面向虚拟存储服务系统模型的构建*)

李 博 谢长生 赵小刚 万胜刚

(华中科技大学计算机学院武汉光电国家实验室 武汉 430074)

摘 要 随着企业应用业务的不断复杂和多变,后端的存储服务需要做到按需而变。同时按需动态地整合和重组企业本身的设备和管理也是迫切需要的。为此,我们用提出的 VS^3——虚拟存储服务系统模型来应对了这一目标。从虚拟存储技术入手,将存储服务集成化,使用户直接面对上层的应用管理。该文详细介绍了整个 VS^3 模型的组织框架,设计了 VS^3 模型中的服务发现机制、数据传输机制。同时在减少原有设备成本的前提下,设计了计算和存储资源相分离的策略,从而为合成新应用服务提供了必要的资源准备。

关键词 存储虚拟化,按需存储,服务发现

Constructing Virtual Storage Service-based System

LI Bo XIE Chang-sheng ZHAO Xiao-gang WAN Sheng-gang

(Wuhan National Laboratory for Optoelectronics, College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract As the enterprise applications business changes more complex and more constantly, the back-end storage services framework must be on-demand, while the on-demand dynamic integration and reorganization of enterprises focusing on the equipment and management is urgently needed. So for this goal, we propose the VS^3, virtual storage system model. Using the technology of virtual storage and the storage services application integration, the model allows users facing the upper applications management, directly. The paper details the model of the entire VS^3 framework, and has designed the VS^3 service discovery and data transmission mechanism. Under the premise of the existing equipments cost reduction, we use the computing and storage resources separation strategy to synthesize new applications and services for future.

Keywords Storage virtualization, On-demand storage, Service discovery

1 引言

数据越来越成为当前信息处理的关键,成为了现实中网络世界最宝贵的资源。数据资源的使用相对于应用的需求变得更为复杂、多样。在一个复杂的应用系统中不仅要处理繁重的事务处理信息、在线服务的不同请求,还要满足应用过程中产生大量数据的存储和管理。这使得在企业内部构建的信息资源框架日趋庞大、复杂,以至于管理非常复杂,资源利用率低下。

同时商业环境要求企业能够随需应变,即企业要能够对外界的各种需要做出及时迅速的响应,这不仅包括满足客户需求变化的需要,更要应对供应链变动或竞争态势突变的需要。为了能够实现这种快速响应,企业对于信息的依赖性也越来越大。从总体上来看,这种对信息的依赖,其实是对存储基础设施效能的更大依赖。即企业的存储基础设施必须能够支持企业的成员对各种最新信息进行准确的、实时的访问。因此应用需要我们的存储能够做到随需应变并能提高管理效率,不仅可以提供对业务过程中的变化做出响应的灵活性,而且帮助客户降低总成本。

为此业界进行诸多尝试和探索,提出了很多相关的解决方案,如 BMC 推出基于应用为中心的 ACSM 方案[1]、Sun 的 SunStorageONE 开放框架体系、赛门铁克的 DCF 解决方案、美国 ADIC 倡导的智能存储思想[2]、Network Application、Seagate 和 Intel 三方公布的 DAFS 文件系统(融合了 NAS 和 SAN 的优势)[2]等。国内的研究[3]实现了集群环境下基于 LVM 的存储系统 CVM;文献[4]提出了 NAS 和 SAN 融合方案统一存储网(USN:Unified Storage Network);文献[5]实现了块级别带内存储虚拟化系统 AXUM,通过目标器模式实现存储虚拟化服务、蓝鲸分布式系统[6]等。国外 IBM[7,8]实现了一个资源管理系统 Neptune,它是一个可以动态配置资源的公用计算集群,类似的还有 Oceano[9,10]——主要针对电子商务的公用计算环境,通过结构体系中的控制层实现资源管理。其中文献[11]实现了一个面向分布式对象的虚拟计算环境——DOVE-G,通过网格框架整合了分布对象和网络服务。文献[12]提出了一个用户为中心的服务框架,通过协调不同网络设备和网络资源来提供服务,实现了面向用户的虚拟计算环境。它们都是很好地利用虚拟化技术来实现存储操作的动态管理,资源的动态配置。

通过文献[11,13-15]论述的虚拟计算环境理论,我们把文献[16]的资源服务思想细化到存储系统中,提出了 VS^3 方案,该方案是基于存储虚拟化技术[17]来建立存储服务可集成和按需配置的模型框架。抽象出存储系统中的不同服务,通过服务间的接口和调用关系来动态按需组合和组建这些服务,实现松耦合、功能对称的面向应用管理平台。

*)基金项目:本课题受国家"九七三"重大基础项目(2004CB318203),国家自然科学基金项目(60603074、60603075)资助。李 博 博士研究生,主要研究方向为网络存储系统、存储系统负载等。

本文第1节介绍了数据和存储管理的重要性，并结合当前研究的解决方案和科研动向，提出了本文的 VS³ 模型体系进行简单思路。第2节具体阐述了模型体系中的核心技术和VSU单元的基本思想。第3节给出了 VS³ 模型体系的详细介绍、各个模块的描述和工作流程。文章最后给出了 VS³ 模型体系的肯定性结论和我们下一步的工作重点。

2 基于虚拟服务的核心技术

从存储系统角度而言，系统中服务主要分为两类：存储服务和计算管理服务。从上一节中，我们给出了这一存储模型的意图，就是利用现有的资源动态配置和组合相应的服务来满足当前对存储的需求。因此从这一角度出发，我们希望利用虚拟化技术来抽象出现有资源所提供的服务。为此我们需要建立自己基本的抽象服务单元——VSU。

2.1 抽象服务单元

VSU(Virtual Service Unit)是该模型体系的基本构成单元，利用该结构我们可以针对系统中的资源进行服务的提取和虚拟化。从功能上讲：1)它抽象出一类具体的服务，并对相应的资源(包含物理和逻辑资源)进行了虚拟化；2)封装了一套自主的管理和调度功能；3)向外界提供统一的应用接口。

为了实现上述功能，我们把 VSU 分为4个层次：资源设备层、虚拟化层、策略层和逻辑服务管理层，如表1。

表1 VSU 的层次

层次	功能
逻辑服务管理层	维护统一的资源逻辑视图，管理和分配逻辑资源，并通过SA来接收和执行请求
策略层	提供适合的策略来实现资源的动态、按需调度；同时实时监控系统的性能变化和策略调度
虚拟化层	把负责管理的资源映射为逻辑资源，同时保证映射操作的平衡和性能
资源设备层	负责管理与配置底层的物理资源，并服务于其它的VSUs；同时监控资源的使用状况等

逻辑服务管理层包含一个服务代理 SA(Service Agent)，可以代表服务声明其服务类型、位置、该 VSU 依赖关系和属性等信息。通过最后的逻辑服务管理层，VSU 可以抽象和虚拟出特定服务，从而屏蔽了底层物理设备和应用协议的各异性。封装的服务，同时也有可能为其他 VSU 提供支撑，从而合成更为复杂的服务。

2.2 服务注册和发现机制

模型中的各类服务通过 VSU 抽象和虚拟出服务与服务之间通过服务注册和发现机制进行有效的整合。为了实现服务资源的透明调用，系统首先通过各个服务的 SA 将各类服务在服务注册中心进行注册，在中心注册其服务的类型、位置和属性等信息。同样，当服务请求者需要某项服务时，也是通过 SA 以单播方式向服务注册中心进行服务查询，并获得所需服务的信息。

2.3 VSU 依赖的抽象关系

一旦模型被确定下来，VSU 之间的依赖抽象也确定了下来，彼此之间的依赖关系也通过 VDA 来说明。这种依赖关系是指当一类服务需要使用另外一类服务时所建立依附关系。如果 VSU-B 需要 VSU-A 提供的服务，我们就说 VSU-B 依赖于 VSU-A 所提供的服务。每个 VSU 通过 SA 记录各自的依赖关系，并把它们统一保存在系统中的 VDA 中。同时也可以修改 VDA 来动态修改 VSU 间的依赖关系，SA 通过周期性地访问 VDA 来保持服务配置的更新。因此，每个 VSU 通过连接 VDA 才能知道它和谁相依赖，需要向某个 VSU 提供服务。图1表述了4个 VSUs 集合，VSU-B 需要 VSU-A 提供的服务。

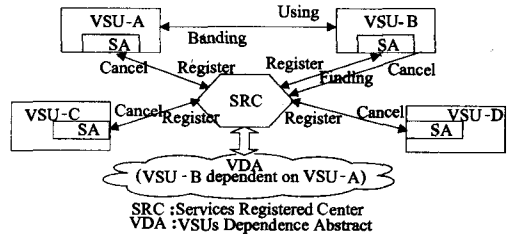

图1 VSU 间的依赖关系

图中描述了4个 VSUs 单元和 VDA，VDA 记录了模型系统中所有服务的依赖关系。每个 VSU 和 SRC 之间可以根据需要注册和注销相应的 VSU 服务。当 VSU-B 需要 VSU-A 提供的服务时，首先应保证 VSU-A 和 VSU-B 都在 SRC 注册，同时双方的 SA 都保存着它们之间的依赖关系，VSU-B 通过本地 SA 记录的依赖关系向 SRC 查询 VSU-A 的服务信息。如果此时 VSU-A 尚未启动或者注册未完成，则 VSU-B 等待，直到 VSU-A 是准备就绪状态，VSU-B 才能向 VSU-A 进行绑定和使用它所提供的服务。

3 VS³ 模型体系

VS³ 就是本文提出的基于虚拟服务存储模型，该模型体系中包含有存储资源池、计算资源池、元数据服务器、绑定服务器、应用服务器、SRC 和 VDA。其中存储资源池、计算资源池、元数据服务器、绑定服务器、应用服务器都是具体中的 VSU 单元。这样彼此间同样满足通过 SRC 的服务发现机制。

3.1 存储资源池 SRP(Storage Resource Pool)

SRP 是模型中的存储资源池，为整个系统提供存储服务，满足了用户和系统中 VSU 提供动态、大容量的存储需求。

SRP 的资源管理层对模型中出现的物理存储设备进行管理。如图2所示，SRP 管理到的物理设备可以是 iSCSI 存储子系统、NAS 存储子系统、RAID 磁盘阵列、磁带库备份系统等。这些物理设备都通过 SRP 的资源管理层实现设备的配置、注册、启动和停止等操作。

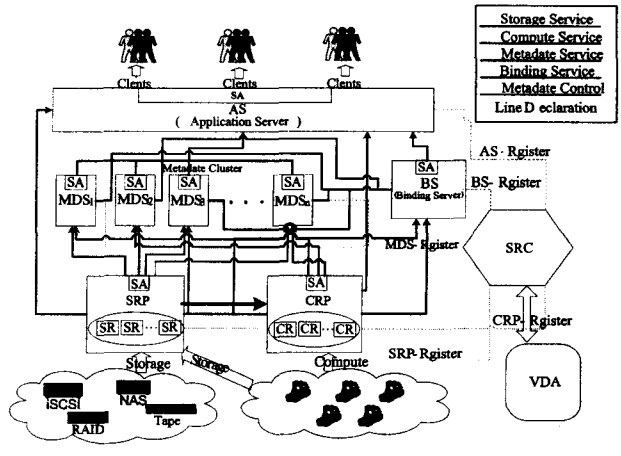

图2 VS³ 模型体系

SRP 的虚拟化层实现了所有物理存储设备的物理地址

到统一逻辑地址空间的映射，从而在 SRP 内部形成了一个个动态分配的逻辑存储资源(SR)。同时这种映射关系基于动态的映射，即在逻辑存储资源 SR 创建时，逻辑资源空间并不能得到真正的物理资源空间的支持，直到需要读写对应的 SR 单元时，虚拟化层才负责动态的映射分配的物理资源，这一机制实现了存储资源最大的使用率。

SRP 策略层实现了对逻辑存储资源 SR 的管理和使用状况的监控。通过上层逻辑服务管理层接收到的服务需求(如容量、IO 性能、安全级别、服务质量等)，动态地配置和组合虚拟化层分配的 SR 来提供相应的存储服务。

SRP 逻辑服务管理层负责接收和执行请求，对下层策略层提供的服务形成统一的带外服务。图中参照模型中统一的 VDA 视图对计算资源池、元数据服务器、绑定服务器和应用服务器提供存储服务。

3.2 计算资源池 CRP(Compute Resource Pool)

CRP(Compute Resource Pool)是模型中的计算资源池，为整个系统提供计算能力的服务，可以创建模型中特定的分布式虚拟计算环境。由于 CRP 的存在结合了 SRP 对系统中存储的支撑，从而动态组合了该模型中多个应用服务器，包括各个元数据服务器，BS 服务器和 AS 服务器。使得管理用户不用指定具体的设备，就可动态地实现需要的服务器设备，有效地利用了现有的存储和计算资源，节约了系统开发的成本，提高了系统应用扩展性。

CRP 资源设备层管理的资源是系统提供的计算节点和 SRP 中存储的系统数据。这些计算资源和数据的使用状况得到该层的监控和保护。CRP 虚拟化层、策略层通过动态的组合虚拟化出模型需要的特定服务器。资源设备管理掌控的计算资源(一些分布式的计算机节点)和 SRP 提供的存储资源，通过网络块协议把特定的存储资源和某些计算节点进行绑定。在这种情况下，特定的存储资源就可以为虚拟服务器提供本地的存储域来存储系统数据，同时策略层整合相应的节点机来构成独立的处理单元，这样在 CRP 的作用下生成了可用的服务器。一旦虚拟服务器关机，所有这些资源回归各自独立的状态。位于 CRP 最上层的逻辑服务管理层主要负责对虚拟服务器进行管理，监控和调控虚拟服务器的使用和运行状态。在本模型系统中，元数据服务器，BS 服务器和 AS 服务器都是通过这一机制实现的。利用的资源也是系统已有的物理资源，不需要添加任何具体的计算机。同时计算资源和存储资源的分离[18]也加强动态虚拟化性能，为模型日后的升级换代提供了良好的扩展性。

3.3 元数据服务器集群

在模型系统中通过多个元数据服务器和 BS 服务器组成了一个 Metadata Servers Cluster 子系统，为系统提供元数据服务。元数据是文件系统中用来描述数据的数据，通常可以为文件的属性、目录的内容、文件的数据块等。

当用户通过 AS 对存储系统进行文件或者目录操作，AS 会把这一请求发送到 Metadata Servers Cluster，由它来返回文件或目录的属性信息和在 SRP 的位置视图。当 AS 收到 Metadata Servers Cluster 返回的 File/Dir Mapping，根据这一元数据，AS 可以对 SRP 进行实际的数据读取。这一过程分离了数据传输和控制操作，所有的文件数据直接可以在 AS 和 SRP 间进行传输，避免了 MDS 的转发。同时 Metadata Servers Cluster 也可以根据需要对模型中存放的数据进行分片(striping)，分别在 SRP 中存储在不同的 SR 中，平衡文件传输时的负载，提高了模型数据读取的性能。

同时 Metadata Servers Cluster 子系统支持多个 MDS，并通过绑定 BS 实现多个 MDS 协同工作，并向 AS 提供一个完整统一的名字空间。子系统中的 MDS 可以通过绑带控制，注册、挂载到 BS 端，同时 BS 也可以通过撤销控制来释放某个 MDS。多个 MDS 的存在使得存储系统可以处理大量的元数据服务，避免了单一 MDS 出现的瓶颈问题。由 BS 来协调 MDS 间的操作，可以动态决定把当前活跃的元数据分派给空闲的 MDS，平衡了 MDS 间的负载。但模型强调，系统中同一个元数据只能由一个 MDS 来管理，所有的 MDS 管理的元数据不存在重叠，BS 实现了完整的元数据管理。

图 3 中 SRP 为 Metadata Servers Cluster 子系统提供元数据的存储，两者可以通过 Metadata Path 进行读写。

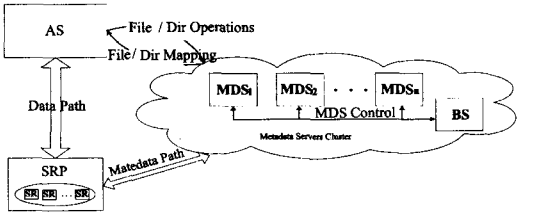

图 3 VS^3 中数据和元数据的操作

应用服务器(Application Server)

AS 在模型系统中是直接面对 clients 的模块，为客户提供用户级服务。Client 通过 AS 访问存储资源时，AS 需要从 BS 指定的 MDS_i 获得 Client 要访问文件或目录的元数据和访问控制信息。AS 根据 Client 的元数据和访问控制信息来决定该 Client 是否对要访问数据有相应的访问权限。若没有则拒绝服务；若有，根据所访问数据的物理位置视图建立数据传输连接。最后，由 AS 向 Client 提供需要访问数据，如图 4。

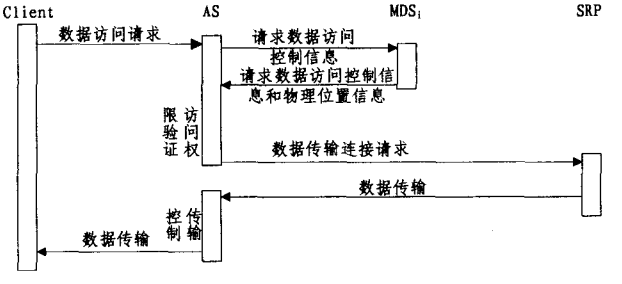

图 4 UML 时序图

如图 2 所示，存储资源池、计算资源池、元数据服务器和绑定服务器为 AS 提供资源服务。SRP 提供存储资源，CRP 提供计算资源，MDS 提供元数据服务，而 BS 提供 MDS 管理服务。AS 监控资源的使用状况，对失效的资源服务进行重建或恢复等操作，并完成用户请求和管理命令到实际资源的传达。

AS 的虚拟化层和策略层实现了服务的按需配置、启动/停止、系统的自动配置、服务的查询、管理资源的动态组合和服务的个性化定制等。

AS 的逻辑服务管理层管理 AS 本身的运作和监控，提供 AS 的启动和定制，同时通过该层包含的 SA 来接收用户的请求，用户访问权限的验证和向下层策略层提供用户存储请求的指标。

(下转第 51 页)

2000个),平均每个节点发送的报文数量会随网络规模的增大而减少,所以造成了网络规模越大节点能耗越低的现象。

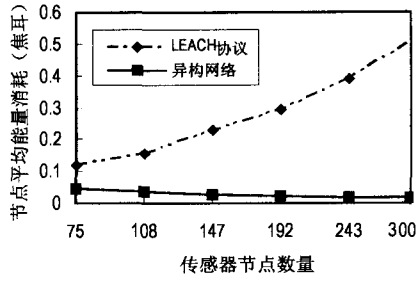

图2 不同网络规模下异构网络和LEACH协议节点平均能量消耗对比结果图

然后,针对分组发送成功率进行了仿真测试,结果如图3所示。从图中可以得出两点结论:(1)异构网络的分组发送成功率总是高于LEACH协议,尤其在大规模的网络场景中;(2)在各种异构网络场景中,分组发送成功率基本上保持在99%左右。

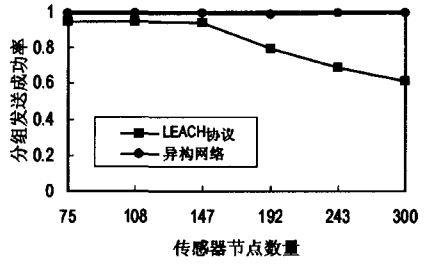

图3 不同网络规模下异构网络和LEACH协议分组发送成功率对比结果图

结束语 针对异构无线传感器网络中异构节点的部署问题(包括异构节点的数量和位置),本文提出了一种基于选址问题理论的解决方法。理论分析和仿真测试的结果表明,该算法可以适用于随机部署的无线传感器网络,能够明显延长网络寿命、提高分组发送成功率。

参考文献

[1] ns-2 Network Simulator. http://www.isi.edu/nsnam/ns/

[2] Cerpa A, Elson J, Estrin D, et al. Habitat monitoring: application driver for wireless communications technology // ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean. Costa Rica, April 2001

[3] Heinzelman W, Chandrakasan A, Balakrishnan H. An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on Wireless Communications, 2002, 1(4): 660-670

[4] Kumar R, Tsiatsis V, Srivastava M B. Computation Hierarchy for In-Network Processing // The 2nd Intl. Workshop on Wireless Networks and Applications. San Diego, CA, Sept. 2003

[5] Mainwaring A, Polastre J, Szewczyk R, et al. Wireless Sensor Networks for Habitat Monitoring // Intl. Workshop on Wireless Sensor Networks and Applications (WSNA '02). Atlanta, GA, Sept. 2002

[6] Rhee S, Seetharam D, LiuS. Techniques for Minimizing Power Consumption in Low Data-Rate Wireless Sensor Networks // IEEE Wireless Communications and Networking Conference. Atlanta, GA, March 2004

[7] Wang H, Estrin D, Girod L. Preprocessing in a Tiered Sensor network for Habitat Monitoring // IEEE Conf. on Acoustics, Speech, and Signal Processing. Hong Kong, China, April 2003

[8] Yarvis M, Kushalnagar N, Singh H, et al. Exploiting Heterogeneity in Sensor Networks. IEEE InfoCom, Miami, FL, 2005

[9] 束金龙,闻人凯. 线性规划理论与模型应用. 北京:科学出版社, 2005

(上接第33页)

结束语 本文研究了如何在开放的、可扩展的网络应用环境下建立按需动态配置的存储服务管理。基于存储虚拟化技术,将文献[16]的资源服务模型引入到了存储领域,提出了VS^3模型,设计了VS^3模型中的服务发现机制、数据传输机制。同时在减少原有设备成本的前提下,设计了计算和存储资源相分离的策略,从而为合成新应用服务提供了必要的资源准备。

下一步研究将以此模型为原型构建面向GIS(Geographic Information System,地理信息系统)领域的应用,从现实应用的角度实现VS^3思想,同时进一步为面向存储服务、按需服务提供更为切实的尝试。

参考文献

[1] Golding R, et al. Storage Resource Management for the Enterprise Server. BMC Software, Inc. 2001

[2] 俞建新,等. 网络存储新技术评析. 计算机工程,2006(10)

[3] 李必刚,舒继武,等. 一种基于集群环境的虚拟存储系统研究与实现. 小型微型计算机系统,2006(6)

[4] 张成峰,等. 网络存储的统一与虚拟化. 计算机科学,2006(6)

[5] 王迪,舒继武,等. 块级别的海量存储虚拟化系统. 清华大学学报(自然科学版),2007(1)

[6] 黄华,许鲁,等. 蓝鲸分布式文件系统的分布式分层资源管理模型. 计算机研究与发展,2005(6)

[7] Pazel D P, Eilam T, Fong L L, et al. A Dynamic Resource Allocation and Planning System for a Cluster Computing Utility // Cluster Computing and the Grid, 2002 2nd IEEE/ACM International Symposium. May 2002: 57-57

[8] Shen K, Yang T, Chu L. Clustering support and replication management for scalable network services. Parallel and Distributed Systems. IEEE Transactions, 2003, 14(11): 1168 - 1179

[9] Fong L L, Kalantar M, Pazel D P, et al. Dynamic resource management in an eUtility // Network Operations and Management Symposium 2002, NOMS 2002. 2002 IEEE/IFIP. April 2002: 727-740

[10] Appleby K. Oceano-SLA based management of a computing utility // Integrated Network Management Proceedings, 2001 IEEE/IFIP International Symposium. May 2001: 855-868

[11] Kim H L, Jeong C S. Distributed Object-Oriented Virtual Environment using Web Services on Grid // Computer and Information Technology, 2006. CIT '06. The Sixth IEEE International Conference. Sept. 2006: 66

[12] Zhu Zhenmin, Su Xiaoli, Li Jintao, et al. A User-centric Service Framework for Pervasive Computing // Pervasive Computing and Applications, 2006 1st International Symposium. Aug. 2006: 1-4

[13] Rousselle P, Tymann P, Hariri S, et al. The virtual computing environment // High Performance Distributed Computing, 1994 Proceedings of the Third IEEE International Symposium. Aug. 1994: 7-14

[14] Casselman S. Virtual computing and the Virtual Computer // FPGAs for Custom Computing Machines, 1993 Proceedings. IEEE Workshop. April 1993: 43 - 48

[15] Xu Huiying, Zhao Jianmin, Zhu Xinzhong, et al. Web Services Based On Grid Technology // Computer Supported Cooperative Work in Design, 10th International Conference. May 2006: 1-4

[16] 王敏,等. 一种虚拟化资源管理服务模型及其实现. 计算机学报,2005(5)

[17] 吴松,金海. 存储虚拟化研究. 小型微型计算机系统,2003(4)

[18] 马一力,等. 存储与计算的分离. 计算机研究与发展,2005(3)