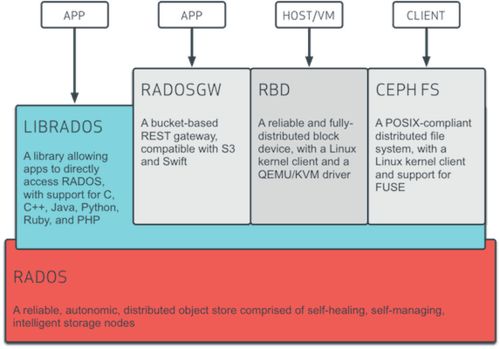
1. Ceph基础架构组件



最底层的是RADOS(Reliable,Autonomic Distributed Object Store)，RADOS自身是一个完整的分布式对象存储系统，它具有可靠、智能、分布式等特征，Ceph的高可靠、高可拓展、高性能、高自动化都是由这一层来提供的，用户数据的存储最终也都是通过这一层来进行存储的，RADOS可以说就是Ceph的核心。

RADOS系统由两部分组成，分别式OSD和Monitor.

基于RADOS层的上一层是LIBRADOS，LIBRODS是一个库，它允许应用程序通过访问该库来与RADOS系统进行交互，支持多种编程语言。

基于LIBRADOS层开发的又可以看到有三层，分别式RADOSGW，RBD和CEPH FS.

RADOSGW：一套基于当前流行的RESTFUL协议的网关，并且兼容S3和Swift.

RBD：RBD通过Linux内核客户端和QEMU/KVM驱动来提供一个分布式的块设备。

CEPH FS：通过Linux内核客户端和FUSE来提供一个兼容POSIX的文件系统。



1. 对象存储、块存储、文件系统的区别

地址：<https://www.talkwithtrend.com/Article/178247>

<https://www.jianshu.com/p/cc3ece850433>

<https://www.cnblogs.com/garfieldcgf/p/12125237.html>

**块存储**（DAS/SAN）通常应用在某些专有的系统中，这类应用要求很高的随机读写性能和高可靠性，上面搭载的通常是Oracle/DB2这种传统数据库，连接通常是以FC光纤（8Gb/16Gb）为主，走光纤协议。如果要求稍低一些，也会出现基于千兆/万兆以太网的连接方式，MySQL这种数据库就可能会使用IP SAN，走iSCSI协议。通常使用块存储的都是系统而非用户，并发访问不会很多，经常出现一套存储只服务一个应用系统，例如如交易系统，计费系统。典型行业如金融，制造，能源，电信等。

***文件存储（NAS）***相对来说就更能兼顾多个应用和更多用户访问，同时提供方便的数据共享手段。毕竟大部分的用户数据都是以文件的形式存放，在PC时代，数据共享也大多是用文件的形式，比如常见的的FTP服务，NFS服务，Samba共享这些都是属于典型的文件存储。几十个用户甚至上百用户的文件存储共享访问都可以用NAS存储加以解决。在中小企业市场，一两台NAS存储设备就能支撑整个IT部门了。CRM系统，SCM系统，OA系统，邮件系统都可以使用NAS存储统统搞定。甚至在公有云发展的早几年，用户规模没有上来时，云存储的底层硬件也有用几套NAS存储设备就解决的，甚至云主机的镜像也有放在NAS存储上的例子。文件存储的广泛兼容性和易用性，是这类存储的突出特点。但是从性能上来看，相对SAN就要低一些。NAS存储基本上是以太网访问模式，普通千兆网，走NFS/CIFS协议。

***对象存储，***前面说到的块存储和文件存储，基本上都还是在专有的局域网络内部使用，而对象存储的优势场景却是互联网或者公网，主要解决海量数据，海量并发访问的需求。基于互联网的应用才是对象存储的主要适配（当然这个条件同样适用于云计算，基于互联网的应用最容易迁移到云上，因为没出现云这个名词之前，他们已经在上面了），基本所有成熟的公有云都提供了对象存储产品，不管是国内还是国外。对象存储常见的适配应用如网盘、媒体娱乐，医疗PACS，气象，归档等数据量超大而又相对“冷数据”和非在线处理的应用类型。这类应用单个数据大，总量也大，适合对象存储海量和易扩展的特点。网盘类应用也差不多，数据总量很大，另外还有并发访问量也大，支持10万级用户访问这种需求就值得单列一个项目了（这方面的扫盲可以想想12306）。归档类应用只是数据量大的冷数据，并发访问的需求倒是不太突出。另外基于移动端的一些新兴应用也是适合的，智能手机和移动互联网普及的情况下，所谓UGD（用户产生的数据，手机的照片视频）总量和用户数都是很大挑战。毕竟直接使用HTTP get/put就能直接实现数据存取，对移动应用来说还是有一定吸引力的。对象存储的访问通常是在互联网，走HTTP协议，性能方面，单独看一个连接的是不高的（还要解决掉线断点续传之类的可靠性问题），主要强大的地方是支持的并发数量，聚合起来的性能带宽就非常可观了。***断点续传***

***独立的互联网存储服务一般都是做对象存储的，因为块存储是给计算机用的，对象存储是给浏览器等 HTTP 客户端用的。独立服务所提供的存储系统，访问都来自互联网，自然是做对象存储；与之相对应，大部分类 AWS 的主机服务商都会提供一个块存储服务搭配主机服务。***



***块存储***就像超跑，根本不在意能不能多载几个人，要的就是极限速度和高速下的稳定性和可靠性，各大厂商出新产品都要去纽北赛道刷个单圈最快纪录，千方百计就为提高一两秒，跑不进7分以内都看不到前三名。（块存储容量也不大，TB这个数量级，支持的应用和适用的环境也比较专业（FC+Oracle），在乎的都是IOPS的性能值，厂商出新产品也都想去刷个SPC-1，测得好的得意洋洋，测得不好自动忽略。）  
***文件存储***像集卡，普适各种场合，又能装数据（数百TB），而且兼容性好，只要你是文件，各种货物都能往里塞，在不超过性能载荷的前提下，能拉动常见的各种系统。标准POXIS接口，后车门打开就能装卸。卡车也不挑路，不像块存储非要上赛道才能开，普通的千兆公路就能畅通无阻。速度虽然没有块存储超跑那么块，但跑个80/100码还是稳稳当当.  
而***对象存储***就像海运货轮，应对的是"真.海量"，几十上百PB的数据，以集装箱/container（桶/bucket）为单位码得整整齐齐，里面装满各种对象数据，十万客户发的货（数据），一条船就都处理得过来，按照键值（KeyVaule）记得清清楚楚。海运速度慢是慢点，有时候遇到点网络风暴还不稳定，但支持断点续传，最终还是能安全送达的，对大宗货物尤其是非结构化数据，整体上来看是最快捷便利的。

从访问方式来说，块存储通常都是通过光纤网络连接，服务器/小机上配置FC光纤HBA卡，通过光纤交换机连接存储（IP SAN可以通过千兆以太网，以iSCSI客户端连接存储），主机端以逻辑卷（Volume）的方式访问。连接成功后，应用访问存储是按起始地址，偏移量Offset的方法来访问的。  
而NAS文件存储通常只要是局域网内，千兆/百兆的以太网环境皆可。网线连上，服务器端通过操作系统内置的NAS客户端，如NFS/CIFS/FTP客户端挂载存储成为一个本地的文件夹后访问，只要符合POXIS标准，应用就可以用标准的open，seek, write/read,close这些方法对其访问操作。  
对象存储不在乎网络，而且它的访问比较有特色，只能存取删（put/get/delete），不能打开修改存盘。只能取下来改好后上传，去覆盖原对象。//因为中间是不可靠的互联网啊，不能保证你在修改时候不掉线啊。所谓你在这头，对象在那头，所爱对象隔山海，山海不可平。

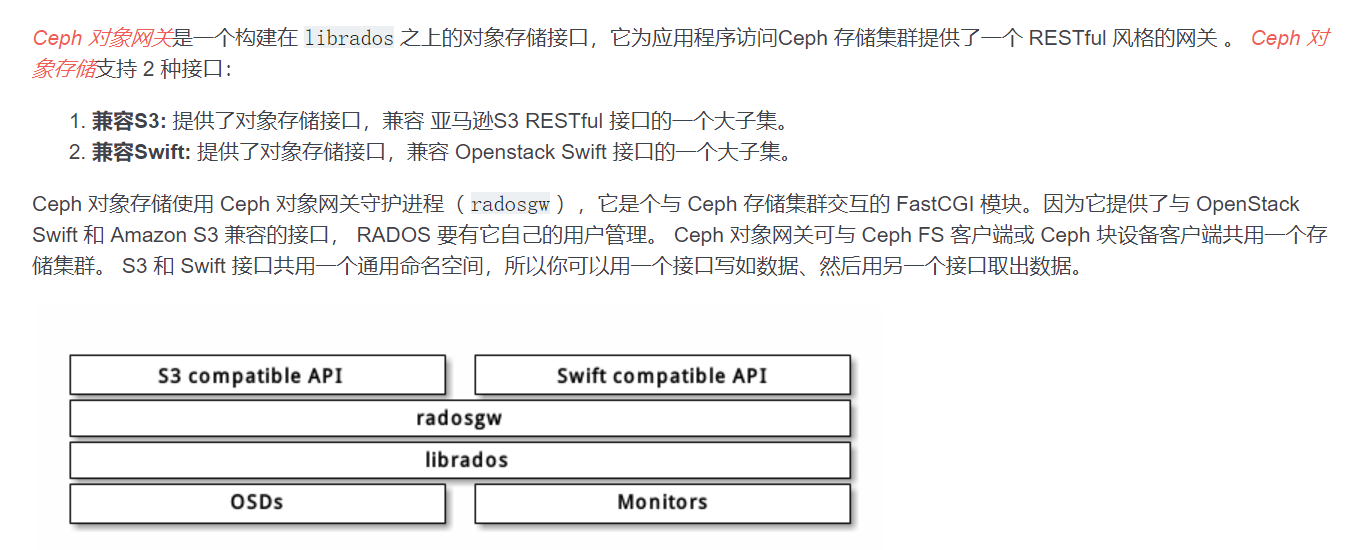
另外再说一点分布式存储的问题，以上三种存储都可以和分布式概念结合，成为分布式文件系统，分布式块存储，还有天生分布式的对象存储。  
对象存储的定义就把元数据管理和数据存储访问分开在不同的节点上，多个节点应对多并发的访问，这自然就是一个分布式的存储产品。而分布式文件系统就很多了，各种开源闭源的产品数得出几十个，在不同的领域各有应用。至于分布式的块存储产品就比较少，也很难做好。我个人认为这个产品形态有点违和，分布式的思想和块存储的设计追求其实是冲突的。前面讲过，块存储主要是图快，一上分布式肯定严重拖后腿，既然都分布开了，节点之间的通信必然增加额外负担，再加上CAP，为了保持一致性牺牲响应速度，得到的好处就是扩展性。这就像把超跑弄个铁索连环，哪里还可能跑出高速？链条比车都重了，穿起来当火车开吗？  
而文件存储原来也就是集装箱货车，大家连起来扮火车还是有可行性的。

**块存储可用于绝大部分通用业务场景下的数据存储**

**对象存储常用于互联网业务网站搭建、动静资源分离、CDN加速等业务场景。**

**文件存储适合企业部门间文件共享、高性能计算、容器服务等业务场景。**

1. 所发生的
2. Ceph对象存储：



所有 Ceph 部署都始于 [Ceph 存储集群](http://docs.ceph.org.cn/glossary/" \l "term-21)。基于 RADOS 的 Ceph 对象存储集群包括两类守护进程：term:对象存储守护进程（ OSD ）把存储节点上的数据存储为对象； term:Ceph 监视器（ MON ）维护集群运行图的主拷贝。一个 Ceph 集群可以包含数千个存储节点，最简系统至少需要一个监视器和两个 OSD 才能做到数据复制。

Ceph的核心组件包括Ceph OSD、Ceph Monitor和Ceph MDS

**Ceph OSD**：Objec Storage Device，它的主要功能是存储数据、复制数据、平衡数据、恢复数据等，与其他OSD间进行心跳检查等，并将一些变化情况上报给Ceph Monitor。一般情况下一块硬盘对应一个OSD，由OSD来对硬盘存储进行管理，当然一个分区也可以成为一个OSD。为了提高读写性能，还引入了Journal盘，允许Ceph OSD功能很快做小的写操作。

**Ceph Monitor**：负责监视Ceph集群，维护Ceph集群的健康状态，同时维护着Ceph集群种的各种Map图，比如OSD Map、Monitor Map、PG Map和CRUSH Map，这些Map统称为Cluster Map，Cluster Map是RADOS的关键数据结构，管理集群种的所有成员、关系、属性等信息以及数据的分发，比如当用户需要存储数据到Ceph集群时，OSD需要先通过Monitor获取最新的Map图，然后根据Map图和Object id等计算出数据最终存储的位置。

**Ceph MDS**：Ceph MetaData Server，主要保存的文件系统服务的元数据，但对象存储和块存储设备是不需要使用该服务的。

Radosgw:

Librados:

OSDs:

Monitors:

对象网关提供的功能：

* 用户管理：用户和权限
* 配额管理：允许再用户级别、用户拥有的bucket级别设置配额。配额包括一个bucket内允许的最大对象数和最大存储容量，大小单位是兆字节。
* 用量：会记录每个用户的用量数据，还可以通过指定日期范围来跟踪用户的用量数据。

索引查询

对象存储适用于大数据、IOT、备份归档等场景，具有EB级别的容量和三大存储中最高的数据可靠性。