智能存储及管理技术探析*

韩德志[1], 刘　昊[2]

(1. 暨南大学 计算机系, 广东 广州 510632; 2. 河南信阳农业高等专科学校, 河南 信阳 464000)

摘　要: 为了实现高性能、低成本管理和适应信息爆炸性增长的需求, 存储系统应向存储智能化、管理自动化、存储容量无限方向发展。介绍了目前智能存储技术领域的研究热点及存储网络工业协会(Storage Network Industry Association, SNIA) 的主动存储管理规范(Storage Management Initiative-Specification, SMI-S)。

关键词: 智能存储管理; 智能存储; 主动存储管理规范; 基于对象存储

中图法分类号: TP393.07　　**文献标识码**: A　　**文章编号**: 1001-3695(2005)02-0005-03

Technology of Intelligent Storage

HAN De-zhi[1], LIU Hao[2]

(1. Dept. of Computer, Jinan University, Guangzhou Guangdong 510632, China; 2. Henan Xinyang High Agriculture College, Xinyang Henan 464000, China)

Abstract: In order to implement high performance, low cost management, and the need of explosive information, great capacity storage system should develop to intelligent storage and automated management. Introduced research hotspot for intelligent storage technology, and storage management initiative specification drawn by Storage Network Industry Association.

Key words: Intelligent Storage Management; Intelligent Storage; Storage Management Initiative-Specification; Object-based Storage

网络上的数字化信息呈爆炸性增长, 图灵奖获得者 Jim Gray 在其获奖演说中指出: 由于互联网的发展, 未来十八个月新产生的数据量将是有史以来数据量之和。据估计, 对 Internet 存储容量的需求每三个月就要翻一番。Internet 上无穷无尽的信息资源需要海量存储系统来存储。据 IDG 估计, 2003 年 IT 业在数据存储方面的花费将占整个硬件成本的 60%。与此同时, 伴随着企业信息存储要求的快速增长, 对存储管理的要求也相应变化着。过去, 企业多采用简单的直连存储方式, 像数据的保存与恢复, 仅借助大容量硬盘、磁带机或者磁盘阵列即可满足单一需求, 那时, 管理多限于对某个或某类产品间的简单数据调用。现在, 随着企业存储系统用途广泛拓宽和存储容量的增加, 在企业内部出现了多种存储方式并存, 如直连 DAS、附网存储 NAS、存储区域网 SAN。每种存储方式中, 设备各异, 并且多种操作平台并存。面对复杂多变的存储系统, 作为网管人员, 不得不更多地考虑各种存储设备与系统之间的互操作性, 有时为了解决某个问题, 需要调用很多程序, 打开很多个窗口, 往往工作效率低下。同时, 用户需要对这些异构的存储平台进行统一管理, 希望能够随时掌握其运行状况, 并根据应用的不同需求合理地调配系统内的存储资源(用户存储数据的最终目的是让所存放的数据为业务带来价值增值, 如果只能实现对数据的简单存储和访问, 那么无法挖掘数据的深层价值)。这要求存储管理必须能够动态地适应, 智能地处理, 以改变以前封闭的、以设备和平台为中心、面向硬件设备的被动管理模式, 逐步向开放、集中、可管理异构环境和分布式环境的智能化管理模式转变。实现这种转变, 需要统一的存储管理规范, 使各个存储硬件, 存储管理软件设计者和生产厂家按照这种统一的规范进行设计和生产。同时, 还需要设计智能存储设备, 即存储设备本身具有可用性、可维护性和自进化性。

目前, 在智能存储方面已出现了令人鼓舞的趋势。存储网络工业委员会(Storage Networking Industry Association, SNIA) 已在推行和完善主动存储管理规范(Storage Management Initiative-Specification, SMI-S), UC Berkeley 大学正在研究智能存储系统(Intelligent Storage, IStorage)。同时, 基于对象的智能存储技术也成为研究热点。

1　IStorage 技术

IStorage 是 UC Berkeley 大学计算工程系 David A. Patterson 教授领导的课题组所进行的一个研究项目[1~3]。该研究旨在从可用性、可维护性和自进化性方面来改进海量存储系统的性能, 以达到智能存储和智能管理目的。这里, 可用性是指当系统的硬件或软件出现故障时, 系统仍能在满足服务质量 QoS 要求的情况下继续工作; 可维护性是指不管系统怎么扩充和多么复杂, 都要求最少的人工参与管理; 自进化性是指系统的性能、可维护性和可用性应随它们的改进/升级/扩充而自动提高。IStorage 是使用专用的智能硬件和软件来构造海量的存储系统。其硬件主要由智能存储设备砖(Intelligent Device Brick) 和智能底盘(Intelligent Chassis) 组成。智能存储设备砖是由磁盘、快速嵌入式 CPU、内存和网络接口 NI 构成, 智能存储底盘是可扩展的(Scaleable) 和可冗余的(Redundant), 由快速网络

收稿日期: 2004-03-09; 修返日期: 2004-10-13

基金项目: 国家自然科学基金资助项目(60173034)

+UPS 组成。

IStorage 的硬件结构的优点有: ①分布式的处理(Shared-Nothing) 。这使系统能抵挡局部的失败。②智能设备能监视其本身的运行状态 Own“health ”、自我测试、管理失败、收集指定应用的性能数据和执行应用工作代码。这提供了自维护和自调节的基础。③即插即用, 带电插入的存储砖, 容易配置、扩充。④通过传感器检测出影响系统整体性能的环境问题, 如温度、振动、湿度、电源和侵入。⑤在每个存储节点上通过安装独立的诊断处理器。其功能有: 提供远程访问控制能力, 对节点的远程控制台访问通过选择运行节点的引导代码来实现; 为异常的情况收集、存储和处理环境数据; 非易变的“飞行记录功能 Flight Recorder ”; 所有节点的诊断处理器通过一个独立的诊断网络连接起来。⑥内建容错注入能力。独立节点的电源控制能力, 按需的网络划分/隔离, 通过诊断网络由诊断处理器和交换机进行管理, 预主动硬件自省的使用(小部件的自动检测、错误恢复机制的被控测试) 。

IStorage 的软件结构的特点有: ①全分布的、不共享代码(Sharing-Nothing Code) 。它可无限扩充系统, 避免了在存储集群中使用前端机易造成单点失效的缺陷。②冗余的数据存储。它满足了高可用性、简单的自测的需求; 同时在应用对象级的复制能容易使用数据一致性策略, 并且有更多机会实现数据的存储位置优化。③灵活的存储接口。由需求驱动提交数据到应用、自动地适应生产者和消费者在性能方面的变化。④重新激活自省功能。使用统计技术来确定正常的行为, 并据此来测试偏离的情况; 一旦非正常的行为被测试出来, 基于策略的驱动自动适应到这种非正常的行为(开始依靠管理人员来指定策略, 最后系统学会自己解决问题) 。

目前, 可用性、可维护性和自进化性是对基于服务器智能存储系统提出的主要挑战, 它甚至比性能更重要。IStorage 项目正在研究把 AME(Availability, Management, Evolutionary) 带入大规模、密集型存储系统中, 从而实现存储智能化, 管理自动化。

2 主动存储管理规范(SMI-S)

存储网络往往包含了来自多个厂商的设备, 这些设备都需要自己的管理软件, 并且这些软件在功能、通用性、安全性和可靠性等方面都难以满足企业的业务发展所需。曾有厂商试图通过私下交换管理软件 API 的方式克服这一问题, 不过小范围的交换无法彻底解决存储网络的互操作性问题。总而言之, 设备之间因缺少标准的制约而使 SAN 的管理成本非常高昂, 这已成为 SAN 发展过程中的一大瓶颈。而事实上, 根据调查结果, 系统管理员对存储管理实现标准化的呼声也越来越高。

在 2002 年中, SNIA 开始主推 SMI-S[4] , 希望对存储网络的管理提供一个统一的标准, 实现主动的存储管理。Bluefin 规范是 SMI-S 标准的基础, 它是一套由多家厂商(Brocade, Cisco, EMC, Hitachi Data Systems, IBM, HP, Dell, Network Appliance, Qlogic, VERITAS 和 SUN 等公司) 联合提交的基于对象的 API, 将存储管理软件和各种存储设备在该 API 上统一, 用来主动发现、监控和管理来自不同厂商的存储设备。SMI-S 的技术基础是由 DMTF(Distributed Management Task Force, 分布式管理任务组) 制定的 CIM(Common Information Model, 通用信息模块) 和 WBEM(Web-Based Enterprise Management, 基于 Web 的企业管理标准) 。SMI-S 的目标是, 在存储网络中的存储设备与存储管理应用之间, 提供一个标准化的通信接口方式, 从而使存储管理实现厂商无关性(Vendor-neutral) , 提高管理效率、降低管理成本。同时, 存储产品的开发和测试成本将降低, 产品投入市场的时间将变短, 客户的满意程度将增加, 其结果生产商的利润也将增加。SMI-S 使用存储管理规范把存储管理和管理对象统一起来, 如图 1 所示。

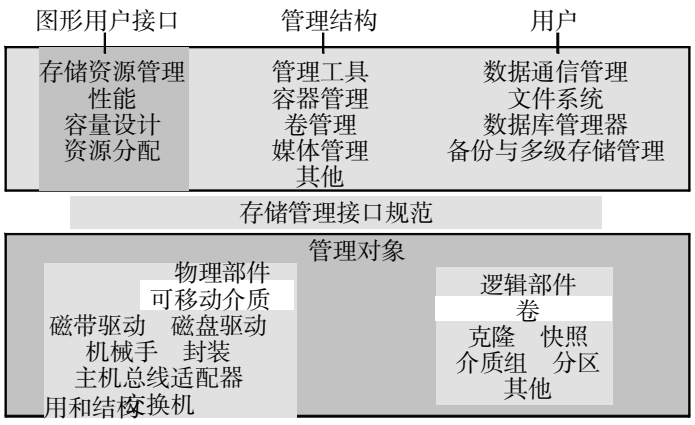

图 1 SMI 模型

图 2 显示了使用 CIM/WBEM 技术的 SMI-S 管理应用环境。从图 2 中可看出, SMIA 的 SMI-S 环境是由一个单一对象模型组成, 这一单一对象模型对应到存储网络中每一类型部件。这些模型定义普通的属性和标准特性的行为。管理行动使用单一管理协议和传输机制进行通信, 这种传输机制能在几个互连设备中的任何一个中执行。在这种模型中还包括一组统一的发现、安全和持久命名服务。为了使多个管理应用协同存在一个网络中, SMI-S 使用一个锁管理器来阻止不同应用的相互干涉。包含 SNIA SMI-S 环境模型和协议是独立于平台, 能够对应到任何平台下开发的应用, 能使运行在任何平台下应用互操作。CIM/WBEM 技术使用一个有效的人和机器都可读的被叫做管理对象格式语言(Managed Object Format, MOF) 到精确指定的对象模型。编辑器能被开发读 MOF 格式文件, 并且自动产生数据类型定义, 并且 GUI 构建是被插入管理应用中。

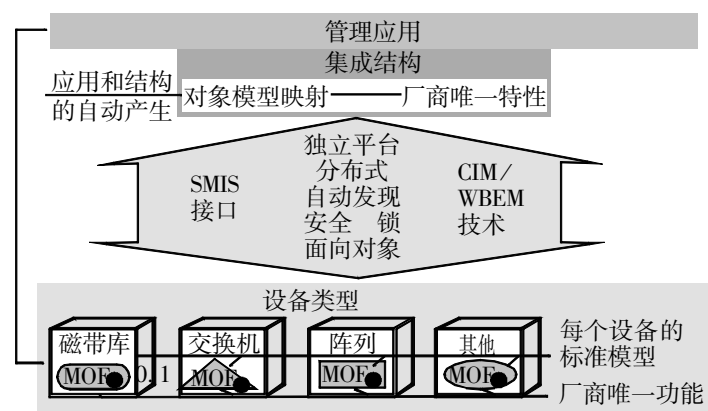

图 2 使用 CIM/WBEM 技术的 SMI-S 管理应用环境

3 基于对象的智能存储技术

对象是一个具有像文件一样接口的存储容器, 它能有效地表征 NAS 和 SAN 结构的融合。它具有 NAS 的优点(高级粗像能实现跨平台的数据共享, 也能实现基于策略的安全) , 同时也具有 SAN 的优点(直接的数据访问和基于交换网络的设备可扩展性) [5] 。尽管按空间分配和数据访问, 对象表现像文件一样, 但它仅作为存储应用容器提供服务(如文件系统和数据

库) , 它实现任何被希望增加接口功能(如锁功能) 和查询机制(如目录服务) 。一个对象是可变长的并能被用来存储任何类型的数据, 如文件、数据库记录、医学图像、多媒体信息; 一个简单的对象甚至能被用来存放整个文件系统或数据库。存储应用决定对象中存储的内容, 是通过一个类似文件系统的具有丰富内容的接口来实现的。并且, 因为对象可以动态地变大或缩小, 存储设备是负责对象全部内部空间的管理。对象是由数据、用户可访问的属性和被设备管理的元数据组成。存储在对象中的数据对于对象存储设备来讲是不透明的, 并且只占对象中的一部分。用户可访问的属性描述是对象的特性, 它们中的一些是不透明的, 而另一些是透明的。被设备所管理的元数据信息是任何额外信息, 它由管理对象物理存储目的存储设备维护。

基于对象的存储设备(Object-based Storage Device, OSD) 可以多种形式存在。范围从一个单一的磁盘驱动器到具有阵列驱动器的存储控制器。OSD 可以是随机访问设备, 也可是可写设备; 磁带设备和光介质也可被用来存储对象。OSD 与基于块的设备的主要区别不是介质, 而是接口。基于块的文件系统能被粗略地分为两部分: 用户部件和存储部件。用户部件是用逻辑数据结构表征用户应用, 如文件、目录及为访问这些数据结构的一个接口; 存储部件映射数据结构到物理存储。如图 3 所示, 在基于对象的存储中, 用户部件不变, 而存储管理部件是被下迁到存储设备, 设备接口由块接口变成对象接口。下迁元数据到存储设备, 使得不同存储应用之间的数据共享变得容易。基于对象存储的明显的优点是实现跨平台的数据共享和应用级的数据安全性; 另一个优点是, 优化了存储设备管理数据的物理布局, 从而为设备的自管理提供了机会。

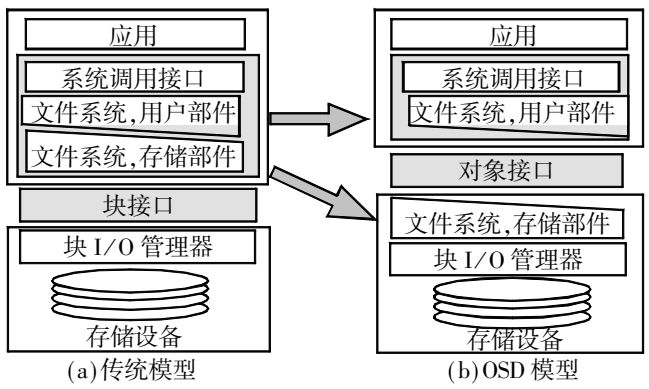

图 3　从文件系统下迁存储管理到存储设备

在对象存储中, 存储设备能理解设备中块之间的一些关系, 并能用这些信息有效地组织数据和预见需求。对象属性包含对象的动态信息和静态信息, 它们暗示了对象的行为, 如期望的读/写速率, 最可能的访问模式(串行或随机) , 对象的期望生存时间等。访问这样的属性可使存储系统更好地组织和服务数据。基于属性的学习所得到的环境知识, 使对象存储向智能存储方向发展。ABACUS[6] 是在分布式文件系统的基础设计的基于对象的智能存储系统。它将存储对象部件分成一个个的功能单元, 这些与特定数据流相关的功能单元能在客户端和存储服务器之间来回移动, 寻求自动的性能和配置管理(通过移走文件系统和应用编程者平衡负载及系统配置负担) , 从而简化文件系统的开发。ABACUS 提供两个基本的需求: 直觉的编程模型(Intuitive Programming Model) 及没有用户参入的灵活和智能的功能分割(Flexible and Intelligent Partitioning Function) 。直觉的编程模型使用户在编程时, 不必考虑每个硬件和环境组合的优化问题; 灵活和智能的功能分割使系统能获得优化的性能, 并且应用能获得最少的执行时间。

ABACUS 由一个可编程模型和一个实时系统组成。可编程模型鼓励编程者把一个应用(读或写等操作) 分成一个个小部件, 每个部件在一个数据流中执行一个小功能步; 实时系统连续观察每个功能部件的执行情况和整个系统(客户端和存储设备端) 资源的可用性。根据这些把功能部件分配到最好的网络节点(客户端和存储设备端) 。在 ABACUS 应用中, 部件对象可以是固定的也可以是可移动的, 可移动对象部件能在应用开始时自适应的约束在客户端或存储设备端, 也可实时地在客户端和存储设备端移动。可移动对象提供一个查检点/恢复机制。固定对象是在一个应用的设计时间, 由设计者决定其所在地点(客户端或存储服务器端) , 地点一旦决定, 在整个运行中不能改变。当一个系统运行时, 应用由一个通信移动对象图表示, 每个对象部件包含状态信息和一个接口, 该接口提供一个调用外部对象的方法。对象相互之间通过这种方法调用, 导致数据在它们之间流动。基于对象的部件应用如图 4 所示。其固定对象包括客户端的控制台对象(Console Object) 和存储服务器端的存储对象(Storage Object) 。控制台对象包含一个部分应用, 这部分应用必须维持在应用所开始的节点上, 它作为同用户或与启动节点上系统剩余部分的接口。控制台对象对应到操作系统中的负责管理打开文件表和实现 VFS-Layer 功能。文件系统对象, 如 Cache 对象、RAID 功能对象、加密功能对象、压缩功能对象和目录管理功能对象等, 都是可移动的对象, 可在客户端和存储服务器端实时地移动。被封装在 C + + “NASDobject”类而住留在存储服务器上的存储对象, 是一个不可移动的对象。

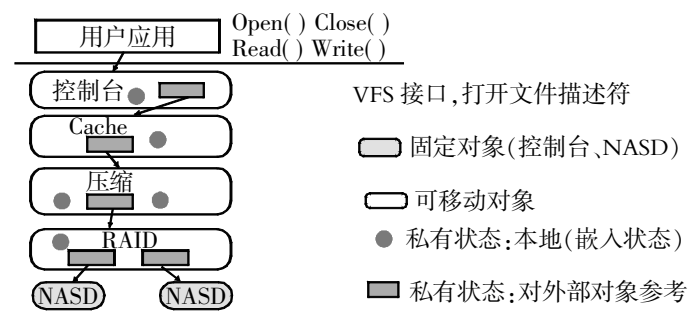

图 4　基于对象存储部件的应用

为了有效地实现基于对象的密集型存储应用, ABACUS 提供两种抽象: 移动对象和移动对象管理器。为了改进存储效率、更好地共享资源, 实现存储功能或增强存储策略, 移动对象管理器在一个给定的节点上分组实现同一类型的多个移动对象。ABACUS 的实时系统由一个迁移和单元透明调用子系统(简称捆绑管理器, Binding Manager) 及一个资源监视和管理子系统组成(简称资源管理器) 。捆绑管理器负责移动对象透明的单元地址建立, 并负责为面向对象迁移的方法调用重定向, 它扮演一个对象迁移的角色, 在每一个过程调用和从一个移动对象返回时都要通知资源管理器。对象管理器使用通知来收集关于对象之间移动的字节及关于移动对象使用资源的统计信息。因此, 它监视整个存储集群的资源的可用性(节点负载、在网络连接的可用带宽) 。ABACUS 的资源管理器搜集每个对象资源的使用情况和每个节点资源可用统计。资源使用统计是被组织作为在对象之间的定时数据流图。在一个给定服务器上的资源管理器寻找执行迁移, 通过迁移(下转第 13 页)

[15] C V Hollot, *et al*. Analysis and Design of Controllers for Aqm Routers Supporting TCP Flows[DB/OL]. http://www1.cs.columbia.edu/~misra/pubs/TAC_special.pdf. 2003-09-11/2004-03-10.

[16] C V Hollot, V Misra, D Towsley, *et al*. On Designing Improved Controllers For Aqm Routers Supporting Tcp Flows[C]. Proceedings of INFOCOM 2001, Anchorage, Alaska, 2001. 1726-1734.

[17] M Vojnovic, *et al*. Global Fairness of Additive-increase and Multiplicative-decrease With Heterogenous Round-trip Times[C]. Proceedings of IEEE INFOCOM 2000, 2000.

[18] Paganini F, Doyle J, Low S. Scalable Laws for Stable Network Congestion Control[DB/OL]. http://www.ee.ucla.edu/~paganini, 2001-09/2004-03-10.

[19] Hespanha J P, Bohacek S, O Braczka K, *et al*. Hybrid Modeling of TCP Congestion Control[A]. Hybrid Sys2 Tems: Computation and Control[C]. Berlin: Springer Verlag, 2001. 291-304.

[20] 任丰原, 林闯, 刘卫东. IP 网络中的拥塞控制[J]. 计算机学报, 2003, 26(9): 1025-1034.

[21] Ratul Mahajanb, Sally Floydc. Controlling High Bandwidth Flows at the Congested Router[R]. AT&T Center for Internet Research at ICSI (ACIRI) TechReport TR-01-001, 2001.

[22] Ao Tang, *et al*. Understanding CHOKe [DB/OL]. Proceedings of IEEE Infocom, San Francisco, CA, 2003. http://citeseer.ist.psu.edu/570417.html, 2003-04/2004-03-10.

[23] Mascolo S, Casetti C, *et al*. TCP Westwood: End-to-End Bandwidth Estimation for Efficient Transport over Wired nd Wireless Networks [C]. Proceedings of ACM Mobicom, Rome Italy (2001). To Appear in ACM Journal on Wireless Networks (WINET), Special Issue on Wireless Networks with Selected Papers from MOBICOM 2002.

[24] Aaron Falk, Ted Faber, *et al*. Experimental Measurements of the eXplicit Control Protocol (XCP) [DB/OL]. http://www.ana.lcs.mit.edu/dina/XCP/, 2004-02/2004-03-10.

[25] C Jinand, D Wei, *et al*. Fast TCP: From Theory to Experiments[DB/OL]. http://netlab.caltech.edu/FAST/, 2003-09/2004-03-20.

[26] Cheng Jin David X, *et al*. Fast TCP: Motivation, Architecture, Algorithms, Performance. The Proceedings of IEEE Infocom, Hong Kong [EB/OL]. http://netlab.caltech.edu/fast, 2004-03.

[27] Tom Kelly. Scalable TCP: Improving Performance in Highspeed Widearea Networks [DB/OL]. http://www-lce.eng.cam.ac.uk/.ctk21/scalable/, 2002-09/2004-03-10.

作者简介:

闫友彪(1974-), 男, 湖南常德人, 硕士, 主要研究方向为计算机网络、计算机图形学; 陈元琰(1961-), 男, 福建仙游人, 教授, 博士, 主要研究方向为计算机图形学、计算机网络; 罗晓曙(1961-), 男, 湖北应城人, 教授, 中国科学技术大学近代物理系博士生, 主要研究方向为非线性动力学; 陶小梅(1980-), 女, 广西桂林人, 硕士, 主要研究方向为计算机网络。

(上接第 7 页)可使访问服务器的全部应用有最小的平均完成时间。在一个 ABACUS 集群中, 有两类节点服务器和客户。服务器是一个至少包含一个基于存储对象的节点, 而客户是访问服务器上的存储对象的节点。ABACUS 与存储有关的部分可分解为可移动对象如图 5 所示。

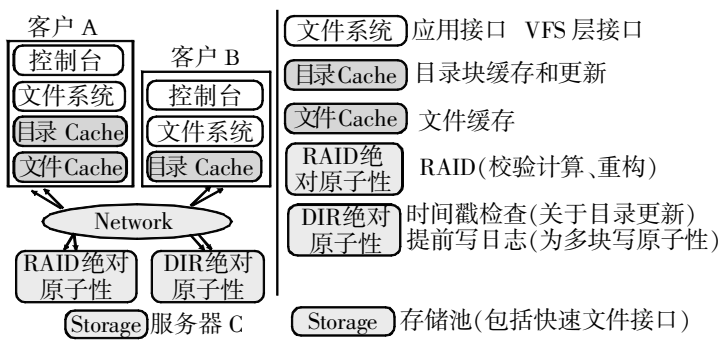

图 5　ABACUS 被分解成可迁移部件实例

ABACUS 把操作系统分解成与用户有关的部分和与存储有关的部分, 通过采用基于对象的设计方法把操作系统和文件系统与用户有关的部分固定在客户端, 底层与存储有关的部分(盘、NASD 等对象)固定在存储服务器端, 而存储有关的中间部分设计成可移动对象。通过存储服务器和客户端移动中间对象来减少网络传输和平衡整个网络负载, 从而实现智能存储特性。

4　结束语

网络信息爆炸性的增长, 要求信息存储系统朝无限的带宽、无限的容量和无限的处理能力(Infinite Bandwidth, Infinite Capacity, Infinite Processing Capability), 即"3i"方向发展。这对信息存储系统的体系结构设计和相关的文件系统设计, 以及相关的存储资源管理和 I/O 优化调度算法等方面的设计提出了极大的挑战。只有向存储智能化, 管理自动化, 即向智能存储网络系统方向发展, 才能迎接这种挑战。这要求未来的存储设备必须是智能存储设备, 即能实现自配置、自保护、自优化、自诊断和自管理。使用智能存储设备单元加上智能存储硬件接口及其环境能实现智能存储系统, 用基于对象存储技术所设计的硬件和软件环境同样可构成具有智能特性的存储系统。而使用 SMI-S 标准构建的 SAN 存储系统, 具有主动化和智能化的存储管理特性。目前, SNIA 正在将 SMI-S 标准推行到 NAS 存储及基于 IP 的存储领域中, 界时融合 SAN, NAS 和基于 IP SAN 的广域存储系统, 可实现主动化和智能化的存储管理。

参考文献:

[1] STORE Group. ISTORE Introduction[EB/OL]. http://istorage.cs.berkeley.edu/, 1999.

[2] ISTORE Group. ISTORE Software Architecture[EB/OL]. http://istorage.cs.berkeley.edu/, 1999.

[3] ISTORE Group. ISTORE: An Introspective Storage Architecture Service Application[EB/OL]. http://istorage.cs.berkeley.edu/, 1999.

[4] SNIA. Emerging Standards: Easing the Complexity of Managing Storage[EB/OL]. www.snia.org/smi, 2002.

[5] Mesnier M, Gregoory R. Object-based Storage[J]. IEEE Communications Magazine, 2003.

[6] Khalil S Amiri. Scalable and Manageable Storage System[EB/OL]. www.ece.cmu.edu/research, 2000.

作者简介:

韩德志(1966-), 男, 博士研究生, 主要研究方向为网络存储系统、基于 IP 的 SAN; 刘昊(1967-), 男, 讲师, 主要研究方向为网络数据库。