优化的信息中心虚拟化实施方案①

刘德文, 倪 明, 颜爱良

(中国电子科技集团 第三十二研究所, 上海 200233)

摘 要: 基于传统信息中心建设中存在的一些缺点, 提出了一种基于 vSphere 架构的优化的信息中心虚拟化实施方案. 相对于其他的虚拟化实施方案, 本方案主要从可靠性和存储设备 I/O 性能两方面进行了优化. 该方案从服务器、网络和存储设备等三个方面对系统可靠性进行了优化, 分析了影响存储设备 I/O 性能的几大因素, 并分别从是否采用多路径输入输出、Write-back cache 写策略、磁盘 I/O 块大小、服务器性能等几个方面对系统 I/O 性能进行了优化. 最后通过对整个系统的测试, 验证了这一方案的优势.

关键词: 信息中心; 可靠性; I/O; 多路径; write-back

Optimization Scheme of Information Center Virtualization

LIU De-Wen, NI Ming, YAN Ai-Liang

(The 32nd Research Institute, China Electronics Technology Group Corporation, Shanghai 200233, China)

Abstract: Based on the defects of traditional information center construction, this paper presents an implementation of virtual optimization vSphere architecture scheme based on informationcenter. Compared with other embodiments of virtualization, the scheme is mainly optimized from two aspects of reliability and storage I/O performance. The scheme optimizes the system reliability from three aspects of the server, network and storage equipment. It analyzes the influence factors of storage device's I/O performance. These factors are separately whether to use the multipath input and output, write-back cache write strategy, I/O disk block size and service performance etc. Finally, through the test of the whole system, we proved the advantages of the scheme.

Key words: information center; reliability; I/O; multi-path; write-back

1 引言

随着 IT 行业的迅猛发展, 各种信息服务在数量和种类上快速增长, 而更多的企业也正在将业务向无纸化办公发展, 这使得信息服务的需求越来越大. 数据中心为信息服务提供了运行的平台, 而对于信息服务的需求也就是对于新一代数据中心的需求.

本文正是基于上海某研究所信息中心改造项目, 该研究所信息中心采用的是传统的信息中心建设方式, 各项应用呈现孤岛式陈列方式, 独立地分布于不同的服务器上. 随着企业的不断发展, 这种建设方式以越来越不能满足公司的需求, 主要存在如下这些缺点: 1)管理复杂度高、扩展性差. 为了保障各项服务的独立性, 不同的服务通常部署在不同的服务器上, 各种服务器性能不一, 内容庞杂, 容易造成单点故障频发, 同时造成系统管理员维护量大, 管理困难. 此外, 孤岛式数据中心扩展性差, 当扩展新业务时需要部署专门的网络、服务器和存储等基础设施, 容易形成数据孤岛[1]; 2)资源利用率差, 空闲资源多. 服务器的配置往往根据业务高峰期的计算请求而申购高额的配置, 所有服务器的 CPU、Memory、Disk Storage 等资源平时利用率并不高, 且空闲的资源并不能用于其他的应用服务[2]; 3)数据容灾困难, 可靠性差. 在孤岛式信息中心中, 如果某个服务器突然发生宕机故障, 该服务器提供的服务也就相应停止, 无法做到持续运作. 此外, 庞杂的软硬件环境是系统的快速备份和有序恢复都带来了不小的难度, 数据中心管理人员需要在不同

① 收稿时间:2014-07-10;收到修改稿时间:2014-08-25

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的设备之间准备不同的备份和恢复方法, 难以做到统一数据备份.

伴随着虚拟化技术的不断发展, 基于虚拟化技术搭建的信息中心已经可以很好地解决以上这些缺点. 计算资源和存储资源的的虚拟化可以实现对这些 IT 资源的集中化管理, 一方面用户可以从繁重、复杂的 IT 资源管理中解脱出来, 简化了资源的管理, 优化了系统的扩展性; 另一方面, 虚拟化技术实现的服务整合与按需供给的新模式是资源利用率得到显著提高, 单位服务的能耗量也大幅度降低. 此外, 虚拟化提供的克隆和动态迁移等技术为信息中心实现了动态可伸缩性、高可用性和负载均衡等特性, 提高了信息中心的可靠性.

2 相关概念

2.1 服务器虚拟化技术

当前市场上最流行的服务器虚拟化产品主要有两种, 分别是 Openstack 和 VMware vSphere. 其中 Openstack 是开源免费的, 且拥有庞大并不断增长的社区, 具有的良好的发展和应用前景. 但 Openstack 更适合大规模应用架构, 功能还不够丰富, 性能稳定性有待认证. 相较而言, vSphere 虽然是商业的, 但 vSphere 的功能更加完善, 可靠性更好, 且更适合小型架构. 因此本方案中采用 vSphere 架构进行服务器虚拟化.

vSphere 是 VMware 公司推出的一套服务器虚拟化解决方案. vSphere 的核心组件为 VMware ESXi, 它可以独立安装和运行在裸机上, 不需要依存于宿主操作系统. VMware ESXi 主要用于调配物理服务器中内存、CPU、存储及网络等各种硬件资源, 并将这些资源分配到运行在其中的个虚拟机中. vSphere 提供 vCenter Server 用于整合和集中管理各种虚拟资源, 并对这些资源进行按需分配. 此外, 在 vSphere 中, vCenter Server 还提供 vMtion、Storage vMotion、DRS、HA、Fault Tolerance 等高级功能.

2.2 网络存储相关技术技术

2.2.1 双机热备

双机热备系统就是对于重要的服务, 使用两台服务器共同执行同一服务, 在运行的过程中实时互相备份, 以冗余提高系统的可靠性. 当一台服务器出现故障时, 另一台可以立即发现故障并接管那台服务器以承担服务任务, 从而在不需要人工干预的情况下, 自动保证系统能持续提供服务[3].

2.2.2 RAID 技术

独立磁盘冗余阵列(RAID,Redundant Array of Independent Disk), 简称磁盘阵列[4]. 其基本思想就是把多个相对便宜的硬盘组合起来, 成为一个磁盘阵列组, 使性能达到甚至超过一个价格昂贵、容量巨大的磁盘. RAID 可以充分发挥多块磁盘的优势, 提升整体的读写速度, 增大存储容量, 保证数据安全性. 通过对多块磁盘空间的数据管理方法, 定义了几种 RAID 级别, 主要包括 RAID0, RAID1, RAID10, RAID3, RAID5, RAID6 等. 目前使用最多的是 RAID5, RAID5 存在校验数据, 允许一个磁盘出现故障, 不会造成数据丢失和服务中断, 具有较高的可靠性.

2.2.3 iSCSI 多路径存储

iSCSI 是当下最流行的存储区域网络标准, 它具有价格便宜、管理方便和传输速度快等优点. 此外, iSCSI 还提供多路径存储访问的机制, 在各个路径之间建立负载平衡、故障切换等带宽聚合应用, 提供更可靠的存储网络环境, 同时还能提升 iSCSI 的可用带宽.

3 方案设计

数据中心的虚拟化就是打破原始物理结构之间的隔断, 将物理资源转变为逻辑上可直接调控管理的资源. 总体架构如图 1 所示, 整个架构主要包括三个部分: 存储层、网络层、服务器层. 存储层将磁盘矩阵集中起来进行统一的管理, 通过虚拟化的技术使前端主机脱离对后端存储设备的依赖, 使虚拟层作为存储服

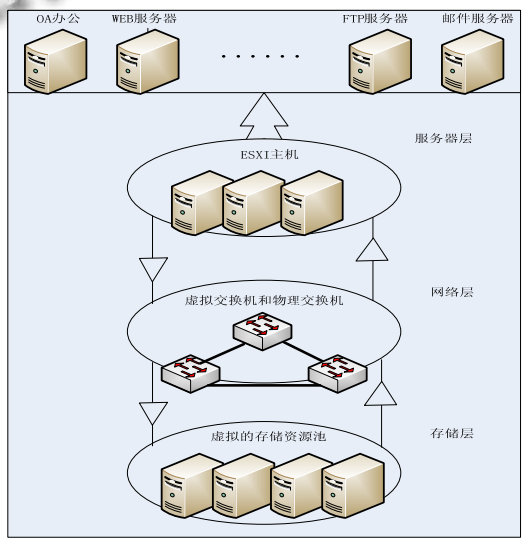

图 1 总体架构图

© 中国科学院软件研究所 http://www.c-s-a.org.cn

务的基础, 从而带来更高的资源利用率、无缝数据迁移等效果. 网络层通过 vSphere 虚拟交换机和物理交换机的结合使用, 不仅提高了网络带宽、保证了网卡的负载均衡, 同时还对网络的各个环节提供了冗余, 保证了网络层的可靠性. 服务器层利用 vSphere 虚拟化技术实现对计算资源的简化管理和按需分配, 不仅提高了资源管理的效率, 还提高了资源的利用率. 以下将分别从这三个方面对整个设计方案进行分析.

3.1 服务器模块

为适应不同应用对服务器性能需求的不同, 本方案采用了两种性能不同的服务器. 其中 E3 服务器共有 6 个节点, 每个节点拥有 4 核 CPU, 32G 内存; E5 服务器共有 4 个节点, 每个节点拥有 8 核 CPU, 48G 内存. 为两种服务器分别建立资源池, 不同的应用建立在不同的资源池内, 可以达到提升计算资源利用率的目的.

软件方面, 每个节点上安装 ESXi 组件, 实现对服务器计算资源的虚拟化, 方便统一高效管理. 在一台 E3 服务器的一个节点上安装 vCenter Server 作为私有云管理机主节点, 同时在另一台 E3 服务器的一个节点上也安装 vCenter Server 作为私有云管理机从节点. 两个节点之间以纯软件的方式实现 active-standby 双机热备, 当主节点发生故障或主节点所在主机发生故障而无法提供服务时, 将自动切换到从节点继续提供服务.

3.2 存储模块

本方案采用两台存储设备双机热备的方式来提供存储, 存储设备架构图如图 2 所示, 图中左边为存储设备 A, 右边为存储设备 B. 每台存储设备内使用 RAID 控制器做成 SCSI 磁盘阵列, RAID 级别为 RAID5. 每两台存储设备间实现 1:1 的双机热备, 使用 heartbeat 心跳线和千兆网口实现故障侦测, 使用 DRDB 和直连的万兆光纤实现数据同步.

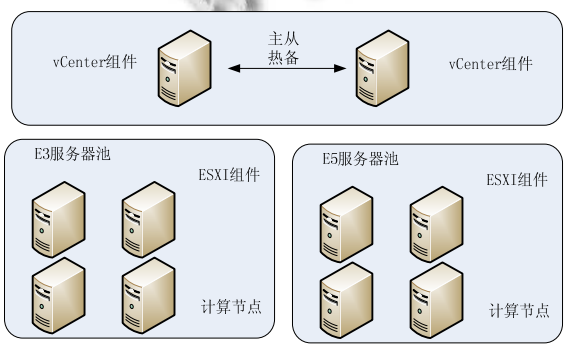

图 2　服务器架构图

本存储方案采用存储设备内部的 RAID5 和之间的双机热备实现了存储设备的冗余, 确保了共享存储的可靠性. 这种存储方案与文献[5]中的双控制器 RAID 方案相比, 虽然磁盘利用率比较低, 但可靠性更高, 本方案即使在一台存储设备完全失效的情况下, 仍可以正常运行, 而文献[5]中则做不到这一点. 此处双机热备的另一个好处在于, 由于 RAID 控制器中的 cache 采用的是 Write-back 策略(Cache 写策略有 Write-back 和 Write-through 两种, Write-back 写策略能有效提高存储 I/O 速度[5]), 如果不实用双机热备, 当存储设备发生故障掉电, cache 中的脏数据将来不及写回磁盘, 从而造成 cache 中的数据丢失. 如果采用双机热备, 数据将同时向两个存储设备的 cache 中读和写, 即使一台设备发生故障, 另一台依然能够正常工作, 避免了数据丢失. 最后, 由文献[6]可知, vSphere 5 上的数据块大小固定为 1MB/块, 而存储设备上的磁盘块大小可设置为不同的大小, 块大小不同, 存储设备 I/O 的速度也不同. 经测试, 将存储设备上的块大小设置为 1MB/块时, 存储设备 I/O 速度最快, 因此本方案中将存储设备中磁盘块大小设置为 1MB/块.

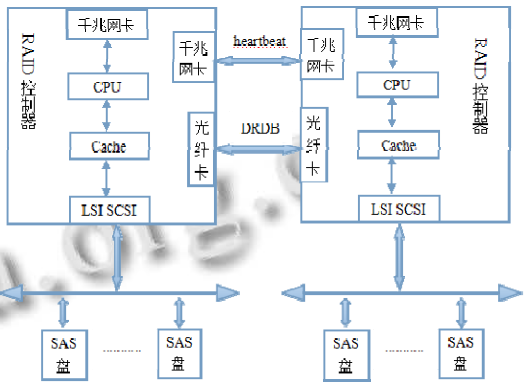

图 3　存储设备架构图

3.3 网络模块

为了使数据中心的网络拓扑更加清晰规范, 方便工作人员维护和管理, 本方案中将网络链路分为三类, 每种链路负责不同的功能. 这三种链路分别为存储链路、管理链路和业务链路. 其中存储链路负责访问 iSCSI SAN 存储, 对存储 I/O 性能有要求. 管理链路用来连接 vCenter 和每台主机, 负责对整个系统的计算资源和存储资源进行管理, 对整个系统至关重要, 需要保证其可靠性, 但该链路只传输管理命令, 数据流通量较小, 因此对 I/O 性能没有严格要求. 业务链路负责

© 中国科学院软件研究所 http://www.c-s-a.org.cn

管理各项应用服务, 由于各项应用分布于不同的主机上, 同时为了防止不同的应用服务间相互干扰, 需要为不同的应用分配不同的网段, 在此本方案使用分布式虚拟交换机进行集中管理.

如图 4 所示, 本方案中将用于业务链路的两个 48 口的交换机通过万兆堆叠模块实现堆叠, 将用于管理链路和存储链路的两个 96 口交换机通过两个聚合的万兆堆叠模块实现堆叠. 这样做的好处有三个方面, 首先可以使两个 48 口交换机和两个 96 口交换机分别作为一个整体交换机, 便于统一管理; 其次增加了用户端口, 由于在交换机之间建立了一条较宽的宽带链路, 这样每个实际使用的用户带宽就有可能更宽; 最后, 堆叠交换机结合冗余链路可以实现交换机的冗余, 即将同一个应用的两条冗余链路连接到两台堆叠交换机的其中一个, 当其中的一台交换机发生故障, 另一个交换机仍能保证应用的正常运行.

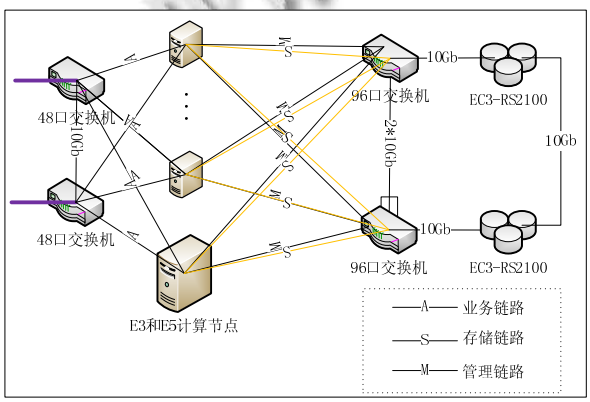

图 4 网络拓扑图

图 4 中的服务器 E3 和 E5 上均有 6 个千兆网口. 本方案中使用 2 个千兆网口实现存储链路, 2 个网口之间实现多路径输入输出[7], 当多个业务并发访问共享存储时, 可显著地提高存储 I/O 速度, 相较于文献[2]中的单路径输入输出方式, 可以更好地适应信息中心实际应用中对存储设备 I/O 性能的要求. 管理链路使用 2 个千兆网口, 将两个网口做成 Nic team, 形成冗余, 分别将两条冗余的链路连接到两台堆叠的 96 口交换机上, 确保管理链路的可靠性. 在 vCenter 中创建分布式虚拟交换机, 将每台主机的剩余 2 个千兆网口做成 Nic team 并实现网口聚合, 用作业务链路, 分别将两条冗余的链路连接到两台堆叠的 48 口交换机上, 确保业务链路的可靠性.

4 测试结果

实验环境为: 一台 E3 服务器、一台 E5 服务器、两台 RS2100 存储设备, 每台提供 10TB 的存储容量、两台 48 口的 H3C 交换机和两个 96 口 H3C 交换机. 按图 1 所示搭建好环境后, 在 vCenter Server 中创建一个包含 6 个 E3 服务器节点的 E3 集群, 和一个包含 E5 服务器节点的 E5 集群, 并开启集群的 HA 功能.

4.1 可靠性测试

(1)服务器可靠性测试包括三项, 如表 1 所示.

表 1 服务器可靠性测试

测试方法	测试结果
将 vCenter 主节点断电	vCenter 从节点接管服务
将 vCenter 所在服务器断电	vCenter 从节点接管服务
在 E3 集群中断开一个服务器节点	虚拟机自动迁移到集群中的其它节点上上

(2)存储设备可靠性测试包括三项, 如表 2 所示.

表 2 存储设备可靠性测试

测试方法	测试结果
拔掉一台存储设备链路	存储仍然有效
拔掉一台存储设备中的硬盘	存储仍然有效
拔掉一台存储设备的电源	存储仍然有效

(3)网络设备可靠性测试包括两项, 如表 3 所示:

表 3 网络设备可靠性测试

测试方法	测试结果
拔掉一台 48 口交换机电源	业务虚拟机正常
拔掉一台 96 口交换机电源	存储仍然可用

4.2 I/O 性能测试

此处采用在共享存储不同卷之间克隆虚拟机的方式来测试 I/O 速度, 使用 iostat 工具来查看实时 I/O 速度. 本方案中, 影响 I/O 性能的因素主要有 4 个: 磁盘块大小、cache 写策略和是否采用多路径存储. 以下分别对这三个方面进行测试:

(1)在 E5 集群内, 设置 cache 写策略为 write-through, 不使用存储多路径, 在共享存储不同卷之间克隆一台虚拟机时, 不同磁盘块大小对存储 I/O 性能的影响如下图所示:

(2) 在 E5 集群内, 将磁盘块大小设置为 1MB/块, 不使用 iSCSI 存储多路径时, 如果 cache 采用的是 Write-back 写策略, I/O 速度约为 32MB/s; 如果采用的是 Write-through 写策略, I/O 速度约为 87MB/s. 可见 cache 的写策略对于 I/O 速度的影响十分巨大.

(3) 在 E5 集群内, 将磁盘块大小设置为 1MB/块, cache 写策略设置为 Write-back, 通过在存储设备不同

© 中国科学院软件研究所 http://www.c-s-a.org.cn

卷之间克隆多台虚拟机来模拟多应用并发. 图 4 显示了在不同应用并发数下, 采用 iSCSI 多路径输入输出和不采用 iSCSI 多路径输入输出是的 I/O 速度对比. 由图可知, 采用 iSCSI 多路径输入输出可以有效地提高存储区域网络的 I/O 速度.

(4) 将磁盘块大小设置为 1MB/块, Cache 写策略设置为 Write-back, 使用 iSCSI 存储多路径, 通过在存储设备不同卷之间克隆多台虚拟机来模拟多应用并发. 在不同的应用并发数下, 分别在 E3 和 E5 集群内测试 I/O 速度. 测试结果如图 5 所示:

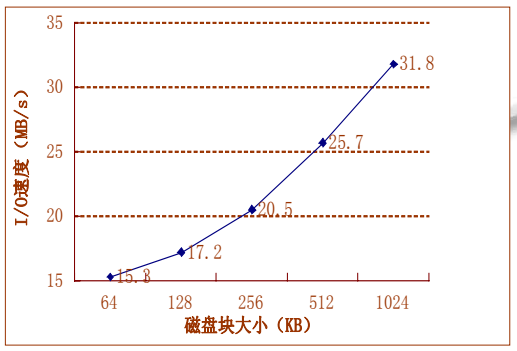

图 5　磁盘块大小对 I/O 速度的影响

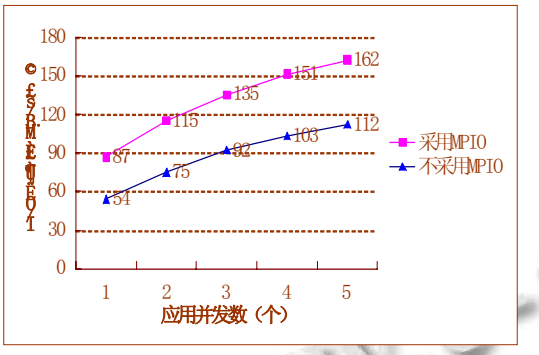

图 6　多路径输入输出对 I/O 的影响

4.3 设计方案效果分析

采用本方案的虚拟化架构对信息中心进行优化之后, 针对传统信息中心建设中存在的缺点, 本方案具有以下优势:

(1) 高可靠性. 由可靠性测试结果可知, 由于对可能引发系统宕机的各个模块均设有冗余, 整个系统十分可靠. 此外, 结合 vSphere 架构提供的虚拟机克隆、动态迁移、高可用性等优点, 可以很好地克服传统信息中心建设中容灾困难, 可靠性差的缺点.

(2) 存储 I/O 性能高. 由测试结果可知, vSphere 架构提供的 iSCSI 多路径输入输出和 RAID 控制器 Write-through 写策略均能有效提高存储 I/O 性能. 此外, 合适的磁盘块大小和好的服务器性能也可以进一步提升存储 I/O 性能.

(3) 提高资源利用率. 在进行虚拟化整合后, 利用 vSphere 架构的 DRS 功能实现服务器虚拟机的负载均衡, 利用 vSphere 架构的 Storage vMotion 功能实现存储的负载均衡. 在保证正常提供服务的前提下, 大大提高了 CPU、内存和存储的利用率.

(4) 简化管理和可扩展性. vSphere 架构的 vCenter 组件实现了对计算资源和存储资源的集中管理, 针对不同的应用需求实现按需分配, 极大地简化了信息中心的管理与维护. 同时, 如果需要添加新的设备, 只需要通过网络接入, 然后加入 vCenter 管理中心即可, 扩展十分方便.

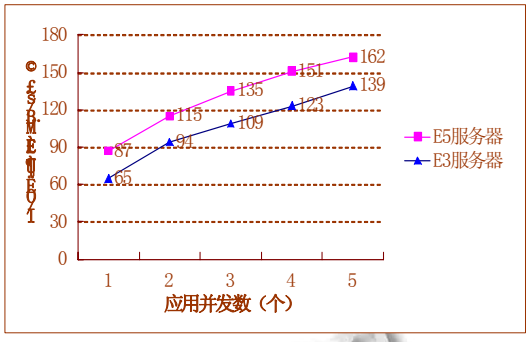

图 7　服务器性能对 I/O 速度的影响

5 总结

本文提出了一种优化的信息中心虚拟化实施方案, 该方案对传统信息中心建设中存在的各项缺点进行了优化. 尤其是在提高系统可靠性和存储设备 I/O 性能方面, 可靠性方面, 从服务器模块、存储设备模块和网络模块三个方面对系统的可靠性进行了优化. 测试结果证明, 整个系统的可靠性十分高. 足以应付实际应用中的一些不确定的故障. I/O 性能方面, 系统从服务器性能、存储多链路、磁盘 I/O 块大小和 RAID 控制器 cache 写策略等 4 个方面对 I/O 性能进行了优化. 测试结果表明 I/O 性能均有明显的提高.

参考文献

1 施庆.基于 VMware vSphere 的高校数据中心虚拟化建设研究[硕士学位论文].上海:复旦大学,2012.

2 储久良,李玲.虚拟化技术在高校数据中心的应用.实验室研

© 中国科学院软件研究所 http://www.c-s-a.org.cn

究与探索,2012,31(12):67–71.
3 刘晓洁,黄永佳.基于 Linux 的双机热备系统的实现技术.计算机应用研究,2007,24(4):255–257.
4 赵亮.高性能磁盘阵列(RAID)关键技术的研究[硕士学位论文].成都:国防科技大学,2002.
5 严亮.双控制器 RAID 系统的研究与实现[硕士学位论文].武汉:华中科技大学,2012.
6 VMware.What's New in VMware Virtual SAN. VMware Technical White Paper, 2014.
7 VMware.Multipathing Configuration for Software iSCSI Using Port Binding. VMware Technical White Paper, 2014.
8 黄昊晶,崔志明.一种以 vSphere 为核心的私有云基础架构设计方案.微电子学与计算机,2011,28(4):38–41.
9 缪军海,朱兰娟,吴智铭.RAID 中 Cache 的设计与十年.微型电脑应用,2001,17(4):29–31.
10 焦繁.论软件 SAN 存储多路径的实现方法.特别关注,2012,12(下):30–32.
11 胡嘉玺.虚拟智慧 VMware vSphere 运维实录.北京:清华大学出版社,2011.
12 谢阳,史有群,陶然,潘乔.基于虚拟化技术的教学云平台构建与管理.计算机与现代化,2013,8:218–221.

www.c-s-a.org.cn
www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn