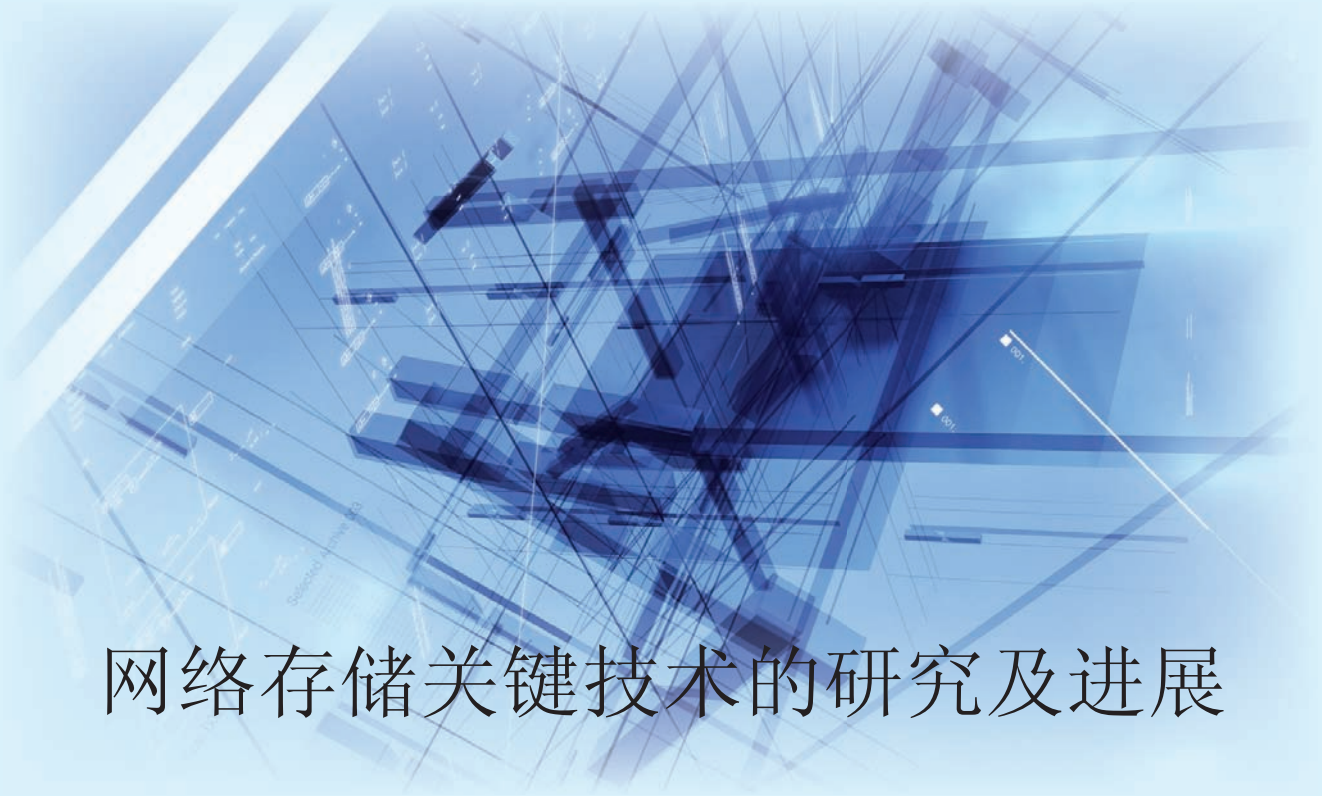

网络存储关键技术的研究及进展

冯　丹　　华中科技大学

【摘　要】文章介绍了几种典型的网络存储系统结构，包括直接连接存储DAS、附网存储NAS、存储区域网SAN以及面向对象存储OBS，并比较了其特点，对近年来发展迅速的基于对象的存储系统和云存储相关技术的研究现状和发展趋势进行了深入分析。

【关键词】网络存储　　存储系统　　云存储

处理、传输和存储是当今数字信息技术的三大基石，计算设施、网络设施以及存储设施合在一起，成为以互联网为代表的现代信息社会的基础设施。以往的存储设备隐藏在服务器中，但随着数字化信息的爆炸性增长，存储系统已成为独立的部分，而且在当前信息设施投资中比例已超过50%。随着数据量的进一步快速增长，对存储的需求持续高涨，存储成为整个IT业发展的新动力，新的技术和产品层出不穷，市场已经十分巨大，而且还在迅速扩展。

收稿日期：2009年6月3日

1　几种典型网络存储系统结构

自计算机诞生以来，随着大规模集成电路技术的发展，计算机系统中CPU计算速度以每年40%~100%的速率提高，而磁盘寻道速度仅以每年7%的速率增长[2]，因此，计算机系统中CPU/主存与存储子系统之间的访问速度差距越来越大。为了弥补这个差距，各国学者把注意力放在了存储系统并行访问能力的提高上，技术上最先获得突破的是磁盘阵列技术（RAID，Redundant Array of Inexpensive Disks）。RAID通过将数据分条存储在不同磁盘上，采用交叉读写技术，利用时间和空间的重叠，对多个磁盘进行并行

I/O访问，从而大大提高了访问性能。

上世纪九十年代以来，随着网络技术的发展与处理能力的大幅提高，传统的单机数据处理方式被依附在网络上的以数据为中心的数据处理方式所取代，使存储系统与网络系统结合起来，产生了网络存储系统。基本的网络存储系统结构包括传统以服务器为中心的直接连接存储（DAS，Direct Access Storage）、附网存储（NAS，Network Attached Storage）、存储区域网（SAN，Storage Area Network）以及最近研究的——面向对象存储（OBS，Object-Based Storage）。

1.1 直连存储DAS

传统的直连存储DAS结构中，将具有块接口的存储设备（如磁盘、阵列）通过专用I/O通道，直接连接到文件服务器上，存储设备相当于服务器的一部分，由服务器提供存储管理与对外服务。在DAS结构存储系统中，数据的传输是以服务器为中心的，可以方便地集中管理数据，具有比较好的数据安全性。但是，客户访问存储系统中的数据时，数据需要在存储设备和服务器间多次转发，尽管文件服务器并不关心数据内容，通常也不对数据本身进行处理，但数据请求与传送都需要文件服务器的介入。当大规模用户进行数据访问时，给服务器的存取转发控制带来非常大的开销，使得文件服务器成为了整个系统中的性能瓶颈，对系统整体读写性能与可扩展性产生很大影响。

1.2 附网存储NAS

NAS是一种以数据为中心的存储结构，存储子系统不再通过专用I/O通道附属于某个服务器，而是通过专门系统的定制，将通用服务器上无关的功能去掉，只保留存储相关功能，可以看成是一台专门负责存储的“瘦”服务器，具有比DAS更高的读写性能。NAS提供文件级数据访问，支持NFS与CIFS网络文件协议，实现异构平台之间的数据级共享，在文件级别上建立安全机制也很容易。但是，NAS没有从根本上改变服务器/客户机的访问方式，因此当客户端数目或来自客户端的请求较多时，NAS服务器仍将成为系统的瓶颈。

1.3 存储区域网SAN

SAN对前两种存储系统结构进行了比较大的改进，真正地将存储子系统从服务器上分离出来独立地连接在高速专用网上的，是一种以网络为中心的存储结构，目前典型两种结构是基于光纤通道的FC-SAN和基于IP网络的IP-SAN。客户通过高速专用网与存储设备连接在一起，通过虚拟化软件进行存储系统的集中管理，具有较好的扩展性。SAN中的服务器专门用来存放元数据，元数据描述了数据本身的属性，完成文件到存储设备物理块的映射。客户在访问存储系统时，通过从元数据服务器得到的元数据，直接访问存储设备，避免了传统服务器因转发带来的延迟，使得SAN具有较高的性能。但是，由于SAN是以块为访问接口，其安全性管理非常有限；随着存储设备与客户数量的提高，元数据服务器的负载过大，成为系统中的潜在瓶颈。

1.4 基于对象的存储OBS

OBS采用对象接口，大小动态可变，吸收了NAS与SAN的优点，既有“块”接口的快速，又有“文件”接口的便于共享；同时克服了NAS与SAN的不足，削弱了影响性能提升的环节，系统具有较高的性能与可扩展性。OBS最显著的特点就是把数据存储的物理视图下放到对象存储设备（OSD，Object-Based Storage Device）上，元数据服务器上只负责维护全局逻辑视图，用户在进行数据传输时，直接与OSD通信，元数据服务器没有直接干预，从而大大减轻了元数据服务器的负担，弱化了系统中的瓶颈环节。

图1描述了上面四种存储系统的结构，可以看到，在SAN与OBS中客户都是可以与存储子系统直接传递数据，而OBS系统将文件存储的物理视图下放到存储设备上，避免了SAN中元数据服务器形成的瓶颈。

2 基于对象的存储系统

基于对象存储中最基本的概念就是对象（Object），它是一种数据的逻辑组织形式，是容纳了长度可变的数据块和可扩展的存储属性的基本容器，提供与文件类似的访问方式，例如打开、关闭、读/写等。

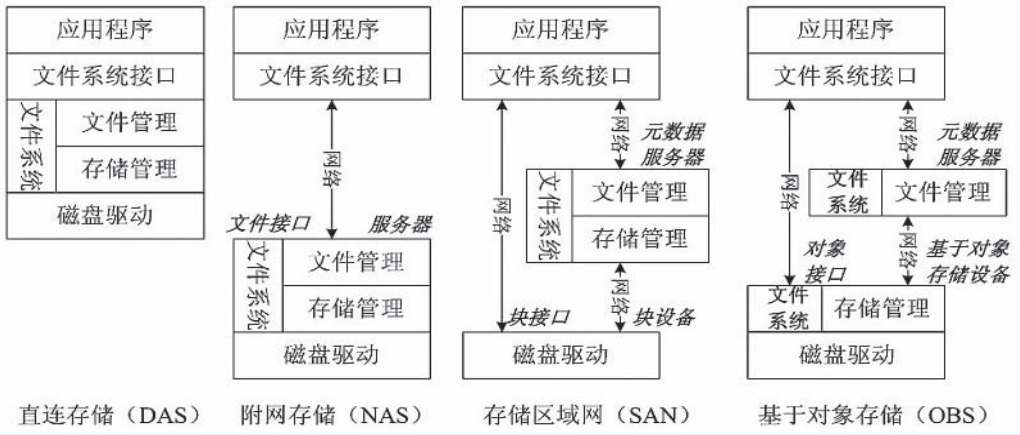

图1　四种存储系统结构图

基于对象存储系统的最基本思想就是将存储系统中的数据交互分为数据通路与控制通路两部分，是一种带外传输方式，由客户、MDS与OSD组成了三方架构，如图2所示。其中的控制通路负责数据的控制、管理功能，由MDS实现；数据通路负责数据的存储管理功能，数据的存储管理功能被分离出来交由智能的OSD实现。通过数据通路与控制通路的分离，将元数据服务器与存储设备分离，大大减轻了读写过程中元数据服务器的负担。客户进行文件访问时，首先通过控制通路从MDS得到该文件在OSD集群上的分布信息与访问授权命令，之后通过数据通路直接访问OSD。系统通过高速网络连接起来，共同协作为用户提供服务，不存在明显的瓶颈环节。

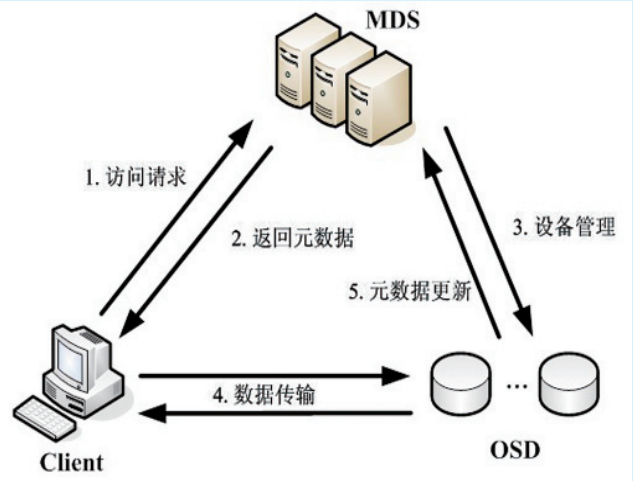

图2　MDS、OSD与客户端之间的交互

2.1 MDS、OSD与客户端交互过程

MDS、OSD与客户端之间的交互过程为：

（1）用户将文件读写请求提交客户端，客户端将要访问的文件名、数据长度等信息以一定的格式传递给MDS；

（2）MDS收到客户端请求，对客户进行安全检测，通过后，将该次访问涉及到的元数据与授权证书发送给客户；

（3）MDS通知OSD有客户即将与其进行数据传输，做好数据发送或接收的准备，并将相关的安全认证信息告诉此OSD；

（4）客户使用MDS反馈的元数据和安全授权，与相应服务器直接进行数据交换；

（5）文件访问结束，OSD向MDS通知数据传输完毕，并提交对象更新信息，MDS进行相应元数据的更新。

2.2 各种原型系统

基于对象的存储系统具有性能高、可扩展性好等诸多优点，它的设计理念一经提出，便成为学术界与工业界关注的热点，众多机构和厂商先后推出了各具特色的原型系统。

（1）NASD

基于对象的存储的研究最早起源于卡内基梅隆大学并行数据实验室的NASD（Network-Attached Security Disks）项目，基本思想是将处理器集成到磁盘驱动器，使它具有一定的智能，能够独立管理其自身的安全、存储和网络通信。NASD采用一种命令接口结构将文件管理器的大部分工作转移到磁盘上。对于流量大、要求迅速响应的操作，如读/写，可直接对磁盘驱动器进行操作；对于带宽依赖小的操作，如全局目录空间的管理，则由专门的文件服务器处理。NASD的提出从系统架构上给出了基于对象存储的原始模型。

（2）Lustre

Lustre是Cluster file system公司在SAN存储结构基础上加以改进，得到的基于对象存储的高性能存储系统。Lustre对象存储系统就是由客户端（Client）、存储服务器（OST，Object Storage Target）和元数据服务器（MDS）三个主要部分组成，三部分通过高速互联网连接，是一种典型的三方存储架构。Lustre提供标准的POSIX接口，客户端可以透明地访问Lustre系统中的数据，而无需知道这些数据的具体位置。Lustre具有较高的读写性能与可扩展性，已经在美国多个国家实验室的高性能计算中得到应用。

（3）ActiveScale

ActiveScale是Panasas公司推出的一个商业化的对象存储系统，应用于大规模Linux集群环境，该系统硬件环境由StorageBlade、DirectorBlade和高速互连网络组成。StorageBlade就是对象存储体系结构中的OSD，是具有一定智能性的存储设备，可以管理存储刀片上的数据；DirectorBlade充当MDS的角色，具有很强的处理能力与高速的互连网接口。整个系统的架构与工作流程与传统对象存储系统模型一致，同时设计了专用的Panasas文件系统，使ActiveScale在Linux集群环境中性能表现优异，可扩展性良好。

（4）Storage Tank

Storage Tank是IBM公司参考对象存储的设计思想，在SAN存储架构上设计出的异构可扩展存储系统。Storage Tank采用缓存策略在客户端缓存文件元数据和数据，提高了访问速度并减少了MDS的负载；采用分布式元数据管理方法，将文件系统元数据划分为多个文件集，不同的文件集可以分配到不同的元数据服务器处理，缓解元数据服务器瓶颈问题。将基于策略分配、卷管理和文件管理引入到数据存储领域，提供了一个基于锁的数据一致性的模型，使每个客户端可以像使用本地文件系统一样使用分布式文件系统。

（5）HUST-OBS

HUST-OBS是华中科技大学开发的基于对象的存储系统，其主要特点是：引入主动存储对象概念，结合磁盘阵列技术构成基于对象的存储设备OSD（Object based Storage Device），它既可独立使用提供存储服务，也可作为对象存储系统的存储节点；OSD根据对象属性历史记录自动分析负载规律，通过自学习和策略触发机制，实现存储主动服务；多OSD构成系统时，由存储系统实现自我组织与管理，针对不同应用环境实现自我优化调节，使系统整体性能最佳；利用对象“封闭”特性，使系统具有安全性；通过数据、节点冗余等提高系统可靠性，在故障出现时实现快速恢复。即HUST-OBS自组织对象存储系统具有如下特性：自组织（Self-Organize）、自调节（Self-Adjust）、自安全（Self-Secure）和自愈（Self-Heal）。

3 云存储

云存储（Cloud Storage）是在云计算（Cloud Computing）概念上延伸和发展而来的。与云计算类似，它是指通过集群应用、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虚拟化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。

云存储概念一经提出，就得到了众多厂商的关注和支持。Amazon推出Elastic Compute Cloud（EC2：弹性计算云）云存储产品，为用户提供互联网服务同时提供更强的存储和计算功能。内容分发网络服务提供商CD Networks和云存储平台服务商Nirvanix发布了一项新的合作，并宣布结成战略伙伴，以提供云存储和内容传送服务集成平台。微软推出了提供网络移动硬盘服务的Windows Live SkyDrive Beta测试版。近期，EMC宣布加入道里可信基础架构项目，致力于云计算环境下关于信任和可靠度保证的研究，IBM也将云计算标准作为全球数据备份中心的3亿美元扩展方案的一部分。

云存储可能获得成功的原因在于其具有相对较低的价格，而且具备可扩展性、灵活性、访问便利性和厂商选择性多等优点。例如，目前企业级存储区域网SAN的存储设备每GB成本约为20美元，相比之下，云计算存储设备每GB的成本仅为1美元，其成本也远低于EMC的Symmetrix或者Clariion阵列、IBM DS8000、DS4000和NetApp阵列等的存储设备。在可扩展性和灵活性方面，以EMC近日推出的一款被命名为Atmos的基础信息管理应用软件为例，它能辨认

世界各地网页的数据副本，可以通过策略和自我配置使用户管理其分支机构。Atmos可以提供基础信息管理服务，因此数据会按照不同客户的需求以不同的进度被分类发送到客户手中，从而可以灵活有效地为客户服务。

云存储作为一项新的存储技术，有望把管理及保护数据的负担转移给云存储提供商，有效降低大规模存储系统的总体拥有成本。但云存储服务进入实用化还有不少研究开发工作要做。最近Amazon的Simple Storage Service（S3：简单存储服务）出现一些故障，导致S3离线3个小时；Google GMAIL的偶发故障也使得其用户逐渐开始关注对数据内容的本地备份等。2009年4月6号开始的SNW（Storage Networking World）详细地分析了云存储平台方面的技术趋势和创新，包括使用大众化存储节点组成的网格架构、全局命名空间，也包括云存储对存储基础架构的需求、Web服务描述语言（WSDL）如何描述存储基础架构的功能、基于Web的文件访问协议等等。因此云存储发展道路上，还要重点解决性能、安全性、可靠性和支持性等多方面问题，为用户提供多样化的存储服务。

4 结束语

本文介绍了几种典型的网络存储系统结构，分析了基于对象的存储系统和云存储相关技术的研究现状和发展趋势，可为网络存储研究和应用提供一定参考。

参考文献

[1] D A Patterson, G A Gibson, R H katz. A Case for Redundant Array of Inexpensive Disks (RAID) [J]. ACM SIGMOD, June 1998: 109-116.
[2] 张江陵，冯丹. 海量信息存储. 北京：科学出版社[M]，2003: 105～107.
[3] Garth A Gibson, Rodney Van Meter. Network Attached Storage Architecture, Communications of the ACM[J]. Nov, 43(11): 2000: 37-45.
[4] Tom Clark. IP SANs: A Guide to iSCSI, iFCP, and FCIP Protocols for Storage Area Networks[M]. Addison-Wesley, 2002: 1-10.
[5] Mesnier M, Ganger G. R, Riedel E. Object-based Storage[J]. Communications Magazine, IEEE, 2003, 41(8): 84-90.
[6] Mesnier M, Ganger G R, Riedel E. Object-based Storage: Pushing More Functionality into Storage[J]. Potentials, IEEE, 2005, 24(2): 31-34.
[7] Garth A Gibson, David F Nagle, William Courtright, et al. NASD Scalable Storage Systems[M]. USENIX 1999, Linux Workshop. Monterey, CA, America: USENIX, 1999: 1-6.
[8] Cluster File Systems Inc. Lustre: A Scalable, High-performance File System. Cluster File Systems, Inc Whiter Paper[EB/OL]. http://www.clusterfs.com. 2002: 1-12.
[9] Tang Hong, A Gulbeden, Zhou Jingyu, et al. The Panasas ActiveScale Storage Cluster- Delivering Scalable High Bandwidth Storage[J]. In: Proceedings of the ACM/IEEE SC2004 Conference on Supercomputing. 2004. 53-62.
[10] J Menon, D A Pease, R Rees, et al. IBM Storage Tank-A Heterogeneous Scalable SAN File System[J]. IBM SYSTEMS JOURNAL, 2003, 42(2): 250-261. ★

【作者简介】

冯　丹：华中科技大学教授，博士生导师，现任华中科技大学计算机学院副院长、973项目“下一代互联网信息存储组织模式与核心技术研究”首席科学家、863重大项目“海量存储系统关键技术”总体专家组组长。主要从事网络存储系统、分布式并行存储系统、容错等方面的研究工作。