云存储架构深度解析：分布式架构和对象存储技术

云存储的实现方式有多种，本文提出用分布式架构以及对象存储构建云存储。针对云存储的应用需求，分析了分布式架构和对象存储的实现方法和优势。

中国电信广州研究院 | 郑文武

云存储是应用层面和业务层面的概念。广义的云计算包括了云存储。但云存储服务作为实体是单独存在的。以提供计算能力为主的狭义的云计算，其使用的存储和云存储并不相同，例如，提供计算服务的云计算，其内部使用的存储一般都是传统的存储，甚至比传统的存储还要简单。其分布式节点全部为计算节点，没有独立的存储节点。

本文讨论的云存储，是指提供云存储服务的存储系统。理论上，云存储服务可以由云存储软件加上任何普通存储设备实现，但实践中，云存储软件和物理设备是紧密结合在一起的，厂家提供的云存储产品一般是硬件设备，在这些硬件设备上运行相应的软件，而并非提供单独的软件。

分布式架构

云存储首先在网络物理架构上采用分布式架构。如下图1所示。

云存储存储将数据存放到多个节点上，节点数目一般达到几十、几百。这些节点一般通过IP网络进行连接。当节点数目较多，并且节点分布地域较广时，可以由多个节点组成站点，多个站点再组成分布式存储系统。

分布式架构使云存储建立在众多节点而不是单个节点之上。一般而言，这些节点可以分为2类，一类节点是传统存储厂家常用的中高端存储，另一类节点是基于X86服务器。应该说，后者更具备云存储的本质。

目前的分布式存储系统一般都是基于MapReduce原理。云存储本质上也是云计算，不过这种云计算具有以下特点：计算量较小而数据量非常大，每个节点的存储空间较大。

图2描述了分布式存储数据和控制流。云存储输入输出数据过程如下：

1)客户端向门户/管理节点发起数据存储请求。在很多云存储系统中，可能并没有专门的门户/管理节点，门户/管理节点只是普通的存储节点。

2)门户/管理节点根据客户端数据特点以及各存储节点的负载状态，将数据进行分拆，交由多个存储节点进行处理、存储。

3)门户/管理节点完成数据分拆和任务分派后，数据在客户端和存储节点直接流动而无需经由门户/管理节点转发。

4)存储节点完成对数据进行一些校验、加密（如果有必要的话）处理，将数据写入到存储介质。或将数据从存储介质上读出，进行解密（如果有必要的话）。

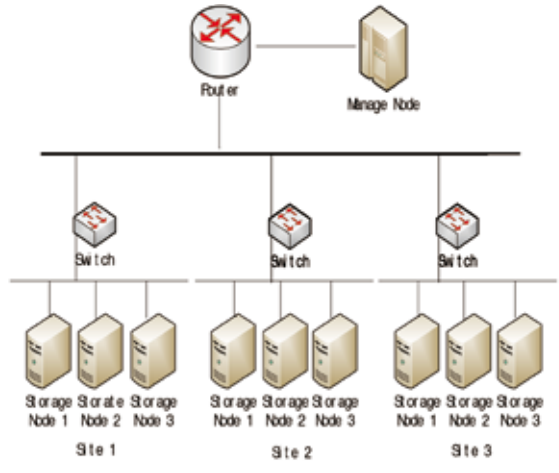

图1 云存储网络物理架构

分布式架构的优势

云存储的分布式实现，使其具备以下优势：非常高的可靠性、较高的性能和非常高的伸缩性。我们下面探讨为什么分布式具有这样的优势。

一、非常高的可靠性

云计算的理念就是不相信单个节点具有很高的可靠性，云计算认为单个节点出现故障是常态。云存储也遵循这个理念。多节点为云存储保护数据提供了便利。传统存储对数据的采用RAID5、RAID0等方式保护。这些保护方式处于较低层次，保护方式的种类也较少。而云存储的数据保护措施，是由运行在操作系统之上的程序所实现，处于较高层次。目前云存储一般支持ErasureCode编码存储，将数据分为N+M块，其中M为校验数据。只要任意N块可用，即可恢复数据。可用看出，云存储的数据保护方式更加灵活，保护的力度更加强大。并且，保护措施可以跟随最新技术随时改进。

因此可以说，基于分布式架构的云存储的可靠性要高于传统存储。

二、较高的性能

云存储的性能度量，本质上也可也归结为带宽和IOPS。从带宽上看，传统的中高端存储磁盘数量很多，如果充分能够利用磁盘的性能，则传统存储可也拥有非常大的带宽。但是受制于接口，传统存储的带宽和其具备的磁盘数量并不相称。接口成为传统存储带宽的瓶颈。

云存储的分布式架构，突破了传统存储的接口限制。云存储每个节点的接口数量可以接近传统存储，所有节点的接口相加，则接口数量远远超过传统存储。因

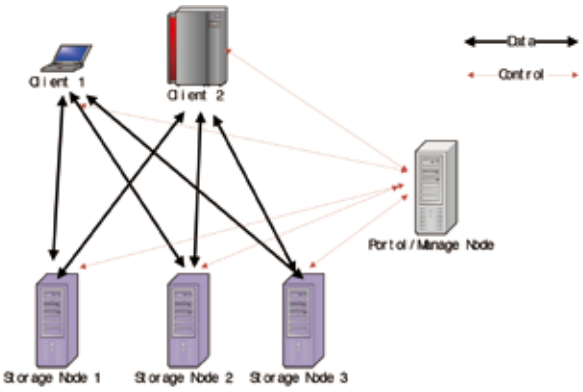

图2 分布式存储数据和控制流

此云存储的带宽理论上可以数十倍、数百倍于传统存储，但新的性能瓶颈产生了，就是网络带宽。所以云存储的带宽没能达到节点所能够提供的理论带宽。但一般也超过了传统存储的带宽。因此，云存储比较适合对带宽需求较大的应用。

由于分布式架构的接口数量众多优势并不在磁盘小数据操作中体现出来，并且传统存储的CPU是适合存储操作的专业CPU，操作系统也是精巧的适合存储操作的定制操作系统，而云存储存储节点的CPU为通用CPU，操作系统一般也是通用操作系统，所以对于瓶颈为CPU处理能力的IOPS性能，云存储比起传统的高端存储处于劣势。

三、非常高的伸缩性

伸缩性也是云存储超越传统存储的一大优势。云存储的分布式节点彼此之间是松耦合关系，由于数据冗余和保护，删除节点不会造成数据丢失或出错。因此增加或删除一个节点非常方便。

前面已经讨论过，在网络带宽范围内，云存储的带宽取决于存储系统的接口数量，而接口数量和节点数量呈线性关系，所以系统的带宽和节点数量在一定范围内保持良好的线性关系。

同时，云存储的节点大部分也是通用的X86服务器，因此和云计算一样，云存储的计算能力也具有良好的伸缩性，取决于计算能力的IOPS因而也具有良好的伸缩性。

所以，云存储比传统存储具有更好的伸缩性。

对象存储原理

云存储的存储数据很多是小文件，海量的小文件的数据存放、搜索、定位如果用传统存储提供的块和文件两种存储方式，则十分耗费时间。为提高存储性能，云存储采用了第三种存储技术：对象存储技术。

对象存储引进了2个新的概念：Ojbect（对象）和OSD

对象是包含文件数据及相关属性信息，可以进行自我管理。对象包括以下内容：一是ID，用于标识对象；二是数据，对象的主体；三是元数据，描述对象存放方式、存放位置的数据；四是属性，根据需要而定义的其他描述对象的数据，如QoS等。

对象可以维护自己的属性，简化了存储系统的管理任务，增加了灵活性。对象大小可以不同，可以包含整个数据结构，如文件、数据库表等。对象时存储的基本单元。

对象包括多个类型，如图3所示。

OSD是包含对象的集合。每个OSD都是一个智能设备，具有自己的存储介质、处理器、内存及网络系统等，负责管理本地的Object，是对象存储系统的核心。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使用Object对所保存的数据进行管理。它将数据存放到磁盘的磁道和扇区，将若干磁道和扇区组合起来构成Object，并且通过此Object向外界提供对数据的访问。每个Object同传统的文件相似，使用同文件类似的访问接口，包括Open、Read、Write等。但是两者并不相同，每个Object可能包括若干个文件，也可能是某个文件的一部分，且是独立于操作系统的。除了具体的用户数据外，OSD还记录了每个Object的属性信息，主要是物理视图信息。这些信息存放到OSD上，大大减轻了元数据服务器的负担，增强了整个存储系统的并行访问性能和可扩展性。

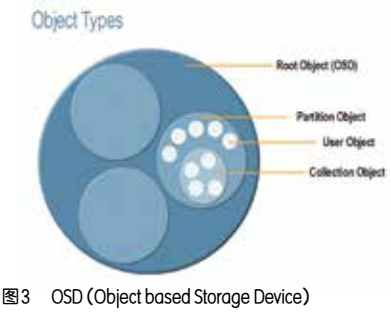

图3 OSD（Object based Storage Device）

而传统存储的底层数据以块状方式组织、管理，上层通过树状结构查找到数据对应的底层的数据块，这个查找过程对于海量的小文件非常耗时，查找定位数据的时间一般要超过实际数据读写的时间。

对象存储通过Hash等索引方法，直接将数据定位到存储介质的具体位置，无需通过逐层查找。在物理上是扁平结构。但在逻辑上，对象存储也提供了虚拟的树状结构的逻辑视图，而且这种虚拟的树状结构对于不同的用户可以不同。

对象存储系统提供给用户的也是标准的POSIX文件访问接口。接口具有和通用文件系统相同的访问方式，同时为了提高性能，也具有对数据的Cache功能和文件的条带功能。同时，文件系统必须维护不同客户端上Cache的一致性，保证文件系统的数据一致。

采用对象存储技术的云存储数据存储过程如下：客户端发出读写请求；文件系统向元数据服务器发送请求，获取要读取的数据所在的OSD；然后直接向每个OSD发送数据读取请求；OSD得到请求以后，判断要读写的Object，并根据此Object要求的认证方式，对客户端进行认证，如果此客户端得到授权，则对数据进行读写。

对象存储的突出优势就是加大数据存储位置的定位时间，加快了小文件的存储速度。而云存储的存储文件相当一部分为小文件，因而对象存储技术极大地提高了云存储的性能。

本文分析了分布式架构和对象存储技术。从中可以看出，在很多情况下采用分布式架构和对象存储技术，可以很好满足云存储的四个要求：足够的性能、高可靠、海量数据和低成本。

编辑 / 王熙 wangxi@bjxintong.com.cn