CHINA OpenInfra Days

CHINA OpenInfra Days | IT大咖说 知识共享平台

OPENSTACK基于CEPH的容灾备份

证通云　高级工程师

宋柏森

songbaisen999@qq.com

SZZT证通电子
SZZT ELECTRONICS

以互联网为基础平台、以安全支付为核心技术、以IDC云计算及智能终端为支撑的智慧城市基础设施服务商和应用方案提供商，引领智慧生活。

创立时间：1993年
上市时间：2007年（代码：002197）

注册资本：5.2亿
总资产：50亿

员工
2100人

证通云=IDC数据中心+云计算+高可信+大数据+人工智能

智慧城市/金融/政务/制造行业应用

人工智能 · 深入理解政务、金融、制造等业务需求，提供丰富的机器学习算法库

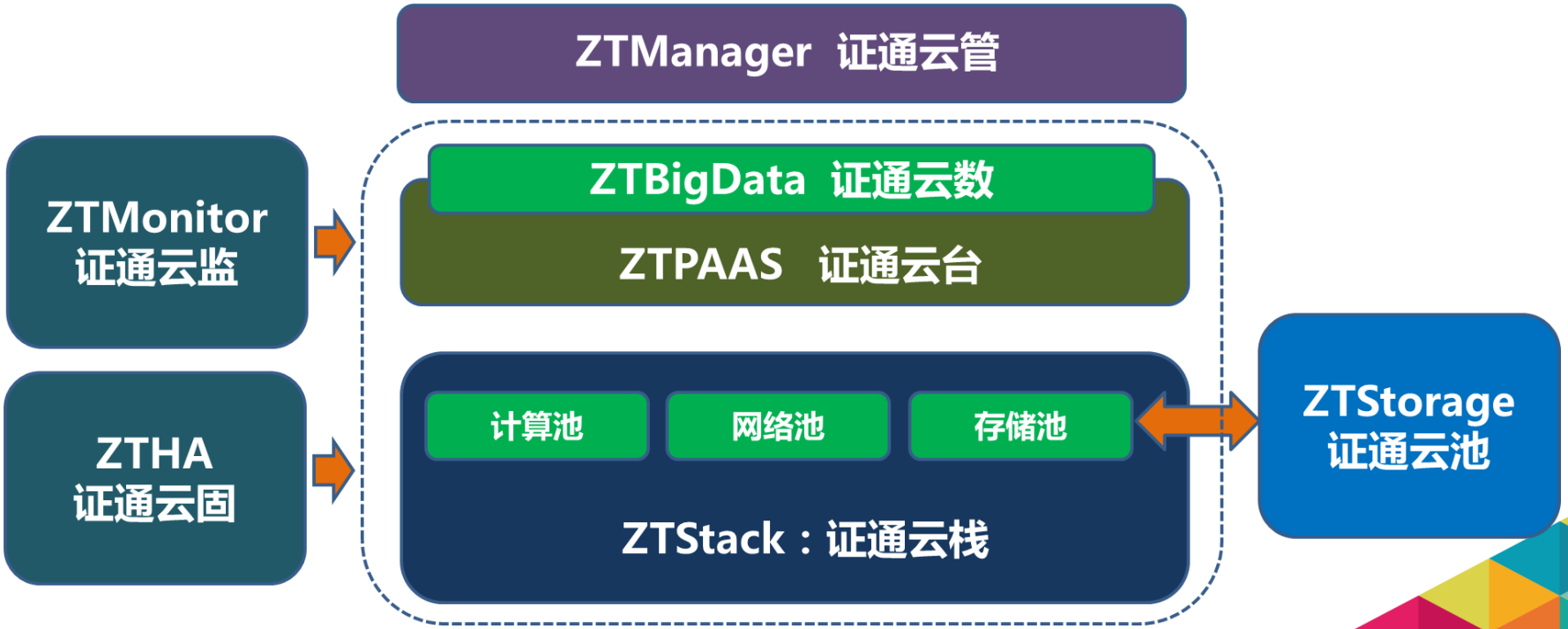
大数据 · 提供HDFS、Hbase、Hadoop、Spark、 NOSQL、 MapReduce以及大数据分析挖掘服务

高可信 · 拥有独立知识产权和专利，并通过可信云云主机服务认证

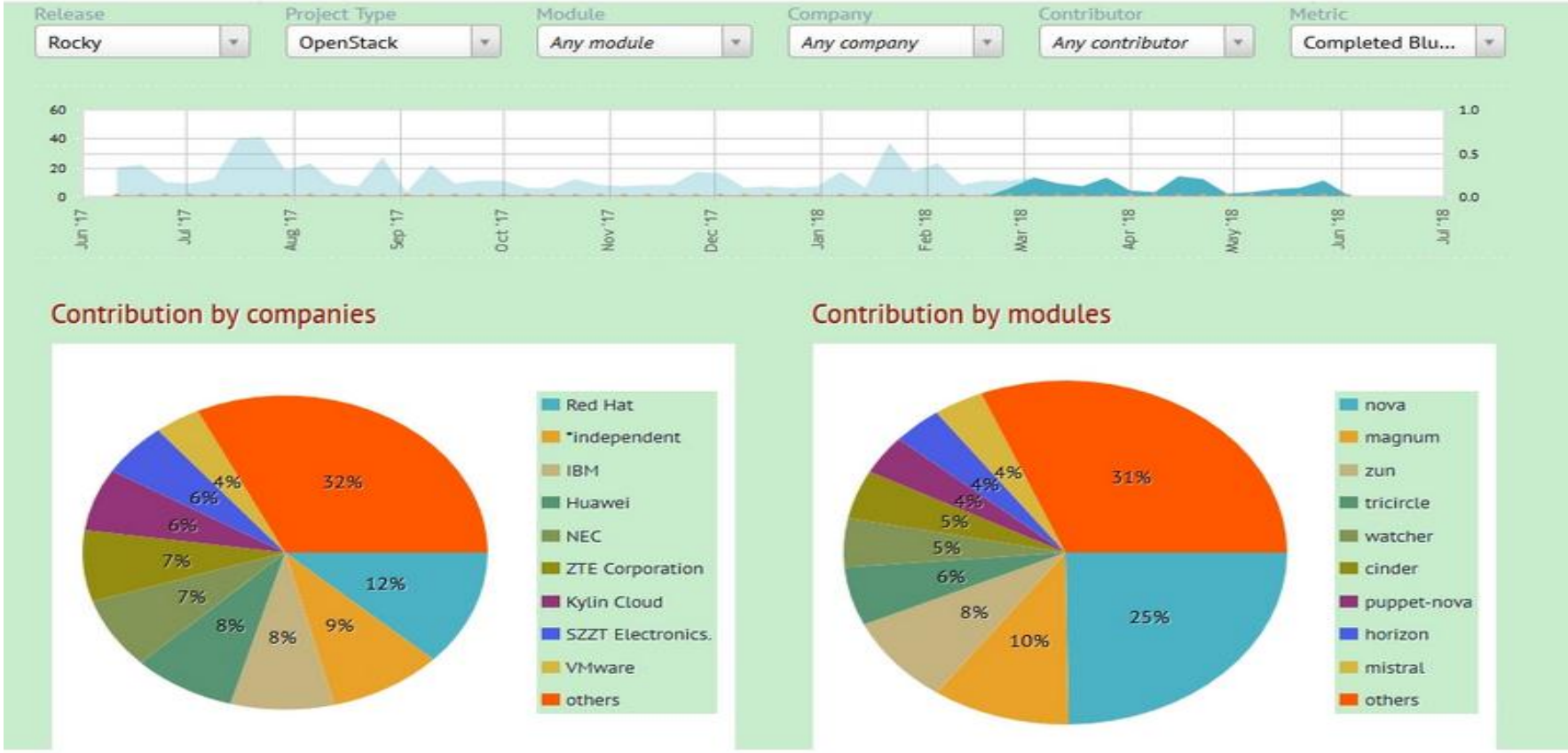
云计算 · 帮助客户实现从传统IT向云计算的转型，为客户提供完整的云计算解决方案

IDC数据中心 · 证通五大数据中心，可提供服务器托管、服务器租用、互联网接入、云计算、云存储及其他各类电信增值服务

证通云（ZTCloud）是云计算基础设施提供商，打造智慧城市平台服务。致力于为IDC数据中心，金融、政务等行业用户提供安全可靠、性能卓越、按需、实时的IAAS&PAAS平台。立足于证通电子IDC数据中心，提供云主机、云存储、云安全、云灾备等基础设施服务平台和行业解决方案。



拥抱开源


Release
Rocky
Project Type
OpenStack
Module
Any module
Company
Any company
Contributor
Any contributor
Metric
Completed Blu...
60
40
20
0
1.0
0.5
0.0
Jun '17
Jul '17
Aug '17
Sep '17
Oct '17
Nov '17
Dec '17
Jan '18
Feb '18
Mar '18
Apr '18
May '18
Jun '18
Jul '18
Contribution by companies
32%
12%
9%
8%
8%
7%
7%
6%
6%
4%
Red Hat
*independent
IBM
Huawei
NEC
ZTE Corporation
Kylin Cloud
SZZT Electronics.
VMware
others
Contribution by modules
31%
25%
10%
8%
6%
5%
5%
4%
4%
4%
nova
magnum
zun
tricircle
watcher
cinder
puppet-nova
horizon
mistral
others

五大数据中心

目前已拥有东莞石碣云数据中心、东莞旗峰云数据中心、广州南沙云谷数据中心、深圳光明数据中心和长沙证通云谷数据中心等五大数据中心。总机架规模达到1.6万个，在行业内排名Top3之内。

长沙证通云谷数据中心（3500个机柜）

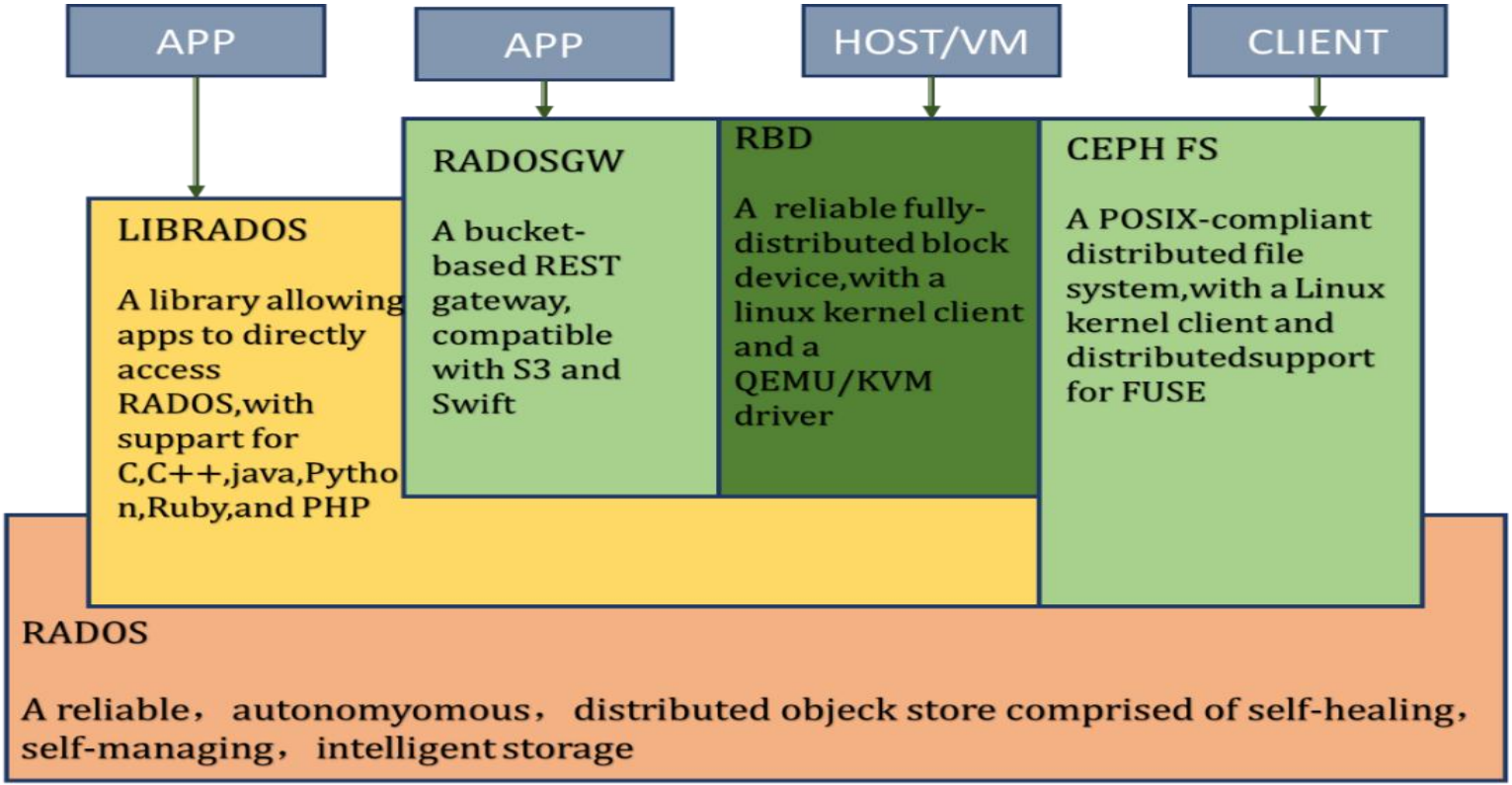
广州南沙云谷数据中心（3680个机柜）

东莞旗峰云数据中心（3400个机柜）

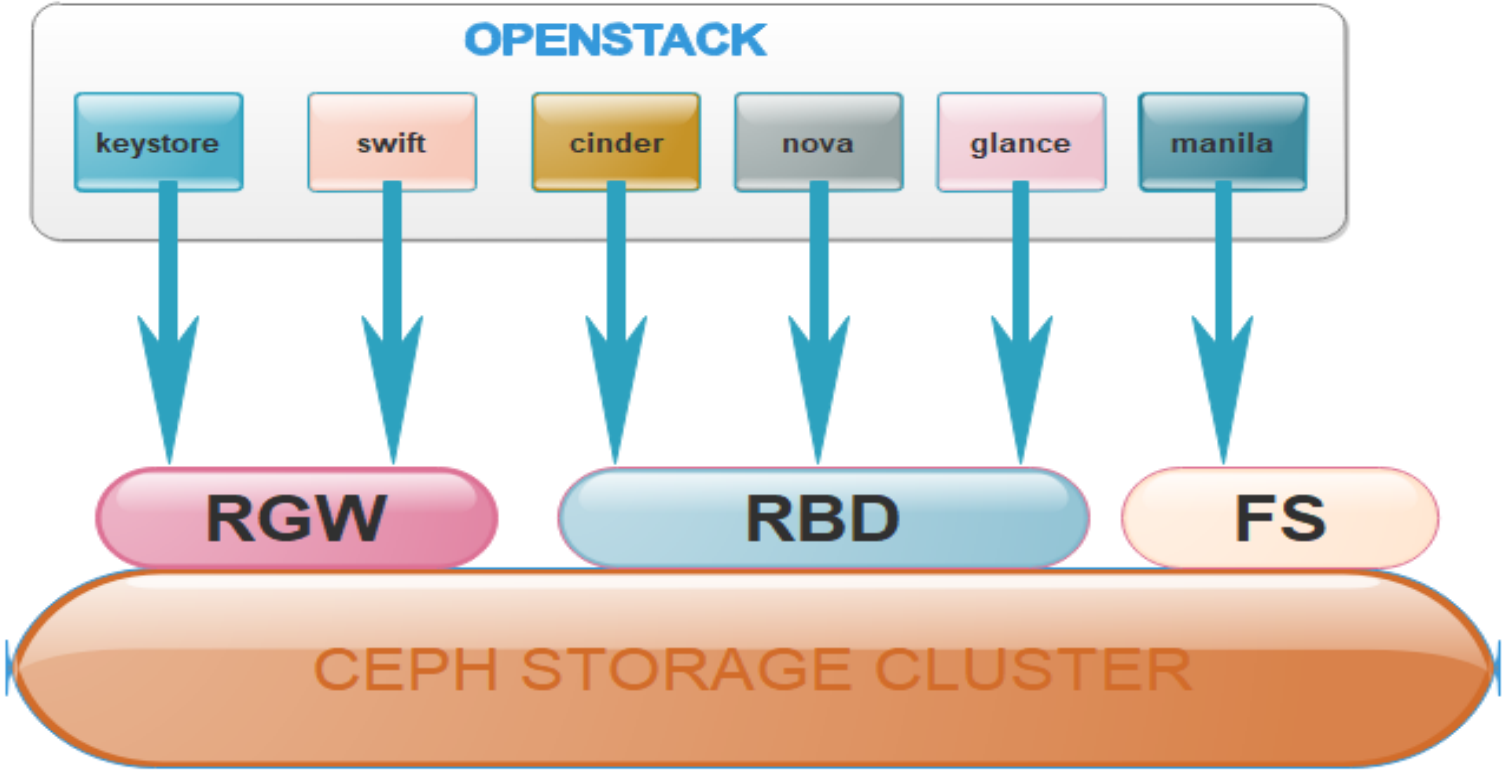
东莞石碣云数据中心（900个机柜）

深圳光明数据中心（3400个机柜）

CEPH 简介


APP
APP
HOST/VM
CLIENT
LIBRADOS
A library allowing apps to directly access RADOS,with suppart for C,C++,java,Python,Ruby,and PHP
RADOSGW
A bucket-based REST gateway, compatible with S3 and Swift
RBD
A reliable fully-distributed block device,with a linux kernel client and a QEMU/KVM driver
CEPH FS
A POSIX-compliant distributed file system,with a Linux kernel client and distributedsupport for FUSE
RADOS
A reliable， autonomyomous， distributed objeck store comprised of self-healing，self-managing， intelligent storage

OPENSTACK与CEPH



ceph为openstack后端存储的配置

- 主机的系统盘，最好支持双盘raid1。
- 每个主机的硬盘转速和配置最好一致，防止木桶效应。
- 主机的电源要支持冗余
- 前后端万兆网络
- SSD或者NVME做日志盘加速
- 网络端口要做bond
- 超融合环境要做绑核
- 主机硬盘支持热插拔
- 交换机做堆叠
- MON上SSD盘

ceph的灵魂crush算法

- 伪随机算法，任意节点相同的输入得到相同的输出。
- 算法稳定性，目前的抽签2算法可以保证，扩容节点和添加节点，尽量只影响相关的节点数据迁移。
- 可以保证数据按照硬盘的大小近似均衡的分布。
- 可以自定义数据的分配规则，指定数据存储的方向。
- 无中心节点，不需要查表，只需少量元数据OSD MAP和CRUSH MAP即可计算出数据的存储位置。
- 可以自定义数据的安全级别提升数据的安全性。

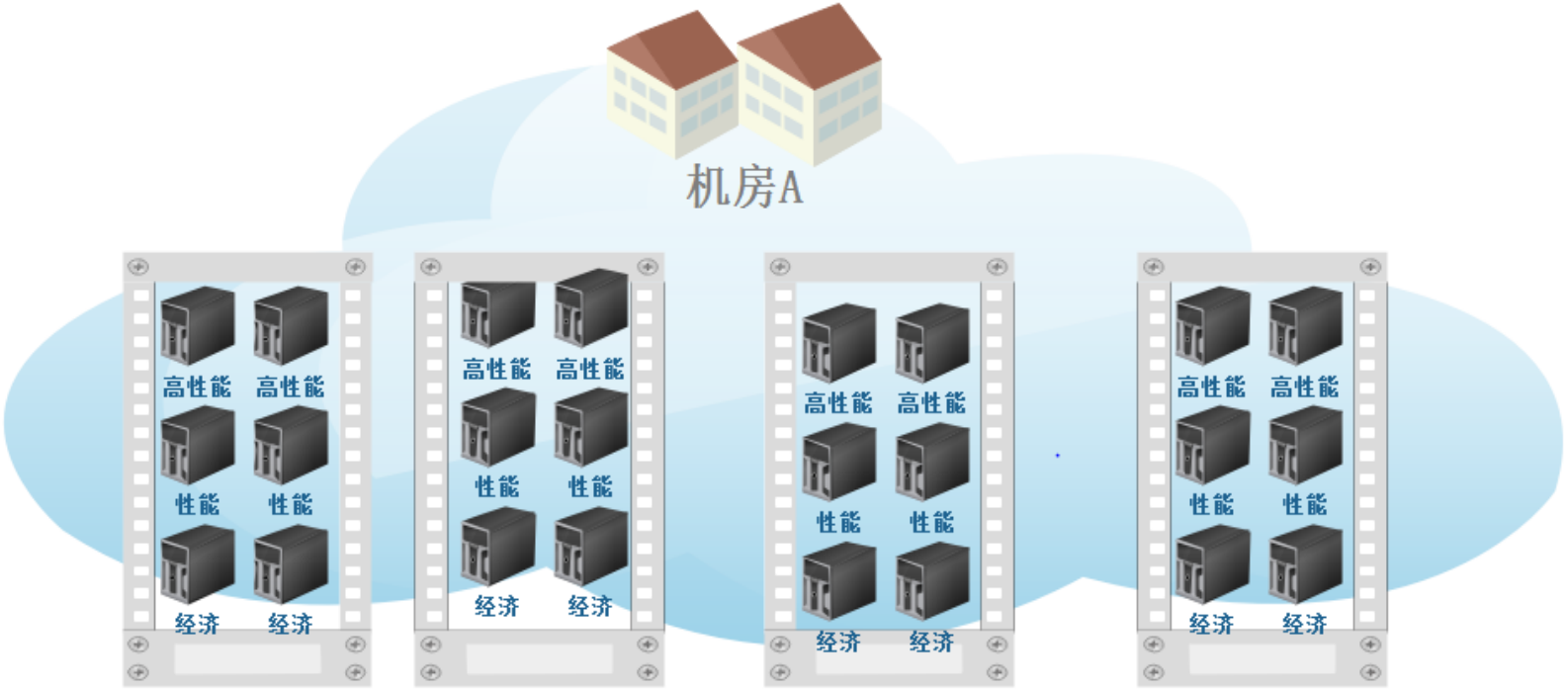
利用CRUSH提升数据的安全性

- CEPH默认是按照主机级别进行容灾

- 大的集群可以通过修改CRUSH配置可以提升安全级别到机架级别

- 更大的集群配置副本分散到机房的级别

ceph+cinder多后端打造不同存储性能资源池



数据迁移粒度，是否该迁移数据？

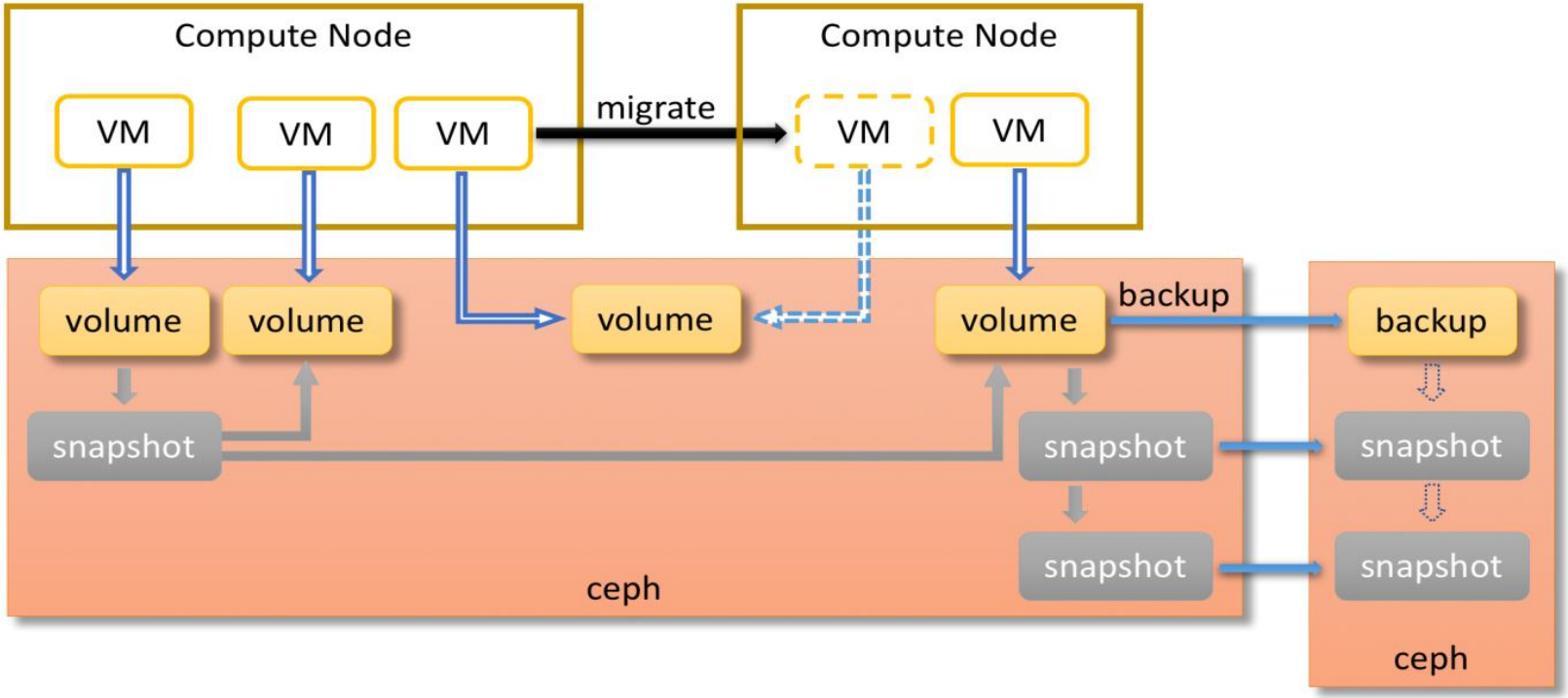
mon_osd_down_out_subtree_limit

```
root@controller170:~# ceph daemon /var/run/ceph/ceph-mon.controller170.asok config show|grep mon_osd_down_out_subtree_limit
    "mon_osd_down_out_subtree_limit": "rack",
```

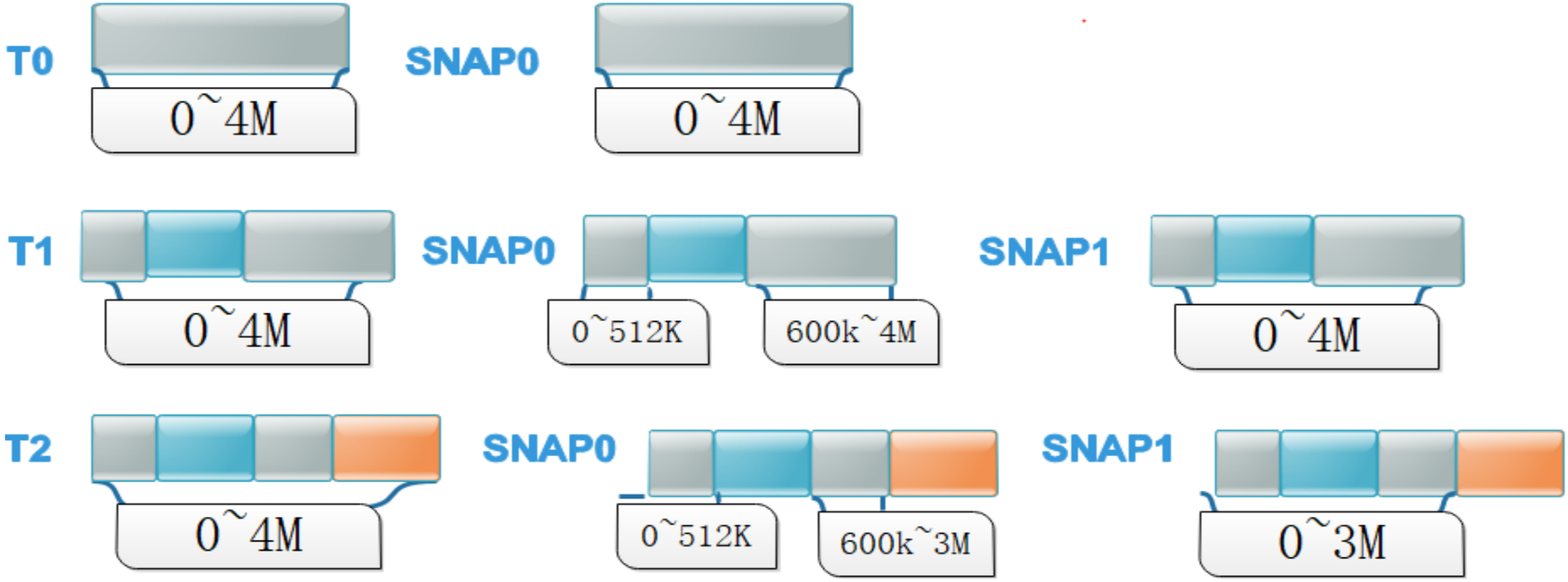
修改CRUSH后，使用虚拟主机，系统重新上电后，是否该重新加入主机？

```
root@ceph1:/home/share/dist# ceph daemon /var/run/ceph/ceph-mon.ceph1.asok config show|grep osd_crush_update_on_start
    "osd_crush_update_on_start": "true",
```

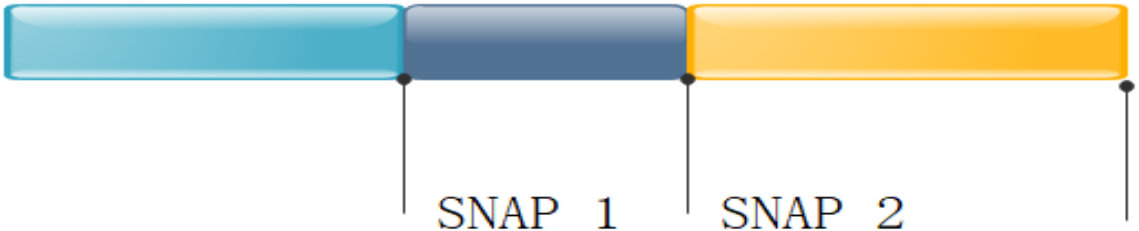
ceph+cinder backup



ceph rbd快照COW



基于CEPH快照的增量备份



rbd export -diff rbd/image@snap1 rbd/image

rbd export -diff rbd/image@snap2 rbd/image

rbd export -diff rbd/image@snap3 rbd/image@snap2

ceph基于cinder backup备份

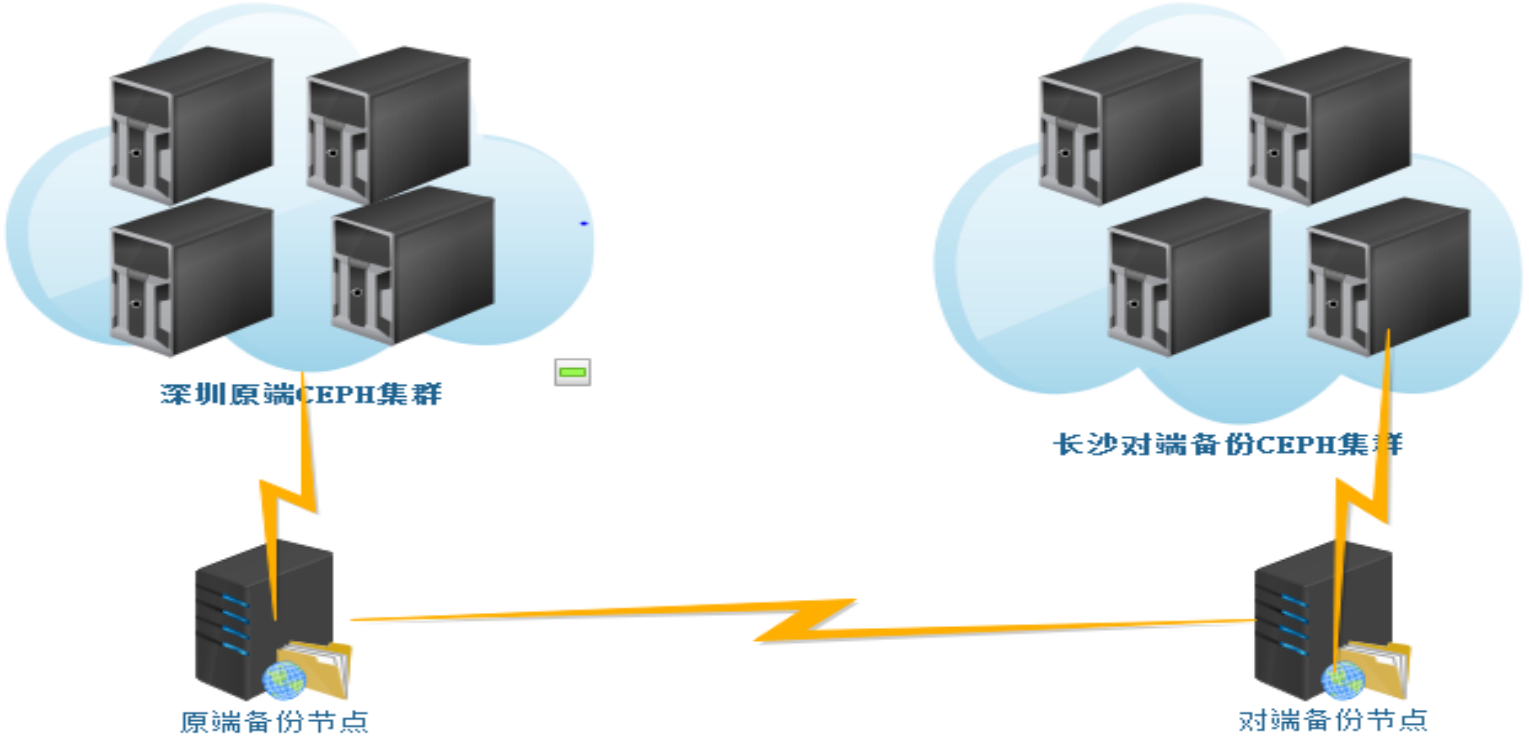
优点：

- 与openstack相融合，此备份是openstack可以感知的。

缺点：

- 需要管理手动触发，备份实时性不强。

基于CEPH的异地冷备方案



异地冷备系统的设计

- 数据的传输采用sftp加密传输的方式，保证数据在传输过程中的安全性。卷在导入和导出的时候进行MD5的检测，MD5不一致的重新备份。数据在传输前使用压缩算法中压缩比较高和时间最少的gzip算法，保证网络传输中数据的最少以及压缩时间最短。
- 保存在备份节点的卷，如果在原端证通云池集群中已经不存在了，说明已经在原端证通云池集群中被管理员删除掉了，这个时候在备份节点将这个卷放入垃圾回收池，并打上进入垃圾回收池的时间。为了节省空间考虑，垃圾回收池只使用一副本，其它的默认使用三副本，如果备份节点条件有限，也可以使用两副本。在垃圾回收池中的卷，默认保存两周，两周后彻底清理掉，这个保存时间，用户可以通过配置文件进行修改。
- 每次在备份成功后，删除前一次备份的打的快照信息，保持整个集群的快照数目，不会无限增加。
- 考虑到进行备份的节点，也可能因为掉电损坏导致无法使用。因此，配备节点数据都保持到CEPH集群中，这样，当备份节点损坏，而无法启动时，可以通过任意节点，启动服务，通过集群来获取之前的配置数据，继续进行服务。
- 用户可以配置备份的开始和结束时间，当在备份的时间段内，集群有recovering或者backfill，以及unclean等非健康状态时停止备份。

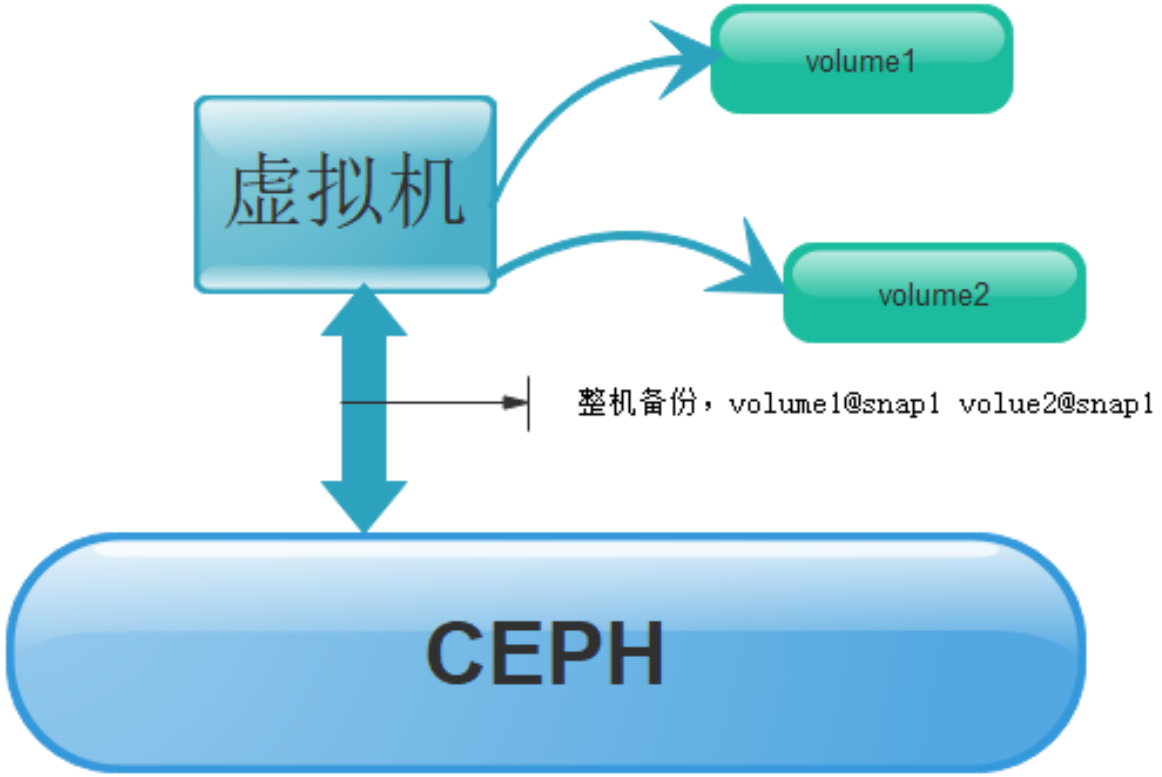
rbd export v2

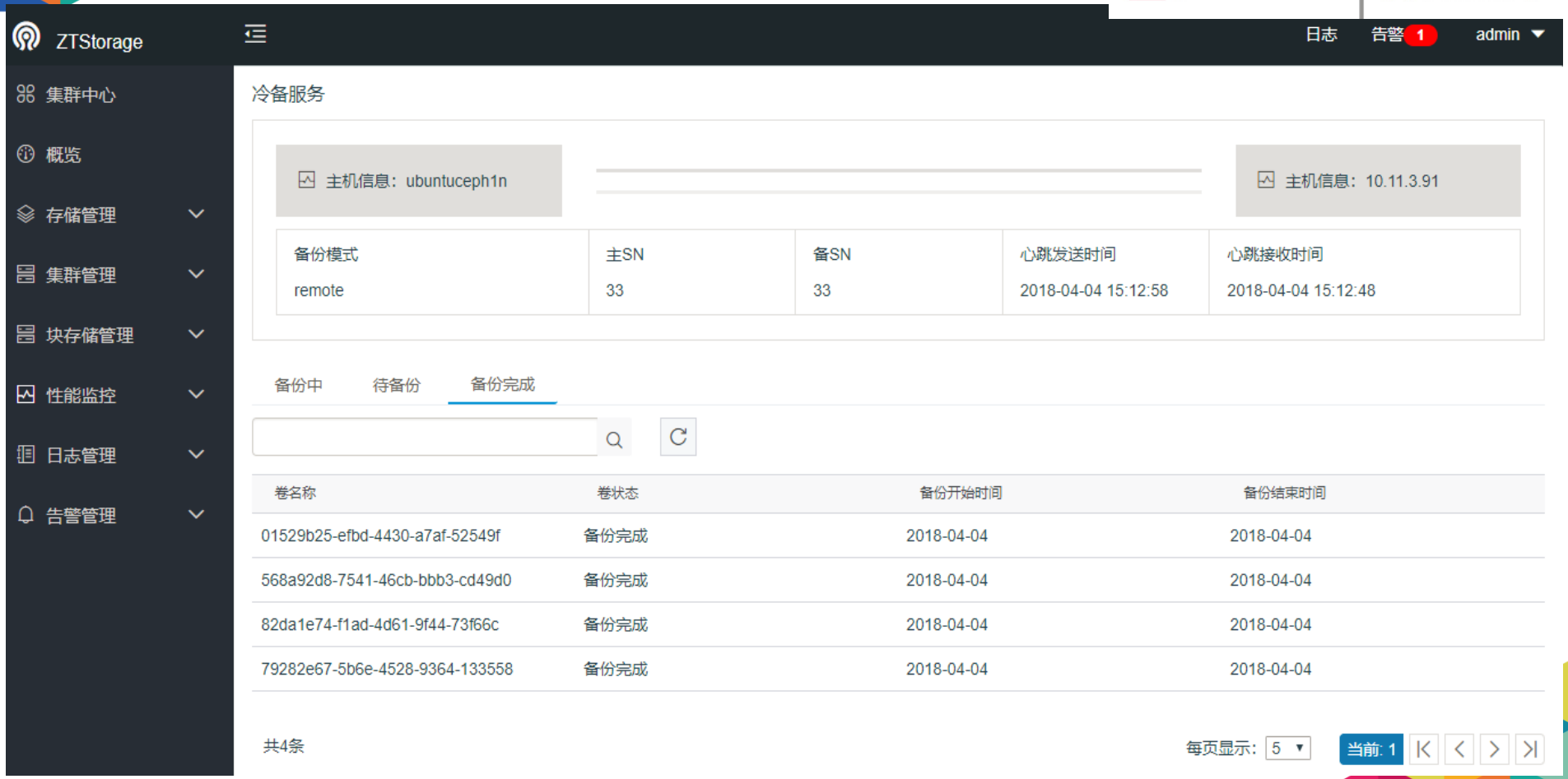
- 对于jewel版本进行卷的导入和导出，只是对原卷进行读取，之后将原卷写入一个文件。并没有将卷的一些重要的元数据导出，例如卷开启的特性，卷创建的快照信息等。

- luminous版本rbd的export提供了--export-format 2,可以将卷的元数据信息导出。

一致性快照

- 一致性快照用于整机备份，一个虚拟机通常挂载了多个卷，对虚拟机做整机备份时所有卷快照的应处于同一时间点，才能保证数据恢复的可靠性；
- 对CEPH添加支持一致性快照的能力， 对上层发起的一致性快照请求会尽量保证多个卷的快照属于同一个时间点；
- CEPH在CINDER挂载和解挂的时候，加入一致性快照组和离开一致性快照组，一致性快照组的多个卷执行pause IO后，再更新快照信息操作，待组内所有的卷都创建完快照，解除pause IO,尽量保证了多个卷快照时间点的一致性。


volume1
虚拟机
volume2
整机备份，volume1@snap1 volue2@snap1
CEPH


ZTStorage
日志
告警 1
admin
集群中心
概览
存储管理
集群管理
块存储管理
性能监控
日志管理
告警管理
冷备服务
主机信息：ubuntuceph1n
主机信息：10.11.3.91
备份模式
remote
主SN
33
备SN
33
心跳发送时间
2018-04-04 15:12:58
心跳接收时间
2018-04-04 15:12:48
备份中
待备份
备份完成
卷名称
卷状态
备份开始时间
备份结束时间
01529b25-efbd-4430-a7af-52549f
备份完成
2018-04-04
2018-04-04
568a92d8-7541-46cb-bbb3-cd49d0
备份完成
2018-04-04
2018-04-04
82da1e74-f1ad-4d61-9f44-73f66c
备份完成
2018-04-04
2018-04-04
79282e67-5b6e-4528-9364-133558
备份完成
2018-04-04
2018-04-04
共4条
每页显示：5
当前: 1

基于ceph rbd的增量备份

优点：

- 当两个CEPH集群网络不通时，可以通过此方式进行备份。
- 通过设计高可用的冷备系统，满足异地冷备份的需要。

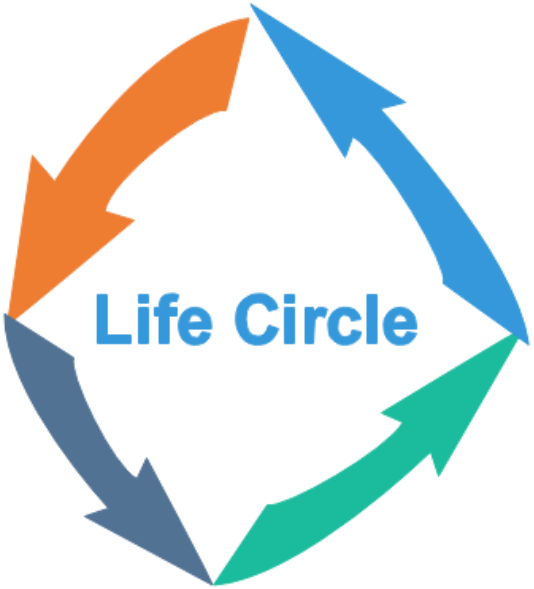
缺点：

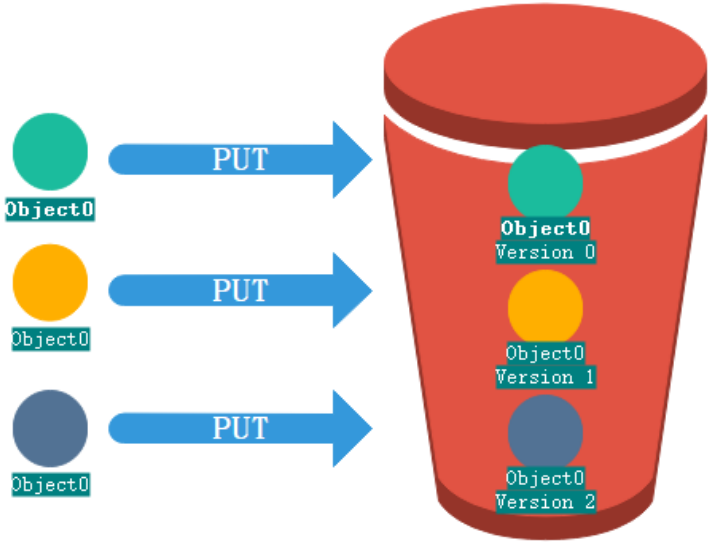
- 由于备份不是通过openstack的，恢复需要人工参与，尤其在openstack测发生误删除的情况。
- 备份实时性不强。
- 因为创建的快照不是经过openstack，导致备份的卷无法在openstack侧删除。

S3/SWIFT 对象存储备份的优势

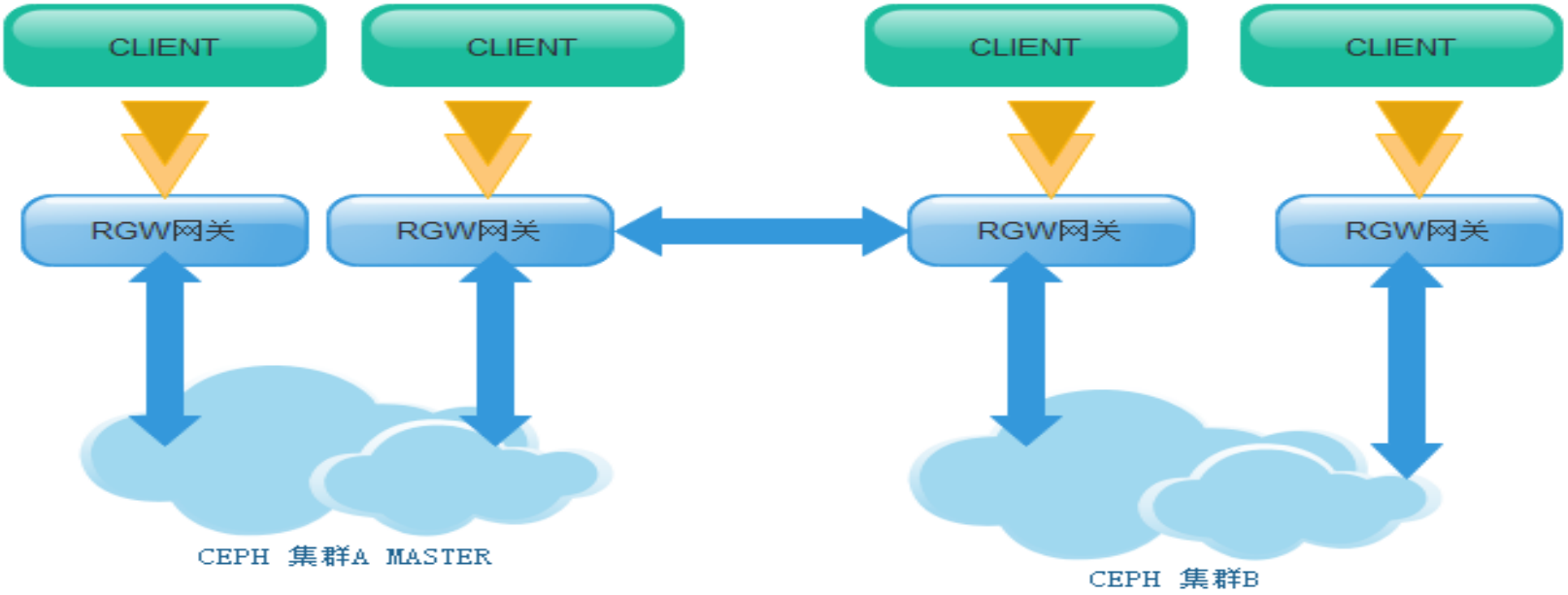
- 万物皆对象，满足海量非结构化数据存储。
- 所有对象都在一个扁平的存储空间，没有目录树结构。
- 所有的对象都保存在桶中。用户可以创建不同的桶，设置不同的访问权限。
- 丰富的API和接口支持，RESTAPI、JAVA、PYTHON等。
- CEPH S3/SWIFT本身还支持多活容灾，增强数据的安全性。
- 主流的公有云服务商都提供了对象存储的API，可以将数据备份到公有云上。利用公有云的安全性和稳定性。

S3/SWIFT的优势

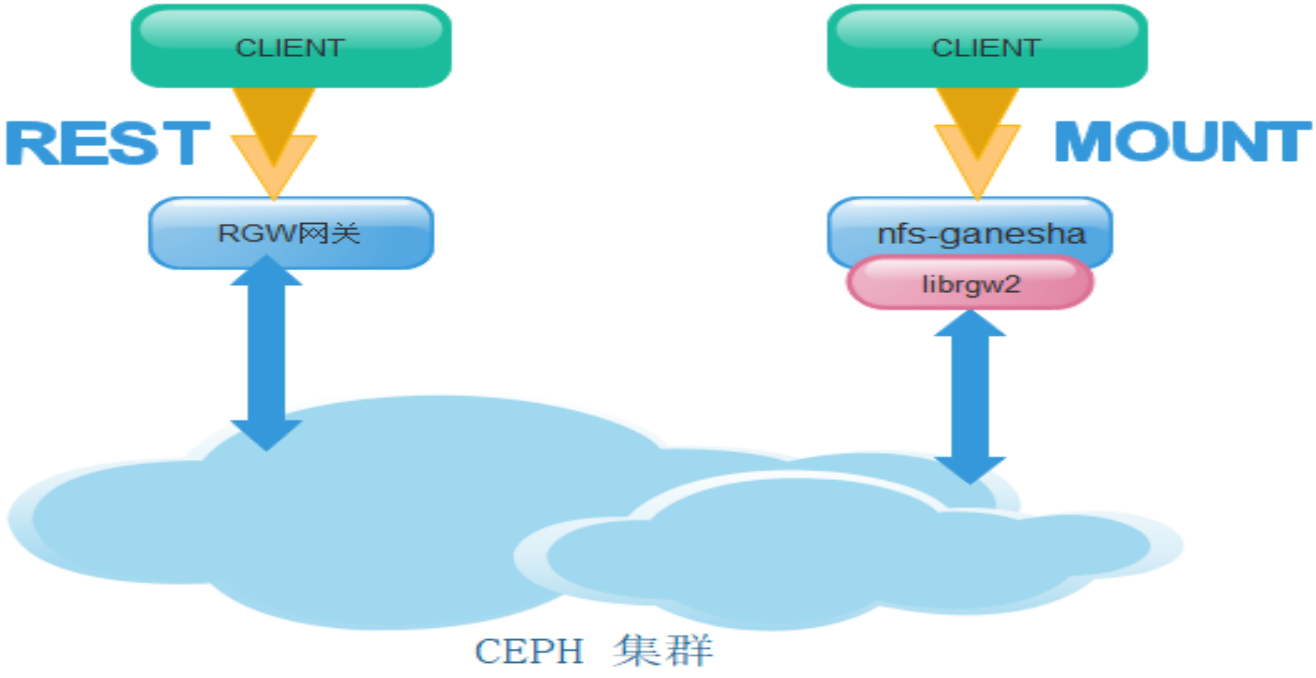

Life Circle


Object0
PUT
Object0
PUT
Object0
PUT
Object0
Version 0
Object0
Version 1
Object0
Version 2

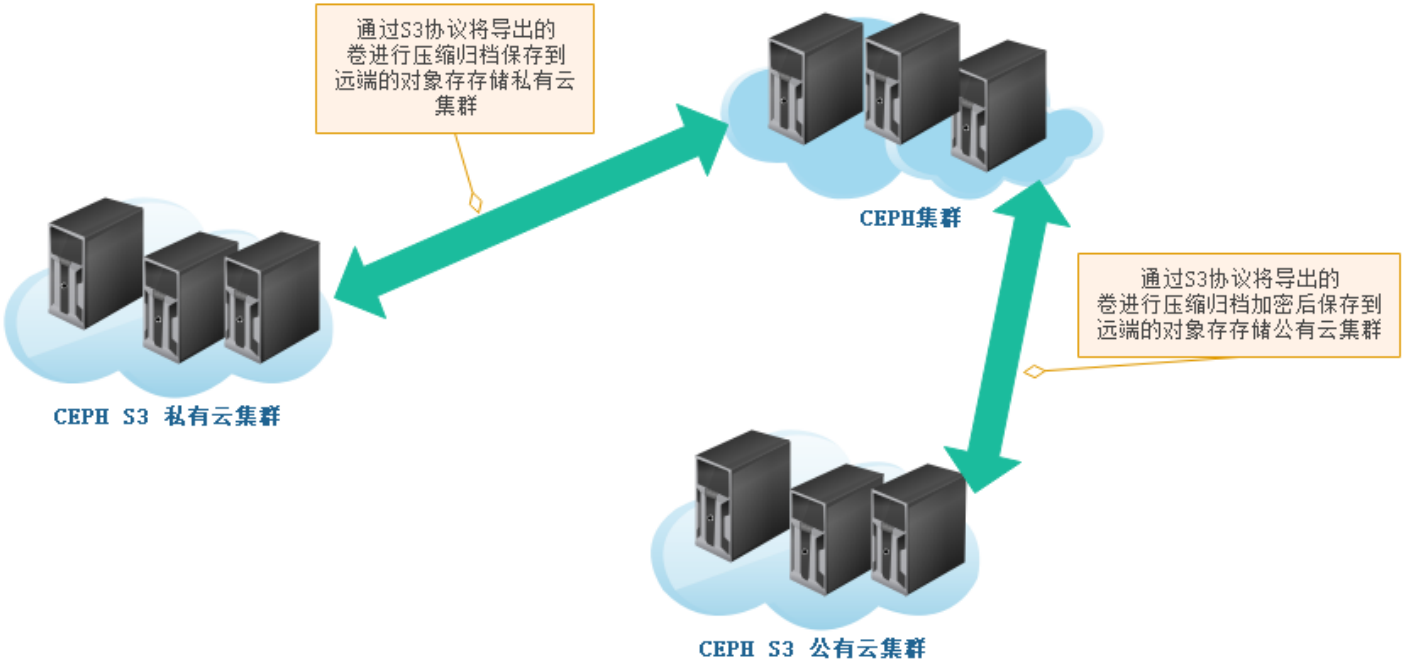
S3/SWIFT的优势


CLIENT
CLIENT
CLIENT
CLIENT
RGW网关
RGW网关
RGW网关
RGW网关
CEPH 集群A MASTER
CEPH 集群B

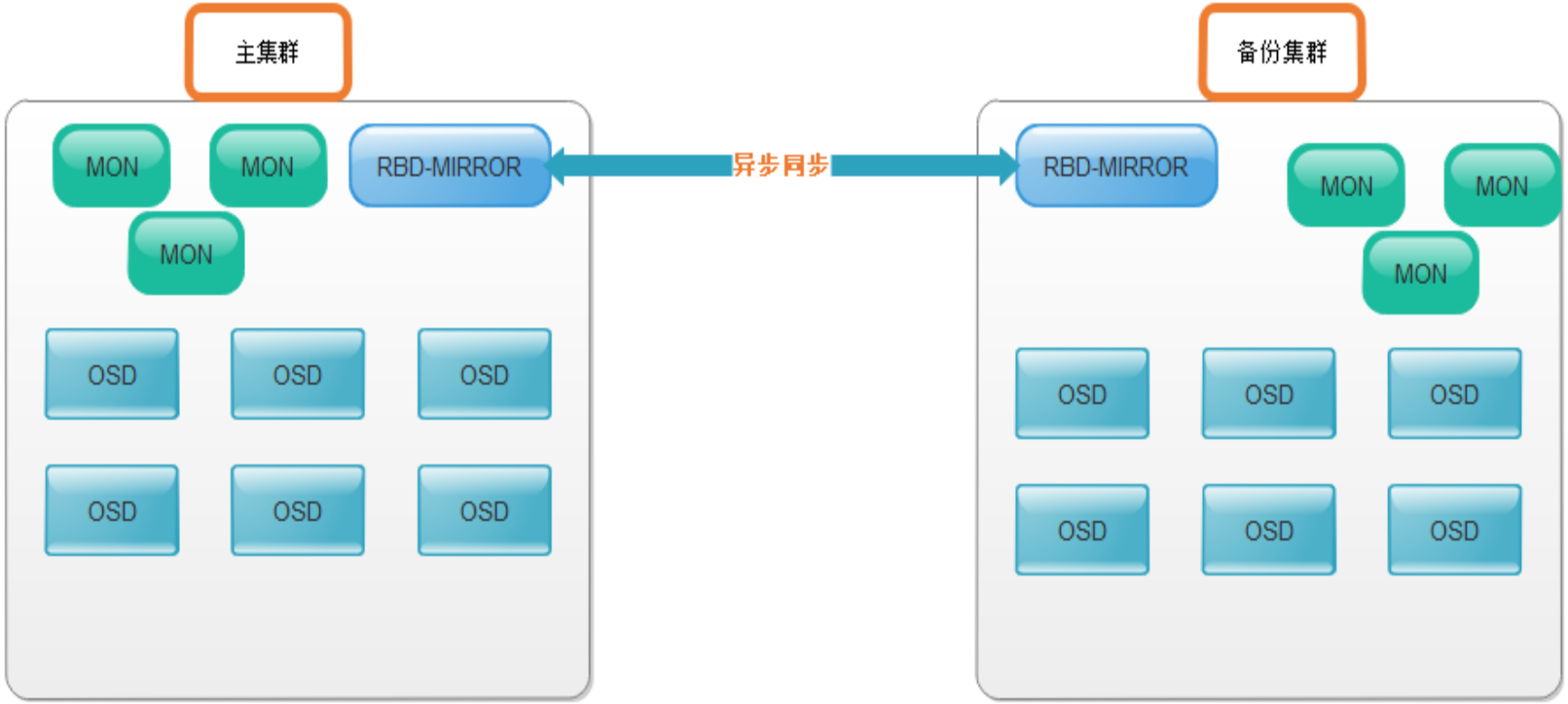
S3/SWIFT的优势



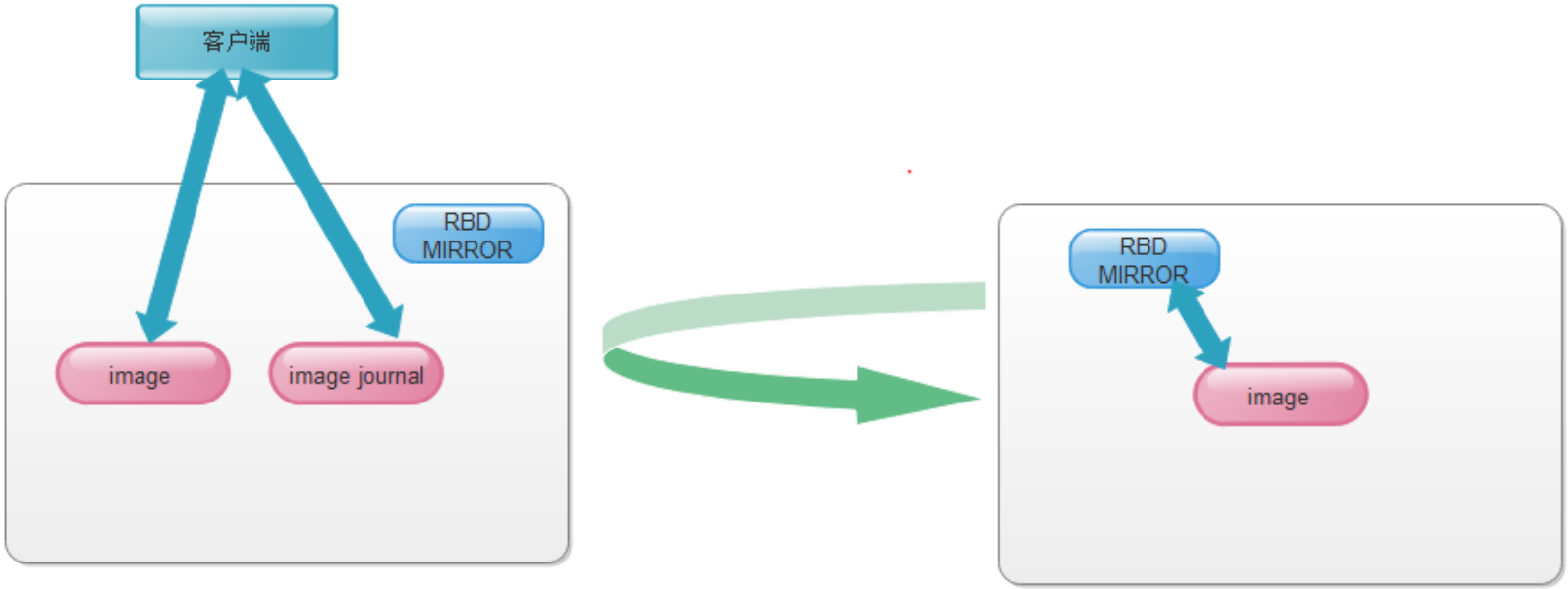
CEPH利用S3/SWIFT进行备份归档



rbd mirror


主集群
MON
MON
MON
RBD-MIRROR
OSD
OSD
OSD
OSD
OSD
OSD
异步同步
备份集群
RBD-MIRROR
MON
MON
MON
OSD
OSD
OSD
OSD
OSD
OSD

rbd mirror基于journal实现的备份



rbd mirror

- 目前支持两种模式的备份方式，一个是POOL模式，另一个是image模式。
- 可以手动升级或者降级一个卷的模式，可以强制远端的安装本地的卷进行还原。
- 对于脑裂需要用户自己手动解决，判断那个卷是正确的。
- L版本之前，只能运行一个mirror进行，L版本后可以运行多个mirror进程。

改进的rbd mirror

- 对于两个集群间为万兆或者更高网络的情况下，修改备份方式从拉到推的方式，保证强一致性。
- 对于主集群删除的卷，备份集群不删除，使用trash，备份mirror加回收线程，两周后清理掉。
- 多mirror进行备份的情况，添加新的策略，基于卷的大小，从大到小顺序分配，提升备份效率。

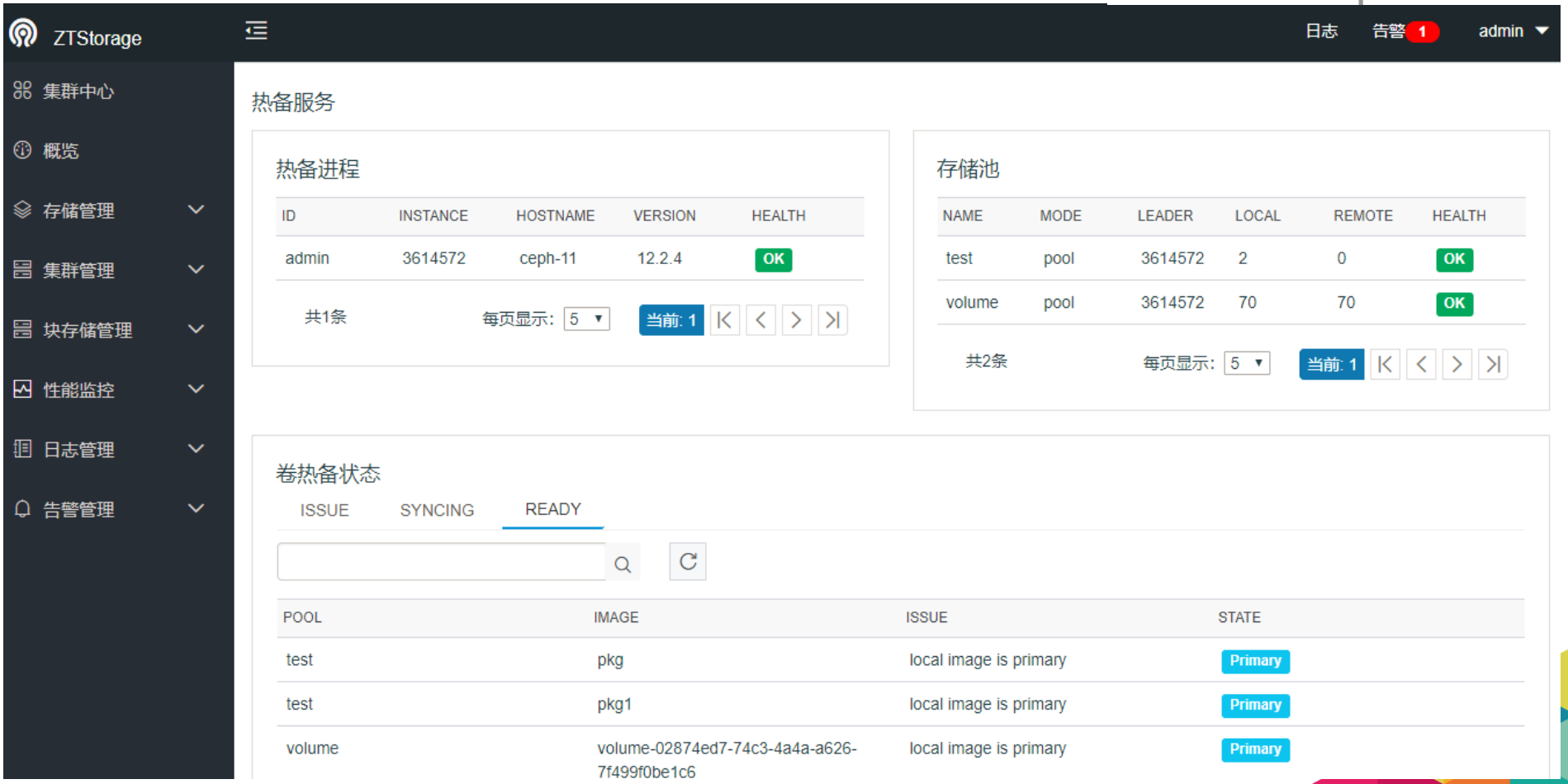
RBD MIRROR备份方式

优点：

- 可以方便的在远端建立一个影子集群，并且发生故障时，可以通过命令自动进行恢复。
- 对网络要求不高，在网络延迟的情况下，也可以正常工作。
- 社区的DASHBOARD支持，WEB API接口支持。

缺点：

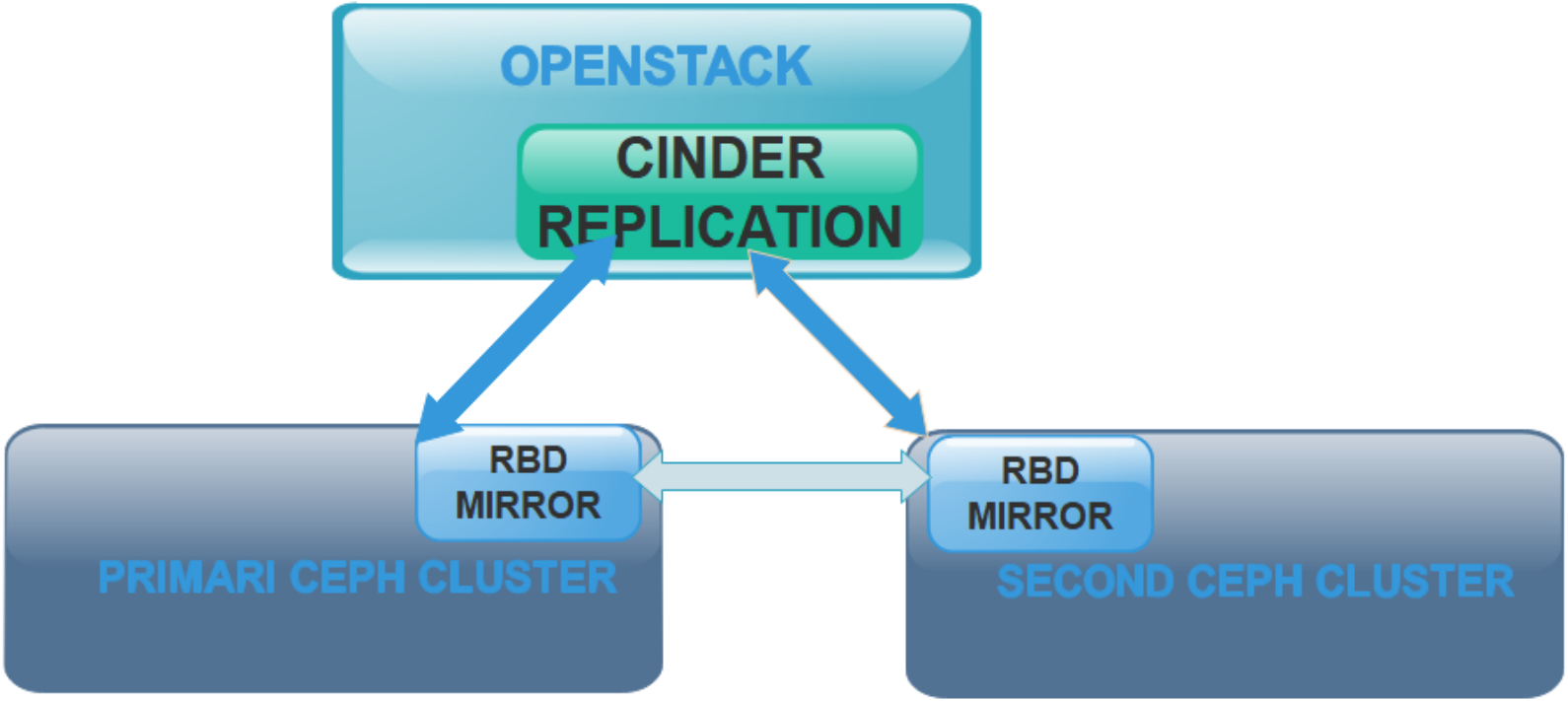
- 需要两个CEPH集群的网络互通。
- 原生的mirror为异步备份。


ZTStorage
日志
告警 1
admin
集群中心
概览
存储管理
集群管理
块存储管理
性能监控
日志管理
告警管理
热备服务
热备进程
ID INSTANCE HOSTNAME VERSION HEALTH
admin 3614572 ceph-11 12.2.4 OK
共1条
每页显示: 5
当前: 1
存储池
NAME MODE LEADER LOCAL REMOTE HEALTH
test pool 3614572 2 0 OK
volume pool 3614572 70 70 OK
共2条
每页显示: 5
当前: 1
卷热备状态
ISSUE SYNCING READY
POOL IMAGE ISSUE STATE
test pkg local image is primary Primary
test pkg1 local image is primary Primary
volume volume-02874ed7-74c3-4a4a-a626-7f499f0be1c6 local image is primary Primary

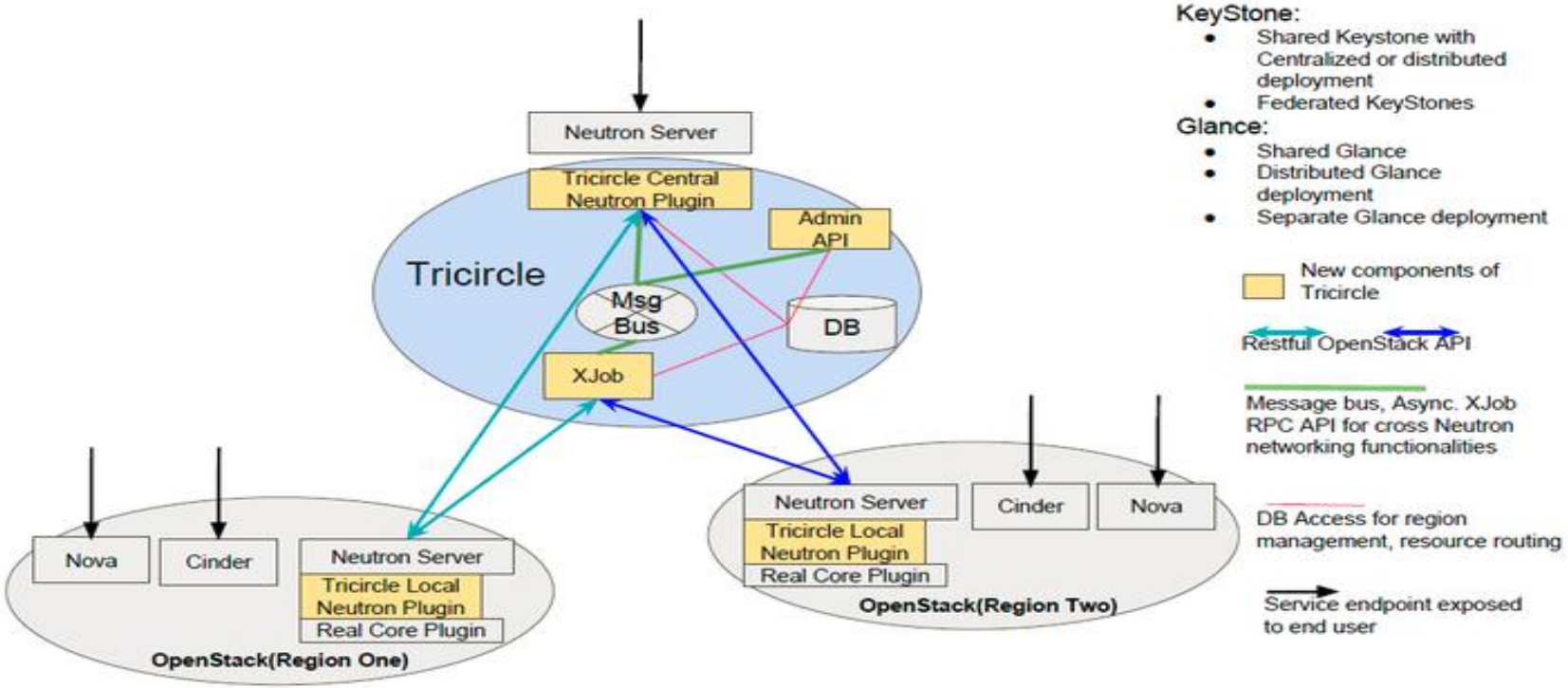
CINDER REPLICATION

- cinder replication是社区为了解决卷的异地灾备切换而开发的功能。目前很多厂商的driver已经支持了该功能。
- ceph的rbd mirror正好与此功能需求相匹配，并且已完了相关的driver开发。
- 此功能推出时间不久，目前主要在实测阶段。

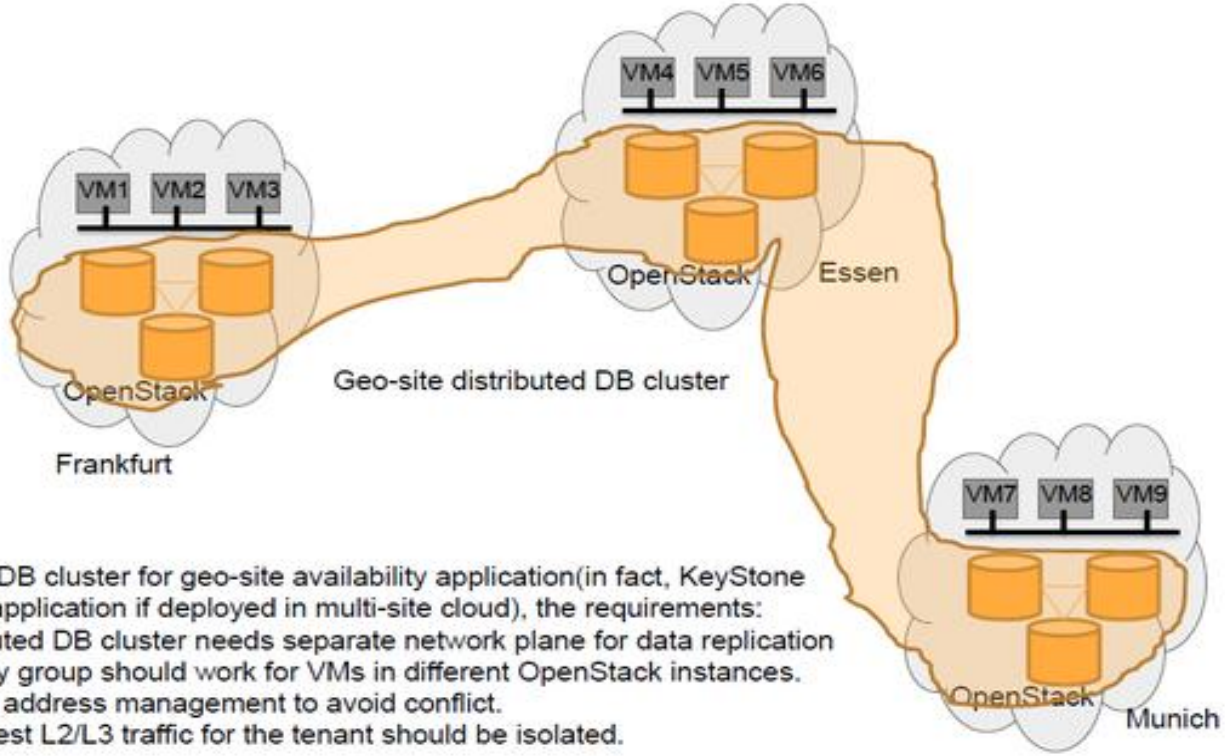
CINDER REPLICATION+RBD MIRROR



Try tricircle?



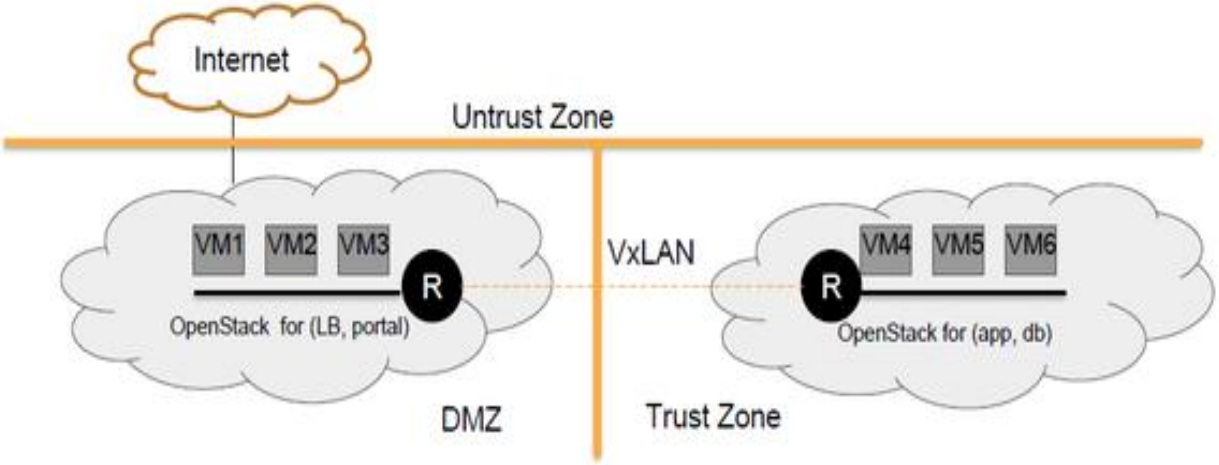
引用tricircle wiki:https://wiki.openstack.org/wiki/Tricircle



Distributed DB cluster for geo-site availability application(in fact, KeyStone the typical application if deployed in multi-site cloud), the requirements:

1. Distributed DB cluster needs separate network plane for data replication
2. Security group should work for VMs in different OpenStack instances.
3. IP/Mac address management to avoid conflict.
4. East-west L2/L3 traffic for the tenant should be isolated.

引用tricircle wiki:https://wiki.openstack.org/wiki/Tricircle



Financial application has different requirements from security aspect, separate OpenStack instance in different zone, tenant level networking automation requirements:

1. East-west L3 traffic for the tenant should be isolated, use VxLAN to connect the routers in different OpenStack instance.
2. Security group should work for VMs in different OpenStack instances.
3. IP/Mac address management to avoid conflict.

引用tricircle wiki:https://wiki.openstack.org/wiki/Tricircle

- 1 通过TRICIRCLE多云级联技术，实现多个OPENSTACK进行级联，对于OPENSTACK多云，进行统一的资源管理和调度。通过插件的方式融入，可以无缝对接开源的OPENSTACK版本，便于后面的推广和部署运维。
- 2 TRICIRCLE多云的分布式网络QOS策略，实现多云间的OPENSTACK的分布式网络QOS的实现。由于整体的网络带宽有限，通过实现多云间的分布式网络QOS可以满足和提升用户的云服务质量。
- 3 多云间的TRICIRCLE的资源的统一管理和分配，保证多云间的OPENSTACK的资源分配的唯一性和可复用性。例如，保证多云间的OPENSTACK的虚机分配的IP或者MAC都是唯一的，并且保证虚机释放资源后，任意其它云的OPENSTACK的虚机可以复用这些资源。
- 4 多云间的OPENSTACK的SFC（Service Function Chain）实现，基于软件定义网络，可以通过SFC使得网络数据不再通过IP地址来传输，而是通过特定的路径来动态的建立服务链，满足不同的租户可以按照不同的顺序向不同的服务模块发送数据。

CHINA
OpenInfra Days

Thank You !