  Ceph 作为软件定义存储的代表之一，最近几年其发展势头很猛，也出现了不少公司在测试和生产系统中使用 Ceph 的案例，尽管与此同时许多人对它的抱怨也一直存在。本文试着整理作者了解到的一些使用案例。

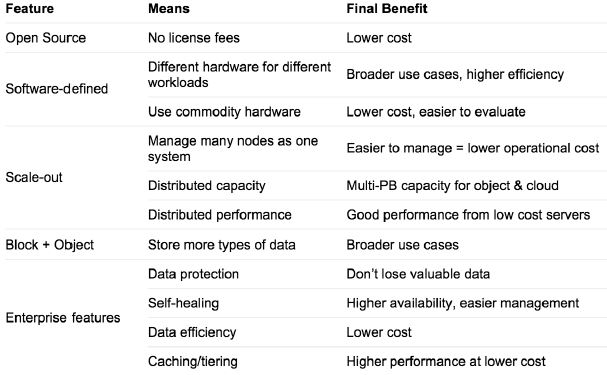
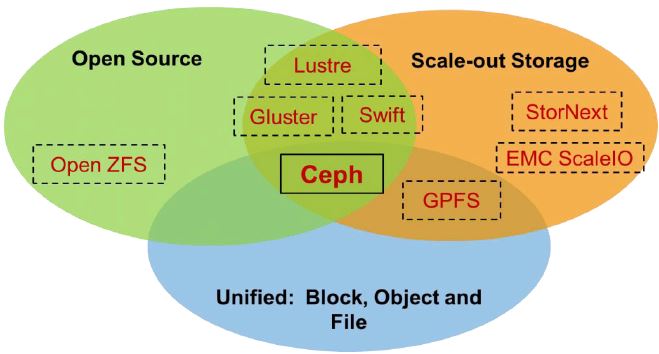
**1. 携程（Ctrip）**

携程所使用的各种存储的现状：

* 商业存储：
  + SAN(HP/ HPS) , 1+ PB, 数据库
  + NAS (HW) , 800+ TB, 文件共享
* 开源存储
  + GlusterFS, 1+ PB, 数据库备份
  + FastDFS,   1+ PB, 海量照片
  + HDFS,      10+ PB, 大数据

而在不久的将来，随着公司业务的发展，携程需要的存储容量需要扩大到10倍以上。

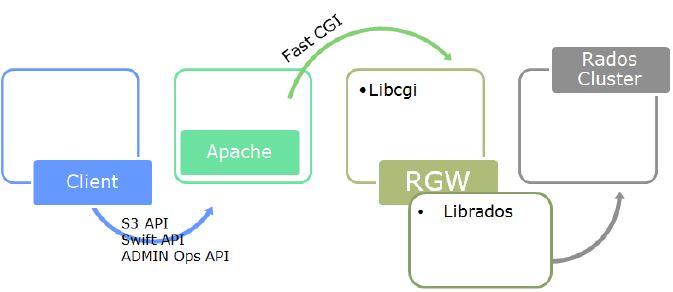
携程选择Ceph的理由：低成本 + SDS + Scale-out + 统一存储 + 企业特性

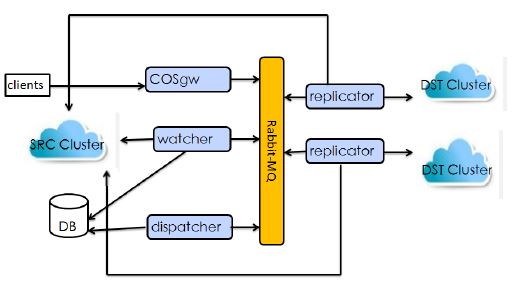
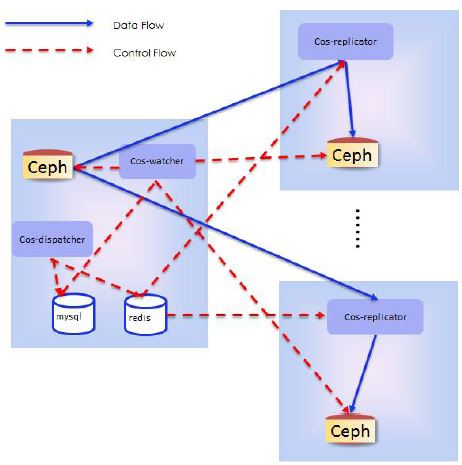
携程目前的Ceph集群的配置：

* CephVersion: 0.94.2，H release
* Object Storage: RGW + Swift API
* SDK: Python/ Java/ C#/ Ruby
* OS: Centos 6.4
* 硬件：CPU(2 channels & 32 Core)、Mem128GB、disk（12\*3TB/SATA disk +2\*256GB raid1 SSD）、NIC(4\*Gigabit LAN, bond 2 in 1 pair)

RGW 使用架构：



携程有在数据中心之间的同步数据的需求。在研究了 CRUSHmap、Radosgw-agent、Federate gateway （不稳定、不灵活（只支持Zone 之间同步）、不易扩展）后，其自研了COS方案，它具有稳定、灵活、扩展性等特点：

下一步的计划：

* Database on Ceph (Dev & QA Farm)
* Openstack/ DockerIntegrate with Ceph
* IT “Dropbox”

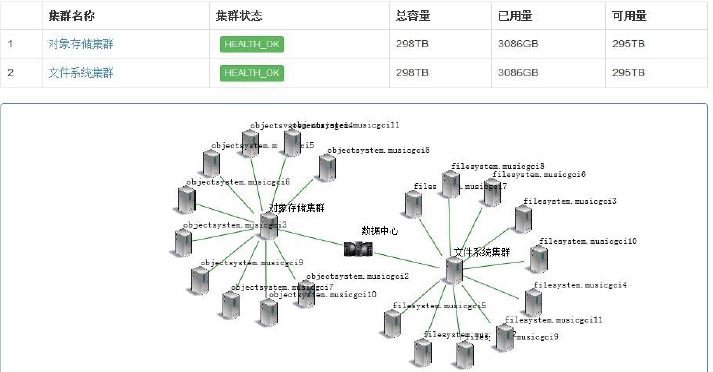
资料来源：携程在 2015/10/18 SH Ceph Day 上的分享。

楼主点评：

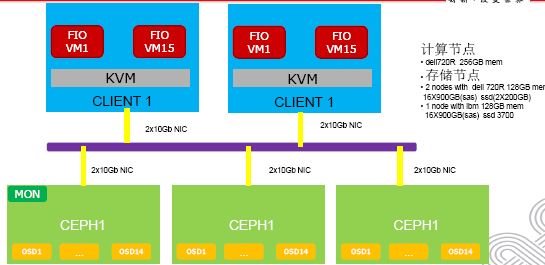
* 与互联网公司的通常做法一致：慎重选择、细致测试、分布使用（往往从开发测试环境中使用开始）、开源的不够用就自研
* 希望携程能有更多的分享和回馈社区

**2. 联通研究院**

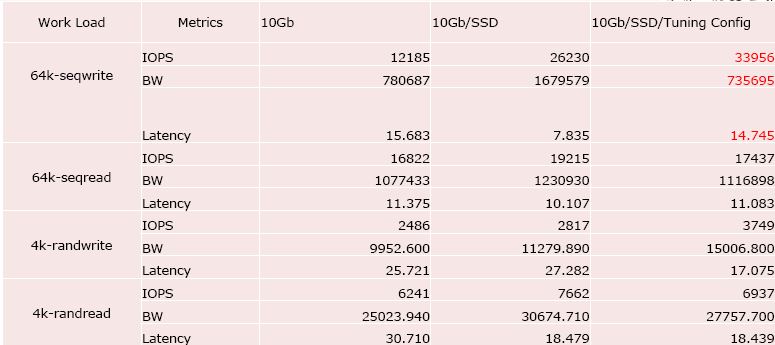
中国联通研究院在使用Ceph对象和文件存储：



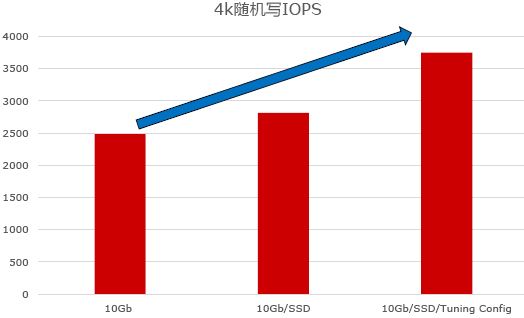
该集群还比较小，更多的是处于做大规模使用前的准备阶段。其测试环境：



测试结果：



他们认为SSD对性能提升的效果一般：



资料来源：联通研究院在 2015/10/18 SH Ceph Day 上的分享。

楼主点评：

* 尚处于小规模测试和试用阶段
* 使用的测试方法或者调优手段可能没到位，不然性能提高不会那么少

**3. 宝德云（PLCloud）**

宝德云使用Ceph的理由：

* Pure Software
* Open Source, Commercial Support
* Unified Storage: RBD, RGW, CephFS
* Scale Out
* Self Healing
* Replication and Erasure Coding
* Integrate well with OpenStack

宝德云的用法：OpenStack + Ceph （RDB，CephFS）+ Docker

* 所有 OpenStack 存储都放在 Ceph 上
* 18\*(5 OSD+1SSD) / CephRBD / CephFS
* 785VM / 4vCPU32GB per VM
* Ubuntu14.04 / Docker1.6.1 / 150+ Containers per VM
* All VM Mount CephFS
* Mount VM Directory as Container’s Data Volume
* Boot 1 VM < 5s
* Boot 1 Container < 1s
* Boot 150+Containers < 120s
* Ceph Rados Gateway driver for Docker Registry

Map RBD device inside DockerContainer

* CephFS as Data Volume

CephFS as NAS Storage

* Run Ceph in Containers

使用案例：宝德云上的爱慕流媒体服务

* Run media web/app/dbvmover OpenStackand CephRBD
* Use CephRGW as media resource storage
* Put video TransportStream/jpg file via c-language programme
* Manage resource via python-swiftclient
* 400+KB per video tsfile
* Reserved video ts/jpg file 7 days or 30 days
* Allow media server temporary access to objects
* Provide media service for Internet and Intranet User

资料来源：宝德云在 2015/10/18 SH Ceph Day 上的分享。

楼主点评：

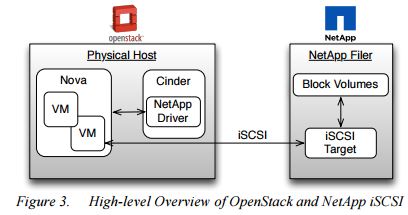
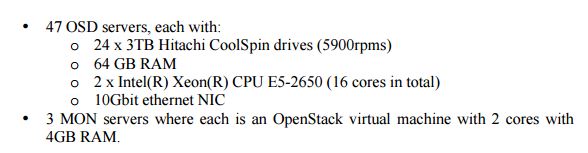
* 够大胆（到目前为止 CephFS 还不稳定呐）、够与时俱进（什么东西新就用什么）
* 没说清楚怎么支持爱慕的超大流数据

**4. CERN （欧洲核子研究委员会）实验室（**[**来源**](http://research.nesc.ac.uk/files/CERN_OpenLab_Report.pdf)**）**

**4.1 测试环境**

CERN 的一些实习生搭了一套环境，使用 NetApp Cinder driver 和 Ceph，进行性能比较。

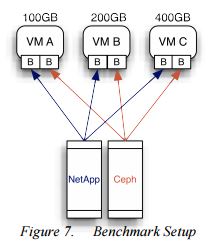
NetApp 环境（适应iSCSI驱动）：                               Ceph 集群：

FAS2040 Storage Systems

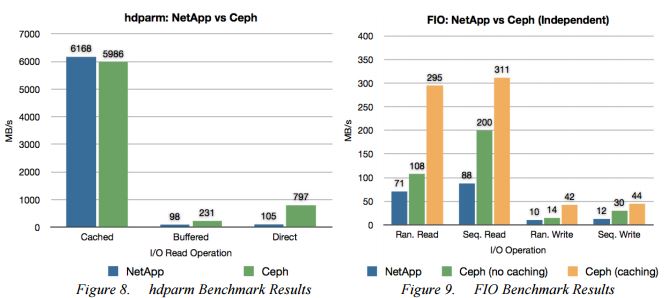
* Data ONTAP 8
* 52 Disks

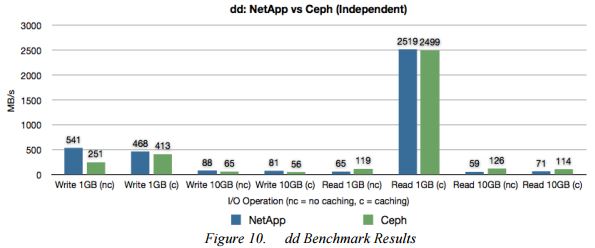
Benchmark 环境：



做法：在两个存储上分别创建100G，200G，400G的卷，分别挂载到三个虚机上，使用 hdparm、Flexible I/O Tester 和 dd命令作为测试工具。

**4.2 测试结果**

（FIO使用的是 writeback 缓存机制）



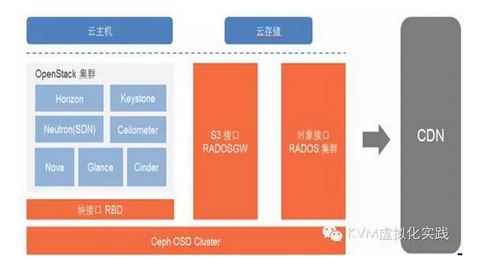
结论：

（1）读上，Ceph 比 NetApp 更快；写上，两者差不多。

（2）Ceph 使用缓存的话，对 I/O 性能影响很大。writeback 能较大地提交性能，而writethrough 只能轻微地提交性能。

（3）对单个卷使用不同的条带化参数，能提交其性能。该功能会在 Cinder 中实现。

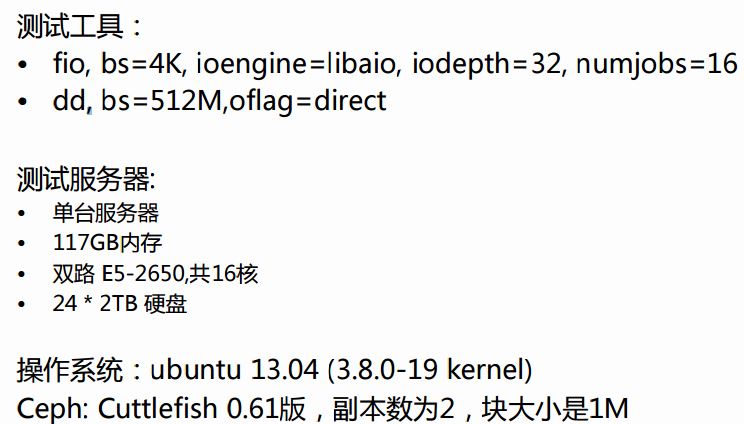
**5. 乐视云（**[**来源**](http://www.chinacloud.cn/show.aspx?id=21835&cid=81)**）**



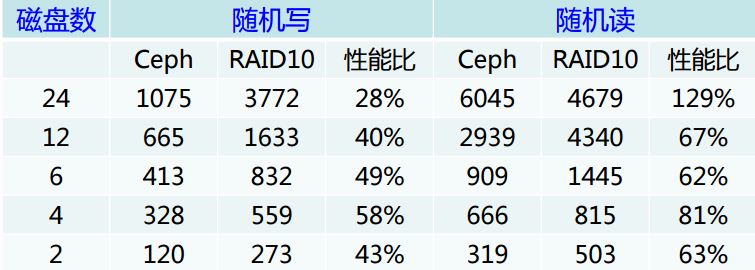
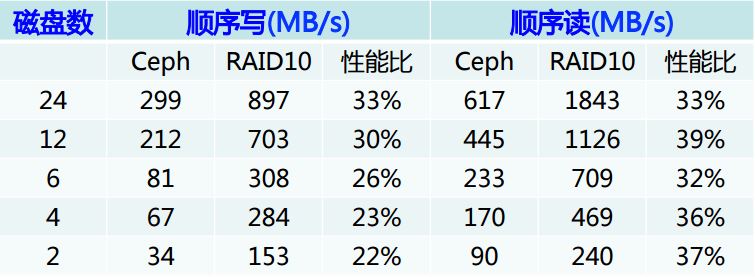
    乐视采用了 Ceph RBD 作为 统一存储，OpenStack使用的Cinder，后端接的是Ceph，Glance也是共享Ceph存储。同时还提供了 S3 对象存储，用作于 CND 源站，存储乐视网的视频以及客户需要分发的资源。S3 也是全国分布式部署，用户可以就近上传，再推送到北京。目前乐视云 OpenStack 规模已达 900 个物理节点，对象存储的数据达到数PB。乐视认为，“ceph 数据分布，性能方面都很不错，crush算法是它的亮点“。

**6. UnitedStack （**[**来源**](http://share.csdn.net/uploads/5245537a6d571/5245537a6d571.pdf)**）**

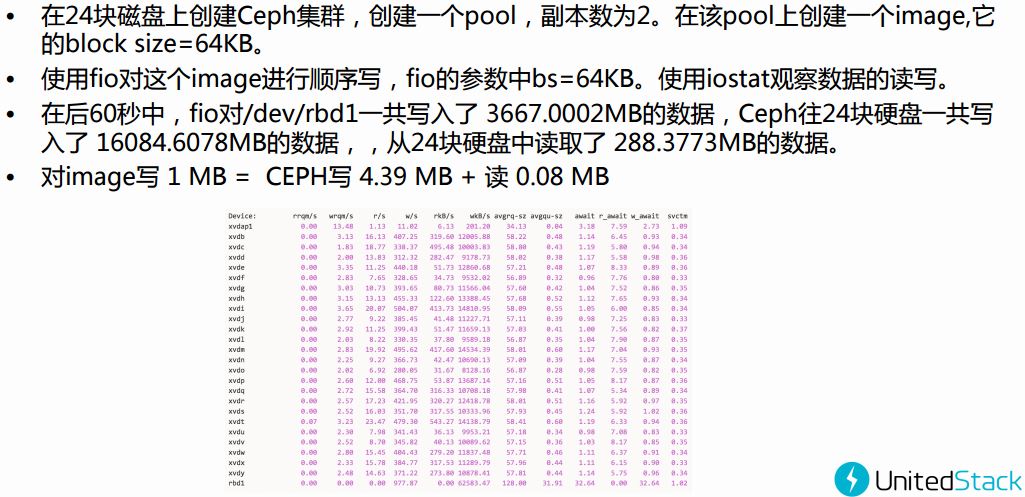
    国内的 UnitedStack 应该说对 Ceph 做出了很大的贡献，包括很多的 bug fix 和新的功能，同时，也使用 Ceph 搭建了很大规模的云。[他们 Ceph 团队的 blog](https://www.ustack.com/category/blog/ceph-blog/) 非常值得关注。下面是他们的测试报告：



IOPS：                                                                                  吞吐率

写惩罚：



结论：

