

高可用的网络备份系统的设计与研究^{*}

谢长生¹ 韩德志² 李小玉¹

(华中科技大学计算机学院 武汉430074)¹ (暨南大学计算机系 广州510632)²

摘要 随着网络数字信息爆炸性增长和关键应用需求,大容量、高可用性的存储系统成为存储领域的研究热点。针对这种情况,我们将多个NAS融合成NAS簇(NASC: NAS cluster),以满足海量存储需求;为保证高可用性的需求,我们开发了一个NASC网络备份系统(简称NASCC),使用户应用关键性数据可同时镜像到NASC中的两个或更多个NAS节点,也可将NAS中备份数据恢复到用户指定的设备。并且,当某个NAS节点出现故障时,其冗余备份NAS自动接替其备份/恢复任务,保证用户备份/恢复数据的高可用性。本文将重点描述NASCC的设计和实现。

关键词 附网存储, 簇, 堆叠式文件系统, 网络备份, 高可用性

Study and Design on NAS Copy System of High Availability

XIE Chang-Sheng¹ HAN De-Zhi² LI Xiao-Yu¹

(School of Computer Science, Huazhong University of Science and Technology, Wuhan 430074)¹

(Computer Department of Jinan University, Guangzhou 510632)²

Abstract With the data information of volatile increase and the demand of application, greater capacity, high availability storage system has been the study hotspot of network storage. To solve the above-mentioned issues, we merge multi-NAS storage server into a big NAS cluster (NASC: NAS cluster) by storage virtualization. To guarantee high-availability and high-security for user's data, we have designed a NASC copy system, it can effectively copy user's data to NASC's NAS, or resume data from NASC's NAS. The paper emphatically discusses the design and implementation of NASC copy system.

Keywords Network attached storage, Cluster, Stackable file system, Network copy, High availability

1 引言

网络数据信息爆炸性地增长及应用数据高可用性和安全性的要求,刺激了网络备份技术的发展。目前,网络备份技术正受到很多用户的关注,不少用户已建立或已开始建立自己的网络备份系统。数据备份是存储系统最重要应用之一,也是保证存储网络高可用性的必要手段。网络备份有多种实现形式,可从不同的角度对备份进行分类,从备份模式来看,可以分为物理备份和逻辑备份。物理备份又称为“基于块(block-based)的备份”或“基于设备(device-based)的备份”。逻辑备份也可以称作“基于文件(file-based)的备份”。网络备份多与镜像技术相结合来实现。为实现海量存储,我们采用聚集技术将多个支持iSCSI协议的NAS^[1,2],融合成一种NAS簇(NAS Cluster,简称NASC)^[3];为保证NASC的安全性和高可用性,我们设计并实现了一种NASC网络备份系统(简称NASCC)。NASCC除通过备份/恢复保证用户数据的安全性

和可用性外,还通过高可用技术来保证整个NASC中用户所备份数据的高可用性。NASCC对文件I/O请求采用的是逻辑备份,而对块I/O请求采用的是物理备份。

NASCC是基于网络环境下的高可用性数据备份/恢复系统,它有以下特点:(1)采用镜像方式,同时将用户数据写入NASC中的两个NAS中(一个称为主备份NAS,一个称为从备份NAS),这可保证用户数据的安全性;(2)可以按照用户的需要,方便、安全、完整地将数据从主备份NAS中恢复到用户指定的机器中,以方便用户的访问;(3)当主备份NAS出现故障时,从备份NAS可即时接替主备份NAS,或将数据恢复到主备份NAS中。这样可以保证用户数据的高可用性;(4)根据一定的策略,定期对NASC中的文件进行配客管理,以提高整个NASC系统的存储空间利用率。本文将从NASCC的设计机制,以及关键设计模块入手,论述NASCC的设计与实现。

^{*}基金项目:本文受国家自然科学基金项目的资助,编号:60173034。谢长生 博士导师,研究方向:网络存储系统,基于IP的SAN。韩德志 博士研究生,研究方向:网络存储系统,基于IP的SAN。

- 7 Huang L, Wu Z, Pan Y. A Scalable and Effective Architecture for Grid services' Discovery. In: Proc. of WWW 2003
- 8 Chander A, Dawson S, Lincoln P, et al. NEVRATE: Scalable Resource Discovery. In: Proc. of CCGrid 2002
- 9 Lu Q, Cao P, Cohen E. Search and Replication in Unstructured Peer-to-Peer Networks. In: Proc. of ACM SIGMETRICS 2002
- 10 Li W, Xu Z, Dong F, et al. A Grid Resource Discovery Model based on the Routing-Transferring Method. In: Proc. of 3rd Intl. Workshop on Grid Computing, 2002
- 11 <http://rfc-gnutella.sourceforge.net/src/rfc-0-6-draft.html>
- 12 Rao A, Lakshminarayanan K, Surana S, et al. Load Balancing in Structured P2P Systems. In: Proc. of the 2nd Intl. Workshop

- on Peer-to-Peer Systems (IPTPS '03), Heidelberg: Springer-Verlag, 2003
- 13 Byers J, Considine J, Mitzenmacher M. Simple Load Balancing in Distributed Hashing Tables. In: Proc. of the 2nd Intl. Workshop on Peer-to-Peer Systems (IPTPS '03), Heidelberg: Springer-Verlag, 2003
- 14 Palmer C R, Steffan J G. Generating Network Topologies that Obey Power Laws. In: Proc. of the IEEE Globecom'00, San Francisco, 2000
- 5 Stoica I, Morris R, Karger D, et al. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications In: Proc. of ACM SIGCOMM 2001

2 NASCC 的整体结构

图1为 NASCC 系统的软件结构图。从图中可以看出网络备份系统在逻辑上分为五个子系统:备份管理器,目的备份代理和源恢复代理程序,存储资源管理模块,NAS 状态检测模块以及备份客户端程序。

网络备份系统的核心是备份管理器:(1)它负责在各个代理程序和客户端程序之间进行交互和通讯,因此针对每个代理程序和客户端程序,备份管理器也具有相应的子模块与之对应,进行通讯和处理;(2)它保证整个 NASC 中用户所备份数据的高可用性,这主要是通过备份管理器中的 HA 模块和 NAS 状态检测模块协同完成的。如图1所示,备份管理器又包括备份/恢复引擎,智能任务调度模块,用户处理模块,存储资源处理模块,以及数据库的各种接口,智能配额管理模块,HA 模块和状态检测模块等。备份管理器的核心部分是智能任务调度模块,HA 模块和智能配额管理模块。

备份管理器在逻辑上提供备份用户的日常管理和提供用户所备份数据的高可用性,如创建、删除、停用/启用、改名、修改密码等;备份策略的定义、分配;系统状态的监控;资源的添加、删除、分配;各种任务的查询、暂停、取消;检测 NASC 中各个 NAS 的运行状态和接受 NAS 发来的出错信息,启动从备份 NAS 或启动 NASC 的内部恢复线程(将从 NAS 的数据恢复到主 NAS 中),保证用户所备份数据的高可用性。

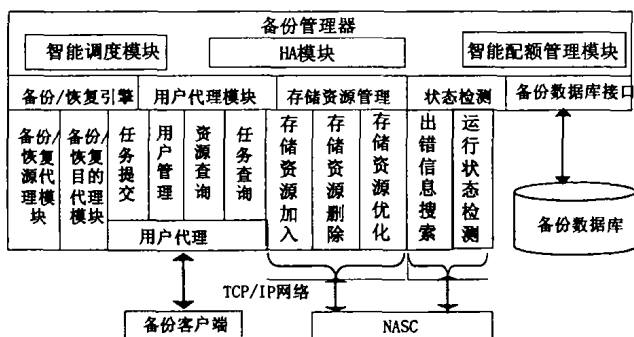


图1 NASC 的整体结构

智能调度模块接受用户的备份请求,并根据整个集群 NAS 的负载情况,在适当的时候创建备份引擎线程,启动备份过程,并将备份过程中的审计信息保存到备份数据库中。同时,智能调度模块也根据整个 NAS 集群的负载情况,均衡地在各 NAS 之间分配备份/恢复任务。

智能配额管理模块根据整个 NAS 系统中的文件的重要性和变化情况,将不同的文件分配不同空间配额。智能配额管理模块能大大提高整个 NAS 集群系统的可利用空间。

HA 模块根据状态检测模块发来的 NAS 出错信息和状态信息,启动从备份 NAS 接替主备份 NAS,或启动内部备份线程将从备份 NAS 的数据恢复到主备份 NAS 中。

备份客户端程序主要完成普通用户的备份/恢复功能,如备份任务的定义和提交;恢复任务的定义和提交;任务状态查询,审计信息的查询和删除。

备份/恢复引擎主要是具体执行一个备份/恢复任务。存储资源和用户管理模块主要是为备份工作进行前期准备。而备份/恢复任务提交主要把用户的备份任务用统一的格式进行归档,交给备份任务管理器,由备份/恢复引擎根据备份任务具体需求完成备份工作。

具体实现时,存储资源代理、恢复源代理和备份目的代理程序集成为一个程序,同时驻留在 NASC 的元数据服务器

上;而备份客户端程序,恢复目的代理,备份源代理集成为一个程序,安装在客户机上;HA 模块和状态检测模块集成为一个程序,同时驻留在元数据服务器上,在备份期间或备份完成后保证备份数据的高可用性。NASC 备份系统采用的方案是实现 C/S/S 方式,也就是 client, metadata server, storage server 方式,这样真正体现了 NAS 簇存储的特点,使得备份系统具有更好的可用性,也更利于扩展。

3 备份系统涉及的文件系统规范

为了访问某一物理文件系统(如 UFS、NFS 等),进程发出的系统调用请求先被翻译为 VFS 调用,然后 VFS 再将调用请求发给适当的物理文件系统(见图2(a)所示)。此时 VFS 被认为是上层文件系统,而物理文件系统被认为是下层文件系统,上下两层文件系统的调用是通过虚拟节点接口(Vnode Interface)实现的。虚拟节点(Vnode)的提出促进了文件系统功能模块化,并产生了堆叠式虚拟节点(Stackable Vnode)^[4]。

堆叠式文件系统(Stackable File System)(如图2(b))是一种面向对象的文件系统^[4],它所提供的对象就是堆叠式虚拟节点,堆叠式虚拟节点是一个抽象的对象,代表了存放在文件系统中的对象。堆叠式文件系统具有很强的模块性,可扩展性;对物理文件系统未作任何修改,在开发时容易调试;堆叠式文件系统作为系统模块加载到系统的核心层,而不是用户层,它避免了传统驻留在内核中的文件系统(如 UFS)与底层的设备驱动紧密结合使得开发和移植的难度很大这一缺点,同时也避免了用户层文件系统过多的上下文切换和数据拷贝降低了该类文件系统的性能的缺陷。所以,我们在开发 NASC 备份文件系统时,备份管理器是采用堆叠式文件系统规范开发的,而其它模块则在应用层开发。

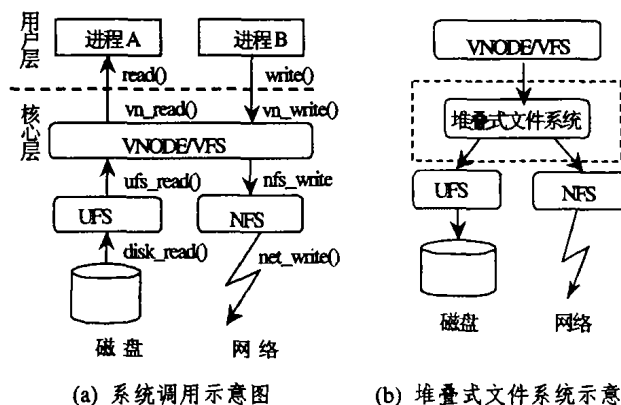


图2

4 NASCC 系统的设计与实现

NASCC 系统必须完成以下功能:第一,基于存储网络备份,能够按照特定的用户应用需求和策略,把需要备份的数据镜像到 NASC 中指定的 NAS 中。第二,实现备份数据恢复,能够按照用户的需要(例如源系统数据出现错误,损坏时),方便、安全、完整地的主备份 NAS 中恢复数据到源系统或者用户指定的目的地。第三,当主备份 NAS 出现故障或数据被损坏时,由从备份 NAS 接替主备份 NAS 或将数据恢复到主备份 NAS 中,这样保证备份数据的高可用性。

4.1 备份管理器的设计

在 NASC 元数据服务器上,我们设计一个备份管理器,它接受用户的备份请求,并根据整个集群 NAS 的负载情况,在适当的时候创建备份引擎线程,启动备份过程,并将备份过

程中的审计信息保存到备份数据库中。备份管理器的核心部分是智能调度模块、HA 模块和智能配额管理模块。

4.1.1 智能调度模块的设计 如图3所示,智能任务调度模块负责整个备份程序的正常运转,当一个客户通过客户端程序登录服务器以后,本模块将该用户信息记录,加入登录用户队列。当该用户提交一个备份/恢复任务的时候,智能任务调度模块首先检查提交的任务是否可以备份/恢复,如果是立即备份/恢复则调用备份/恢复引擎模块处理程序,否则写入数据库中的备份/恢复任务表。

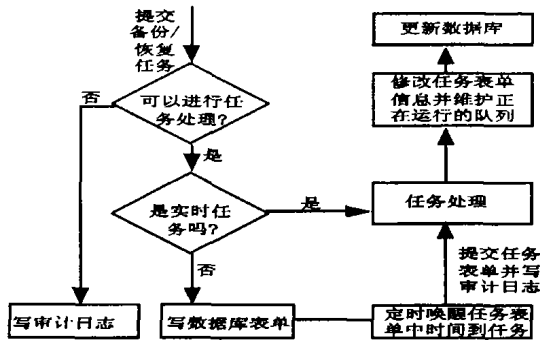


图3 智能调度模块的工作流程

通过在智能调度模块中设置一个定时器,定时唤醒任务调度模块,任务调度首先检查数据库的备份/恢复任务预定执行时间是否来到,如果到达预定执行时间,任务调度模块将备份/恢复任务从数据库取出,加入任务运行队列中,同时修改任务的状态为正在运行。然后调度备份/恢复引擎模块,开始执行备份/恢复数据工作。然后任务调度模块遍历运行队列,检查是否有运行中的任务已经执行完成,如果发现执行完成的任务,将它们从队列中取出,改变它们的状态为已经完成,将审计信息写入数据库。

4.1.2 智能配额管理模块 智能配额管理模块根据整个NAS系统中的文件的重要性和变化情况,将不同的文件分配不同的空间配额。用户频繁访问并且经常变化,对用户又很重要的文件分配高的空间配额;用户频繁访问而经常不变的文件分配中级空间配额;用户很少访问又很少变化的文件,压缩存放并分配低的空间配额;对可以再生的文件分配最低的空间配额。当智能配额管理模块与智能调度模块结合后能大大提高整个NAS集群系统的可利用空间,以及提高备份/恢复效率。

4.1.3 HA 模块 HA 模块是一个应用进程程序模块,它同NAS的状态检测设置了一个4k的共享内存,状态检测模块定时把所检测到的NASC中NAS状态信息放入共享内存中,而HA模块定时读取共享内存中的信息并对其进行分析,如果发现某个NAS出现故障,则立即启动其对应的从备份NAS接替其的I/O备份/恢复请求;如果发现某个NAS中的数据被破坏,则立即创建一个内部恢复线程,从从备份NAS中将数据恢复。这样可充分保证NASC中用户所备份数据的高可用性。

4.2 备份代理程序模块的设计

因为NASC的每个NAS都可能成为恢复源和备份目的的设备,因此设计了一个备份/恢复代理程序,它融合了备份源和备份目的两个部分的功能。另外,为了使实现更加灵活的备份功能,在NASC中,任何一个NAS能够同时执行多个备份任务,也可同时执行多个恢复任务。也就是说NASC中的NAS和客户端之间是一种多对多的关系,即一个NAS可以为多个备份任务提供备份目的,同时可以为多个恢复任务提

供恢复源。

这个备份代理程序的程序结构如图4,它有两个任务队列,一个是备份源任务队列,它记录当前NASC中并行执行的多个备份任务信息。数据镜像线程将每个备份任务的数据同时镜像到两个NAS节点中。恢复源任务队列它存放恢复任务线程接收的恢复任务,数据恢复线程将根据恢复任务队列中的信息将备份NAS中指定的信息恢复到客户端。

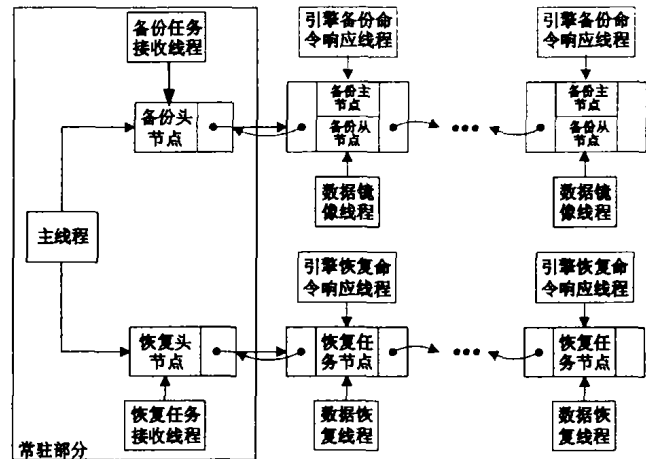


图4 备份代理的程序结构图

4.3 数据备份/恢复协议及备份数据高可用性协议

为了实现数据备份,首先需要定义一个DBP(Data Backup Protect)协议,它规定了整个备份过程的各个步骤,并定义了备份管理器、备份源和备份目的之间的通讯接口,以及它们之间的交互过程。在备份/恢复过程中,可以用iSCSI协议以块的形式传输备份/恢复数据,也可用FTP协议以文件的形式传输备份/恢复数据。

为了保证整个NASC系统的高可用性,高可用协议主要定义NASC中的NAS和状态检测模块之间的通讯接口和通讯方式。在系统正常运行,每个NAS定期向状态检测模块发心跳信息,状态检测模块也定期向每个NAS发应答信息,它们都是通过UDP方式。并且每个NAS也将数据出错信息即时地通告给状态检测模块,状态检测模块将所了解的NAS信息即时写入共享内存,以便HA模块采用相应的策略保证整个NASC系统的高可用性。

4.4 备份/恢复引擎模块

备份引擎模块和恢复引擎模块是两个对称过程模块,它们的实现是类似的,这里具体讨论备份引擎模块。备份引擎模块的功能是在一个备份/恢复任务需要执行时,智能任务调度模块启动一个新的备份/恢复线程来协调备份源代理程序和备份目的代理程序实现备份任务的完成。同时,备份引擎还负责将备份/恢复过程中的审计信息存入数据库。所以,备份引擎的主要功能是和备份源代理程序及备份目的代理程序通讯,实现备份过程中的数据和信息交互。

4.5 备份数据库

备份数据库有三个子库,一个子库用于记录还未有响应的用户备份/恢复信息;一个子库用于记录已完成的备份/恢复信息;另一个子库用于存放整个NAS集群中文件的配额管理信息。

5 NASCC系统与元数据服务器文件系统的集成

块I/O客户和文件I/O客户都可通过NASC元数据服务器,把数据备份到某一NAS服务器中,也可从NAS服务器中恢复数据。在图5中,NASC-UIOS-FS是NASC的统一IO

空间模块^[3],它向用户提供一个统一存储视图;NASC-LOADEQ-FS是NASC的负载均衡模块^[3],它时刻监视每个NAS存储服务器运行的进程数和各种资源的使用情况,并据此来调整整个NASC中各个NAS服务器的负载;NASC-COPY-FS是本文论述的NASC网络备份文件系统。NASC-UIOS-FS、NASC-LOADEQ-FS和NASC-COPY-FS都是作为堆叠式文件系统与NASC文件系统集成的。在图5中,我们只画出了NASC备份系统服务器端程序与文件系统集成的情况,而备份客户端程序模块与文件系统的集成与服务器端类似。

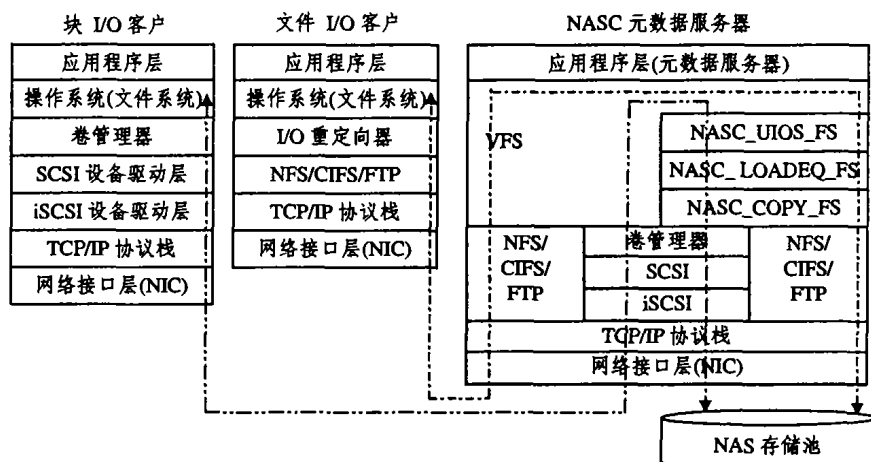


图5 NASCC与NASC元数据文件系统的集成

表1 实验机器的配置

	CPU	内存	OS	硬盘	NIC/HBA
Host 1	Pentium4 1.7G	256MB	Linux 7.1	IBM 60G	AGE-1000SX (NIC)
Host 2	Pentium4 1.6G	256MB	Linux 7.1	Maxtor 40G	AGE-1000SX (NIC)
Host3	Pentium4 1.6G	256MB	Wondows2000	Maxtor 40G	AGE-1000SX (NIC)
iNAS	Pentium4 1.5G	1GB	Linux 7.1	IBM 60G	AGE-1000SX (NIC)

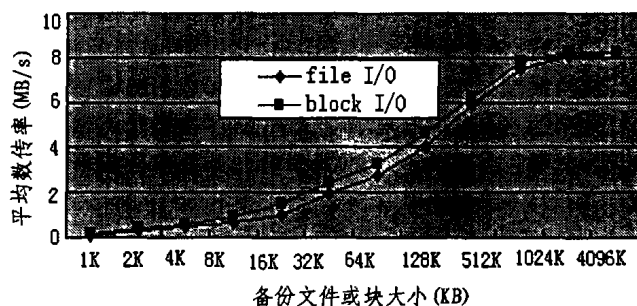


图6 NASC备份平均数据传输速率

我们主要测试用户发出不同文件或块大小请求命令时,NASC的平均数据传输速率,其测试结果如图6所示。

从图6可以看出,对小文件或小块数据备份性能很低,而大文件的备份效率比较高,这是因为不论文件大小,每个文件或块备份所用控制命令的时间是一样的,要想提高小文件或小块数据的备份效率,就必需考虑减少执行控制命令所用时间的大小。针对此种情况,一方面我们在客户端采用聚集I/O的方式,另一方面我们在NASC的元数据服务器端采用聚集I/O的方式,使备份尽量发生是大文件或大块数据。

在对NASCC高可用性实验中,用上述试验中同样的设备,我们在客户端先发多个备份I/O数据,NASCC同时镜像备份到两个iNAS中;然后,我们再从客户端向NASC的元数据服务器不断发恢复请求,在NASC响应恢复请求时,我们人为地关掉其中的一个iNAS,从元数据服务器端测试的结果

6 性能测试与分析

由于NASC中的NAS是同时支持块I/O请求和文件I/O请求^[1],因此在测试NASC备份系统时,我们分文件I/O请求备份和块I/O请求备份两种情况进行。具体的试验环境配置如表1所示。其中:Host1是元数据服务器,Host2是支持iSCSI协议的客户机,它发块I/O请求,Host3是发文件I/O请求的客户机,iNAS是我们自己开发的支持iSCSI协议的NAS服务器,本次试验用了两台iNAS。

可以看出,恢复速度稍有下降,然后又能恢复正常。从这可以看出,NASCC能对备份数据提供高可用性。

结论与展望 NASC是将多个NAS通过存储虚拟化技术和负载均衡技术,融合成一个大的NAS簇。它的主要用途是为用户提供海量、高可用性和高效的数据备份/恢复系统,针对此种情况,我们采用了堆叠式文件系统技术,设计了NASC网络备份系统NASCC,实验结果显示,NASCC对大文件或大块数据的备份/恢复有很高的效率,对小文件或小块数据的备份/恢复效率很低。虽然,我们在客户端和服务器采用了小I/O聚集技术,这对小文件或小块数据的备份/恢复效率有所提高,但效果不很显著。下一步,我们将对整个NASC系统软件进行优化,对NASC网络备份系统进行改进,并将在NASC元数据服务器端采用DCD技术^[5],以使对小文件或小块数据备份请求达到更好的性能。

参考文献

- 1 韩德志,谢长生,傅湘林,等.一种基于iSCSI技术的附网存储服务器系统的设计与实现.计算机研究与发展,2004(1)
- 2 谢长生,傅湘林,韩德志,等.一种基于iSCSI的SAN的研究与实现.计算机研究与发展,2003(5)
- 3 韩德志,秦运华.附网存储机群的研究与实现.小型微型机系统(已录用)
- 4 Heidemann J S, et al. File System Development with Stackable Layers. ACM Transactions on Computer Systems, 1994, 12(1): 58~89
- 5 Hu Y, Yang Q. DCD-disk caching disk: A new approach for boosting I/O performance. In: Proc. of the 23rd Intl. Symposium on Computer Architecture, (Philadelphia, Pennsylvania), May 1996. 169~178