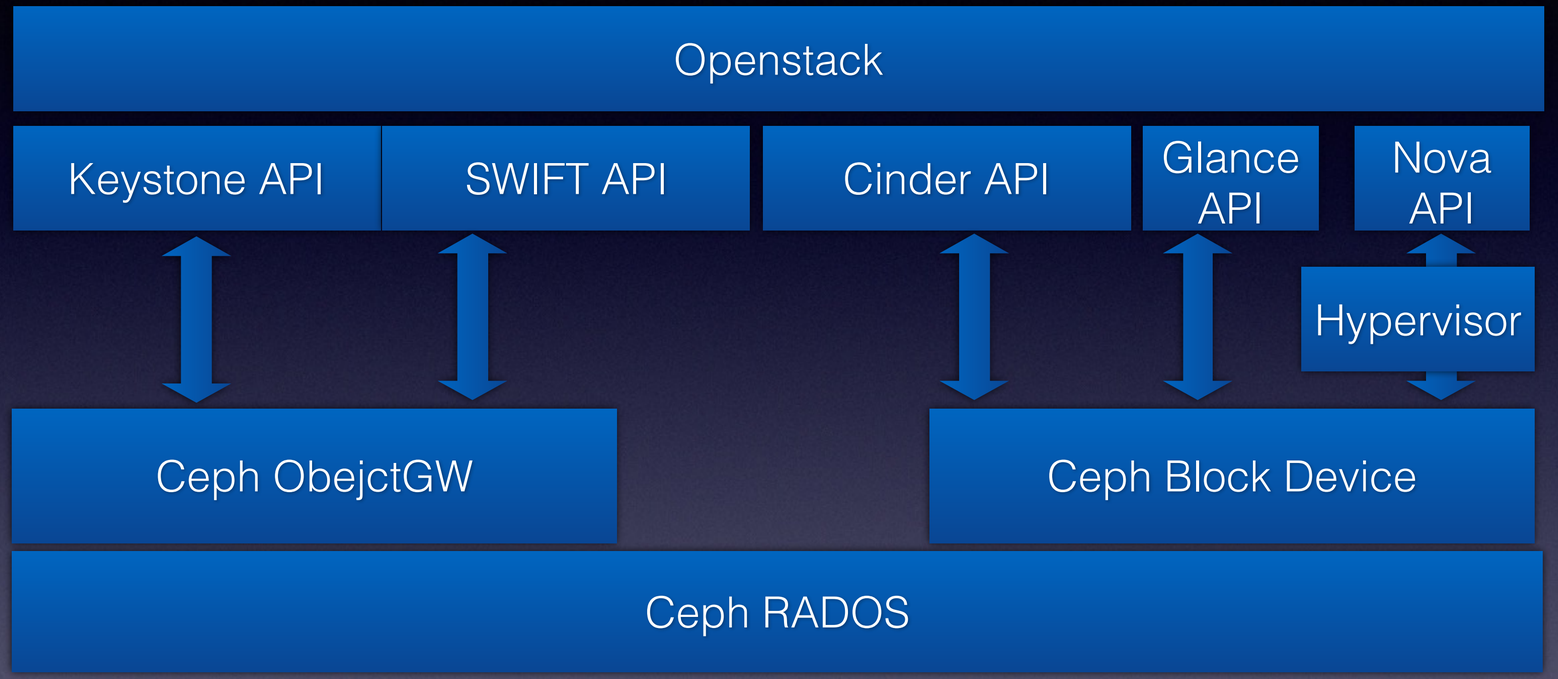
举个例子，高性能计算环境应用Ceph

弹性 可靠 可用



高性能计算环境应用Ceph

高性能存储环境从直连本地存储，到分布式，到可扩展统一存储



高性能环境的数据存储具有一定的特殊性，计算过程、计算结果、数据分析都对应不同的存储场景，Ceph在此场景中优势明显，但尚待商业的压力测试。

在该应用中，Ceph的开放性及API的粒度都有助于我们进行不同层面的定制化。

云存储领域中，不少企业都直接或间接的使用了Ceph或者Ceph的设计方法，在数据一致性和系统可靠性上，获得了一定的宝贵经验。

谢 谢