一种融合NAS和SAN技术的存储网络系统*)

傅湘林 谢长生 曹 强 刘朝斌

(华中科技大学计算机学院外存储系统国家重点实验室 武汉430074)

A Storage Network System that Merges NAS and SAN

FU Xiang-Lin XIE Chang-Sheng CAO Qiang LIU Zhao-Bin

(National Storage System Laboratory, Huazhong University of Science and Technology, Wuhan 430074)

Abstract With the increasing explosively of data in network, more and more people pay attention to the networked storage, currently the main technology of networked storage is NAS (Network Attached Storage)and SAN (Storage Area Network). They are different, and they are not competed but complement, they are used in different occasion. To reduce the TOC (total of cost), people hope to merges the two technology: design a united storage network, it can supply the virtues and overcome the drawbacks of the NAS and SAN. This paper analysis and discuss the topic.

Keywords NAS(Network Attached Storage, SAN(Storage Area Network), Merge, USN(United Storage Network)

1 传统的NAS和SAN

为适应数据爆炸性的增长,先后出现两种非常重要的网络存储技术:NAS和SAN。按照存储网络工业协会(SNIA)的定义:NAS是可以直接联到网络上向用户提供文件级服务的存储设备。NAS是一种存储设备,有其简化的实时操作系统,它将硬件和软件有效地集成在一起,用以提供文件服务。目前采用的协议是NFS和CIFS,其中NFS应用在UNIX环境下,最早由SUN microsystem开发,而CIFS应用在NT/Windows环境下,是由Microsoft开发。NAS的结构及采用的协议使得NAS具有以下优点:异构平台下的文件共享;充分利用现有的LAN网络结构,保护现有投资;容易安装,使用和管理都很方便,实现即插即用;广泛的连接性,可以适应复杂的网络环境;较低的总拥有成本等……。

实际应用中NAS也表现出一些缺陷:(1)在文件访问的速度方面:NAS采用的是File I/O方式,在提出请求的客户端,File I/O请求先经过整个TCP/IP协议栈封装,再经过网络传输。到达NAS后,经过TCP/IP协议栈将封装的File I/O命令解封装,到达NAS的文件系统,再对存储设备进行读写。数据取出来之后要经过类似的与之相反的过程,这带来巨大的网络协议开销,因此NAS的文件访问速度相对SAN而言很低,不适合对访问速度要求高的应用场合,如数据库应用,在线事务处理等。(2)在数据备份方面:需要占用LAN的带宽,浪费宝贵的网络资源,严重时甚至影响客户应用的顺利进行。(3)在资源的整合和NAS的管理方面:NAS只能对单个存储(单个NAS内部)设备之中的磁盘进行资源的整合,目前还无法跨越不同的NAS设备,难以将多个NAS设备整合成一个统一的存储池,因而难以对多个NAS设备进行统一的集中管理,只能进行单独管理……。

按照SNIA定义,SAN是一种利用Fibre Channel等互联协议连接起来的可以在服务器和存储系统之间直接传送数据的网络。SAN是一种体系结构,它是采用独特的技术(如FC)构建的与原有LAN不同的一个专用于存储的网络,存储设备和SAN中的应用服务器之间采用Block I/O的方式进行数据交换。独特的体系结构和构建技术使得SAN具有如下优点:高性能、高速存取,目前光纤通道可提供2Gbps的带宽,新的10Gbps的标准也正在制定之中;高可用性:网络用户可以通过不止一台服务器访问存储设备;集中存储和管理:可以整合各种不同的存储设备形成一个统一的存储池,向用户提供服务;可扩展:服务器和存储设备相分离,两者的扩展可以独立进行;支持大量的设备,理论上具有1500万个地址;实现LAN-free backup,数据备份不占用LAN带宽;等……。尽管如此,SAN仍有自身的缺陷:如异构环境下的文件共享方面,SAN中存储资源的共享一般指的是不同平台下的存储空间的共享,而非数据文件的共享;目前SAN主要是采用FC技术,其构建、维护、管理都非常困难,而且基于FC的互联设备和存储设备都非常昂贵,这些都大大增加了企业的总拥有成本;由于不同厂商的FC的具体实现不同,相互之间设备的互操作性很难解决;其连接距离也限制在10km左右……,这些都极大地阻碍了SAN的普及和推广。

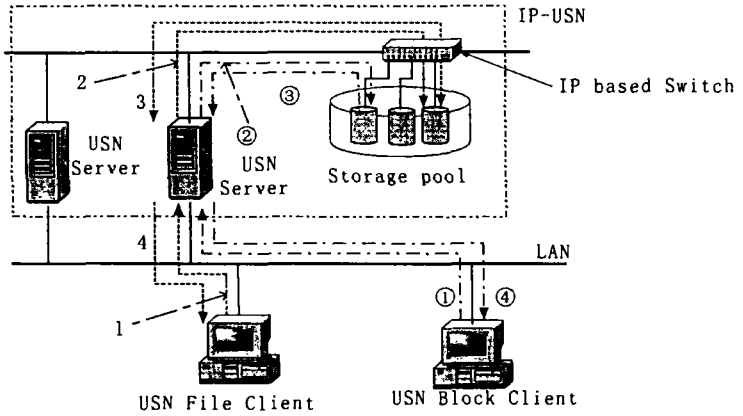
传统的NAS和SAN是两种不同的技术,但两者并非相互竞争,它们应用在不同的场合,是相互补充的,两种结构可能因满足不同的需求而同时存在。当然这也带来了一些弊端:两者构建技术不同,企业要求同时构建和管理两种不同的存储结构,增加了实现和管理成本,同时也不能很好地做到存储资源的整合等。NAS/SAN的融合技术正是在这种情况下应运而生。为此,本文试图从两个方面:以NAS技术为主和以SAN技术为主来进行NAS/SAN的统一,以构建一个统一的存储网络,即统一存储网USN(United Storage Network),它融合了NAS和SAN两种技术的优点而克服了各自的缺点。

2 基于NAS的统一存储网络USN

传统的NAS是可以直接附网的存储设备,其上含有操作

*)本文受国家自然科学基金项目(编号:60173043)和国家"973"重大基础项目(编号:G1999033006)资助。谢长生 教授,博士生导师,研究方向为基于网络的存储系统;计算机高速接口与通道;采用新原理的超高密度、超高速存储技术。傅湘林,曹 强,刘朝斌 博士生,研究方向为基于IP的存储区域网。

系统(文件系统)和存储子系统,两者的关系是紧耦合的关系,存储子系统从属于单个的 NAS 设备,其可扩展性是有限的。而基于 NAS 技术的统一存储网络 USN 是将操作系统(文件系统)和存储子系统分开,如图1所示,其文件系统功能部分单独构成一个 USN server,用以保存元数据,同时完成文件系统的功能;其存储子系统采用外置式的 SAN 结构,通过网络和 USN server 相连,从而构成一个专用于存储的网络,并通过存储虚拟化可以将 USN 中的各种存储设备虚拟、整合为一个统一的存储池。采用这种结构,系统可以同时提供 File I/O 和 Block I/O 两种类型的服务,全部的存储空间逻辑上也可以相应划分为提供文件服务的 File space 和提供数据块服务的 Block space 两部分。图1中,USN File Client、USN Block Client 分别为发出 File I/O 请求和 Block I/O 请求的客户端,USN server 为该统一存储网络的服务器,图2为整个系统大致的软件实现示意图,其中 USN server 的开发是构建整个 USN 系统的关键。发出 File I/O 请求的客户端 USN File Client 不用加装任何软件,直接通过 NFS/CIFS 向 USN 发出 File I/O 请求,而客户端 USN Block Client 必须能够支持 iSCSI 功能,USN 通过 iSCSI 技术向 USN Block Client 提供 Block I/O 服务。



注:1、2、3、4是 Client(USN File Client)向系统提出文件读写请求的数据流程。
①、②、③、④是 Client(USN Block Client)向系统提出数据块读写请求的数据流程。

图1 基于 NAS 的统一存储网

当存储网络提供 File I/O 服务时,其数据读写过程是:1 USN File Client 通过 NFS/CIFS 向 USN server 发出 File I/O 请求,其工作方式和传统的 NAS 相同;2 File I/O 请求经过 USN server 上的文件系统转化为 Block I/O 请求,从而直接操作 SAN 结构中的存储设备;3 存储设备将相应的数据块返回给 USN server;4 数据块经过 USN server 上的文件系统组合成文件,再通过 NFS/CIFS 提供给 USN File Client。在这种工作方式中,USN 同时具备 NAS 和 SAN 的优点:由于采用的是以 NAS 的方式和网络相连,其工作方式和传统的 NAS 设备相同,因而具有了传统 NAS 的大多数优点:可以实现异构的文件访问和共享,通过 IP 网络访问,可以充分利用现有的网络环境,具有广泛的连通性并适应复杂的网络环境;其存储子系统部分采用 SAN 结构通过网络相连,因而又具有 SAN 结构高可扩展性、高可用性,以及存储资源的集中和统一管理等优点。存储设备可以象传统 SAN 一样进行扩充,存储资源可以集中管理和整合,服务器和存储设备之间实现高速数据访问等。

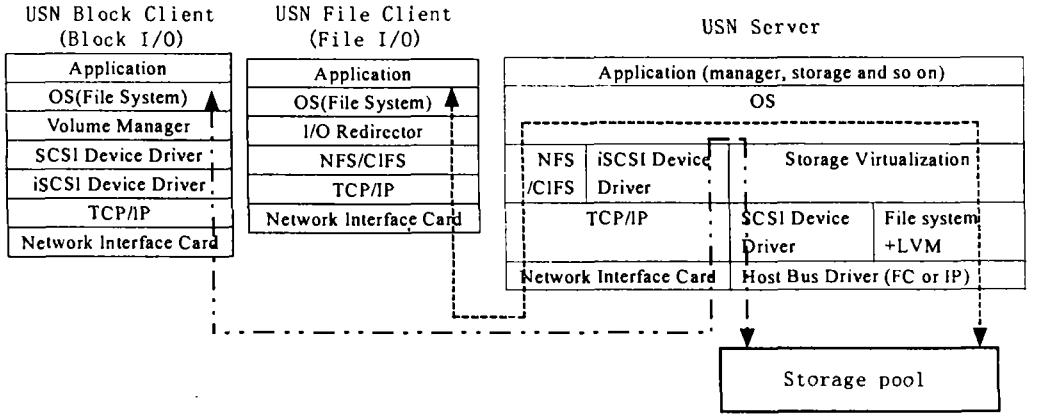


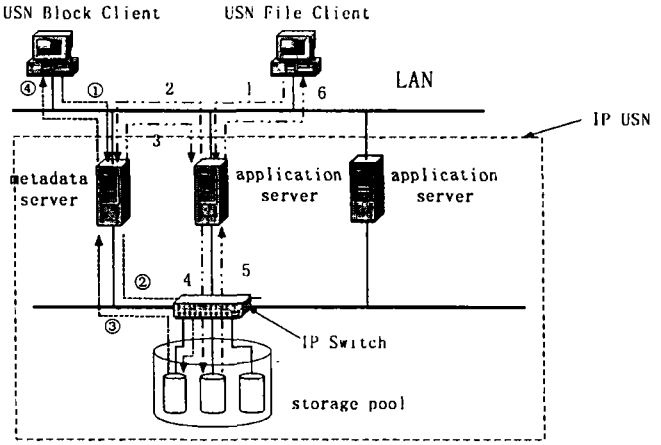
图2 基于 NAS 的 USN 的软件实现

存储网络提供数据块服务时,采用的是 iSCSI 技术,此时 Client 上软件的层次结构如图2所示(提出块 I/O 请求的 USN Block Client),在 SCSI 设备驱动层和 TCP/IP 协议栈之间存在一个 iSCSI 设备驱动层,负责将 SCSI 命令封装成 TCP/IP 数据包以便在 TCP/IP 网络上传输,封装后的命令到达目的端(USN server)后,再解封装得到封装前的 SCSI 命令,从而对存储设备进行读写。如图1所示,具体的数据读写流程为:①USN Block Client 上的应用程序发出的 Block I/O 命令(SCSI 命令)经 iSCSI 设备驱动层和 TCP/IP 封装后成为 iSCSI 命令,在 IP 网络上向 USN server 传输;②iSCSI 命令达

到 USN server 之后，经解封装，恢复成封装前的 SCSI 命令，USN server 利用这些 SCSI 命令对存储设备发出 Block I/O 请求；③ 所需的数据块由存储设备返回 USN server；④ 数据块经 USN server 中的 iSCSI 设备驱动层和 TCP/IP 协议栈封装后在 IP 网络上传输返回 USN Block Client，经 USN Block Client 解封装并由 USN Block Client 上的文件系统组合成文件。采用 iSCSI 提供 Block I/O 服务时，USN Block Client 和 USN server 各完成文件系统的部分功能，其中 USN Block Client 完成 File I/O 请求和 Block I/O 请求的转化，文件与数据块的映射等功能，而目录操作则必须由 USN server 完成，以维护整个存储池中整个 Block space 的一个全局、统一的目录视图，构成一个单一的存储系统映象。正因为 USN Block Client 和 USN server 各完成文件系统的部分功能，所以 Block space 上所有文件的元数据应该同时保存在 USN Block Client 和相应的 USN serve 上。

这种结构无论控制信息还是数据信息都必须通过该 USN server，因此这是一个潜在的性能瓶颈，而且系统在此处也是一个单点故障，为此需要采取冗余措施（如图1所示）。

3 基于 SAN 技术的统一存储网络 USN



注：1、2、3、4、5、6是 Client（USN File Client）向系统提出文件读写请求的数据流程。
①、②、③、④是 Client（USN Block Client）向系统提出数据块读写请求的数据流程。

图3 基于 SAN 技术的统一存储网

如图3所示为基于 SAN 技术的统一存储网络和传统的 SAN 不同，USN 中有两种服务器：对外提供 File I/O 服务的应用服务器和专用的元数据服务器（同时向外提供 Block I/O 服务）。类似 SAN 结构，USN 将文件系统和存储子系统分开，把文件系统功能部分从各种不同的应用服务器中剥离出来，集中到一个专用的服务器上，该服务器因为保存所有存储子系统中数据的元数据，并向应用服务器提供元数据服务，因此也称为元数据服务器。除此之外，元数据服务器还直接向外提供 Block I/O 服务，因此这种结构也可同时提供 Block I/O 和 File I/O 两种服务。图4，5分别是应用服务器和元数据服务器上的软件实现示意图。

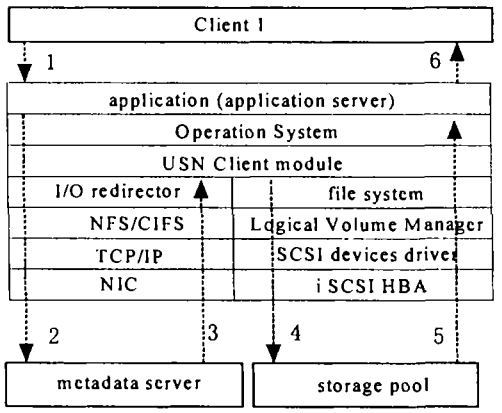


图4 应用服务器的软件实现

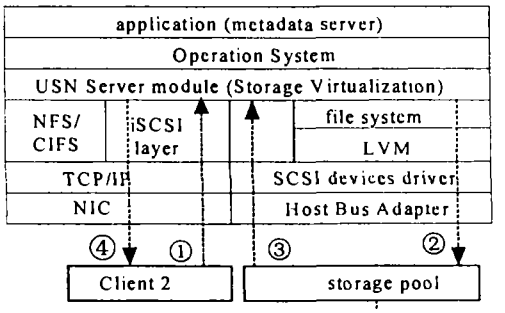


图5 元数据服务器的软件实现

提供 File I/O 服务的应用服务器上都安装有一个专用的客户端软件模块 USN Client Module。由该模块截获应用程序提出的文件请求，并将其通过 LAN 发送到元数据服务器，相应的元数据返回到客户端后，也由该软件处理，使得应用服务器在获取相应的元数据后可以采用 Block I/O 的方式来访问存储设备，该应用服务器上的软件结构如图4所示。其数据读写的流程为：1、USN File Client 首先向相应的应用服务器发出文件读写请求；2、文件读写请求到达应用服务器，此时应用服务器上还没有关于存储设备的信息，也不了解所需读写数据在存储设备上的存放信息（即文件的元数据）。应用服务器上的 USN Client Module 截获这个文件请求，后向元数据服务器发出元数据读请求；3、元数据服务器接到这个请求时，一般要完成如下几方面的工作：① 进行必要的安全论证：给予用户一个可以进行读写的权限；② 提供相应的锁机制：避

免其他客户同时访问或更新这个文件以致破坏文件的一致性;③ 返回文件的元数据;4、应用服务器取得元数据之后,以Block I/O 的方式直接向存储池中相应的存储设备发出Block I/O 请求;5、存储设备完成相应的读写操作,读写数据块并返回给应用服务器;6、应用服务器上的文件系统将得到的数据块组合成文件,并通过 NFS/CIFS 提供给 USN File Client 使用。

基于 SAN 的 USN 融合了 NAS 和 SAN 技术的优点但又克服了各自的缺点:在客户机看来,由于访问方式采用的是NAS 的文件 I/O 的方式,整个存储系统相当于一个 NAS 文件服务器,因此具有 NAS 的一些优点:如可以提供异构环境下的文件访问和文件共享,适应复杂的网络环境等。另外,文件系统和存储子系统之间采用 SAN 的结构,因而在可扩展性、可用性、资源整合、集中管理等众多方面具有 SAN 的优点。和基于 NAS 的 USN 不同,在这种体系结构中,控制信息和数据信息具有不同的通道,因此避免了可能的性能瓶颈。虽然整个系统只有一个元数据服务器,但由于只是处理元数据,因此不会对整个系统的性能构成太大的影响。当然,也可以采用冗余措施来降低元数据服务器的负载或提高系统的可用性。

除提供 File I/O 服务,USN 还可以利用 iSCSI 技术提供Block I/O。其原理和基于 NAS 的统一存储网络相同,具体的数据读写过程如图3、5所示。

总结 融合 NAS 和 SAN 技术的统一存储网以其优异的性能吸引了越来越多的注意力。目前的研究热点就是将NAS/SAN 各自的优点整合到一个系统中,使得单一系统同时具有 NAS/SAN 两者的优点而克服各自的缺点。首先应该具备 SAN 的优点:在服务器和存储设备之间采用高速的Block I/O 方式或具有 Block I/O 的性能;存储设备与存储设备之间采用 LAN-free 的方式备份数据;可以对存储资源进行集中管理,利用存储虚拟化技术,对存储空间进行整合,形成一个统一的存储池……;其次要具备 NAS 的优点:能够通过局域网(IP 网络)、以标准的 NFS 和 CIFS 协议在异构的环境下实现文件级的访问与共享;适应复杂的网络环境;构建、维护、管理、使用简单方便;降低总拥有成本。所以理想中的存储网络就应该统一在现有的 IP 网络的基础上,使得任何平台下的客户在得到授权的情况下都可以在任何时候,从任何地点访问放在异处的信息。

参 考 文 献

1 IBM redbook. IP Storage Networking: IBM NAS & iSCSI Solutions

2 Satran J. et al. iSCSI. http: // search. ietf. org/internet-drafts/draft-ietf-ips-iscsi-09. txt

3 (美)Marc Farley 著,孙功星等译. SAN 存储区域网络. 机械工业出版社

4 谢希仁编著. 计算机网络. 电子工业出版社

5 Brocade whitepaper: comparing the Storage Area Network and Network Attached Storage

6 Winter corporation whitepaper: Scalable Networked Storage: the convergence of Network Attached Storage and Storage Area Network with the Highroad

7 HTACHI whitepaper: Planning for Network Network Attached Storage and Storage Area Network Convergence

(上接第71页)

致 TCP 流能随即增加发送速率。因此 TCP 流的有效吞吐率随着速率的增加也相应增加,而多播流则随着分层的取消其有效吞吐率也相应降低。在 t=60时,多播流和 TCP 流已趋于稳定状态。对比图3和图4,可以看出采用分层迁移的多播流较有较大 Leave 时延多播流更能给 TCP 流提供好的公平性,还有较快的聚合速度。

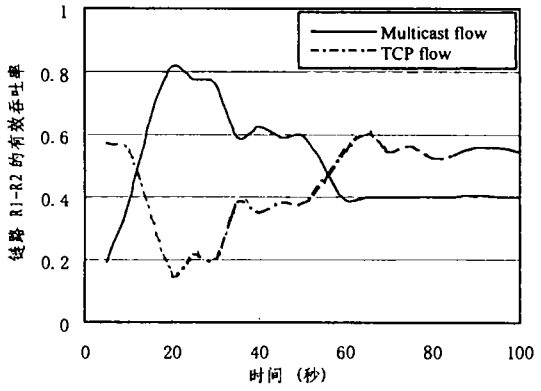


图4 采用分层迁移多播流与 TCP 流的吞吐率

结束语 针对 IGMP Leave 时延过大导致分层多播拥塞控制方案存在拥塞响应慢的问题,本文提出了分层迁移的概念和方法。分层迁移方法用快速的 Join 过程替代慢速的 Leave 过程,不仅解决了拥塞响应慢的问题,而且无需对现有的 IGMP 协议、路由器以及多播路由协议作任何修改。仿真结果不仅显示了分层迁移方法的可行性和有效性,还表明它能有效改善分层多播流对 TCP 流的公平性。

参 考 文 献

1 McCanne S, Jacobson V, Vetterli M. Receiver-driven layered multicast. In Proc. ACM SIGCOMM, Stanford, CA, August 1996. 117～130

2 Deering S. Host extensions for IP multicasting. RFC 1112, August 1989

3 Fenner W. Internet group management protocol, version 2. IETF RFC 2236, January 1997

4 Vicisano L, Rizzo L, Crowcroft J. TCP-like congestion control for layered multicast congestion control scheme. In SIGCOMM 2000, Stockholm, August 2000

5 Rizzo L. Fast group management in IGMP. In Hipparch Workshop, London, June 1998. 32～41

6 Byers j, et al. FLID-DL: Congestion for layered multicast. In: Proc. 2nd Int'l Wkshp. Networked Group Commun., Palo Alto, CA, Nov. 2000

7 Nonnenmacher J, Biersack E W. Optimal multicast feedback. In: Proc. IEEE Infocom, San Francisco, USA, March 1998

8 Li X, Paul S, Ammar M. Multi-Session Rate Control for layered Video Multicast. In: Proc. of Symp. on Multimedia Computing and Networking (MMCN'99), (San Jose, CA), Jan. 1999