存储虚拟化技术的研究

谭 生 龙

(湖北经济学院 计算机学院 武汉 430205 华中科技大学 计算机科学与技术学院 武汉 430074)

摘要:存储虚拟化技术是基于网络的存储管理技术,它屏蔽大量异构设备的差异性,向用户提供简单的逻辑存储访问接口;它简化了存储管理,优化了系统性能,提高了存储设备的利用率。本文从存储虚拟化的概念入手,详细分析了存储虚拟化的模型结构、分层、采用的协议等方面的内容,介绍了当前虚拟化技术的最新进展,讨论了基于网络的虚拟存储技术的优点和不足。

关键词:存储虚拟化 存储区域网络(SAN) 网络附加存储(NAS) 网络存储 iSCSI

Research on Storage Virtualization Technologies

TAN Shenglong

(Dept. of computer, Wuhan University of Economics, Wuhan, 430205, China,
Dept. of Computer Science and Technology, Huzhong University of Science and Technology, Wuhan, 430074, China)

Abstract: Storage virtualization techno logies is the network - based storage management techno logies; it screens the differences of heterogeneous storage appliances, supply users with simple logic storage access interface; it simplifies storage management, optimizes the storage performance and improve the availability of storage devices. We first introduced the conception of storage virtualization and analyzed the model of storage virtualization, its hierarchy and popular protocols adapted by storage virtualization. We also discussed the latest tendency of storage virtualization and pointed out their advantages and disadvantages of those techno logies.

Keywords: storage virtualization, storage area network(SAN), network attached storage(NAS), network - based storage, iSCSI

随着全球信息的爆炸式的增长,存储技术越来越为业界所关注,据报道估计,截止在 2008年底全球所产生和复制的数据量已达到 250Exabytes(2 ×1021bits),预计到 2010年这个数字将增长 4倍,达到 1000Exabytes[1]。数据的持续快速增长带来的数据膨胀的压力使越来越多的企业把数据存储作为重要项目来管理,从而带来存储管理技术快速发展,而存储设备的差异性使高效管理这些设备面临诸多困难;存储虚拟化技术就是在如何提高存储设备的管理效率,如何整合不同类型的存储资源,如何向用户提供统一的访问接口等前提下提出的;目的在于解决异构存储系统在兼容性、扩展性、可靠性、容错容灾等方面的问题。虚拟存储管理系统屏蔽了不同平台下具有不同属性的存储设备的差异性,向用户提供可以任意分割和扩展的基于虚拟卷的存储系统,该系统具有良好的可扩展性、稳定性、可用性和高性能,用户可以在线增减存储容量,屏蔽不同类型存储设备的差异性,能动态进行负载均衡,向用户提供简单统一的虚拟访问接口,存储服务可以跨越多磁盘或多分区,存储设备基于网络而独立于地域分布,系统支持多种标准协议并向用户透明。

存储技术的发展经历了传统的以磁盘磁带为基础的本地直接存储(DAS),以扩展存储容量为目的的 JBOD(Just a Bunch of Disks)存储、以网络附加存储 NAS(Network Attached Storage)和存储区域网络 SAN(Storage Area Network)为基础的网络存储和基于互联网的以提供存储服务为目的的云存储的发展过程。

本文于 2009 - 09 - 16收到。

1 虚拟化的存储系统

存储虚拟化是通过虚拟卷映射、流数据定位、数据快照、虚拟机等[2]技术实现异构存储设备的统一管理以及存储位置无关性而提出的,目的在于屏蔽存储管理中的一系列复杂问题而向用户提供简单透明统一的存储访问模式。目的是为了解决存储需求不可预见的持续膨胀式增长、以适应网络存储系统变得越来越庞大和复杂、众多而异构的存储设备如何有效的统一管理和高效的利用,屏蔽不同存储设备的差异性而提供简单而统一的访问方式。

传统的存储系统通常是直连到 PC或服务器上,基于 PC的存储属于私有存储,它不方便集中统一管理,不方便共享、扩展、备份及访问控制,没有扩展性且利用率低;基于服务器端的存储在提供存储服务时需要占用服务器资源,在系统负载很大时会显著降低存储服务的性能,而且基于服务器端存储不是专有系统,操作系统没有经过优化,不能提供高效的存储服务。

对存储数据的访问模式可以分为基于文件的访问模式和基于块设备的访问模式;基于文件的访问模式方便文件的共享和安全控制;基于块设备的访问模式便利数据库等大规模应用的数据访问与传输,实现这两种访问模式的技术分别为基于文件存储的网络附加存储 (NAS)和基于块设备访问的存储区域网络 (SAN)。

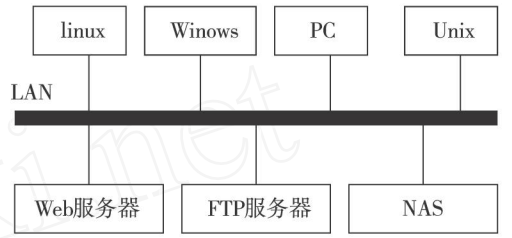

图 1 网络附加存储体系结构

(1)网络附加存储 NAS。网络附加存储是组建在局域网上的以文件共享为目的网络存储服务器,它与通用目的的服务器相比,具有高性能的文件共享特性。该服务器的操作系统是为文件共享而经过特殊定制和优化的,具有小、快而高效的内核;特别是在文件检索,文件共享,文件存储,文件访问与服务等方面具有很高的吞吐率。另外,该操作系统还精简了与提供文件服务无关的很多功能而只保留和配置有与文件服务相关的功能和软件。网络附加存储支持大量的网络文件协议 (如 FTP, NFS, SMB /CIFS等),具有通用而开放的访问接口,支持多种标准通用协议,可为不同平台下的应用所调用。

多个网络附加存储 NAS网络可以进一步相互连接形成更大网络附加存储网络[3],这也体现了NAS具有较好的扩展性。不过,它只提供文件级别的虚拟化而不能提供块级别的虚拟存储服务从而限制它进一步发展。

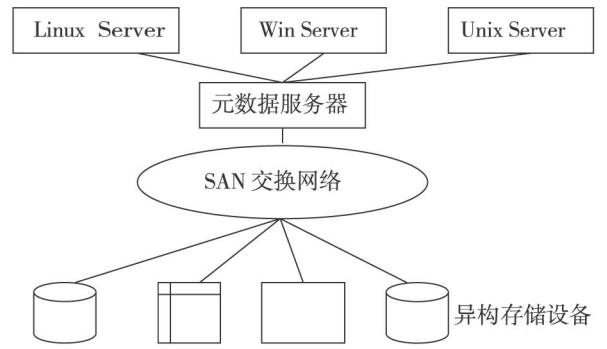

图 2 对称的 SAN存储虚拟化结构

(2)存储区域网络 SAN。存储区域网络 SAN的结构是基于网络的多服务器共享多存储设备,使存储设备独立于服务器而直接链接到可进行高速存取访问的高性能局域网上,这种独立于服务器且通过网络直连的存储设备称为存储区域网络,它具有高扩展、高可靠和高性能等优点,但它具有不兼容多操作系统,对异构的存储设备的虚拟化管理需要在 SAN基础上进行附加设计[4]等缺点。根据虚拟化设备所处的位置,又可以把存储区域网络 SAN细分为对称的 SAN存储虚拟化结构和非对称的 SAN存储虚拟化结构。

2 存储虚拟化系统的结构

存储虚拟化技术存在两种结构模式即对称结构和非对称结构。对称结构是指在储存设备和应用服务器的数据路径上实现存储设备的虚拟化,其数据和控制信息共用同一传输路径,其虚拟化功能通过运行在虚拟化控制器上的虚拟化管理软件实现,这种结构也称为带内存储虚拟化技术。虚拟化管理软件通过映射

技术实现异构存储体到虚拟存储池中的逻辑存储单元 (LUN)的映射,实现主机可以访问的逻辑卷,在主机端通过屏蔽协议端口等底层信息实现逻辑存储单元 (LUN)到逻辑盘符的映射。虚拟化管理控制器很容易成为系统的瓶颈,故采用包括数据缓存、预取、主动存取多种策略来提高的数据传输率,且它的安全性需要仔细的设计。

非对称结构的虚拟存储由存储网络中的一台独立并装有虚拟化管理软件的服务器实现存储虚拟化功能,该服务器完成存储设备的逻辑映射、存储分配、数据安全保障等元数据的管理,应用服务器首先通过访问元数据服务器获取映射后虚拟设备,然后通过数据通路直接访问存储设备,因此实现了数据和指令在不同的路径上的传递,这种结构也称为带外存储虚拟化技术。

3 存储虚拟化的实现方式

存储虚拟化可以在三个层次上实现:基于主机的虚拟化、基于存储设备 (存储子系统)的虚拟化和基于存储网络的虚拟化。

(1)基于主机的虚拟化也称为基于服务器的虚拟化,是通过在服务器操作系统中嵌入或添加虚拟层来实现设备虚拟化的,该方法不需要添加特殊的硬件而只需安装具有虚拟化功能的软件模块,它以驱动程序的形式嵌入到应用服务器的操作系统中,呈现给操作系统的是逻辑卷 (Logic Volume Management),通过逻辑卷把分布在多机上的物理存储设备映射成一个统一的逻辑虚拟存储空间,逻辑卷管理系统实际上是一个从物理存储设备映射到逻辑卷的虚拟化存储管理层,它可实现系统级和应用级的多机间存储共享。

运行在服务器上的虚拟化软件需要占用服务器的 CUP、内存、带宽等开销,对操作系统的依赖性较大,使得虚拟化系统不能兼容不同的平台,移植性较差。但是基于主机的虚拟化最容易实现的,一般只需在应用服务器端安装卷管理驱动模块就可以完成存储虚拟化的过程,具有成本低、同构平台下性能高的特点。

(2)基于存储设备的虚拟化一般在存储设备的控制器中实现,又称为存储控制器的虚拟化。由于该虚拟化的实现方法直接面对具体的物理设备,在性能上达到最优,由于该虚拟化逻辑被集成到设备内部,存储虚拟化的管理简单方便而对用户透明,但由于这种虚拟化技术没有统一标准,一般只适用于特定厂商的产品,异构产品间很难实现存储级联,所以这种存储虚拟化产品的可扩展性易受到限制。另外,由于厂商的限制,用户对存储设备的选择面也很窄,如果没有第三方的虚拟化软件提供底层屏蔽服务从而实现存储级联和扩展,则该系统的扩展性就很差,但近期也有一些研究成果采用基于目录的虚拟化方式[5]来克服这些不足。

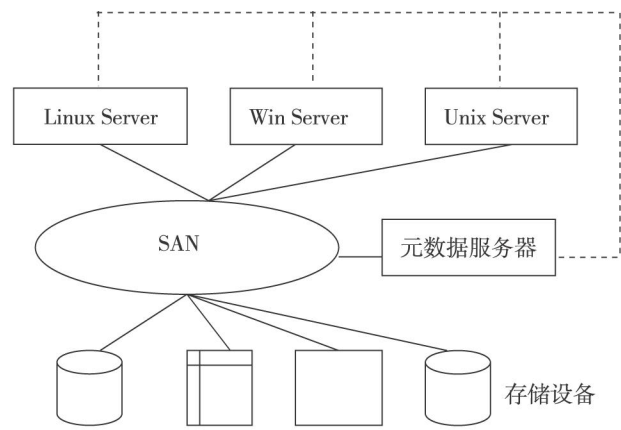

图 3 非对称的 SAN存储虚拟化结构

(3)基于网络的存储虚拟化。基于网路的虚拟存储化技术是当前存储虚拟化的主流技术,它当前在商业上具有较多的成功产品。典型的网络虚拟存储技术主要包括网络附加存储 NAS(Network Attached Storage)和存储区域网络 SAN(Storage Area Network)。由于这两种系统的体系结构、通信协议、数据管理的方式不同,所以 NAS主要应用于以文件共享为基础的虚拟存储系统中,而 SAN主要应用在以数据库应用为主的块级别的数据共享领域。存储区域网络 SAN是当前网络存储的主流技术。虚拟化存储的实现可以分布在从主机到存储设备之间路径的不同位置上,由此可把基于网络的存储虚拟化细分为基于交换机的虚拟化、基于路由器的虚拟化、基于存储服务器端的虚拟化。

(1)基于交换机的虚拟化。基于交换机的虚拟化是通过在交换机中嵌入固件化的虚拟化模块层来实现的,由于在交换机中集成有交换和虚拟化功能,交换机很容易成为系统的瓶颈,并可能产生单点故障。不过这种结构不需要在服务器上安装虚拟化软件,可以减少应用服务器的负载,也没有基于存储设备或者主机

环境的安全性问题,在异构环境下有较好的互操作性。

(2)基于路由器的虚拟化。基于路由器的虚拟化是将虚拟化模块集成到路由器中,使存储网络的路由器既具有交换机的交换功能,同时具有路由器的协议转换功能,它把存储虚拟化的范围由局域网范围内的虚拟存储扩展到了广域虚拟存储。近年来,基于路由器的虚拟化技术得到了长足的发展和广泛的应用,例如基于 iSCSI的虚拟存储技术[6,7]等,它为广域网下的云存储夯实了底层结构。

(3)基于专用元数据服务器的虚拟化。基于专用元数据的虚拟化是在存储网络中接入一台专用的元数据服务器来完成存储虚拟化工作,属于带外虚拟化方法。

元数据服务器提供基于网络虚拟存储服务,它负责映射不同的物理设备,形成整个虚拟设备存储池的全局统一数据视图,并负责与驻留在各个应用服务器上的虚拟化代理软件进行通信,各应用服务器上的虚拟代理软件负责管理存储访问视图和 I/O通信并实现数据访问重定向;该代理软件具有实现数据高速缓存和数据预存取功能,并具有维护本地存储视图和元数据的功能,可以缓存和暂存本地存取的元数据信息,并保持与专用元数据服务器的数据一致性,通过数据访问的局部性减少访问元数据服务器的次数从而可以显著的提高存储吞吐率。

(4)基于局域网的存储虚拟化。基于局域网的虚拟化技术也称为基于 IP存储虚拟化,它是当前在虚拟存储领域最活跃的研究热点之一。基于 IP存储虚拟化技术产生很多成功产品,特别是 10Gb/S以太网的出现,更是加速了局域网虚拟化的快速发展,其中支持局域网的协议包括 FCP, iFCP, SCSI, iSCSI, vSCSI[8], InfiniBand[9]等,它们都是基于 TCP/IP的数据存储访问协议 (如图 4所示)。

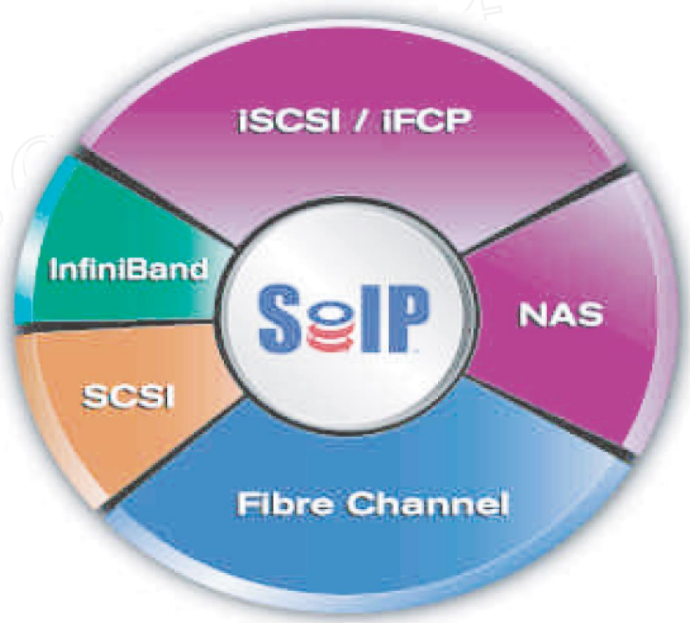

图 4 基于 IP的存储访问协议

基于网络的大规模虚拟存储技术将是今后一段时间内虚拟存储化技术的主要研究热点,其中基于 iSCSI协议的网络存储被认为是继续推动存储区域网 (SAN)快速发展的关键技术[10],该协议通过 IP协议封装 SCSI命令,把大型存储设备接入网络,使基于 iSCSI协议的存储设备可以分布在局域网、广域网和互联网上,从而实现独立于地理位置的数据存储、数据备份和数据检索;特别是 10Gb以太网的迅速普及和缩短访问延迟的远程内存直接访问技术 (RDMA)的快速发展[11,12,13],将会加速基于 IP的虚拟存储技术的进一步快速发展。

(5)基于互联网的存储虚拟化。基于互联网的虚拟化是存储技术的最高形式。它采用集群技术、网格技术、覆盖网技术、P2P技术以及分布式文件系统等技术实现将全球范围内不同类型的存储设备通过虚拟化技术整合起来,向外提供统一的虚拟内存和硬盘的功能。虽然基于互联网的虚拟化的发展还处在起步阶段,但一些研究成果已经显现,如由 Jun Wang, Xiaoyu Yao等人提出的基于成熟的 TCP/IP协议的 SAN技术,采用 iSCSI协议及分层缓存机制实现对基于广域网的存储服务器的高速访问[14]。基于互联网的存储虚拟化 (例如存储云)实际上是一种为用户提供存储服务的虚拟化技术[15]。

4 实现存储虚拟化的关键技术

实现存储虚拟化系统的关键是实现众多异构存储设备到统一虚拟存储资源的视图映射,通常在用户和存储设备路径上加入存储管理部件来实现虚拟化,它屏蔽了不同类型、不同特性的物理设备,实现大量异构

存储资源的整合,向用户提供方便访问、任意划分、在线扩容、安全稳定的虚拟存储系统。

实现虚拟化存储系统需要解决的一些关键技术包括:

(1) 异构存储介质的互联和统一管理。存储虚拟化的核心任务是兼容多种属性的存储设备,屏蔽它们间不同的物理特性并向用户提供统一的虚拟逻辑设备访问方式,由网络连接的各种物理存储设备以虚拟卷的形式向用户呈现,而用户关注的是存储容量和数据安全策略,而存储容量的物理分配则对用户透明的,存储虚拟化管理系统及其所兼容的协议屏蔽了连接到存储网络中的各类设备的差异性,简化了逻辑存储设备的管理、配置和分配,并向用户提供在线划分、扩展、配置存储和在线增加与更替存储设备的虚拟化存储管理技术。

(2) 数据的共享冲突与一致性。数据共享是存储虚拟化的主要功能之一,基于网络的虚拟存储对数据共享访问提出了很高的要求,存放在不同物理存储器中的数据拷贝为操作系统间及操作系统和数据仓库间的数据共享带来便利,但同时必须仔细设计锁机制算法、备份分发算法以及缓存一致性技术来保证数据的完整性。

(3) 数据的透明存储和容错容灾策略。数据的透明访问需要虚拟存储屏蔽存储设备的物理差异性,由系统按照资源的特性及用户的需求自动调度和利用存储资源,便于用户在逻辑卷的基础上对数据进行复制、镜像、备份以及实现虚拟设备级的数据快照等功能。虚拟存储系统必需按照数据的安全级别建立容错和容灾机制,以克服系统的误操作、单点失效、意外灾难等因素造成的数据损失。系统必需对用户透明地的实现多种机制下的数据备份、数据系统容错和灾难预警及自动恢复等策略。

(4)性能优化和负载均衡。存储系统应该从全局的观点并根据不同存储设备的特性来优化存储系统,应该根据不同存储的存储响应时间、吞吐率和存储容量来安排多级存储体系结构,实现数据的多级高速缓存和数据预取功能。根据用户的需求安排不同的存储策略实现对数据的按需存取,仔细设计 I/O均衡策略,根据具体的物理设备合理分配用户的 I/O请求,使用条带化方法、数据分块、时空负载区分、数据主动存取和数据的过预取[16,17]策略来提高数据的访问效率,为了进一步提高访问效率,也可以采用基于存储对象的存储主动服务[18]策略来提高数据的主动预测服务。

(5)数据的安全访问策略。基于网络的存储必需对访问加以控制,数据被越权访问和恶意攻击是虚拟存储系统必需要避免的,透明的存储服务所带来的数据安全性必需由虚拟化管理软件来实现,其实现安全访问的策略是多样的,如基于密钥的认证管理及数据加密策略,以及在存储体之上增加一层可信的管理层节点[19]等都是可行的方法。

(6)高可靠性和可扩展性。高可靠和可扩展性是虚拟存储系统必需具备的特性,系统应该采用高效的故障预测、故障检测、故障隔离和故障恢复技术来保证系统的高可靠性。虚拟存储系统应该在不中断正常存储服务的前提下实现对存储容量和存储服务进行任意扩展,透明的添加和更替存储设备,虚拟存储系统还应该具有自动发现、安装、检测和管理不同类型存储设备的能力。

5 结束语

本文主要介绍了存储虚拟化的内容和实现虚拟化的主要方法,重点探讨了当前存储虚拟化的研究热点和实现存储虚拟化的主流技术,详细分析了存储区域网络 SAN和网络附加存储 NAS方面的相关技术,并进一步讨论实现大规模虚拟化存储网络的一系列相关问题,最后对实现网络存储所涉及的一些关键技术进行了深入的讨论。

参 考 文 献

[1] ROGER W. Future hard disk drive systems[J]. Journal of Magnetism and Magnetic Materials, 2009, 321(6): 555 - 561

[2] CHARLES M, SID S. Online Storage Virtualization: The key to managing the data explosion: Proceedings of the 35th Hawaii International Conference on System Sciences(HICSS - 3502) [C]. [S. l.]: [s. n.] 2002.

[3] YOSHIKO Y, SHINICHI K, ATSUSHI E, et al. Concept and Evaluation of X - NAS: a Highly Scalable NAS System: Proceed-

ings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Techno logies (MSS'03) [C]. [S. l.]: [s. n.] 2003.

[4] LI Bigang, SHU Ji - wu, ZHENG Wei - min. VICS: a Storage Virtualization Management System for SAN: International Worksho Pon Storage Network Architecture and Parallel I/Os[C]. [S. l.]: [s. n.] 2005.

[5] NIKOLAI J, ARUN M, Chaitanya Patti1 et al. RAIF: Redundant Array of Independent Filesystems: 24th IEEE Conference on Mass Storage Systems and Technologies (MSST'07) [C]. [S. l.]: [s. n.] 2008.

[6] VEENA T, CHANCHAL G. Optimizing iSCSI Storage Network: A Direct Data Transfer Scheme Using Connection Migration: Internalional conference on oriental astronomy (ICON'08) [C]. [S. l.]: [s. n.] 2008.

[7] FUMITO I, HIROYUKI O, YOSHIHIRO N, et al. On Maximizing iSCSI Throughput using Multiple Connections with Automatic Parallelism Tuning: Fifth IEEE International Worksho Pon Storage Network Architecture and Parallel I/Os (SNAPI'08) [C]. [S. l.]: [s. n.] 2008.

[8] FEND Dan, YE Jun, SHI Zhan. A Protocol vSCSI for Ethernet - based Network Storage: 21st International Conference on Advanced Networking and Applications (AINA'07) [C]. [S. l.]: [s. n.] 2007.

[9] DSCOTT G. A Scalable, High Performance InfiniBand - Attached SAN Volume Controller: Second International Worksho Pon High Performance I/O Systems and Data Intensive Computing (HiperIO'08) [C]. [S. l.]: [s. n.] 2008.

[10] EMD Storage Introduction to IPStorage: Overview, Methods and Advantages of the iSCSI (Internet SCSI) Protocol (White Paper). [S. l.]: [s. n.] 2007.

[11] ETHAN B, ROBERT R. Implementation and Evaluation of iSCSI over RDMA: Fifth IEEE International Worksho Pon Storage Network Architecture and Parallel I/Os (SNAPI'08) [C]. [S. l.]: [s. n.] 2008.

[12] KO M, CHADALAPAKA M, HUFFERD J, el al. Internet Small Computer System Interface (iSCSI) Extensions for Remote Direct Memory Access (RDMA). RFC 5046 (Stan dards Track) [S], Oct. 2007.

[13] 张勇, 刘景宁. 结合 iscsi协议的 VISA存储系统的研究与实现 [J]. 微计算机应用, 2007, 28 (1):

[14] WANG Jun, YAO Xiao - yu, CHRISTOPHER M, et al. A New Hierarchical Data Cache Architecture for iSCSI Storage Server [J]. IEEE Transactions on Computers, 2009, 58 (4), 433 - 447

[15] ROBERT L, GU Yun - hong, MICHAEL S, et al. Computer and storage clouds using wide area high performance networks [J]. Future Generagtion Computer Systems, 2009, 25 (2): 179 - 183

[16] DENG Yuhui, WANG Frank. Exploring the performance impact of stripe size on network attached storage systems [J]. Journal of Systems Architecture, 2008, 54 (8): 787 - 796.

[17] RAO Jia, BU Xiang - ping, XU Cheng - Zhong, et al. VCONF: A Reinforcement Learning Approach to Virtual Machine Auto - configuration: The 6th international conference on autonomic computing and communications (ICAC '09) [C]. [S. l.]: [s. n.] 2009.

[18] ZENG Ling - fang, FENG Dan, WANG Fang, et al. Storage Active Service: Model, Method and Practice. Proceedings of the Japan - China Joint Worksho Pon Frontier of Computer Science and Techno logy (FCST'06) [C]. [S. l.]: [s. n.] 2006.

[19] ZHANG Min, ZHANG De - sheng, XIAN He - qun, et al, Towards A Secure Distribute Storage System: International conference on advanced computing techno logies (ICACT'08) [C]. [S. l.]: [s. n.] 2008.