SUSE Enterprise Storage Introduction

February 2019

Arthur Yang
Sales Engineer

Ceph Community

2004年

Ceph项目开始，提交了第一行代码

2010年

Linus Torvalds将Ceph client合并到内核中，

使Linux与Ceph磨合度更高

2012年

拥抱OpenStack，

进入 Cinder项目成为重要的存储驱动

2014年

Ceph乘上了OpenStack的春风，各大厂商积极参与其中

同时Inktank公司被RedHat公司1.75亿美金收购

2015年

成立Ceph顾问委员会，成员包括SUSE

Canonical、CERN、Cisco、Fujitsu、Intel、SanDisk

Ceph社区活跃度

Commits by Contributor

Show 10 entries Search

#	Contributor	Commits
1	Sage Weil	7457
2	Sage Weil	4563
3	Sage Weil	3708
4	Kefu Chai	2285
5	Jason Dillaman	2164
6	Yehuda Sadeh	1969
7	Yehuda Sadeh	1670
8	John Spray	1653
9	Danny Al-Gaaf	1633
10	Josh Durgin	1563

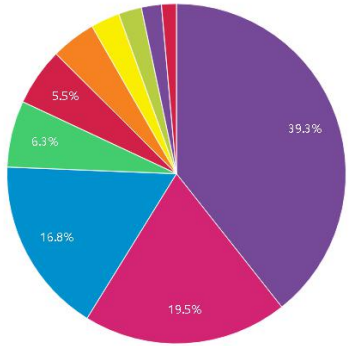
Showing 1 to 10 of 1,198 entries Previous Next

Commits by Company

Show 10 entries Search

#	Company	Commits
1	Red Hat	26405
2	Inktank	13097
	*independent	11274
3	SUSE	3692
4	DreamHost	2861
5	ZTE Corporation	1859
6	Intel	1479
7	Mirantis	1283
8	XSky	949
9	Deutsche Telekom	899

Showing 1 to 10 of 95 entries Previous Next



SUSE Contributor

Commits by Contributor

Show 10 entries Search

#	Contributor	Commits
1	Danny Al-Gaaf	742
2	Nathan Cutler	516
3	Abhishek Lekshmanan	432
4	Ricardo Dias	344
5	Mykola Golub	222
6	Joao Eduardo Luis	163
7	Igor Fedotov	147
8	Volker Theile	138
9	Tiago Melo	131
10	Stephan Müller	114

Showing 1 to 10 of 47 entries Previous Next

SUSE社区活跃度

Commits by Contributor

Show 10 entries

Search

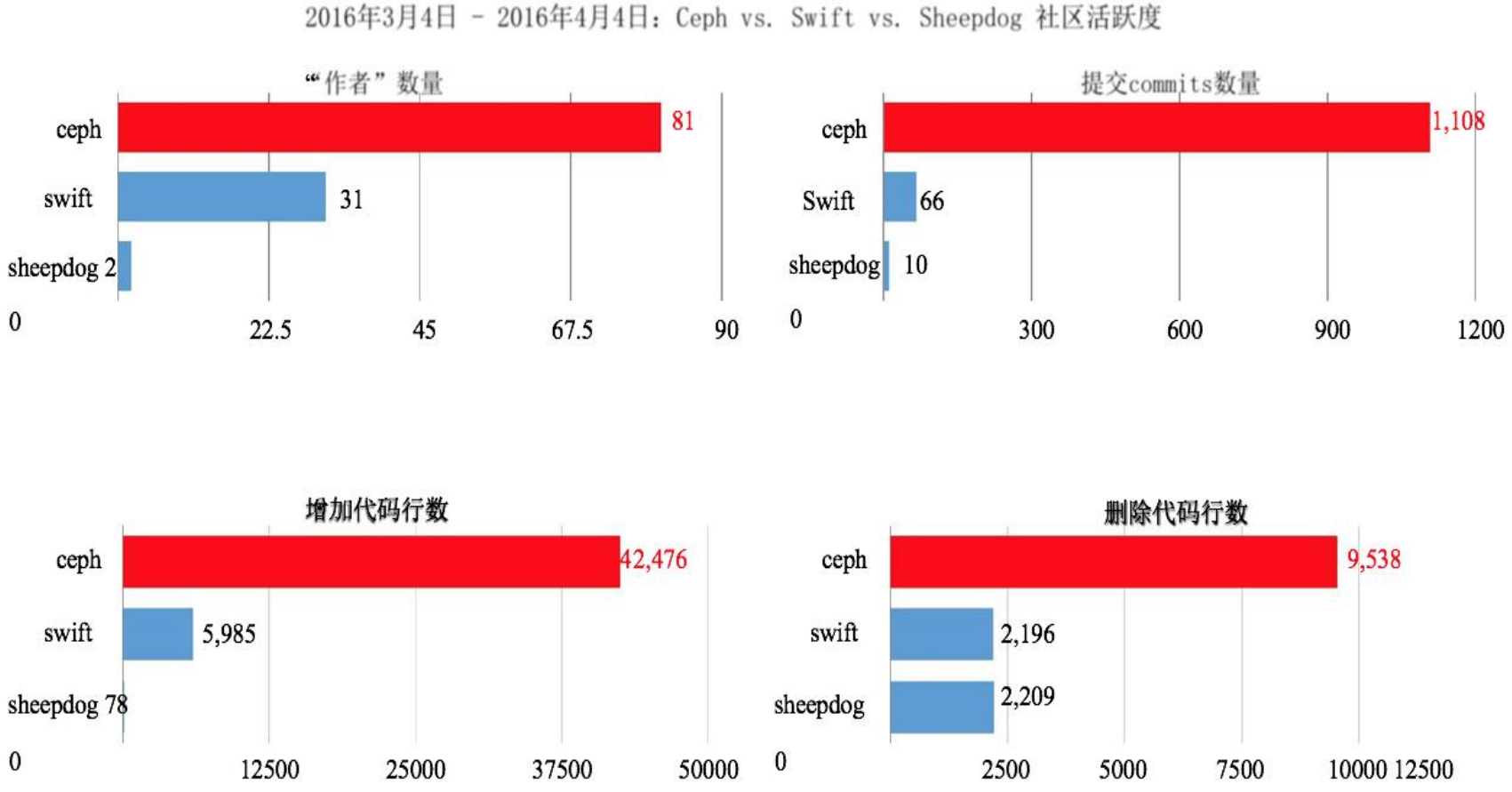
#	Contributor	Commits
1	Danny Al-Gaaf	742
2	Nathan Cutler	483
3	Abhishek Lekshmanan	358
4	Ricardo Dias	253
5	Joao Eduardo Luis	156
6	Ricardo Marques	97
7	Mykola Golub	97
8	Volker Theile	94
9	Igor Fedotov	86
10	Tiago Melo	86

Showing 1 to 10 of 43 entries

Previous Next

Contribution Summary

Commits: **3738**
LOCs: **481941**
Do not merge (-2): **0**
Patch needs further work (-1): **0**
Looks good (+1): **0**
Looks good for core (+2): **0**
Approve: **0**
Abandon: **0**
Change Requests: **0**
Patch Sets: **0**
Draft Blueprints: **0**
Completed Blueprints: **0**
Filed Bugs: **0**
Resolved Bugs: **0**
Emails: **0**
Translations: **0**

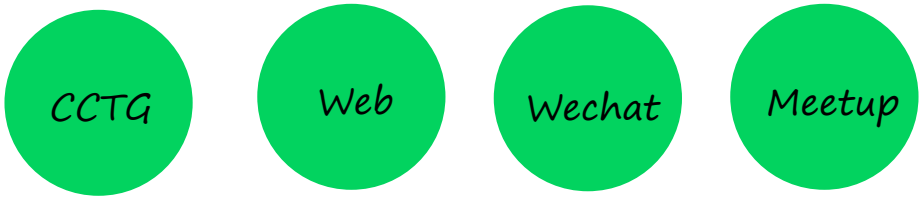

2016年3月4日 - 2016年4月4日：Ceph vs. Swift vs. Sheepdog 社区活跃度
“作者”数量
ceph
swift
sheepdog 2
81
31
0
22.5
45
67.5
90
提交commits数量
ceph
Swift
sheepdog
1,108
66
10
0
300
600
900
1200
增加代码行数
ceph
swift
sheepdog 78
42,476
5,985
0
12500
25000
37500
50000
删除代码行数
ceph
swift
sheepdog
9,538
2,196
2,209
0
2500
5000
7500
10000
12500

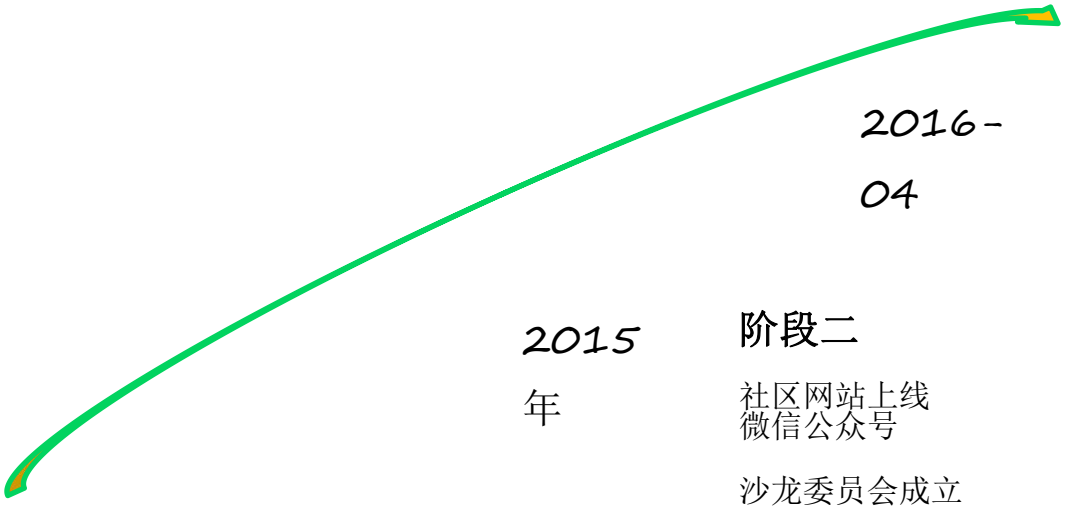
Ceph生态系统

Ceph使用C++语言开发，目前Ceph 已经成为最广泛的全球开源软件定义存储项目，拥有一个得到众多IT 厂商支持的协同开发模式。作为开源项目，Ceph遵循LGPL协议。

Ceph China Community

成立于2014年7月，包括线上的翻译小组、社区网站、 QQ群/微信群、订阅号和线下沙龙。Ceph中国线下沙龙已经成功举办多次，邀请国内一线工程师专家讲最实际最落地的采坑经验，首次提出Ceph中国行布道之旅，全面深入的在中国多座城市进行Ceph & OpenStack布道。2016年获得Ceph社区官方认可，目前Ceph中国已经成为国内最专业的Ceph技术交流社区平台。





2014年

阶段一

建立Ceph中国社区QQ群

2015年

阶段二

社区网站上线
微信公众号

沙龙委员会成立

2016-04

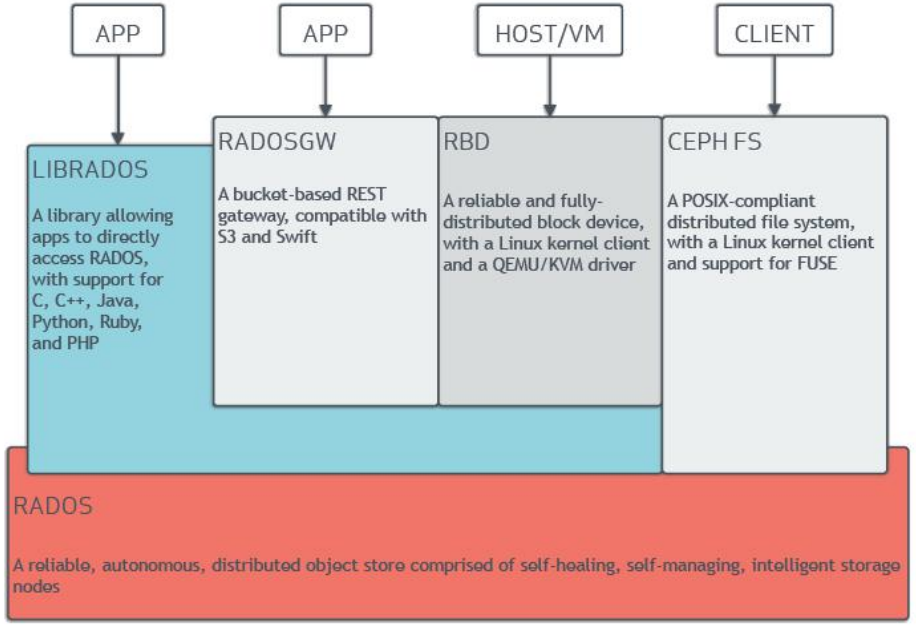
阶段三

Ceph文档本地化

获得Ceph官方认可

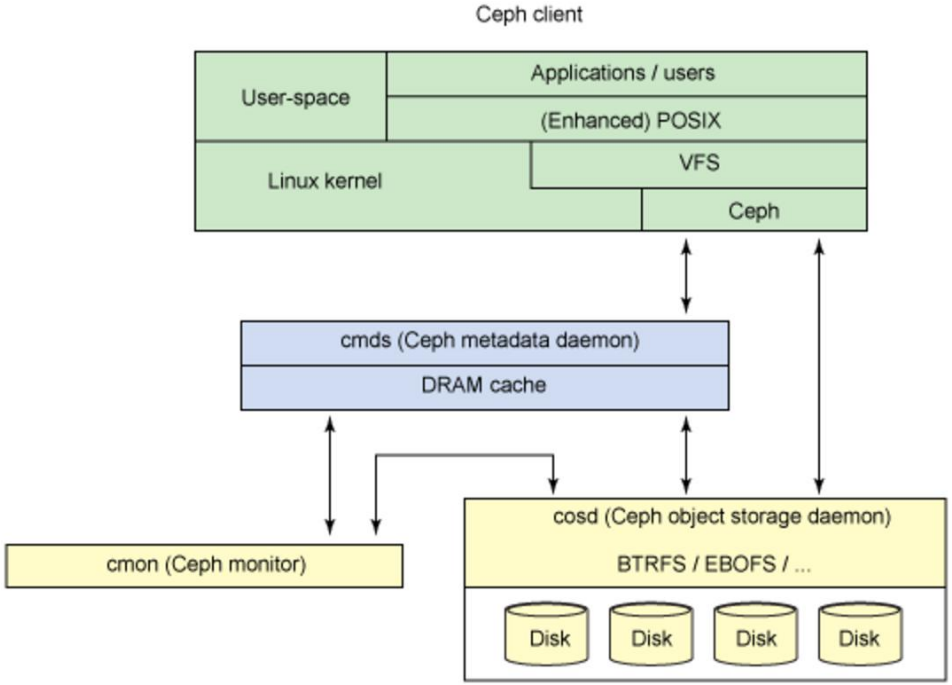
加入中国开源云联盟

Ceph Architecture

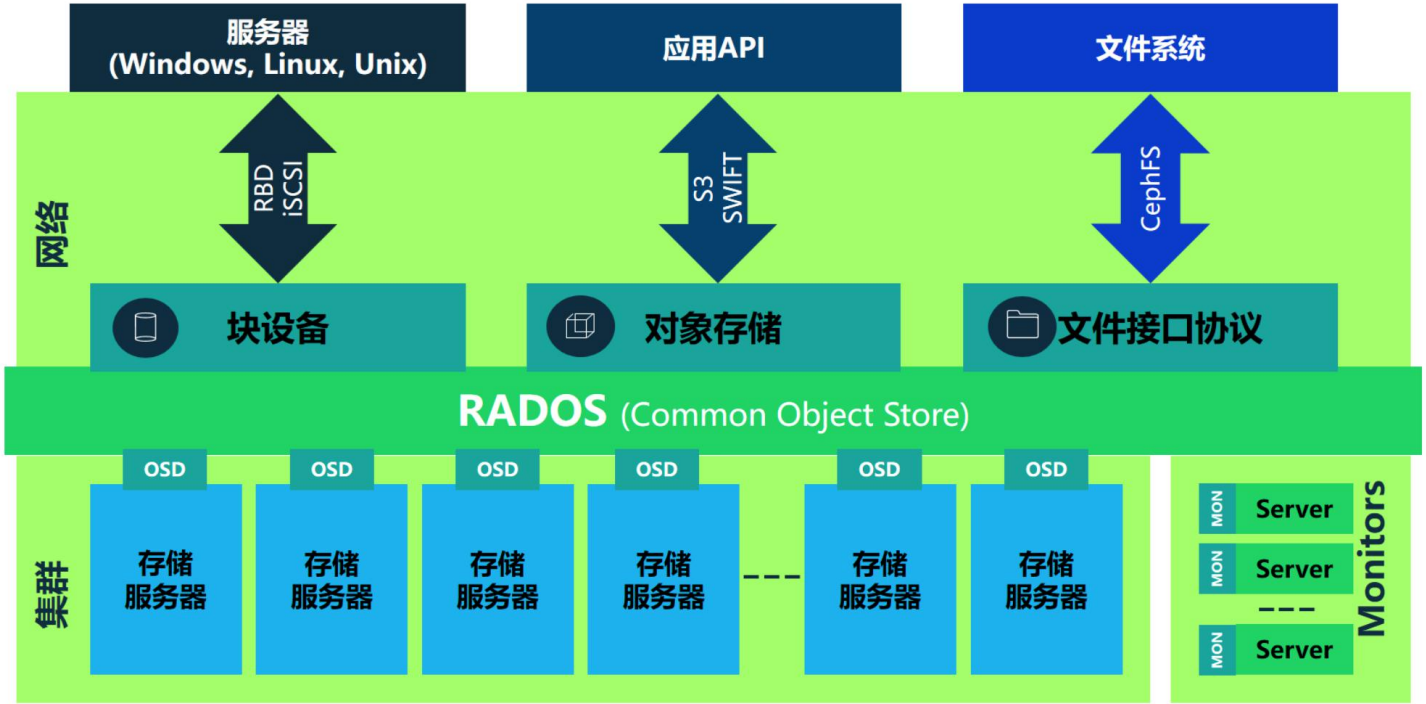


- 提供块设备存储
- 提供对象存储
- 提供文件系统存储
- 一个ceph集群至少包含一个ceph monitor、两个OSD守护进程、一个MDS(文件系统)
- 采用CRUSH 算法

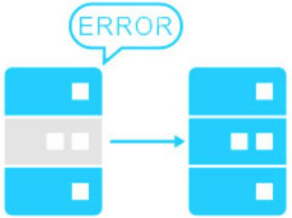
Ceph整体架构


Ceph client
User-space
Applications / users
(Enhanced) POSIX
Linux kernel
VFS
Ceph
cmds (Ceph metadata daemon)
DRAM cache
cmon (Ceph monitor)
cosd (Ceph object storage daemon)
BTRFS / EBOFS / ...
Disk
Disk
Disk
Disk

Ceph逻辑架构

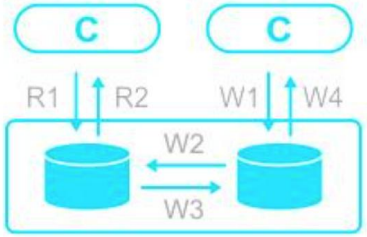

服务器
(Windows, Linux, Unix)
应用API
文件系统
网络
RBD
iSCSI
S3
SWIFT
CephFS
块设备
对象存储
文件接口协议
RADOS (Common Object Store)
集群
OSD
存储
服务器
Monitors
MON
Server

Ceph性特



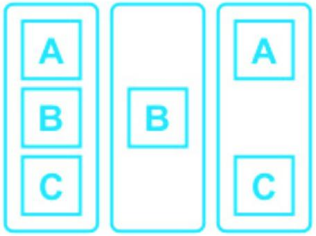
快速重构

硬盘或服务器故障后，可在全集群范围内自动并行重建数据；仅重建实际数据，不影响数据可用性，重建速度比传统SAN高一个数量级以上，大大减少重建过程中磁盘损坏造成数据丢失的概率。



强一致性复制协议

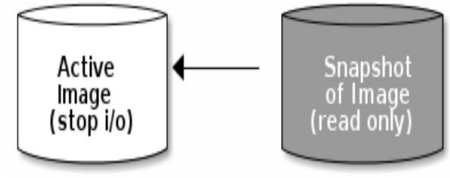
写数据如果成功，该数据的多个份副本必然是一致的，读取时，可从任一副本读到正确数据



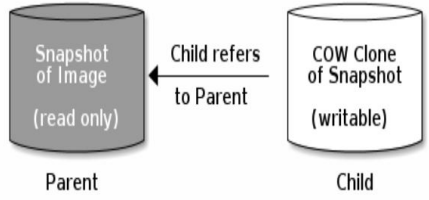
数据高可用

根据可靠性要求可配置2副本或多副本（3副本情况下，数据可靠性达7个9以上），数据跨磁盘、跨节点或跨机柜分布，自动隔离故障节点或磁盘，不会因某个磁盘、节点、机柜故障导致数据丢失或者不可访问

Ceph数据保护

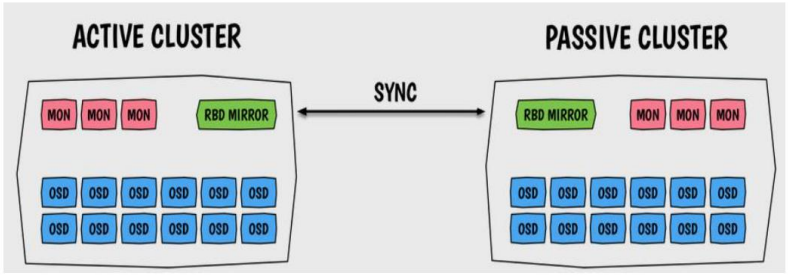


快照



克隆

纠删码

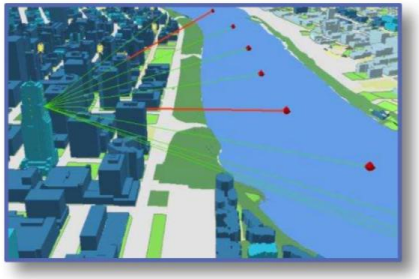


镜像复制

Ceph对象存储

企业数据量的90%+为

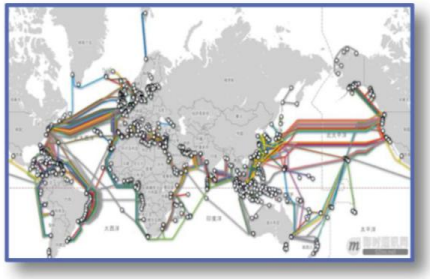
非结构化数据



全球数据量每年增长超50%，大多数为图片、视频、文档等非结构化数据

数据存储

跨越地域限制



分布式应用、广域网上的数据共享、移动访问，使得数据的存取和访问跨越地域边界

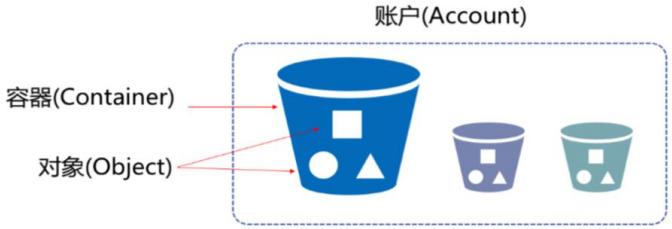
越来越多的

动态并发访问

“互联网+”应用和数据中心虚拟化需要动态并发访问，文件访问模式变为上传/下载/更新

对象存储

- 对象 =文件本身 + 元数据(文件的属性)
- 结构扁平
- 根据账户/容器/对象ID定位对象
- 程序通过HTTP协议, RESTful API访问对象

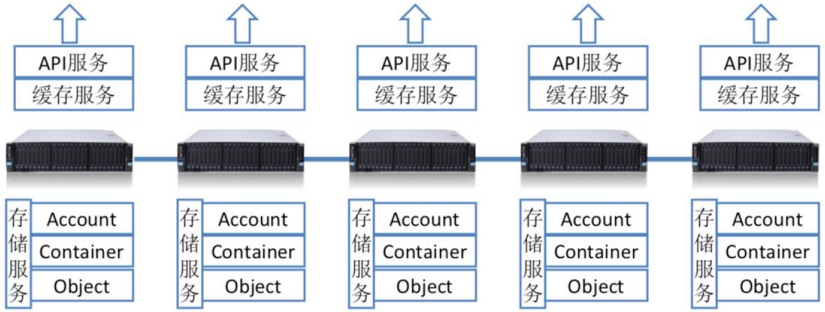


创建/更新文件（都是HTTP请求）：
PUT account1/container1/object1 HTTP/1.1
filename=platmap.mp4
Content-Type: video/mp4; Content-Encoding: gzip
Host: myserver.com

下载文件：
GET account1/container1/object1 HTTP/1.1

删除文件：
DELETE account1/container1/object1 HTTP/1.1

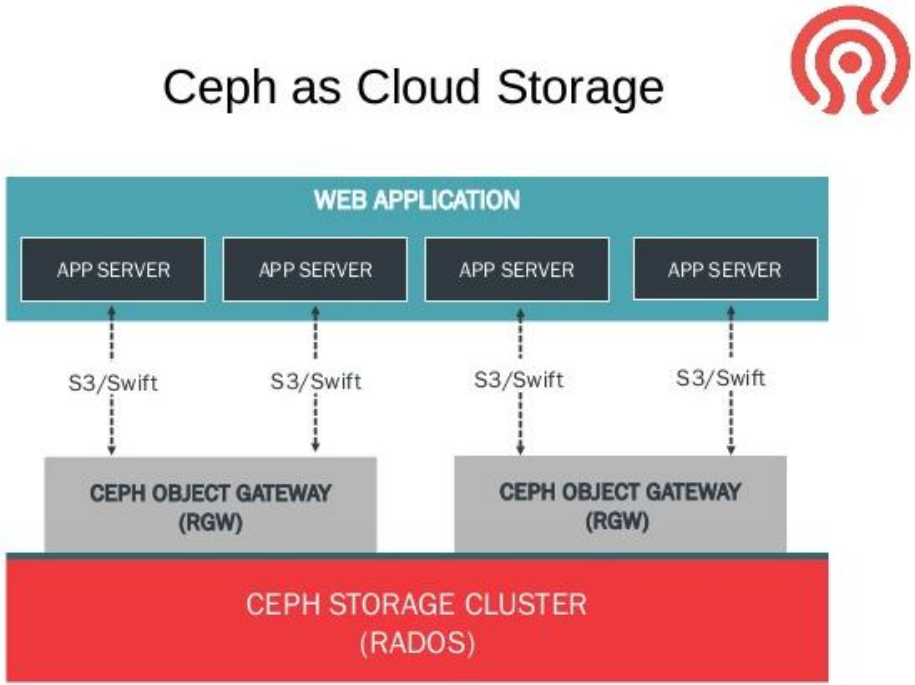
OpenStack Swift对象存储



- OpenStack最早的两个项目之一、六个核心服务之一
- 百PB级实际案例
- 无集中节点或组件，全分布式架构，可用性、可靠性极高

- 支持跨机房、跨地域部署，实现单集群全国、全球范围分布
- 支持多租户
- 支持对接Hadoop、Spark
- 支持OpenStack各种备份数据存储，帮助OpenStack环境实现灾备和双活。

RGW全称Rados Gateway，是ceph封装RADOS接口而提供的gateway服务，并且实现S3和Swift兼容的接口，也就是说用户可以使用S3或Swift的命令行工具或SDK来使用RGW。
RGW对象存储也可以作为docker registry的后端，相对与本地存储，将docker镜像存储到RGW后端可以保证即使机器宕机或者操作系统crush也不会丢失数据。



RGW用法

radosgw-admin user info --uid=mona

radosgw-admin bucket stats

S3用法

s3cmd ls

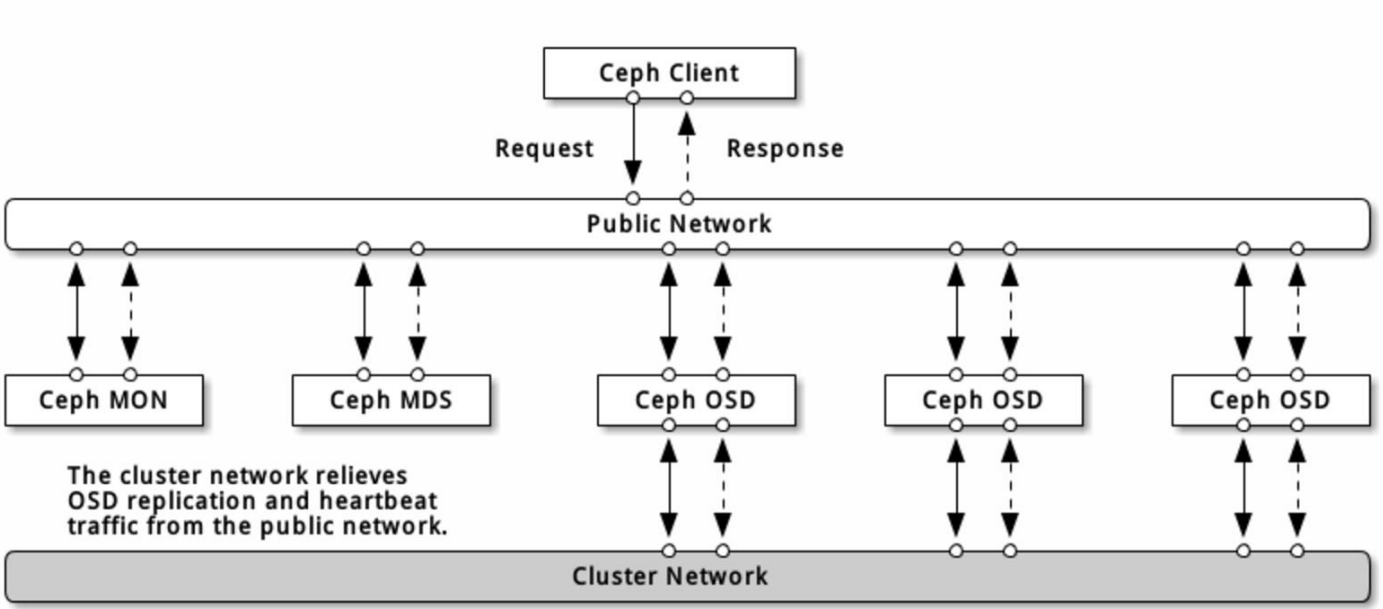
安装s3cmd

```
apt-get install python-setuptools
git clone https://github.com/s3tools/s3cmd.git
cd s3cmd/
python setup.py install
```

Swift用法

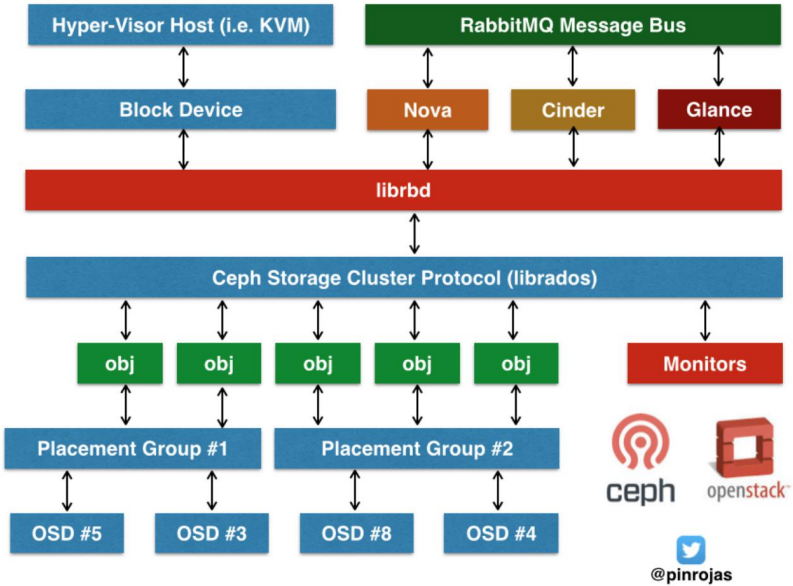
swift -V 1.0 -A http://localhost/auth -U mona:swift -K secretkey post example-bucket
swift -V 1.0 -A http://localhost/auth -U mona:swift -K secretkey list

Ceph网络规划

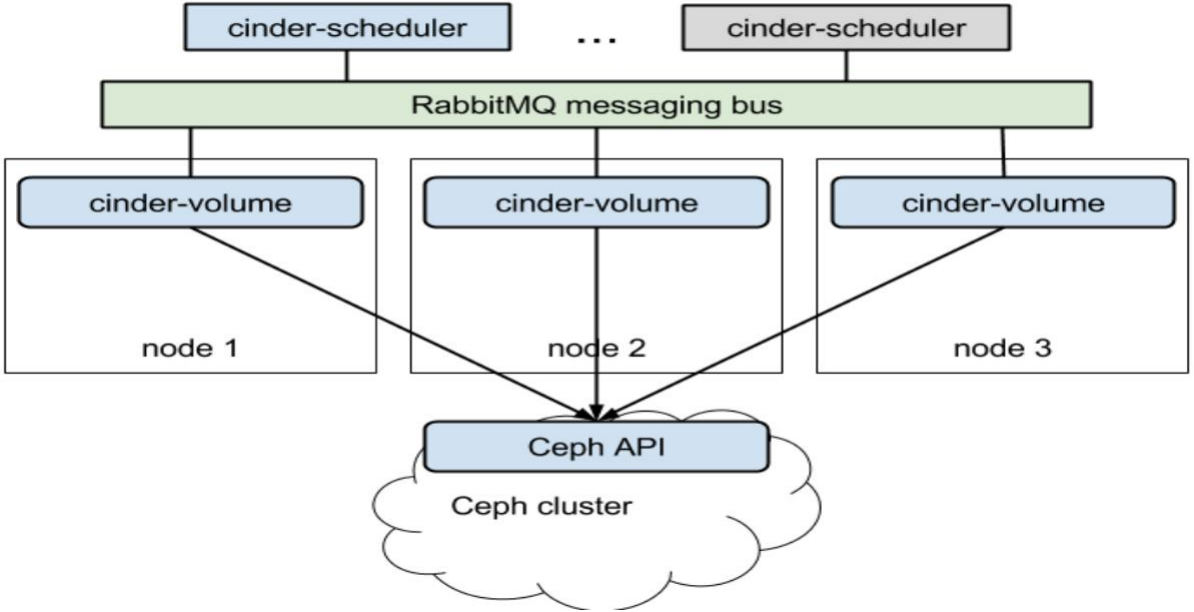

Ceph Client
Request
Response
Public Network
Ceph MON
Ceph MDS
Ceph OSD
Ceph OSD
Ceph OSD
The cluster network relieves
OSD replication and heartbeat
traffic from the public network.
Cluster Network

结合点

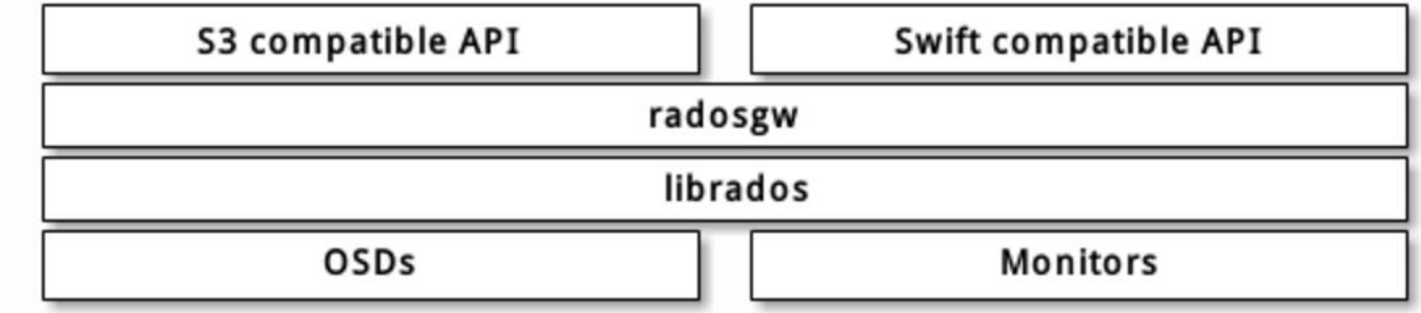
- openstack与ceph块设备结合
  - images 将glance后端与cep
  - volumes 将cinder后端与c
  - vms 将虚拟机文件存储到c
- openstack与swift对象存储结



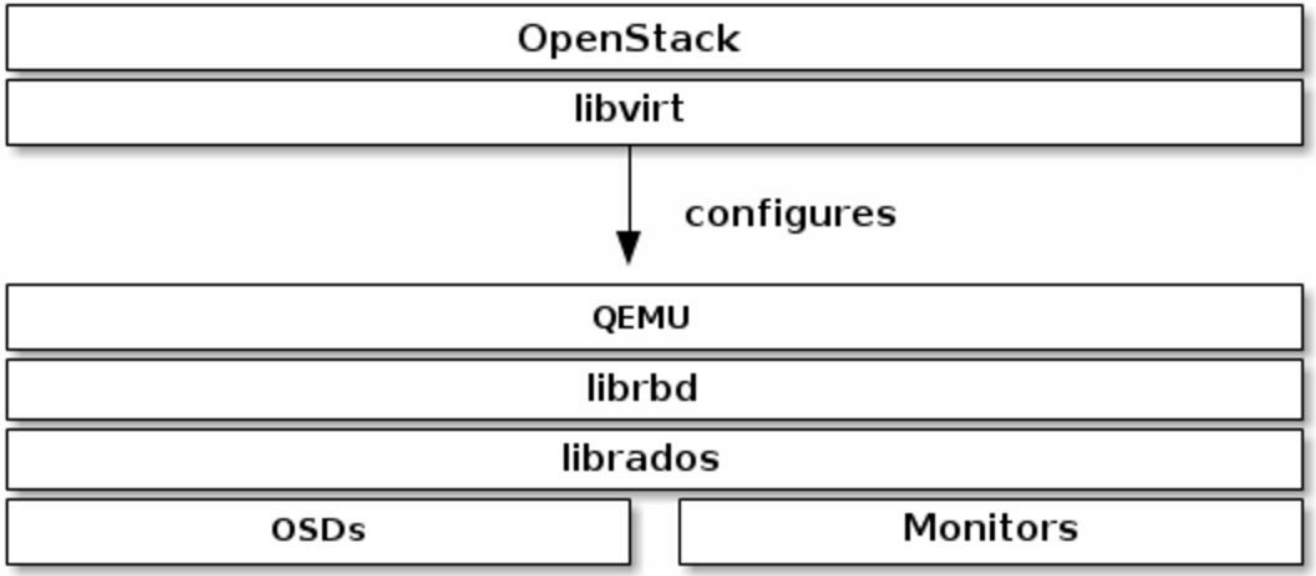
Ceph与OpenStack Cinder



Ceph与OpenStack Swift

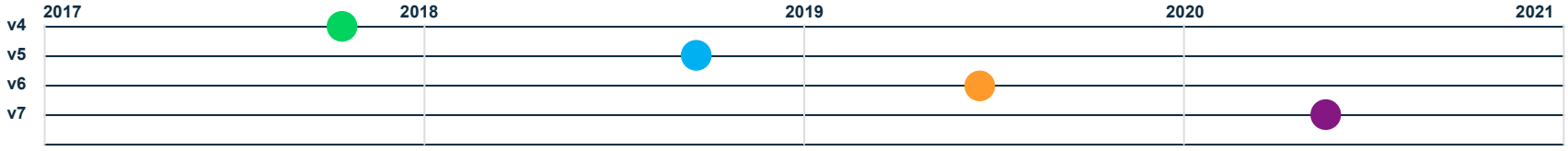


Ceph与OpenStack Compute



SUSE Enterprise Storage

SUSE Enterprise Storage RoadMap



v5

Built On
- Ceph Luminous release
- SUSE Linux Enterprise Server 12 SP3

Manageability
- openATTIC phase 2
  - ✓ Grafana monitoring dashboard
  - ✓ Prometheus event alert - email
- DeepSea (Salt) phase 2
  - ✓ Online Filestore to BlueStore

Interoperability
- NFS Ganesha
- NFS access to S3 buckets
- CIFS Samba**
- CephFS Multi MDS support

Availability
- Erasure coded block and file

Efficiency
- BlueStore back-end
- Data compression

V5.5

Built On
- Ceph Luminous release
- SUSE Linux Enterprise Server 12 SP3

Manageability
- Internationalization
- Usability enhancements
- Predefined profiles
- Autonomous data balancer**
- DeepSea (Salt) phase 3

Interoperability
- Non SUSE RBD and CephFS clients
- CIFS/Samba
- AppArmor security module

Availability
- Asynchronous iSCSI replication**
- Multisite RADOSGW N+1 with N/N-1

Efficiency
- BlueStore/RocksDB optimizations

V6

Built On
- Ceph Nautilus release
- SUSE Linux Enterprise Server 15

Manageability
- Ceph-Mgr dashboard based on oA
- Single sign-on
- Automatic metric reporting phase 1
- IPv6
- CephFS directory quotas
- Autonomous PG determination**

Interoperability
- RDMA back-end**
- QoS

Availability
- Sync to external cloud via S3
- CephFS snapshots
- Asynchronous iSCSI replication
- Asynchronous file replication**

Efficiency
- Cache tiering enhancements
- SW cache layer
- Data deduplication**

V7

Built On
- Ceph “O” release
- SUSE Linux Enterprise Server 15 SP1
- SUSE CaaS Platform**

Manageability
- Ceph Mgr dashboard phase 2
- Integration with Kubernetes**
- Automatic metric reporting phase 2
- Last good configuration rollback
- Self-healing placement groups
- Autonomous PG determination

Interoperability
- Containerized control plane**
- RDMA back-end

Availability
- Asynchronous file replication

Efficiency
- Data deduplication
- OSD Daemon optimizations

** Items are tech preview

* Information is forward looking and subject to change at any time.

SUSE 企业存储

一个可以随时DIY 的企业级分布式云存储

文件服务(NFS,CIFS,CephFS)

块服务(iSCSI, RBD)

对象服务(S3,Swift,Librados)

SUSE 企业存储软件

SUSE 企业Linux 操作系统

AMD ARM DELL IBM hp HUAWEI H3C

intel lenovo cisco Sugon inspur 浪潮 FiberHome

分布式存储架构

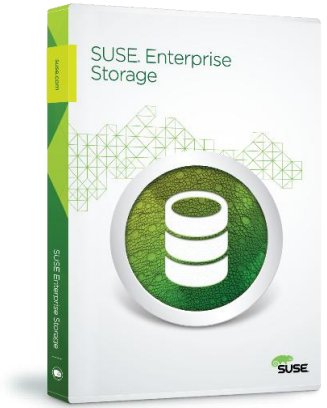
分布式存储架构

一款存储，适配多种场景

ECS （对象存储）

ISILON（文件存储）

VSAN （块存储）



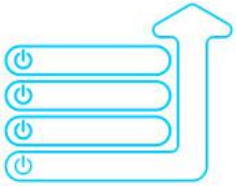
性能型

容量型

数据管理特性

PB级存储容量的集群

SES架构具有良好的横向扩展能力，满足用户业务和数据量不断增长的需要

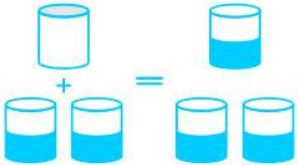


按需平滑扩展

避免叉车式升级，保护基础设施投资，大大降低扩容成本。

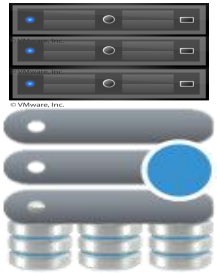
动态在线扩容

动态增加、删除存储和计算资源，不影响业务可用性和数据可靠性



自动数据均衡

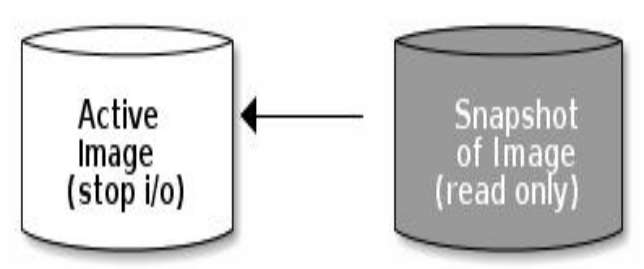
动态增加、删除节点或者磁盘，能自动迁移和均衡数据，以优化性能，提升可靠性。



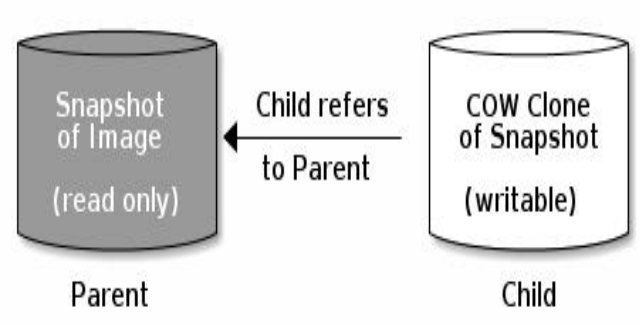
数据迁移

SES支持跨代节点并存， 无需数据迁移。支持在线节点退出

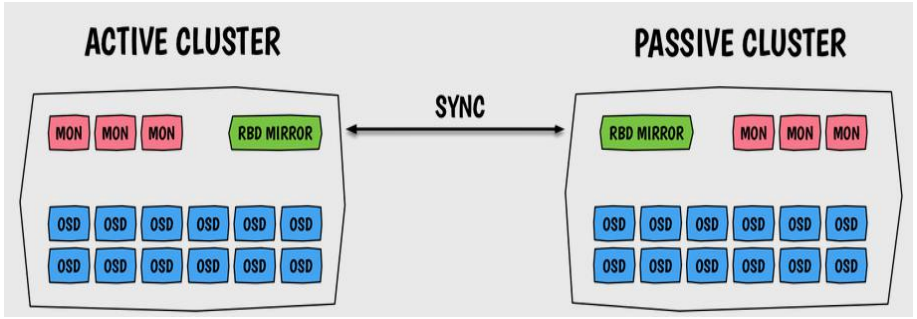
数据保护特性



快照

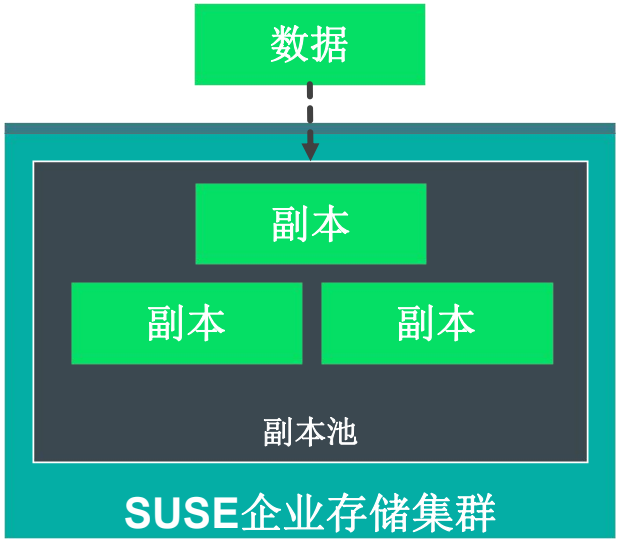


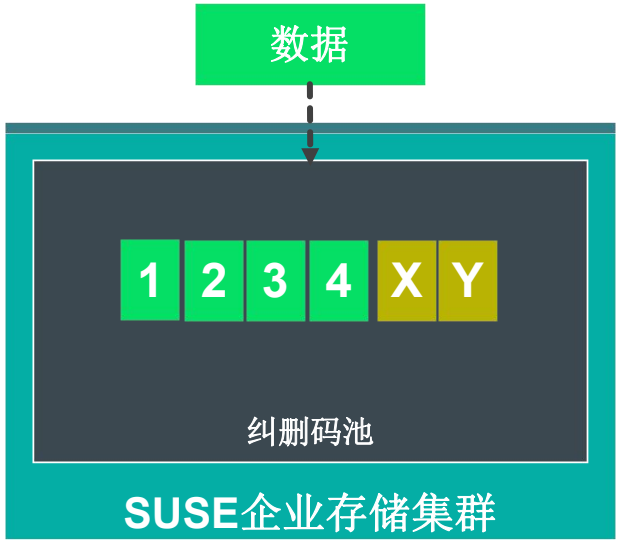
克隆



异步复制

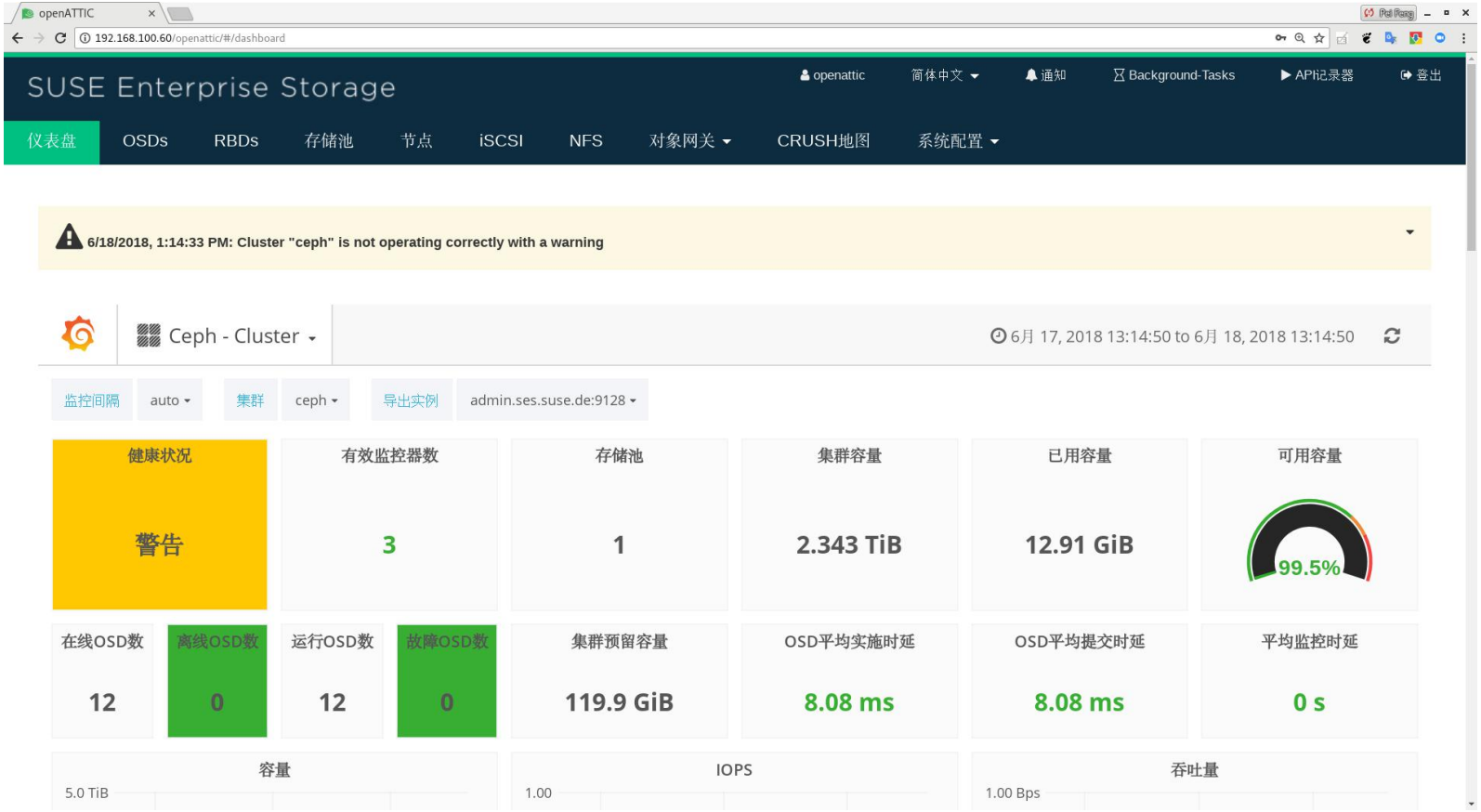
多种数据保护，构建稳固持久的存储





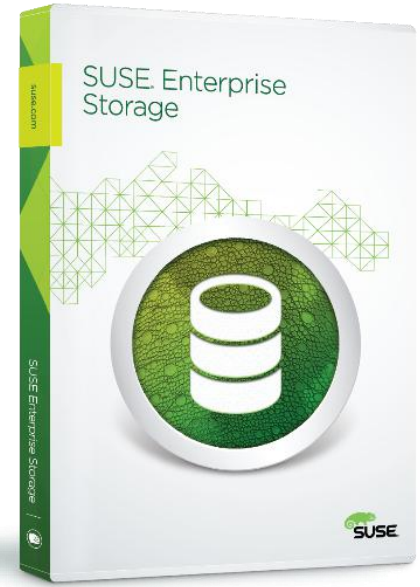
- 20倍的数据重建速度
- 10倍的磁盘容错能力

本地化管理界面


openATTIC
192.168.100.60/openattic/#/dashboard
SUSE Enterprise Storage
openattic
简体中文
通知
Background-Tasks
API记录器
登出
仪表盘
OSDs
RBDs
存储池
节点
iSCSI
NFS
对象网关
CRUSH地图
系统配置
6/18/2018, 1:14:33 PM: Cluster "ceph" is not operating correctly with a warning
Ceph - Cluster
6月 17, 2018 13:14:50 to 6月 18, 2018 13:14:50
监控间隔
auto
集群
ceph
导出实例
admin.ses.suse.de:9128
健康状况
警告
有效监控器数
3
存储池
1
集群容量
2.343 TiB
已用容量
12.91 GiB
可用容量
99.5%
在线OSD数
12
离线OSD数
0
运行OSD数
12
故障OSD数
0
集群预留容量
119.9 GiB
OSD平均实施时延
8.08 ms
OSD平均提交时延
8.08 ms
平均监控时延
0 s
容量
5.0 TiB
IOPS
1.00
吞吐量
1.00 Bps

SUSE 的与众不同

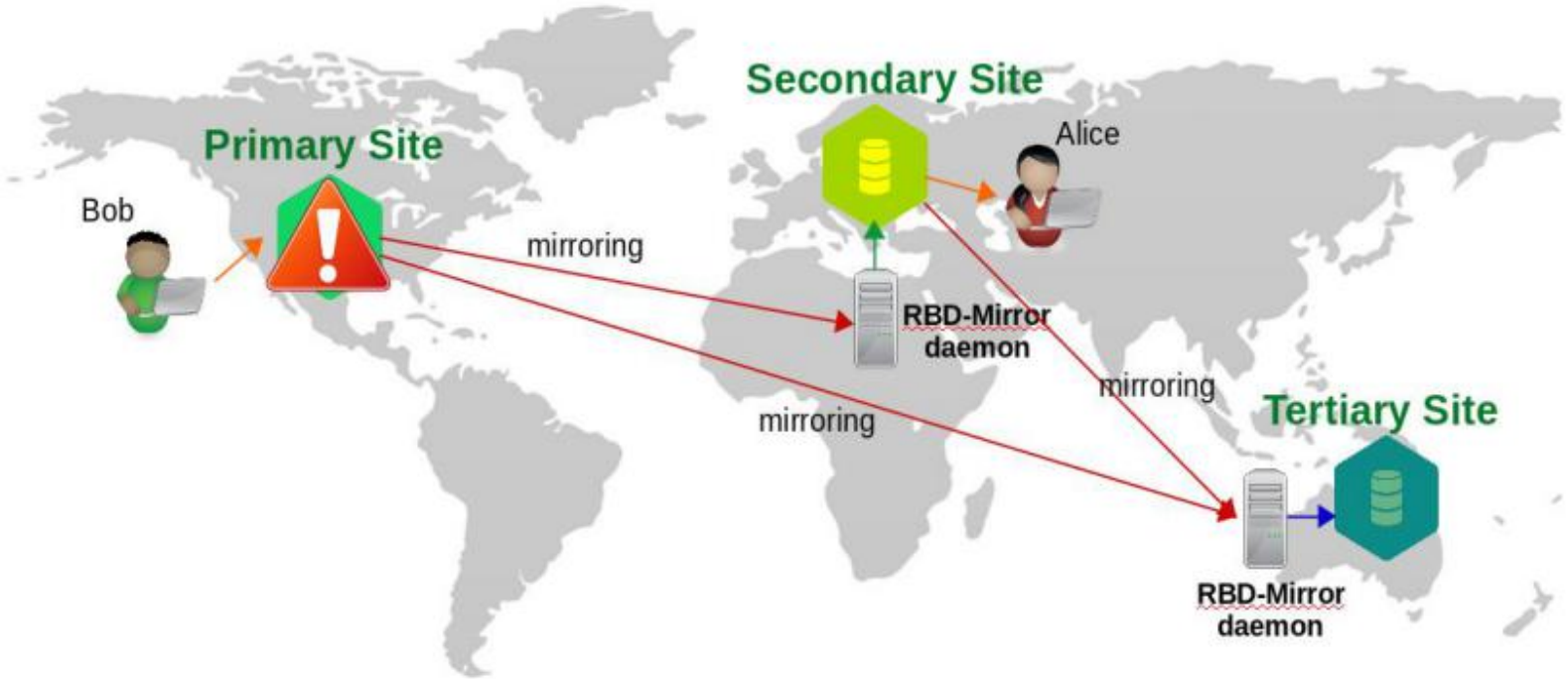
- 25+年开源技术支持
- 本地化管理界面
- 软件代码全开源
- 最开放的接口iSCSI
- 最稳固的底层平台 SUSE Linux
- 最全的硬件兼容



AMD ARM DELL IBM hp HUAWEI H3C

intel lenovo cisco Sugon inspur 浪潮 FiberHome

异步复制/异地容灾


Secondary Site
Primary Site
Alice
Bob
mirroring
RBD-Mirror
daemon
mirroring
mirroring
Tertiary Site
RBD-Mirror
daemon

SUSE企业存储部署

配置repo

配置salt

#准备安装环境
salt-run state.orch ceph.stage.0

#发现硬件配置
salt-run state.orch ceph.stage.1
#分发配置到所有节点
salt-run state.orch ceph.stage.2

#装备集群
salt-run state.orch ceph.stage.3

#装备服务
salt-run state.orch ceph.stage.4

```
admin:/ # deepsea stage run ceph.stage.0
Starting stage: ceph.stage.0
Parsing ceph.stage.0 steps...

[parsing] on master
          |_ ceph.stage.0

[parsing] on admin.suse.com
          |_ ceph.salt-api
             ceph.updates
             ceph.metapackage
             ceph.sync
             ceph.repo

[parsing] on *
          |_ ceph.updates
             ceph.repo
             ceph.sync
             ceph.packages.common
             ceph.metapackage
             ceph.mines

Parsing ceph.stage.0 steps... ✓

[init]    validate.setup.......................................... ✓ (4s)
          |_ cmd.shell(/usr/bin/zypper info ceph) on
             admin.suse.com..................................... ✓ (0.7s)
             node2.suse.com..................................... ✓ (0.5s)
             node3.suse.com..................................... ✓ (0.5s)
             node1.suse.com..................................... ✓ (0.6s)
[init]    advise.salt_run......................................... ✓ (0.0s)
```

SUSE企业存储三节点部署

默认ceph设置OSD节点需要4个，如果OSD节点小于4个，需要修改配置文件
#vim /srv/modules/runners/validate.py

```
if (not self.in_dev_env and len(storage) < 4) or (self.in_dev_env and len(storage) < 1):
    msg = "Too few storage nodes {}".format(",".join(storage))
```

#装备集群
salt-run state.orch ceph.stage.3

SES 5.0的案例 —— BMW

背景介绍：

1、200+套SUSE Linux （Just in China）

2、包含BMW私有云平台（IaaS、PaaA for Web/DB/SAP）、ERP/CRM、数据库、OA等---Host : 1,500 Xen running on SUSE Linux Enterprise Server

--VM : 1 vCPUs/1 GB to 16 vCPUs/64 GB in production and include up to 8 TB of application storage

--Storage : Fibre-Channel SAN, NFS / SUSE Cloud and Ceph (Rados later)

遇到的问题：

1，缺少敏捷性和可移植性。

2，使用NPIV或者FC，灵活性缺乏

3，移植和必要的重定义在FC SAN上还是一个issue。

4，NFS虽然能解决一些问题，但是还不是一个完善的分布式的存储解决方案

方案优势：

1、良好的性能，可靠性，高性价比适应所有的工作负载

2、无单点故障，稳定可靠，数据冗余备份，可实现快速恢复

3、无需新的技术，只要了解一些新的操作和一些简单地模型即可；

规模：

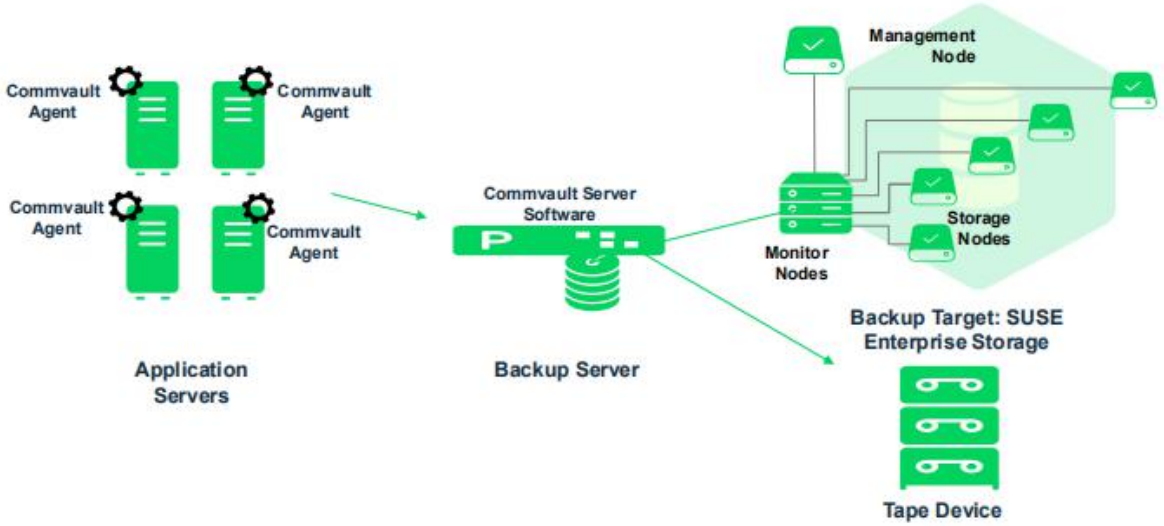
•8 Storage nodes (to be extended soon)
–16x physical cores (HT) / 256 GB memory
–10x 1.2 TB HDD + 2x 100 GB SSD (for journals)
•SUSE Storage 2.0

SES 5.0的案例 —— 备份和归档

某保险公司使用**Commvault**进行 保单 备份 （合同+身份证扫描件）

约30TB的原始数据，每年新增3T；归档和数据库共用存储设备，造成浪费和隐患

最终选择 SES 和 Dell 服务器，通过S3接口，实现归档数据单独备份



谢 谢!