文件系统了解及对比选型

分布式文件系统（Distributed file system，DFS），或是网络文件系统（Network File System），是一种允许文件透过网络在多台主机上分享的文件系统，可让多机器上的多用户分享文件和存储空间。分布式文件系统解决的最大的问题是资源共享的问题，因此分布式文件系统最大的特点是多个客户端可以访问相同的服务端

1. 专业名词

缩写	全称	注释
SDS	Software Defined Storag	软件定义存储技术
RAID	Redundant Array of Independent Disk	磁盘阵列技术
DG	Disk Group	硬盘组
VD	Virtual Disk	虚拟磁盘（等效磁盘组）
Mirroring	-	镜像技术，又称为复制技术(Replication)
Striping	-	条带（可提供并行的数据吞吐能力）
Erasure Code	-	纠删码（提供高可用性和高速读写能力）
RADOS	Reliable，Autonomic，Distributed Object Store	可靠的自主的分布式对象存储
OSD	Object Storage Device	对象存储设备
MDS	Metadata Server	元数据服务
MTTR	Mean Time To Restoration	平均恢复时间

2. 存储方式

块存储、文件存储、对象存储

2.1. 块存储

就好比硬盘一样，直接挂在到主机， 一般用于主机的直接存储空间和数据库应用(MySQL)的存储

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

2.2. 文件存储

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

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

2.3. 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），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。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

2.4. 性能对比

- **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。**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
- **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
- **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

2.5. 访问方式

- **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
- **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

3. 文件系统

开源分布式存储系统对比，还有很多其他，这里只列举这几款

存储系统	Ceph	Swift	HDFS	FastDFS	Ambry	MinIO
开发语言	C++	Python	Java	C	Java	Go
开源协议	LGPL	Apache	Apache	GPL3	Apache	Apache
存储方式	对象/文件/块	对象	文件	文件/块	对象	对象
在线扩容	支持	支持	支持	支持	支持	-
冗余备份	支持	支持	支持	支持	支持	-
单点故障	不存在	不存在	存在	不存在	不存在	-
易用性	一般	一般	一般	简单	简单	简单
跨集群	不支持	-	不支持	部分支持	不支持	-

存储系统	Ceph	Swift	HDFS	FastDFS	Ambry	MinIO
适用场景	大中小文件	大中小文件	大中文件	中小文件	大中小文件	大中小文件

对比了 Github 的 Star，MinIO 增长的很快，而且官方还有中文文档提供，中小企业使用不错

参考

- 为什么要有文件系统
- 开源分布式存储系统笔记
- 开源分布式存储系统的对比
- 常见分布式文件存储介绍、选型比较、架构设计
- 超简单，MinIO搭建私有化文件服务

上次更新时间: 2023-01-30 10:31:31