**0. 名词缩写**

RAID：独立冗余磁盘阵列 Redundant Arrays of Independent(Inexpensive) Disks

RMW：Read Modify Write

RCW：Read Construct Write

SCSI：小型计算机系统接口 Small Computer System Interface

iSCSI：Internet 小型计算机系统接口 Internet Small Computer System Interface

SATA：串行高级技术附件 Serial Advanced Technology Attachment

AHCI：串行 ATA 高级主控接口 Serial Advanced Host Controller Interface

NVMe：非易失性存储器主控制器接口 Non-Volatile Memory express

IOPS：每秒读写操作次数Input/Output Operations Per Second

RTO：恢复时间目标 Recovery Time Object

RPO：恢复点目标 Recovery Point Object

PCI：外设组件互联标准 Peripheral Component Interconnect

SCM：存储级内存 Storage Class Memory

OSD：基于对象存储设备 Object-based Storage Device

COW：写时复制 Copy On Write

CDB：命令描述符 Command Descriptor Block

MTTR：平均恢复时间 Mean Time To Restoration

MTTF：平均失效时间 Mean Time To Failure

SNTL：SCSI to NVMe Translation Layer

LSN：Logical Sector Number

PSN：Physical Sector Number

**1. 数据去重技术**

Deduplication（去重，消重）是近年来存储业界非常热门的一个技术，无论是Primary Storage，还是备份系统，抑或是云存储比如百度迅雷的网盘，都需要考虑dedup来减少冗余，降低成本。当然了各种系统对dedup的需求是不一样的。Primary Storage对性能要求高，特别是IOPS指标，这样很难实现在线的dedup，往往只会采用离线的dedup在闲时进行去重。备份系统则是各有千秋，去重备份系统的领头羊DataDomain是采用榨干CPU能力的在线dedup，由于有一些独特的技术，能做到吞吐量每小时数T。而网盘服务中，多用户可能存在重复文件，但各用户的带宽有限，往往是采用文件级去重，客户端在上传文件之前就进行文件hash计算，只传输云端没有的新文件。

Dedup从技术上说，分文件级去重和块级去重。文件级的去重，最原始是对整个文件做一个hash取得key值，实现简单，效果一般。如上所述，在多用户环境中重复文件很多的条件下，能取得一定的去重效果。但是在实践中，往往完全相同的文件还是不多见的。例如网盘中的电影文件，往往同一个电影同一个文件，有些用户获取文件时，部分数据不全（往往出现在BT下载中种子没了的情况下）。这时候文件还是可以流畅观看的，但是如果用全文件级别的hash，一点差异就会导致完全不同的hash值。而且动辄数G的文件做全文的hash消耗太大。据我所知，有些网盘服务是在大文件的某些特定偏移取出特定大小的数据块，然后算出多个hash值。如果两个文件的多个hash值匹配，就判断两个文件是相同文件。

块级的去重，除了能消除重复的文件，还能在同一个文件有部分变化时，只保留变化部分。而实现块级的去重，数据如何分块是关键。数据分块方法分为固定长度分块和可变长分块。固定长度分块实现起来也相当的简单，但是如果文件中有小小的新增或者删除改动，会造成旧的分块失效，从增删数据位置开始，所有的分块hash值都改变。

所以，变长分块才是最终解决之道。变长分块要达到的目标就是，当数据中有新增或者删除的数据时，变动的数据只影响周围1-2个数据块，只有这1-2个数据块的hash值发生改变，而后续更多的数据块还是按照原来的方式分块，hash值不变。这样一来，可以想到的解决方法是，变长分块要根据数据内容特征来分块。

**2. ReRAM**

ReRAM代表电阻式RAM，将DRAM的读写速度与SSD的非易失性结合于一身。ReRAM基于忆阻器原理，致力于商业化ReRAM的企业包括惠普、Elpida、索尼、松下、美光、海力士等等。

* SLC = Single-Level Cell ，即1bit/cell，速度快寿命长，价格超贵（约MLC 3倍以上的价格），约10万次擦写寿命
* MLC = Multi-Level Cell，即2bit/cell，速度一般寿命一般，价格一般，约3000---10000次擦写寿命
* TLC = Trinary-Level Cell，即3bit/cell，也有Flash厂家叫8LC，速度慢寿命短，价格便宜，约500次擦写寿命，目前还没有厂家能做到1000次。

**3. 文件存储、块存储与对象存储的异同**

通常来讲，所有磁盘阵列都是基于Block块的模式，而所有的NAS产品都是文件级存储。

* 块存储

以下列出的两种存储方式都是块存储类型：

（1） DAS（Direct Attach STorage）：是直接连接于主机服务器的一种储存方式，每一台主机服务器有独立的储存设备，每台主机服务器的储存设备无法互通，需要跨主机存取资料时，必须经过相对复杂的设定，若主机服务器分属不同的操作系统，要存取彼此的资料，更是复杂，有些系统甚至不能存取。通常用在单一网络环境下且数据交换量不大，性能要求不高的环境下，可以说是一种应用较为早的技术实现。

（2）SAN（Storage Area Network）：是一种用高速（光纤）网络联接专业主机服务器的一种储存方式，此系统会位于主机群的后端，它使用高速I/O 联结方式， 如 SCSI, ESCON 及 Fibre- Channels。一般而言，SAN应用在对网络速度要求高、对数据的可靠性和安全性要求高、对数据共享的性能要求高的应用环境中，特点是代价高，性能好。例如电信、银行的大数据量关键应用。它采用SCSI 块I/O的命令集，通过在磁盘或FC（Fiber Channel）级的数据访问提供高性能的随机I/O和数据吞吐率，它具有高带宽、低延迟的优势，在高性能计算中占有一席之地，但是由于SAN系统的价格较高，且可扩展性较差，已不能满足成千上万个CPU规模的系统。

* 文件存储

通常，NAS产品都是文件级存储。

NAS（Network Attached Storage）：是一套网络储存设备，通常是直接连在网络上并提供资料存取服务，一套 NAS 储存设备就如同一个提供数据文件服务的系统，特点是性价比高。例如教育、政府、企业等数据存储应用。

它采用NFS或CIFS命令集访问数据，以文件为传输协议，通过TCP/IP实现网络化存储，可扩展性好、价格便宜、用户易管理，如目前在集群计算中应用较多的NFS文件系统，但由于NAS的协议开销高、带宽低、延迟大，不利于在高性能集群中应用。

* 对象存储

总体上来讲，对象存储同兼具SAN高速直接访问磁盘特点及NAS的分布式共享特点。

核心是将数据通路（数据读或写）和控制通路（元数据）分离，并且基于对象存储设备（Object-based Storage Device，OSD）构建存储系统，每个对象存储设备具有一定的智能，能够自动管理其上的数据分布。

对象存储结构组成部分（对象、对象存储设备、元数据服务器、对象存储系统的客户端）：

（1）对象

对象是系统中数据存储的基本单位，一个对象实际上就是文件的数据和一组属性信息（Meta Data）的组合，这些属性信息可以定义基于文件的RAID参数、数据分布和服务质量等，而传统的存储系统中用文件或块作为基本的存储单位，在块存储系统中还需要始终追踪系统中每个块的属性，对象通过与存储系统通信维护自己的属性。在存储设备中，所有对象都有一个对象标识，通过对象标识OSD命令访问该对象。通常有多种类型的对象，存储设备上的根对象标识存储设备和该设备的各种属性，组对象是存储设备上共享资源管理策略的对象集合等。

（2）对象存储设备

对象存储设备具有一定的智能，它有自己的CPU、内存、网络和磁盘系统，OSD同块设备的不同不在于存储介质，而在于两者提供的访问接口。OSD的主要功能包括数据存储和安全访问。目前国际上通常采用刀片式结构实现对象存储设备。OSD提供三个主要功能：

a. 数据存储。OSD管理对象数据，并将它们放置在标准的磁盘系统上，OSD不提供块接口访问方式，Client请求数据时用对象ID、偏移进行数据读写。

b. 智能分布。OSD用其自身的CPU和内存优化数据分布，并支持数据的预取。由于OSD可以智能地支持对象的预取，从而可以优化磁盘的性能。

c. 每个对象元数据的管理。OSD管理存储在其上对象的元数据，该元数据与传统的inode元数据相似，通常包括对象的数据块和对象的长度。而在传统的NAS系统中，这些元数据是由文件服务器维护的，对象存储架构将系统中主要的元数据管理工作由OSD来完成，降低了Client的开销。

（3）元数据服务器（Metadata Server，MDS）

MDS控制Client与OSD对象的交互，主要提供以下几个功能：

a. 对象存储访问。

MDS构造、管理描述每个文件分布的视图，允许Client直接访问对象。MDS为Client提供访问该文件所含对象的能力，OSD在接收到每个请求时将先验证该能力，然后才可以访问。

b. 文件和目录访问管理。

MDS在存储系统上构建一个文件结构，包括限额控制、目录和文件的创建和删除、访问控制等。

c. Client Cache一致性。

为了提高Client性能，在对象存储系统设计时通常支持Client方的Cache。由于引入Client方的Cache，带来了Cache一致性问题，MDS支持基于Client的文件Cache，当Cache的文件发生改变时，将通知Client刷新Cache，从而防止Cache不一致引发的问题。

（4）对象存储系统的客户端Client

为了有效支持Client支持访问OSD上的对象，需要在计算节点实现对象存储系统的Client，通常提供POSIX文件系统接口，允许应用程序像执行标准的文件系统操作一样。

**4. 对块存储、文件存储、对象存储这三者的本质差别的理解**

（1）块存储

典型设备：磁盘阵列，硬盘

块存储主要是将裸磁盘空间整个映射给主机使用的，就是说例如磁盘阵列里面有5块硬盘（为方便说明，假设每个硬盘1G），然后可以通过划逻辑盘、做Raid、或者LVM（逻辑卷）等种种方式逻辑划分出N个逻辑的硬盘。（假设划分完的逻辑盘也是5个，每个也是1G，但是这5个1G的逻辑盘已经于原来的5个物理硬盘意义完全不同了。例如第一个逻辑硬盘A里面，可能第一个200M是来自物理硬盘1，第二个200M是来自物理硬盘2，所以逻辑硬盘A是由多个物理硬盘逻辑虚构出来的硬盘。）

接着块存储会采用映射的方式将这几个逻辑盘映射给主机，主机上面的操作系统会识别到有5块硬盘，但是操作系统是区分不出到底是逻辑还是物理的，它一概就认为只是5块裸的物理硬盘而已，跟直接拿一块物理硬盘挂载到操作系统没有区别的，至少操作系统感知上没有区别。此种方式下，操作系统还需要对挂载的裸硬盘进行分区、格式化后，才能使用，与平常主机内置硬盘的方式完全无异。

优点：

a. 这种方式的好处当然是因为通过了Raid与LVM等手段，对数据提供了保护。

b. 另外也可以将多块廉价的硬盘组合起来，成为一个大容量的逻辑盘对外提供服务，提高了容量。

c. 写入数据的时候，由于是多块磁盘组合出来的逻辑盘，所以几块磁盘可以并行写入的，提升了读写效率。

d. 很多时候块存储采用SAN架构组网，传输速率以及封装协议的原因，使得传输速度与读写速率得到提升。

缺点：

a. 采用SAN架构组网时，需要额外为主机购买光纤通道卡，还要买光纤交换机，造价成本高。

b. 主机之间的数据无法共享，在服务器不做集群的情况下，块存储裸盘映射给主机，再格式化使用后，对于主机来说相当于本地盘，那么主机A的本地盘根本不能给主机B去使用，无法共享数据。

c. 不利于不同操作系统主机间的数据共享：另外一个原因是因为操作系统使用不同的文件系统，格式化完之后，不同文件系统间的数据是共享不了的。例如一台装了WIN7/XP，文件系统是FAT32/NTFS，而Linux是EXT4，EXT4是无法识别NTFS的文件系统的。就像一只NTFS格式的U盘，插进Linux的笔记本，根本无法识别出来。所以不利于文件共享。

（2）文件存储

典型设备：FTP、NFS服务器

为了克服上述文件无法共享的问题，所以有了文件存储。

文件存储也有软硬一体化的设备，但是其实普通拿一台服务器/笔记本，只要装上合适的操作系统与软件，就可以架设FTP与NFS服务了，架上该类服务之后的服务器，就是文件存储的一种了。主机A可以直接对文件存储进行文件的上传下载，与块存储不同，主机A是不需要再对文件存储进行格式化的，因为文件管理功能已经由文件存储自己搞定了。

优点：

a. 造价交低：随便一台机器就可以了，另外普通以太网就可以，根本不需要专用的SAN网络，所以造价低。

b. 方便文件共享：例如主机A（WIN7，NTFS文件系统），主机B（Linux，EXT4文件系统），想互拷一部电影，本来不行。加了个主机C（NFS服务器），然后可以先A拷到C，再C拷到B就OK了。（例子比较肤浅，请见谅……）

缺点：

读写速率低，传输速率慢：以太网，上传下载速度较慢，另外所有读写都要1台服务器里面的硬盘来承担，相比起磁盘阵列动不动就几十上百块硬盘同时读写，速率慢了许多。

（3）对象存储

典型设备：内置大容量硬盘的分布式服务器

对象存储最常用的方案，就是多台服务器内置大容量硬盘，再装上对象存储软件，然后再额外搞几台服务作为管理节点，安装上对象存储管理软件。管理节点可以管理其他服务器对外提供读写访问功能。

之所以出现了对象存储这种东西，是为了克服块存储与文件存储各自的缺点，发扬它俩各自的优点。简单来说块存储读写快，不利于共享，文件存储读写慢，利于共享。能否弄一个读写快，利 于共享的出来呢。于是就有了对象存储。

首先，一个文件包含了了属性（术语叫metadata，元数据，例如该文件的大小、修改时间、存储路径等）以及内容（以下简称数据）。

以往像FAT32这种文件系统，是直接将一份文件的数据与metadata一起存储的，存储过程先将文件按照文件系统的最小块大小来打散（如4M的文件，假设文件系统要求一个块4K，那么就将文件打散成为1000个小块），再写进硬盘里面，过程中没有区分数据/metadata的。而每个块最后会告知你下一个要读取的块的地址，然后一直这样顺序地按图索骥，最后完成整份文件的所有块的读取。

这种情况下读写速率很慢，因为就算你有100个机械手臂在读写，但是由于你只有读取到第一个块，才能知道下一个块在哪里，其实相当于只能有1个机械手臂在实际工作。

而对象存储则将元数据独立了出来，控制节点叫元数据服务器（服务器+对象存储管理软件），里面主要负责存储对象的属性（主要是对象的数据被打散存放到了那几台分布式服务器中的信息），而其他负责存储数据的分布式服务器叫做OSD，主要负责存储文件的数据部分。当用户访问对象，会先访问元数据服务器，元数据服务器只负责反馈对象存储在哪些OSD，假设反馈文件A存储在B、C、D三台OSD，那么用户就会再次直接访问3台OSD服务器去读取数据。

这时候由于是3台OSD同时对外传输数据，所以传输的速度就加快了。当OSD服务器数量越多，这种读写速度的提升就越大，通过此种方式，实现了读写快的目的。

另一方面，对象存储软件是有专门的文件系统的，所以OSD对外又相当于文件服务器，那么就不存在文件共享方面的困难了，也解决了文件共享方面的问题。

所以对象存储的出现，很好地结合了块存储与文件存储的优点。

（4）最后

为什么对象存储兼具块存储与文件存储的好处，还要使用块存储或文件存储呢？

a. 有一类应用是需要存储直接裸盘映射的，例如数据库。因为数据库需要存储裸盘映射给自己后，再根据自己的数据库文件系统来对裸盘进行格式化的，所以是不能够采用其他已经被格式化为某种文件系统的存储的。此类应用更适合使用块存储。

b. 对象存储的成本比起普通的文件存储还是较高，需要购买专门的对象存储软件以及大容量硬盘。如果对数据量要求不是海量，只是为了做文件共享的时候，直接用文件存储的形式好了，性价比高。

**5. 大数据时代的云存储技术**

（1）什么是云存储

云存储是在云计算(cloud computing)概念上延伸和发展出来的一个新的概念，是指通过集群应用、网格技术或分布式文件系统等功能，将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统。当云计算系统运算和处理的核心是大量数据的存储和管理时，云计算系统中就需要配置大量的存储设备，那么云计算系统就转变成为一个云存储系统，所以云存储是一个以数据存储和管理为核心的云计算系统。

（2）云存储分类

按照服务对象可以把云存储分类如下：

a. 公共云存储。公共云存储可以以低成本提供大量的文件存储。供应商可以保持每个客户的存储、应用都是独立的、私有的，公共云存储可以划出一部分来用作私有云存储。

b. 私有云存储。通过私有云存储，一个公司可以拥有或控制基础架构，以及应用的部署。私有云存储可以部署在企业数据中心或相同地点的设施上。私有云可以由公司自己的 IT 部门管理，也可以由服务供应商管理。

相比传统存储模式，云存储的租购模式显得更加灵活方便，其实，企业自己构建一个数据中心需要购买硬件等必备设施及复杂多变的维护管理。

c. 混合云存储。这种云存储把公共云和私有云结合在一起。主要用于按客户要求的访问，特别是需要临时配置容量的时候。从公共云上划出一部分容量配置一种私有云，可以帮助公司面对迅速增长的负载波动或高峰时很有帮助。

尽管如此，混合云存储带来了跨公共云和私有云分配应用的复杂性。

（3）云存储具有的优势

作为大数据时代的云存储技术，与传统存储技术相比，具有如下明显优势：

a. 灵活方便。对于中小企业或个人用户，完全可以将数据的创建与维护委托给云服务提供商，而只是租用云服务提供商的服务即可，用户不必考虑存储容量、 存储设备类型、 数据存储位置，以及更多的关于数据的可用性、 可靠性和安全性等繁杂的技术层面。避免了购买硬件设备及技术维护而投入的精力，节省下来的大量时间可以用于更多的工作业务发展。

b. 成本低廉。就目前来说，企业在数据存储上所付出的成本相当大，因为企业要建立一套存储系统不仅需要购买硬件等基础设施，同时，系统维护还需要专门的人员，企业的存储空间及管理费用都面临巨大挑战，企业必须更新或增加各种数据存储和管理设备，而且还要负担逐渐高涨的管理成本。为了减少这种成本压力，好多企业将大部分数据迁移至云存储上，所有的升级、维护等管理任务均由云存储服务提供商来完成。因此，可以将数据存储与管理的成本降到最低，同时，还能获得最优良的数据存储服务。

c. 量身定制。随着大数据的到来，传统的存储模式已不再适用企业的数据存储，企业急需一种新的存储方式来满足企业数据存储的个性化需求。私有云即可满足企业这一个性化需求。云服务提供商可以专门为企业客户提供一种量身定制的云存储服务解决方案，也可以由企业自己的 IT 机构部署一套私有云服务架构。 企业不仅可以得到最优质的个性化服务，还能在一定程度上降低安全风险。

（4）云存储发展需要考虑的主要问题

由于云存储具有传统数据存储模式不具备的诸多优势，越来越多的中小企业正在将自己的数据中心逐渐转移至云端。而大型企业除了租用公共云存储服务以外，也开始着手建立自己的私有云存储数据中心。 但是，云存储仍处于快速发展阶段，云存储要想得到广泛应用，还有待进一步完善和改进。

a. 安全问题。数据存储在云计算中心，安全问题始终是用户最为关注的问题。用户的敏感信息和个人隐私数据如何才能保证其具有安全性和私密性，云存储服务提供商应根据分布式文件中可能存在的安全威胁和安全需求，来制定相应的安全策略，以便在计算机中实施相应的保护机制，确保用户数据避免丢失或被窃。

b. 可靠性问题。可靠性是数据存储系统最基本也是最关键的一项指标。数据的备份复制是云存储管理中心必须考虑的问题，而且数据的备份复制可以大大提高云存储系统的可靠性和性能，同时能够增强系统的容错能力。如果云存储数据遭到破坏，云存储服务提供商应该能够快速全面地恢复数据，确保用户数据的完整性。

（5）结束语

云存储不仅仅是存储，更多的是应用，是云计算时代的一场存储革命，随着云存储的安全性、 可靠性、 实用性等存储技术的不断成熟，人们对存储技术的认识不断提高，云存储成本的不断降低，一定会形成一个安全可靠、实用快速的云数据存储与访问系统。云存储必将广泛应用于各行各业，云存储必然是未来数据存储的发展趋势。