HZ BOOKS
华章科技

PEARSON

由来自EMC 官方的世界级虚拟化技术顾问亲自撰写，EMC高级副总裁Chad Sakac鼎力推荐

既详细讲解vSphere 5的核心技术和工作机制，又系统阐述利用vSphere 5构建一个更加可靠、易于管理、节约成本、更加环保的虚拟数据中心的完整过程和各种技术细节

VMware vSphere 5: Building a Virtual Datacenter

VMware vSphere 5

虚拟数据中心构建指南

（法）Eric Maillé　René-François Mennecier　著

姚军　等译

机械工业出版社
China Machine Press

更快地交付更多的服务，并提供更大的灵活性，这是每一位数据中心管理员的工作职责之一。他们不仅要高效地处理大量的数据，面对史无前例的复杂性，还必须在更低的预算和更少的资源下完成这一切工作。用VMware的vSphere 5进行数据中心虚拟化是实现这些目标和加速云服务迁移的最佳途径。本书详细讲解了在数据中心环境中评估、规划、实施和管理vSphere 5的所有实用知识。

本书的两位作者都是来自EMC官方的顶级数据中心虚拟化顾问，以数据中心管理员和专家的视角，首先介绍了vSphere 5基本的功能和优势、vSphere的架构、vCenter Server和ESXi 5.0等核心组件；然后转向实施，给出了详细的示例、解决方案和从他们的丰富经验中提取的最佳实践，分享了对预算、安排和规划的实用观点，正确架构的选择，以及vSphere和现有数据中心要素（包括服务器、存储、群集、网络基础架构和业务持续性计划）的整合；最后提供了一个完整的实战案例：一个用于支持特定业务目标的数据中心虚拟化项目。

本书主要内容包括：

- 评估环境中数据中心虚拟化的潜在好处
- 组织和管理到虚拟化数据中心的顺利迁移
- 预测数据中心虚拟化相关的特定挑战和风险
- 在优化稳定性、灵活性、伸缩性和成本中权衡
- 为环境选择最佳的安装/配置选项
- 有效地将vSphere 5虚拟化与现有数据中心要素联系起来
- 从vSphere 5强大的新数据中心特性中得到更多价值
- 提供存储以有效地在目前和未来支持你所部署的VM
- 管理有限的内存和其他服务器约束
- 利用新的服务持续性和高可用性选项
- 将备份架构作为降低成本的手段

PEARSON

PEARSON
www.pearson.com

客服热线：(010) 88378991, 88361066
购书热线：(010) 68326294, 88379649, 68995259
投稿热线：(010) 88379604

读者信箱：hzjsj@hzbook.com
华章网站：www.hzbook.com
网上购书：www.china-pub.com

上架指导：计算机/虚拟化
ISBN 978-7-111-41677-7
9 787111 416777 >
定价：59.00元

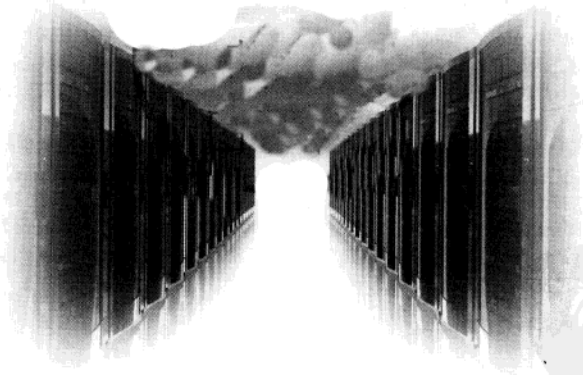
华章程序员书库

VMware vSphere 5: Building a Virtual Datacenter

VMware vSphere 5

虚拟数据中心构建指南

（法）Eric Maillé　René-François Mennecier　著
姚军　等译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

VMware vSphere 5虚拟数据中心构建指南 /（法）麦里（Maillé, V.），（法）门内尔（Menecier, R. F.）著；姚军等译. —北京：机械工业出版社，2013.3
（华章程序员书库）
书名原文：VMware vSphere 5：Building a Virtual Datacenter

ISBN 978-7-111-41677-7

I. V… II. ①麦… ②门… ③姚… III. 虚拟处理机–指南 IV. TP338-62

中国版本图书馆CIP数据核字（2013）第039336号

本书是来自EMC官方的VMware vSphere 5虚拟数据中心构建指南，由EMC 官方的两位世界顶级虚拟化技术顾问亲自撰写，EMC高级副总裁Chad Sakac鼎力推荐。既详细讲解vSphere 5的核心技术和工作机制，又系统阐述利用vSphere 5构建一个更加可靠、易于管理、节约成本、更加环保的虚拟数据中心的完整过程和各种技术细节。

全书共8章：第1章系统介绍了服务器虚拟化技术及其使用要素、虚拟化环境规范、虚拟化的好处及其3个阶段，以及整个虚拟化生态系统；第2章介绍了vSphere 5的演变历程和各种架构组件；第3章详细讲解了vSphere 5的存储形式、技术、组件和机制；第4章讲解了任何虚拟化解决方案中都必需的服务器和网络组件及其技术细节；第5章介绍了高可用性及其恢复计划；第6章介绍了vSphere 5中的备份方法和技术原理；第7章讲解了利用vSphere 5构建虚拟数据中心的实施细节；第8章则通过一个完整的案例演示了构建和管理一个虚拟化项目的流程和技术细节。

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）
责任编辑：谢晓芳
北京市荣盛彩色印刷有限公司印刷
2013年3月第1版第1次印刷
186mm × 240 mm · 13.5印张
标准书号：ISBN 978-7-111-41677-7
定 价：59.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

译 者 序

每个 IT 经理可能都有过这样的经历，在每年年终的时候，公司批下来的预算又减少了，但是业务部门对服务水平的要求却逐年上升，在这种时候，可能大家都有“巧妇难为无米之炊”的感叹。

我们该怎么办？也许很多人的答案就是据理力争，要求追加预算，但雪上加霜的是，机房的可利用空间和电力供应都到了极限，已经再也容不下新的设备了。此外，许多机构因为法律规范和公司的要求，还需要建立业务持续性计划和灾难恢复计划，在这种情况下，无论从财政上还是从技术上来看，IT 部门似乎都已经山穷水尽、无以为继了。可以预见，在全球经济遭遇寒冬的时候，这种现象还会延续，甚至更加严重。

技术的不断发展给我们带来了曙光。虚拟化技术在这时走到了台前，这种诞生之初只能用于测试和开发的技术经过多年的发展，已经不断完善成熟，并成为了一个热门话题。回头审视我们的数据中心，有多少服务器上强大的 CPU 和海量的存储处于闲置状态？如何更好地整合这些资源，从而更好地节约成本、空间和电力消耗，是摆在我们面前的一大难题。这些广泛的需求刺激着软硬件制造商，他们不断提出新的解决方案，软件巨头们纷纷抛出自己的虚拟化产品，其中 VMware 公司的产品牢牢占据着市场的领导地位，而 Intel、AMD、IBM、NEC、日立等硬件制造厂商也都在自己的产品中加入了对虚拟化的支持。

虚拟化在技术上已经成熟，成功的案例也不胜枚举，但是，成功实施虚拟化并不是一蹴而就的，尽管相对于物理环境，虚拟化成功地简化了服务器部署、备份等工作，从而使系统管理员可以更加轻松地制定业务持续性计划、部署各种关键应用，但是从一开始，这一工作就需要精心的规划、认真的准备、审慎的实施以及持续有效的管理。毕竟，大型数据中心是现代企业的核心，任何错误的举动都可能给整个公司带来严重的后果。

本书是两位顶级虚拟化专家的力作，Eric Maillé 和 René-François Mennecier 作为 EMC 的高级专家，不仅十分熟悉 VMware 产品的组成、特性和优势，而且有着丰富的虚拟化项目实施管理经验，他们在书中详尽地介绍了 VMware 产品的各种特色，以及最新版本 vSphere 5 在虚拟化数据中心构建中提供的各种功能以及丰富的工具集，由浅入深地介绍了虚拟化数据中心中服务器、存储、网络、高可用性、容错直至灾难恢复计划等各个方面的规划、部署和管理方法，在本书的最后，还通过一个实际的项目，具体介绍了物理数据中心向虚拟化数据中心迁移的规划过程。

对于所有志在摆脱前文所述的财务和技术困境的 IT 经理来说，本书是不可多得的指南。在翻译本书的过程中，译者也受到了很大的启发，回想过去多年管理数据中心的经历，更希

望将本书介绍给同行们，使大家能够尽快地从物理环境管理的“泥沼”中走出来，建设一个更加可靠、易于管理、节约成本，也更有利于环境的新型数据中心。

本书的翻译工作主要由姚军完成，徐锋、陈志勇、刘建林、白龙、方翊、陈霞、林耀成等也为翻译工作做出了贡献。由于译者水平所限，书中难免出现一些错误，请广大读者多加批评指正，在此也感谢机械工业出版社的编辑们对翻译工作的大力支持。

译　者

序

数据中心（datacenter）的管理是当今IT部门面临的问题和挑战的缩影，它令人左右为难：要在逐年缩减的预算内提供更多的服务。作为关键的业务生态系统，数据中心变得越来越复杂，它的各个组成部分更加盘根错节。在这样的复杂度之下，企业计算还必须保持灵活性，并且必须应对为公司带来收入的业务需求。

与此同时，生成的数据量也在不断地增加。据估计，从现在开始的10年内生成的信息总量将是目前信息总量的44倍，要管理这样的数据流，我们需要10倍数量的服务器。

在这些挑战之外，法律上对业务持续性和数据保护计划的强制要求也加重了IT团队的负担。由于这些额外的约束，许多数据中心受到空间不足的影响，在业务线需要新机器来满足新需求的时候会导致严重的问题。能源是另一个关注点，因为数据中心如果管理不善，可能会达到供电能力的上限。有些公司发现自己无法增加数据中心的电量，必须寻求替代的解决方案。

我们也必须考虑自己：用户。我们有许多种通信方法（平板电脑、PC、智能手机），我们想要并且期待能够立即访问到信息（社会化网络、短消息、互联网）。很难理解，在我们的公司里为什么必须等待几天才能访问（有时甚至无法访问）在企业环境之外可以很快（大部分是免费的）访问的服务。

对于以一般公众为目标的服务（如Facebook），了解用户开立一个账户然后开始用标准技术使用这些服务有多么容易是很重要的，这些标准的技术和传统企业使用的技术没有任何不同。在日常使用中所不同的是这些技术的使用方式——完成处理的方式和采用自动化方法的程度。只有完全理解了这些相关的整体知识，加以协调并使其自动化，我们才能够为处理数据的新方式打下基础。在这种背景下，我们还必须理解，目前在IT部门中使用的常规方法必须改进，信息系统的变革是很有必要的。

服务器虚拟化是这种变革的关键，并且通常被作为应对当前挑战的基本解决方案。它能够建立一个高效的技术平台，以支持业务需求，为公司内部用户提供服务，同时降低成本和电力消耗。这种技术与网络和存储虚拟化相结合，形成了新一代的数据中心。加上自动化和自助服务，就构成了“IT即服务”（IT as a service，也称为云）的基础。

VMware vSphere 5服务器虚拟化解决方案是推进这种变革的核心组件。它提供了一个敏捷、灵活、可伸缩的环境，能够迁移到云计算服务。云计算的确是IT行业将来最大的机遇，利用云计算，最终用户能够访问一系列完全虚拟化的、基于数据中心的服务，而没有必要经历传统技术中十分棘手的部署过程。

为此，我们必须掌握服务器虚拟化技术及其各个组成部分。这些组件最终决定了这个技术平台的稳定性、灵活性和可伸缩性，对这一点的理解也是绝对必需的。技术上的复杂性可能造成难以解决的选择问题。这就是为什么要理解这些组件的相互作用，以及如何在数据中心里以优化的方式使用 VMware vSphere 5。

在本书中，Eric Maillé 和 René-François Mennecier 对 VMware vSphere 5 的运行机制进行了精要的解释，同时介绍了在数据中心生产环境限制下的最佳实践。本书的讲解透彻而生动，提供了许多例子和作者根据专业经历提出的建议。

我强烈推荐本书，它正是所有对这一复杂而引人入胜的主题感兴趣的读者所寻求的实用指南。

——Chad Sakac

EMC 高级副总裁兼全球系统工程经理，vExpert 和 vSpecialist 专家、virtual Geek（http://virtualgeek.typepad.com）的创始人和作者

前　言

我们热衷于虚拟化已经很多年了，对这一领域发生的任何事件都非常关注。我们又一次被 VMware 的最新版本及其提供的可能性深深打动。通过 vSphere 5，VMware 再次展现了创新的能力，这一解决方案从可用性、性能和使用灵活性上都达到了极限。但是这个快速发展的领域以及该软件繁多的功能可能令人迷惑。因此，对于我们来说，很有必要以书籍的形式对理解这种技术所需的信息进行综合，说明在数据中心中有效利用这一解决方案的方法。

我们希望奉献一本全面的书籍，能够涵盖我们认为最重要的主题，目标是利用我们在所服务公司的日常工作中得到的经验，提炼重要的信息和建议，帮助读者对 vSphere 5 有总体的了解。

但是我们仍然必须做出选择。我们决定不采用过于技术化的风格，而是旨在编写一本大部分读者（而不是少数专家）都能理解的书籍。

第 1 章是专门为了帮助读者理解 VMware vSphere 5 的功能而写。后续的章节解释这一技术与数据中心各个要素（服务器、存储、网络）之间的联系，以及备份和服务持续性（本地和远程）的各种方案，其中包括介绍 vSphere 5 安装和操作管理的一章（第 7 章）。

第 8 章介绍在大型公司的数据中心，按照管理层清晰定义的目标，迁移到虚拟化环境的一个实用案例。我们详细地介绍不同阶段中的步骤、建立的目标架构、实施这一项目的方法以及获得的好处。这一章很实用，强调了这类项目的难点和风险点，并概要介绍了实施成功、平稳迁移的一种方法。

本书主要面对的读者是负责信息系统基础架构项目的人员：系统 / 网络 / 存储管理员、项目经理、咨询师、架构师、销售专员、宣讲师等。对于为 IT 生涯做准备，需要理解虚拟化的学生和未来的工程师来说，本书也是很有吸引力的。

免责声明

尽管我们属于 EMC 集团，但是对于所讨论的技术，我们尽力保持客观和中立。本书旨在真实地表达我们从实践中得到的真知灼见。我们的目标是提供自认为必不可少的信息，帮助读者根据对事实的了解和他们的需求做出选择，而不偏向于某种特定的技术。我们所表达的是自己的观点，EMC 集团和书中所述技术的利益相关方都无须为这些观点负责。

致谢

我们首先要感谢那些在本书的写作过程中帮助过我们的朋友，这些朋友使我们在专业领域中更进一步。你们的帮助是无价的，特别要感谢以下各位：

Denis Jannot、Lothaire Lesaffre、Cody Hosterman、Philippe Audurieau、Rolland Zanzucchi、Jérôme Fontanel、Hervé Gilloury、Jean-Baptiste Grandvallet、Hervé Oliny、Emmanuel Bernard、Philippe Rolland、Mickaël Tissandier、Emmanuel Simon、Nicolas Viale、Jean-Louis Caire、Jean Strehaiano、Gwenola Trystram、Philippe Chéron 和 Bernard Salvan。

十分感谢 VMware 法国公司的技术主管 Sylvain Siou 的贡献和忠告，感谢 Richard Viard 在 VMware 相关的各个领域中展现出来的技术造诣。还要感谢我们的虚拟化生态系统网络：Eric Sloof、Iwan Rahabok、Ilann Valet、Olivier Parcollet、Julien Mousqueton、Jérémie Brison、Raphaël Schitz、Vladan Seget 和 Damien Peschet。我们还要感谢技术审校者 Thomas Keegan 和文稿编辑 Richard Carey 及 Keith Cline，感谢他们帮助我们将本书介绍给以英语为母语的 VMware 社区。

特别感谢 Chad Sakac 在百忙之中为我们的书撰写序言。

我们希望听到你的意见

作为本书的读者，你是我们最重要的批评者和评论者。我们珍视你的意见，希望知道我们在哪些方面做得好，哪些方面需要改进，你希望我们出版哪些领域的书籍，以及你愿意与我们分享的真知灼见。

作为 Pearson 的联合出版商，我们欢迎你提出意见。你可以直接发送电子邮件或者信件，让我们知道你是否喜欢这本书——以及帮助我们改进工作的建议。

请注意，我无法帮助你解决本书相关技术领域的问题。但是，我们有一个用户服务组，关于本书的具体技术问题将会转发给他们。

当你撰写邮件的时候，请务必包含本书的书名和作者，以及你的姓名、电子邮件地址和电话号码。我将认真地查阅你的意见并与本书的作者和编辑分享。

电子邮件地址：VmwarePress@vmware.com

邮寄地址：David Dusthimer
Associate Publisher
Pearson
800 East 96th Street
Indianapolis，IN 46240 USA

目　录

第1章
从服务器虚拟化到云计算

在过去几年中，由于技术的变化和快速增长的服务及资源需求，企业信息系统有了很大的改变。计算机资源的需求从未达到如此的高度。利用新的通信手段（如智能手机、社会化网络和即时消息），用户希望从任何地方，在任何时候都能立即访问信息。而且，公司管理层需要高水平的服务，在有限的预算条件下支持企业及业务需求。

1.1 虚拟化：IT 变革的核心

IT 经理面临对信息系统进行变革和现代化改造的巨大压力。为了满足日益增长的需求，同时控制成本，就必须实施服务器虚拟化。服务器虚拟化已经成为现代计算的基础，为云计算（cloud computing）铺平了道路。我们已经进入了一个新计算时代，曾经用于测试和开发的技术现在已经用在关键应用上。

在 2012 年，VMware vSphere 5 是企业环境中部署得最多的服务器虚拟化解决方案。它已经成为生产环境中的标准。Microsoft、Citrix 和 Oracle 也都提供了替代的方案，但是从硬件和软件上说，它们与数据中心的各个组成部分的集成度和兼容性都无法和 VMware vSphere 5 相比。

为了从 VMware 的许多优势中获益，IT 团队必须理解这种技术和数据中心各个组成部分之间的相互作用。这些团队必须发展跨学科技能，并采纳利用该技术潜能和加速与数据中心集成所必需的方法学。虚拟化环境主要对系统管理员和网络、存储及安全团队产生影响。从早期就让这些成员加入，他们就能够支持这一项目而不是阻碍它的实施，他们必须具备自我培训的能力和改变传统方法的动力，这样的改变需要培训、网上信息和 Web 研讨会的支持。在项目的一些阶段，请求服务供应商分享他们的知识是很有帮助的。

这是重新思考工作流程和规程的一个机会。可以规划一次概念验证（Proof Of Concept，POC）来确定解决方案的功能特征，但是考虑到这一技术已经在各种活动和各类业务中证明了自己的成功，这一过程现在似乎已经没有意义。

管理团队也必须了解这一点，他们必须意识到与机构和工作流程相关的变化，以及项目成功所需要的投资。

有了这些支持，机构各个层面上的每个人都应该意识到与虚拟化相关的问题和深层次变化，以及公司由此得到的利益。

1.1.1 服务器虚拟化

服务器虚拟化是掩盖硬件设备物理资源的一个抽象层（abs traction layer），为系统提供与实际形式不同的资源。从本质上说，硬件资源是有限的。服务器虚拟化突破了这些限制，开启了一个具有潜力的新世界。服务器虚拟化并不是一个新概念，它是 IBM 在 20 世纪 60 年代为其大型主机系统创造的概念。在很长的时间内，它似乎是 x86 环境所无法采用的技术，但是 VMware 在 1998 年取得了成功。

VMware 技术将整个 x86 服务器虚拟化为一个逻辑实体——虚拟机（virtual machine，VM）。如图 1-1 所示，低级别的虚拟化层次（称为虚拟化管理器 hypervisor）可以在单个物理机器上运行多个操作系统。

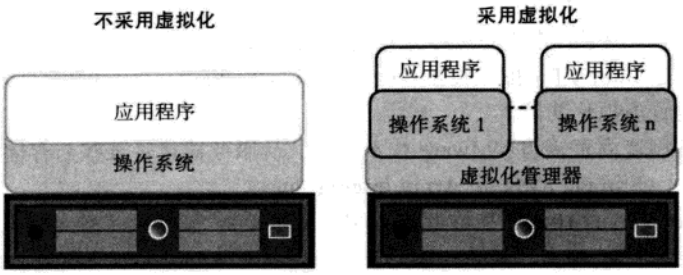


图 1-1 没有虚拟化，一台物理机器上只能运行一个操作系统，而虚拟化允许在一台机器上同时使用多个操作系统

虚拟化管理器使得操作系统独立于所处的硬件，为在单台机器上整合各种基于服务器的服务带来了许多可能性。

1.1.2 采用服务器虚拟化的要素

后面几节将介绍在 IT 环境中采用服务器虚拟化应该考虑的因素。

1. 资源的浪费

在实体环境中，据估计有 80% 的数据中心服务器平均使用率低于 10%。而从占地面积、耗能和散热方面来说，数据中心都将达到极限。这种浪费的根源在于许多公司都在运行单个应用程序的 x86 服务器上投入了大量资金。这使得物理服务器数量剧增，考虑到它们的实际使用，运行成本简直是天文数字。据估计，管理员花费大约 70% 的时间支持或者维护无法给公司带来任何价值的运营活动。

注意： 与购买服务器的成本相比，管理、执行等间接成本和电力消耗成本已经成为巨额的费用，最多的时候，它们高达服务器初始成本的 4 倍。

维护服务器基础架构运行条件的高成本浪费了企业的资金，造成运营效率低下，不利于创新和管理新项目的能力。

与这种浪费对立的是，管理层向 IT 经理施加压力，要求他们在保证一定服务水平的同时，在相同（甚至更少）的 IT 预算下应对不断增加的请求量。虚拟化是满足这类需求，同时降低成本并且保持公司 IT 系统最新的一种方法。

2. 服务器虚拟化中的技术

服务器技术在过去几年有了很大的进步，已经具备了多核 64 位处理器和更大的内存管理能力。现在，在一个能够容纳几十个操作系统的服务器上安装单个操作系统显然是没有道理的。

虚拟化技术最大限度地利用了多核处理器，而且可以达到很高的整合水平。现在的服务器在性能上达到 4 年前的 10 至 12 倍。这大大地增强了与 AIX、SUN 和 HP-UX 等 UNIX 服务器竞争的能力。除了一些特殊的配置以外，SAP 或者 Oracle 等策略性应用程序在虚拟化的

x86 环境中都能工作得很不错。

硬件中也已经集成了许多支持原生虚拟化的功能特性：

- 处理器通过 Intel VT（虚拟技术 Virtualization Technology）和 AMD V（Virtulization）拥有了内置虚拟化功能。
- 存储设备制造商直接与 VMware 技术接口，从服务器接管某些与存储相关的任务（称作 VAAI：vStorage APIs for Array Integration）。
- 诸如 Cisco Nexus 1000V 和最近的 IBM DVS5000V 等一些网络交换机能够简化这种环境中的网络管理。

从这几个例子可以看出，硬件和软件公司已经开始开发完全集成和利用虚拟化技术潜力的产品。

1.1.3 虚拟化环境规范

考虑了实施虚拟化环境的要素之后，你应该熟悉这种环境的基本规范。

1. 更改数据中心模型

用封装在文件中的虚拟实体代替物理服务器，这改变了数据中心当前的模型。在虚拟化之前，使用的是由许多小规模物理服务器组成的分布式模型。通过虚拟化，这种模型变成围绕单个服务器站点的集中化整合模型。存储成为支柱，为了部署虚拟机器，它必须提供高性能和数据安全解决方案。网络也起着关键的作用，因为一旦选择云计算模型，公司就完全依赖网络和互联网连通性。因此，网络带宽必须足够。

这一变革迫使企业重新定义目前的基础架构。

2. 虚拟机

在虚拟环境中，管理员管理的是虚拟机（Virtual Machine，VM）。VM 包含了一个物理服务器的全部内容：操作系统（称作客户操作系统）、应用程序和数据。

至于基础架构，VM 与物理服务器完全相同，不需要进行应用移植。管理员可以很精确地微调 VM 的大小。这种配置粒度使他们能够为 VM 提供所需的资源。

VM 相互完全隔离（操作系统、注册表、应用程序和数据）。如果一个 VM 遭到病毒感染或者操作系统崩溃，不会危及其他 VM。目前，VM 之间的屏障还从未打破过。

VM 的所有组件都包含在文件中，这称为封装（encapsulation）。封装简化了备份、复制、灾难恢复计划过程以及新环境的迁移，带来了很好的使用灵活性。

而且，VM 完全独立于它们所在的硬件。在传统的物理环境中，操作系统与所安装的硬件紧密联系。（这种环境很庞大，要求每个服务器都需要有一个安装了特定驱动程序的主映像。）而在虚拟化环境中，虚拟化层总是为 VM 提供相同的虚拟硬件（图形卡、SCSI 卡等），所以可以创建相同的 VM 而不需要考虑底层硬件。这减少了专用于每种类型硬件的多种主映像的创建，使大规模部署变得更加简单。

注意： 许多公司在老的服务器上部署关键应用程序，对这些服务器的升级被无限延后的原因就是迁移的复杂性。利用虚拟化，这些服务器可以简单地转换为虚拟服务器，不需要重新安装操作系统或者应用程序。

---

3. 灵活性

灵活性是虚拟化环境的主要特征之一。VM 可以从一台物理服务器以完全透明的方式转移到另一台服务器上。这减少了服务中断并简化了管理员在执行计划维护或者向新平台迁移时的日常管理。将负载分配给不繁忙的服务器也很容易。

4. 即时部署

虚拟化革新了传统服务器管理方法和服务器的运行。利用即时部署能力，新服务器能够在几分钟内投入使用，而在传统的物理环境中这可能要花费数周。这完全改变了时限性，使公司能够很快地适应与业务相关的变革，例如，合并、收购和新服务或者新计划的实施。特殊的需求能够很快满足，也能改进用户服务。

5. 将资源集中到群集

ESXi 主机服务器能够聚集到称为*群集*（cluster）的实体中，这时的虚拟环境可以作为一个整体而不是单独的单元来管理。因为实体资源在群集中共享，可以获得高级的高可用性功能。在高活动率的时候，负载可以自动地在群集中所有服务器上共享。这简化了管理员的工作并保证了应用程序的服务水平。

6. 服务质量

为了保证每个 VM 都能访问所需的资源，可以实施*服务质量*（Quality of Service, QoS）。QoS 可以在多个级别上设置——VM、ESXi 主机服务器或者群集。

1.1.4 虚拟化的好处

服务器虚拟化提供了无法抗拒的直接好处，主要有：

- 成本降低：成本的降低是企业的主要关注点。虚拟化是降低 CAPEX（资金支出）和 OPEX（运营支出）的方法之一。数据中心内的服务器越少，意味着投资越低、维护成本越低、占地面积越小，电力消耗和散热的成本也越低。降低电力消耗和散热的成本是最根本的好处，因为这方面的费用最多可达基础架构硬件成本的 3 倍。而且，电费可能逐年增加，限制甚至降低电力消耗也就更加关键。

如图 1-2 所示，服务器虚拟化是降低数据中心电力消耗的一种方法。

2000 ~ 2010 年间数据中心的电力消耗增长曲线在 2006 ~ 2007 年开始降低。在 2000 ~ 2005 年间，电力消耗翻了一番（从 300 亿千瓦时 / 年增长到 600 亿千瓦时 / 年），而在 2005 ~ 2010 年间，增长率只有 56%，这种情况是由三个因素造成的：经济危机减缓了投资；数据中心的电力消耗得到了更好的控制；大规模采用虚拟化，从根本上减少了服务器的数量。

- **服务等级协议（SLA）的改进**：利用 vSphere 5 的先进功能，可以非常简单地实现高可用性解决方案，同时又将管理员从耗时的群集解决方案中解放出来。而且，实现活动恢复计划和备份操作也比原来简单得多，这解决了过度狭窄的备份窗口以及相关的时间约束问题。
- **灵活性**：虚拟化能够很轻松地适应增长的需求，不管这些需求是业务需求、与合并收购相关的企业需求还是用户需求。灵活性增强了反应能力和创新能力，因为任何项目必需的专用基础架构可以简单地建立，而不需要大量的投资。

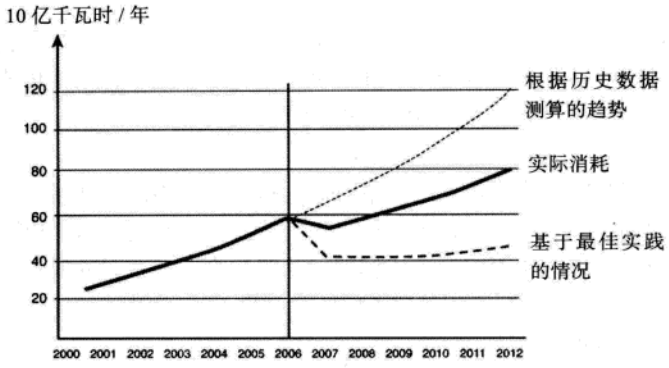


图 1-2 通过服务器虚拟化，数据中心电力消耗减少的情况（以上图形摘自 2011 年 8 月 Jonathan G. Koomey（斯坦福大学）的研究

- **操作效率**：虚拟化的固有功能极大地简化了程序化维护操作、迁移的各个阶段和软件更新，而在物理环境中这些操作都极其耗费资源。
- **自动化**：这使管理员可以解放出来，从事其他工作，并允许用户独立地使用资源，而不需要来自 IT 团队的帮助。自动化减少了不提供附加值的重复工作，对于成本的降低起到关键的作用。成本降低和信息系统自动化水平存在直接的联系。
- **用户**：用户得到的益处最多，因为他们将 IT 当作一个服务消费，而不需要通过 IT 团队。
- **标准化**：应用统一标准的能力更好地满足法规的要求，为日常运营提供更有效的过程。

但是，认为实现这一环境就能解决所有的问题也是个错误。在虚拟化展现出所有潜力之前，还需要经历多个阶段。

1.2 虚拟化的各个阶段

我们来看看图 1-3 中虚拟化的 3 个阶段，这几个阶段能够带来成功的云计算服务。

1.2.1 第 1 阶段：IT 合理化

用新一代服务器替换高耗能的旧服务器，对基础架构进行合理化（rationalize）。这降低了数据中心的电力消耗和占用空间，其中的某些数据中心已经达到使用寿命。而且，利用 vSphere 的分布式电源管理（Distributed Power Management，DPM），服务器可以根据群集资源需求关闭或者重启，从而优化电源使用。

1. 好处

虚拟化过去常常用在非关键性测试和开发环境中。这样，开发人员能够根据需要使用多种操作系统和应用程序版本。管理员可以在不需要网络或者存储团队的情况下测试生产环境。IT 经理在这一阶段中可以虚拟化直接由他们控制的服务器（例如，基础架构服务器），域控制器、打印服务器和 Web 服务器很容易进行虚拟化。在这一阶段中，基础架构虚拟化渗透率可以达到 0% ~ 30%。

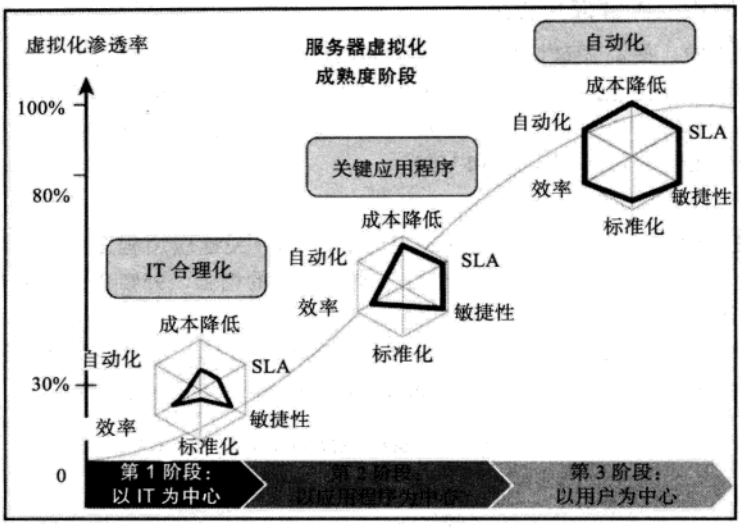


图 1-3　第 1 阶段以 IT 服务直接控制的基础架构服务为中心，第 2 阶段以业务关键应用程序为中心，第 3 阶段则关注提供给用户的服务

这一阶段能让各个团队熟悉新环境。利用即时部署（instant provisioning），这些团队能更好地对新需求做出反应，因为新的服务器可以在几分钟内做好准备，而不需要像物理环境那样花费几周，这也同时改进了对用户的服务。

使用快照（snapshot）还可以简单地更新应用程序，因为如果应用程序出现故障，可以快速地回溯。

2. 挑战

一般来说，期望从虚拟化技术中得到的成本节约在合理化阶段中还不能实现。确实，在这一阶段必须购买新服务器、存储设备和 VMware 许可证，还必须培训各个团队。这是一个投资阶段（investment phase）。在这一开始阶段之后，有些管理者得出结论：这种技术不能降低成本。这一点都不奇怪。实际的成本节约发生在下一阶段，那时这种技术已经得到完善并且自动化。之后，成本将得到可观的降低。

在合理化阶段，整合度很低（每个服务器 5 ~ 10 个 VM），所以性能问题不明显。大部分问题集中在存储空间。由于部署 VM 很简单，出现了一种称为虚拟机蔓延（VM sprawl）的现象，这种现象应该得到控制，否则会导致整个环境失控。为了保持控制能力，必须采用严格的 VM 管理规则。下面是规则的一些例子：

- 自动关闭（删除）在规定时间内（几个星期）不活跃的 VM。
- 必须识别连续数天处于休眠状态（启动但是没有活动）的 VM，以便了解其不活跃的原因。如果这个 VM 已经被人遗忘，应该关闭并最终删除。
- 必须识别生产 VM 并且使其与其他 VM 分开。例如，可以创建文件夹来分类 VM。

1.2.2 第 2 阶段：关键应用程序

在第 2 阶段中，提高关键应用程序的服务水平并显著改进操作效率。虚拟化现在可以大规模地部署到信息系统中。工作重心转移，第 1 阶段的中心是基础架构，而现在的中心是应用程序。关键应用程序的例子包括 Microsoft Exchange、Lotus Notes、SAP、Oracle、Microsoft SQL、SharePoint、Citrix 和 VDI。

1. 好处

目前，除了少数特殊的应用程序负载之外，大部分关键应用程序在 VMware 环境中都可以无限制地运行。在 vSphere 5 中，得益于 VM 性能的增强，95% 的应用程序负载可以在 VM 中安全地运行。

如图 1-4 所示，2010 年 1 月 ~ 2011 年 4 月间的观察证明，这些应用程序在虚拟化应用程序中运行得很好。注意 Microsoft SharePoint 在 VMware 下的高采用率（67%）。

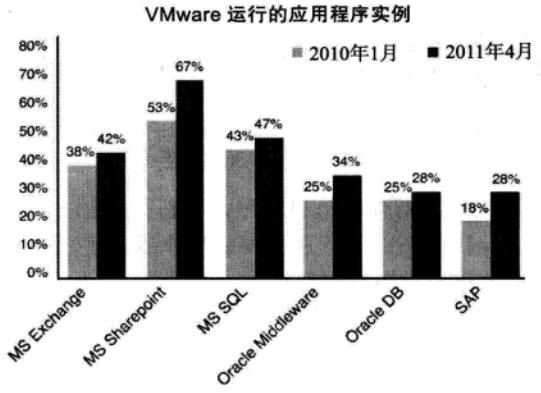


图 1-4 大部分传统关键应用程序在 VMware 中的运行情况

注：研究数据涉及 VMware 的客户。

2. 利用复制快速部署应用程序

IT 管理员虚拟化 Oracle 或者 SAP 的原因之一是：这是复制应用程序的最简单方法。在物理服务器中，复制应用程序的过程是低效率的根源之一。IT 管理员必须配置每个应用层，包括硬件、操作系统和应用程序，这很费时且可能造成潜在的配置错误。虚拟化为快速部署创建优化的“黄金映像”，简化了复制过程。而且，应用程序还可以打包成一个 vApp，包含多个预先配置了不同应用层（例如，Web、应用程序和数据库）的虚拟机。这个包可以按照需求快速部署到生产（也可以是测试和生产前）用的基础架构中。

3. 简化高可用性的实现

在物理环境中，高可用性是很令人头痛的。每个应用程序需要特殊的高可用性方法，这就需要昂贵的许可证、专用的备用基础架构以及配置和管理每个解决方案（Oracle RAC 和 Data Guard、Microsoft 群 集、Data Base Mirroring [SQL Server]、Database Availability Group[DAG]

Exchange Server…）的熟练人员。使用 vSphere 技术的标准化方法可以替代这种昂贵的方法，也没有其他技术那么复杂。有些公司利用虚拟化方法，用 vSphere HA、vSphere FT 或者 App-Aware HA（这是一个允许用户插入来自 Symantec 或者 Neverfail 的第三方 App-Aware 产品的 API）等集成的高可用性解决方案代替费时的群集解决方案。而且，因为 vMotion 和 Storage vMotion 的使用，计划维护操作不再需要中断所涉及的服务，从而提升了服务水平。

4. 保证预留资源

在这一阶段中，虚拟化渗透率达到 30% ~ 80%。

5. 挑战

因为与关键应用程序相关，这一阶段非常敏感。根据 2011 年 7 月国际数据公司（IDC）的研究，许多公司发现图 1-5 中列出的原因降低了它们的部署速度。

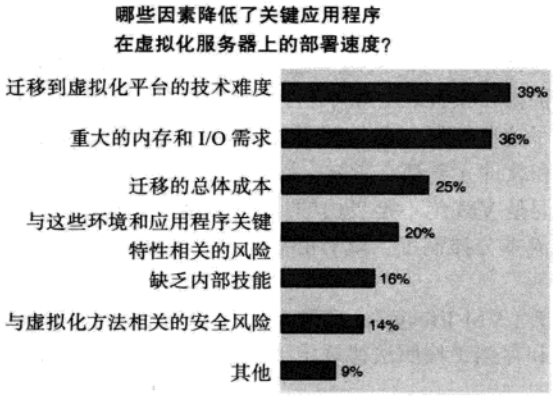


图 1-5　可能影响应用程序部署的因素

这一阶段通常需要能够有效地克服技术障碍的方法。这些问题与性能、备份和灾难恢复计划（Disaster Recovery Plan，DRP）相关。这一阶段的挑战在于实现足以支持应用程序负载的目标架构。现有的方法必须加以修改，因为它们通常不适合于虚拟化环境。

有些公司在这种改变上可能面临强大的内部阻力。因此，关键是让所有 IT 团队尽早加入到项目中来，并在整个转换的过程中给予他们支持。

存储架构也是必不可少的，因为它们是虚拟机的宿主。由于数据量可能非常大，优化数据管理变得很重要。数据压缩、精简配置（thin-provision）和备份重复数据删除都很适合于这些环境。必须考虑虚拟化环境的动态特性，对现有架构进行修改。

对于管理员来说，获得多种技能以及新环境中的培训都是不可或缺的。他们的角色和任务必须演变，最终拥有广阔的视野，很好地理解虚拟技术各个组成部分与数据中心的相互作用。

1.2.3　第 3 阶段：自动化

在第 3 阶段中，利用自动化（automation）降低成本。自动化能够节约管理员的时间，使他们能将更多的精力投入到信息系统的更新和改进中。现在，70% 的维护和 30% 的创新

这一比例已经颠倒过来。自动化是管理增长和满足每年提出的新要求的关键。自动化拉近了业务需求和计算机资源的距离，不需要求助IT团队，支持创新并提供不受财务问题困扰启动新计划的可能性。确实，许多项目中止的原因都是IT预算的缺乏。

这一阶段还包括过程的工业化和为达到高整合水平的资源优化。这些工作能够证明前两个阶段的投资是正确的。

注意： 一个管理员在物理环境中通常能管理50台物理服务器，在虚拟环境中能管理多达200 ~ 300个VM，实施自动化之后可以管理多达数千个VM。

因为基础架构问题和应用服务水平在前两个阶段已经解决，考虑到过程的实施和部署自动化，这个阶段可以将重点放在用户需求上。利用服务目录，用户可以独立地消费IT资源而无需依赖大量IT建议或者工作。这就是所谓的IT即服务（IT as a service）。

1. 挑战

在第3阶段中，IT团队面对的挑战和过去不同。以前遇到的问题与技术或者基础架构问题相关，例如，维护操作条件、更新平台以及各类迁移（应用程序、硬件、操作系统等）。虚拟化解决方案减少了这些问题，但是公司现在必须面对组织和管理问题。服务器虚拟化对数据中心的所有硬件和软件元素都有影响，必须有合适的管理原则，确定具有决策权的人。

自动化很方便，但是VM绝不是自由的。采用严格的规则并定义一个VM管理策略绝对有必要。如果建立了成本分摊制度，就有必要定义制度的标准（例如，按照消耗量或者VM数量）。

VM生命周期管理（VM lifecycle management）的各个方面在确定哪个VM可以删除以及哪个VM必须保留和存档（按照法律规定）时，都是必须考虑的。

2. 迈向云计算

虚拟化的第3阶段（也就是最后一个阶段）也是引入云计算的一个阶段。云计算是一个广泛的主题，需要单独的一本书来介绍，所以这里不作详细说明。

云计算的定义多种多样，下面是来自NIST（美国国家标准和技术学会）的一个有趣的定义：

云计算是一种可以在最小的管理投入或者服务提供商交互的情况下快速部署和发布的计算模型，能够提供对一个可配置计算资源池（例如，网络、服务器、存储、应用和服务）的普遍、方便、按需的网络访问。

实际上，云计算是利用计算资源的一种新方式，目标是通过提供全自动计算服务，降低IT基础架构的复杂度。

在这个模型中，信息系统从一个成本中心变成一个利润中心，可用的资源可以通过订阅或者收费服务重新计算。用户通过提供服务目录的一个网关来消费计算资源。从技术上和财务上讲，虚拟化技术的出现使得人们可以不考虑物理组件，从而使这种新模型成为可能。

云计算模型的主要好处是激活的速度和提供的敏捷性，它能快速地做出反应，满足公司的业务需求。使用可用的服务，几乎可以立刻创建各种项目，这改变了日常的工作方法。

这种模型的主要问题之一涉及对数据所处位置和所有权的理解。这是云服务供应商的问题，它们必须解释数据所在的位置，以及规范是否能够有效地服从于购买此类服务的国家或者

团体的法律和规定。数据流向和所有权必须以容易理解的方式编写文档并接受各方的验证。

在企业内部创建的云称为私有云（private cloud）。由外部供应商（服务提供商，service provider）提供的服务称为公共云（public cloud）。在发生私有云无法满足的活动激增时，可以在规定时间内分配来自公共云的附加资源，这称作混合云（hybrid cloud）。

公共云是外部提供商（服务提供商）向最终用户提供的服务。提供商和客户在最低性能或者交付时间上达成一致，这称作服务水平协议（Service-Level Agreement，SLA）。客户不再需要管理内部 IT 基础架构，可以将重点放在核心业务上，因此这种模型最适合于中小型企业（Small and Medium Business，SMB），在这种企业中 IT 通常不是核心。服务水平（通常是每周 7 天每天 24 小时具有 DPR、备份等）和安全水平提供的用户体验好于大部分 SMB 目前的水平，主要是因为内部能力和财务方法的相关问题。运营成本明显低于内部解决方案，而且用户总是能够访问最新的软件版本。加之，计算资源变得灵活，它们能够适应活动量的波动和季节性特征。

云计算存在多种模型，包括：

- **软件即服务**（Software as a Service，SaaS）：用户对底层基础架构没有控制权。他们直接使用在云中找到的应用程序。例如：Mozy、Zimbra、Gmail、Google Apps、Microsoft Office 365、Cisco WebEx）。
- **平台即服务**（Platform as a Service，PaaS）：用户用编程语言或者服务提供商提供的工具，部署他们创建或者购买的应用程序。他们不控制底层基础架构。例如，SpringSource、Microsoft Azure。
- **基础架构即服务**（Infrastructure as a Service，IaaS）：用户自己部署处理器、内存、网络和存储等资源。他们保留对操作系统和应用程序的控制。例如，EMC、VMware、Cisco、Amazon Web Services、IBM、CSC、Verizon）。

1.3 虚拟化生态系统

下面几节提供虚拟化生态系统和各种解决方案的快速概览。

1.3.1 服务器虚拟化

必须区分裸机虚拟化产品和主服务器上的（称为基于主机的（host based）产品。服务器上基于主机的虚拟化应用可以用于测试，但是决不能用于生产。如果基于主机的版本投入生产，副作用是灾难性的，但是它们作为测试环境是很有趣的。

这类产品中著名的有：

- Microsoft Virtual Server 2005、Vitrual PC
- VMware server
- VMware Workstation、VMware Player、VMware ACE、VMware Fusion（Mac 版本）

在裸机（bare-metal）虚拟化应用程序中，虚拟化层直接安装在硬件上，安装从 CD-ROM 启动开始，与传统操作系统的安装方式相同。只有这些解决方案经过了优化，可以用于生产环境。它们和操作系统一样安装于硬件上。

下面是主要的裸机虚拟化解决方案：

- VMware vSphere
- Microsoft Hyper-V
- Citrix XenServer
- Oracle VM
- Red Hat KVM

尽管 Citrix XenServer、Oracle VM、Red Hat KVM 效果不错，但是市场关注的是 Microsoft 和 VMware 之间的争斗。VMware 在虚拟化市场上处于领先，在技术上也走在 Microsoft 的前面。

VMware 目前在大公司中确立了很高的地位，在这方面它处于准完全的统治地位。它在生产环境中达到了很高的成熟度，并且很好地与数据中心的各种软硬件元素集成，这是一个很大的优势。

Microsoft 提供 Hyper-V，这个解决方案很适合于当今的 SMB，但是无法为大公司提供 VMware 的所有高级功能。

1.3.2 桌面虚拟化

工作站虚拟化使每个用户远程登录到位于数据中心的一个虚拟机。远程访问也可以通过许多硬件解决方案实现，如传统 PC、便携电脑、低速终端甚至智能手机，不需要在客户端工作站上作任何配置。

下面是目前最有名的解决方案：

- VMware View
- VMware WSX（新的 VMware 项目：通过 HTML5 技术的 Windows 桌面）
- Citrix XenDesktop
- NEC Virtual PC Center
- Quest vWorkspace
- Systancia AppliDis Fusion
- Neocoretech NDV

对于便携电脑和台式机，客户端 – 虚拟化管理器类型的解决方案可以直接安装在现有硬件之上，这些解决方案可以让多个相互隔离的 VM 运行于单个 PC 上。有了这样的解决方案，用户可以使用多个不同且完全隔离的环境（例如，一个映像用于专业用途，另一个用于私人用途），并且可以通过集中的管理控制台进行控制。

现有的解决方案有：

- Citrix XenClient
- Virtual Computer NxTop（现在是 Citrix 的一部分）

1. 性能和容量规划

当在虚拟环境中使用关键应用时，具备用于监控性能和容量规划的工具是很重要的。

以虚拟环境为中心的主要软件工具包括：

- VMware vCenter Opterations Management Suite

- Quest vFoglight（已被 Dell 收购）
- Orsyp Sysload
- vKernel（被 Questr Software 公司收购，因此现在归属 Dell）
- Veeam One
- Xangati

2. P2V 转换工具

迁移项目需要使用将物理机器转换为虚拟机器的工具。这种迁移称作*物理－虚拟转换*（Physical to Virtual，P2V）。

下面是最著名的 P2V 工具：

- VMware vCenter Converter
- Vizioncore vConverter
- HP Server Migration Pack
- Plate Spin Migrate（Novell）

3. 备份

传统备份解决方案可用于虚拟环境。许多制造商提供了非常适合这类环境的解决方案。下面是最流行的产品：

- EMC Avmar
- Symantec NetBackup
- Veeam Backup
- IBM Tivoli Storage Manager TSM
- VMware Data Recovery
- Quest vRanger Pro
- Comm Vault
- Acronis vmProtect

这些备份解决方案一般不需要代理，它们与 VMware 提供的 API 接口，提供无缝集成。某些解决方案允许在虚拟机中进行文件级恢复并提供重复数据删除。

4. 云套件

几年前新推出的刀片服务器管理通过重新组合同一个机架上的多台服务器，简化了它们与数据中心的集成（减少电缆、减少占地面积、减少电力消耗等）。云套件的想法与此类似，但是打包了硬件和软件级别的解决方案，可用于整个数据中心。整个套件可以立即投入使用，集成了部署和调整工具，目标是减少虚拟化环境在工业化平台上的交付时间。所有主要的 IT 厂商都提供了基于这个模型的产品。

私有云方面的产品包括：

- Vblock（EMC/VMware/Cisco 联盟）
- FlexPod（NetApp/VMware/Cisco 联盟）
- IBM Cloudburst

- ❑ Oracle Exadata/Exalogic
- ❑ Dell vStart
- ❑ HP CloudSystem

公共云方面的产品包括：

- ❑ Google
- ❑ Amazon
- ❑ Microsoft
- ❑ Salesforce.com

5. 云数据存储

对于这种新服务，任何内容都能自动地存储于云中，所以数据可以从任何地方、任何类型的终端（例如，PC、智能手机、平板电脑）访问。这些解决方案广泛流行。与改动发生之后，数据立刻在不同的终端之间自动同步（这也加强了数据的安全）。这使得多个用户需要修改相同文档的工作更容易协同进行。

注意： *我们用 Dropbox 服务编写本书，用这种方法很容易共享我们的文件。*

下面是此类著名的解决方案：

- ❑ Dropbox
- ❑ Oxygen
- ❑ VMware Project Octopus
- ❑ Apple iCloud
- ❑ Ovh hubiC
- ❑ EMC Syncplicity

1.4 美好的明天

正如你所看到的，当今的数据中心虚拟化环境有着庞大的生态系统。这些解决方案支持计算的根本变革，即用更少的资源提供更高的效率。但是，仅靠虚拟化并不能解决 IT 人员面临的所有问题，实际上，如果没有合适的管理，虚拟化可能造成一些问题。因此，在规划虚拟化策略时，重要的是建立 VM 生命周期，恰当地管理资源分配，减少虚拟化蔓延的现象。从一开始就与各个方面的 IT 人员协作，包括存储、网络和安全小组，你的虚拟化项目就能够成功。

今天的挑战与前 10 年不同，因为需求也不一样了。即时新闻和社会化网络生成了难以管理的海量数据。因此，必须为转向云计算服务铺平道路，构建新一代的数据中心，迎接新的挑战。

第2章
vSphere 5的演变和架构组件

2.1 VMware概述

2.2 vSphere 5许可证

2.3 vSphere 5的新增功能

2.4 现有功能

2.5 单独销售的软件

2.6 vSphere 5技术架构

2.7 安全性

2.8 发展的解决方案

在深入研究 VMware 新的 vSphere 5 产品的软件组件和架构之前，我们先来看看这家公司及其虚拟化产品的演变过程。

2.1 VMware 概述

VMware 由如下几位知名人士于 1998 年创立：

Diane Greene，总裁兼总经理（1998 ~ 2008 年）

Mendel Rosenblum 博士，首席科学家（1998 ~ 2008 年）

Scott Devine，主任工程师（1998 年起）

Edward Wang 博士，主任工程师（1998 ~ 2009 年）

Edouard Bugnion，首席架构师（1998 ~ 2004 年）

2004 年，EMC 以 6.25 亿美元的价格收购了 VMware。在收购时，该公司的销售额已经达到约 1 亿美元，拥有 370 名员工。2011 年，销售额接近 37 亿美元，到 2012 年，公司的雇员已经超过 9000 人。

Pat Gelsinger 于 2012 年 9 月就任 VMware 的 CEO（此前，Paul Maritz 从 2008 年 7 月起担任这一职务）。

2.1.1 VMware 产品线

在过去几年，VMware 的产品线有了很大的扩展。早年，该公司主要提供虚拟化管理器，而现在的重心已经扩展到云计算服务、管理工具、协作工具和应用程序。本书的焦点是 vSphere 5 的数据中心虚拟化平台，并将详细介绍作为数据中心管理基础的组件，书中提及的其他组件仅是为了提供信息。图 2-1 提供了 VMware 目前产品的一个概览。

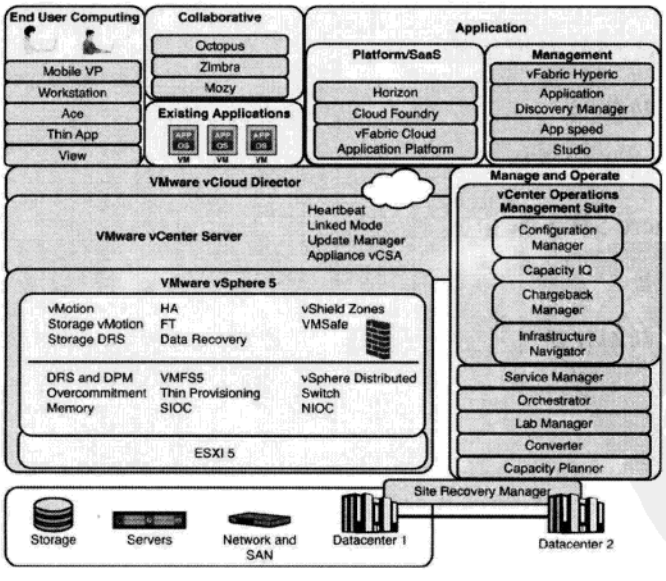


图 2-1 VMware 产品线

VMware 的策略明显是面向云计算解决方案，包括了图 2-2 中所示的元素。

2.1.2 VMware 的发展

VMware ESX 的第一个版本于 2001 年发布，vSphere 5 是其第 5 代产品。每个新版本带来的革新和技术跨越都值得回顾。图 2-3 提供了这些革新的时间轴。

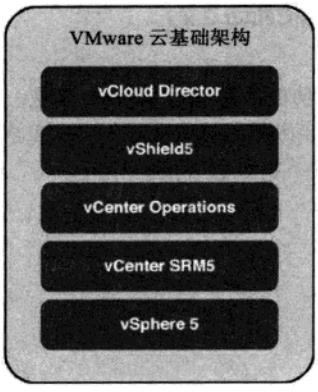


图 2-2 VMware 的云基础架构元素

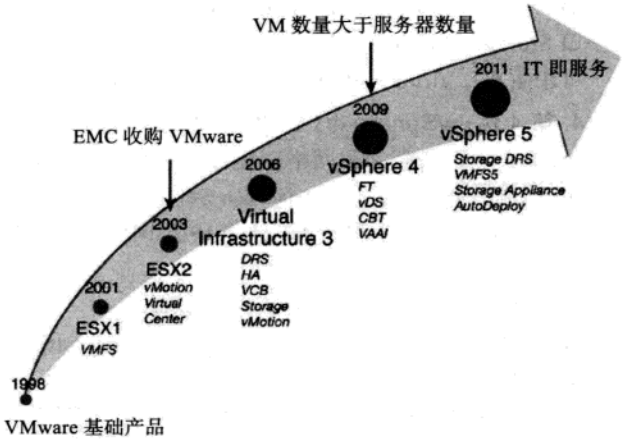


图 2-3 到 vSphere 5 为止的产品发展

1. 第 1 代（1998 ~ 2003 年）

ESX1 是 x86 平台的第一个虚拟化管理器，提供本地存储。此时还没有集中化的能力，服务器必须逐个管理。这一版本适用于低负载机器。这个虚拟化管理器用于测试和开发环境，没有直接的竞争对手。

可用版本：VMware ESX 1.x

2. 第 2 代（2003 ~ 2006 年）

vMotion 是第一种能够动态移动虚拟机器的技术，Virtual Center 能够集中管理多个物理服务器。虚拟化不再限于测试，也可以用于生产环境。VMware 在全球有超过 20 000 个客户，在这个市场上完全处于统治地位。竞争对手很弱，但是出现了有趣的开源解决方案，如 KVM 和 Xem。2004 年 EMC 收购了 VMware。

技术跨越：

- vMotion
- Virtual Center

可用版本：VMware ESX 2.x 和 Virtual Center 1.x

3. 第 3 代：Virtual Infrastructure 3（2006 ~ 2009 年）

虚拟化管理器不再相互独立。它们将资源聚合成统一的实体，提供很高的服务水平，具备高可用性和负载分配能力，由 Framework VCB 进行备份。VMware 在全球大部分大企业中

成为必不可少的资产。Citrix 于 2007 年以 5 亿美元的价格收购 Xen 加入了竞争。Microsoft 推出了 Hyper-V，进入这一领域。

技术跨越：

- ❑ VMware HA（High Availability，高可用性）
- ❑ VMware DRS（Distributed Resource Scheduling，分布式资源调度）
- ❑ VCB（VMware Consolidated Backup，VMware 整合备份）
- ❑ Storage vMotion（从第 3.5 版开始）

可用版本：Virtual Infrastructure 3（VMware ESX 3.x、Virtual Center 2.x）

4. 第 4 代：vSphere 4（2009 ~ 2011 年）

vSphere 4 扩展存储和网络功能并提供更多的可用性和安全功能。尽管 VMware 在这一市场处于统治地位，但是 Citrix Xen Server，尤其是 Microsoft 推出的 Hyper-V R2，都加剧了竞争。2009 年是重要的一年，生产环境中的 VM 在数量上超过了物理服务器。

技术跨越：

- ❑ VMware FT（Fault Tolerance，容错）
- ❑ vNetwork 分布式交换机（vNetwork Distributed Switch，）
- ❑ 各种 API
- ❑ 变更数据块跟踪（Changed Block Tracking）
- ❑ 数据恢复（Data Recovery）

可用版本：vSphere 4（VMware ESX4/ESXi4、vCenter Server 4）

5. 第 5 代：vSphere 5（2011 年以后）

现在，vSphere 5 自身的定位是为云计算优化的数据中心虚拟化平台。存储级的多项改进以及与最重要的软件和硬件制造商之间的稳固关系，使得 VMware 成为最能满足部署关键应用程序重要需求的解决方案。许可证策略也进行了修改，根据 VM 中配置的处理器和内存颁发许可证。放弃服务控制台得到更轻量级的 ESXi。许多新产品补充进来，如 vCenter Operations Management Suite、SRM 5、vShield 5 和 vCloud Director 1.5。

技术跨越：

- ❑ Storage DRS
- ❑ 新的存储 API
- ❑ VMFS-5（Virtual Machine File system 5，虚拟机文件系统 5）
- ❑ vStorage Appliance
- ❑ vSphere Replication for SRM 5

可用版本：vSphere 5（VMware ESXi 5、vCenter Server 5）

表 2-1 展示了最后三个版本的对比。

表 2-1 功能比较：VMware V13（vSphere 4 和 vSphere 5）

功能	VMware V13	vSphere 4	vSphere 5
主机服务器			
ESX 版本	ESX 32 位	ESX 4 64 位	ESX i5 64 位
最大主机内存	256GB	1TB	2TB
管理			
集中管理	Virtual Center	vCenter Server 4	vCenter Server 5
vCenter 链接模式		是	是
主机配置		是	是
DPM（分布式电源管理）		是	是
vApps		是	是
更新管理器		是	是
主机更新		是	是
vCenter Appliance			是
自动部署			是
映像构建器			是
高级功能			
vMotion	是	是，具有 EVC	是，具有 EVC
Storage vMotion	是（快照）	是（DBT）	是（镜像）
DRS（分布式资源调度器）	是	是	是
存储 DRS			是
可用性			
HA	是（AAM 代理）	是（AAM 代理）	是（FDM 代理）
FT		是	是
存储			
VMFS	VMFS3	VMFS3	VMFS3 和 VMFS-5
最大 LUN 尺寸	2TB	2TB	64TB
精简配置		是	是
存储 I/O 控制（SIOC）		从 vSphere 4.1 开始	是（SAN 和 NAS）
vSphere Storage Appliance		否	是
存储 API		从 vSphere 4.1 开始	是，VAAI2 和 NAS VAAI
备份			
备份 API	VCB	是，VADP	是，VADP
变更数据块跟踪		从 vSphere 4.1 开始	是
备份软件		是，Data Recovery	是，Data Recovery
网络			
vNetwork 分布式交换		是	是
VMDirectPath		是	是
网络 I/O 控制（NIOC）		从 vSphere 4.1 开始	是
网卡组合（NIC Teaming）	是	是	是
安全性			
vShield Zones		是	是
VMsafe		是	是

2.2 vSphere 5 许可证

下面几节提供重要的许可证信息，你应该在采用 vSphere 5 解决方案之前熟悉它们。

2.2.1 vSphere 5 版本

下面的 vSphere 版本面向中小型企业（Small and Medium Business，SMB）：

- ❑ vSphere Essentials
- ❑ vSphere Essentials Plus

vSphere Essentials 和 vSphere Essentials Plus 用于小规模部署。使用这些版本可以管理 3 个主机服务器，每个主机服务器使用两个物理处理器，最多可以使用 192GB 的 vRAM。这些版本集成了一个 vCenter Server Foundation 许可证。

免费的虚拟化管理器只能使用 vSphere 基本虚拟管理器功能。它可以透明地升级到更高级的 vSphere 版本。和付费版本不同，vSphere 虚拟化管理器不能由 vCenter Server 管理，只能通过与主机直接连接的 vSphere Client 管理。

注意： 一些研究显示，只有 10% 的 SMB 部署了虚拟化服务器。为了简化 VMware 在 SMB 中的部署，VMware 提供了 VMware Go，这是一个基于 Web 的免费服务，引导用户安装和配置 VMware vSphere。这个服务使用户能在几次点击内虚拟化其服务器，是完成虚拟化的简单方法。

对于中大型企业，有三个版本：

- ❑ vSphere standard
- ❑ vSphere Enterprise
- ❑ vSphere Enterprise Plus

表 2-2 比较了三个版本的特性。

表 2-2 功能比较：Standard、Enterprise 和 Enterprise Plus 版本

功　能	Standard	Enterprise	Enterprise Plus
ESXi 5.x	√	√	√
vCenter 代理	√	√	√
虚拟 SMP（vCPU）	8	8	32
vRAM 池 /CPU	32GB	64GB	96GB
更新管理器	√	√	√
VMFS-5	√	√	√
映像配置文件	√	√	√
vStorage API for DATA Protection	√	√	√
精简配置	√	√	√
vSphere HA	√	√	√
数据恢复	√	√	√
vMotion	√	√	√
Hot Add	×	√	√

（续）

功　能	Standard	Enterprise	Enterprise Plus
vSphere FT	×	√	√
vShield Zones	×	√	√
Storage vMotion	×	√	√
DRS	×	√	√
DPM	×	√	√
vStorage API for Multipathing	×	√	√
vStorage API for Array Integration	×	√	√
虚拟串行口集中器	×	√	√
分布式 vSwitch	×	×	√
主机配置文件	×	×	√
存储 I/O 控制	×	×	√
网络 I/O 控制	×	×	√
配置驱动存储	×	×	√
存储 DRS	×	×	√
自动部署	×	×	√
View Accelerator	×	×	√

注意： vSphere 4 中的 Advanced 版本不再提供，由 Enterprise 版本替代。只要维护协议是最新的，使用 Advanced 版本的客户就会自动升级到 vSphere Enterprise。

2.2.2 许可模式

如图 2-4 所示，VMware vSphere 5 按照处理器和 vRAM 授权赋予许可证。每个 VMware vSphere 5 处理器许可证自带一定数量的 vRAM 容量——可以配置到虚拟机的内存容量（称作 vRAM 池）。这一内存容量代表了 vCenter 中（启动的）VM 可配置的总内存。

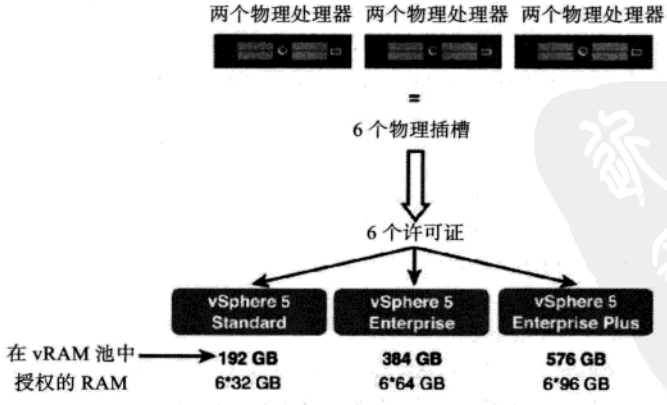


图 2-4 vSphere 许可模式示例

为了进一步解释图 2-4 中所示的模式，下面举一个例子：一个标准 vSphere 许可证授权池中每个 CPU 插槽可以使用 32GB vRAM。如果有三个双处理器（两个插槽）的物理服务器，就需要 6 个 vSphere Starndard 版本许可证。这 6 个许可证允许配置好（并启动）的 VM 拥有 vRAM 池中总共 6×32=192GB 内存。例如，如果每个 VM 配置了 2GB 内存，就意味着在池中允许有 96 个 VM 运行。扩展 vRAM 很容易：只要向内存池中添加更多的标准 vSphere 许可证。

注意： vRAM 的计算是根据年度平均使用率进行的，在一定时期内可以超过理论容量（在这个例子中是 192GB），只要平均值低于该容量就没有关系。（vCenter 会提出警告，但是不会妨碍性能，功能也不会受到影响）。如果平均值超过该容量，就必须解决问题，最大限度地利用 vSphere 的全部容量。

注意： 如果配置的 VM 内存超过最大 vRAM 值（对于 Enterprise Plus 是 96GB），需要增加一个许可证。例如，一个配置了 256GB 内存的 VM 仅需要购买一个许可证。

对于每个 CPU 的内核数量不再有 vSphere 4 中存在的限制（根据不同许可证，每个 CPU 6 ~ 12 个内核）。主机服务器的 RAM 也不再有限制。

vSphere 5 和 vCenter Server 的许可证分开销售。许可证密钥由 25 位字母数字组成。

2.2.3 vCenter Server 5.0 许可证

每个 vCenter Server 实例需要一个许可证。有三种可用的许可证：

- VMware vCenter Server for Essentials：与 vSphere Essentials 及 Essentials Plus 集成的许可证。
- VMware vCenter Server Foundation：用于小规模部署，最多支持 3 个 ESX 主机服务器（双处理器）。
- VMware vCenter Server Standard：用于大规模部署，不限制 ESX 主机服务器数量。

vCenter Server 许可证包括：

- vCenter 链接模式（仅用于 vCenter Server Standard）
- vCenter Orchestrator（仅用于 vCenter Server Standard）
- vCenter Server
- vSphere Client
- VMware vSphere Web Client
- VMware vSphere Update Manager
- VMware ESXi Dump Collector
- VMware Auto Deploy
- VMware vSphere Authentication Proxy
- vCenter Host Agent Pre-Upgrade Checker

图 2-5 展示了 vSphere 5.0 安装屏幕以及可用的不同产品和工具。

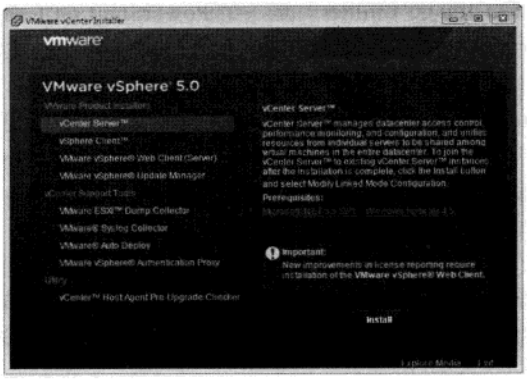


图 2-5 vSphere 5.0 安装页面

2.3 vSphere 5 的新增功能

vSphere 5 带来了许多新功能（据公布超过 200 项），其中许多功能与虚拟架构的关键部分——存储相关。下面是需要记住的一些关键功能：

- ESX4/ESXi 4 虚拟化管理器被更轻量级的仅有 144MB 大小的 ESXi 5 版本代替，该版本没有服务控制台。这个更轻量级的版本需要的更新较少，漏洞也更少。
- VM 现在可以使用新的虚拟硬件（版本 8），提供高水平的性能，支持最多 32 个 vCPU 和 1TB RAM，以及支持极高应用程序负载的 I/O 能力。
- 文件系统从 VMFS-3 升级到 VMFS-5，去除了卷容量方面的一些限制。
- 新的存储机制——vSphere Storage Appliance（VSA）帮助小规模部署从高级功能中得到好处（例如，vSphere HA、vSphere DRS、vSphere DRM、vSphere FT 和 vMotion）而不需要投资共享存储设备。
- 重新编写了 vSphere 高可用性（HA）。从 Legato 自动化可用性管理器（Automated Availability Manager）购买而来的代理被故障域管理器（Fault Domain Manager，FDM）代理所代替。
- 重新开发了 Storage vMotion 机制。该技术过去基于数据块变更跟踪，当 I/O 负载很大时这会造成很多困难，在融合阶段有出现故障的危险。在 vSphere 5 中，Storage vMotion 使用的新机制具备主机级别的独立驱动器，简化了融合工作。
- vSphere Storage DRS 是一项新功能（用于存储的 DRS 等价功能），能够根据可用空间和 I/O 负载均衡数据存储设备的负载。引入了一个新的概念：数据存储群集（datastore cluster）。
- vSphere 配置驱动存储（Profile-Driven Storage）能够创建判断存储性能的规则。利用这些规则，VM 可以与应该使用的存储服务选项关联，不需要关心其准确的位置。
- Linux SUSE 中的新 vCenter Server Appliance（vCSA）简化了小规模环境的部署。

- 映像构造器（ESXi Image Builder）是一个 PowerShell CLI 命令集，用于创建由最新驱动程序和更新预先配置的 ESXi 映像（管理员常用的一类主映像，又称黄金映像）。这样创建的映像可以根据定义好的策略加以使用和预先配置。
- 自动部署（Auto Deploy）简化了大规模部署，ESXi 映像可以加载到内存中而不是实际地安装到服务器硬盘，这称为*无状态* ESXi（stateless ESXi）。部署服务器时，自动部署根据与主机配置文件相关的一个*预启动执行环境*（Preboot Execution Environment，PXE）架构进行。
- 使用主机配置文件（Host Profile）功能，ESXi 服务器的工业化部署可以由预先定义的安装和配置参数执行。这样，就可以实现一个配置策略，保证 ESXi 服务器基础架构并符合规定的策略。可以使用主机配置文件配置网络、存储、vSwitch、许可证密钥等。

注意：*自动部署可以和映像构建器和主机配置文件关联，自动化大量服务器上定制映像的部署。*

- vSphere 5 引入了一个新的*命令行接口*（Command-Line Interface，CLI），简化了管理员的工作（过去管理员需要使用多种命令行工具）。VMware 提供了一个工具，以相同的语法规则提供统一的命令。
- 存储 I/O 控制支持*网络文件系统*（Network File System，NFS）。
- ESXi VM 交换空间可以移到本地*固态驱动器*（Solid-State Drive，SSD）或者共享的存储空间。在必须进行内存交换的情况下，这样可以降低对性能的影响。

2.4 现有功能

下面是 vSphere 当前功能的总结。有些功能是必不可少的，将在后续的章节中深入研究。vSphere 环境的动态特性用如下功能简化了维护操作和迁移阶段的工作：

- vMotion 能够在一个群集中的物理服务器之间移动 VM。这种迁移只发生在管理员的人工干预之后。*增强型 vMotion 兼容性*（Enhanced vMotion Compatibility，EVC）确保了不同代 vMotion 处理器之间的兼容性。
- **分布式资源调度器**（Distributed Resource Scheduler，DRS）自动化了 vMotion，在 ESXi 服务器之间分配 VM 的工作负载。
- **分布式电源管理**（Distributed Power Management，DPM）与 DRS 相结合，在更少的物理服务器上运行 VM，降低数据中心的电力消耗。
- Storage vMotion 是管理员的人工干预，允许从一个数据存储的 VM 到同一个存储设备或者不同存储设备中的另一个数据存储的虚拟磁盘热迁移，这种迁移不需要中断服务。存储 DRS 自动化了分配数据存储工作负载的操作。

高可用性和高服务水平由如下特性保证：

- **vSphere HA（高可用性）**确保了在 ESXi 主机服务器故障时，所有连接到该主机的 VM 立刻自动重新在群集中其他 ESXi 主机上重启。

- **vSphere FT（容错）**提供极高的可用性。当ESXi主机服务器故障时，被FT保护的VM不会经历服务中断或者数据丢失。它们继续在群集中的其他服务器上运行。与vSphere HA不同，HA中的VM会重启，而vSphere FT对于应用程序和用户都是完全透明的。
- **vCenter Data Recovery**是vSphere 5的基础架构备份解决方案，完全集成到vCenter Server。它基于vStorage API for Data Protection，允许时运行中的VM进行热备份。

服务质量（QoS）由如下特性提供：

- **存储I/O控制（SIOC）**从vSphere 4.1开始出现。通过在发生争用时分配I/O负载以提供数据存储级QoS。

注意：不应该混淆存储DRS和SIOS。当争用发生在数据存储级时，SIOC保证关键的VM获得一定的带宽和延时，而存储DRS通过将VM从一个数据存储迁移到另一个数据存储来分配数据存储的负载。

- **网络I/O控制（NIOC）**是SIOC在网络上的等价物。它提供VM之间与网络流量相关的QoS。其他分享VM资源的QoS通过共享、保留和资源池获得。

安全通过如下特性保证：

- **vShield Zones**是保护VM的虚拟防火墙，以一种用具的形式出现，也可以用于分析所有网络流量。vShield Zones防火墙处于vSwitch级别，使用阻止或者允许某个端口或者协议、网络流量的规则。
- **VMware VMsafe**是专用于保护VM的一个API。使用VMsafe的软件编辑器不需要在VM安装代理就能拥有保护VM的选项。

有效的存储管理通过如下特性实现：

- 使用**精简配置**，分配给VM的磁盘空间不需要从物理上保留。磁盘空间随着VM中数据量的增加而动态分配。这有利于存储空间的优化。
- **Volume Grow**允许在不关闭VM的情况下动态扩展现有VMFS。
- **热添加（Hot Add）特性**允许为VM热添加CPU和内存等组件。

注意：这项功能必须在VM高级参数中激活。（必须关闭VM才能激活参数。）支持大部分Windows 2008服务器和一些Windows 2003客户操作系统。

- 使用**热扩展（Hot Extend）**功能，可以在不关闭VM的情况下动态扩展虚拟磁盘。

网络管理通过如下特性实现：

- **vSphere标注交换机（vSphere Standard Switch，vSS）**为每个ESXi提供一个虚拟交换机。
- **vSphere分布式交换机（vSphere Distributed Switch，vDS）**是一个可以在几个ESXi服务器直接共享的虚拟交换机。与网络相关的交换策略可以实现并应用到所有的数据中心服务器。vDS允许VM保持其网络属性，方法是从一个服务器移动到另一个服务器，简化管理。

- **虚拟串行口集中器**（Virtual Serial Port Concentrator）允许建立 VM 和串行 IP 设备之间的通信。可以通过使用 Telenet 或者安全 Shell（Secure Shell，SSH）的网络来连接 VM 串行口。
- View Accelerator 允许从每个 ESXi 服务器上的某个 VM 映像中删除内存页面的重复数据。这能加速大量相同 VM 的启动，解决了启动风暴问题（同时启动大量 VM 引起的密集活动），因而对虚拟桌面基础架构（Virtual Desktop Infrastructure，VDI）环境很有用。

2.5 单独销售的软件

下面几节描述的软件产品可以单独购得。

2.5.1 vCenter SRM 5

SRM 5（vCenter Site Recovery Manager 5）是一个业务恢复解决方案，对可能发生在生产站点的事故提供了简单的保护。SRM 确保了灾难恢复计划（Disaster Recovery Plan，DRM）的集中化管理，自动化在应急站点上恢复生产的过程。通过 SRM 5，管理员能够在不影响生产的情况下进行转移测试，还可以用它进行计划迁移操作。这个版本的新特性如下：

- 集成主机级别的复制（仅异步模式）
- 集成的故障恢复（在前一版本中没有）
- 启用计划迁移

2.5.2 vCenter Converter

物理－虚拟（Physical-to-Virtual，P2V）转换工具能将物理机器转换为虚拟机器。VMware 提供了 VMware Converter 转换工具。这个工具很适合转换少量机器。对于大规模迁移，市场上的其他工具更为适合，因为它们能够工业化这一过程，提供更丰富的功能。

注意： *在 vSphere 5 中，VMware Converter 并未以 vCenter 插件的形式提供，所以必须在独立模式下安装。*

2.5.3 vCenter Operation Management Suite

VMware 持续投入研究和开发力量，改进 vCenter Operation Management Suite 这一有力的新工具，认为它是所有云计算项目的组成部分之一。这个套件的价值是为 vSphere 环境带来可见性，将管理员从手动操作中解放出来的全自动功能，以及主动事故管理。这个工具还通过使用基于配置策略的方法确保了合规性，对管理员部署虚拟化基础架构也有帮助。这个解决方案提供了如下特性：

- 云计算环境或者扩展的生产环境所需的功能和配置管理
- 性能管理功能和根源分析功能
- 环境总体可见性，可关联配置更改和性能异常

- 用于IT服务的成本分摊功能
- 跨越过去数月的报告功能
- 基础架构和应用程序依赖拓扑图

vCenter Operation Management Suite支持管理员的日常工作，并提供实际输入，帮助实现最佳的虚拟化基础架构规划。图2-6展示了vCenter Oprations界面。

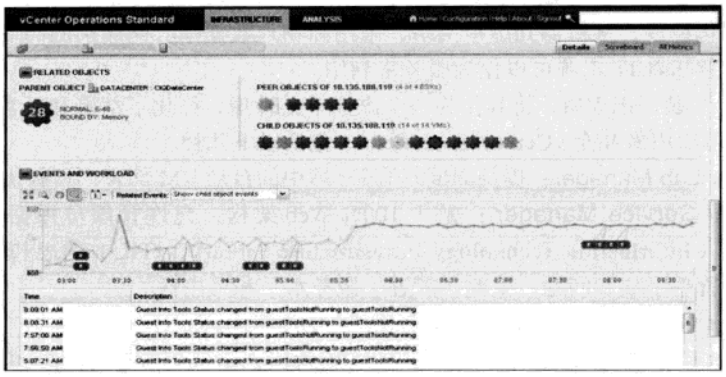


图2-6 vCenter Oprations界面

注意：服务提供商是对这个工具感兴趣的主要利益相关方，因为这个工具使它们能够以最优化的方式利用基础架构，从而使它们能够达到很高的整合水平并且更加节约成本。

vCenter Oprations有4个版本：Standard、Advanced、Enterprise、Enterprise Plus。Enterprise Plus版本将分析延伸到物理环境，如表2-3所示。

表2-3 vCenter Operations版本

vCenter Operations Management Suite				
	Standard	Advanced	Enterprise	Enterprise Plus
使用范围	小环境	大环境	虚拟和云基础架构	云和异构环境
vCenter Oprations Manager	√仅性能	√	√	√
vCenter Infrastructure Navigator			√	√
vCenter Chageback Manager			√	√
vCenter Configuration Manager			√仅vSphere主机	√

根据不同版本，包含如下特性。

- **vCenter Chageback Manager**：根据每个VM使用的资源分摊成本。数据来自vCenter Server。
- **vCenter Configuration Manager**：根据预先定义的标准和公司策略进行配置管理，在服务器和软件级别验证合规性。
- **vCenter Infrastructure Navigator**：自动发现应用程序服务、关系可视化和虚拟计算

机、存储及网络资源上的应用程序依赖映射。它能够进行基础架构和操作的应用程序感知管理，帮助管理员更好地理解更改的影响。

注意： vCenter Capacity IQ 是用于确定虚拟环境使用情况的容量管理工具。这个工具现在完全集成到 vCenter Oprations Management Suite，因此它无法单独销售。

- ❑ vCloud Director：全虚拟化数据中心的一个配置和应用程序部署解决方案。它管理所有虚拟化组件，从机器到虚拟网络，采用和物理数据中心同级别的高安全性。这个解决方案意味着 IT 资源可以作为服务来利用。
- ❑ vApp：封装一组 VM，将其作为一个虚拟单元管理，简化了部署和日常管理。

其他软件可以用来补充 vCenter Operations，包括以下几种。

- ❑ vCenter Lab Manager：提供测试 / 开发环境中的自动 VM 部署，具有快照 / 回滚功能。
- ❑ vCenter Service Manager：基于 100% Web 架构，允许管理员根据信息技术基础架构库（Information Technology Infrastructure Library，ITIL）的思想管理公司的过程——管理事故和问题、更改、配置、服务水平和可用性。
- ❑ vCenter Orchestrator：自动化某些任务，创建工作流。
- ❑ vCenter Capacity Planner：提供配置容量情况，以确定未来所需的物理资源。

2.6 vSphere 5 技术架构

从技术上讲，vSphere 5 是作为虚拟基础架构运行的一套组件。vSphere 5 提供许多新的特性，但是其架构（如图 2-7 所示）大体与 vSphere 4 相同。

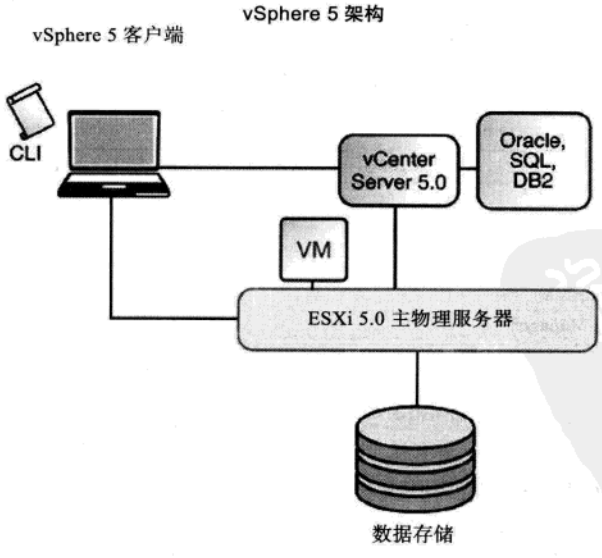


图 2-7 vSphere 5 基本架构

架构中的基本组件如下。

- vCenter Server：提供基础架构中所有 ESXi 服务器的集中化管理，由数据库服务器支持：Oracle、Microsoft SQL 、IBM DB2 和 Microsoft SQL Server 2008 Express。
- ESXi 5.0 主机服务器：基础架构的核心。

ESXi 5.0 是虚拟化管理器，是一个虚拟化层，可以在一台物理机器上运行多个操作系统。ESXi 服务器直接连接到 vCenter Server，其操作由 vSphere Client 接收。

- 数据存储：基础架构存储空间。

数据存储是数据中心物理存储资源的虚拟表现形式。它通过提出 VM 存储的标准模型，简化了市场上各种存储类型和解决方案的复杂特性。

- vSphere Client：基础架构的用户界面。

vSphere Client 是必不可少的组件，因为它是创建、管理和监控 VM 及 ESXi 主机所必需的界面。vSphere Client 可以直接连接到一台 ESXi 服务器或者 vCenter Server。

- vSphere Web Access：Web 界面。

这个 Web 界面是用 Adobe Flex 开发的，在虚拟和 ESXi 机器上进行基本管理和配置操作。允许在基于 Windows 和 Linux 的平台上从一个简单的浏览器来管理虚拟环境而不需要安装 vSphere Client。

- 虚拟机（VM）：代表一台服务器及其操作系统、应用程序和数据。

下面几节将深入研究 vSphere 5 的两个重要组件的架构细节：vCenter Server 5 和 ESXi 5 虚拟化管理器。

2.6.1 vCenter Server 5

vCenter Server 是虚拟基础架构的集中管理工具。它监控和管理 ESXi 主机服务器和 VM 中的所有活动。每个 vCenter Server 5 安装能够管理最多 1000 个主机服务器和 100 000 个 VM。图 2-8 展示了 vCenter Server 5 控制台。

图 2-9 说明了 vCenter Server 5 的架构。

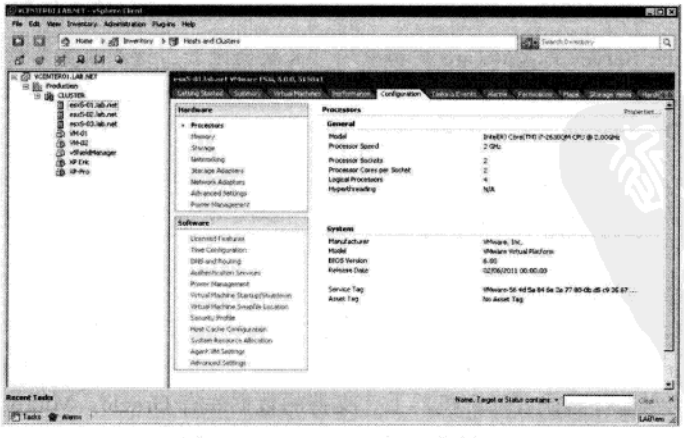


图 2-8 vCenter Server 5 控制台

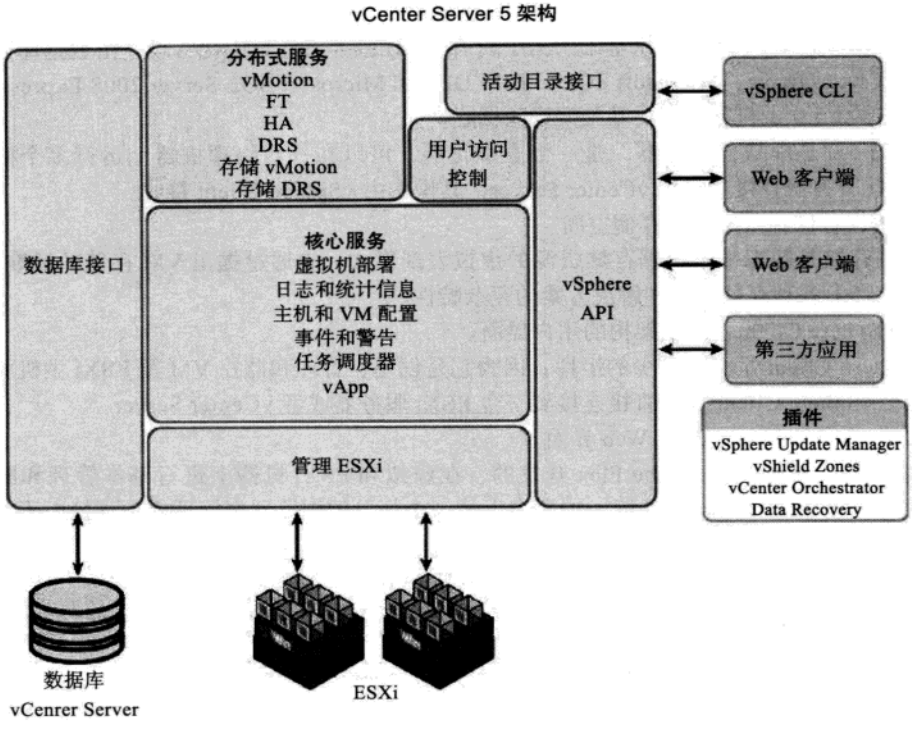


图 2-9 vCenter Server 5 架构框图

vCenter Server 5 的组件有用户访问控制、基本服务、分布式服务、插件和接口。vCenter Server 可以安装在 VM 或者物理服务器上，但是只能安装于 64 位 Microsoft Windows 环境中。对于小规模环境，也可以使用一个称为 vCSA 的 vCenter Appliance，该用具基于 Linux SUSE。

建议： 因为 vCenter Server 是 VMware 架构中不可或缺的元素，这个服务器需要很高的服务水平。因此，为了从 VMware 的 HA 特性中得到好处，VMware 建议将其安装在 VM 上，而不是物理服务器上。对于 vCenter Server VM，建议禁用 DRS 并赋予 HA 高优先级。

vCenter Server 是虚拟基础架构的集中化管理工具。VM 可以在没有 vCenter Server 的情况下工作，但是如果有些功能没有它就无法运行。表 2-4 详细地列出了停止 vCenter Server 对各个组件的影响。

1. vCenter Server 数据库

vCenter Server 数据库是虚拟基础架构数据库。每个 VM、主机、用户等的状态都存在于 vCenter Server 数据库中。该数据库在 vCenter Server 安装时进行安装和配置，可以位于 vCenter Server 的本地机器或者远程机器上。支持的数据库有 Oracle、Microsoft SQL 和 IBM DB2。Microsoft SQL Server 2008 Express 可以用于最多 5 个 ESX 和 50 个 VM 的小规模部署。

表 2-4 vCenter Server 中断：受影响的组件

组件	vCenter Server 中断的影响
虚拟机	在 14 天内不会产生后果。之后，许可证会阻止 VM 启动，关闭的 VM 无法重新启动
ESXi 服务器	只能用 vSphere 客户端直接连接 ESXi 主机。因为没有更多的服务控制台，所以不能使用 Web 界面
vMotion/Storage vMotion	不可用
vSphere DRS	不可用
vSphere 插件	不可用
vSphere HA	代理保持正常运行，能够启动故障切换。容许控制不可用

2. vCenter 链接模式

vCenter 链接模式能够连接到基础架构中的任何 vCenter Server 实例，虚拟化和管理组成群组的所有对象。这简化了管理员的工作，他不需要连接到每个 vCenter Server 实例，能够在一个客户管理控制台中拥有对基础架构的全局视图。管理员一次性定义各个角色，就能将它们应用到群组中的所有 vCenter 实例。

这种模式使用 Microsoft 活动目录应用模式（Active Directory Application Mode，ADAM）（随 vCenter Server 自动安装，基于轻量级目录访问协议（Lightweight Directory Access Protocol，LDAP））存储，并在不同 vCenter Server 实例间同步数据。这种模式可以在安装 vCenter Server 或者集成后安装时进行配置。下列数据在不同实例之间复制：

- 登录信息（IP 地址和端口）
- 证书
- 许可证信息
- 用户角色（每个用户可以查看和操作自己有权限的实例）

如果几个域之间存在信任关系，链接模式组中的 vCenter Server 实例可以位于不同的域。

注意： vCenter 链接模式组不是一个 vCenter Server HA 解决方案。它方便了从单个控制台对基础架构各 vCenter 实例进行管理。要确保 HA 能力，可以使用 vCenter Server Heartbeat。

3. vCenter Server Heartbeat

vCenter Server Heartbeat 是为 vCenter Server 提供 HA 的一个特性。它还提供了 vCenter Server 和其他服务中断时的故障切换管理。故障切换在主机服务器无响应时触发。在这种情况下，被动服务器立即接管主动服务器的角色。主机服务器和辅助服务器可以由两个物理服务器、两个 VM 或者一个物理服务器和一个 VM 组成。

服务中断可能由物理服务器停转、网络相关问题、SQL 数据库或者应用本身引起。

vCenter Server Heartbeat 保护如下服务：

- vCenter Server
- ADAM

- ❑ vCenter Management Web Server
- ❑ 更新管理器
- ❑ Guided Consolidation Service
- ❑ Orchestrator

4. vSphere 更新管理器

- ❑ vSphere 更新管理器（VUM）能够进行不同 vSphere 版本更新和补丁的集中和自动化管理。

更新管理器可以用于如下工作。

- ❑ 升级 ESXi 主机（VMkernel）
- ❑ 安装和更新第三方软件（例如，Nexus 1000v、PowerPath/VE）
- ❑ 升级 VM 虚拟硬件和 VMware Tools
- ❑ 更新到新版本的 VMFS

VUM 执行如下任务。

- ❑ 直接连接到 VMware 网站搜索 ESX 相关补丁，以收集最新的补丁。
- ❑ Update Manager 收集的信息用于定义基线。基线有两类：升级基线（upgrade baseline）定义主机服务器、VM 或者虚拟用具要求的版本级别；补丁基线（patch baseline）定义必须应用的更新级别。
- ❑ vSphere Update Manager 分析 ESX 主机服务器和 VM 的状态，将其与管理员定义的基线对比。分析完成后，不符合的机器将被标记为需要升级补丁。

1）对虚拟机应用补丁

为了减少应用补丁导致 VM 停止服务的风险，更新管理器可以在应用 VMware Tools 或者虚拟硬件更新之前获得 VM 状态的一个快照，这些快照的存储时间由管理员定义。

2）对 ESX 主机服务器应用补丁

vSphere Update Manager 还允许在使用 VMware DRS 的 ESX 主机服务器上进行无中断补丁应用。它将主机置于维护模式并将 VM 热迁移到其他主机，然后应用补丁。在补丁应用之后，主机离开维护模式，ESX 服务器上的 VM 可以回到生产模式。然后，VUM 为群集的下一台主机应用补丁。

注意： 使用 vSphere Update Manager 5 时，不再允许维护 VM 应用程序补丁和 OS 补丁，为了进行这些操作，你必须使用软件交付工具，例如 Microsoft 的系统中心配置管理器（System Center Configuration Manager，SCCM）或者 Windows 服务器更新服务（Windows Server Update Services，WSUS）或者 IBM 的 Landesk 或 Tivoli。

---

5. vCenter API

VMware 为 vCenter Server 提供 API。软件编辑器和集成器可以开发自己的解决方案并提供增值产品和功能，补充 VMware 所提供的产品。

6. vCenter Server 插件

插件由客户端和服务器端组成。安装服务器插件时，它注册到 vCenter Server，可以通过

vSphere Client从客户端下载。VMware提供某些可选插件，包括更新管理器、站点恢复管理器、数据恢复和自动部署。

7. vCenter Server Appliance

vCenter Server Appliance（vCSA）是一个预包装的64位SUSE Linux Enterprise Server 11。它包含了一个嵌入式数据库（DB2 Express），能够管理最多5个ESXi服务器和50个VM，也可以连接到外部Oracle或者IBM DB2数据库。

利用这一用具就不需要购买Windows许可证，从而降低了总拥有成本（total cost of Ownership，TCO）。部署操作也得以简化，需要做的所有工作就是将开放虚拟化格式（Open Virtualization Format，OVF）文件导入到vSphere 5平台。更新也因为更新管理器而变得更加简单。在管理员的层面，与vSphere Client的连接没有什么差别。

用具的配置可以通过Web界面完成，验证则通过活动目录或者网络信息服务（Network Information Service，NIS）进行。

vCSA支持所有经典的VM功能（例如，HA、快照或者vStorage API for Backup）。

vCSA VM配置如下：

- 两个vCPU
- 8GB vRAM
- 1个vNIC
- 两个vDisk
  - vDISK 1：5.3GB“精简盘”
  - vDISK 2：25GB“原盘”

在第一版中，vCSA有如下限制：

- 不支持第三方软件插件，也不支持SQL Server数据库。不支持链接模式和IPv6。
- vCenter Server Appliance与vSphere和vCenter许可证分开销售。

2.6.2 ESXi 5虚拟化管理器

ESXi 5是组成vSphere 5基础架构核心的虚拟化管理器。它是允许多个操作系统运行于一台物理机器上的虚拟化层。这个144MB的版本被称作“轻量”是因为它没有任何服务控制台。有些制造商将ESXi直接集成到服务器内部存储、USB key或者SD卡。

这个版本的优势如下：

- ESXi提供轻量级的144MB架构，而前一个版本具备服务控制台，大小为2GB。
- 加强安全性，需要的维护操作（如OS补丁和更新）更少。
- ESXi可以加载到内存而无需启动盘（这要归功于自动部署）。

1. ESXi 5组件

图2-10提供了ESXi 5组件架构的框图。

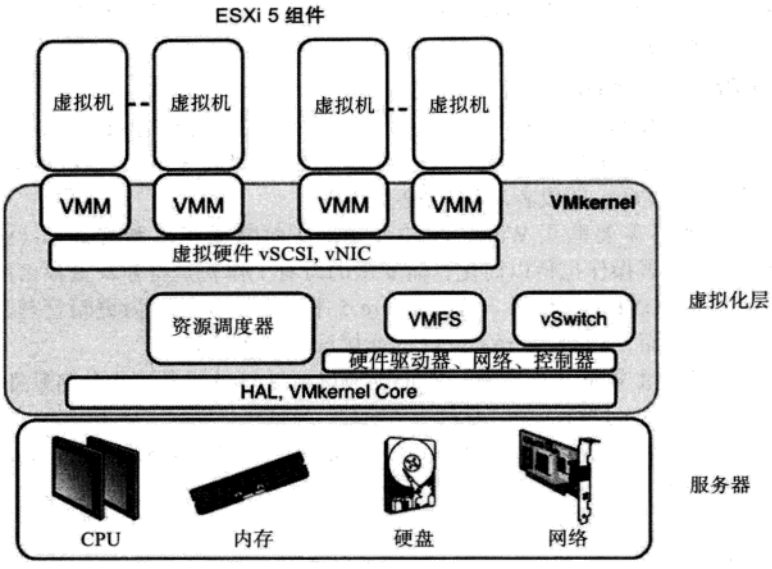


图 2-10 ESXi 5 组件

ESXi 5.0 包含下列主要组件：

- ❑ 虚拟化层
- ❑ VM

2. 虚拟化层

虚拟化层有两个组成部分：虚拟机监视器（Virtual Machine Monitor，VMM）和 VMkernel。

虚拟化层始终为给定的 VM 提供相同的虚拟硬件（vNIC、vSCSI），对 VM 隐藏各种不同的硬件组件。该层允许 VM 并发操作，并负责主机服务器的资源共享。

每个 VM 拥有自己的 VMM 实例。VMM 执行所有虚拟 CPU 指令，它确保 VM 内存和主机系统内存之间的通信。VMM 拦截来自 VM 的 I/O 请求并将它们提交给 VMkernel。VMM 还控制启动时的保证最低分配量（例如内存、磁盘），以及它们之间的隔离。

VMkernel 是虚拟化的核心和推动力，它是完全由 VMware 开发的（ESXi 5 为 64 位系统）。VMkernel 控制和管理服务器的实际资源，它用资源管理器（Resource Manager）排定 VM 顺序，为它们动态分配 CPU 时间、内存和磁盘及网络访问。它还包含了物理服务器各种组件的设备驱动器——例如，网卡和磁盘控制卡、VMFS 文件系统和 vSwitch。

注意： 许多人认为 ESX 和 VMkernel 是基于 Linux 发布版本的，因为它们能够使用 Linux 命令提示符（从早期版本的服务控制台）。事实并非如此，VMware 也非常清晰地表明这一观点：VMkernel 是一个专利产品。但是，早期版本的服务控制台确实是 Red Hat Enterprise 的一个修改版本，可能用来启动 ESX。还要注意，ESXi 自己能够启动，不需要服务控制台。

注意： ESX 旧版本中的服务控制台是一个命令行接口，可以授权访问 VMkernel，修改和配置 ESX 主机服务器参数，也可以用于加强 ESX。为了安全性和稳定性的原因，这个服务控制台被从虚拟平台中删除，代之以更加轻量的 ESXi 版本。

3. 虚拟机

虚拟机（VM）由客户操作系统（Operating System，OS）和虚拟硬件组成（见图 2-11）。

主机（或称 ESXi 主机）是安装 ESXi 的主物理服务器。客户操作系统是安装在虚拟机上的操作系统。虚拟硬件由虚拟网卡（virtual NetworkIng Card，vNIC）、vSCSI 和 vCPU 等虚拟组件组成。

图 2-11 客户 OS 和虚拟硬件

vSphere 5 中推出的第 8 版虚拟硬件提供如下支持：

- VM 中可有 32 个 vCPU 和 1TB vRAM
- 一个 3D 虚拟卡（用于支持 Windows Aero）
- 通过 vSphere Client 以客户模式连接 USB 设备的可能性
- 一个 USB 3.0 控制器
- 智能卡阅读器支持
- UEFI BIOS
- 用于 BIOS 启动顺序的配置 API
- E1000e 高性能网卡（仅可用于某些操作系统）
- 通过图形界面配置多核配置（没有高级参数）

硬件版本早于第 8 版的 VM 可以在 ESXi 5.0 主机中正常工作，但是无法使用所有功能。例如，使用版本 7，无法使用 32 个虚拟处理器。表 2-5 概述了虚拟硬件版本和各代 ESXi 主机之间的兼容性，图 2-12 展示了 Virtual Machine Version 8 Hardware（第 8 版虚拟机硬件）选项卡。

表 2-5 ESXi 主机和兼容的 VM 硬件版本

	与虚拟硬件的兼容性			兼容的 vCenter Server 版本
	版本 8	版本 7	版本 4	vCenter Server
ESXi 5.0	创建、编辑、运行	创建、编辑、运行	编辑、运行	vCenter Server5.0
ESX/ESXi 4.x	不支持	创建、编辑、运行	创建、编辑、运行	vCenter Server 4.x
ESX Server 3.x	不支持	不支持	创建，编辑，运行	VirtualCenter Server 2.x 及更高版本

VM 封装在文件中，这些文件包含了 VM 中运行的所有硬件和软件的状态。组成 VM 的文件如图 2-13 所示。

下面的列表简单地描述了图 2-13 中所示的文件。

- vmdk：对应一个元数据文件。这个虚拟磁盘描述（可编辑文件）提供指向 .flat-vmdk 文件的链接。
- flat-vmdk：最重要的文件，因为它是 VM 的虚拟磁盘，包含了 VM 的所有文件：操作系统、应用等。

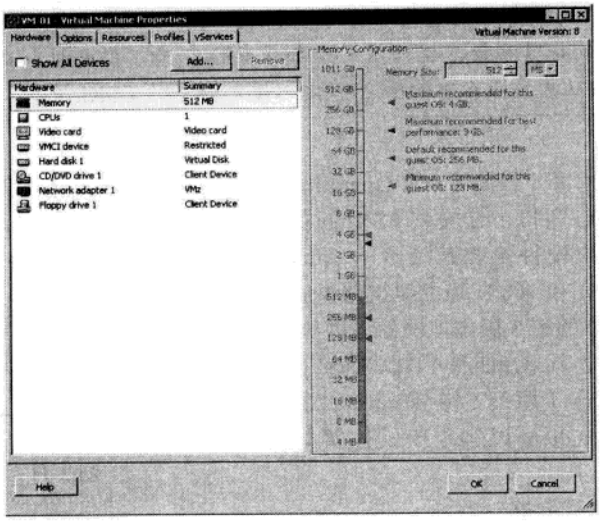


图 2-12　第 8 版虚拟机硬件选项卡

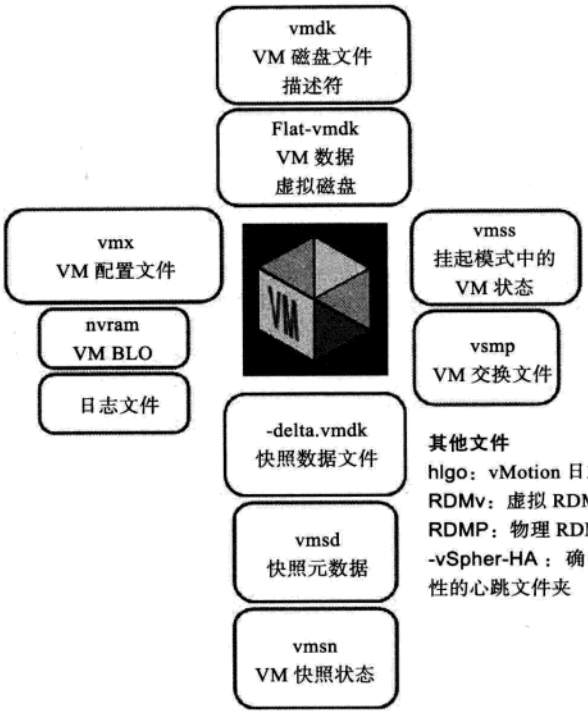


图 2-13　VM 文件

- vmx：包含 VM 的所有配置信息和实际参数，如内存大小、硬盘大小、网卡信息和 MAC 地址。它是 VM 创建时首先创建的文件。
- nvram：包含 VM BIOS 状态。
- log：跟踪 VM 的活动。存在多个日志文件，它们对于诊断问题很有用。VMware 支持服务使用这些文件。VM 关闭和重启或者文件达到高级配置选项中定义的最大大小时会创建一个新的日志文件（vCenter 高级选项：log.rotate size 和 logkeepOld）。
- vmss: 在 VM 挂起时创建。这个文件包含了活动内存的全部内容。当 VM 重新投入服务时，vmss 文件的内容返回到服务器的 RAM，创建与 VM 挂起前相同的工作环境。
- vswp：在 VM 启动时自动创建。该文件作为 VM 的内存交换。如果无法创建该文件，VM 就无法启动。

快照初始化时，创建如下文件类型。

- -delta.vmdk：在获得一个 VM 快照时创建和使用。在快照创建的时候，原始的 vmdk 文件进入静止（quiesced）状态，并被改为只读状态。原始文件中不再写入更多的数据。例如，00000#.vmdk 包含了与快照相关的元数据。
- vmsd：包含快照信息和元数据，包括相关 vmdk 和 vmsn 的名称。创建一个包含所有快照信息的文件。每个快照都创建一个 vmsn 文件。文件名使用递增的数字。例如，Snapshotxxx.vmsn 包含某个快照的状态，该快照包含了建立快照时运行中的 VM 状态。

其他文件如下所示。

- hlog：vMotion 日志文件。
- RDMv：虚拟兼容性模式中的原始设备映射。
- RDMp：物理兼容性模式中的原始设备映射。
- vSphere HA：包含特殊文件的一个特殊文件夹。例如，host-xxx-hb、Poweron 和 Protectedlist 是 vSphere HA 用于心跳和受保护 VM 一致性的文件。
- mutx：用作模板。

提示： *不正确地编辑这些文件可能导致 VM 失效。最好是保持这些文件的原貌。如果必须编辑，预先保留备份是必不可少的。*

---

4. VMware Tools

使用 VMware Tools，客户操作系统的虚拟硬件就能完美地与 ESXi 集成。VMware 添加了如下组件：

- 优化的驱动程序，如 vmxnet、LSI Logic SCSI 和 SAS
- 用于 VM 静止快照的同步（SYNC）驱动程序
- 内存气球驱动程序（vmmemctl）
- VM 心跳
- 时间同步
- 干净地关闭 VM 的可能性
- 向 Perfmon 添加动态链接库（Dynamic Link Library，DLL）的可能性

VMware Tools 极大地改进了图形显示和鼠标移动，并添加了有用的特性，例如，在 VM 启动之后添加脚本的选项。VMware Tools 必须在每个操作系统上安装。

2.7 安全性

下面的小节讨论与 vSphere 5 安全性相关的重要特性。

2.7.1 vShield Zones

VMware vShield 是一套安全虚拟用具和 API，用于与 vSphere 一起工作，保护虚拟数据中心免遭攻击和误用。vShield Zones 由一个管理器和一个虚拟用具组成，管理器提供管理界面，可进行策略部署，而虚拟用具提供如图 2-14 所示的安全性。这一用具自动集成到 vCenter Server。各个分区（zone）和外部世界之间的所有活动都得到监控，根据采用的策略过滤网络帧。

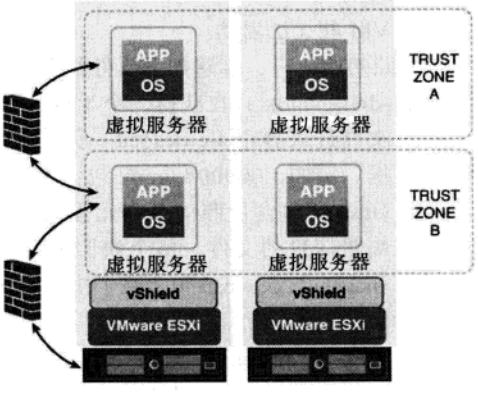


图 2-14 vShield Zones

vShield 套件包括 vShield Zones、vShield Edge、vShield App 和 vShield Endpoint。

- vShield Zones：为 VM 之间的流量提供防火墙保护。对于每条 Zones 防火墙规则，你可以指定源 IP、目标 IP、源端口、目标端口和服务。vShield Zones 是一个 VM 保护虚拟防火墙，采用用具的形式，也可以用于分析网络流量。利用 vShield Zones，可以创建逻辑分区，例如保证互联网流量与服务器内部流量隔离的非军事区（demilitarized zone，DMZ）。因为有了 vShield Zones，不再需要为 DMZ 创建专用的 ESXi 服务器。可以创建包含数百个 VM 的独立分区，而不需要管理复杂的 vSwitch 和 VLAN 配置。vShield Zones 防火墙位于 vSwitch 级，使用阻止或者允许某些端口或者协议 / 网络流量的规则。
- vShield Edge：提供网络边界安全性和网关服务，用端口组、分布式端口组或者 Cisco Nexus 1000V 隔离 VM。vShield Edge 提供常见的网关服务，如动态主机配置协议（Dynamic Host Configuration Protocol，DHCP）、虚拟专用网（Virtual Private Networking，VPN）、网络地址转换（Network area Translation，NAT）以及负载均衡，将独立的末端网络连接到共享的上联网络。vShield Edge 的常见部署包括 DMZ、VPN 外联网和多租户云环境，为虚拟数据中心（Virtual Datacenter，VDC）提供边界安全性。
- vShield App：一个内部 vNIC 级防火墙，允许你创建访问控制策略而无需考虑网络拓扑。vShield App 监控 ESXi 主机所有出站和入站流量，包括同一个端口组的 VM 之间的流量。vShield App 包括流量分析和基于容器的策略创建。
- vShield Endpoint：一个基于内省的防病毒解决方案。vShield Endpoint 使用虚拟化管

理器从外部扫描客户 VM，不需要代理。vShield Endpoint 避免资源瓶颈，同时优化内存的使用。

2.7.2 需要监控的组件

从安全的角度看，下列 ESX 组件必须受到监控：

- 虚拟化层
- 虚拟机
- 网络

1. 虚拟化层安全性

VMkernel 专用于支持 VM，而不用于任何其他目的。VMkernel 接口严格地限制在用于管理 VM 的 API。VMkernel 的保护已经得到加强，以保证其完整性。磁盘保护和完整性技术利用了硬件之素，如可信任平台模块（Trusted Platform Module，TPM）。

2. 虚拟机安全性

VM 互相隔离。这确保多个 VM 即使在共享物理资源的情况下仍能并发而安全地运行。如果一个 VM 崩溃或者遭到病毒感染，其他 VM 不会受到影响。

3. 网络安全性

网络是所有系统中最敏感和最脆弱的元素之一，它必须得到保护。VM 在共享服务器资源（CPU 内存）方面是相互隔离的，但是可能通过网络进行通信。和任何物理网络一样，网络的这部分必须用 VLAN、DMZ 等技术加强安全。还可以在 vSwitch 端口组级别上配置和实施安全策略。

2.8 发展的解决方案

VMware 解决方案是行业中最为成熟和成功的。在 1998 年，它曾经为机构部署和管理服务器的方法带来变革。从简单的虚拟化管理器开始，VMware 已经构建一个全面的产品系列，远远超出了基础的虚拟化管理器。有了组成 VM 的许多组件，管理工具变得至关重要，VMware 在这个领域也是出类拔萃的。

vSphere 5.0 是一个云操作系统，也是构建虚拟数据中心的基础。它虚拟化了整个 IT 基础架构，如服务器、存储和网络，集合这些异构资源，并将这些缺乏灵活性的基础架构转化为虚拟化环境中简单且可以统一管理的一组元素。

你可能会认为，具有如此之多的特性，VMware ESXi 的规模肯定很大。但 ESXi 5 的大小只有 144MB。VMware 已经尽可能地从虚拟化管理器中删除多余的代码，减少攻击面，从而改进安全性。看到这样高级的设计，就不难理解 VMware vSphere 持续地在这个市场上占据绝对的领先地位。

第3章
vSphere 5中的存储

存储通常是虚拟化架构中最关键的部件，在系统性能和可扩展性方面起着重要的作用。它必须支持 VM 的活动，并可以进行升级以满足未来的需求。在某些项目中，花在设计存储架构的时间可能占到全部工作量的 60%。因此，最好的解决方案必须根据业务约束、目标和分配的预算进行选择，因为不同的存储解决方案的成本有显著的不同。

3.1 存储的表现形式

因为 vSphere 5 提供了广泛的存储选项，知道各种选项提供的特性以及理解物理环境中的传统存储之间的交互和 vSphere 与这种环境的整合（见图 3-1）都是很重要的。

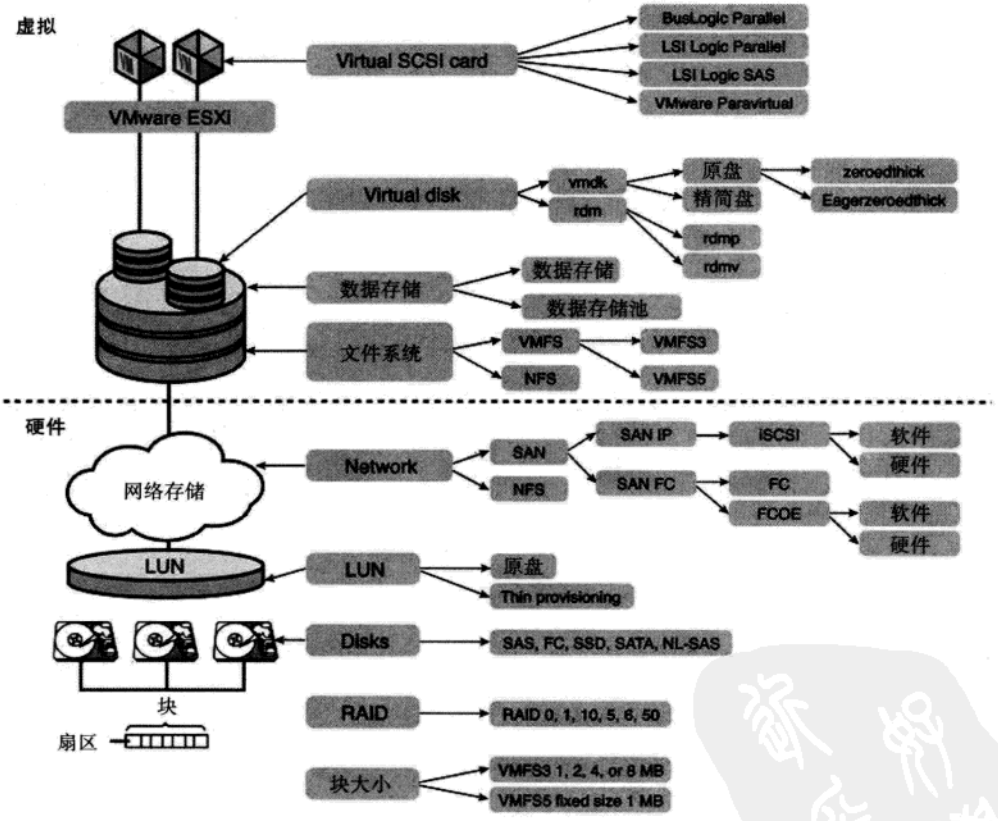


图 3-1 传统上由存储管理员操纵的实际对象（下方）与 VM 管理员操纵的部分（上方）之间的交互方式

3.2 可用的存储架构

VMware 支持多种存储协议，这可能使各个公司难以了解最适合其需求的选项。虽然这种灵活性和自由度可能是件好事，太多的选项还是使决策变得困难甚至无从着手。几年前，生产环境唯一可行的选择是存储局域网络（Storage area Network，SAN）光纤通道（Fibre Channel，FC），但现在协议之间的差别已经不那么重要，必须考虑多种标准。图 3-2 展示了支持的协议。

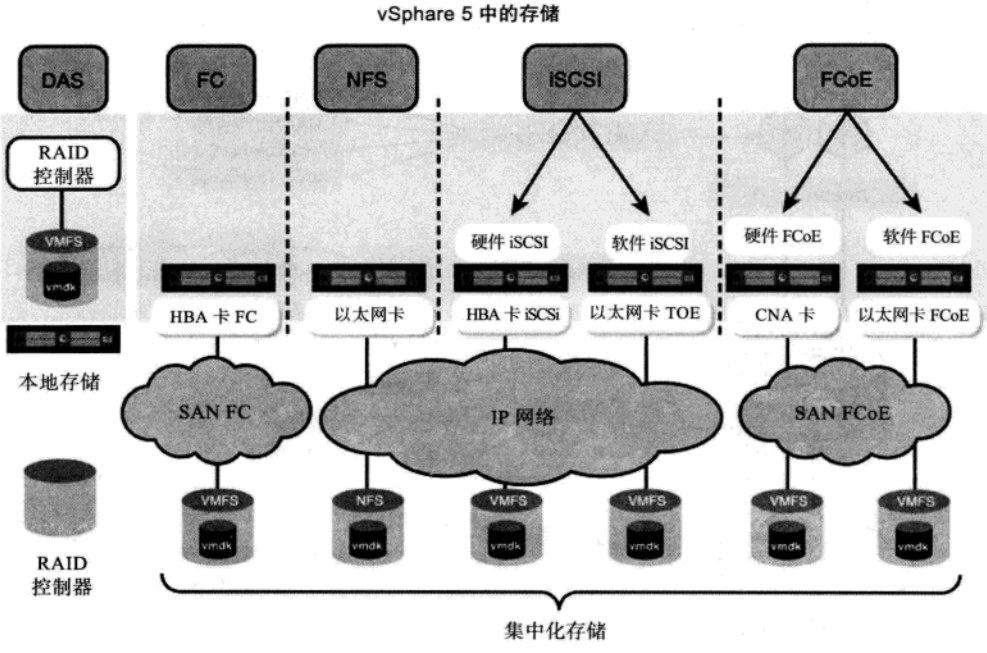


图 3-2 局部和集中化的存储架构

虚拟环境中可以使用如下的存储选项（在创建图 3-3 中的数据存储时选择）：

- 本地存储：硬盘直接连接在服务器中或者作为直接连接存储（Direct-Attached Storage，DAS），即直接连接到服务器的磁盘阵列。
- 集中化存储：存储在服务器外部。ESX 支持如下协议：
  - 光纤通道（FC）
  - 互联网小计算机系统接口（Internet Small Computer System Interface，iSCSI）软件或者硬件启动器
  - 网络连接存储（Network-Attached Storage，NAS）使用的网络文件系统（NFS）
  - 以太网光纤通道（Fibre Channel Over Ethernet，FCoE）软件或者硬件启动器

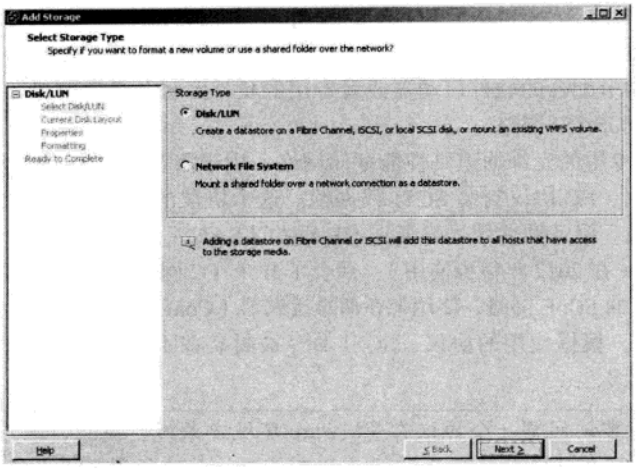


图 3-3 在创建数据存储时必须选择存储类型

3.2.1 本地存储

本地存储在安装 ESXi 虚拟化管理器时很常用。当 ESXi 服务器是独立的且不处于群集中，这个存储空间可以用于操作系统映像文件（以 ISO 文件的形式提供）或者非关键性的测试和开发 VM。因为本地存储的定义通常是非共享的，服务水平很低，所以关键的生产 VM 应该避免放在这种存储上。除非使用 vSphere Storage Appliance，否则 vMotion、分布式资源调度器（DRS）、高可用性（HA）和容错（FT）都无法使用。

3.2.2 集中存储

在集中式架构中，vSphere 可以工作于群集中，通过使用 vMotion、DRS、HA、FT 和站点恢复管理器（SRM）等高级特性改进服务水平。而且，这些架构类型提供出色的性能，并添加了 vStorage APIs for Array Integration（VAAI），将某些与存储相关的任务交给存储阵列，减轻主机服务器的负担。

NAS 存储服务器基于在 NFS 级别上访问数据的客户 / 服务器架构。这种协议称为文件模式（file mode），使用公司标准的以太网络，可以采用 1GbE（1Gbps）或者 10GbE（10Gbps）的网卡。

其他协议通过在被称为存储区域网络（Storage-Area Network，SAN）的专用网络中使用 SCSI 命令，提供主机服务器和存储之间的直接 I/O 访问（也称为块模式，block mode）。在 VMware 中，块模式相对文件模式的优势是原始设备映射（Raw Device Mapping，RDM）卷可以归属于 VM。VMware 在这种架构中使用虚拟机文件系统（Virtual Machine File System，VMFS）。

注意： 在 VMware 中，NFS 和 VMFS 之间存在明显的不同。NAS 服务器利用 NFS 管理文件系统并依赖于 ESXi 网络层（问题由网络团队解决），而 VMFS 直接由 ESXi 存储层管理。

SAN有基于IP的SAN和基于FC的SAN等不同类型：

- SAN IP（称作iSCSI）：通过TCP/IP网络封装SCSI命令（SCSI over IP）。你可以使用与标准网卡配合的软件启动器或者专用的硬件主机总线适配器（Host Bus Adapter，HBA）访问iSCSI网络。
- SAN FC：专用的光纤通道高性能存储网络，用于需要直接和顺序访问数据的高级I/O能力的应用。FC协议封装SCSI数据帧。这个协议的开销很小，因为SCSI包采用原生方式发送。服务器使用光纤通道HBA访问SAN。
- SAN FCoE（在2012年很少使用）：聚合了IP和FC网络。FCoE使用光纤通道技术，但是用于聚合的FCoE网络，使用聚合网络适配器（Converged Network Adapter，CNA）。

如图3-4所示，根据使用的协议，SCSI命令被封装在不同的层次中。使用的层次越多，主机级别的开销越多。

注意： 许多公司经常询问哪一个协议在VMware环境中最好。显然，这是很难回答的问题，就像在没有任何背景的情况下询问你两点之间的最佳行程一样。很明显，前往零售商店的最佳交通方式和出外度假是不一样的。因此，在回答有关最佳协议的问题之前，你需要知道总体的环境，以及基础架构、IT团队技能和虚拟化应用的类型（关键与非关键）、性能期望、财务考虑等信息。

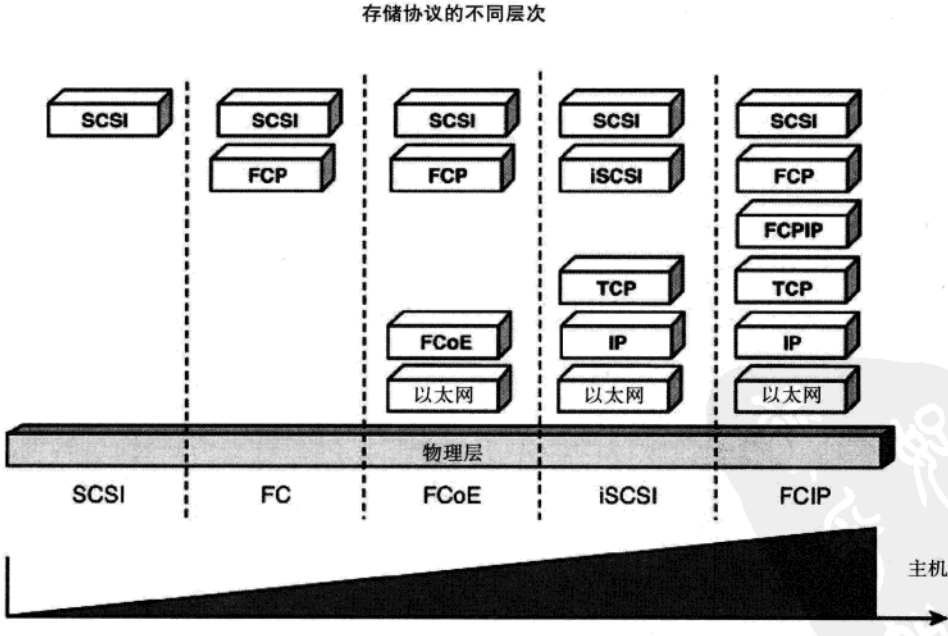


图3-4 不同协议中的SCSI命令层次

3.3 存储网络

前一节已经作过解释，可能使用的网络有两种：IP 以太网（NAS 或者 iSCSI 模式）和 FC 网络 (FC 或者 FCoE)。

3.3.1 IP 存储网络

这类网络原来的设计不是用于提供高性能存储的，而是在网络不同要素之间传递信息。因此，它不适合于需要高性能的应用，例如数据库应用。IP 网络处于 OSI 第 3 层，所以它可以路由，适合于长距离网络互联。FC 网络位于第 2 层，因此无法路由。目前，IP 网络的吞吐量达到 10GbE，未来将达到 40GbE 和 100GbE。

IP 网络的问题在于因为如下因素而导致“丢包”：

- 物理线路上的信号衰减
- 路由问题
- 缓冲区溢出（接收器无法处理输入流）

TCP/IP 协议允许丢包重传（如果发送数据没有得到接收方的确认），但是这对性能有显著影响。

另一个问题是，在一个 IP 包中只能传输有限的数据量，称作最大传输单元（Meximum Tranomission Unit，MTU）。这个数量（有效载荷，payload）对于以太网数据包为 1500 个字节。超过 1500 个字节的数据在发送之前必须分段。每当网卡接收到一个包，就向主机发送一个中断以确定接收。这增加了主机的负载和 CPU 周期（称作开销，overhead）。随着发送包数增加，路由变得更加复杂和费时。

注意： 在整合的虚拟环境中，需要考虑这种开销。它不应该恶化主机服务器的性能，主机服务器的处理能力应该专用于应用。

为了减少这种数据帧碎片，创建了超长帧（jumbo frame）。这种帧允许传递大于 1500 个字节的包（MTU 最大为 9000 个字节）。超长帧在效率改进中起到了很大的作用，一些研究显示，它将 CPU 开销减少了 50%。MTU 必须在整个传输链条中激活和兼容，包括物理交换机、网卡、电缆等。

注意： 9000 字节的超长帧替换了 6 个标准的 1500 字节以太网包，主机的 CPU 周期消耗减少了 5 倍。

不过，需要当心的是，如果出现问题，源和目标之间的 MTU 越大，重传的包也就越大，这会降低性能，增加延迟。为了最大限度地利用超长帧，网络必须稳定并且很好地实现。

IP 存储网络的优点是比 SAN FC 设备便宜，以太网络已经部署，所以在某些情况下，需要的实现工作较少，这种网络更容易使用。而且，IT 团队对这种技术已经有多年的使用经验。

1. VMware 中的 iSCSI

在 VMware 环境中，iSCSI 协议在 2006 年开始才被支持。如果以最优化的方式部署，这种协议能提供非常好的性能。IP 网络由存储团队之外的团队管理。

优点：iSCSI已经得到许多活动部门的采用，因为它使用了公司的TCP/IP网络进行块模式访问，不需要投资FC设备。因此，它更容易安装，从而对于某些环境来说是理想的选择。使用传统的以太网络，意味着覆盖更长的距离，而不需要特殊的转换设备（例如FC-IP转换器），例如，用于复制。实施这种技术所需的技能是网络技能，而不是高级的存储技能。

缺点：测试证明，iSCSI协议是消耗CPU资源最多的协议。因此，监控CPU的使用很重要，应该在网络部署时加以考虑。

我们建议如下的最佳实践：

- ❑ 只在架构能够启用超长帧（MTU 9000），从而最大限度地利用该协议的时候，使用iSCSI才有价值，这种情况能够提供出色的性能。启用超长帧必须在传输链的两端进行。
- ❑ iSCSI HBA卡在使用10GB连接时必不可少，应该尽可能实现链路聚合，以提供高性能和冗余，以防故障。
- ❑ 建议从物理上分隔iSCSI存储网络和标准IP网络。如果不可能，应该使用虚拟局域网（Virtual Local-Area Network，VLAN）隔离传输流。
- ❑ 使用具有TCP减负引擎（TCP Offload Engine，TOE）的网卡，从主机接管某些与iSCSI层相关的指令，减少开销。
- ❑ 实施服务质量（QoS），为传输流设置优先级。在vSphere中，可以使用存储I/O控制（Storage I/O Control，SIOC）功能实现。
- ❑ 网络丢包是获得好的iSCSI网络性能所面临的主要挑战之一。丢包可能是因为错误的网络配置或者错误的线路质量引起的（例如，在千兆链路上使用5类线而没有使用6类线）。

2. VMware中的NFS

网络文件系统（NFS）是NAS使用的协议，从2006年起得到ESX的支持。它通过网络提供文件系统级的存储共享。VMware支持NFS over TCP第3版。和某些说法相反，测试显示在正确实现的情况下，这个协议的性能很好。因此，某些条件下在虚拟环境中可以使用这种协议。超长帧（MTU 9000）的启用允许传输8192（8KB）NFS数据块，很适合于这个协议。默认情况下，ESXi主机可以安装8个NFS，并且可以扩展到256个NFS。如果将最大NFS安装数增加到超过默认的8个，确保一定要增加Net.TcpipHeapSize和Net.TcpipHeapMax。这些值在高级配置中，控制堆存储的数量（以兆字节为单位），这些存储分配用于管理VMkernel TCP/IP网络连通性。

- ❑ ESXi 5.0：将Net.TcpipHeapSize设置为32
- ❑ ESXi 5.0：将Net.TcpipHeapMax设置为128

注意：默认情况下，精简配置是NFS数据存储上创建的虚拟磁盘所用的格式。

优点：和iSCSI一样，NFS使用标准的TCP/IP网络，非常易于实施，不需要专用的存储架构。这是最便宜的解决方案，也不需要特殊的存储技能。NAS往往提供重复数据消除功能，这能减少存储空间需要量。

缺点：在这里描述的所有解决方案中，NFS的性能最低，但是接近于iSCSI。它所利用

的主机服务器 CPU 比 FC 协议多，但是少于 iSCSI 软件。所以可以想象，它可以用于以下生产环境：VM 要求第 2 层和第 3 层应用的平均性能。

注意： 在 vSphere 5 中，这个协议不支持 NFS 启动或者 RDM 的使用。

我们建议如下的最佳实践：

- 每个 NFS 卷使用 100 ~ 400 个 vmdk 文件。对于最大尺寸为 64TB 的 NFS 卷来说，最大可能逻辑单元号（Logic Unit Number，LUN）为 256。（制造商可能提供文件系统支持限额的信息，通常是 16TB）
- 使用专用交换机或者 VLAN 分隔专门的存储网络和以太网络。
- 启用流控制
- 使用具有大的端口缓存的专用交换机，启用超长帧
- 启用生成树协议
- 使用 10Gb 网络（强烈推荐）
- 使用全双工 TOE 卡降低 ESXi 主机服务器负载
- 为 NFS 和 iSCSI 流量使用专用的交换机或者 VLAN，将存储流量与其他网络流量分离

3.3.2 光纤通道网络

从根本上说，光纤通道网络是专用于存储的，提供对块模式数据的直接无丢失访问。这种网络是为高性能存储设计的，通过缓冲区信用阀值（buffer credit，一种用于调整 SAN 数据流的缓存）等高级机制得到很低的延迟。FC 协议通过专用的光纤通道网络封装 SCSI 包。速度为 1Gbps、2 Gbps、4 Gbps、8 Gbps 或者 16Gbps。FC 包携带的有效载荷为 2112 个字节。这种存储网络通过光纤通道交换机在服务器和存储设备之间传送数据。图 3-5 中展示的 SAN 实现了存储整合，提供了高伸缩性。

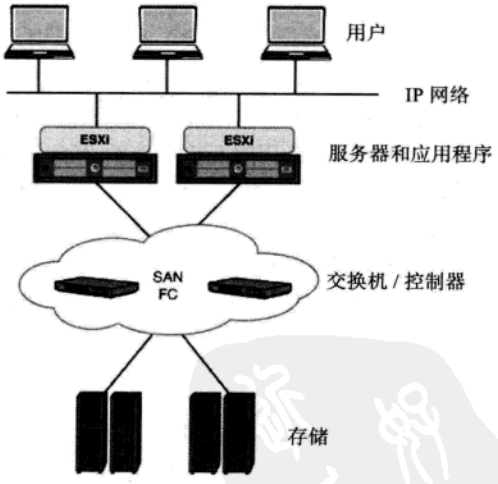


图 3-5 光纤通道 SAN 架构略图

1. VMware 中的 SAN FC

光纤通道（Fibre Channel，FC）是 VMware 支持的最高级协议。这也是它成为生产环境中最常用协议的原因。

优点：目前为止，FC 似乎是性能最高的协议，和 NFS 和 iSCSI 相比，它所使用的主机服务器 CPU 资源也最少。该协议能实现很高的性能，而且因为这种技术是无丢失的，因此网络是可预测的。这种协议适用于所有流行的应用，对于数据库或者企业资源计划（Enterprise Resource Planning，ERP）等 I/O 密集型应用很理想。

缺点：FC是最昂贵的解决方案，因为它需要建设特殊的存储架构，并且需要投资HBA卡、FC交换机、小封装热插拔收发器（Small form-factor pluggable transceiver，SFP）和电缆。而且，实现这种解决方案更为复杂，需要专门的存储技能。培训是必需的，还要学习管理SAN的新术语，例如LUN masking、Zoning WWN和Fabric。

我们建议如下的最佳实践：

- ❑ 为了减少损坏的链路，在每个服务器上插入多张HBA卡，并使用多存储阵列访问路径。
- ❑ 使用负载均衡软件（如ESXi Round Robin或者EMC PowerPath/Virtual Edition），优化服务器和存储之间的路径管理。
- ❑ 使用符合ALUA，与VMware的VAAI API兼容的存储阵列。
- ❑ 在群集所有成员之间使用相同数量的路径，群集中的所有主机服务器应该看到相同的卷。
- ❑ 服从ESXi群集成员和存储之间的连接兼容性矩阵。

注意： 我们曾经遇到一些管理员，他们将服务器直接连接到控制器存储阵列，以节约交换机的成本！这样的做法失去了FC SAN提供的冗余性，不建议这么做。

- ❑ 在所有连接中使用相同速度的交换机，避免在SAN中造成竞争点。

示例： 某公司有具备FC 8Gb端口的刀片服务器。SAN核心交换机速度为4Gbps。在核心交换机上出现一个明显的竞争点，对所有连接的元素都有影响。正确的做法是强制新设备采用4Gbps的速度。

- ❑ 检查FC交换机和HBA的固件级别，遵循存储阵列制造商的使用说明。

2. VMware中的SAN FcoE

以太网光纤通道（FCoE）代表着几个不同领域的融合：以太网络（TCP/IP）、存储用的SAN（SAN FC）和群集所用的InfiniBand（IPC）。这意味着，现在可以在这些不同的协议上使用一种类型的接口卡、交换机电缆和管理接口。FC帧被封装在以太帧中，提供了比TCP/IP更有效的传输。

FCoE帧携带2500个字节的有效载荷。目标是实现和FC类似的以太网无丢失性能。这通过加强物理网络可靠性和一些改进（尤其是关于QoS的改进）来实现。FCoE需要专用的设备，也需要启用超长帧（2180字节）。通过流控机制消除了拥堵。

因为FCoE在2012年中还相对不常见，我们对这种协议在VMware环境中的优缺点还缺乏实际的经验。

3.3.3 哪个协议最适合你

在我们的经验中，SAN FC是虚拟生产环境中管理员首选的协议管理器。据估计，有70%的客户在VMware生产环境中使用SAN FC。然而，具有超长帧的10GbE的出现，使得SAN IP基础架构易于实现，同时保持着能够满足某些情况的性能水平。除了技术标准之外，最好的选择还要根据现有架构和预算来决定。

总结如下：

- ❑ SAN FC应该是需要高性能（第1层和第2层）的应用（如数据库应用）首选的方案。

- iSCSC 可以用于 2 层应用。有些公司在 iSCSI 中使用 IP 进行远程数据复制，这很有效且能控制成本。
- NAS 可以用于网络服务，例如基础架构 VM——域控制器、DNS、文件或者非关键应用服务器（第 3 层应用），也可以用于 ISO 映像、模板和 VM 备份存储。

3.4 VMFS

虚拟机文件系统（VMFS）是由 VMware 开发的文件系统，专用于群集虚拟环境和大文件存储，并为此作了专门的优化。VMFS 的结构使其可以在单个文件夹中存储 VM 文件，简化了 VM 的管理。

优点：传统文件系统只允许单个服务器获得存储资源的读 / 写访问权。VMFS 是所谓的*群集文件系统*，允许多个 ESXi 主机服务器同时读 / 写存储资源。为了确保多个服务器不会同时访问同一个 VM，VMFS 提供*磁盘锁*（on-disk locking）系统。这一系统确保 VM 在某个时刻只与单个 ESXi 服务器协作。为了管理访问权，ESXi 使用 SCSI 保留技术修改元数据文件。这种锁的持续时间很短，避免整个 LUN 的 I/O 操作不会被 ESXi 服务器和 VM 独占。不要频繁进行 SCSI 保留之所以重要，也正是因为它们会损害性能。

ESXi 在如下情况使用 SCSI 保留：

- 创建一个 VMFS 数据存储
- 扩展 VMFS 数据存储
- 启动 VM
- 获得文件锁
- 创建或者删除文件
- 创建模板
- 从模板部署一个 VM
- 创建新 VM
- 用 vMotion 迁移 VM
- 扩张文件（例如，VMFS 快照文件或者精简配置虚拟磁盘）
- 使用 HA 功能（如果服务器出现故障，释放磁盘锁，允许另一个 ESXi 服务器重启 VM 并将磁盘锁用于自己的目的）

注意： VAAI 特殊特性之一——硬件辅助锁减少了 SCSI 保留。这个 API 将锁的活动直接交给存储阵列控制器。

3.4.1 VMFS-5 规范

vSphere 5 推出 VMFS 5，最大容量为 64TB。表 3-1 介绍了 VMFS 第 3 版到第 5 版的发展。

VMFS-5 提供了比 VMFS-3 更高的容量限值，因为它的地址表被拓展为 64 位。（VMFS 提供 32 位地址表，容量限值为 256 000 个 8MB 的块（2TB））。在 VMFS-5 中，块的大小固定为 1MB，最大容量为 64TB。在 VMFS-3 中，块尺寸在 1MB ~ 8MB 之间，在块大小太小

时可能导致虚拟磁盘最大容量问题。（例如，1MB 的块使 VMDK 文件的大小限制在 256GB，为了更大的文件尺寸，该卷必须用正确的块大小重新格式化）子块从 64KB 变成 8KB，因此 VMFS-5 可以管理 1KB 的小文件。

表 3-1 VMFS-3 与 VMFS-5 的对比

功能	VMFS-3	VMFS-5
最大容量	2TB	64TB
块大小	1MB、2MB、4MB 或 8MB	1MB
子块	64KB	8KB
小文件	不支持	1KB

你还应该注意如下两点。

- ❑ 必须为每个 LUN 创建一个 VMFS 数据存储。
- ❑ VMFS 保存一个事件日志。这保证了数据完整性，在问题发生时能够快速恢复。

3.4.2 从 VMFS-3 升级到 VMFS-5

VMFS-3 与 VMFS-5 兼容。从 VMFS-3 升级到 VMFS-5 得到支持，并且可以在不中断服务且 VM 同时运行的情况下发生。但是，最好是创建新的 VMFS 卷，因为 VMFS-3 到 VMFS-5 的升级有如下限制。

- ❑ 块保持原始的大小（可能大于 1MB）。不同块大小的数据存储之间的复制操作无法从 VAAI 全复制特性中获益。
- ❑ 子块仍为 64KB。
- ❑ 创建新的 VMFS-5 卷时，文件最大数量仍然保持为 30 720，而不是 10 万个文件。
- ❑ 仍然使用主引导记录（Master Boot Record，MBR）类型分区，但是在卷大于 2TB 时自动变为 GUID 分区表（GUID Partition Table，GPT）。

3.4.3 VMFS 数据存储签名

每个 VMFS 数据存储都有一个全局唯一标识符（Universal Unique Identifier，UUID），用于标识 VMFS 数据存储所在的 LUN。这个 UUID 必须是唯一的。如果两个 VMFS 使用同一个 UUID 同时进行安装，ESXi 无法知道在哪个卷上执行读 / 写操作（操作将随机发送到每个卷），这可能导致数据损坏。vSphere 会发现这种情况并加以避免。

注意： UUID 保存在 VMFS 文件系统的头信息里，根据 4 个变量生成：日期、时间、ESXi MAC 地址的一部分以及存储阵列中的 LUN 标识符。这确保了环境中该值的唯一性，并形成了 VMFS 卷元数据的一部分。

当 VMFS LUN 被复制、快照或者复制时，创建的 VMFS LUN 完全和原来相同，包括 UUID。为了利用新的 VMFS LUN，可以赋予一个新的签名，或者使用如下选项（如图 3-6 所示），在特定条件下保留原来的签名：

注意： 卷签名适用于 FC 或者 iSCSI 中的 VMFS 文件系统，但是不适用于 NFS 卷。

- 保留现有签名：这个选项能够保留相同的签名和复制的数据存储的安装。为了避免UUID冲突，这一安装仅在源VMFS LUN被卸下（或者移除）时进行。

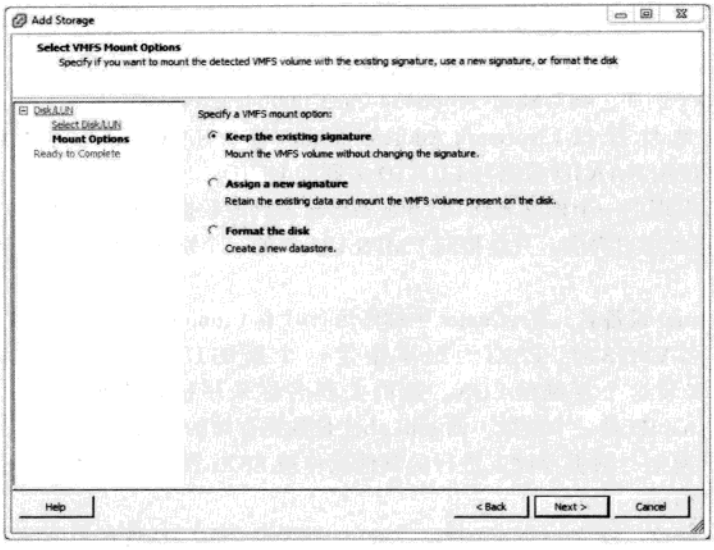


图 3-6 重新安装复制或者快照LUN时提供的选项

- 指定一个新的签名：重新签名VMFS时，ESXi为LUN副本指定一个新的UUID和名称。这就可以用不同的标识符同时安装两个VMFS数据存储（原始卷和副本）。注意，重新签名是不可逆的。要记得进行数据存储重扫描，以更新ESXi中的LUN。
- 格式化磁盘：这个选项完全重新格式化各个卷。

注意： 如果VMFS数据存储包含VM，重新签名可能带来一定的后果。确实，每个VM的配置文件（vmx、vmsd和vmdk文件）根据UUID值指定VM虚拟磁盘所在的数据存储。在卷重新签名的情况下，这些文件中的UUID值不再正确，因为它们用旧UUID指向以前的VMFS。VM必须在vCenter中重新注册，以整合新的UUID，数据中心、资源池和网络映射也必须重新设置。

1. 作为DRP一部分的VMFS卷重签名

在实施灾难恢复计划（DRP）和复制卷改变签名的情况下会生成一个新的UUID。卷上记录的VM的vmx和vmdk配置文件指向原来的UUID，而不是新的卷。因此，DRP计划中的所有VM必须从vCenter的库存中手工删除，然后重新记录，以便恢复新的UUID。这是一个麻烦的过程，在手工操作时可能导致处理错误。

站点恢复管理器（SRM）第5版提出的宝贵想法之一是自动化这种工作流程以简化过程和避免错误。利用SRM 5，复制卷在备份站点上被重新签名，配置文件自动引用正确的UUID，使VM指向新的复制卷。每个受到保护的VM都与指定的虚拟磁盘相关联。

注意： 使用 RDM 卷时，手工操作更加复杂，因为 RDM 的 VMFS 指针不复存在。SRM 也能自动用重新记入库存的新 VM 重新映射这些卷。

2. 技术细节

在这种环境中，VMFS 卷以如下方式表示：

- 用 UUID（例如，487788ae-34666454-2ae3-00004ea244e1）。
- 用网址地址授权（Network Address Authority，NAA）ID（例如 naa.5000.xxx）。vSphere 使用 NAA ID 检测与 LUN ID 关联的 UUID。
- 用 ESXi 发现的一个标签名和 vCenter server 看到的数据存储名（例如，myvmfsprod）。这个名称由用户提供，只是指向 VMFS UUID 的一个别名，但是用它更容易找到数据存储。
- 用 VMkernel 设备名，在 vCenter 中称作运行时名（runtime name），例如，vmhba 1:0:4。

当重新签名 VMFS 时，ESXi 为副本指定一个新的 UUID 和新的标签名，并和原始 VMFS 一样安装这个复制的 LUN。新的关联名称采用格式类型 snap，例如，snapID-oldlabel，其中 snapID 是一个整数，而 oldLabel 是数据存储的原名。

除了快照和复制，数据存储上进行的其他操作被 ESXi 视为原始存储的一个副本，因此需要管理员采取行动。

- LUN ID 修改：修改 LUN ID 时，vSphere 发现这个 UUID 现在与新设备关联。
- 修改 SCSI 类型：例如，从 SCSI-2 改为 SCSI-3。
- 为某些系统激活 SPC-2 符合性：例如，EMC Symmetri 需要这种激活。

3.4.4 重新扫描数据存储

每次在 ESXi 或者存储级别上进行存储相关的更改，都有必要重新扫描存储适配器，纳入新配置。这能更新可见数据存储列表和相关的信息。

每次执行如下任务时都必须重新扫描：

- 在 SAN 级别上修改分区（zoning），这会影响 ESXi 服务器
- 在 SAN 中创建新 LUN，或者进行重签名
- 在存储阵列中修改 LUN 屏蔽
- 重连电缆或者光纤
- 在群集级别上更改主机

默认情况下，VMkernel 扫描 LUN 0 ~ LUN255。（记住，主机引入的最大 LUN 数量是 256）为了加速扫描过程，可以在高级参数中设置较低的值：Disk.MaxLUN（例如，在图 3-7 中设置为 64）。

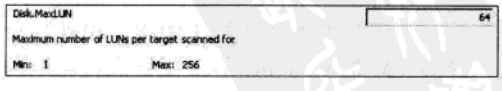


图 3-7 进行数据存储扫描

注意： 你还可以右键点击数据中心、群集或者包含相关主机的文件夹，启动数据存储的重新扫描。

3.4.5 对齐

对齐（Alignment）是需要考虑的重要问题。堆叠多个层次可能造成不对齐的分区，如图 3-8 所示。图 3-9 展示了与此相对的对齐分区。

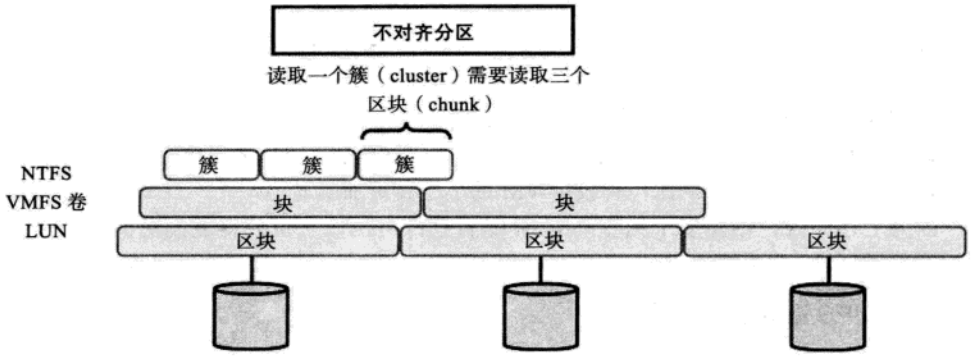


图 3-8 不对齐分区

RAID 栈中最小的单位称为区块（chunk）。下一层是 VMFS，它使用 1MB 的块（block）。在上面格式化的 NTFS 使用 1 ~ 64KB 的块（称作*磁盘簇*，disk cluster）。如果这些层次没有对齐，读取一个簇就可能意味着读取两个覆盖 3 个不同硬盘上 3 个区块的数据块，这可能影响写入，从而降低性能。

分区对齐时，一个簇只需要读取一个块，而块与区块对齐。这种对齐很重要，在 VMware 环境中，不对齐可能使性能下降 40%。

在 Microsoft 环境中，Windows Server 2008 自动对齐，而旧的操作系统必须用 Diskpart 工具对齐。参见软件制造商的说明书。

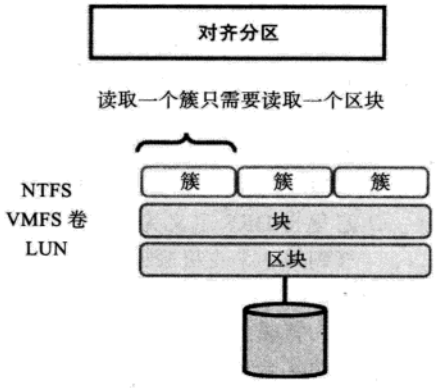


图 3-9 对齐分区

3.4.6 增加容量

卷增长（Volume Grow）功能可以动态地扩展现有的 VMFS，而不需要关闭 VM（最多 32 次扩展）。当物理存储空间添加到一个 LUN，现有的数据存储可以在不关闭服务器或者相关联的存储的情况下扩展。这补充了允许动态 LUN 扩展的存储阵列选项。扩展虚拟磁盘（vmdk）的存储空间也可以使用 VMDK 热扩展（Hot VMDK Extend），在没有快照的情况下，以持续模式进行。建议将扩展放在具有相同性能的磁盘上。

Vmdk 扩展和磁盘可用空间的可见性取决于 OS 的机制及其文件系统。根据 OS 版本，可能需要第三方工具来扩展系统分区，Windows 2003 就是这种情况。更多的信息参见 VMware 知识库：KB1004071。

3.4.7 可以创建单个 64TB 卷来保存所有 VM 吗

vSphere 5 中 LUN VMFS-5 的最大大小为 64TB。从理论上讲，可以创建一个非常大的 64TB 卷。因为存储阵列集成了 VMware API（VAAI），它们提供出色的卷访问性能。但是，我们不建议采用这种方法，原因如下：

- 隔离环境绝对是不可或缺的，生产、测试、接收和备份应该有自己的专用环境和 LUN。重要的是不要混合不同的 I/O 配置（例如随机和顺序访问），也不要根据 VM 活动实施负载均衡（尽管存储 DRS 允许负载均衡）。
- 在迁移期间，大型卷的迁移比多个小型卷的迁移更复杂，因此可能分成多个阶段进行。
- 大型卷损坏时，影响比包含较少 VM 的小型卷更加明显。

因为上述原因，创建多个独立的 LUN 是首选的方法。它也使得复制更简单（例如，允许只对关键环境进行保护）。

3.4.8 VMFS 配置最佳实践

建议采用如下最佳实践：

- 一般来说，应该创建 600GB 到 1TB 之间的 VMFS 卷，每个卷使用 15 至 20 个活动的 vmdk（不要超过 32 个）。（一个 VM 可能有多个活动 vmdk）
- 对于需要高性能的环境，如 Oracle、Microsoft SQL 和 SAP，最好使用 RDM 模式。
- VMware 建议在 NFS 上使用 VMFS，因为 VMFS 提供完整的功能集，可以为 I/O 密集应用使用 RDM 卷。
- 为了避免争用，不要在一个 LUN 上连接超过 8 个 ESX 服务器。
- 避免在同一个 VMFS 上放置多个具有快照的 VM。
- 避免将 DRS 定义为激进（aggressive），因为这会频繁触发 VM 从一个主机服务器迁移到另一个主机服务器，从而频繁发生 SCSI 保留。
- 将生产 LUN 与测试 LUN 分开，在专用的 LUN 上存储 ISO 文件、模板和备份。
- 在为新磁盘配置 OS 后对齐 vmdk 分区。
- 避免集合多个 LUN 以组成一个 VMFS，因为不同环境无法分离（生产环境、测试环境和模板），这会增加争用的风险，以及更频繁的保留。
- 避免为每个 VM 创建一个 VMFS，因为这会增加 LUN 数量，使管理更加复杂，同时限制了 256 个 LUN 或者 256 个 VM 的扩展。

3.5 虚拟磁盘

和传统硬盘一样，虚拟磁盘包含 OS、应用和数据。VM 的虚拟磁盘由一个 vmdk 文件或者一个 RDM 卷表示。

3.5.1 VMDK

vmdk 是最重要的文件，因为它们是 VM 的虚拟磁盘，所以必须受到安全保护。在 vSphere 5 中，vmdk 的最大大小为 2TB（更准确地说是 2TB 加上 512 个字节）。虚拟磁盘包

括两个文件：一个扩展名为 .vmdk 的描述符，和一个包含数据、扩展名为 -flat.vmdk 的文件，你可以在命令行接口（见图 3-10）或者图形用户界面（见图 3-11）中看到。

- vmdk 文件对应于一个元数据文件，这是虚拟磁盘的说明（某些维护支持所需要的可编辑文件）。这个文件提供了指向 -flat.vmdk 文件的链接，并包含关于 UUID 的信息。（参见本章前面的 3.4.3 节）。
- -flat.vmdk 文件对应于虚拟磁盘及其内容。

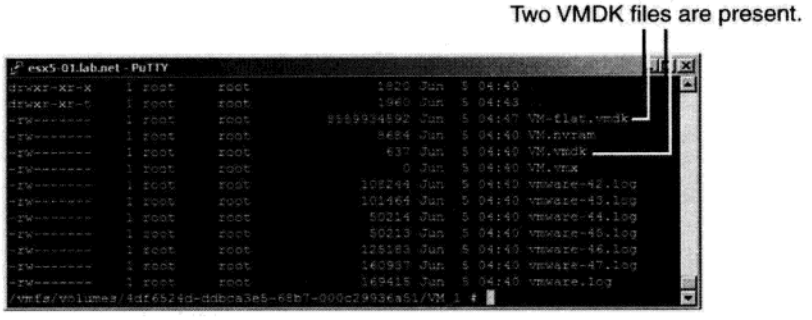


图 3-10 命令行接口显示 vmdk 文件

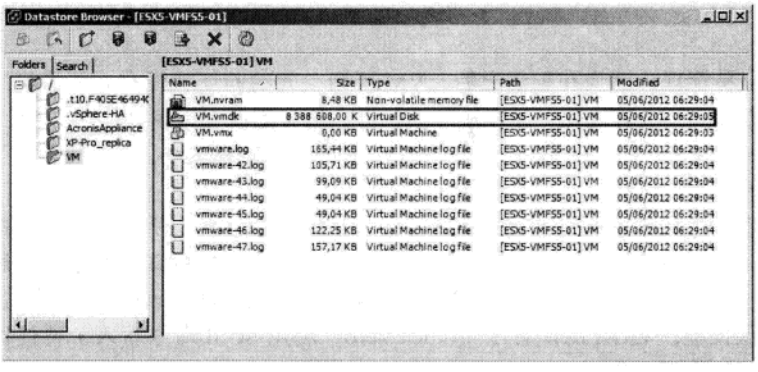


图 3-11 vCenter 的 GUI 显示一个 vmdk 文件，与虚拟磁盘大小相同

3.5.2 磁盘类型

创建 VM 时，可以使用如下磁盘类型：厚盘（thick disk，延迟置零或者置零）或者精简盘（thin disk），其选项参见图 3-12。表 3-2 比较了这些磁盘类型的优点。

表 3-2 磁盘类型和各自的优点

	好 处	何时使用
延迟置零厚盘	创建较快，但是第一次写入性能较低	创建 VM 时的标准选项
置零厚盘	创建时间较长，但是第一次写入性能较好	复制 VM 或者从模板部署 VM 时使用这一模式
精简盘	创建很快，但是写性能不如其他模式	NFS 数据存储默认使用这一模式

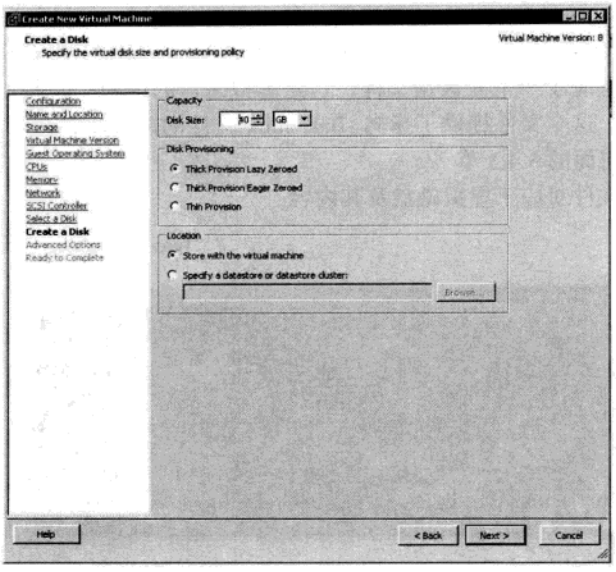


图 3-12　可选的磁盘类型

1. 厚盘

厚盘更容易管理，因为在配置之后，VM 可用空间的验证就没有必要进行了。但是，这意味着由于磁盘空间没有优化，存在额外的代价。这类磁盘支持容错（FT）特性。

在厚盘中，vmdk 文件的大小等于创建 VM 时配置的磁盘大小。

厚盘有两种格式。

- 延迟置零（lazy zeroed 或者 zeroed）：这是默认的格式。所有磁盘空间都被分配，但是原来在磁盘级别上写入的数据不被删除。存储空间中的现有数据不被删除而是留在物理磁盘上。擦除数据和块置零（格式化）只在第一次写入磁盘的时候进行，这会稍微降低性能。VAAI 的块置零（block zero，利用 SCSI 命令写入）特性极大地减轻了这种性能降低现象。
- 置零（eager zeroed）：所有磁盘空间被保留；数据完全从磁盘上删除，磁盘创建的时候就进行块置零（格式化）。创建这样的磁盘花费更长的时间，但是因为以前的数据被删除而增强了安全性。与延迟置零厚盘比较，它在写入磁盘的时候性能要好得多。

对于需要高性能的应用建议使用厚盘格式。使用这个模式的简单方法之一是在配置 VM 磁盘的时候选择容错等群集支持特性。

创建新的 VM 总是比复制或者部署模板更快。

2. 精简盘

一些研究表明，40% ~ 60% 的磁盘空间在分配之后从未使用过。使用精简盘选项（称为精简配置）时，VMFS 上保留的空间等于磁盘上实际使用的空间。这个空间的大小动态增加，

存储空间得以优化。

示例： 创建一个 20GB 文件，但是只使用 6GB。
在精简盘中，vmdk 文件占用的存储空间为 6GB，而使用厚盘，vmdk 文件使用 20GB 存储空间。

在这种模式下性能低下，因为空间在请求时动态分配，磁盘块需要置零。精简盘对避免存储空间浪费有帮助，但是需要特别小心管理，以确保存储空间不会短缺。Out of Space API 允许主动监控和告警，可以避免这种情况发生。

注意： 在实施复制的时候精简 LUN 非常有用，因为第一次同步只复制磁盘上使用的数据。对于厚配置的 LUN，所有数据都必须被复制（即使数据块为空）。用精简配置的 LUN 进行的初始同步工作量大大减小。

注意： 避免将基于存储阵列的精简配置与精简模式的 vmdk 磁盘组合使用，因为分清不同的磁盘变得很困难，很容易出现解读错误。

你可以使用如下任何一种方法将磁盘从精简配置转换为厚盘：

- 使用 Datastore Browser 中的 Inflate 选项。
- 使用 Storage vMotion 将磁盘类型修改为厚盘，如图 3-13 所示。

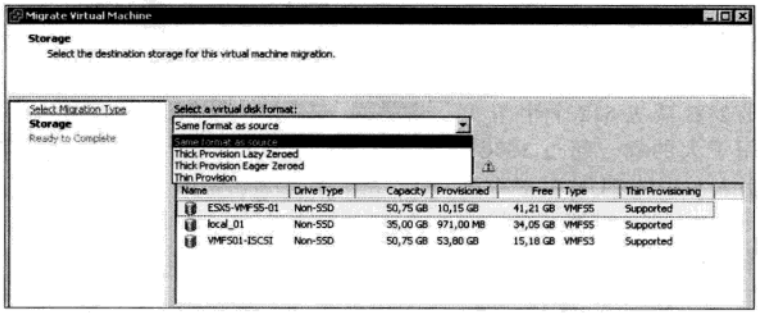


图 3-13 使用 Storage vMotion 界面修改磁盘类型

3. 模式

虚拟磁盘有三种模式：

- 独立持久（independent persistent）：VM 的所有磁盘写入都实际写入磁盘（在 vmdk 文件中）。即使重启，修改也被保留。这种模式提供最佳的性能。
- 独立非持久（independent nonpersistent）：VM 启动之后进行的所有更改在关闭时全部被撤销。修改被写入一个文件，记录 VM 文件系统级别的所有更改。在这种模式下，重启 VM 意味着回到参考 VM。性能不是很好。
- 快照（snapshot）：这种模式能够返回前一个状态。

注意：遵循安全规则和相关的最佳实践，避免非持久磁盘。当 VM 重新启动，非持久磁盘无法分析日志，因为一切都回到初始状态，这会在安全问题出现时阻碍调查和更正措施。

3.5.3 原始设备映射

使用原始设备映射（Raw Device Mappint，RDM）格式，可在 ESX 服务器中引入原始存储卷。这种模式主要用于如下情形：

- 使用 Microsoft 群集时（MSCS 或者 Windows 2008 Server 下的 Windows 服务器故障切换群集，这是唯一支持的模式）
- 采用基于阵列的快照时
- 为了高性能（数据库类型）而直接在 VM 中引入卷时
- 为 VM 引入大型 SAN 卷时（从 300TB 起），避免长 P2V 卷转换到 vmdk

RDM 采用存储在 VMFS 数据存储中的一个文件（指针的一种类型）的形式，作为 LUN 卷的代理。

图 3-14 说明了 vmdk 和 RDM 之间的不同。

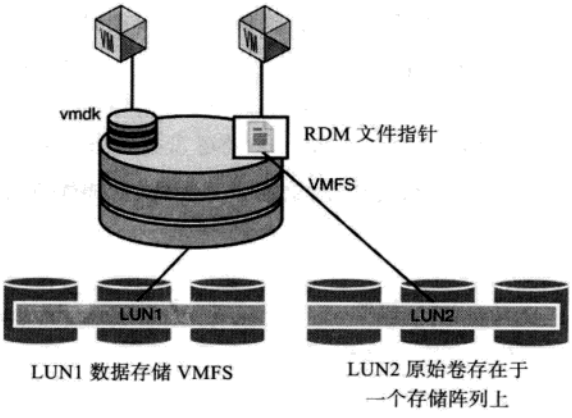


图 3-14 vmdk 与 RDM 格式的对比

RDM 格式以两种模式存在：RDMv（虚拟兼容模式）和 RDMp（物理兼容模式）。

1. RDMv 磁盘

RDMv 磁盘的最大大小为 2TB（准确地说是 2TB 减去 512 个字节）。RDMv 主要用于大的卷。超过 300GB 时，为 VM 引入专用的 LUN 可能很有趣。确实，vmdk 是一个很容易被移动的文件，但是当它很大时，移动可能也更加复杂。在这种情况下，较好的做法是引入原始卷，并使用存储阵列功能移动卷。

RDMv 在 VMFS 上创建一个文件，作为 VMFS 和直接与 VM 连接的 LUN 之间的代理。这使虚拟化管理器能够拦截 I/O 并在需要时进行记录。RDMv 授权 VM 快照（但不是存储阵列快照）以及复制及模板的创建。

2. RDMp 磁盘

RDMp 磁盘的最大大小为 64TB。这种类型的磁盘不允许虚拟化管理器拦截 I/O。这意味着无法采用 VM 快照（但是可以实现基于阵列的快照），也不可能创建复制或者模板。

一般来说，RDMp 磁盘用于通过存储阵列快照功能，将生产数据库上的数据引入到测试服务器中，也可以用于 MSCS 群集。使用 MSCS 时，共享磁盘不能共享 OS 的虚拟控制器。

有些公司对将应用迁移到虚拟环境犹豫不决。利用 RDMp，这种改变可以缓慢而可靠地

完成，因为如果虚拟环境中的测试无法令人信服，公司可以自由地返回物理环境。对于虚拟环境没有正式支持的应用（例如旧版的 Oracle），RDMp 可以用于提供简单的方法，复制软件发布者支持的物理环境中的问题。

RDMp 磁盘无法像传统 VM 那样备份。两种模式提供的功能见表 3-3。

表 3-3 RDMv 和 RDMp 磁盘的对比

RDM 类型	vMotion	Storage vMotion	文件名	VM 快照	存储阵列级快照
Rdmv	是	是	rdm.vmdk	是	不建议
Rdmp	是	是	rdmp.vmdk	是	是

3.5.4 OVF 格式

当前的虚拟磁盘格式是 vmdk（VMware 使用）和虚拟硬盘（Virtual Hard Disk，vhd，Microsoft Hyper-v 和 Citrix XenServer 使用）。

开放虚拟机格式（Open Virtual Machine Format，OVF）不是一种虚拟磁盘格式；它是一种文件格式，其特性方便了各种虚拟化和虚拟化管理器平台之间的互操作性。OVF 文件包括参数和虚拟硬件设置、先决条件及安全属性等元数据。OVF 包不仅限于一个 VM，可以包含多个 VM。OVF 文件可以加密和压缩。

OVF 模板由如下文件组成。

- MF：一个清单文件，负责验证 OVF 模板的完整性，确定其是否被修改。
- OVF：一个 XML 文件，包含虚拟磁盘的有关信息。
- vmdk：VMware 中的一个虚拟磁盘，但是这个文件可以使用不同的格式，以便提供虚拟化管理器的互操作性。VMware 规范允许不同类型的虚拟磁盘。

注意： 为了简化 OVF 文件导出中各个项目的移动和操纵，可以使用开放虚拟化用具（Open Virtualization Appliance，OVA）格式，将多个文件组合成一个文件。OVA 文件等同于 TAR 文件，实际上可以改名，使用 .tar 扩展名来代替 .ova 扩展名，以便用于典型的存档应用。

你可以从 http://solutionexchange.vmware.com/store/category_groups/19 下载包含 OVF 操作系统和应用解决方案的预配置虚拟用具。

3.6 数据存储

在 VMware 中，存储空间被称作数据存储（datastore）。数据存储是保存 VM、模板或者 ISO 映像的存储资源的虚拟表现形式。数据存储隐藏了不同技术和存储解决方案的复杂性，为 ESX 服务器提供一个统一的模型，无须考虑实现的存储类型。数据存储的类型有 VMFS 和 NFS。

注意： VMware 的最佳实践建议适当地将用于存储模板或者 ISO 映像的数据存储与用于 VM 的数据存储分离。我们还建议监控数据存储的可用空间。应该始终有至少 25% 至 30% 的可用空间。这些空间对快照或者备份操作以及 VM 交换来说是必需的。缺乏空间可能造成严重的后果，并且可能影响虚拟环境的总体性能。

数据存储群集又称数据存储池（Pool Of Datastore，POD）是一组集合起来形成单个实体的数据存储，如图 3-15 所示。创建数据存储群集时，可以使用存储 DRS。

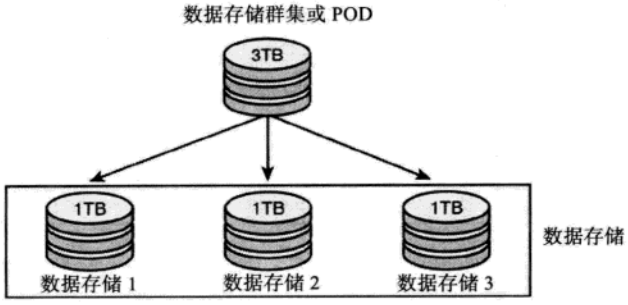


图 3-15　数据存储池（POD）

数据存储群集可能包括来自不同存储阵列（从性能和容量的角度看）的卷，而且可以混合不同的 VMFS（VMFS-3 和 VMFS-5），但是通常不建议这么做。在数据存储群集中不支持混合 VMFS 和 NFS 卷。

3.7　Storage vMotion

Storage vMotion 允许不同存储空间之间 VM 虚拟磁盘的热迁移。组成 VM 的所有文件从一个数据存储迁移到同一个存储阵列或者不同存储阵列中的另一个数据存储不会造成服务中断。存储阵列可以来自不同的制造商。

注意： vMotion 从一个物理服务器上将 VM 迁移到另一个服务器，但是不移动组成 VM 的文件。Storage vMotion 移动虚拟磁盘。这两个操作不能同时在同一个 VM 上进行，除非关闭这个 VM。

3.7.1　何时使用 Storage vMotion

Storage vMotion 用于存储阵列的预防性维护操作，对于购买新的存储阵列也可能很有用，因为它不需要服务中断。迁移很容易以完全透明的方式进行。这将管理员从这一在传统物理环境中常常很麻烦和敏感的任务中解放出来。Storage vMotion 允许管理员更换存储阵列制造商和迁移 VM，而不需要复杂的兼容性矩阵。

注意： 在存储级活动很少时使用 Storage vMotion 是首选。用 Storage vMotion 进行迁移之前，必须确认源和目标 ESXi 服务器之间有足够的存储带宽。

3.7.2　Storage vMotion 的工作原理

vSphere 5 中对 Storage vMotion 进行了一些改进。过去曾经使用多种技术。在 vSphere4.1 中，用脏数据块跟踪（Dirty block tracking）在源和目标之间复制磁盘数据块：

全复制，然后仅向目标发送修改过的块。（脏数据块跟踪是变更数据块跟踪模式的一种形式）这种技术的问题是切换到目标 VM 的时间以及源 VM 中大量 I/O 负载下的故障风险。在 vSphere 5 中，如图 3-16 所示，Storage vMotion 建立 VM 的全复制，然后使用一个镜像驱动程序在源和目标 VM 之间分离写入修改的数据块。

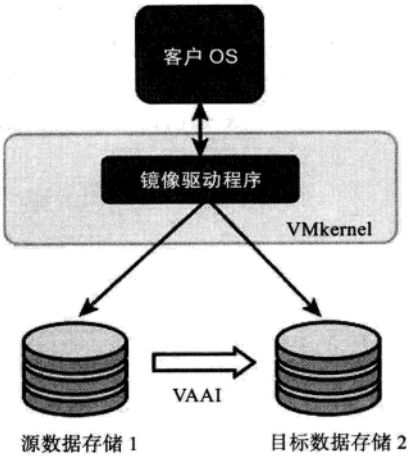


图 3-16 Storage vMotion 使用一个镜像驱动程序

I/O 镜像对于连续的磁盘复制是首选的方法，因为它的优点是即使在目标 VM 速度缓慢的情况下也能保证迁移成功。迁移将更简短，更可预测。

使用 Storage vMotion 时会发生如下现象：

1）VM 的工作文件夹被复制到目标数据存储。

2）VM 的一个映像（称作影子 VM，shadow VM）使用复制的文件在目标数据存储上启动。影子 VM 处于暂停状态。

3）Storage vMotion 激活一个驱动器（称作镜像驱动器，mirror driver）将已经复制的数据块镜像写入目标数据存储。

4）目标数据存储的 VM 磁盘文件复制完成，同时 I/O 被镜像。

5）Storage vMotion 暂停源 VM 并将正在执行的源 VM 传送给影子 VM。

6）旧的文件夹和 VM 磁盘文件被从源数据存储中删除。

注意：原始文件只在目标文件被正确写入且发送了确认消息之后才被删除，确保了操作的成功。

Storage vMotion 在企业版中可用，可以用于具有快照的 VM，也支持链接复制的迁移。

3.8 存储 DRS

存储 DRS（SDRS）能够自动化选择 VM 使用的数据存储，有利于更平衡的性能和更有效的存储空间利用。这节约了管理员的时间，他们不再需要花费时间选择所用的数据存储。为此，数据存储被集中到数据存储群集中。

SDRS 负责如下操作：

- VM 初始定位
- 根据如下因素在数据存储之间均衡负载：
  - 存储空间的使用
  - 根据延时确定的 I/O 负载

初始定位发生在 VM 创立、移动或者复制的时候。根据集群数据存储的已用空间和 I/O

负载，SDRS 提供特定的数据存储来保存 vmdk。

3.8.1 数据存储负载均衡

负载均衡每两个小时根据已用空间，每 8 个小时根据最近 24 个小时的历史数据确定的 I/O 负载进行。如图 3-17 所示，在数据存储超过用户定义的磁盘已用空间（默认为 80%）和 I/O 延迟（默认为 15 毫秒）阈值时，SDRS 做出迁移建议。

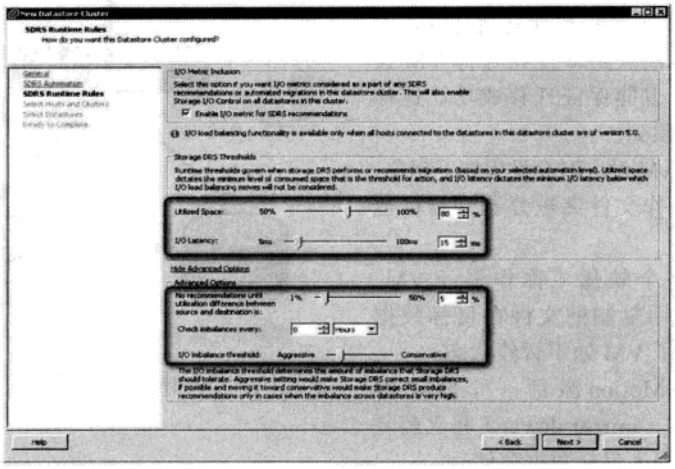


图 3-17 SDRS 界面显示负载均衡信息

自动化有多个级别：

- ❑ 手工（默认）
- ❑ 自动化
- ❑ 规划（定时）。例如，规划模式在备份期间很有趣，它不必移动虚拟磁盘，因而可以在备份操作期间禁用 SDRS。
- ❑ 数据存储维护模式。数据存储的维护模式从数据存储中删除所有 vmdk，并将它们分布到群集的其他数据存储中。

这时，你可以问，“SDRS 如何检测数据存储 I/O 负载？”

SDRS 使用 SIOC 功能和注入器机制选择最佳的目标数据存储。注入器用于随机地“注入”I/O 以确定每个数据存储的特征，这能确定与每个数据存储的响应时间和延时。

3.8.2 亲和性规则

如图 3-18 所示，可以应用多种亲和性规则。

- ❑ VM 内 vmdk 亲和性：所有相同的 vmdk VM 被放在同一个数据存储中。
- ❑ VM 内 vmdk 反亲和性：这条规则可以用来确保 vmdk 被放在不同的数据存储上。这条规则很有用，例如，这条规则可以用于分离数据库 VM 的日志磁盘和数据磁盘。该规则适用于 VM 中的所有或者部分磁盘。

- VM-VM 反亲和性：不同 VM 被放在不同的数据存储中。这提供了 VM 在数据存储故障时的冗余性。

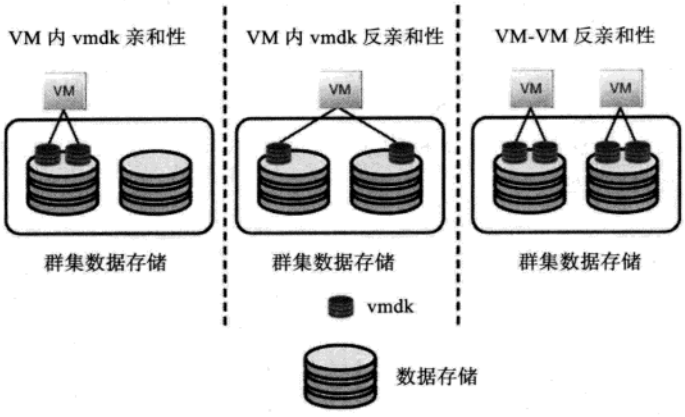


图 3-18 亲和性规则

SDRS 目前有如下限制：

- SRM 不支持 SDRS。
- SDRS 只能用于 ESXi5 或者更高版本的主机。

3.8.3 配置驱动存储

配置驱动存储（Profile-driven storage）维护 VM 和规定存储需求的相容性。这种功能消除了初始定位错误，通过自动化简化了管理员的日常管理工作。管理员建立包含存储特征的配置文件。这些配置可以使用 vSphere Storage *存储检测* API（VASA）实施，或者与用户定义的指标关联（例如，Gold、Silver、Bronze）。

VM 配置文件在部署、创建、迁移、复制等期间使用。如果 VM 被放在提供存储配置文件定义的容量的存储空间中，这个存储就是相容的。配置驱动存储补充了 SDRS 的初始定位和 vmdk 的自动化迁移。

3.9 存储 I/O 控制

资源共享带来了新的挑战。非关键 VM 不应该独占可用资源。磁盘共享只解决了一部分问题，因为共享只在和单个 ESXi 主机相关时建立，只在 ESXi 主机级别的争用发生时使用。后一种情况是不相干的，因为位于另一个 ESXi 上的 VM 可以使用较大而优先级较低的共享。图 3-19 说明了使用和不使用*存储 I/O 控制*（SIOC）的存储共享。

为了有效地管理 I/O 资源分配，它必须独立于 VM 的位置。这一问题必须通过在 ESXi 群集级别上共享数据存储的访问资源来解决。这就是 SIOC 的作用，它在群集级别而不是 ESX 级别上实现共享服务质量（QoS）。SIOC 监控数据存储的 I/O 延迟。当延迟达到阈值

（默认设置为 30 毫秒），数据存储出现拥塞，SIOC 加以干预，根据每个 VM 定义的共享规则分配可用资源。较低优先级的 VM 精简 I/O 队列。当且仅当在访问数据存储时出现存储 I/O 争用，则发生共享。使用 SIOC 确保了最重要的 VM 在任何情况下（即使出现拥塞）都有足够的资源。

使用这种与 VM 数据存储访问相关的 QoS，管理员可以放心地整合环境。即使在高活动率的情况下，最关键的 VM 仍然有必要的资源。

SIOC 可以在数据存储的属性对话框中激活（见图 3-20）。注意，目前 SIOC 不支持多扩展的数据存储和 RDM 磁盘。

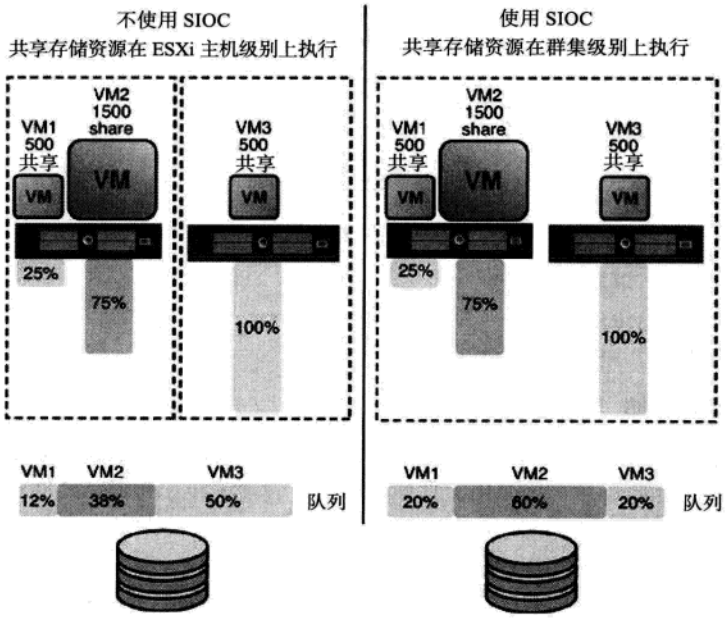


图 3-19　使用和不使用 SIOC 的存储

VMware 建议对不同磁盘类型使用不同的阈值。

- ❑ 光纤通道：20 毫秒至 30 毫秒
- ❑ 串行连接 SCSI（SAS）：20 毫秒至 30 毫秒
- ❑ 固态驱动器（SSD）：10 毫秒至 15 毫秒
- ❑ 串行 ATA（SATA）：30 毫秒至 50 毫秒

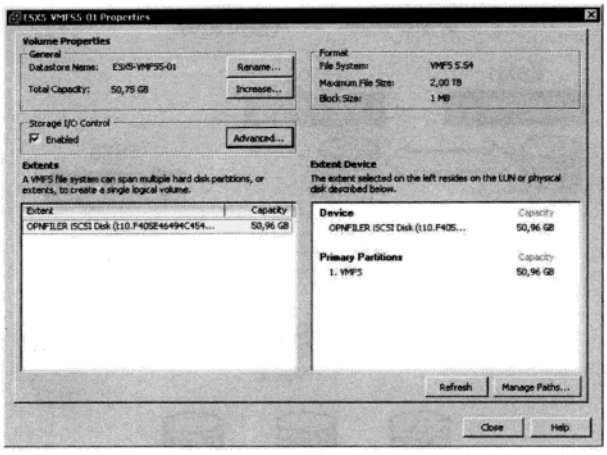


图 3-20 数据存储属性对话框中启用了 SIOC

3.10 vSphere Storage Appliance

vSphere Storage Appliance（VSA）是为中小型企业（SMB，20 至 35 个 VM）设计的一种用具，允许通过使用 HA、DRS、DPM、FT 和 vMotion 等高级特性，在更低的成本下访问共享存储。这种用具只能用于 vSphere 5，以 VM 的形式部署在每个 ESXi 服务器上（以 3GB 的 OVF 文件形式分发）。VSA 占用 ESXi 服务器本地磁盘的可用空间，并显示一个由 ESXi 服务器复制的 NFS 卷。

在另一个 ESXi 服务器上复制本地存储，确保了主机服务器退出服务时的冗余。当一个 VSA 节点退出服务时，VSA 管理器将 IP 地址和共享存储切换到复制的 VSA。这一步不需要中断数据存储 VM 的服务。

VSA 支持群集中的 2 至 3 个 ESXi 服务器，在两节点配置中最多可支持 25 个 VM，三节点配置下最多可配置 35 个 VM。

因此，VSA 有两种部署配置：使用一个 VSA 的两台 ESXi 服务器以及安装在 vCenter 上的 VSA 群集服务，或者如图 3-21 所示，使用一个 VSA 的三台 ESXi 服务器。

VSA 管理器（作为 vCenter Server 中安装的插件）是 VSA 群集的管理界面。它能够监控群集的状态，进行维护和 VSA 节点替换操作。

VSA 用具有如下最低需求：

- 6GB RAM
- 4、6 或者 8 个相同的磁盘（相同大小和特征），配置为 RAID 5、RAID 6 或者 RAID 10
- 4 个 1GB 的网卡
- 物理交换机上配置两个 VLAN

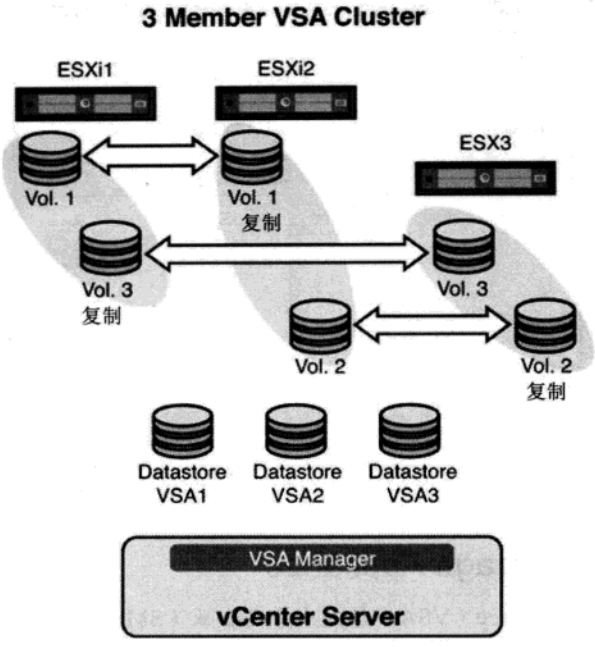


图 3-21 VSA 部署配置

因为 vCenter Server 提供 VSA 管理，所以它必须放在 VSA 群集之外，在群集外的一个 VM 上或者在专用的物理服务器上。注意，这是 VMware 建议将 vCenter Server 放在物理服务器上的唯一情况。

安装 VSA 相当简单和快速（大约 10 分钟），不需要特殊的存储技能。

3.11 VMware 存储 API

VMware 提供的 API 允许管理员和发布者扩展 vSphere 5 功能。

3.11.1 vStorage API for Array Intergration

vStorage API for Array Intergration（VAAI）是一组应用编程接口，提供 VMware 和存储阵列制造商之间的互操作性，以更智能的方式与 VMware 通信。有些任务负载可以转移到存储阵列，减轻 ESXi 主机的负载。

注意： 处理器制造商已经在芯片中集成了 Intel VT 和 AMDV 指令，减少高消耗的 CPU 侦听。处理器制造商对服务器所做的正是 VAAI 对存储阵列所做的。这些 API 现在对于获得高级别的整合似乎是必不可少的。

表 3-4 列出了 vSphere 4.1 中的 VAAI 和 vSphere 5 中的 VAAI2。

表 3-4 VAAI 功能：vSphere 4.1 和 vSphere 5

VAAI vSphere 4.1	VAAI2 vSphere 5
块	
硬件辅助锁	空间用尽
硬件加速置零	空间回收
硬件加速复制	
NAS	
不可用	全复制 扩展统计 空间保留

下面是表 3-4 中列出的各种特性的简单说明。

- 硬件加速锁：没有这个 API，SCSI 保留就会在全局 LUN 级别上完成。有了这个 API，SCSI 保留工作在块级别而不是 LUN 级别上完成，这样与 SCSI 保留相关的问题较少，而且减少了 VM 启动的时间，在虚拟桌面基础架构（VDI）项目中更是如此。
- 硬件加速置零：没有这个 API，创建数据存储时，“置零”由服务器完成，服务器向存储阵列发送 SCSI 命令。有了这个 API，ESX 服务器初始化一个命令，存储阵列负责重复这个操作并在结束时通知 ESX 服务器。这减少了 ESXi 服务器和存储阵列之间的流量。
- 硬件加速复制：没有这个 API，复制操作从 ESX 服务器向存储阵列进行。有了这个 API，数据由存储阵列的阵列中移动，没有通过服务器。这减少了 ESXi 服务器的负载和数据迁移所需的时间。

在 vSphere 5 中，为 VAAI 2 定义了一些新的概念：

- 死空间回收（Dead Space Reclaim）：当虚拟磁盘被删除，或者精简配置 LUN 上使用 Storage vMotion 将一个虚拟磁盘从数据存储中迁移到另一个数据存储之后，可以恢复不再使用的空间。ESXi 5.0 通过 VAAI 命令将释放数据块的有关信息发送给存储系统，然后存储系统恢复这些数据块。
- 精简配置空间用尽（Thin Provisoning Out of Space）API：预防精简配置 LUN 上的存储空间问题。
  - 精简配置 LUN 报告：在 vCenter 中可以识别使用的存储阵列。
  - 超过限额：当数据存储中超过容量阈值时，在 vCenter 中显示警告。
  - 空间用尽行为：VM 在写入之前确定空间是否足够。如果存储空间已满，在 vCenter 中显示警告信息，然后 VM 暂停（其他 VM 继续运行）。

NAS VAAI 存储定义了如下概念。

- 全文件复制：NAS 可以冷方式进行 vmdk 文件的复制和快照操作，类似于 VMFS 块复制（全复制）。
- 扩展统计：可以看到 NFS 数据存储上已经消耗的空间。
- 空间保留：允许为 NAS 存储创建厚配置模式 vmdk 文件。

3.11.2 vSphere 存储 API：存储感知

vStorage API for Storage Awareness（VASA）是一个存储检测 API，可直接从 vCenter 进行存储阵列相关信息的虚拟化，这些信息包括复制、RAID 类型、压缩、重复数据消除、精简或者厚格式、磁盘类型、快照状态和性能（IOPS/MBps）。此外，vStorage API 可用于配置驱动存储。

3.12 多路径

多路径（multipathing）可以定义为使用冗余组件（如适配器和交换机）以创建服务器和存储设备之间逻辑路径的一种解决方案。

3.12.1 可插入存储架构

可插入存储架构（Pluggable Storage Architecture，PSA）是一组 API，允许存储制造商在 VMkernel 层中直接插入代码，从而开发第三方软件（如 EMC PowerPath VE），提供与存储阵列技术直接相关的更高级负载均衡功能。但是 VMware 也提供标准的基本多路径机制，即原生多路径（Native MultiPathing，NMP），这些功能分布在以下 API 中：存储阵列类型插件（Storage Array Type Plug-in，SATP）负责与存储阵列通信；路径选择插件（Path Selection Plug-in，PSP）提供路径之间的负载均衡。

如图 3-22 所示，VMware 提供三种 PSP。

- 最近使用（Most Recently Used，MRU）：选择 ESXi 启动时发现的第一个路径。如果这一路径不可访问，ESXi 选择替代路径。
- 固定：使用设计为首选路径的专用路径。如果没有配置，则使用启动时发现的路径。这一路径无法再使用时，随机选择可用路径。当该路径再次可用时，ESXi 又会使用固定首选路径。
- 循环（Round Robin，RR）：自动选择所有可用路径，以循环的方式将 I/O 发送到每条路径，这能实现基本的负载均衡。PSA 协调 NMP 操作，第三方软件协调多路径插件（MPP）软件。

图 3-22 VMware 提供的 PSP

NMP 循环路径选择策略有一个 I/O 操作限制参数，控制每条路径切换到下条路径之前发送的 I/O 操作数量。默认值为 1000，因此，NIP 默认在向给定路径发送 1000 次 I/O 之后转向另一条路径。调整循环路径选择 I/O 操作限制，能够显著地改进某种工作负载下的性能（例如联机事务处理 [online transaction processing，OLTP]）。在随机和 OLTP 工作负载环境中，将循环路径选择参数设置为较低的数字可以得到最好的吞吐率，但是对于顺序工作负载，降低该值不会得到同样显著的改进。因此，有些硬件存储公司建议将 NMP 循环路径选择 I/O 操作限制参数设置为较低的值（可以设置为 1）。

第三方软件解决方案使用更高级的算法，因为循环选择算法有一个局限性，在进行自动分配的时候没有考虑路径级别上的实际活动。有些软件建立动态负载均衡，设计为在任何时候都使用所有路径，而不是像循环路径算法那样，在同一时间仅用一条路径来负担所有 I/O 负载。

3.12.2 模式

访问共享存储空间的数据是虚拟环境的基础。VMware 强烈建议实施多种 LUN 访问路径。最小值是两条路径，但是 VMware 建立使用四条路径。多路径提供冗余的 LUN 访问路径，从而减少了服务中断。路径不可用时，会使用另一条路径，这不会造成服务中断。这些切换机制被称作多路径 I/O（MultiPath I/O，MPIO）。

在 VMware 中，如图 3-23 所示，存储可以采用不同的模式。

- 主动 / 主动：在给定时刻，一个 LUN 同时连接到多个存储控制器。I/O 可以同时来自多个控制器。
- 主动 / 被动：在给定时刻，一个控制器拥有一个 LUN（从属 LUN）。只要 LUN 链接到该控制器，其他控制器就不能向它发送 I/O。
- ALUA：对 LUN 的访问不是直接的（无优化的），而是通过辅助控制器以不对称方式发生。

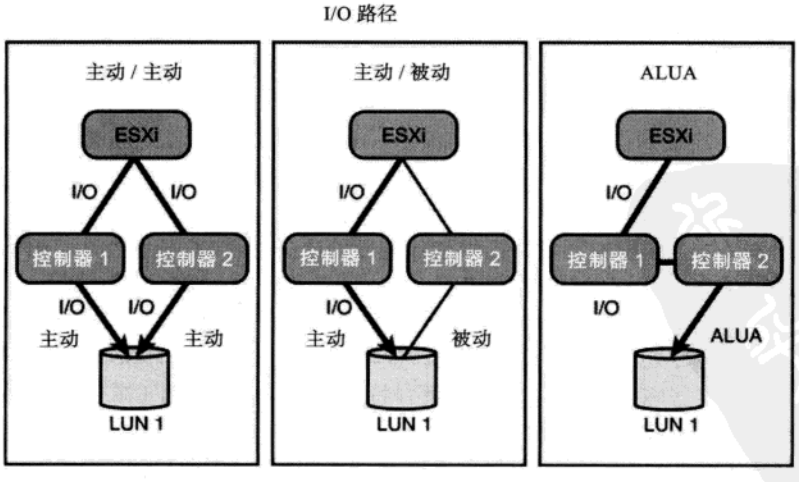


图 3-23 存储模式

3.13 磁盘技术考虑因素

本小节研究在决定环境中使用的磁盘技术时需要考虑的因素。

3.13.1 支持的磁盘类型

正如你已经看到的，存储架构很重要，磁盘技术扮演着重要的角色。ESXi 支持各种磁盘，包括 SSD、SAS、FC、SATA、NL-SAS、IDE、USB 和 SCSI。

可用选项很多，可以根据多种标准选择技术。如表 3-5 所示，从磁盘技术的角度，可以考虑许多参数：速度可以表现为每分钟转速（RPM）、每秒 I/O 次数（I/O per second，IOPS）和传输带宽。

表 3-5 各种磁盘类型的平均速度参数（可能变化）

磁盘	RPM	IOPS
SSD	N/A	3000
SAS	15K	180
SAS	10K	130
NL-SAS	7.2K	100
SATA	5.4K	50

固态盘（SSD）是由闪存组成的高性能盘片，它们是非机械性的，比较不容易遭遇故障，消耗的电力和发热量都远小于传统磁盘。它们的访问时间很短，具有很高的 IOPS（3000IOPS），对于读取操作很理想，但是对于大量写入操作并不是很合适。

这种存储通常用于日志文件（例如数据库日志）。它们通常用于扩展控制器的缓存。（EMC 称作为快速缓存盘，而 Netapp 称作闪存缓冲盘）。在 VMWare 环境中，这些高性能盘对于存储 VM 的交换内存很理想。它们在磁盘活动波峰出现的时候对负载的吸收很有用，例如，在 VDI 环境中，所有 VM 同时启动的时候。（这种现象称为启动风暴，boot storm。）磁盘大小通常为 100GB、200GB 和 400GB。800GB 容量很快也将出现。

串行连接 SCSI（Serial Attached SCSI，SAS）磁盘用于替换光纤通道磁盘。这些磁盘以点对点的方式直接连接到控制器。转速很高——10 000RPM 或者 15 000RPM。它们对于随机访问应用很理想，且能够处理 8 字节至 16 字节的小规模 I/O（通常是数据库）。数据流是双向的，目前的磁盘大小为 300GB、600GB 和 900GB。

目前，SAS 最适合于虚拟环境，提供最佳的性价比。尽管 FC 磁盘在生产环境中仍然随处可见，但是用 SAS 替代它们是一个趋势。

近线 SAS（Near-Line SAS，NL-SAS）磁盘使用安装在 SAS 接口上的 SATA 磁盘机制。它们与 SATA 相比的好处是全双工数据传输，读写可以同时进行，而 SATA 同时只能进行一次读或者写操作。这些磁盘提供 SCSI 命令解释特性，例如命令队列（减少读磁头移动），并且提供比 SATA 更好的错误和报告控制。

串行 ATA（Serial-ATA，SATA）磁盘能够管理大的容量——2TB、3TB，很快可以达到 4TB。建议将它们用于大文件的顺序传输（例如备份和视频文件），但是不适合于有大量随机

I/O 的数据库应用（如 Oracle、Microsoft SQL、MySQL）。ATA 是单向传输的，同一时间只能进行一次读或者写操作。SATA 是否适合于关键生产 VM 取决于存储阵列制造商，可以从制造商那里得到建议。SATA 磁盘适合于测试 VM 或者 ISO 映像、模板或者备份存储。

3.13.2 RAID

表 3-6 列出了建议的 RAID 类型及相关的传统用途。

表 3-6 RAID 类型和传统用途

	写	读	用 途	保 护
RAID0	优秀	优秀	实时工作站	无（条块）
RAID1	优秀	优秀	DB 日志文件、操作系统、ESXi 虚拟管理器	镜像
RAID5	好	很好	DB、ERP、web 服务器、文件服务器、邮件	奇偶校验
RAID6	普通	很好	存档、备份、文件服务器	双奇偶校验
RAID10	优秀	优秀	大数据库、应用服务器	条块 + 镜像

3.13.3 存储池

在物理环境中，一个 LUN 专用于一个服务器，从而专用于特定的应用。在这种情况下，可以设置参数采用适合于应用（顺序或者随机）的 RAID 级别。这种方法不适合于虚拟环境。因为虚拟环境的动态特性，根据应用保持相同的 LUN 归属逻辑变得很难。VM 是多变的，可以从一个数据存储转移到另一个数据存储。RAID 级别无法保持相同。有些制造商建议使用存储池来代替专用的 RAID 级别，这是首选方法，因为它提供了出色的性能并简化了管理。

3.13.4 自动磁盘分层

只有 20% 的 LUN 数据被频繁访问。统计还显示，80% 的数据在两周之后就不被使用。通过自动分层，频繁使用的数据自动地放在高性能的 SSD 或者 SAS 磁盘上，而较少使用的数据存储在 SATA 或者 NL-SAS 等性能较低的磁盘上。

3.13.5 性能

因为资源池和各种层次（例如，应用、虚拟化管理器、存储阵列），在虚拟环境中监控性能很复杂。以 IOPS 度量的速度和以 MBps 度量的带宽取决于数据存储所在磁盘的类型和数量。存储活动应该进行监控，以根据这些标准确定是否形成等待队列（队列长度，queue length）。在虚拟化管理器或者 vCenter 的层面上，识别争用的最可靠及最简单性能指标是设备访问时间。

所有 HBA 的读写访问时间应该低于 20 毫秒。另一个应该监控的指标是停止磁盘（Stop Disk）值，它指出了相关存储无法吸收的活动，从而说明争用的出现。该值应该始终设置为 0。如果它的值高于 0，负载应该重新均衡，这有两种原因。

- 该存储设备上的 VM 活动过多。
- 存储没有正确配置。（例如，确保没有分区问题，到存储的所有路径都可用，活动很好地分布到所有路径上，存储缓冲没有被强制清除）

3.13.6 其他建议

下面是改进磁盘性能的其他建议。

- 使用固态缓存允许大量的磁盘 I/O。这种缓存可以作为一种杠杆，因为大部分的读写 I/O 活动都发生在缓存中。数据库需要很多 4 字节、8 字节、16 字节的随机访问 I/O 操作，而视频文件备份服务器需要高速度和大数据块（32 字节、64 字节、128 字节或者 256 字节）。
- 不应该在同一个磁盘上混合顺序访问和随机访问。如果可能，I/O 应该按照类型分割（读、写、顺序、随机）。例如，3 个存放事务型数据库的 VM 应该各有 3 个数据存储：
- 一个 RAID 5 数据存储用于 OS，分离 OS 意味着 VM 可以在不从 RAID 5 中夺取数据库可用 I/O 的情况下启动。
- 如果读 / 写比例为 70%/30%，采用一个 RAID5 数据存储来存放数据库。否则，RAID 类型应该改变。数据库通常使用 70% 的随机读类型事务。
- 一个用于日志的 RAID 1 数据存储，因为日志的写入是顺序的（对于高写入率的大型数据库可使用 RAID 10）。

3.14 设备驱动程序

图 3-24 显示了 ESXi 提供给客户 OS 的 SCSI 控制器。在可用的选项中，BusLogic 为旧的操作系统提供更好的兼容性。LSI Logic Parallel 在某些操作系统上提供更高的性能，但是必须添加驱动程序。

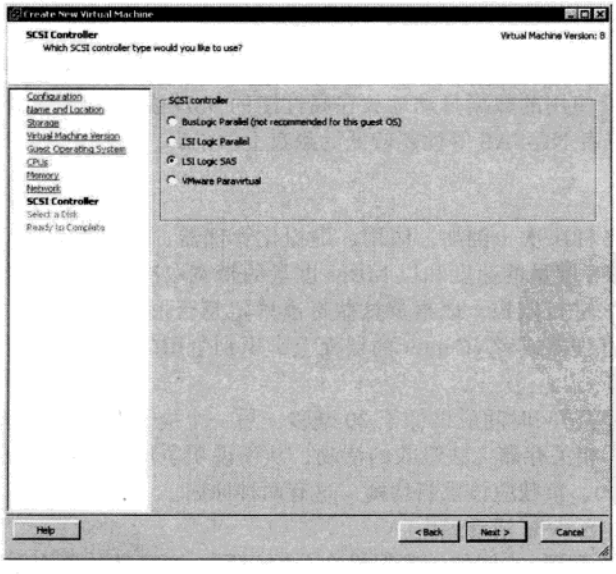


图 3-24 标准 SCSI 控制器选项

VMware Paravirtural 设备驱动程序主要用于大负载的情况。它直接访问虚拟化层。这个驱动程序被称作 Paravirtural（半虚拟）是因为它翻译来自客户 OS 的请求，将它们直接发送给虚拟化管理器。这减少了虚拟机监控器（VMM）的请求拦截，改进了性能。这个选项只能工作于 Windows Server 2003、Windows Server 2008 和 RHEI 5。

另一种可用控制器 VMDirectPath I/O 是一种存储 I/O 驱动程序，允许直接访问物理设备而不用通过 ESXi 虚拟化层，从 VM 中的所有驱动程序原生功能和性能中获益。

3.15 存储是基础

在任何架构中，有一个强大的基础是关键，在虚拟化中更是如此。许多可用选项使得虚拟化在远程办公室或者最高性能的数据中心中都很流行。vSphere 5 提供更多的阵列感知特性，确保虚拟化管理器能够与存储通信，从而为此打下了基础。vSphere 智能地使用这些信息，使 VM 离开性能低下的存储卷。由于选项很多，你可能会认为配置非常困难，但是 vSphere 提供了确保最优配置的工具。

第4章
服务器和网络

大部分 IT 专业人士考虑虚拟化时，想起来的第一件事往往是服务器。服务器的复杂性及其重要性，使其成为首要的考虑，但是服务器上的虚拟机（VM）如果没有网络连通性，通常也无法完成其意图。服务器似乎令人畏缩，但是有了本章提供的信息，你很快就能掌握它。网络常常是网络团队的事情，但是作为虚拟化专业人士，你需要有效地与他们沟通。本章帮助你理解任何虚拟化解决方案中都必需的服务器和网络组件。

4.1 ESXi 服务器

服务器由几千个部件组成，但是在虚拟化中通常只提到三个。在第 3 章中提到的存储是数据和配置所在的地方。内存是正在使用的数据所在的地方，往往也是可用资源中最珍贵的。CPU 是最后一个重要组成部分，ESXi 具有复杂而高效的调度算法，最大限度地提高处理器资源的效率。

4.1.1 内存管理

因为内存通常是最有限的资源，ESXi 可以进行多种资源节约操作，确保你能够最好地利用服务器硬件。ESXi 采用的许多内存管理技术是独有的，使得 VMware 虚拟化管理器能比其他解决方案在有限的内存中进行更多的工作。

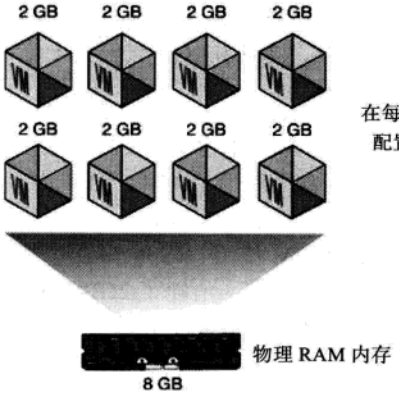


图 4-1 利用内存过量配置，可以在只有 8GB 物理 RAM 内存的服务器上配置 8 台内存各为 2GB 的虚拟机，且不会有损失性能

1. 内存过量配置

虚拟化提供的资源共享比起传统架构有一定的优势，最显著的就是 ESXi 主服务器内存的管理。配置后的虚拟机内存可能超过物理内存（RAM），这被称作内存过量配置（memory overcommitment）。内存过量配置使高度的服务器整合成为可能，如图 4-1 所示。

使用内存过量配置时，ESXi 必须使用技术从一个或者多个 VM 中回收内存。这些技术被称为透明页面共享（Transparent Page Sharing，TPS）、气球（ballooning）、交换（swap）和内存压缩（memory compression）。

2. 透明页面共享

ESXi 主机运行多个具有相同操作系统的 VM 时，有些内存页面可能相同。TPS 进程扫描内存，如果发现相同的页面，则保存一个副本并让 VM 使用这个页面，然后释放重复的页面。这一过程对 OS 完全透明，OS 并不知道自己与其他 VM 共享内存页面，如图 4-2 所示。

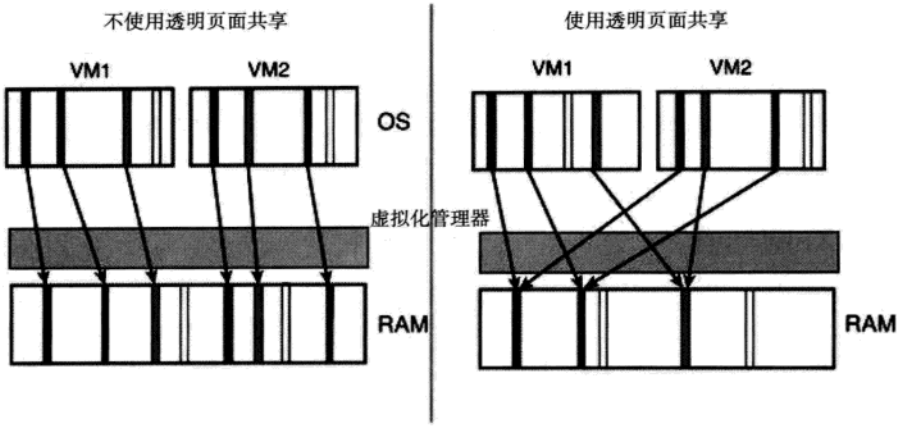


图 4-2 内存页面（使用和不使用透明页面共享）

例子： 考虑用同一模板制作的 10 个具有 1GB 内存的 VM。在所有 VM 启动后，内存页面相同。
如果没有 TPS，10 个 VM 占据的内存为 10×1GB=10GB。
使用 TPS，只使用 1GB 空间，节约了 9GB。这种技术增加了可观的内存空间。

3. 气球

气球技术从 VM 中取得不需要的内存，将其赋予在使用内存过量配置时需要内存的 VM。

气球利用了如下原理：在生产环境中，只有一部分内存密集使用，即系统正常工作必需的活动内存。其余内存（闲置内存，idle memory）很少使用。因为内存是共享的，闲置内存可以回收供其他 VM 使用。vSphere 通过气球利用这些未使用内存。

在图 4-3 的例子中，消耗了 2GB 内存，但是只有 25% 的内存是活动的。其余内存在启动时被分配，但都没有使用。在传统环境中这没有关系，但是在共享资源环境中，未使用内存可以回收供其他 VM 使用。这就是气球的作用。

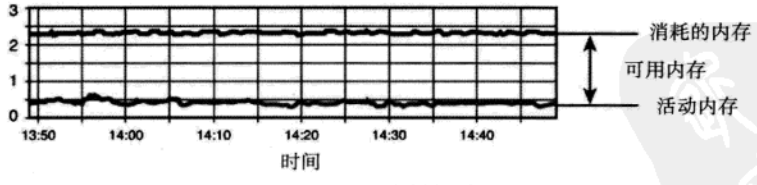


图 4-3 气球可以使用的闲置内存

气球是如何工作的？安装 VMware Tools 后，VM 中有一个气球驱动程序（vmmemctl）。这个驱动程序没有外部接口，它与 ESXi 直接进行专门的通信。当 ESXi 服务器想要回收内存时，它给气球“充气”，增加对 VM 的压力。客户 OS 调用自己的内存管理算法释放较少使用的内存页面（闲置模式下的内存），如图 4-4 所示。

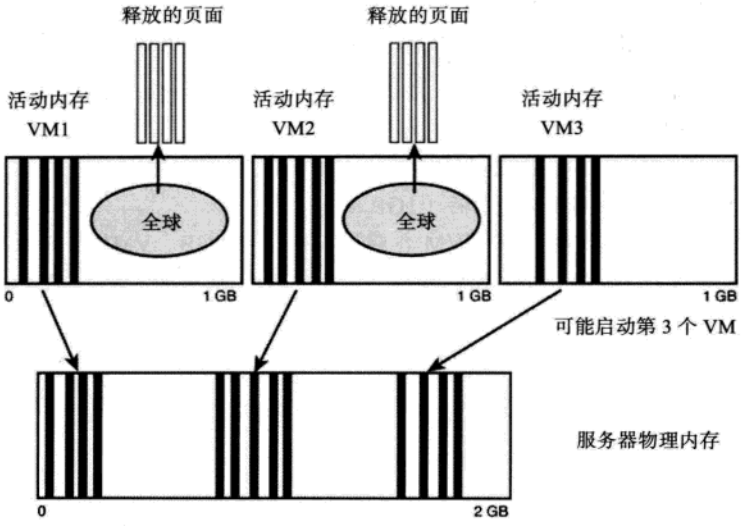


图 4-4 当第 3 个 VM 启动时，其他 VM 的气球驱动程序充气并对 VM 加压，释放出闲置的内存页面

使用气球时，ESXi 必须确定在某些 VM 中给气球驱动程序充气以回收内存。但是哪个 VM 会释放内存呢?

这就是共享（share）概念的作用。共享是与其他 VM 相对的重要性或者优先级，有 4 个级别：低、中、高和自定义（使用 1:2:4 比例或者自定义比例）。ESXi 优先从共享级别设置为低的 VM 中回收内存，赋予共享设置为高的 VM。共享设置为高的 VM 内存最后回收。

示例： 一台 ESXi 服务器有 4GB RAM。两个 VM 各配置了 4GB 内存。VM1 被设置为高（比率为 4），VM2 被设置为低（比率为 1）。

根据这个共享设置，比例的计算如下：

VM1=4/5×4GB，包括取自 RAM 的 3.2GB

VM2=1/5×4GB，包括取自 RAM 的 800MB

如果添加了第 3 台共享设置为高（比率为 4）的 VM，结果如下：

VM1=4/9×4GB=1.77GB

VM2=1/9×4GB=444MB

VM3=4/9×4GB=1.77GB

共享设置的主要局限是缺乏关于使用活动内存的标准。因此，共享设置为高的 VM 可能积累非活动的内存，而共享设置为低的 VM 如果使用大量活动内存，将因此遭到惩罚。

ESXi 引入闲置内存税（idle memory tax）来解决这个问题。这种税的作用很简单：优先回收空闲内存，不管 VM 的共享级别是什么。思路是空闲页面将被课税，而活动页面不会。本书不详细讨论闲置内存的计算，重要的是理解 ESXi 使用内存页面的样本概率方法。它直接这样做，无须在 VM 的操作系统中干预。每个 VM 都在一个可配置的间隔内单独采样。默

认情况下，ESXi 每 30 秒采样 100 个页面。

默认情况下，闲置内存税参数被设置为 75%，对应的是 75% 的未使用内存回收率。

示例： 一台 ESXi 服务器有 4GB RAM。每个 VM 配置为 4GB 内存（内存共享设置为常规）。这里有两个 VM，一个有密集的内存活动，另一个没有。

VM1= 活动目录，具有 1GB 活动内存（3GB 闲置）

VM2=SQL，具有 3GB 活动内存（1GB 闲置）

因为共享被设置为常规，每个 VM 各获得 50%，即 2GB。VM1 还有 1GB 可用。

闲置内存税授权虚拟化管理器回收 VM1 非活动内存的 75%（确切地说是 1GB 的 75%，即 750MB）。

在闲置内存税的作用下，VM1=1.25GB，VM2=2.75GB。

通过禁用气球，可以阻止回收某个 VM 的内存。为此，这个 VM 的保留级别必须配置为与配置内存相同。

4. 交换

当 VM 启动时，创建一个扩展名为 .vswp 的交换文件。只有能够创建和访问这个文件时才能启动 VM（这意味着应该注意可用的存储空间）。

ESXi 使用这个交换文件以防使用过量配置且气球驱动程序无法从其他 VM 回收内存的情况。在这种情况下，ESXi 使用磁盘作为内存。注意，这种技术应该永远不使用，因为磁盘访问会大大影响性能。交换的唯一好处是内存不足的时候 VM 不会崩溃。

提示信息： 硬盘访问时间以毫秒（ms，千分之一秒）计算，而内存访问时间以纳秒（ns，十亿分之一秒）计算，两者相差 100 万倍。也就是说，如果内存花费 1 秒钟访问信息，硬盘就要花费 11.5 天（1000 000/60 秒 ×60 分钟 ×24 小时）。

交换文件有如下特征。

- ❑ 扩展名为 .vswp。
- ❑ 文件大小等于配置的内存减去保留内存。
- ❑ 默认情况下，该文件放在 vmx 配置文件所在文件夹下。（这可以在内存的高级选项中修改）。
- ❑ 该文件在 VM 关闭时立即删除（可以禁止删除）。
- ❑ 如果没有足够内存创建该文件，VM 无法启动。

5. 内存压缩

在虚拟化管理器决定在硬盘上交换页面之前，有一个中间级别：内存压缩。ESXi 压缩页面，并将它们存储在受保护的内存空间中。对压缩内存的访问快于磁盘访问，没有太多的性能损失。当一个虚拟页面必须交换时，ESXi 首先试图在 2KB 或者更小的块上面压缩。

注意： 在 vCenter 的高级参数中，可以定义压缩缓存的最大大小并停用内存压缩。

6. 改变大小

为了从高水平的整合中获益，必须使用内存过量配置。这可以从如下原则入手，基本的内存配置规则是每个物理核心至少有 4GB 至 8GB RAM。

示例： 具有 12 个核心的服务器应该至少有 48GB 至 96GB 内存。

可以用多种方法确定可操作的 VM 数量：

- 对于测试、开发和试生产应用采用优化的方法，遵循如下规则：VM 中配置的内存总数是服务器物理 RAM 的 2 倍。

示例： 具有 32GB RAM 的服务器可以有 32 个 VM，每个配置 2GB 内存（或者 64 个配置 1GB 的 VM）。

- 对于敏感应用采用保守的方法，利用物理环境下使用的内存量来确定 ESXi 服务器上需要的内存总量，然后增加 10%。

示例： 32GB RAM 的服务器可以有 30 个 VM，每个配置 1GB 内存。

下面是需要记住的一些最佳实践：

- 必须避免内存交换。但是，使用 SSD 磁盘保存存储交换文件是减少性能下降的好办法。
- 支持 TPS，它能避免性能下降。为了得到最优的结果，在同一个服务器上使用同类 VM（Windows 或者 Linux）。使用模板创建 VM，能够提供相同的基础。
- 如果活动率很高，确保所有 VM 都有足够的内存。如果气球活动很多，客户 OS 进行交换，就会发生内存资源问题。

注意： 有些环境（例如 SAP）密集使用内存，建议不要使用内存过量配置。在这种情况下，使用与配置内存相同的保留内存。

- 用与活动相符的参数配置 VM，不要配置过多或者过少的内存。

4.1.2 处理器

在解释工作原理之前，我们先定义一些术语：

- 1 个中央处理单元（Central processing unit，CPU）=1 个物理处理器 =1 个插槽
- 1 个物理内核 =1 个物理处理器中的 1 个内核
- 超线程（hyperthreading）能在每个物理内核上创建两个逻辑实例
- 逻辑 CPU（LCPU）对应 VM 可用及可配置的逻辑 CPU 数量（在 ESXi 中）。每个 ESXi 服务器最多支持 160 个 LCPU。

示例： 两个具有 6 个内核、启用超线程的物理处理器对应于 24 个 LCPU。理论上，可以配置一个具有 24 个 vCPU 的 VM 或者 4 个具有 6 个内核的 vCPU 的 VM。但是，好的做法是单个 VM 上配置的 LCPU 不要超过 ESXi 上可用 LCPU 数量的 70% ~ 80%。在 24 个 LCPU 的情况下，VM 配置的最大 vCPU 数量应该为 18 个。

- vCPU 是 VM 中配置的虚拟 CPU。
- vCore 是 vCPU 中的一个内核。

在 vSphere 5 中，可以配置多达 32 个 vCPU 的 VM，每个 ESXi 主机可以有多达 2048 个 vCPU。vCPU 中的 vCore 数量也可以配置，但是每个物理内核的 vCPU 数量限制为 25 个。

注意： vCore 级别的配置很有趣，因为有些操作系统和应用基于 vCPU（而不是 vCore 的数量）。例如，VM 可以配置成 1 个 vCPU 和 4 个 vCore 而非 4 个 vCPU，因为这对性能没有影响但是能节约许可证，还提供了摆脱某些操作系统最大处理器数的限制。例如，Windows 2008 Server 标准版限制 4 个 vCPU（256 个 LCPU）。

创建一个 VM 时，虚拟插槽数量和每个虚拟插槽的内核数量必须输入，如图 4-5 所示。

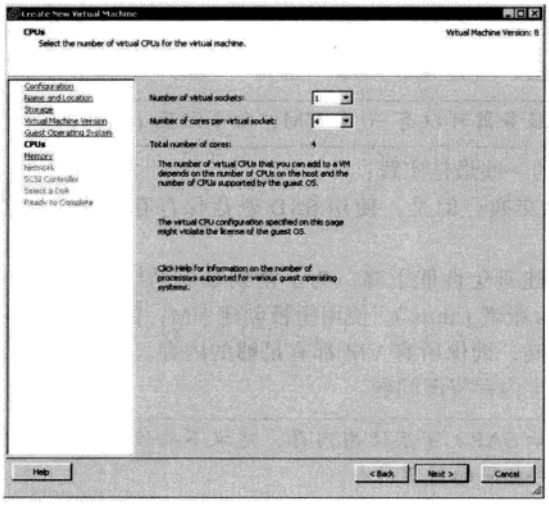


图 4-5　选择虚拟 CPU 数量

1. 处理器管理

从客户操作系统虚拟处理器（称作 vCPU）发往 ESXi VMkernel 的指令被 VMM 拦截。在固定时间间隔内，VMkernel 动态地在服务器的不同处理器（或者多核处理器的内核）中分配 VM 工作负载。因此，VM 指令根据每个处理器的工作负载从一个处理器（或者内核）转移到另一个处理器。处理器是服务器组件中唯一未被虚拟层屏蔽的。客户 OS 能够发现所运行的服务器中找到的物理处理器的类型和型号。图 4-6 说明了这一概念。

2. 多内核和虚拟化

Intel 的研究显示，利用频率的提高增加 13% 的性能，在耗电上要增加 73%。然而，增加一个核心并将频率降低 20%，性能可以增加 70%，而耗电仅增加 2%。如果再增加两个核心，总耗电仅增加 6%，而性能增加了 210%！

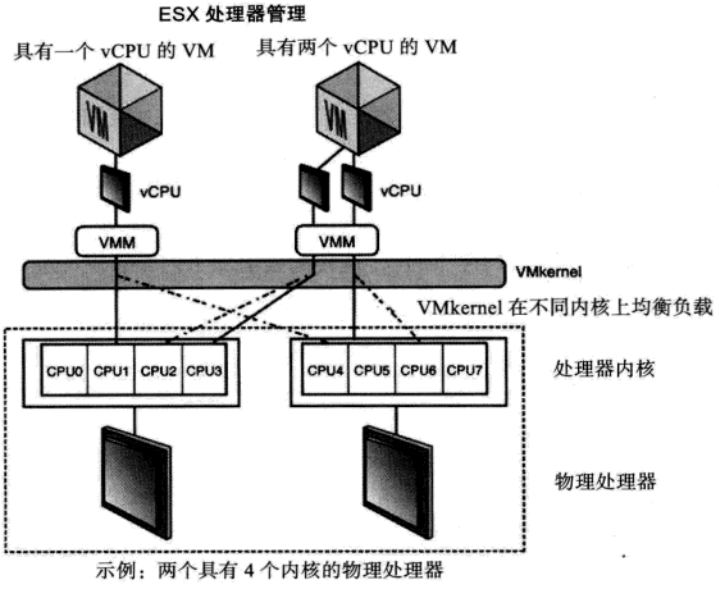


图 4-6 处理器管理——4 个虚拟处理器（内核）和两个物理处理器

使用多核心导致可观的耗电下降，并且提供很好的性能。虚拟化是最好地利用多内核提供的可能性的技术之一，因为 ESXi 能够像管理物理处理器一样地管理核心。

这使得用大量核心达到很高的整合率成为值得思考的方案。现在，有些处理器有 6 个内核，但是在几年内，处理器可能有几十个甚至上百个内核，能够达到更高的整合水平。

3. 对称多处理和虚拟 SMP

对称多处理（Symmetric MultiProcessing，SMP）和虚拟 SMP（vSMP）是指操作系统通过并行执行任务（称为多线程），同时使用多个不同的处理器，这能够在处理器之间均衡负载。理论上，这似乎非常有趣；而在实际中，需要注意几个方面的问题。

如果开发能够并行执行任务的应用程序，具有多个物理处理器的服务器就能利用 SMP 并从中获益。但是在实践中，除了少数据库类应用（Microsoft SQL、Oracle、IBM DB2、SAP）和商业 / 科研应用之外，多线程应用程序还很少。如果应用开始时不是这样设计的，就需要重新开发。

注意： 旧版本 ESX（ESX 2）中，配置两个 vCPU 的 VM 要求同一时候要有 2 两个可用的处理器才能管理任务（称作 CPU 调度）。在更新版本中（ESX 3、4 和 5），VMware 引入了松弛协同调度（relaxed co-scheduling），即使同一时候没有两个可用物理 CPU，也能调度两个 vCPU 的 VM。尽管如此，不要为 VM 分配没有必要的多个 vCPU，这一点始终很重要。

在某些情况下，使用 vSMP 甚至不利于性能，因为配置为两个 vCPU 的 VM 需要同时有

两个可用的处理器才能管理任务。在共享环境中，这有发生争用的危险。

一般的原则是，在使用 SMP 之前，最好是咨询软件制造商。

4. vCPU

客户 OS 使用被称为 vCPU 的虚拟处理器。在 vSphere 5 中，VM 可以配置为可利用 SMP 的 1 个至 32 个 vCPU。（注意，客户 OS 必须能够支持 SMP。）

如果同一台服务器上运行多个应用，可以配置具有多 vCPU 的 VM。这能够改进性能，因为程序可以同时在不同的处理器上运行。在实践中，系统先决条件（例如，OS、服务包）可能导致各类应用不兼容，无法共存。而且，这种配置增加了内存管理冲突的风险。

因为这些原因，在少数情况下，如果 VM 中运行专门开发的应用或者多种应用，最好是限制 VM 中 vCPU 的数量。

注意： Windows 2003 Server 和更早的版本根据系统中的 CPU 数量使用不同的 HAL（硬件抽象层）。对于单个 vCPU，使用单处理器（uni-Processor，UP），对于多处理器则使用对称多处理（SMP）。Windows 在从 UP 转换为 SMP 时自动改变 HAL 驱动程序。但从 SMP 转换为 UP 则非常复杂。所以，最好是从一个 vCPU 开始，如果只有一个 vCPU 使得性能低下，再按照需要添加 vCPU。如果你从多个 vCPU 减少为一个 CPU，仍然会在 OS 中使用多处理器 HAL。这会比使用正确 HAL 的系统性能更低，且消耗更多的 CPU。Windows 2008 Server 不受这个问题的影响，因为它对 UP 和 SMP 使用相同的 HAL。

---

5. 超线程

超线程是在一个物理处理器或者内核上创建两个逻辑内核实例，从而在核心中并行执行任务，提高效率。这种功能在前一代处理器中已经消失，但是 Intel Nehalem 处理器又重新集成了它。

更多关于性能结果的信息可以访问 www.vmware.com/products/vmmark/results.html。

6. 虚拟化技术

为了能够在同一个硬件上同时使用不同的操作系统，虚拟化管理器使用了多种不同的技术：完全虚拟化、半虚拟化和硬件辅助虚拟化。

要理解这三种技术的工作原理，必须理解 x86 处理器的架构。

如图 4-7 所示，x86 处理器架构提供 4 个特权级别（Ring 0 至 Ring 3，其中 Ring 0= 内核模式，Ring 3= 用户模式）。

执行级别（ring）定义了程序的执行特权。程序安装的级别越低，对系统的控制越强。OS 具有最高的控制级别，它运行于 Ring 0，直接访问资源。

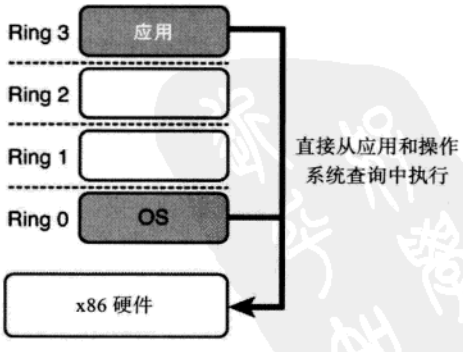


图 4-7　x86 特权级别

应用运行于最高的 Ring 3，它们无法修改在较低层次执行的程序。应用不能停止 OS，而 OS 能够停止应用。

Ring 1 和 Ring 2 定义比 Ring 0 低的特权。

在这种架构内，虚拟化管理器是如何处于硬件和 OS 之间，运行于 Ring 0 的呢？

这一难题的克服归功于全虚拟化和半虚拟化技术。

1）全虚拟化

VMware 在 1998 年开发了二进制翻译（binary translation）技术，首先解决了这一问题。利用二进制翻译，虚拟化管理器能够放在 Ring 0 中，操作系统则被移到编号较高的级别（Ring 1），这使它们得到比应用更高的特权，如图 4-8 所示。

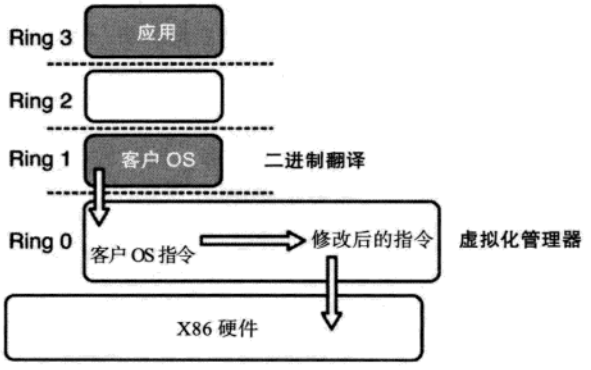


图 4-8 二进制翻译示例

因为操作系统被设计为运行于 Ring 0，它们通常会检查所在级别，因为有些指令只能在源或者目标在 Ring 0 时才能执行。ESXi 拦截某些请求进行二进制翻译，使客户 OS 不知道它们在系统中的实际位置。

二进制翻译修改来自客户 OS 的一些指令，然后将它们发送给物理处理器进行处理。

使用这种技术的好处是，在客户 OS 的核心层（节点）不需要任何修改，因为二进制翻译发生在处理器二进制代码级别。

优点：主机 OS 没有意识到被虚拟化，也不需要修改 OS。这带来了与许多操作系统的兼容性。

缺点：二进制翻译需要 CPU 上的额外工作（开销）。

2）半虚拟化

半虚拟化（paravirtulization）由 Xen 开发，是使用多种操作系统的另一种技术。它包括对客户操作系统（节点层：核心）进行修改，使它们能够在 Ring 0 之外运行，如图 4-9 所示。

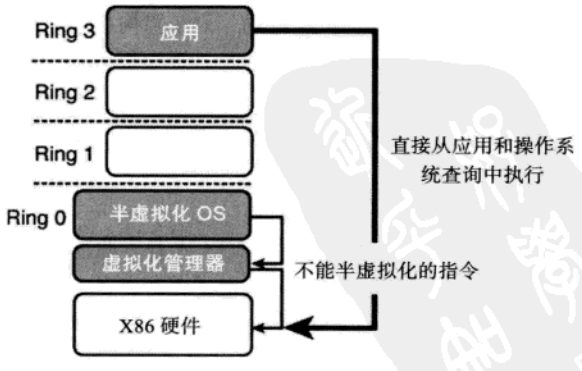


图 4-9 半虚拟化示例

客户操作系统知道自己的虚拟化，在向硬件发送某些低级指令之前对指令进行修改。因此，没有指令拦截和二进制翻译。

这种半虚拟化技术很有效，但是需要对客户 OS 的修改，这并不总能做

到，特别是对某些 Windows 版本。半虚拟化技术简化了虚拟化管理器所执行的任务，虚拟化管理器具有某些处理器和内存指令集的特权。

优点：性能很接近于纯物理环境。

缺点：实现很复杂，因为内核必须修改。与 OS 的兼容性很差。

注意： VMware 使用半虚拟化驱动程序引入了半虚拟化技术的某些特性。这些特殊驱动程序（VMware 开发）能够感知虚拟化层，更容易与虚拟化管理器通信。它们由某些客户操作系统的 VMware Tools 添加，能够改进性能。

3）硬件辅助虚拟化

为了简化虚拟化管理器的任务，同时避免将 OS 放在设计以外的 Ring 级别或者修改 OS 内核，具有 Intel VT 技术的 Intel 处理器和具有 AMD-V 技术的 AMD 处理器引入了一个新的执行模式——*硬件辅助虚拟化*。

这一模式提供了对应于 Ring 0 以下的一个根级别，和对应于 Ring 0 至 Ring 3 的常规级别，如图 4-10 所示。特权级别直接访问硬件，这能直接接收一些限制二进制翻译的客户 OS 的指令。

虚拟化管理器工作于根模式，拥有最高级别的控制权。客户操作系统工作于 Ring 0。它们占据的空间和原始设计一样。

没有必要修改客户 OS，在大部分情况下也不必使用二进制翻译。（但是，对于某些指令集仍然需要二进制翻译。）

这种新的根级别大大减少了开销。这种指令集的变革也使得 VM 之间的物理资源共享更加流畅。

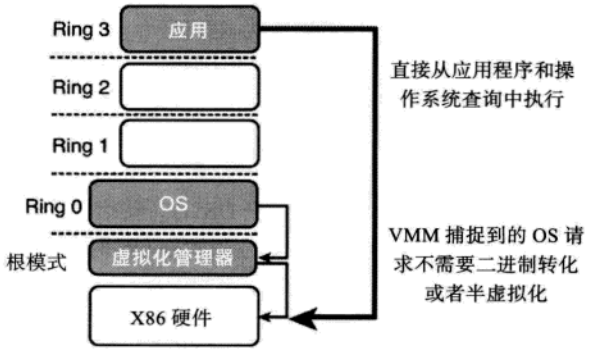


图 4-10　硬件辅助虚拟化

由于处理器中的硬件辅助，x86 架构摆脱了一些技术负担，提供了非常接近于原生环境的虚拟环境性能（根据不同情况，接近程度为 90% ~ 95%）。

7. CPU 指标

有许多可用的指标。在 vCenter Server 中，以下两个指标特别有趣：CPU 使用和 CPU 就绪时间。

1）CPU 使用（平均）

这一指标存在于服务器、VM 或者资源池级别，确定 CPU 使用的时间比例（见图 4-11）。

应该监控这一指标，它给出了 VM 是否独占 CPU 大部分时间的指示。监控 ESXi 的物理处理器（pCPU）使用也很重要，它们平均不应该超过 75%。

2）CPU 就绪时间

一个有趣但鲜为人知的指标，也被称作*就绪时间*（ready time）。在具有共享资源的虚

拟环境中，多个 VM 可能同时使用物理 CPU。VMkernel 中的资源管理器根据 VM 请求分配 CPU 时间并确定 VM 应该运行于哪一个核心。有些 VM 必须等待物理 CPU，就绪时间是 VM 在 CPU 上运行之前等待的时间量。

注意： 如果处理器核心超载，资源管理器将 VM 留在同一个核心上，以便使用 CPU 缓存上的数据，这能够造成较短的 CPU 就绪时间。不过，如果等待时间太长，管理器也可能决定将 VM 迁移到另一个内核。

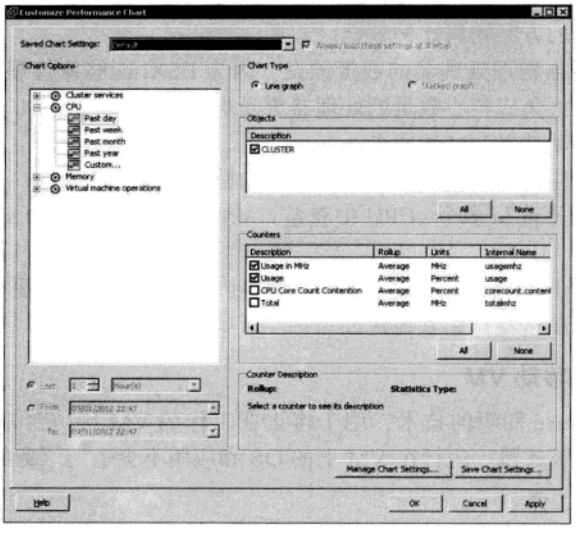


图 4-11 CPU 性能图标设置

虽然对于服务器来说就绪时间的积累很正常，但是应该监控该值，不应超过如下限值：

- 对于一个 vCPU，10% 或者 2000 毫秒
- 对于两个 vCPU，20% 或者 4000 毫秒
- 对于四个 vCPU，40% 或者 8000 毫秒

该值不是唯一需要考虑的，但是必须定期监控。这能很好地指示最终配置问题（调度亲和性、vSMP 的不正确使用，或者不合适的 VM 共享 / 限制设置）和基础架构中 VM 不正确部署。

3）好的做法

每个 ESXi 主机服务器可以有几个 VM？这是很难回答的问题，因为答案取决于环境和应用负载，不过有几个可用的参考点。一台服务器上可以运行的 VM 数量取决于 ESXi 服务器中可用的内核（不考虑超线程）数量。整体来说，可以进行如下的容量规划：

- 中等负载下，每个内核 2 至 4 个 vCPU
- 大负载下，每个内核 1 个 vCPU

示例： 一台具有 12 个内核的服务器允许运行 24 至 48 个 vCPU。这等同于中等应用负载下 24 ~ 48 个各配置 1 个 vCPU 的 VM。对于高活动率应用程序，则代表着 12 个 vCPU（例如，12 个各配置 1 个 vCPU 的 VM 或者 3 个具有 4 个 vCPU 的 VM）。

其他值得考虑的好选择包括：

- 从成本角度来说，单处理器（1 个插槽）是有利的，但是提供的整合率低于多 CPU 系统。双处理器（两个插槽）是目前虚拟化中最常用的，因为它们提供很好的价格 / 整合度比。四处理器（4 个插槽）提供了对虚拟化的高附着率。八个处理器（8 个插槽）及更多处理器的方案销售比例较低。
- 许可证部分根据物理处理器的数量颁发。因为 ESXi 能够像管理物理处理器一样管理核心，具有最大集成核心数量的处理器更受欢迎。
- 根据形态和活动部署 VM 很重要。
- 研究显示，启用超线程能改进多线程应用的性能。
- 因为有些应用不能从多个 vCPU 中获益，VM 应该配置最少的 vCPU。咨询制造商以得到建议。
- 频率较高的处理器能改进性能。但是，对于某些工作负载，更大的缓存能够提高性能。咨询软件制造商得到其他建议。

4.1.3 用 vMotion 移动 VM

vMotion 是 VMware 发明的技术，用于将正在工作的 VM 从一台 ESXi 主机服务器完全透明地迁移到另一台服务器。运行在 VM 上的 OS 和应用不会遭遇服务中断。

1. vMotion 如何工作

使用 vMotion 迁移时，只移动 VM 的状态和内存（及其配置）。虚拟磁盘不移动，保留在相同的共享存储位置中。VM 迁移完成后由新的主机管理。vMotion 仅能工作于集中存储架构（例如 SAN FC、SAN FCoE、iSCSI 或者 NAS）或者 vStorage Appliance 。

触发 vMotion 时，活动内存通过网络，以如下的步骤传送到目标主机（见图 4-12）。

1）vCenter 服务器确认 VM 处于稳定状态。

2）VM 的内存状态和内容通过 VMkernel 端口发送到目标服务器。vMotion 获取一系列内存快照，并连续地将这些快照传送到目标服务器。

3）目标服务器上的复制完成之后，vCenter Server 解锁并挂起源 VM，使目标服务其能够执行磁盘锁来控制它。

4）因为网络层也由 VMkernel 管理，vMotion 确保在迁移后，VM 的网络标识（如 MAC 地址和 SSID）得到保留。活动网络连接也能保留。

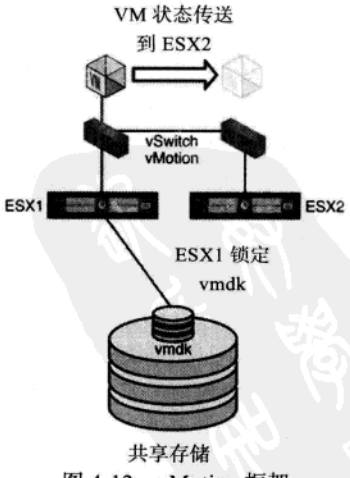


图 4-12 vMotion 框架

5）VM 继续其在目标主机上的活动。

6）VM 在 ESX1 上运行。其 vmdk 虚拟磁盘在共享存储上找到。vCenter 触发 vM 到目标服务器的迁移。VM 状态被复制到第 2 台服务器（见图 4-13）。

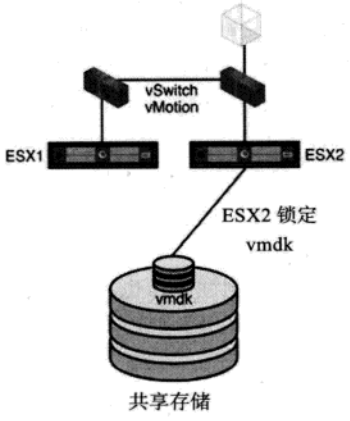


图 4-13 VM 状态复制后，ESX1 服务器去掉磁盘锁，ESX2 服务器锁定 vmdk 供自己使用。vmdk 虚拟磁盘在迁移期间不移动

2. 何时使用 vMotion

vMotion 主要用于计划维护操作，例如固件升级或者添加组件（例如内存）。这使得将 VM 迁移到另一个服务器、执行维护操作、在操作完成之后将 VM 切换回原始服务器成为可能。

3. 需求和最佳实践

- vMotion 产生 SCSI 保留，从而在短时间内锁定 LUN。因此，不应该频繁使用它，因为这可能导致磁盘访问性能的损失。
- 避免花费太多时间在指令集差异过大的处理器上使用 vMotion。最好是尽可能选择同代和同家族的处理器。
- 为了正常工作，vMotion 需要许多先决条件，将在后面的小节中列出。

1）存储

vMotion 只有在两个服务器（源服务器和目标服务器）都能访问的共享存储阵列上才有效。

2）网络

vMotion 需要至少一个千兆网卡。所有源和目标主机服务器必须配置为一个专用 VMkernel 端口组。进行迁移时，vCenter Server 根据给定的名称将 VM 分配到端口组。这就是主机之间使用一致的端口组名称的重要性。

注意： 在 vSphere 4 中，只能为 vMotion 配置单个 GbE 网卡。这一限制在 vSphere 5 中已经被去除，可以有最多 16 个 1GbE 网卡和 4 个 10-GbE 网卡。这也减少了迁移 VM 所需要的时间，特别是对于密集内存活动的 VM。每个主机上支持的并发 vMotion 实例在单个 GbE 网络上是 4 个，在单个 10GbE 网络上是 8 个。

CPU 在 vMotion 活动中起到的作用很有限。根据处理器类型、型号和代次，指令集可能不同，所以不同处理器之间的兼容性需要验证。VMware 在提供改进的处理器兼容性上付出了很大的努力。增强 vMotion 兼容性（EVC）能够屏蔽某些差异，提供了具有不同代次处理器的服务器之间的兼容性。（参见 VMware 知识库：KB1992 和 KB1003212。）

注意： ESXi 5 引入了虚拟硬件第 8 版。该版本与 ESXi 5 之前的版本不兼容，所以使用虚拟硬件第 8 版的 VM 只能迁移到 ESXi 5 及更新版本的主机服务器。

为了支持 vMotion，VM 的虚拟硬件版本必须兼容。为了安全，可以在 vMotion 传输中启用加密（在 vCenter 高级选项中）。

4. 增强 vMotion 兼容性

因为每一代处理器都有多种新功能，EVC 确保群集中的所有主机服务器为 VM 提供相同的指令集，确保 vMotion 中不同代次处理器之间的兼容性。

为此，EVC 使用基线（baseline）。基线启用群集中所有处理器都支持的功能集定义，是所有处理器的共同特征。

ESXi 直接使用 Intel VT Flex Migration 和 AMD-V Extended Migration 处理器和技术，只展示共同的指令集，屏蔽可能造成兼容性问题的指令。（基线和屏蔽的指令集参见 VMware 知识库：KB1003212。）

EVC 不会影响性能，也不会影响核心数量或者缓存大小。唯一可能降低性能的是不使用的一些指令（例如 SSE 4.2）。

注意： *为了使用 EVC，群集中的所有主机必须使用来自同一个制造商（Intel 或者 AMD）的处理器，并在 BIOS 中启用 NX/XD 功能。*

在群集中一旦启用 EVC，所有主机服务器自动配置，以对应规定的基线。不符合先决条件的主机服务器不能成为群集成员。

4.1.4 分布式资源调度器

分布式资源调度器（Distributed Resource Scheduler，DRS）是在生产环境中，群集中不同主机服务器之间共享工作负载，自动化使用 vMotion 的一种方法。

DRS 收集群集主机服务器的使用信息，提供改进工作负载分配的 VM 定位建议。当某台服务器是群集的一部分且启用 DRS 时，在两种情况下会发出建议：

- 初始定位：VM 启动时发生初始定位。

注意： *DRS 和高可用性（HA）可以联合使用。当 HA 重启 VM，初始定位的 DRS 推荐在群集中使用的最佳主机服务器。*

- 负载均衡：当 VM 运行，DRS 在群集的不同服务器之间分配 VM 负载以优化资源。这种迁移可以根据定义的标准自动或者手工（在管理员验证之后）进行。

注意： *启用 vMotion 改进的兼容性功能（EVC）时，可以使用容错（Fault Tolerance，FT）功能。*

1. DRS 规则

在启用 DRS 的 ESXi 群集中，可以设置规则，始终将 VM 放在同一台服务器上（亲和性规则），始终运行不同服务器上的 VM（反亲和性规则），或者始终在特定服务器上运行 VM（主机亲和性）。

图 4-14 展示了类型设置，描述如下。

- Keep Virtual Machine Together（将虚拟机放在一起）：允许实施 VM 亲和性。这个选项确保使用 DRS 迁移 VM 时，VM 保留在 ESXi 群集中的同一台服务器上。这一设置的主要优点与 VM 之间的性能考虑有关（例如某台服务器链接到一个数据库时）。同一台主机上的 VM 之间的通信非常快速，因为这种通信发生在 ESXi 服务器内部（不需要通过外部网络）。
- Separate Virtual Machines（分隔虚拟机）：允许实现 VM 反亲和性。这个选项确保 VM 位于不同的 ESXi 服务器上。这种配置主要用于高可用性考虑（例如，使用 Microsoft MSCS 群集或者 Windows Server Failover Clustering for Windows 2008）。使用这条规则，VM 处于不同的物理服务器上。若一台 VM 损坏，可以确保应用运行于另一个 VM。不应该将反亲和性规则用于管理性能，其重点在于可用性。
- Virtual Machines to Hosts（虚拟机－主机）：允许利用主机亲和性，将 VM 放在特定的主机服务器上，可以指定 VM 在某个主机上运行。这样可以微调群集中 VM 和主机服务器之间的关系。可以将群集的一部分专用于 VM。

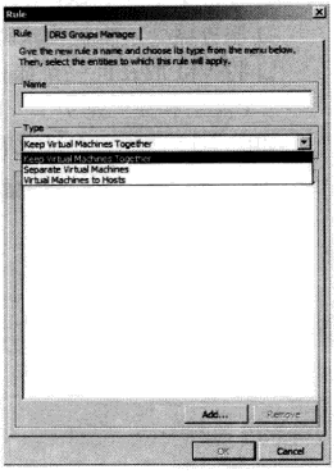


图 4-14 设置 DRS 规则

可能的设置如下：

- Must Run on Hosts in Group（必须在组中的主机上运行）
- Should Run on Hosts in Group（应该在组中的主机上运行）
- Must not Run on Hosts in Group（不能在组中的主机上运行）
- Should not Run on Hosts in Group（不应该在组中的主机上运行）

注意：Must Run 参数表示指令是强制的，直接影响 vSphere HA 和 vSphere DPM 的使用。表 4-1 列出了两种 Must 参数对 vSphere HA 的好处。

vSphere HA 在主机故障之后定位 VM 时跳过某些亲和性（或者反亲和性）规则。这不适用于 HA 系统运行的 Must Run 和 Must Not Run 参数。

表 4-1 Must 参数与 vSphere HA

DRS			
类型	亲和性	反亲和性	vSphere HA 是否尊重这一规则?
VM-VM	将 VM 放在一起	分割 VM	否
VM- 主机	应该在组中的主机上运行	不应该在组中的主机上运行	否
	必须在组中的主机上运行	不能在组中的主机上运行	是

2. 自动化

vSphere DRS 或多或少地根据定义的参数自动化迁移 VM。vSphere DRS 激活时进行如

下工作。

- ❑ 在群集中的 ESXi 主机服务器之间重新分配 CPU 或者内存负载。
- ❑ 服务器处于维护模式时迁移 VM。
- ❑ 将 VM 放在同一个服务器主机（亲和性规则）或者分隔到两个不同的主机服务器上（反亲和性规则）。

如图 4-15 所示，自动化有不同的级别。

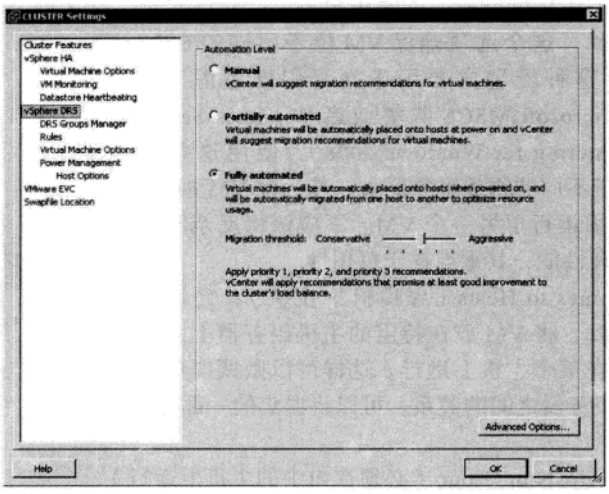


图 4-15　自动化级别

- ❑ Manual（手工）：DRS 做出建议，但是在没有得到管理员验证时不会在主机服务器上放置 VM。
- ❑ Partially Automated（半自动化）：初始定位由 DRS 自动完成，但是运行中的 VM 迁移只有当管理员在 vCenter 中验证建议之后才进行。
- ❑ Fully Automated（全自动）：初始定位和运行中的 VM 迁移都自动进行。这种自动化迁移基于 5 个建议级别所对应的阈值，从保守（5 星）到激进（1 星）：
  - ❑ 第 1 级：保守，5 星。迁移只在规则得到尊重的时候或者主机处于维护模式时进行。
  - ❑ 第 2 级：4 星。迁移只在符合第 1 级或者迁移带来显著改进的时候才进行。
  - ❑ 第 3 级：3 星。迁移只在符合前两个级别或者迁移能带来好的改进时进行。
  - ❑ 第 4 级：2 星。迁移只在符合前 3 个级别或者迁移能带来中等的改进时进行。
  - ❑ 第 5 级：1 星。迁移只在符合级别 1 至 4 级别的所有建议或者迁移带来少量改进时进行。

保守的设置产生的迁移较少，只在亲和性规则不受尊重的时候进行。第 5 级在群集个服务器之间产生的 VM 迁移更为频繁。

在某些情况下，有些 VM 可能采用不同于群集别定义的自动化级别，如图 4-16 所示。

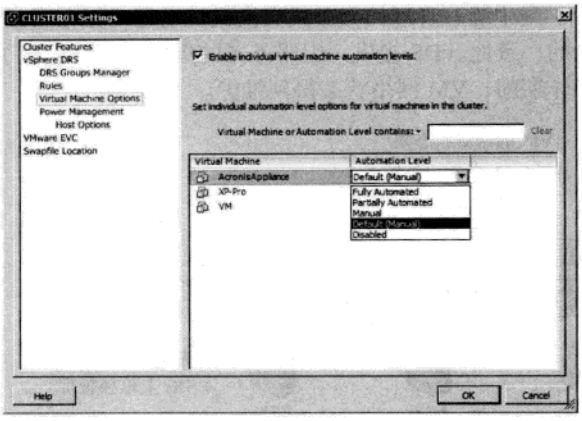


图 4-16 管理员也可以在 VM 级别上指定自动化级别

可能的选择如下：

- Fully Automated（全自动）
- Partially Automated（半自动）
- Manual（手工）
- Disabled（禁用）

4.1.5 vSphere 分布式电源管理

分布式电源管理（DPM）功能允许以最少主机服务器运行的方式布置 VM，降低数据中心服务器消耗的电力。DPM 与 DRS 结合，将 VM 从很少使用的服务器上移走，从而减少工作服务器的数量，从电源供应和空调需求两个方面减少电力消耗。

DPM 使用智能电源管理接口（IPMI）等远程服务器电源开启技术，或者集成关机（iLO）等远程访问卡、局域网唤醒（Wake-on-Lan）等功能。

ESXi 还可以利用高级处理器功能（如 Intel Enhanced Speedstep 或者 AMD PowerNow!），根据实际需求改变处理器频率。

4.2 网络

如图 4-17 所示，vSphere 5 中的网络管理通过以下两种技术选择之一完成：

- vSphere 标准交换机（Standard Switch，也称为 vSS）是一种在每台 ESXi 主机上单独管理的简单配置虚拟交换机。vSS 的使用很简单，但是在大规模的环境中，因为每次配置修改都必须在每个 ESXi 上复制，增加了管理员的工作负担。vSS 的另一个缺点与使用 vMotion 的 VM 迁移相关。因为 VM 的网络状态被重新初始化，监控和故障检修更为复杂。
- vSphere 分布式交换机（Distributed Switch，vDS）是在一组 ESXi 服务器上（每个

vDS 最 350 台）集中统一管理的分布式交换机。它还确保在服务器之间移动 VM 时某种配置和安全的一致性。vDS 提供一组服务器的集中管理，简化了操作。和 vSS 不同，在 vMotion 操作期间，VM 网络状态得到维护。

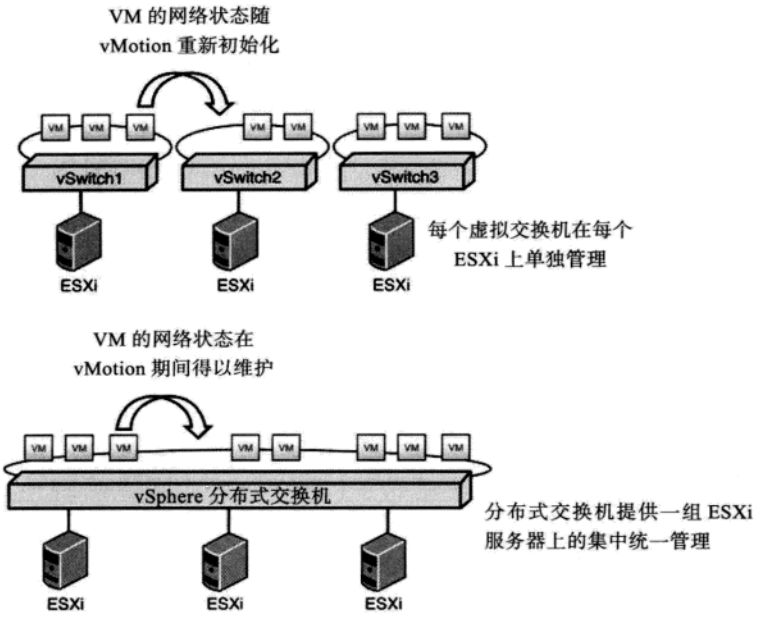


图 4-17　vSphere 5 的标准和分布式交换机选项

还存在一些第三方解决方案，例如 Cisco Nexus 1000V。这是一种分布式交换机技术，具有高级功能，为物理环境和虚拟环境提供统一的 Cisco 管理，对于大规模环境可能有用。

注意： 使用 vSS 或者 vDS 虚拟交换机时，因为它们由 VMkernel 直接管理，网络团队无法直接看到 vSwitch 的内部配置。Nexus 1000V 允许网络团队控制整个网络配置，并将其应用到 VM。

IBM 也有一个分布式虚拟交换机，称作 System Networking Disributed Virtual Switch（系统网络分布式虚拟交换机）5000V。Cisco Nexus 1000V 和 IBM DVS 5000V 都构建于 vDS 之上，所以需要 vSphere 5.0 Enterprise Plus 版许可证。

4.2.1　vSphere 标准交换机

vSS 在 VMkernel 中的第二层上操作（和二层交换机类似），将物理网络组件与虚拟网络组件链接起来。ESX 不负责不同广播域中主机之间 VM 的 vMotion 迁移。

注意： 设计网络架构时，如果你想要 vMotion 或者 DRS 等特性，就必须将主机放在同一个二层 VLAN 中。

虚拟交换机带来了令人印象深刻的使用灵活性，因为它能简单地创建如下高级网络架构：

- 非军事区（DMZ）
- VLAN（ESXi 支持 802.1Q VLAN 标记，VLAN ID 为 1 ~ 4095）
- VM 完全与公司网络隔离，或者连接到不同的网络

可以让多个网络使用同一个 vSwitch，或者为每个网络创建多个不同的 vSwitch。

1. VM 与网络元素的通信

图 4-18 展示了物理和虚拟网络不同元素的概略图。

每个 VM 拥有一个或者多个虚拟以太网卡，称作 vNIC。每个 vNIC 有一个 MAC 地址、一个 IP 地址（vSphere 5 支持 IPv4 和 IPv6），以及到 vSwitch 的一个连接。

注意： 默认情况下，vSphere 自动为 VM 中的每个 vNIC 生成一个 MAC 地址。这个值从 00:0c:29 开始，然后是 VMware 算法生成的 3 个字节。MAC 地址可以在如下范围中手工定义：00:50:56:00:00:00 至 00:50:56:3F:FF:FF。

通过 vNIC，VM 连接到 vSwitch，虚拟交换机有两个特征：一个是虚拟的，另一个是物理的。

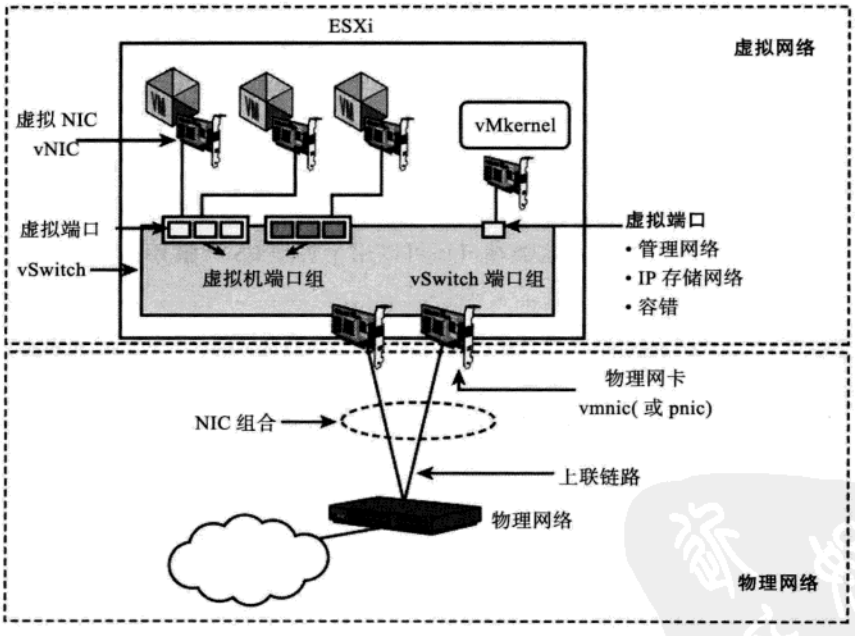


图 4-18 物理和虚拟网络元素

在 vSwitch 的虚拟网络端，有虚拟端口和端口组。VM 的 vNIC 连接到与端口组关联的 vSwitch 的虚拟端口（可以看做是虚拟的 RJ-45 端口），端口组通常对应于一个虚拟 LAN（VLAN）或者特定的端口组。每个 vSwich 可以有最多 1016 个活动的虚拟端口（加上

VMkernel 自用的 8 个保留端口）。

如图 4-19 所示，两种连接类型可用于定义一个端口组：Virtual Machine 和 VMkernel。

注意： 端口组（part group）的概念只存在于虚拟环境，可以实现有关安全、网络分段和流量管理的策略。它还能改进可用性和优化性能。

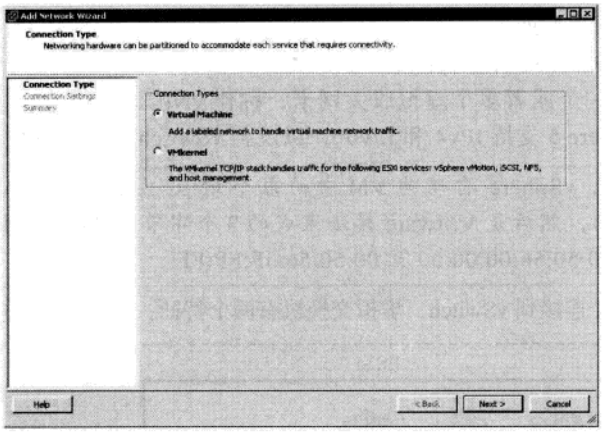


图 4-19　连接类型

1）VMkernel

这个连接用于网络管理、vMotion 流量、IP 存储网络或者 FT。VMkernel 端口需要一个 IP 地址和至少一个物理网卡。这类端口还可以用于管理 ESXi 服务器，和用户接口（如 vSphere Client）通信。

注意： 管理端口和 vMotion 及 FT 端口必须处于不同子网和不同的 VLAN。

2）虚拟机

这个连接被 VM 用来连接到物理环境以及相互通信。在没有连接网卡的情况下，网络完全与服务器外部隔离，VM 之间的通信将发生于 VMkernel 级别（ESXi 主机内部）。

在 vSwitch 的物理网络方面，有物理网卡，称为上联链路（或物理 NIC，pNIC）。这些 pNIC 是物理网络端口。它们连接到 vSwitch 时就变成 vmNIC。

vmNIC 链路可以在 vSwitch 中以两种模式建立。

- 主用适配器：标准交换机使用。
- 备用适配器：在主用适配器损坏时立刻启动。

如图 4-20 所示，vSwitch 可以有一个或者多个物理相关网卡（每个 ESXi 服务器最多可以有 32 个 1GbE 网卡和 8 个 10-GbE 网卡）。在关联多个网卡时，可以将它们组合为单个逻辑卡，提供冗余和工作负载分配。（这被称为网卡组合）。vSwitch 也可以不关联物理网络交换机，从而创建与外部完全隔离的网络。

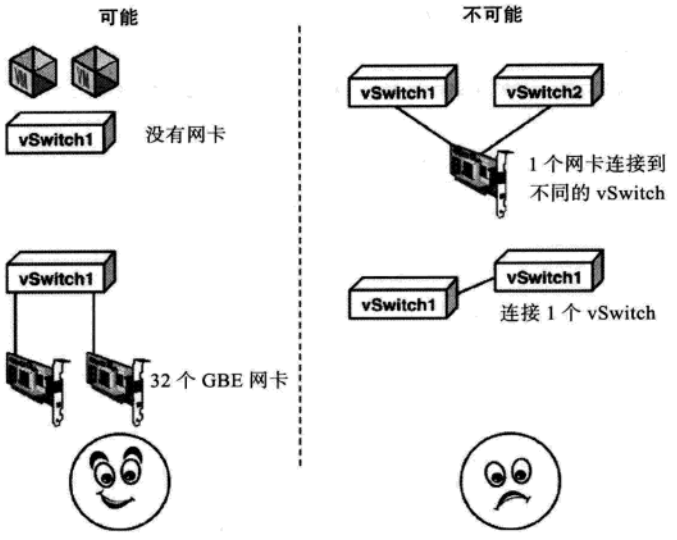


图 4-20 网卡配置

当同一个 ESXi 服务器中的两个 VM 连接到相同的 vSwitch 时，这两个 VM 之间的网络流量在本地进行转发，不占用外部物理网络的带宽。

同一个 vSwitch 上可以有多个网卡。物理网卡（上联链路）无法链接到不同的 vSwitch，一个 vSwitch 也不能链接到另一个 vSwitch，这样能够消除网桥循环。

建议： 分配 pNIC 之后，它们得到了从 0 至 n 的 vmNIC ID。如果主板上集成了两个网络端口和一个双端口 PCI 卡，编号可以在启动物理服务器之前预测，为集成团队提供预先连线规划。vmNIC 根据 PCI ID 编号。这意味着主板的端口 1 为 vmNIC 0，端口 2 为 vmNIC 1，以此类推。值得一试的是在主板上插入一张网卡，改进框架内的冗余性，应付主板转接器或者内部总线的故障，消除单点故障（Single Point Of Failure，SPOF）。

2. 设置虚拟机端口组

如图 4-21 所示，创建一个 VM 端口组时，必须输入端口组名称和可选的 VLAN ID（1 至 4095）。创建之后，必须配置 vSwitch 的端口组策略。

注意： VLAN 允许共享物理网络及其逻辑分段，减少网络设备需求（例如，交换机或者网卡数量）。这也减少了多个物理子网的使用。而且，VLAN 提供更好的流量管理、网络流量分隔和网络之间的隔离。两个物理交换机的互连通过链路聚集（trunking）来完成，可以使 VLAN 存在于多个聚集的交换机上。

设置必须按照如下配置。

❑ General（常规）：这个设置允许配置 vSwitch 可用端口数量（8 ~ 4088 个端口）和最

大传输单元（MTU）（1280 ~ 9000）。见图 4-22。

❑ Security（安全）：可以在 vSwitch 和端口组级别上设置如下策略（见图 4-23）：

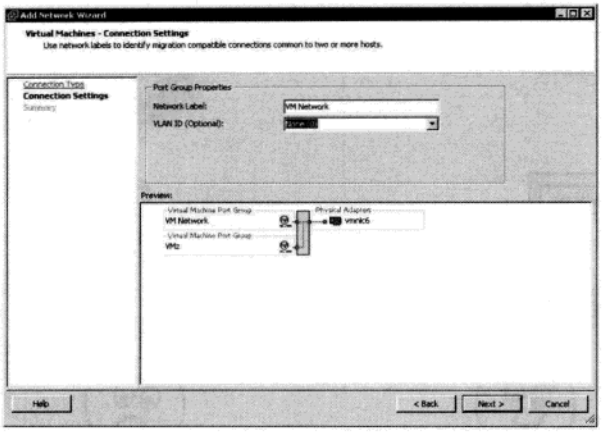


图 4-21　端口组属性设置

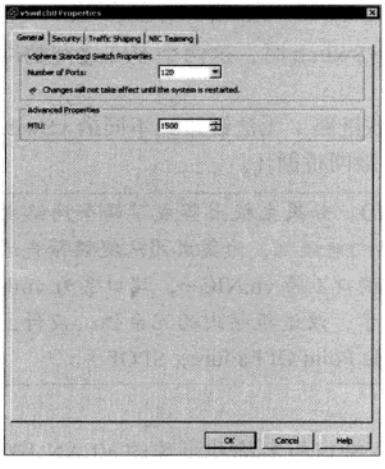


图 4-22　vSwitch 常规属性

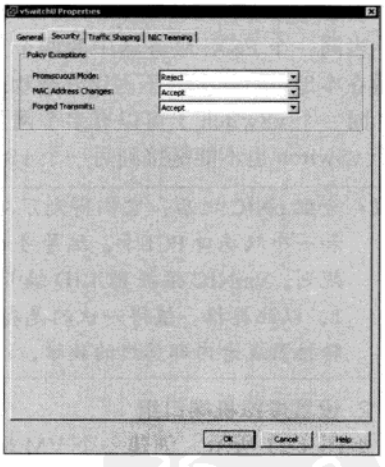


图 4-23　vSwitch 安全属性

❑ Promiscuous Mode（混杂模式）：这种模式（默认下设置为 Reject）禁止 VM 中的 vNIC 查看所连接的 vSwitch 的网络流量。为了安全，不建议启用这个模式。然而，VM 可能专用于网络监控（例如，为了检测入侵）。在这种情况下，启用这个功能是值得的。

❑ MAC Address Changes 和 Forged Transmits（MAC 地址更改和伪造传输）：在 vSphere 中创建一个 VM 时，存在两个 MAC 地址——初始和有效。

vSphere 生成的初始 MAC 地址范围在 00:50:56:00:00:00 至 00:50:56:3F:FF:FF（或者

00:0C:29，可在 vmx 文件中修改）。客户 OS 不能控制这个 MAC 地址。

有效 MAC 地址是用于与网络其他元素通信的地址。默认情况下，这两个地址相同。但是可以在客户 OS 中（网卡的高级设置里）强制设置另一个 MAC 地址。可以在 vmx 配置文件中授权或者禁止改变 MAC 地址，也可以使用 MAC Address Changes 和 Forged Transmits 设置配置客户 OS 中的 MAC 地址。MAC Address Changes 设置与入站流量相关，而 Forged Transmits 设置与出站流量相关。例如，如果 MAC Address Changes 设置为 Reject 且两个 MAC 地址不相等，入站通信将被禁止。

注意： 为了达到高的安全水平，VMware 建议将这些参数设置为 Reject（默认值）。然而，在有些机器从物理转换为虚拟的情况下，因为软件编辑器不再存在或者 VM 中使用的软件不再得到支持，VM 必须强制从过去的物理机器上得到 MAC 地址。在这种情况下，如果允许软件运行的许可证与 MAC 地址关联，必须获得一个非 VMware MAC 地址，该选项就必须设置为 Accept。

注意： 在单播模式的 Microsoft 网络负载均衡环境中，必须将端口组级别的 Forged Transmits 设置为 Accept。（见 VMware 知识库：KB1556。）

- Traffic Shaping（流量整形）：这个设置（见图 4-24）启用交换机级别出站流量的带宽限制。默认情况下该设置不启用。在 vDS 的情况下，流量整形与入站和出站流量相关。
- 网卡组合：网卡组合（见图 4-25）是集中连接到 vSwitch 的多个物理网络交换机。这种组合可以在不同的 pNIC 之间提供负载均衡，如果网卡出现故障可以提供容错。

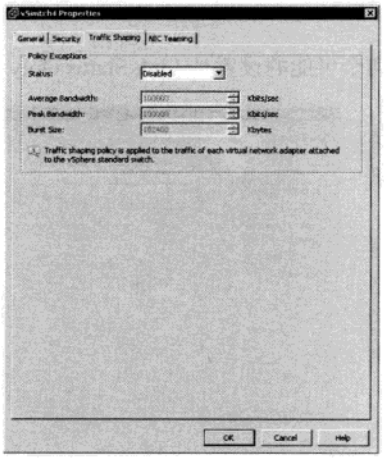


图 4-24 流量整形属性

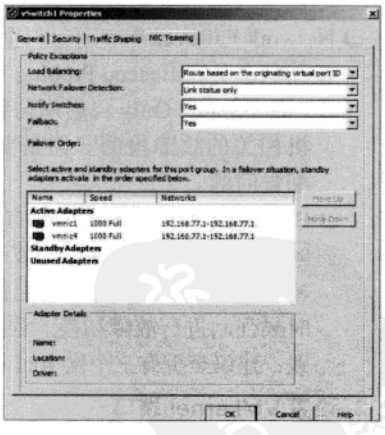


图 4-25 vSwitch 网卡组合属性

还要注意，网卡组合属性页面上有负载均衡选项。负载均衡（Load Balancing）根据连接的数量而非网络流量分布负载。在大部分情况下，只管理出站流量，均衡按照三个不同策略进行。

- Route based on the originating virtual switch port ID（根据来源虚拟交换机端口 ID 进行路由，默认值）：在这种配置下，负载均衡根据物理网卡数量和使用的虚拟端口数量进行。在这种配置策略下，连接到一个 vSwitch 端口的虚拟网卡始终使用相同的物理网卡（vmNIC）。如果物理网卡出现故障，虚拟网卡被重定向到另一个物理网卡。

示例 1：10 个 VM 用连接到 1 个有 5 个物理网卡的 vSwitch。使用网卡组合，负载如下：每个物理网卡上将连接两个 VM。

示例 2：如果 5 个 VM 连接到一个具有 6 个物理网卡的 vSwitch 上，5 个 VM 连接到 5 个不同的物理网卡，另一个网卡只在这 5 个网卡之一出现故障时才使用。注意，端口分配仅在 VM 启动或者故障切换时发生。均衡在 VM 启动时根据端口占有率进行。

注意：谈到网卡组合，重要的是理解这样的情况：例如，两个 1GB 卡创建组合，如果一个 VM 消耗超过一个网卡的容量，将会出现性能问题，因为超过 1GB 的流量不会经过第二个卡，这将会影响共享同一端口的 VM，因为这个 VM 消耗了所有资源。

- Route based on source MAC hash（根据源 MAC 哈希进行路由）：原则与默认值相同，但是根据 MAC 地址的数量进行。
- Route based on IP hash（根据 IP 哈希进行路由）：前两个策略的局限性是虚拟网络使用相同的物理网卡。基于 IP 哈希的负载均衡使用源和目标 IP 地址确定使用的物理网卡。使用这种算法，一个 VM 可以根据目标使用不同的物理网卡通信。这个选项强制将物理交换机端口配置为 EtherChannel。因为物理交换机的配置类似，这个选项是唯一提供入站负载均衡的。
- Network Failover Detection（网络故障切换检测）：两个可能的设置是 Link Status Only（仅链路状态）和 Beacon Probing（信标探测）。
  - Links Status Only：启用与物理网络电缆和交换机相关的故障检测。但是要注意的是，不检测配置问题。
  - Beacon Probing：允许向所有网卡发送以太网广播帧，检测链路状态所没有发现的故障。这些网络帧可使 vSwitch 检测错误配置并在端口被阻塞的时候强制进行故障切换。根据 VMware 的最佳实践，建议至少有三个网卡时才启动这一功能。

3. 设置 VMkernel 端口

VMkernel 端口的配置提供与 VM 端口组相同的设置，加上 IP 地址配置和使用定义。如图 4-26 所示，选择为 vMotion、Fault Tolerance Logging（容错日志）、Management Traffic（管理流量）和 iSCSI Port Binding（端口绑定）。

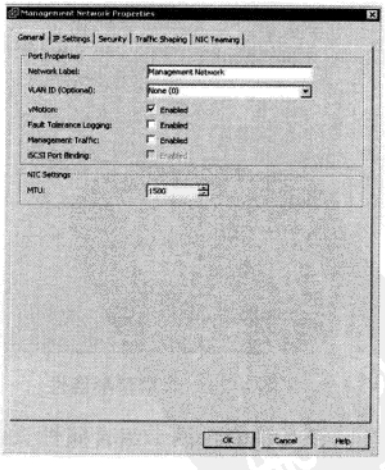


图 4-26　VMkernel 端口设置

4.2.2 vSphere 分布式交换机

vDS 使用由 vCenter Server 配置和管理的 dvSwitch。标准 vSwitch 为单个主机服务器管理网络，而分布式交换机为所有关联的 ESXi 主机服务器工作，vDS 保证在用 vMotion 将 VM 从一台 ESXi 服务器迁移到另一台时，网络配置保持一致，如图 4-27 所示。以同质的方法处理数据中心的 ESXi 服务器，从而统一管理网络。dvSwitch 拥有一个或者多个 dvPortGroups，可以应用全网策略。分布端口组在单一位置（vCenter）中创建，所有主机服务器继承相同的配置。

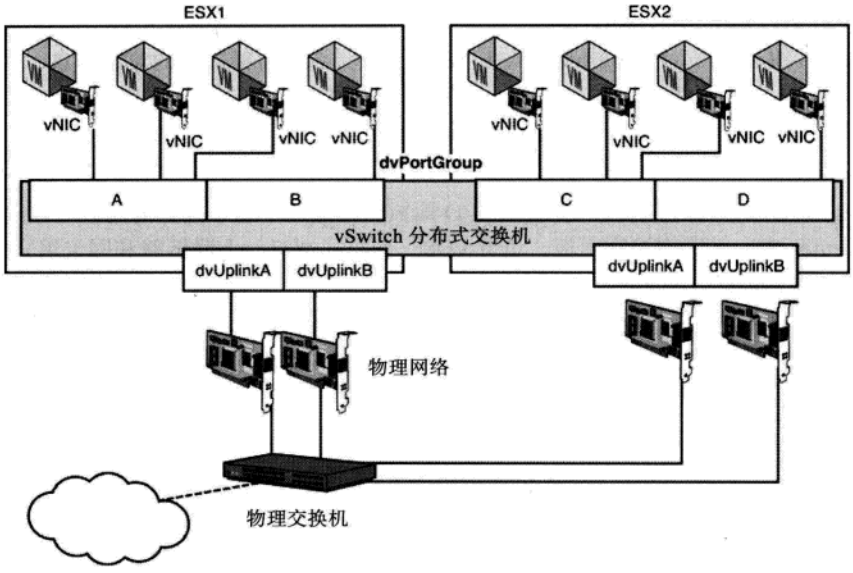


图 4-27 vDS 配置

注意： 使用 vMotion 迁移 VM 时，分布式端口的统计及关联的策略被保留。这简化了调试和故障检修操作。

注意： vDS 允许通过网络 I/O 控制（NIOC）实现服务质量。

dvUplink 为每台主机的物理网卡（vmNIC）提供了一个抽象层。应用到 dvSwitch 和 dvPortGroup 的策略也应用到 dvUplink（不再应用到每台主机中找到的每个网卡）。因此，每个网卡与一个 dvUplink 关联。

vDS 还允许管理 VM 与端口的绑定。这可以三种方式实现。

- 静态绑定：在 VM 连接到分布式端口时为其分配一个端口。静态绑定增加了安全性，并为网络团队提供了更多故障检修的可能性。例如，VM 可以在从一台 ESXi 主机移动到另一台时实施 vMotion，保留流量统计和默认帧，因为当 VM 从一台 ESXi 主机

迁移到另一台时，交换机的端口和 VM 一起“移动”。

- 动态绑定：在 VM 第一次启动，连接到分布式端口之后为其分配一个端口。动态绑定在 ESXi 5.0 中已经显得过时。
- 短暂绑定：没有发生任何端口绑定（动态）。

vDS 还允许使用私有 VLAN（PVLAN）。PVLAN 能够隔离同一 VLAN 中 VM 的流量，增加同一个子网的 VM 之间的安全性。PVLAN 在 DMZ 上非常有用，这些服务器必须从公司外部和内部都能访问。

vSS 仅支持出站流量（出路）的流量整形，与此相反，vDS 支持出站流量和入站流量的整形。

vSphere 5 用 vDS 支持 NetFlow。NetFlow 允许收集源和目标之间的流量信息，为管理员提供了查看 VM 之间网络通信的可能性。这有助于检测入侵、分析（profiling）或者其他恶意进程。所有数据都被发送到一个 NetFlow 收集器。

1. vDS 架构

如图 4-28 所示，vDS 包含控制面板和 I/O 面板。在这两个元素是分离的。

- vCenter Server 管理控制面板，负责 vDS、分布式端口、上联链路和网卡组合的配置，协调端口迁移。
- I/O 面板是隐藏的，它的实现就像每台 ESXi 主机上 VMkernel 中的标准 vSwitch。它的任务是管理流。每台主机上运行一个 I/O 面板代理（VMkernel 进程），负责控制面板和 I/O 面板之间的通信。

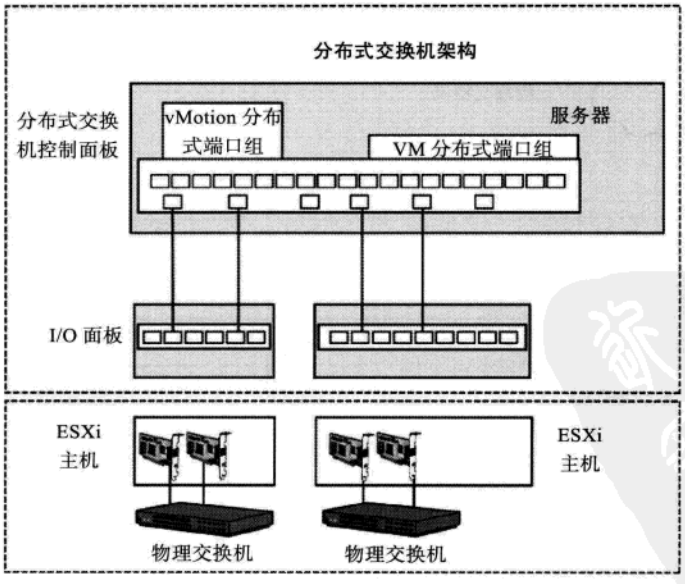


图 4-28　分布式交换机架构

2. 网络 I/O 控制

和存储所用的 SIOC 一样，网络 I/O 控制（NIOC）能够为网络实现服务质量（QoS）。它用 vDS 提供了网络流量管理和对网络 I/O 的控制。对于每种类型的流量和每一组 VM，管理员可以在如下级别上定义共享、限制和 QoS 优先级规则：

- VM
- 管理网络
- iSCSI
- NFS
- FT
- vMotion
- 用户定义
- vSphere 复制

注意： NIOC 在整合的 10-GB 网络基础架构中非常有用，因为它允许不同类型流量的共享和限制。

4.2.3 虚拟网卡

在 VM 中可以配置如下虚拟网卡。

- vlance：模拟 PCNet32 卡，为 32 位客户 OS 提供最佳的兼容性。
- E1000：模拟 Intel 82545EM 千兆以太网卡，与大部分最新客户 OS（如 Windows 2008 Server）兼容。
- E1000e：E1000 卡的改进，可用于虚拟硬件版本 8。
- Vmxnet、vmxnet2 和 vmxnet3：提供最好的性能，并支持超长帧。（这些卡只在安装了 VMware Tools 时可用。）

更多信息请参考 VMware 知识库：KB 1001805。

4.2.4 Cisco Nexus 1000V

Cisco Nexus 1000V 是集成到 VMware 虚拟数据中心的一个虚拟网络用具。这个用具能够将与 VLAN、安全、过滤等相关的网络策略扩展到服务器中的所有 VM，利用现有网络团队在用的标准 Cisco 管理工具，为整个虚拟网络中的物理和虚拟网络提供配置、监控和管理接口。图 4-29 说明了 Cisco Nexus 架构。

如图 4-29 所示，Cisco Nexus 1000V 分布式 vSwitch 由两个元素组成：虚拟管理模块（Virtual Supervisor Module，VSM）和虚拟以太网模块（Virtual Ethernet Module，VEM）。

- VSM：这个模块（以虚拟用具的形式部署）可以从命令行配置虚拟交换机。它可以双倍部署来确保 HA（从主用 VSM 切换到备份 VSM 被称作 Switchover）。每个 VSM 支持 64 个 VEM。
- VEM：这个组件集成在物理网卡和 VM 中的虚拟网卡之间。它代替 VMware 的 vSwitch，负责将包发送到正确的位置，在每台 ESXi 主机上安装，通过 VSM 配置。数据面板代理（Data Plane Agent，DPA）从 vCenter 接收信息。

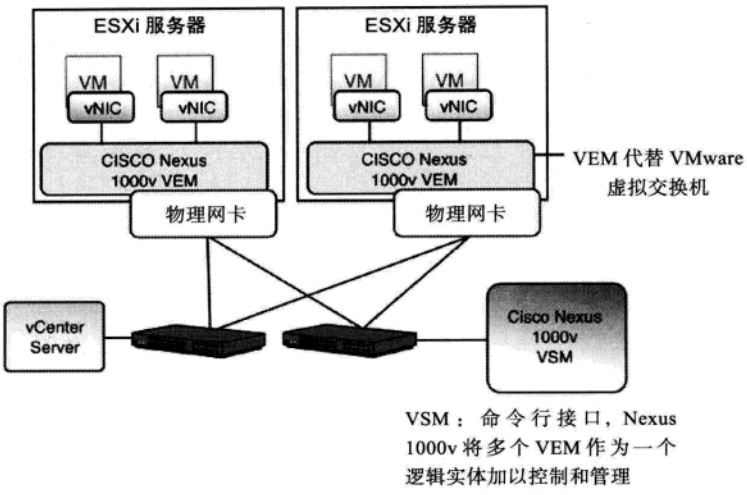


图 4-29 Cisco Nexus 架构

VEM 组件通过三个不同的渠道与 VSM 通信。

- 管理：这个渠道由 ESXi 服务器和 vCenter Server 使用。VSM 使用这个渠道与 vCenter 交流，使端口组可见，这样 VM 就能连接到端口组。管理渠道在安装时以不透明数据（Opaque data）格式，从 vCenter 向 VEM 发送配置数据。
- 控制：这是 VSM 和 VEM 之间的通信渠道。它揭示了 VEM 模块、配置和第二个 VSM 的同步。心跳包在两个 VSM 之间传送（默认情况下，每两秒传送一次，超时为 6 秒）。NetFlow 流也从 VEM 发送到 VSM。
- 包：这个渠道允许信息从网络流向 VSM，以供分析。

注意： 将管理网络的工作放在标准 vSwitch 上能避免你自断后路，也就是说能够防止配置错误。控制和包渠道可以留在 dVS 上。

4.2.5 部署和好的做法

网络可能成为虚拟架构中的瓶颈之一。强烈建议使用最大数量网卡，特别是现在的网卡成本很低，实用性更强。遵循这里的建议有助于确保虚拟基础架构自动管理物理元素的故障，避免任何生产事故的发生。建议最少使用 4 个网卡。

建议： 实施虚拟化项目时，有必要提醒网络团队将 ESXi 服务器整合到数据中心所需要的端口数量。虚拟化基础架构一开始消耗许多端口（为了节约旧的物理机器退役所需的时间）。有时候必须从某些网络端口（为了冗余，至少两个网络端口）开始，直到迁移到 VM 形式的机器上释放端口。定购和整合新的交换机需要时间，必须事先考虑这些延迟，因为它们可能对项目的进展有明显的影响。

还有，使用多个4端口网卡，如果有足够的PCI插槽，将它们分布到不同的PCI端口和转接器。（简单的标准配置使用最少两个端口进行管理，两个端口用于vMotion，两个用于IP存储，两个用于VM。）

下面是其他一些推荐的最佳实践：

- 为安全起见，必须为管理网络、存储网络和VM网络流量创建单独的vSwitch。
- 管理端口可以配置为一个专用的vSwitch和两个物理交换机，或者两个专用的vSwitch和1个物理交换机，但是后者需要两个IP地址（HA检测建议配置）。
- 对于VM端口组，使用最大数量的网卡。（最少需要两个千兆网卡）。
- VM的平均流量和带宽为7MB，通常每个千兆卡可供8 ~ 10个VM使用。当然，这只是一个平均值，取决于工作负载和应用本身。
- 对于VMkernel端口，为vMotionn使用两个网卡。对于iSCSI或者NFS，至少使用两个网卡。
- 在群集的不同ESXi服务器上以一致的方式配置网络。

4.3 虚拟化环境中的应用

在公司里，我们发现应用有不同的关键性等级。有些服务器很关键，例如数据库服务器，因为它们包含了对公司业务需求不可或缺的信息。数据库通常是多层应用的一部分。它们是管理工作流程的参考，特别是在SAP环境中。

很长一段时间，虚拟化用于环境中的测试、开发和接收服务器。现在，采用曲线表明，虚拟化已经普遍用于所有类型的服务器和应用负载。因为vSphere 5 VM支持多达32个vCPU、1TB内存、1 000 000IOPS和36Gbps网络，它们可以轻松地满足绝大部分应用（VMware的说法是95%的应用程序）的性能需求。将关键应用迁移到虚拟化环境不再有任何障碍，因为大部分应用可以毫无问题地虚拟化。

但是，有些应用有着很精确的需求，必须加以考虑，以便用可能的最佳方法将其整合到虚拟化环境中。

4.3.1 Oracle和SQL数据库

- CPU虚拟化只增加少量开销。在主机级别启用超线程。
- 设置内存保留，避免内存过量配置。按照DBA管理员的建议改变内存大小。根据制造商的建议设置VM中的vCPU和内存。
- 网络：使用VMXNET3半虚拟化NIC。VMXNET3为虚拟环境进行了优化，设计用来提供高性能。
- 存储：创建专用的数据存储来为数据库工作负载服务。为使用iSCSI和NFS的基于IP的存储启用超长帧，正确对齐VMFS，为高要求工作负载的Oracle数据文件使用半虚拟化的SCSI适配器，需要高性能的应用可以放在具有专用逻辑单元号（LUN）的专用数据存储上，并且坚持某些基本规则，例如不混合各类I/O磁盘。

- 因为数据库产生的是随机类型的读 I/O，为读写比例为 70% 读 /30% 写的数据库使用 RAID5 或者 5/0。
- 因为日志文件采用顺序写入，为其使用 RAID 1 或者 1/0。首选使用原始设备映射（RDM）模式的磁盘，但是也可以使用 VMDK 磁盘。

注意： *混合 I/O 时，磁头移动过多，延时增加，这些都对性能不利。*

参考制造商对 VM 中 vCPU 数量和内存的建议。必须考虑环境中可能混合多种类型的 VM。因此，为某些关键 VM（特别是数据库）实施 QoS 很重要。

1. vSphere 5 是否支持 Oracle 数据库?

很长一段时间，虚拟化环境中的 Oracle 数据库支持很混乱，解释也各不相同。正式的说法是，VM 只支持 Oracle 数据库 11G 之后的版本（准确地说是 11.2.0.2）。之前的任何版本都没有得到官方支持。

然而，对于更早的版本，向 Oracle 支持部门报告事故且 Oracle 已经知道该问题时，技术支持人员会提供合适的解决方案。如果 Oracle 尚不了解这个问题，技术支持保留要求在物理硬件上重现该问题，证明其与虚拟层无关的权力。在实践中，只有少数的情况需要在物理环境中重现。

2. VMware 中的 Oracle 许可证

对虚拟化环境中的 Oracle 数据库许可证模式必须注意，许可证基于 VM 可以利用的所有处理器。在一个 ESXi 群集中，VM 可能工作于任何群集服务器，这意味着需要群集中所有处理器的许可证。

示例： *在有 3 个节点，每个节点有两个 CPU 的 ESXi 群集中，需要 6 个 CPU 的 Oracle 许可证。但是，为了减少 Oracle 许可证费用，管理员可以适用本章前面介绍的主机亲和性隔离 Oracle 工作负载的 vMotion 域，只为工作负载有“亲和性”的主机服务器提供许可证。*

更多有关许可证的知识，参见 http://bartsjerps.wordpress.com/2011/11/09/oracle-vmware-licensing-cost-savings。

4.3.2 Exchange

对于 Exchange 2010，具有 32 个 vCPU 的 VM 大大超过了 Microsoft 要求的最高配置水平。对于单个简单角色需要 12 个 CPU，对于多角色，需要 24 个 CPU。

例如，对于 Exchange 2007，可以创建多个 VM（配置为 12GB，2vCPU），每个 VM 支持 2000 个用户。这意味着一台物理服务器可以支持最多 16 000 个邮箱（8 个 VM，每个支持 2000 个邮箱）。在物理环境中，单个物理服务器不支持超过 8000 个邮箱。增加服务器的容量不能用于增加同一个服务器中的邮箱数量。

4.3.3 SAP

因为 SAP 应用使用很多内存，VMware 建议不要使用内存过量配置。为了强制这一设置，

必须用 VM 配置的内存作为保留内存配置值。内存必须根据实际使用的内存进行配置。

http://service.sap.com/sizing 上可以找到一个配置内存大小的工具。这个工具能够确定服务器代理插件同步（Server Agent Plugin Synchronization，SAPS）的先决条件。不管物理环境还是虚拟环境，内存大小的配置模式都一样。

SAP 应用的其他一些关注点如下：

- 如果应用提供支持（通常都是如此），配置为多 vCPU 的 VM 启用应用的多线程功能，这能改进性能。
- 关于存储，可以在 SAP 环境中混合 RDM 和 vdmk 磁盘，为每个 VM 首选一个 LUN 映射，以避免 I/O 争用。
- 建议在 VM 中使用 vmxnet 驱动程序以改进性能。
- 在一个具有单个 vCPU 的 VM 中使用 vSphere FT 功能保护 SAP 中央服务（Central Service）组件是个好的选择。

4.3.4 活动目录

活动目录（Active Directory，AD）在 VM 中工作得很好。但是要重点考虑的是时钟同步。所有活动目录活动（例如，修改密码、复制服务等）取决于这些操作发生的时间。Kerberos 是任何活动目录验证请求的核心。它要求通过时间同步服务来同步 AD 客户时钟。Kerberos 的实现容许的最大时间差异是 5 分钟。

对时间的依赖性可能在 VM 中造成问题。当 VM 工作时，它需要 CPU 周期，这会向 VM 提供时间心跳信号。然而，如果 VM 没有活动，它就不会向主机服务器请求 CPU 周期。主机服务器什么也不提供，VM 的时间就会滞后（因为物理主机不向 VM 发送时间心跳信号）。这时候会出现时钟停止（timekeeping）的现象，VM 会遭遇时间的不统一。

解决这个问题可以采用两种方法：

- 使用 Windows 时间服务
- 使用 VMware Tools

更多相关内容可以参考 VMware 知识库：KB1318 和 KB1006427。

下面是建议的 AD 最佳实践：

- 不要暂停 AD VM。
- 不要获得 AD VM 的快照，因为这会导致出错，影响性能。
- 使用 vSphere HA，因为它的优先级高于基础架构中的所有其他 VM，必须设置为高优先级。
- 将 VM 配置为一个 vCPU，一个 vmxnet3 虚拟网卡。
- 使用物理服务器作为 PDC 模拟器。

4.3.5 vSphere 5 环境中的 Microsoft 群集服务

微软群集服务（Microsoft Cluster Service，MSCS）在虚拟环境中的两个 VM 之间可以保留。但是要注意，MSCS 只能在 RDMp 模式（存储区域网络光纤通道 [SAN FC] 或者 iSCSI，

但是不能是网络文件系统 [NFS]）的磁盘上工作，因此不能使用快照和复制。

MSCS 可以用 vSphere HA 代替吗？从理论上说不行，因为两者的粒度不同。MSCS 监控应用，如果它在 HS 中工作则重新启动，而 vSphere HA 监控物理 ESXi 服务器，在服务器宕机（即使它仍能够在某些情况下检测崩溃的 OS 或者应用）时重新启动 VM。

不过，有些公司已经放弃了 MSCS，因为群集的实现方法使得完全利用它很困难（例如，必须开发脚本，测试每一个 OS 更新）。而且，MSCS 要求被保护应用程序是群集感知（cluster aware）的，只有少数应用能做到这一点，而 vSphere HA 能够工作于任何类型的 VM（遗留应用程序）。注意，vSphere HA 和 MSCS 在故障的时候，应用都会停止运行，但是会自动启动备份。

有些公司更愿意实施 vSphere HA。它提供的检测粒度不同，但是日常的管理更简单。

注意： 有些公司考虑，如果 OS 损坏，vSphere HA 就不再有用了。这是真实的情况，但是恢复一个 VM 比重建一台物理机器快得多。如果公司认为这种恢复时间太长，建议实施 MSCS 群集。

4.3.6 改变数据中心

虚拟化已经改变了数据中心，为 x86 计算带来空前的灵活性。内存过量配置允许主机上运行更多 VM，vSphere 提供了气球等技术，帮助最大限度地过量配置内存。VM 的管理能力不仅限于内存，也可以用在 CPU 和网络资源上。vSphere 提供的不仅是增强的管理，还能在物理 ESXi 主机之间移动 VM，改进它们的可用性，这就使服务中硬件升级成为可能，也能通过分配 VM 来管理性能。由于网络交换有多种选择，和机构中的网络专家紧密协作，做出正确选择就变得很重要。

利用这些功能，几乎所有应用程序都能运行于 vSphere 5 VM。Microsoft SQL、Oracle、Exchange、SAP 和活动目录都可以部署在虚拟机上，有了正确的规划，甚至可以比物理环境运行得更好。

第5章
高可用性和灾难恢复计划

应用程序可用性对于公司 IT 系统的正常运作是必不可少的。利用 vSphere 5 的高级特性实现安全的硬件架构，有助于达到高服务水平，在生产场所遭到灾害袭击的时候能够快速恢复业务。

5.1 概述

在这个介绍性的小节之后，本章分为两个主要部分。第一部分讨论数据中心内的本地高可用性。第二部分讨论重大事件之后，生产场所无法再运作时的灾难恢复计划和业务持续性。本章描述了高级特性（例如 vSphere HA[高可用性]、vSphere FT[容错] 和 SRM 5[Site Recovery Manager]）以及它们与基础架构各个组成部分的互动。

5.1.1 恢复点目标 / 恢复时间目标

在数据保护领域，关键的因素是恢复点目标（Recovery Point Objective，RPO）和恢复时间目标（Recovery Time Objective，RTO），它们用来确定故障恢复之后各种可能的解决方案和选择。

RPO 对应于发生破坏时可以接受的数据损失最大数量。每日备份是适用于 24 小时 RPO 的一般技术。对于几小时的 RPO，使用快照和同步复制。RPO 为 0 需要建立同步复制模式，对应于“无数据损失”请求。

RTO 对应于可接受的最大中断时间，取决于重启应用并重新投入服务所需要的时间。位于受保护的远程场所的磁带可以用于 48 小时 RTO。对于 24 小时 RTO，可以从本地站点的磁带进行恢复。对于 4 小时或者更短的 RTO，必须实施多种辅助性技术，例如群集、复制、VMware HA 相关技术、FT、SRM 和存储虚拟化。

虚拟化简化了某些过程，能够缩短 RTO。RTO 取决于实现的技术，关键在于生产站点中断时应用的重新启动和一致性。如果应用程序一致性没有保证，RTO 就会有变化，难以预测。

当然，每个公司都想要确保没有数据丢失，在问题发生时尽快重启生产的解决方案。但是毫无疑问，RPO 和 RTO 的时间越短，解决方案的实施成本就越高。这就是让经理和管理层根据业务需求和约束参与确定 RTO 和 RPO 的原因。

另一个考虑因素是业务影响分析（Business Impact Analysis，BLA），它量化了数据对公司的实际价值。公司投资保护的往往是非关键数据（对管理员可能很重要，但是对公司却不一定），而对关键数据的保护却不足。利益相关方的集体决策能够确定所选解决方案的风险水平。

服务水平协议（SLA）是规定服务水平的契约，是在服务提供商和客户之间正式的具有约束力的协议。在这种协议里，RTO 和 RPO 是重要的因素。

5.1.2 信息可用性

信息系统对公司的运作是必不可少的。它使用户更有生产力，并且拥有交流意见的有效方法（例如，邮箱、协作工具和社会化网络）。信息系统提供了业务活动所需要的应用程序。服务不可用（甚至部分不可用）可能导致严重，甚至不可恢复的收入损失。

统计数字：根据美国国家档案文件署（National Archives and Records Administration，NARA）的统计，数据中心遭受10天或者更久的严重破坏的公司中，93%的公司在第二年中破产。

为了保护公司，最关键的是要实施措施减少服务中断，在生产场所遭受严重事故之后恢复运行系统。

如图5-1所示，系统失效期的大部分原因（79%）来自于计划中的维护——备份操作、硬件添加、迁移和数据提取。这些是可预测的服务不可用情况。其他类型的系统不可用情况则与不可预测的事件相关，如果没有采取可靠的措施和规程，这些预期之外的事件可能造成严重的后果。

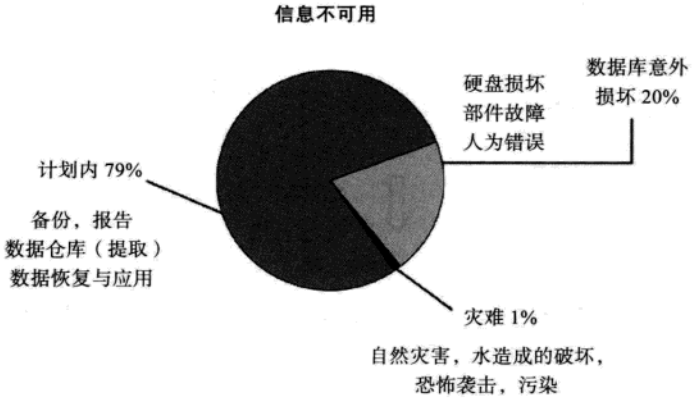


图5-1　系统失效的可预测和不可预测原因

数据可用性计算公式如下：IA=MTBF/（MTBF+MTTR）：

- IA：信息可用性。
- MTBF：平均故障间隔时间，系统或者部件遭遇故障的平均时间。
- MTTR：平均修复时间，维修非服务状态的部件所需的平均时间。（MTTR包括检测损坏部件、规划技术人员干预、诊断部件、获得备件和投产前系统维修的时间。）

信息可用性以百分比度量，是对设定期间业务需求的回应。在这个百分比数字中9的个数越多，可用性越高，一般来说，99.999%以上称为高可用性。

表5-1说明了可用性与每年失效期的关系。

具有99.999%可用性的系统代表每年宕机时间为5.25分钟。注意，这是很短的时间，短于物理服务器重启的时间！

表5-1　根据可用性比例得出的失效期

可用性	每年失效期
98%	7.3天
99%	3.65天
99.8%	17小时31分钟
99.9%	8小时45分钟
99.99%	52.5分钟
99.999%	5.25分钟
99.9999%	31.5秒

5.1.3 基础架构保护

信息系统的保护（见图 5-2）可以分为两类：站点的本地可用性（local availability）和生产站点发生严重事故时所遵循的业务持续性（business continuity）过程。

为了保持本地的高可用性，你可以在某个部件损坏的时候采用如下手段避免服务中断：

- 用硬件冗余消除单点故障（Single Point Of Failure，SPOF）
- 使用群集系统，在服务器故障时将应用恢复到生产状态。
- 保护数据安全，通过备份以避免数据丢失，或者采用复制机制弥补存储的丢失。
- 使用快照，在应用损坏时快速恢复到健康状态。

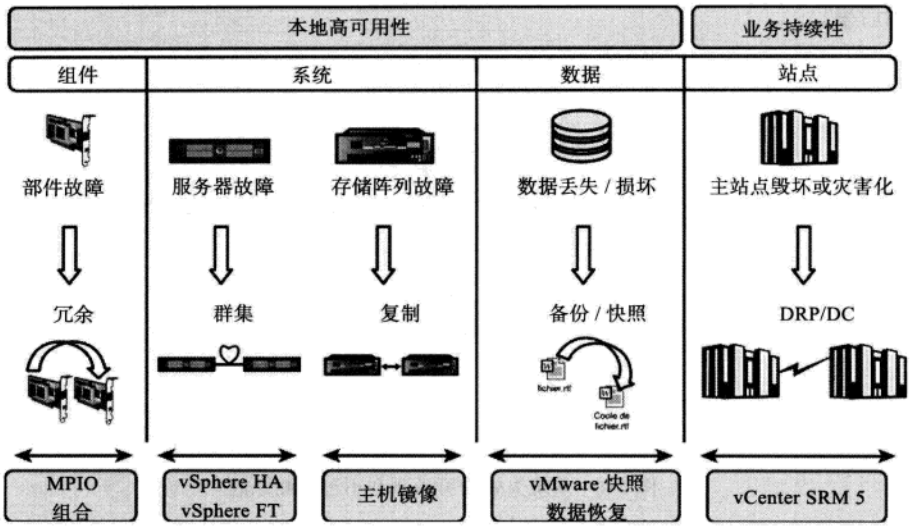


图 5-2 基础架构组件

对于业务持续性，可以实施灾难恢复计划（DRP），在重大事件发生时快速地将生产站点恢复到工作状态。通过部署 DRP，业务持续性计划（Business Continuity Plan，BCP）以及备份站点上的 IT 应急计划，就能在生产数据中心中发生事件时加以处理。

5.2 本地可用性

本节介绍支持本地可用性保护的策略和过程。

5.2.1 消除 SPOF

SPOF 在损坏时会造成信息系统不可用。基础架构虚拟化使得在较少的设备上进行，因为这些硬件上有大量虚拟机（VM），它们变得很关键。一部分硬件的故障可能造成多个 VM 的中断，对信息系统造成可怕的后果。硬件冗余有助于避免这样的后果。要在虚拟化环境中减少服务中断，最佳实践是为所有硬件建立冗余，如图 5-3 所示。

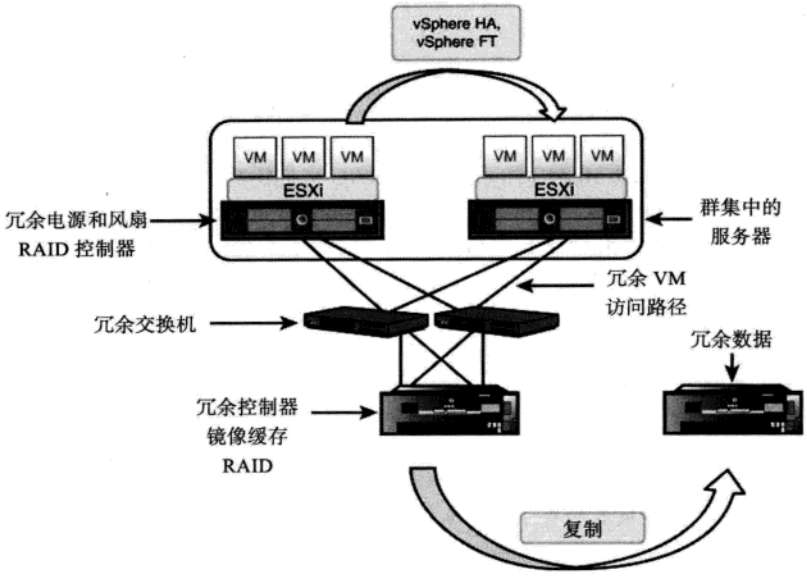


图 5-3　硬件冗余模型

规划硬件冗余时的重要考虑因素如下。

- 硬盘是需要保护的最重要部件，因为它们包含数据，且频繁被访问。为了预防硬盘故障，必须使用 RAID 技术。
- 电源和风扇容易过热并损坏。冗余电源和风扇能够弥补这些问题。
- 尽管不常用，内存也可以用镜像技术保证安全。
- 网卡可以使用 IEEE 802.3ad 协议聚合链路以加强安全。

因为处理器和主板没有机械部件，它们出现故障的可能性较小。但是，可以通过 NEC 的 Fault Tolerant 或者 Stratus Server Systems 等硬件解决方案来保护这些关键部件，这些方案的架构能够提供主板冗余。也可以实施软件解决方案来避免故障，如群集（Microsoft MSCS/WSFC）或者 vSphere FT、vSphere HA 等来提供高可用性方案。

在设计精良的存储区域网络（SAN）架构中，所有部件都有冗余：主机总线适配器（HBA）卡、光纤通道（FC）物理交换机和存储阵列控制器。数据访问安全可以通过使用 vSphere 的原生多路径或者多路径 I/O（MPIO）处理实施冗余访问路径，在连接损坏的时候自动切换来加强。

在存储阵列中，有些阵列制造商提供镜像缓存、冗余控制器和多个磁盘访问路径。

5.2.2　高可用性

高可用性是确保公司中一个或者多个关键应用的本地服务持续性的一组机制。在虚拟化环境中，由于 vSphere 的高级特性，提供 99% 以上级别的服务持续性比物理环境要容易得多。

5.2.3 什么是群集

从根本上说，群集是一组相互连接、组成单个逻辑实体的资源。

群集的目标是提供如下能力：

- 通过IBM PowerHA for IBM AIX、HP Serviceguard for HP UX、Veritas Cluster Server、SUN Cluster、Microsoft MSCS和Oracle RAC来提供高可用性。
- 计算（称作高性能计算，high-performance computing）。

vSphere中的群集是一组具有共享资源，通过vSphere HA、分布式资源调度器（DRS）或者FT提供高可用性和工作负载分配的ESXi服务器（每个群集最多32个ESXi主机和3000个VM）。

vSphere 5还引入了数据存储群集的概念，即一组能够使用Storage DRS根据I/O活动和可用空间分配VM工作负载的数据存储（每个数据存储群集最多有32个数据存储和9000个虚拟磁盘）。

提示： 在一个群集中，最佳实践是统一ESXi的硬件（例如，相同的处理器和内存配置）和软件（例如，相同版本和补丁级别）。这能简化日常管理，以最优的方式使用vMotion和FT等高级特性。如果服务器不完全相同，确保群集的处理器是同一代的。在使用更新管理器之后，验证所有群集成员都进行了所有的更新。

推荐： 群集可以包含最多32台ESXi主机服务器。在我们的经验中，最好是在群集中放置8台ESXi主机。使用超过8台服务器使管理（例如，维护兼容性或者更新硬件）更加复杂。使用少于8台主机灵活性较差，在工作负载分配和可用性上的潜力也较差。

5.2.4 vSphere HA

ESXi主机服务器损坏时，vSphere HA（高可用性）所保护的VM自动重启（检测群集的其他ESXi服务器之后15 ~ 18秒）。作为HA群集一部分的所有服务器必须能够访问同一个共享存储空间。服务中断和应用失效得以减少，因为重启自动进行，不需要管理员的干预。

注意： 如果失效期过长，可以使用vSphere FT（容错）将失效期降低为0。

1. vSphere HA组件

vSphere HA为vSphere 5进行了完全的重写，在区分主机服务器实际故障和简单的网络问题方面做了许多改进。HA不再依赖于域名系统（Domain Name System，DNS）服务器，而直接使用IP地址。

vSphere HA的必备组件如下：

- 故障域管理器（Fault Domain Manager，FDM）：替换了以前的Legato自动可用性管理器（Automated Availability Manager，AAM）。这个代理的任务是与群集其他服务器交流有关服务器可用资源和VM状态的信息。它负责心跳机制、VM定位和与hostd代理相关的VM重启。

- hostd 代理安装在主机服务器上。FDM 直接与 hostd 和 vCenter 通信。这个代理对于 HA 的正常运作是必需的。如果该代理不能工作，FDM 搁置所有 HA 功能，等待代理再次运行。
- vCenter Server 负责在群集 ESXi 主机上部署和配置 FDM 代理。vCenter 向选择的主机服务器发送群集的配置修改信息（例如，在群集添加主机的时候）。

注意： 如果 vCenter 不可用，vSphere HA 是自治的，确保 VM 的重启。然而，没有 vCenter，就不可能修改群集的配置。

注意： 在 VM 的客户 OS 中没有安装任何可帮助 vSphere HA 运作的代理。

2. 主机服务器和从属服务器

创建一个 vSphere HA 群集时，FDM 代理部署在群集的每台 ESXi 服务器上。一台服务器被选为群集的主（master）服务器，其余的都是从属（slave）服务器。主机服务器的任务是监控组成群集的主机状态和检测事故。主机服务器保存受 HA 保护的 VM 列表，负责在群集中安置 VM，在主机服务器损坏的时候重启 VM。它还验证 VM 是否真正重启，与 vCenter 直接交换信息。

如果主机服务器损坏或者重启，会选择另一台主机服务器。主机服务器的选择在群集中 vSphere HA 第一次激活的时候发生，新的选择发生在如下情况：

- 主机服务器故障
- 主机服务器与网络隔离或者被分区
- 主机服务器与 vCenter 失去联系
- 主机服务器进入维护或者待机模式
- 重新配置的 HA 代理

新主机的选择依据是哪一台服务器连接的数据存储最多，如果数据存储的数量相等，则比较哪一台服务器的管理对象 ID 最高。

从属服务器保持它们的 VM 状态为最新，并向主机服务器通知变化。从属服务器还发送心跳信号以监控主机服务器的健康状况，在当前主机服务器故障的时候参与选择主机服务器。如果主机服务器故障，从属服务器负责重启主机服务器的 VM。

注意： 在 vSphere 4 中，HA 群集使用 5 个首选主机，其他的所有主机都是辅助性的。这在使用刀片服务器时造成了重要的局限。确实，当 5 个首选节点都在同一个机架上时，这个机架失效就会导致 HA 无法运作，因为没有可用的主要节点，不可能重启任何 VM。

5 个首选主机的另一个局限是在处于两个数据中心之间的延伸群集中使用 HA。因为首选主机最多为 5 个，且无法知晓它们在何处选择，所以每个数据中心使用的 ESXi 最多为 4 个，以确保至少有一个首选主机可用于其他数据中心以重启 VM。因此，群集被限制在 8 个主机。有了主 / 从概念，这一限制就不存在了，如果主机服务器故障，选择另一台主机服务器，不管它在哪个位置。

3. 心跳信号

如图 5-4 所示，HA 群集的 FDM 代理通过私有信息交换（称作心跳）相互通信。

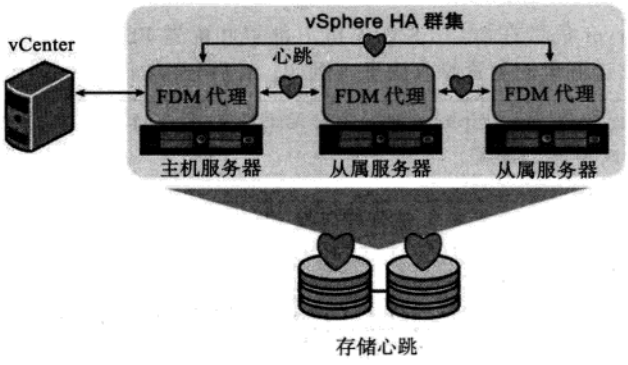


图 5-4　FDM 代理通过心跳通信

心跳一词指的是确定服务器仍然正常工作的一种机制。每个从属服务器向主机服务器发送心跳，主机服务器也向每个从属服务器发送心跳，这种情况每秒发生一次。当主机服务器不再从一台从属服务器接收心跳，就意味着网络通信被破坏，不一定是从属服务器出现故障。为了验证从属服务器仍在工作，主机服务器用两种方法检查其“健康状况”：

- 向从属服务器管理 IP 地址发送一个互联网控制信息协议（Internet Control Message Protocol，ICMP）Ping。
- 在数据存储级别进行信息交换（称作数据存储心跳，datastore heartbeat），如图 5-5 所示。

第二个通信渠道能够区分在网络上被隔离和完全崩溃的从属服务器。

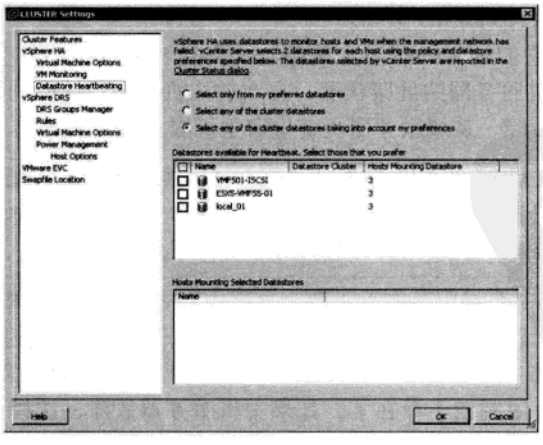


图 5-5　数据存储心跳选项

数据存储心跳依靠 VMFS 卷的元数据（称作心跳区域，heartbeat region）工作，该数据定期更新。为了定期更新这个区域数据，主机必须在该卷上打开一个文件。因此，如图 5-6 所示，HA 创建一个特殊文件，格式如下：host-number host-hb。

每台主机在数据存储上有一个专用的文件。在 NFS 卷上，每台主机每隔 5 秒写入 host-xxx-bb 文件。为了验证这一操作，主机服务器验证该文件是否可用。

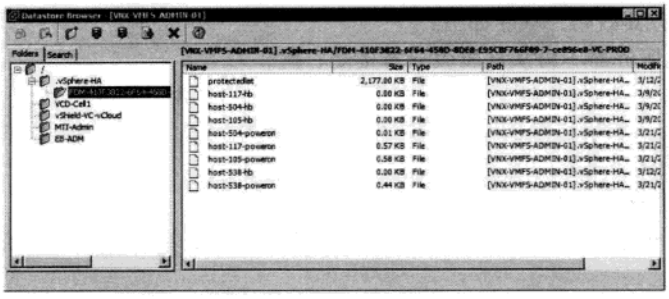


图 5-6 vSphere HA 心跳文件

vSphere 在每个数据中心根目录下创建一个文件夹，用于心跳和受保护 VM 的一致性。正如你在图 5-6 中所看到的，这个文件夹的名称为 .vSphere-HA，里面有几个文件：

- ❑ host-xxx-hb 文件：用于数据存储心跳。
- ❑ host-xxx-poweron 文件：跟踪每个主机运行的 VM。poweron 文件还在从属服务器与网络隔离时通知主机服务器。（如果此参数为 0，意味着没有被隔离；如果参数为 1，就是被隔离。）
- ❑ protectedlist 文件：表示 HA 保护的 VM 列表。主机服务器使用这个文件保存 HA 保护的各个 VM 的库存和状态。

提示： *不要删除或者修改 .vSphere-HA 文件夹及它所包含的文件。这样做可能使 HA 不可用。*

4. 不同的状态

当主机服务器不再能直接与某个从属服务器的 FDM 代理通信，但是心跳数据存储有应答，该服务器仍然正常运作。在这种情况下，可以认为从属服务器从网络上被隔离或者分区。

不再能接收到主机服务器心跳，且主机服务器也无法 Ping 其管理 IP 地址的服务器被认为是隔离的。群集中多台服务器被隔离，但是能通过管理网络进行通信的情况被称为网络分区（Network Partition）。发生这种情况时，第二台主机被选为同一个分区网络中的主机服务器。可以有多个分段，每个分区有一个主机服务器。例如，可以有 3 个具有不同主机服务器的分区。通信被重新建立之后，只会保留一台主机服务器，其余的服务器再次变成从属服务器。

当一台主机被隔离，HA 根据定义的参数 Power Off 或者 Shut Down（如果安装了 VMware Tools）强制 VM 停止，并且在群集的其他主机服务器上启动 VM 的重启。这种行为可以修改，选择 Leave Powered On（选项如图 5-7 所示），VM 可以保持运行，这使 VM 能够继续运作。

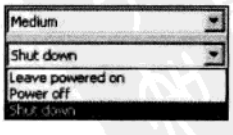


图 5-7 HA 群集隔离响应

推荐： 在 IP 存储架构中（iSCSI 或者 NFS），最好不要使用 Leave Powered On 选项，否则可能发生“脑裂”（split brain）现象，即 VM 可能在两台主机上同时启动，从而导致出错。然而在 vSphere 5 中，这种风险得以减小，因为 HA 在检测到脑裂现象时会自动停止原始 VM。

如果主机服务器和从属服务器没有建立通信（心跳网络和心跳数据存储），后者被宣告为已经损坏。HA 机制在另一台主机服务器上重启 VM。

如果从属服务器故障，顺序如下：

- T0：主机服务器不再从从属服务器接受心跳。
- 3 秒之后，主机服务器发送心跳数据存储，持续 15 秒。
- 10 秒之后，如果心跳网络和数据存储没有应答，该主机被宣告为不可达，主机服务器 ping 从属服务器的管理服务器，持续 5 秒。

这时可能发生两种情况：

- 如果心跳数据存储未配置，主机在 15 秒之后被宣告死亡，根据定义的参数启动 VM 的重启。
- 如果配置了心跳数据存储，主机在 18 秒之后被宣告死亡，根据定义的参数启动 VM 的重启。
- 如果主机服务器故障，顺序会有所不同，因为在 VM 重启之前必须选择新的主机服务器：
- T0：从属服务器不再接受主机服务器的心跳。
- 在 10 秒之后，从属服务器启动新主机服务器的选择。
- 25 秒之后，选择新主机服务器。（该选择根据哪一个主机有最多数据存储做出。）它读取包含所有受 HA 保护的 VM 的 protectedlist 文件。
- 35 秒之后，新的主机服务器启动 protectedlist 文件中没有运行的所有 VM 的重启。

5. 重启优先级

当 HA 机制触发时，vSphere HA 遵循指定的顺序重启 VM。VM 根据图 5-8 所示的 VM 重启优先级重启。可能的值为 High（高）、Medium（中，默认值）、Low（低）和 Disabled（禁用）。

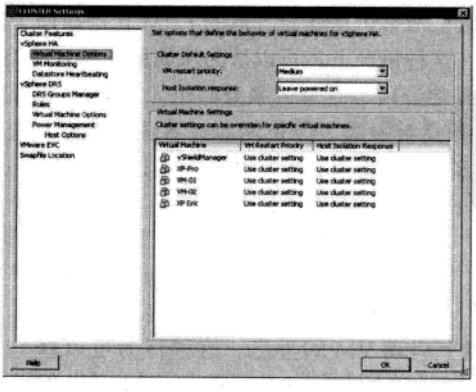


图 5-8　选择 VM 重启优先级

重启顺序非常重要，因为有些服务必须在其他服务之前重启。例如，域控制器必须首先启动。VMware 建议将提供最重要服务的 VM 的重启优先级设置为高。

注意： vSphere HA 不会在一个给定主机上同时发出超过 32 个并发的开机任务。如果一台主机损坏，该主机包含 33 个虚拟主机，全部具有相同的重启优先级，则只会启动 32 个开机请求。不管开机任务成功还是失败，一旦某个任务完成，vSphere HA 将会发出第 33 个虚拟机的开机任务。

具有高优先级的 VM 最先启动。如果某个 VM 被禁用，它在主机故障的时候不会被重启。如果出现故障的主机数量超过了容许控制规范，低优先级的 VM 可能无法重启，因为它们没有必要的资源。

6. VM 和应用程序监控

VM 监控设置（见图 5-9）使用主机服务器和 VM 的 VMware Tools 功能之间的心跳交换确定 VM 是否不能正常工作。这种交换每隔 30 秒执行一次，如果未发生交换，则重启该 VM。

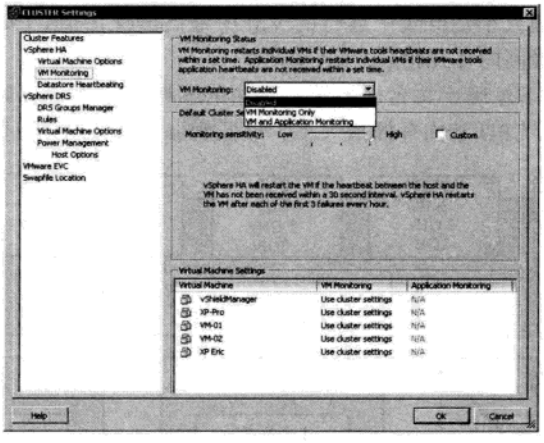


图 5-9　VM 监控设置

注意： 当 VM 中的应用崩溃或者 VM 网卡失效时，VM 不会重启，因为心跳交换是内部的，不会检测到这类事故。但是，如果客户 OS 崩溃，VM 监控会检测到，VM 将被重启。

为了检测客户 OS 上的应用崩溃，vSphere 5 提供一个应用识别 API（vSphere HA 的新模块）。利用它可以为应用开发健康监控解决方案，使 HA 在必要时启动 VM。有可能出现 VM 工作正常，但是应用运行不正常的情况。在 CLUSTER Settings 对话框中选择 VM and Application Monitoring（VM 和应用程序监控），可以在接收不到应用心跳时重启 VM。

Nerverfail（vAppHA）和 Symantec（Application HA）使用了这个 API。这提供了一个增加关键应用服务水平的简单解决方案，很好地代替了难以进行日常管理的群集解决方案，还能管理不能识别群集的应用。

注意：要监控和重新启动能识别群集的应用，可以使用 VM 中的 Microsoft 群集服务器（MSCS）类群集。（先决条件是使用原始设备映射物理 [RDMp] 磁盘。）

7. 确定 HA 群集中可能出现故障的主机数量

当 HA 群集中一台或者多台服务器发生故障时，剩余服务器的所有资源必须能够接管需要迁移的 VM 工作负载。Current Failover Capacity（当前故障切换容量）设置确定 HA 群集中多少主机服务器出现故障时，仍能保证有足够的插槽满足所有运行中 VM 的需求。

为了确定 HA 群集中可以出现故障的服务器数量，HA 使用插槽大小（slot Size）。插槽大小确定每台 ESXi 服务器所能接受的必要 CPU 和内存资源。

插槽大小的计算如下：

- 对于 CPU，插槽大小是群集中最高的 VM 保留值。（如果没有保留，该值默认被设置为 32MHz，但是可以在高级设置 das.vmCpuMinMhz 文件中修改。）

注意：在 vSphere 4 中，默认的插槽大小值为 256MHz。

- 对于内存，插槽大小是群集中运行的 VM 中最高的保留值。（如果没有保留，该值被设置为 VM 的最高内存开销。）

确定插槽大小时，HA 确定每台服务器上可用的插槽大小数量。最大的插槽数等于主机服务器上资源数量除以 CPU 和内存插槽大小。最低值被保留（例子见图 5-10）。

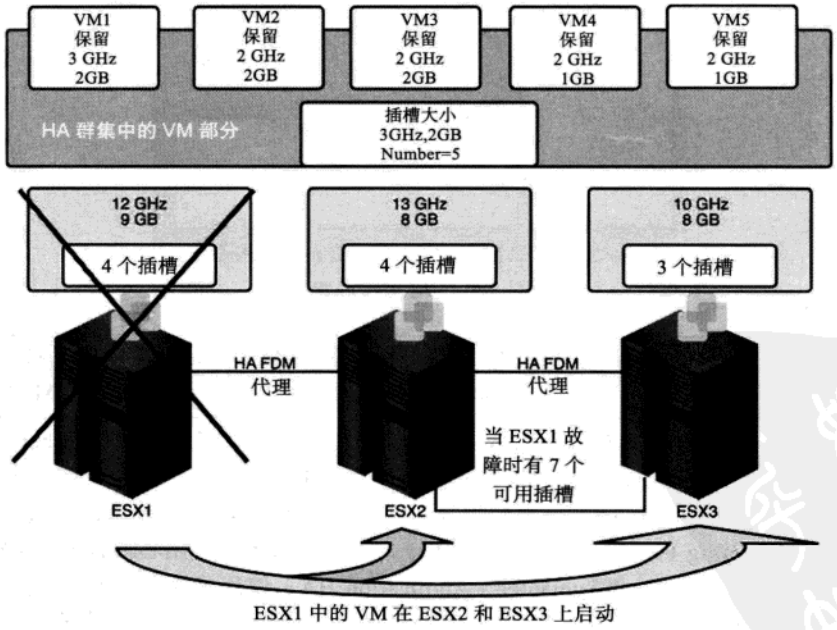


图 5-10　插槽大小示例

示例： 如图 5-10 所示，5 个 VM 正在运行（对应于群集中 5 个必要的插槽）。CPU 最高保留值为 3GHz，内存最高保留值为 2GB，因此插槽大小为 3GHz 和 2GB。

ESX3 服务器的可用插槽数如下：

CPU=10GHz ÷ 3GHz=3

内存 =8GB ÷ 2GB=4

因此，ESX3 的可用插槽为 3 个。

如果 ESX1 故障，ESX2 和 ESX3 能够管理 7 个插槽。

如果 ESX1 和 ESX3 出现故障，ESX2 包含 4 个插槽，无法处理工作负载。

在这种情况下，群集的故障切换容量为 1。这意味着如果一台服务器失效，剩余的两台服务器能够接管 HA 群集中所有运行 VM 的工作负载。

注意： VM 可能包含高于其他 VM 的保留值。这可能改变插槽大小的计算。可以在高级选项 das.slotCpuInMHz 和 das.slotMemInMB 中设置处理器和内存最大值。

8. 容许控制

创建一个 HA 群集时，建议配置主机服务器中允许的故障数量。（群集容许的最大故障主机数量设置可以为 31 个节点。）

管理员可以毫无问题地配置少于或者等于当前故障切换容量的值。如果配置的数量大于当前故障切换容量，容许控制（Admission Control）会加以干预，根据定义的参数允许或者禁止某些行动。

启用容许控制时，如果 VM 违反可用性约束就无法启动。为了确保所有 VM 有足够的资源重启，启动 VM、将 VM 迁移到另一台主机服务器或者增加 VM 的 CPU 及内存保留的操作将被禁止。

禁用容许控制时，VM 即使在违反可用性约束时也能启动。

即使总可用插槽数无法处理所有被迁移 VM，新 VM 也可以启动。在这种情况下，VM 根据重启优先级在群集的可用插槽上启动。这样做的风险是某些 VM 可能找不到可用的插槽。

示例： 这个例子引用了前一个例子：

3 台 ESXi 服务器都有可用插槽。

HA 群集中运行 5 个 VM。

当前故障切换容量为 1。

如果启用容许控制，在保证所有 VM 负载的情况下，还可以启动最多两台 VM（7 个可用插槽）。如果禁用了 VM 的容许控制，就可以启动超过 2 个 VM，但是在这种情况下，无法保证所有 VM 有足够的可用插槽启动。

注意： 如果 VM（准确地说是插槽）的数量超过当前故障切换容量设置的值，vSphere HA 会进入警告模式。如果禁用容许控制，vSphere HA 进入常规模式。

9. 最佳实践

下面是我们建议的 vSphere HA 最佳实践：

- ❑ 因为必须保留资源，使用 HA 降低了整合率。因此，HA 只应该用于真正关键的应用。
- ❑ VMware 建议不要在密集使用 Storage vMotion 或者 Storage DRS 的群集中使用 vSphere HA。
- ❑ 为了增强 vSphere HA 的能力，建议在不同的物理网络交换机上为管理网络使用多个物理网卡（最少两个）。

5.2.5 vSphere 容错

vSphere 容错（Fault Tolerance，FT）比 vSphere HA 更高一级。它提供了非常高的可用性，同时保护 VM 免受主机服务器故障的影响。这意味着，即使服务器突然停止运行，也没有服务中断，没有数据丢失，也没有连通性的损失。

FT 硬件解决方案由 NEC 和 Stratus 等制造商提供。这些公司提供所有部件（甚至主板和处理器）均有冗余的服务器。这些服务器内建的芯片集采用保守的原则，让主用和备用模块在两个独立的主板上同时运行同一组指令。这样，即使某个部件故障也能确保服务持续性。这些服务器用于决不能停止运行的极关键应用。vSphere FT 将这种超高可用性技术用于虚拟环境。FT 只能在共享存储上工作，并且需要特殊的网络配置。

1. vSphere FT 的工作原理

vSphere FT 制作给定 VM 的一个副本。如图 5-11 所示，在主用 VM 上发生的所有事情，都被镜像到备用 VM 上。

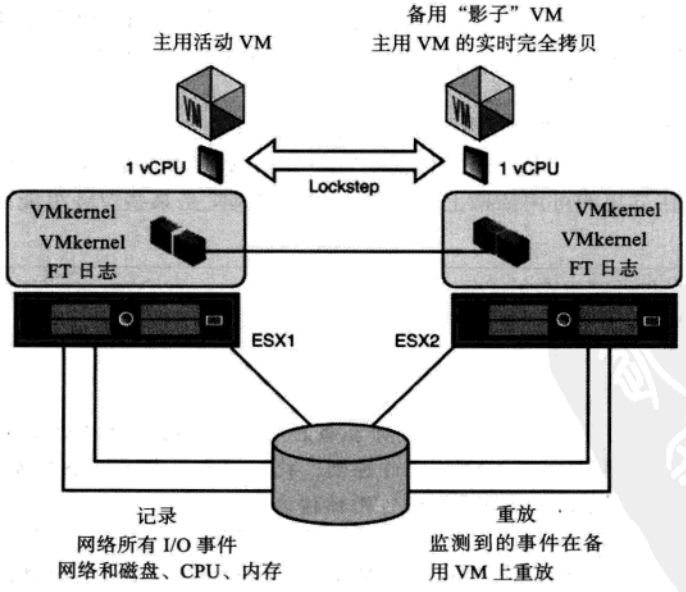


图 5-11 容错的工作方式

vSphere FT 使用 vLockstep 记录 / 重放技术，捕捉主用活动 VM 上的所有事件，如 CPU 指令集、内存状态、I/O，并将它们发送给另一台主机服务器上的备用 VM。这种技术记录所执行的任务并恢复它们。

如果主用服务器失效，保存相同指令集的备用 VM 可以用完全透明的方式接管活动，不会造成服务中断或者数据丢失。

所有信息和日志都通过 VMkernel 的 FT 专用端口（VMkernel FT Logging，vmklogger）发送和接收。

2. vSphere FT 须知

为了使用 FT，客户操作系统应该是标准的，没有内核修改，也不安装特殊的补丁。FT 在 VM 级别上激活。FT VM 及其副本不能运行于同一台主机服务器。FT 使用反亲和性规则，保证两个 VM（主用和备用）永远不会在同一台主机上。这是为了避免一台 ESXi 服务器失效时两个 VM 同时失效。

目前主要的局限性在于，FT 限于单个 vCPU 的 VM。然而，有许多关键应用使用一个 vCPU 就足以应付。我们来看一些 vSphere FT 的用例：

- 小到中型 SQL Server 或者 Exchange 实例
- Web 和文件服务器
- 制造和定制应用
- 基于 SAP NetWeaver 7.0 平台的 SAP ECC 6.0 系统
- BlackBerry Enterprise Server：1vCPU BES 能够支持 200 个用户，100 ~ 200 个电子邮件 / 天

此外，这种技术是资源密集的，VM 的容量被消耗两次（在源 ESXi 上和备用 VM 目标 ESXi 上），需要高性能的专用网络。如果使用多个 VM，建议使用 10GB 网络，这种网络仍然很昂贵。如果目标 ESXi 机器遭遇争用而性能下降，也会影响源 VM 的性能，源 VM 的速度会下降。两个处理器之间的延时必须得到监控。为了避免生产 VM 的性能下降，还必须考虑两台服务器之间的距离、带宽和活动。

与这种功能相关的另一个约束是 FT 目前无法用于具有快照的 VM。这一局限意味着无法通过使用 vStorage API for Data Protection（VADP）（例如 VMware Data Recovery 和类似的软件）进行备份，因为它们的备份过程中使用了 VM 快照。（参见第 6 章中关于 VADP 备份的小节。）要备份 FT VM，必须首先禁用 FT，然后进行备份，在备份过程结束时重新启用 FT。

注意，在激活 vMotion 的改进兼容性（EVC）功能时，vSphere FT 可以与 vSphere DRS 一同使用。

下面是使用 vSphere FT 时需要考虑的其他因素。

- 处理器必须相互兼容以支持 FT 和客户 OS。（支持 VMware FT 的处理器和客户 OS 参见 VMware 知识库：KB1008027。）
- 需要最少两个千兆网卡（建议使用 10GbE 卡）：一个用于 FT Logging，另一个用于 vMotion。
- 所有 ESXi 主机服务器的补丁和版本必须是同一级别。

- 虚拟磁盘模式必须设置为置零厚盘或者 RDMv（仅虚拟兼容性）。
- 不可能使用 Storage vMotion。如果虚拟磁盘需要迁移，必须停止 FT 功能，迁移磁盘，再重新启用 FT。
- 在具有链接复制的 VM 上使用 FT 是不可能的。
- 为了确保冗余和最大的容错保护，VMware 建议在群集中最少有三台主机。在故障切换的情况下，这提供了一台能够容纳新创建的备用 VM 的主机。

5.3 业务持续性

业务持续性（BC）是一组过程和活动，允许准备和实施一些规程，确保事件发生时的信息可用性。灾难恢复计划（DPR）和 IT 应急计划对此也有作用。

BC 的目标是确保公司在影响 IT 系统的灾害发生之后存活下来。它的目标是尽快重启业务（理想的情况下不会造成服务中断），最大限度地减少数据丢失。

在重大危机或者影响 IT 中心的大规模危机发生时，DRP 确保了基础架构的重建以及支持公司运营的应用的重新启动。有了 DRP，目标平台没有共享任何基础架构，从而保证了完全的独立性。备用站点通常位于距离生产站点 30 英里（50 公里）以上的位置。如果主站点不复存在或者无法访问，用户能够被引导到备份站点。通过路由重定向或者指向备份网站的公共地址，通信被切换到备份站点，由其接管生产活动。

IT 应急计划类似于 DRP，但是只考虑同一个 LAN 中的本地恢复。一般来说，源和目标平台所在场地共享电子基础架构、通信线路和用户办公室。

注意： DRP 包含在生产环境重大事件发生时重建基础架构的所有要素。BCP 提供的 RTO 低于 DRP，但是代价更高。

注意： 在 VMware 的术语中，DRP 涉及不同 ESXi 群集、vCenter 和 SRM 的复制。这种技术常用于 VMware。BCP 涉及不同站点之间的高可用性考虑，它需要一个具有扩展 SAN 的架构，提供低延时，与存储虚拟化技术关联。

5.3.1 故障切换起因

触发 DRP 的故障切换的最常见故障起因是水害（例如，来自空调或者管道的水）、电气问题（例如发生通过设备浪涌的故障）、自然灾害（例如洪灾、火灾或者地震）或者蓄意破坏 / 恐怖主义活动。

5.3.2 物理环境中的 DRP 问题

业务重启或者持续性常常涉及错综复杂的需求和过程。在纯物理环境中，这方面的管理极其昂贵和复杂。首先，必须在备份站点上重建源架构，因为生产平台的更新也需要更新目标平台，管理变得更加复杂，给 IT 团队带来沉重的工作负担。

切换到备份站点的规程必须严格测试和验证。例如，必须触发恢复计划的人可能误解文

档，最先起草计划的人可能离开公司，或者在必须应用恢复计划的当天缺席。

因此，必须每年对备份站点的切换进行多次测试。这能使团队训练和验证规程，保证数据完整性。

在物理环境中，这些架构通常是为具有强大财务能力的大型企业设计的，它们的应用关键性值得这样的投入。

5.3.3 vSphere 5 对 DRP 的影响

因为 VM 是从底层硬件抽象而来的，BCP 和 DRP 也得到了简化。因为不必创建相同的备份站点，成本也得以降低。由于 VM 封装在文件中，还因为复制机制，在生产站点上更新一个系统会自动更新备份网站上的系统。在虚拟化环境中，DRP 可以使用 VMware 站点恢复管理器（SRM）测试，这是一个用于重启规程的调度器。这样就减少了错误和不正确的操作。重启根据公司内部应用程序的关键性或者重要性排定优先级，顺序安排进行。

统计数字：根据最新的市场调查，在已经切换到虚拟化的公司中，63% 的公司建立了 DRP。

复杂的 DRP 不再只对大型企业实用，也可以用于所有类型的公司。对于许多客户来说，切换到虚拟环境的主要原因就是更容易建立 DRP。

5.3.4 复制

复制（replication）是在备份站点（本地或者远程）上创建生产数据的精确备份（或者重现）的行为。复制可以用于加强存储阵列安全，但是主要用于业务恢复计划。重复的数据必须是可恢复的，当数据在应用级别上一致时，业务可以重启。这允许从复制的数据（称作复制品 Replica）重启操作。确保一致性是复制技术必须考虑的第一个问题。

在备份网站上重启应用程序所需的时间（RTO）取决于应用的一致性。如果一致性不能保证，就必须验证基本完整性或者使用备份，这就延长了 RTO。

谈到复制，关键的因素是延时和写入修改速度。

1. 复制解决方案

下列技术可以在不同级别上进行数据复制：

- 操作系统：这种类型被称作基于主机的复制，使用 Double-Take、Neverfail 或者 Legato 等复制软件。这种技术的缺点是在 OS 中造成开销，以及恢复是崩溃一致（Crash consistent）这一事实。尽管这种机制的表现不错，但是复制是异步的。因为源机器和目标机器都是活跃的，所以管理操作系统的工作量加倍。重复工具版本必须随着 OS 的发展而更新。这类解决方案的优点是成本相对低。
- 应用恢复：这种技术基于制造商的复制机制。例如，Microsoft SQL（SQL Mirroring、Log shipping，SQL Server AlwaysOn），Microsoft Exchange[本地持续复制（Local Continuous Replication，LCR）、备用连续复制（Standby Continuous Replication，SCR）等]，Lotus Domino（Domino Clustering）和 Oracle（DataGuard）都有确保数据一致性的复制机制。同一个应用中有多种类型的 RPO（同步或者异步）。这类复制

的优点是确保复制数据的一致性，缺点是需要专业知识：每个应用使用不同的复制机制，需要管理员使用多种工具。

注意：有些复制的数据库偶尔需要编写与数据库通信的应用，以便理解和支持应用故障切换。

- 低级（硬件）复制：这种技术往往是性能最好，也最健壮的。通常由存储阵列自身提供。其他解决方案基于用具、服务器中安装的硬件卡或者物理交换机（I/O 分割）。硬件复制的好处是在异构环境中使用单一工具管理复制。因为复制在存储阵列中的数据块级别上进行，所以没有操作系统或者应用兼容性问题。使用这类复制时，必须监控应用一致性。下面是目前可以利用的主要解决方案。
  - EMC：Symmetrix Remote Data Facility（SRDF）、Mirror View、RecoverPoint、VPLEX
  - HDS：TrueCopy Synchronous Remote Replication、Universal Storage Platform VM
  - IBM：TotalStorage、点对点远程拷贝（Peer to Peer Remote Copy，PPRC）Global Mirror 和 Metro Mirror
  - HP：StorageWorks Replication Manager、3PAR Remote Copy
  - NetApp：SnapMirror 和 MetroCluster
  - FalconStor：连续数据保护（Continuous data protector）

vSphere 中使用的复制技术一般是存储阵列复制。然而在 vSphere 5 中，和 SRM5 相关的一种新特性能够在 ESXi 主机级进行复制。这种功能称为基于主机的复制（Host-Based Replication，HBR），在 VM 级（而不是整个逻辑单元号）的粒度上进行复制。

2. 同步复制与异步复制的对比

如图 5-12 所示，有两类复制：同步复制（提供的 RPO 为 0）和异步复制（提供的 RPO 为几分钟）。

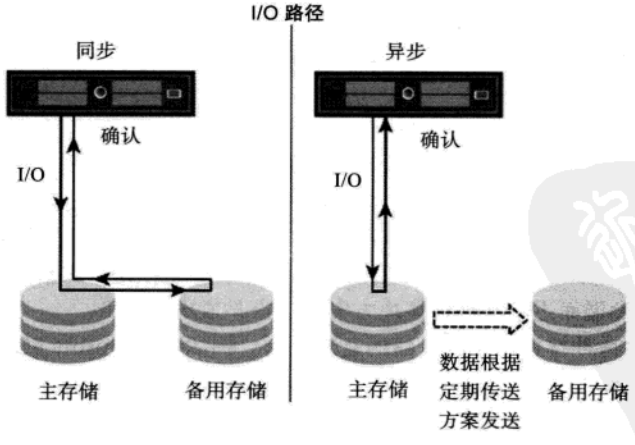


图 5-12　同步复制与异步复制的对比

同步复制确保源阵列写入的每一部分数据都写入到目标阵列。当I/O发送到生产（主）阵列，也会自动发送到目标阵列。主机级的确认仅在I/O写入备用阵列的时候才执行。使用这种机制，数据总是相同的，不会发生数据丢失。如果主站点上发生事件，备用阵列上的数据与主阵列完全相同。

同步复制机制有一个后果。I/O执行任务所需的时间（发送I/O直到接收确认）被称作延时（latency）或者往返时间（Round-Trip Time，RTT）。这一时间必须很短（通常在4 ~ 6毫秒），否则生产应用将会遭遇性能下降。这就是被复制站点的距离受到限制的原因。（理论上的最大距离为125英里（大约200公里），但是实际上这一距离通常不超过几十英里。）

注意： 可能造成很长的延时的因素之一是网站之间的带宽（管道）。设备越少，延时也越短，所以改进这一指标的方法之一是使用阵列－阵列之间的连接，而不通过交换机。

异步复制提供更快的确认，因为确认发生在I/O写入源阵列时。这就不需要等待备用站点返回I/O，所以没有延时问题，也就没有距离的限制。数据本地存储在源阵列上（在规划好的缓存中），然后以几分钟至几小时的间隔定期发送到备用阵列。RPO不为0，而是几分钟到几个小时。

为了正确定义应该使用的复制类型，必须知道写入应用负载或者修改率，如图5-13所示。

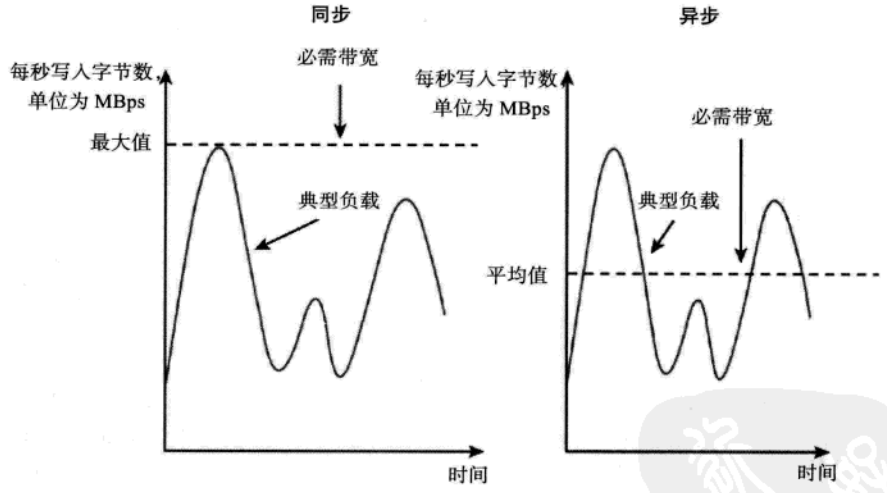


图5-13　对于同步复制，两个站点之间的带宽至少要高于应用负载的峰值，以避免性能问题。如果带宽低于峰值，必须使用异步复制

复制可以通过SAN FC或者SAN IP进行。有些客户在生产环境中使用SAN FC，而为远程复制使用10Gbit网络上的iSCSI。

记住，复制的目标是重现主站点的数据，它不能保护损坏数据的删除错误，因为这种错误会传播到备用站点。（复制品只能代表被复制数据的最后映像。）在这种情况下，必须采用

快照或者备份恢复到健康状态。

其他一些机制能够提供持续数据保护（Continuous Data Protection，CDP）。这类解决方案非常有趣，不仅数据被复制，而且实施了写入日志。通过日志，可以返回到过去的任何时间点，这能保护系统免遭存储失效，而且能够保护存储设备中的数据损坏或者丢失。

5.3.5 SRM 5

VMware 站点恢复管理器（Site Recovery Manager，SRM）是 vSphere 环境中的自动化业务恢复解决方案，是 DRP 的一部分。这个软件独立于 vSphere 5 许可证销售，但是因为它作为插件提供，所以能够完美地与 vCenter 集成。SRM 5 是一个任务调度程序。它自动化了远程站点上的业务恢复（故障切换）和回溯（故障恢复）。利用 SRM，可以定义 VM 的重启顺序，如果源站点和备份站点之间有一个网络交换机，可以改变 VM 的 IP 地址。它使管理员能够在不造成任何服务中断的情况下，用隔离网络（测试 VLAN 或者气泡网络）测试 DRP。SRM5 提供运行时的精确报告，指明每个基本任务的成功或者失败，详细地说明它们的执行情况。接下来，这允许技术团队分析这些结果并验证这个交换过程。SRM 也用于规划维护操作（例如，如果生产数据中心的电源中断），将生产环境从一个数据中心迁移到另一个。

SRM 还能监控被复制的数据存储，如果数据存储中的某个 VM 没有在恢复计划中就会发出警告。

SRM 有两个可用版本——标准版（最多支持 75 个 VM）和企业版（不限制 VM 数量），提供如下功能：

- ❑ vSphere 复制或者存储阵列复制
- ❑ 无服务中断测试
- ❑ 自动故障切换
- ❑ 自动故障恢复（SRM 5 新功能）
- ❑ 计划迁移

1. 架构

图 5-14 说明了具有存储阵列级复制的 SRM 5 典型架构。

SRM 5 的使用需要如下条件：

- ❑ 两个 vSphere 群集，一个 vSphere 生产群集由 vCenter 管理，备份群集由另一个 vCenter 管理。两台 vCenter 相互连接，用 SRM 插件组成对等链接。这样，两台 vCenter 可以互相监控资源可用性和数据中心的物理可用性，还允许 SRM 实例与作为备份的 vCenter 通信，从而操纵 VM 和库存资源等对象。
- ❑ 两台 SRM 5 服务器连接到每个站点的数据库。
- ❑ 两个存储阵列，一个用于生产站点的主阵列和一个复制数据的备份阵列。存储阵列必须在具有存储复制适配器（Storage Replication Adapter，SRA）的 VMware 兼容性矩阵中验证。这些阵列还必须与复制的阵列兼容。如果没有提供快照和复制产品，它们都必须配备附加的快照许可证（用于测试）和复制许可证。
- ❑ 两个存储阵列之间具有数据复制解决方案的站点间链路。

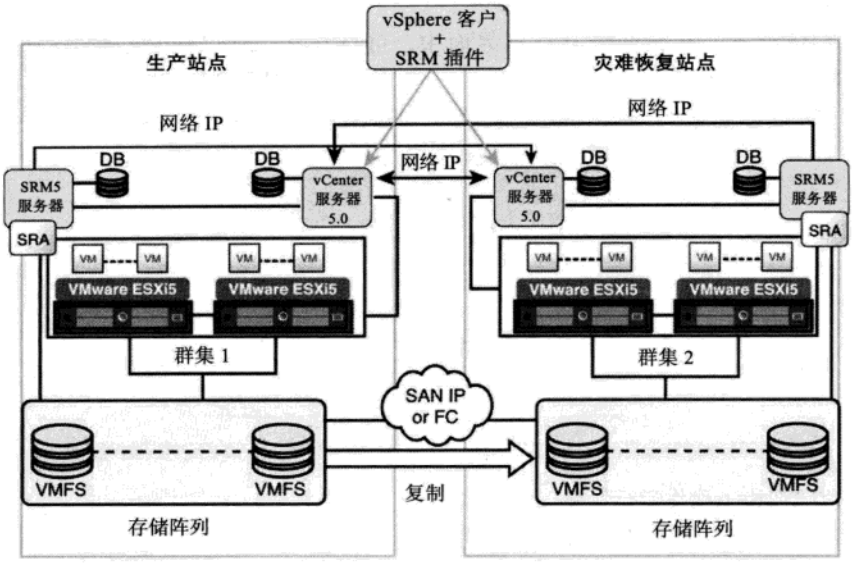


图 5-14　SRM 5 架构

2. 不同模式

SRM 5 可以工作于*主动 / 被动*或者*主动 / 主动*模式。主动 / 主动架构允许在一个站点上运行半数 VM，另一半 VM 则运行于另一个站点。每个站点互为备份。这是首选的模式，因为它减少了备份站点的成本。这种模式也可用于生产环境，而不仅是用在重大灾害的时候。而且，如果遭到灾害，失效只影响 50% 的生产环境，只有 50% 的 VM 需要重启，如果恶化恢复的情况下有些 VM 不是必需的，这个数量还会更少。

以降级模式恢复可以用于限制备份站点的硬件投资。确实，对于 DRP，IT 管理人员可以选择只投资于核心业务必需的关键应用。在这种情况下，应该排定 VM 的优先级，有些 VM 应该排除在恢复计划之外。

3. 存储复制适配器

存储复制适配器（SRA）是 SRM 的重要组件。SRM 不管理存储阵列 LUN。SRA 作为 SRM 和与存储设备 LUN 相关操作的中继。如果一个 VM 跨越多个 VMFS 卷，它还监控不同数据存储之间的一致性。SRA 是确认支持的存储阵列制造商开发的软件，在 VMware 网站上可以找到。通过 SRM，SRA 为 LUN（故障切换 / 故障恢复）快照、复制、断裂（fracture）和提升（promote）操作。

在同一个 SRM 服务器中可以使用来自不同制造商的 SRA。

注意：来自给定制造商的所有阵列不一定都已经验证了与 SRM 5 的集成。请与制造商联系。

4. 故障切换

如图 5-15 所示，在新的恢复计划界面中，可以在 Planned Migration（计划迁移）和 Disaster Recovery（灾难恢复）中选择。

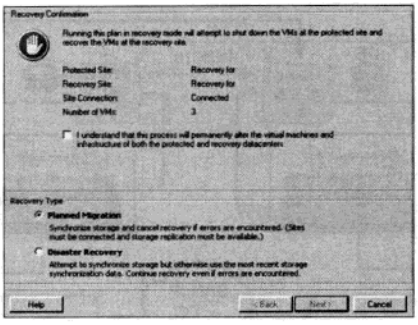


图 5-15 恢复类型选项

启动故障切换（称作运行恢复）的顺序如下：

1）SRM 5 停止第一个站点的 VM（如果主阵列仍然可达）。确实，灾害可能只有部分影响，运行中的 VM 需要“干净”地停止。

2）如果主阵列存在（没有被毁坏），发生镜像反转。

3）SRM 联系 SRA 停止主阵列 LUN 和备份阵列 LUN 之间的复制（称为断裂）。

4）SRM 5 使得备用站点的 LUN 可用于备用站点的主机服务器（执行远程站点服务器上复制的提升）。备份站点的 LUN 变为主用 LUN。

5）备份站点的 ESXi 服务器 HBA 卡执行重扫描以发现 LUN。

6）SRM 将受保护的 VM 输入 vCenter 库存清单，重新在 vCenter 级别上映射 vmx 和合适的资源。

7）SRM 根据配置启动并自动化存储在复制 LUN 上的 VM。VM 逐个重启，两个因素会触发下一个 VM 的重启：

- ❑ VM 的 VMware Tools 特性重启，通知 SRM，并启动下一个 VM。
- ❑ 如果没有安装 VMware Tools，SRM 在超时（默认为 10 分钟）后重启下一个 VM。

提示： 验证 DNS 正常也是必要的。两个站点的 SRM 服务器和 vCenter 必须能够正确地解析自身，并且相互可达，否则 SRM 解决方案无法正常工作。

服务水平有两种类型：已复制数据存储和未复制数据存储。未复制数据存储被排除在业务恢复计划之外，如果生产存储毁坏，需要一个 VM 恢复阶段。

5. VM 启动优先级

在早期版本的 SRM 中，VM 根据三个优先级分组之一启动 VM：高、中和低。SRM 5 有了更多粒度控制，如图 5-16 所示，它提供 5 种优先级分组（优先级组 1 到优先级组 5）。优先级组 1 的 VM 首先启动，该组的所有 VM 启动之后才开始启动优先级组 2 中的 VM。这一

过程持续到所有优先级组 5 的 VM 启动完毕。

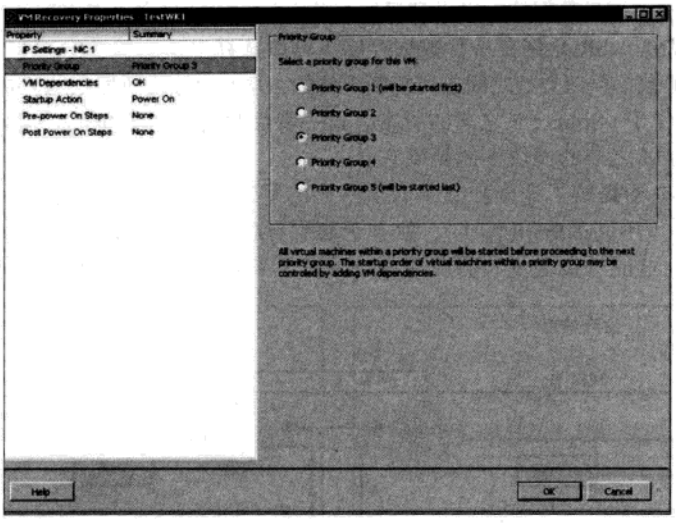


图 5-16 SRM5 优先级组级别

还可以定义一个优先级组中 VM 之间的依赖关系，确保有些 VM 先于其他 VM 启动。这种依赖关系仅适用于同一优先级组中的 VM。

6. 计划迁移

计划迁移在规划维护操作时非常实用（例如，数据中心电源中断）。选中这个选项时，所有 VM 被干净地停止（VMware Tools 关闭 VM）以确保应用一致性。停止后，VM 被同步到最近时间，然后切换到备份站点。

7. 测试

测试模式能在不影响生产的情况下，进行 VM 备份站点的故障切换测试。启动测试模式时，VM 被放在一个隔离网络（测试 VLAN 或者气泡网络），不退出 ESXi 主机。VM 不能从一台 ESXi 主机上与另一台 ESXi 主机通信。如果有些应用需要与其他主机通信，它们必须组合在一起，或者创建一个与基础架构其余部分完全隔离的测试 VLAN。

测试过程开始时，SRA 命令阵列显示运行于备用站点上的镜像快照。这意味着生产环境受到保护，进行测试时保持复制。在测试结束时，ESXi 服务器的快照被删除，不需要重新同步阵列。

使用这些技术隔离网络和存储，测试就可以在业务时间内进行，不会影响到生产。

8. vSphere Replication

SRM 5 有一种复制机制——vSphere Replication（vSR），在服务器级别上进行（一般被称作基于主机的复制）。Replication 的粒度为 VM 级别（最低要求虚拟硬件第 7 版），可以选择被复制的 VM，而不是整个数据存储（传统存储设备复制的情况）。以前的 SRM 版本不包含复制机制。

这种技术是变更数据块跟踪（CBT）的一种形式，需要 VM 的初始完整拷贝。制作拷贝之后，只有修改的数据块（增量数据块）发送到目标存储空间。一次最多可以复制 500 个 VM，总共可以保护 1000 个 VM。vSR 允许在一种数据存储（可以选择本地磁盘、SAN FC、NFS、iSCSI 和 vSA）和任何其他数据存储类型之间进行复制（例如，从 SAN FC 数据存储复制到 NFS）。

vSR 不能工作于 vSphere 5 之前的版本。这一特性只可用于 SRM，不能集成到 vSphere 5（见图 5-17），SRM 之外的模块不能得到它。

vSR 的价值在于提供了完全独立于底层平台的复制解决方案。被复制 VM 可以在异构设备的数据存储中（不同品牌和型号），独立于底层存储协议（只要使用处于 VMware 兼容性矩阵中的协议）。这能够实现复制的虚拟化，开拓了很多可能性，大大增加了管理的灵活性。

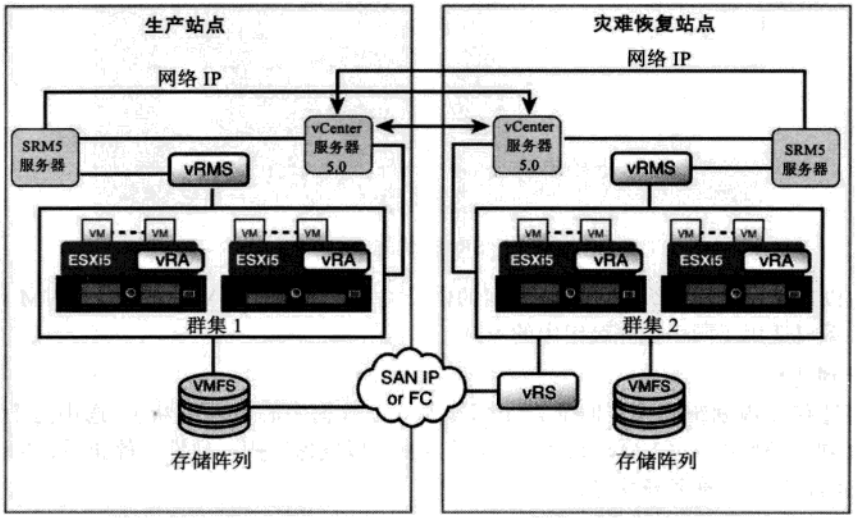


图 5-17　使用 vSphere Replication 的 SRM 架构

HBR 修改为复制建立的架构。需要如下条件：

- 每个站点有一个包含复制配置的虚拟管理用具 vSphere 复制管理服务器（vSphere Replication Management Server，vRMS）。需要一个 SQL Server、Oracle 或者 DB2 数据库。vRMS 在 vCenter 中被记录为一个插件，管理 vRA 代理的部署。
- 每台 ESXi 主机有一些 vSphere 复制代理（vSphere Replication Agents，vRAs）。vRA 负责监控修改标记为“被保护”的 vDisk 中数据块的写入 I/O，并将这些修改发送到备份站点。
- 备份站点上的一台 vSphere 复制服务器（vSphere Replication Server，vRS），接收受保护网站 vRA 的修改。

注意： vRS 对于复制来说是自治的。它在 Center Server、SRM 或者 vCenter 出现故障时也能继续运行。

这种复制限于初始版本，在高 SLA 需求的情况下不建议使用，只支持异步模式，最小 RPO 为 15 分钟，最长为 24 个小时。这些限制使得这种解决方案达不到最佳的效果，在如下情况它会成为一个障碍：

- ❑ 在服务器级别有开销。主机必须完全用于 VM 和应用。
- ❑ 不可能复制物理 RDM 模式、VM FT、链接复制中的磁盘，不支持存储 DRS。
- ❑ 只能复制运行中的 VM。被停止或者挂起的 VM、ISO 映像和模板不能复制。
- ❑ 第一个版本不提供压缩或者加密。

阵列提供的低级复制机制减轻了主机服务器在密集 I/O 任务上的负载。阵列集成了更多健壮的特性，提供更多可能性，如同步复制、故障恢复期间的数据块修改同步（阵列比较数据块）以及减少网站间带宽需求的压缩及重复数据消除技术。

阵列复制提供 LUN 级的粒度，允许创建用于应用一致性的相关分组。

提示： *这种复制解决方案适应小环境，但是不适合于企业级环境。对于大规模的环境，最好使用存储阵列复制机制，以确保未来的伸缩性。*

---

9. 故障恢复

故障恢复（failback）即回溯，是毁坏的主站点恢复之后进行的操作。它能够将 DRP 期间切换的所有 VM 恢复到原始位置。故障恢复比故障切换更复杂，所以正确地配置和执行需要更多的步骤。SRM 以前的版本不提供故障恢复，所以必须使用第三方解决方案（值得注意的是由阵列制造商提供的解决方案）。在这个版本的 SRM 中，故障恢复是标准功能。

10. SRM 5 界面

SRM 5 的界面进行了修改和增强，简化了受保护 VM 的显示和标识（见图 5-18）。

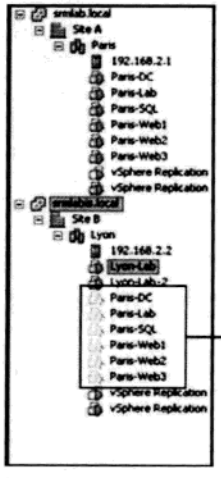


图 5-18 SRM 5 界面中受保护 VM 的显示更简单

在存储阵列级别上，如图 5-19 所示，可以看到复制的方向。

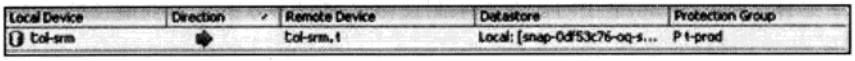


图 5-19　SRM 5 中复制方向的显示

5.3.6　延伸群集

延伸群集（stretched cluster）能够将一个群集扩展到两个物理上不同的站点，提供两个数据中心之间的高可用性，并在两个数据中心之间传播工作负载。

建议添加一个分布式存储虚拟化解决方案（例如 EMC VPLEX 或者 NetApp Metro Cluster），和传统复制解决方案相比，这能为管理员提供更多的灵活性。图 5-20 展示了典型的架构。

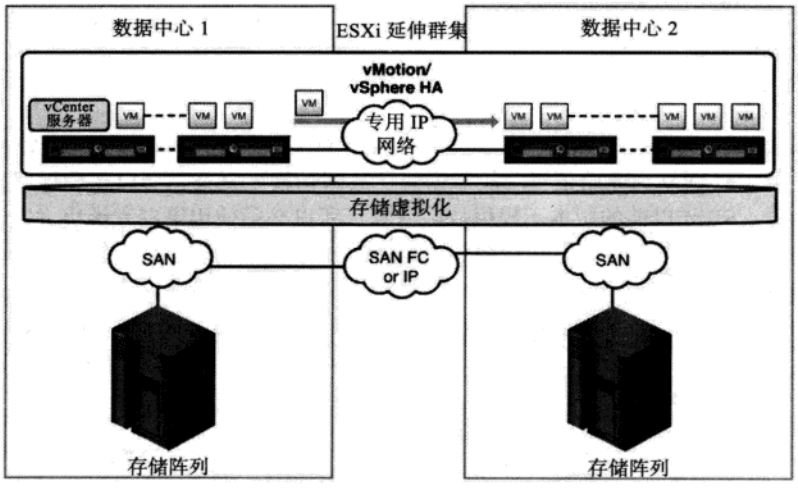


图 5-20　延伸群集架构

在这个架构中，VM 可以使用 vMotion 从一个数据中心迁移到另一个数据中心。存储虚拟化导致延时的减少，能够在没有任何服务中断的情况下将 VM 从一个数据中心迁移到另一个数据中心。因此，可以在一个数据中心的计划维护操作期间得到保护（称作灾难避免，disaster avoidance），例如，一个数据中心的电气故障。当数据中心位于多个大陆时，生产环境还可以移动，以靠近用户。

vSphere HA 也可用于延伸群集，保护生产站点免于完全崩溃，强制 VM 在备份站点上重启。

运行于生产模式时，如图 5-21 所示，来自数据中心 1 的 VM 在数据中心 1 的主机服务器上运行。存储虚拟化解决方案向群集的 ESXi 主机服务器显示虚拟 LUN。每次写入都在两个存储阵列上进行。读操作在最靠近 ESXi 主机服务器的存储阵列中进行。

使用 vMotion 时，如图 5-22 所示，VM 通过专用千兆网络迁移到备用站点。因此，第一个数据中心的部件可以停止，进行维护操作。

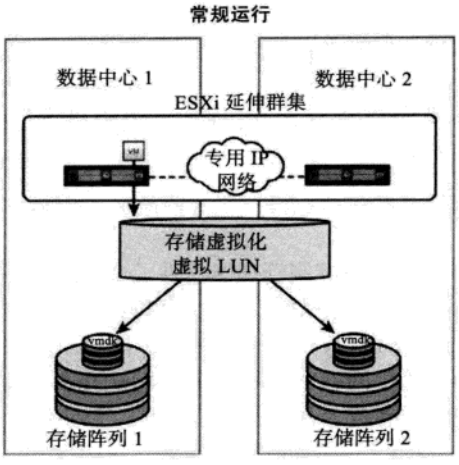


图 5-21 常规运行条件下的延伸群集

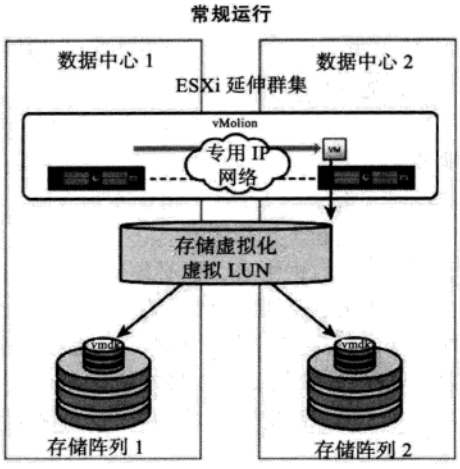


图 5-22 使用 vMotion 在延伸群集架构中进行维护操作

注意： 不使用虚拟化层，可以在延伸群集中执行 vMotion。在这种情况下，迁移到第二个站点上的 VM 访问位于第一个站点上的虚拟磁盘。但是，为了使 VM 能够访问第二个阵列上的 vmdk 虚拟磁盘，必须断裂复制，LUN 必须提升到第二个站点的 ESXi 服务器。

1. 远距离 vMotion

vMotion 传统上在同一个数据中心中使用。在延伸群集的情况下，远距离 vMotion 有如下先决条件。

- 最小带宽 622Mbps 的 IP 网络。
- 为了 vMotion，vSphere 5 服务器之间的最大延时为 5 毫秒，Metro vMotion 的最大延时为 10 毫秒。

注意： VMware 只在源 ESXi 和目标 ESXi 之间的延时小于 5 毫秒 RTT 时才支持 vMotion。然而，在 vSphere5 中，Enterplue Plus 版本有一个新特性 Metro vMotion，支持 10 毫秒 RTT 的延时。

- 源和目标必须在一个具有相同 IP 子网和相同广播域的专用网络上。
- 源和目标 ESXi 服务器必须能够访问 VM 运行的 IP 子网。这是一个重点，为了保持与外部元素的通信，使用 vMotion 迁移时，VM 保持相同的 IP 配置。

注意： 延伸群集不是远距离 vMotion 的先决条件，因为 vMotion 可以在两个不同群集之间进行。然而，在这种情况下，需要两个单独的 vCenter Server，导致 VM 群集属性的丧失（例如，HA 重启优先级和 DRS 设置）。

2. 延伸群集中的 vSphere HA

可以在延伸群集中运行 vSphere HA，在重大事故影响第一个站点时，在第二个数据中心上启动 VM 的重启。在这种环境下，HA 根据定义的优先级重启 VM 。然而，这不能提供经典 SRM 的重启控制粒度（例如改变 IP 地址或者根据工作流重启）。

3. 预先考虑意外情况

对于大部分公司，是否遇到灾害并不重要，重要的是何时遇到灾害。vSphere 提供许多选项最大限度地减少事件的影响，主要的特性是利用 VM 的可移植性。vSphere HA 可以在主机故障时在健康的 ESXi 主机上重启 VM。这是个很好的特性，但是如果你需要保护整个数据中心，它就起不了作用。

对于整个数据中心，你必须使用 VMware SRM 5。站点复制管理器提供了全面的管理和调度解决方案。利用 SRM5，你可以为受保护 VM 分组，修改 IP 地址以保证与目标网络的兼容性。和备份类似，高可用性中最重要的事情是确保 DRP 得到合适的测试。任何未经测试的 DRP 都是无效的，所以你的机构必须确保基础架构、网络、应用程序和业务团队一起验证计划。确保成功的关键是着眼于灾难恢复策略的组织和技术方面。

第6章
vSphere 5中的备份

转移到虚拟化环境提供了对备份架构进行深入评估的机会。备份技术和方法必须适应虚拟环境，因为这种环境涉及的问题独特而复杂，虚拟机（VM）的备份方式通常与传统服务器不同。备份的成功实施揭示了虚拟化的一些好处，也是降低成本的一个手段。反之亦然，不合适的备份方法可能对整个信息系统产生致命的后果。

6.1 备份概述

在详细介绍虚拟环境中可能发生的关键备份问题之前，我们在本节中定义一些基本概念。

6.1.1 什么是备份

备份（backup）是生产数据的精确（物理）拷贝，创建备份的目标是恢复任何损坏的数据。备份拷贝可以用不同的方式创建：

- 复制给定时刻的数据
- 数据复制（即使原始数据修改，备份数据也总是与原始数据相同）

6.1.2 备份的目标

备份有三个主要目标。

- 操作性目标：在意外的数据删除、丢失或者损坏（例如病毒攻击）时恢复数据。
- 存档：为了管理或者法律遵从性的原因保留数据（例如，电子邮件和联系人）。
- 业务恢复计划：为了在重大灾害时恢复数据。

6.1.3 业务影响

始终记住，备份的最重要方面是恢复！我们常常看到IT经理在建立备份系统时没有考虑如何恢复数据。下面是对业务有直接影响，必须回答的重要问题的一个不完全列表：

- 恢复点目标（RPO）是什么？
- 恢复时间目标（RTO）是什么？
- 备份哪类数据（例如，应用、OS、数据库、消息、普通文件、日志文件、数据库日志文件、文件系统或者视频）？
- 使用什么样的备份窗口？
- 何时、在哪里进行恢复？最经常进行的是哪些恢复？
- 何时备份应用程序，备份是否真的保持一致？
- 应该实施什么样的保留策略？应该保存多少每日、每周、每月和年度备份？
- 是否应该异地存放备份？
- 备份的日平均数据修改率是多少？
- 数据是否加密或者压缩？

6.1.4 传统备份方法

备份代理是物理环境中使用的传统方法，包括安装一个应用代理，其目标是在备份之前保持应用一致性，以便产生一个可恢复映像。在指定的时刻，代理通过网络向服务器发送数

据，在选择的介质（如磁带或者硬盘）上备份。

注意： 历史上，备份使用磁带，这种介质的优点是以低廉的成本获得很大的存储容量，可以在外部场所存放且不存在任何运营成本。硬盘可以通过磁带模拟（tape emulation，称作虚拟磁带库，Vitual Tape Library，VTL）使用，它提供了更好的灵活性和快速的访问，且不需要改变现有的备份架构。

在传统物理环境中，你可以使用备份代理，因为应用所在的服务器通常具有可用的物理资源。（服务器资源平均只使用10%，所以具有可用于备份代理的资源。）

这种方法很有效，但是给IT团队带来了严重的工作负担。它需要部署代理、更新和维护。代理消耗许多服务器资源（CPU、内存、网络和存储），恢复通常很困难（往往较慢，且分阶段进行）。这种方法还需要介质管理，经常遭遇处理错误、遗忘等情况。

6.1.5 虚拟环境中的备份问题

虚拟环境中可能发生如下备份问题，在实施时应该考虑：

- 争用的风险
- 文件关键程度
- 大容量
- 改进的服务水平

下面的小节更详细地研究这些问题。

1. 争用的风险

虚拟化的目标之一是实现高整合度，以降低成本。主机服务器的可用资源完全贡献给应用。备份代理大量使用和消耗这些资源。因此，客户OS中不建议使用代理，因为这会严重影响性能，产生争用。

不过，在某些情况下，对于某些应用（例如数据库和消息应用），有必要在客户OS中安装代理，改进设置的精度，增加恢复粒度。例如，利用消息代理，就能够恢复单个电子邮件消息。

2. 文件关键程度

主vmdk文件代表的虚拟磁盘的完整内容包括操作系统、配置、应用和数据。这个文件很关键，管理员的处理错误（例如意外删除）都可能导致灾难。对此应该谨慎考虑，采用快速恢复解决方案。

3. 大容量

在虚拟环境中，备份VM的完整映像是可能的。比起物理环境中传统备份的数据，这个映像需要的容量很大。因此，你部署的技术必须适应这种容量，所以应该设置备份窗口时间表（通常是晚上8时）。

4. 改进的服务水平

建立虚拟基础架构必须显著地改进服务水平，应该正式建立服务水平协议（SLA）。备份数据应该能够直接和快速访问，恢复应该只花费几分钟。

在这种环境中很容易理解，传统的备份方法（在磁带或者 VTL 上）不是最优化的。必须准备高性能技术，利用虚拟化的特殊优势。

6.2 虚拟环境中的备份方法

本节研究虚拟计算中的不同历史和现行备份方法。

6.2.1 VMware 整合备份简史

因为 VM 封装在文件中，可以备份 VM 的完整映像。VMware 为此在 VMware 基础架构 3（VMware infrastructure，VI3）中开发了 VMware 整合备份（Consolidated Backup，VCB），但是这一功能现在已经被放弃。VCB 不是一个软件，而是用于 VM 执行备份操作的一个过程（称作框架）。它减少了网络流量，因为备份服务器直接访问存储区域网络（SAN）逻辑单元号（LUN），释放主机服务器上的负载。

这种解决方案理论上似乎有效，但是在实践中 VCB 有许多缺点，因为 vmdk 文件需要先复制后才能传送给备份服务器，所以需要很大的缓存空间。此外，这种框架产生不稳定性，而且需要难以开发且功能非常有限的脚本（例如，没有重复数据消除和增量模式，文件恢复步骤难以实施等）。

为了消除这些限制，从 vSphere 4 开始，VCB 被 vStorage API for Data Protection（VADP）所替代。

6.2.2 vSphere 5 中的方法

在 vSphere 5 中，目前有两种备份方法，使用如下工具：

- VADP
- 客户 OS 中的代理

虚拟化改变了备份，因为使用这些方法，恢复可以在多种级别上进行：整个 VM、VM 中的一个文件或者备份应用中的数据。

现在，我们将研究这些方法的工作原理，以及何时使用它们。

1. VADP

通过这些 API，可以在不使用客户 OS 的备份代理的情况下备份 VM 文件的一个映像。备份是“热”实施的（没有任何服务中断），提高了应用程序服务水平。备份应用程序制造商可以开发基于 VADP 的解决方案。

注意： EMC Avamar、VMware Data Recovery、Veeam Backup、Quest vRanger Pro、Symantec NetBackup 等软件原生集成这些 API。

vStorage API 的优点是它们创建 VM 的“安装点”以供备份，服务器可以在不需要缓存空间（VCB 需要缓存）的情况下使用这些安装点。备份得以简化，不再需要脚本的开发和维护。如图 6-1 所示，这一任务可以迁移到专用的服务器（称作代理服务器），将主机服务器的所有资源留给 VM 和应用。

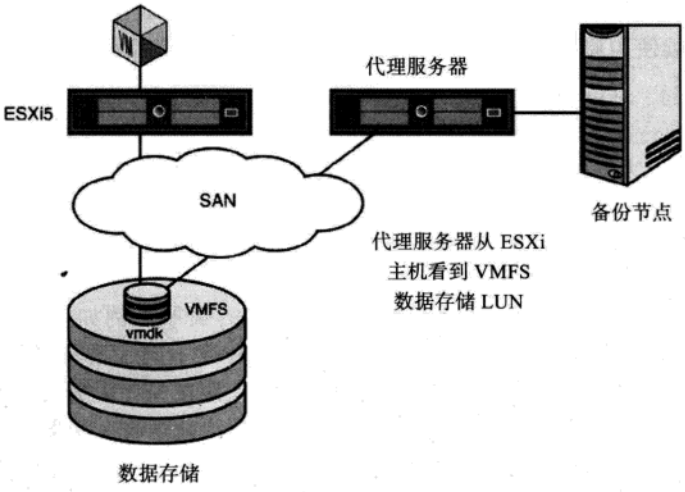


图 6-1 代理服务器必须能访问 VM 所用的数据存储（LUN VMFS、NFS 和 RDM）。备份通过 LAN 或者 SAN 进行

使用 VADP 备份 VM 时，需要本章后面讨论的快照机制。使用 VADP 进行备份保证了某些 Windows 环境中的应用一致性，这归功于卷影拷贝服务（Volume Shadow Copy Service，VSS）。而且，利用变更数据块跟踪，备份和恢复很快发生（见本章后面相关的小节）。这减少了所需的网络带宽，因为只有修改的数据块发送到备份服务器。

VADP 可以进行不同级别的恢复。有些制造商利用 API 提供 VM 完整映像的恢复（全 VM）或者客户 OS 中单独文件的恢复（文件级恢复）：

- 全 VM：允许完整恢复 VM，以及所有相关文件。这种模式的优点是提供整个 VM 的直接备份，包括 OS 和应用，也可以恢复整个虚拟机。必须注意空间使用量，如果没有使用压缩或者重复数据消除等技术，空间使用量可能极大。
- 文件级恢复：可以恢复客户 OS 中的文件夹和文件而无须恢复整个 VM 映像。启动文件级恢复时，代理服务器显示 VM 卷为安装点，可以仅恢复需要的文件。代理服务器上不需要本地卷，因为它是一个安装点。

注意： 文件级恢复只能在某些备份软件（不是全部）应用上进行。创建备份时，有些软件创建以后恢复所需的索引，因为它们能够安装备份。具体的特性可以咨询制造商。

有些制造商结合 VADP 和其他 API（如 VMware 提供的 VLX API）提供相同的能力。

2. 用代理备份

这种传统方法主要用于数据库类应用的 VM。这些 VM 是例外，需要在虚拟机中保留代理，因为在备份期间要进行日志和数据库的管理。代理还能进行恢复粒度的控制和记录，而全 VM 模式必须恢复整个虚拟机才能找到丢失的数据。例如，使用代理对 Exchange 之类的消息服务器很有意义，在这种服务器中恢复邮件比恢复整个数据库更优先。使用代理还有一

个好处，不需要修改现有的过程。具有精心开发的基于代理备份解决方案的机构在一开始往往在 VM 中继续使用原来的方案。随着时间的推移，它们可能最终迁移到更优化的基于 VADP 的解决方案。

6.3 快照

本节提供快照的概述以及如何在虚拟环境中使用它们。

6.3.1 阵列快照与 vSphere 快照的对比

首先（也是最重要的），必须区分存储阵列提供的快照（例如 EMC 的 SnapView 和 NetApp 的 SnapVault）和 vSphere 提供的快照。它们的粒度不同：存储阵列的快照在整个 LUN（包含多个 VM 的 VMFS 卷）上完成，而 vSphere 快照在 VM 级别上完成。

存储阵列快照使用指针确定从上次快照以来那些数据块被修改，以及何时写入和提交修改；VM 快照冻结 vmdk 文件并创建另一个文件来写入修改。

注意： 快照从严格意义上讲不是备份，因为源数据从物理上没有被复制到另一个位置。如果数据毁坏，仅使用快照不可能恢复它。

6.3.2 VM 快照的优点

快照是虚拟化的主要优点之一，对于管理员非常有用。快照在 VM 处于健康状态和重大修改进行之前获取。它能建立一个简单回溯解决方案，以防应用程序迁移、新服务包或者补丁安装等操作之后出现故障。

快照也可以由 vSphere 5（GUI 或者脚本）在备份软件应用通过 VADP 使用它们的时候启动。

1. 快照工作方式

快照捕捉某个给定时点的关键 VM 状态信息，如：

- ❑ 磁盘状态
- ❑ VM 设置
- ❑ 内存状态（可以撤销）

快照触发时，VM 上的所有活动都被挂起。（理想状况下，快照在 VM 的活动较少时进行。）重要的是在截取快照之前确保应用状态一致。（参见 6.4 节。）

如图 6-2 所示，在快照创建的时刻，web.vmdk 文件被冻结，处于只读状态。原始文件中不会再写入数据。所有修改将写入新创建的文件 web0001-delta.vmdk。这个文件只反映快照创建之后的差别。重要的是在截取快照之前确保应用程序状态一致。

2. 快照注意事项

在生产环境中，快照必须在规定时间内使用，以便观察不正常工作的部分，实施系统回溯。建议的快照期间不超过一两天。

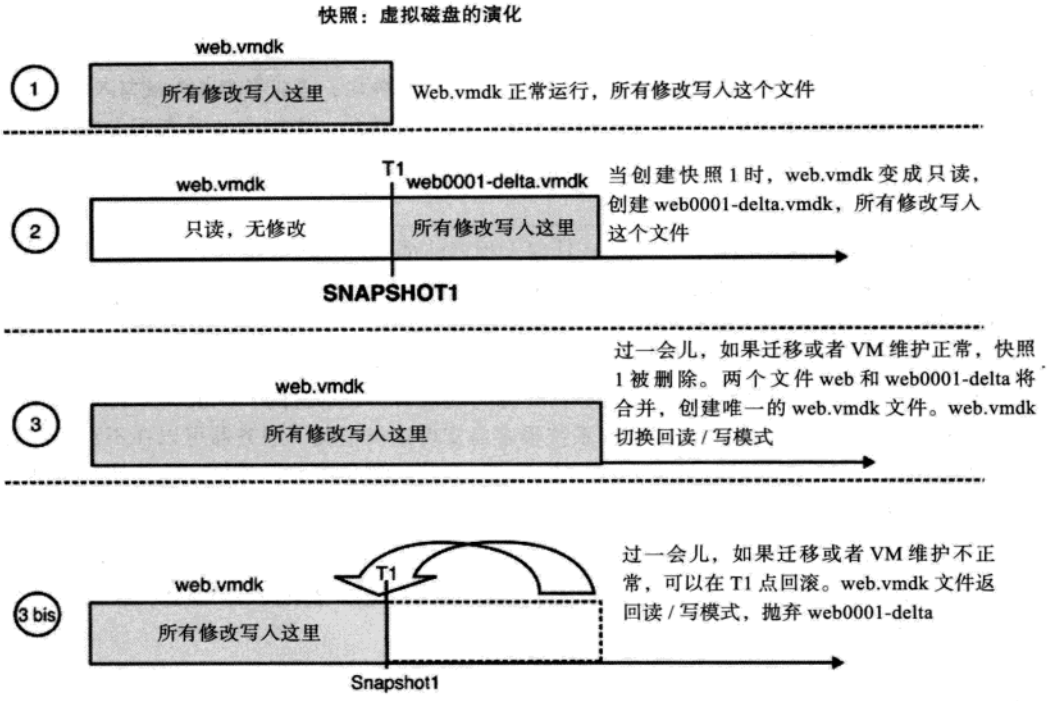


图 6-2 VM 快照过程

不建议长期保存 vSphere 快照，修改的数据越多，快照就会变得越大。如果不作限制，快照可能增长到占据所有数据存储空间，使所有 VM 处于危险之中（例如崩溃或者损坏）。

目前的经验法则是在 24 小时内 VM 总量的快照平均修改率为 20%。

提醒： *vSphere 快照不能在物理兼容性模式（RDMp）磁盘中使用原始设备映射。*

不正确的快照处理可能导致 VM 不可用。因此，必须考虑如下要点：

- 快照恢复是破坏性的，所有修改和 xxxdelta.vmdk 文件都会被永久性破坏。
- 不要手工删除 VM 目录中的 xxxdelta.vmdk 文件。使用 vCenter 中的快照管理器（Snapshot Manager）进行这些操作。

注意： *理解快照的机制和使用是很重要的，强烈建议设计一个规程并坚持执行。如果有疑问，在任何处理之前建立备份。*

6.4 应用一致性

运行高 I/O 操作率的应用程序（例如数据库和消息应用）时，确保应用一致性是备份的

主要难题之一。在共享存储环境中，为了获得高性能和较低的延时，大部分 I/O 活动在阵列控制器的缓存中进行。

注意： 存储阵列缓存组件对于获得高性能是必不可少的。确实，当服务器 I/O 被写入阵列时，I/O 进入缓存之后就发送确认，然后再将 I/O 写入磁盘。这能改变磁盘写入顺序，减少读磁头的移动。这种磁盘写入滞后在缓存数据没有被写入（刷新）磁盘时会造成丧失应用完整性的风险。

为了遵循最佳实践，应用的一致性在任何干预之前都必须（尽可能）得到保证。

备份、复制或者快照等操作进行时，必须注意确保一致的应用状态。突然性的硬件崩溃也可能给应用带来风险。

在 vSphere 中，快照（或者 VM 映像级的备份）是崩溃一致（crash-consistent）的，意味着它和突然停止生产 VM 然后重启服务器（称为硬重启）是等价的。Windows Server 2003 或者 2008）或者 Red Hat Linux 等操作系统很容易支持这类情况，服务器可以在不重新安装任何软件的情况下重新启动。然而，对于应用而言，建议避免崩溃一致性，而使用技术确保文件系统一致或者应用一致状态。

注意： 数据不一致并不一定意味着备份不可用，但是意味着在恢复期间，需要使用日志文件（如果存在）将应用返回到一个连贯的状态（从而增加了 RTO）。注意，在某些情况下，损坏也可能是恢复无法使用。

提示： 为了限制突然硬件崩溃，从而减少应用不一致的风险，强烈建议使用后备电池（在存储阵列中或者保护 RAID 卡）或者不间断电源，在数据刷新（写入）到磁盘上时为设备保持供电。

下面的小节讨论用于确保应用一致性的方法。

6.4.1 卷影拷贝服务

Microsoft 提供卷影拷贝服务。这个 Windows 服务的目标是创建某个时刻的一致（连贯）映像（一致性的时点拷贝，也称为影子拷贝，shadow copy）。使用这个服务，应用程序可以进入应用一致性状态，确保应用在恢复时的完整性。

为了利用这一功能，被备份的应用必须支持 VSS（例如，MS SQL Server、MS Exchange 或者活动目录）；客户 OS 也必须支持 VSS。这个服务通过 VMware Toole 完全适用于 vSphere（例如，在使用快照或者通过 VMware API 制作备份的时候）。VSS 引入了确保应用一致性的机制。当应用向 VSS 接收器发送请求时，接收器停止 VM。所有 I/O 活动被保留，同步所有当前数据。缓存被刷新到磁盘上。

在 vSphere 5 中，Windows 2003 和 Windows 2008 服务器用 VSS 组件确保应用一致性。

以前的版本如 Windows2000 没有 VSS，它们不能确保应用一致性。对于这些版本，必须使用预先冻结和事后解冻脚本（在下一小节中讨论）。

注意：快照生成一个崩溃一致状态。有了 VSS，VM 根据应用和客户 OS，处于应用一致性状态。

VSS 是如何工作的？ VSS 通过应用与备份软件、存储的协调来生成影子拷贝。与 VSS 的交互由客户 OS 上运行的 VMware Toole 提供。VMware 提供一个 VSS 接收器和一个 VSS 快照提供程序（VSS Snapshot Provider，VSP）。当备份过程启动，VMware Tools 启动接收器，VSP 被当作 Windows 服务记录。它在应用停止，可以获取 VM 快照时通知 ESXi。

一致性的不同级别如下：

- 崩溃一致性（等同于硬重启），理论上提供了如下级别上的一致性：
  - 客户 OS 级别上（对于现代 OS 通常没有影响）
  - 在文件级别上（有些文件的修改可能没有被写入磁盘 [建立快照时在缓存中]）
  - 在应用程序级别上（可能改变应用的完整性，通常是一个数据库）
- 文件系统一致性，理论上提供了如下级别上的一致性：
  - OS 级别上（没有影响）
  - 在文件级别上（没有影响；不同缓存中的所有数据被集中并且备份到磁盘上；文件级没有修改）
  - 在应用程序级别上（可能改变应用的完整性，通常是一个数据库）
- 应用一致性，理论上提供了如下级别上的一致性：
  - OS 级别上（没有影响）
  - 在文件级别上（没有影响；不同缓存中的所有数据被集中并且备份到磁盘上；文件级没有修改）
  - 在应用级别上（没有影响；所有数据都被备份，应用被正常关闭，确保其一致性）

一般来说，最好是将崩溃一致性解决方案看做完全不提供一致性。不管业务对快照有何种依赖性，重要的是经常而全面地测试从快照中恢复的能力。

6.4.2 预先冻结和事后解冻脚本

某些环境不能使用 VSS，需要脚本来停止（使用预先冻结脚本）和重启服务（使用事务解冻脚本）。虚拟化管理器通过主机代理接收这些信息，将其传送给 VMware Tools，触发预先冻结脚本。

Windows 机器所用的预先冻结脚本必须放在如下文件夹之一：

C:\Program Files\VMware\VMware Tools\backupScripts.d

C:\Windows\ backupScripts.d

对于 Linux 机器，它必须在如下目录中：

```
/usr/sbin/pre-freeze-script
```

预先解冻脚本能在创建 VM 快照之前关闭某些服务，或者触发管理员想在 VM 中进行的维护。

执行这个脚本之后，请求磁盘静止，这能将缓存的磁盘 I/O 刷新到 VM 磁盘上，达到文件一致。当同步驱动器完成了到 VM 磁盘的 I/O 刷新之后，启动 VM 的快照。被备份的磁盘

被冻结，其上的数据保持一致，VM 在备份期间发生的修改将被暂时地存储在缓存快照文件中。

建立快照之后，触发一个事后解冻脚本（如果有的话）。例如，该脚本重启预先冻结脚本停止的服务。此后，备份可以与生产一致的方式启动。

在 Windows 中，事后解冻脚本在如下目录：

C:\Program Files\VMware\VMware Tools\backupScripts.d

C:\Windows\ backupScripts.d

在 Linux 中，脚本在如下目录：

```
/usr/sbin/post-thaw-script
```

6.5 虚拟环境故障检修

本节研究一些可以用于虚拟环境中备份问题故障检修的工具和方法。

6.5.1 变更数据块跟踪

以前，进行增量或者差分备份时，ESXi 扫描整个 VM 卷，找出修改过的数据块进行备份。从 vSphere 4.1 起，ESXi 服务器配备了 VM 中所有数据块的映射（见图 6-3）。每个数据块带有一个时间戳，指出从前一次备份之后发生修改的数据块位置。识别最后一次备份以来修改的数据块不再需要扫描整个快照。这种技术大大加快了增量备份操作（最多加快 10 倍；见图 6-4 中的图表），减小了备份窗口。这种变更数据块跟踪（Changed Block Tracking，CBT）功能默认被禁用，因为它会在主机服务器级别消耗一些资源。

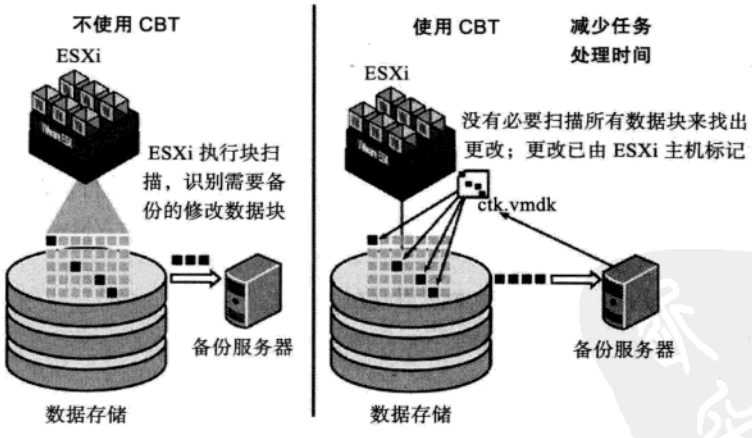


图 6-3 不使用 CBT（左）和使用 CBT（右）的备份过程

对于具有虚拟硬件第 7 版或者第 8 版的 VM，这个功能在启用 CBT 的每个虚拟磁盘的 VM 文件夹中的 -ctk.vmdk 文件（占据几十兆空间）中列出（上次备份以来）修改过的数据块。应用可以向 VMkernel 发送请求，要求发回上次备份以来修改过的虚

拟磁盘数据块。

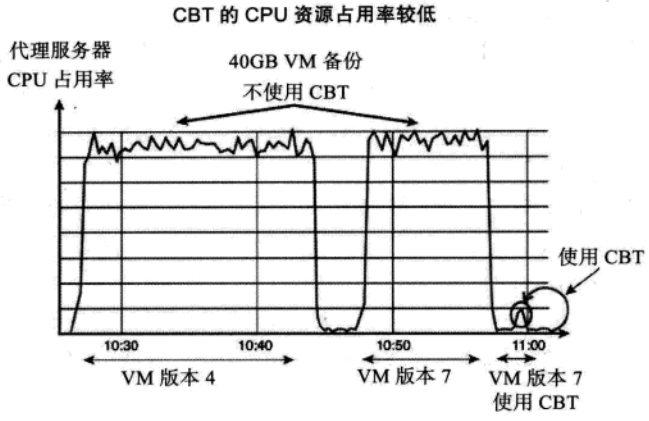


图 6-4 使用和不使用 CBT 的资源占用情况

CBT 可用于 vmdk 和 RDMv（不可用于 RDMp）。必须注意如下限制。

- CBT 可存在于所有 NFS、VMFS 和 iSCSI 数据存储的精简和厚盘配置模式。
- 如果虚拟磁盘连接到共享的虚拟 SCSI 总线或者具有快照的 VM 上，CBT 不能工作。
- CBT 的启用（ctkEnabled）通过 VM 的高级设置完成（只用于虚拟硬件的第 7 版和第 8 版）。
- CBT 启用在 VM 中的 stun-unstun 周期执行之后生效。stun-unstun 由 vSphere 在执行如下操作时触发：
  - 电源开启
  - 挂起后恢复
  - 迁移
  - 创建、删除或者恢复快照

注意： CBT 来源于 vSphere 4 Storage VMotion 技术（称作脏数据块跟踪或者 DBT），该技术使用这种方法快速确定那些变更数据块必须传送。

重复数据删除

为了减少需要的存储空间，必须使用压缩和重复数据删除等高级技术。

重复数据删除对于磁盘上找到的相同文件（或者数据块）只复制一次，大大减少了所需的存储空间。这种技术使备份更加可靠，使备份复制更少使用 WAN。

注意： 重复数据删除绝对是减少浪费的最有效技术。例如，发送一个有 3MB 附件的电子邮件给 100 个收件人可能使用 300MB 空间，使用重复数据删除只消耗 3MB 空间。在某些情况下，如果附件包含了重复信息，需要的空间甚至少于 3MB。

因为 VM 常常是从模板中创建的，磁盘上可能有大量相同的数据块。重复数据删除通过只存储一次冗余数据，能够显著减少需要的存储空间。

源端的重复数据删除在备份过程开始时，数据发送到备份解决方案之前进行。网络带宽得以保留，这对于某些环境是有价值的，但是必须小心，不要过多地消耗主机服务器的资源。市场上有许多产品提供这种技术：EMC Avamar、Symantec PureDisk、Atempo HyperStream 和 CommVault 等。

目标端进行的重复数据删除在备份解决方案级别上进行。这将源服务器从附加的工作负载中解放出来，但是没有减少网络带宽的需求。这种技术上的选择必须根据特定情况的约束做出。

- 在具备完整增量备份的传统环境中，存储容量增长的需求大约为每周 150% ~ 200%。
- 在虚拟环境中的效率引人注目，利用重复数据删除，存储容量增长的需求每周只有 2% ~ 7%。

注意： 市场上不同的解决方案使用的技术也各不相同，重要的是区分在备份工作期间（在备份期间，只发送不同的数据块）进行的重复数据删除和仅在 VM 中进行的重复数据删除。向制造商咨询他们的解决方案对这些情况的处理方式。

6.5.2 无 LAN 备份

在物理机器上，备份通常经过网卡。但是数据量不断增加，使备份窗口扩展。在大部分情况下，生产环境总是需要更多的可用性，所以备份窗口需要压缩。

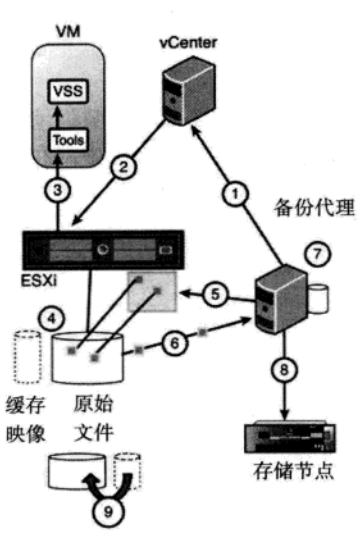
虚拟化可以通过 SAN 链接，而不需要通过网络进行备份（无 LAN 备份），这充分利用了存储和网络的性能，减少备份的耗时。

通过存储网络进行备份的另一个好处是，和物理机器中使用代理进行的备份不同，这对于机器的处理器使用没有影响。这种方法能将 CPU 的处理能力专用于应用，即使在备份期间也是如此。

网络数据管理协议（Network Data Management Protocol，NDMP）代理可将重复数据消除活动的负载转移到一台专用服务器上。

6.6 通过 VADP API 的备份过程

图 6-5 展示了 VMware 通过 VADP API 进行备份过程的各个步骤。



① 备份服务器向 vCenter 服务器发出请求

② vCenter 通知 VM 的虚拟化管理器所有者，这个 VM 开始一次备份

③ 虚拟化管理器将信息转发给 VMware Tools。VMware Tools 与 VSS 链接，使 VM 进入文件系统一致性状态。I/O 刷新 +I/O 冻结

④ 进行 VM 快照：web.vmdk 进入只读模式，备份期间发生的修改暂时存储在缓存映像 -delta-vmdk 中

⑤ VM 准备好备份时，通知备份服务器。备份服务器查询 CBT 映射，立刻找到上次备份以来修改过的数据块

⑥ 备份服务器通过 IP 或者 SAN 网络抓取修改过的数据块

⑦ 重复数据消除操作发生在备份服务器上

⑧ 完成重复数据消除的数据发送到存储节点

⑨ 备份完成后，缓存映像（-delta.vmdk）提交给原始 VM 文件

图 6-5 通过 VADP ADP 的备份过程

6.7 Data Recovery 2.0

VMware Data Recovery（VDR）是基于 Cent OS 5.5 64 位 Linux 的备份用具。这个工具作为插件集成到 vCenter Server。这个插件首先必须安装才能使用，然后必须导入用具。该工具提供了基本备份解决方案，小规模的机构可以用它加强虚拟基础架构安全，但是它不适合于大规模的生产环境。VDR 可以使用重复数据消除功能。它提供了完整性控制、再盘点（确保重复数据消除引擎调整了实际可用数据的目录）和回收（允许根据保留规则检索重复数据消除存储中不应再保留的数据）机制。

下面是 VDR 的一些约束。

- 每个 VDR 用具能够备份最多 100 个 VM，每个 vCenter 最多可以有 10 个 VDR 用具。
- 重复数据消除存储在 CIFS 网络共享上的限制为 500GB，在 vmdk 或者 RDM 格式存储上的限制为 1TB。
- NFS 只有在 ESX/ESXi 服务器引入且 VMDK 直接指定给用具的时候才能作为重复数据消除存储。

为了确保 Windows 中的应用一致性，VDR 使用 6.4.1 节中描述的 VSS 机制。

6.8 备份很重要，恢复更关键

备份 VM 有如此之多的选择，很容易让人忘记备份实际上没什么，只有恢复才最重要！

不管使用 VM 中的代理，还是更现代化、为 VM 优化的解决方案，重要的都是经常测试恢复。不同的解决方案有不同的一致性程度，每一个在恢复策略中都占有一席之地。尤其是快照，能够实现 OS 级别、文件级别或者应用级别的静止。

备份可能使用许多磁盘容量，特别是在用于存档的时候。有些技术（如重复数据消除和压缩）能够提供很大的帮助，VMware 已经以 VADP 的形式将它们加入产品套件中。在很多方面上，VM 的恢复可能比物理服务器的恢复更复杂，也更广泛。然而，如果有了很好的规划，机构能够更快、更完整地恢复基于 vSphere 的解决方案，问题也更少。

第7章
实施vSphere 5

本章将带你经历 vSphere 5 的实施。在你确定架构规模之后，有许多可选择的安装和配置方案。阅读本章之后，你应该能确定哪种配置最适合你。

7.1 确定规模

在实施虚拟基础架构之前，必须确定架构的规模。关于虚拟基础架构的规模常常有一个问题，规模需求的确定取决于许多因素：你有 50 个虚拟机（VM）还是 1000 个 VM？工作负载多大？以及访问类型（随机还是顺序）。然而，不管我们多么力求精确，通常还是需要一些假设。例如，你需要考虑未来（未知）的需求，重要的是，不要设计无法进行故障检修的方案。

为了帮助你了解架构规模的确定和所要创建的目标架构，表 7-1 总结了本书前面章节中提供的指南，可为实施计划打下基础。

表 7-1 实施中的架构规模选项

	最佳实践	最大值
群集		
数量	4 ~ 8 台主机	32 台主机
DRS	DRS 全自动。灵敏度：中等	
服务器		
CPU	双处理器 / 四处理器	160LCPU/ 主机
	常规负载下 2 ~ 4vCPU/ 核心。重负载下每个核心 1 个 vCPU	25vCPU/ 核心 2048vCPU/ 主机
内存	每个物理核心 4 ~ 8GB	2TB RAM/ 主机
网络	最少 6 个 GbE 网卡	32GbE 卡 8 × 10GbE 卡
存储		
LUN VMFS 5	600GB ~ 1TB	VMFS 5: 64TB
		RDMp 64TB
		RDMv 2TB*
	LUN/ 主机 :8	256 LUN/ 主机
		256NFS 安装 / 主机
vmdk	15 ~ 20 个并发的活动 vmdk/ 数据存储	2048 个 vmdk/ 主机
VM		
vCPU	1、2 或者 4vCPU/VM，取决于 VM 的应用	32vCPU
vRAM	1、2 或者 8GB vRAM/VM, 取决于 VM 的应用	1TB/VM
vmdk	OS:40GB 应用：100/200GB。不要超过 800GB。（如果需要更多磁盘，使用 RDM 模式）	2TB*

注：*2TB 减去 512 个字节。

7.2 不同的安装模式

VMware 提供了多种安装 ESXi 的方式。

- 交互式安装：对于少于 10 台服务器的小规模部署，这一安装使用了经典的手工方法。插入 DVD 启动，出现提示的时候回答问题。
- 脚本安装：为了自动化安装，创建一个包含 ESXi 配置的脚本，然后将其放在一个主机服务器能够访问的位置。
- 通过 vSphere Auto Deploy ESXi 安装：这种新方法可以将 ESXi 加载到内存（ESXi 仅占用 144MB），不需要将其物理安装到硬盘上。当启用自动部署时，服务器用主机配置文件（Host Profile）检索配置，执行预启动执行环境（PXE）启动。这种模式用于在非常大的生产环境中工业化快速制作服务器。
- 用 ESXi Image Builer CLI 自定义安装：使用这个工具安装时，可以创建带有最新更新或者补丁、特殊驱动程序的预配置 ESXi 映像（与公司策略相关的母板）。

注意： 有些制造商在某些服务器上提供将 ESXi 直接安装（嵌入式 ESXi）到特殊的存储卡（通常是 SD 或者 USB 卡）的选项。在这种情况下。服务器直接在扩展卡上启动，没有必要使用本地磁盘，减少了部署所需的时间。

7.3 安装前

下面概述推荐的部署 vSphere 5 安装前步骤。

7.3.1 检查列表

下面是安装虚拟基础架构所需的要素检查列表。

- VMware ESXi 5 的最新更新 DVD，或者用 Image Builder CLI 预先配置的 ESXi 5。
- vCenter Server CD（可选）。
- 25 个字符的许可证密钥。（可以在安装之后安装密钥）。从 ESXi 启动之后，该产品可以评估并试用 60 天。
- 在 VMware 兼容性矩阵中列出的一台服务器，至少连接一个网卡。
- 网络设置（IP 地址、子网掩码、主机名称、网关、DNS）。
- 一个具有 5GB 可用空间的数据存储以安装 ESXi（自动部署模式除外）。
- 要定义根用户密码。
- 操作系统安装 CD 或者 ISO 映像，用于安装客户 OS，具有相关许可证。
- 用于远程连接的 Windows PC

7.3.2 先决条件

本节介绍安装 vSphere 5 所需的硬件和软件。

1. ESXi 5 服务器

- 服务器：安装 ESXi 5 之前，必须检查 VMware 硬件兼容性列表（HCL），验证硬件及其内部组件的兼容性。可以在 www.vmware/resources/compatibility/search.php 上找到这个列表。vSphere 5 能够管理很宽泛的硬件组件选择，但是确认兼容性对于避免安

装时令人不快的意外情况是必需的。如果你对这些组件没有把握，咨询制造商。

- ❑ 处理器：VMware ESXi 5 只能安装在具有 64 位处理器、支持 LAHF 和 SAHF 指令的服务器上。
  - ❑ 对于 AMD 处理器，版本必须为 revE 或者更高。
  - ❑ Intel 处理器必须集成虚拟化技术（VT）指令，并在 BIOS 中激活。
- ❑ 内存：最少需要 2GB RAM，最大为每台服务器 2TB。
- ❑ 网络：至少连接一个网卡，每台 ESXi 服务器最多 32 个物理 GbE 网卡（8 个 10GbE 网卡）。
- ❑ 存储：一个受到支持的存储，有 5GB 可用空间用于 ESXi。
  - ❑ 每个服务器最多 256 个逻辑单元号（LUN）和 256 个 NFS 安装点。
  - ❑ 每个 VMFS5 LUN 为 64TB。
  - ❑ RDMv 为 2TB。
  - ❑ RDMp 为 64TB。
  - ❑ 块大小：VMFS5 中为 1MB。

可以在共享存储阵列上安装 VMware ESXi（通常称作从 SAN[存储区域网络] 启动）：

- ❑ 支持从 FC（光纤通道）阵列或者 iSCSI 硬件启动。
- ❑ 不支持从 NFS（网络文件系统）和 iSCSI 软件上启动。

提示： 在服务器本地磁盘上安装 ESXi 是首选方案，最好不要从 SAN 启动，从 SAN 启动的管理更复杂，可能成为处理错误的根源。虚拟化管理器现在很常见，可以很快重新安装。这种方法不会简化灾难恢复计划（DRP）（通常是从 SAN 启动的目标）。本章后面还要讨论使用自动部署的另一种方法。

---

2. vCenter Server

- ❑ 处理器：两个 64 位 CPU 或者一个频率高于 2GHz 的双核 64 位 CPU。
- ❑ 内存：4GB RAM。
- ❑ 磁盘空间：最少 4GB。
- ❑ 网络：1 个网卡（建议使用千兆）。
- ❑ 软件需求：
  - ❑ Microsoft .NET 3.5 SP1 和 Windows Installer 4.5。vCenter Server 需要一个数据库：支持 IBM DB2 或者 Oracle、Microsoft SQL Server。
  - ❑ vCenter Server 能管理最多 1000 台主机和 10 000 个运行的 VM。
  - ❑ 如果没有现成的数据库，可以安装 SQL Server 2008 Express 数据库，但是这将架构限制在最多 5 个 ESXi 和 50 个 VM。支持数据库版本的完整列表请参考 vSphere 兼容性矩阵。
  - ❑ Update Manager 支持 SQL 和 Oracle。
  - ❑ 操作系统：vCenter Server 只能安装在 64 位 Microsoft 产品上：Windows 2003 或者 Windows 2008。（支持的操作系统列表参见 VMware 网站。）

注意： VMware 不支持在活动目录（Active Directory）域控制器上安装 vCenter Server。对于 vCenter Server 要使用静态 IP 地址和有效的域名系统（DNS）而非动态主机配置协议（DHCP）。如果选择 DHCP，验证 DNS 中是否记录了 vCenter Server 的正确值。

提示： VMware 建议在一个 VM 中安装 vCenter Server，以利用虚拟化服务水平。有一种例外情况必须使用物理服务器：在使用 vCenter Storage Appliance（VSA）的时候，因为 vCenter 确保了 VSA 的管理。

3. vCenter Server Appliance

- 磁盘空间：vCenter Server Appliance（VCSA）需要至少 7GB 磁盘，最多 80GB。
- 内存：
  - 少于 100 台主机和 1000 个 VM 的小规模部署：至少 8GB
  - 100 ~ 400 台主机和 1000 ~ 4000 个 VM 的中等规模部署：12GB
  - 多于 400 台主机和 4000 个 VM 的大规模部署：16GB

vCSA 可以部署在 ESX 4.x 或者更高版本的主机上。

4. vSphere Client

vSphere Client 需要 .NET 3.0 SP1 Framework 和 Microsoft Visual J# 2.0 SE。如果在系统中没有找到这些组件，它们将随 vSphere Client 一起安装。vSphere Client 只能安装在 Microsoft 环境中。

表 7-2 列出了建议的 vSphere Client 和 vCenter Server 规格。

表 7-2 vSphere Client 和 vCenter Server 建议规格

	处理器（或者核心）	内存	磁盘
最多 50 台主机和 500 个 VM			
vCenter Server	2	4GB	5GB
vSphere Client	1	200MB	1.5GB
最多 300 台主机和 3000 个 VM			
vCenter Server	4	8GB	10GB
vSphere Client	1	500MB	1.5GB
最多 1000 台主机和 10000 个 VM			
vCenter Server	8	16GB	10GB
vSphere Client	1	500MB	1.5GB

5. 硬件选项

- 服务器：在我们的经验中，机架和刀片服务器最常用于虚拟化项目，因为它们能够显著地减少占地空间。在某些情况下也可以使用塔式服务器，因为它们能提供更多的 PCI 插槽以添加网卡或者光纤通道（FC）卡。每个制造商都提供物有所值的解决方案。我们建议选择著名的品牌，如 HP、Dell、IBM、Cisco、富士通西门子、NEC 或者 Bull。在所有情况下，硬件都必须明确地出现在 VMware 硬件兼容性矩阵中。

- 存储：为了vSphere 5的最优化使用，必须采用支持VMware vStorage API for Array Integration（VAAI）和网络连接存储（NAS）VAAI应用编程接口（API）的存储阵列。如果你计划使用站点恢复管理器（SRM），最理想的是具有存储复制适配器（SRA）模块的阵列。选择具有冗余机制、ALUA模式和处理增长能力的阵列。最著名的制造商有EMC、日立数据系统、NetApp、HP、Dell、IBM和Oracle Pillar Data System。确认VMware HCL兼容性。

提示： 虚拟化延长了架构的生命周期。（生命周期通常为5年。）必须规划解决方案的发展。例如，选择服务器类型时，服务器可以用简单的方式扩展，这一点非常重要。它们需要有空闲的扩展插槽，以添加额外的部件，如网卡、HBA卡和内存。存储也适用同样的原则。一般来说，容量每年必须有20% ~ 30%的提升空间。

7.4 准备服务器

技术的演变很快，所以这里只提供准备服务器的一般建议。

1. 对于BIOS

- 启用所有专用于虚拟化的硬件辅助功能，例如Intel VTx或者AMDv，以及用于内存管理单元（Memory Management Unit（MMU））和客户操作系统之间的Intel EPT和AMD RVI选项。
- 禁用某些服务器上的自动服务器重启（Automatic Server Restart，ASR）功能。
- 禁用所有电源管理设置。
- 禁用无用设备，如串行口、并行口和USB接口。
- 启用超线程。

2. 为了创建RAID群集

- 启用写缓存并使用缓存后备电池或者闪存缓冲。
- 将服务器各个部件的固件升级为最新版本（例如，BIOS、ILO类型管理和远程访问卡、RAID卡、网卡、FC HBA卡、硬盘、FC交换机、存储和网络部件）。
- 内存是最关键的部件之一，在虚拟化环境中一直需要它。VMweare建议测试内存72个小时。Memtest（www.memtest.org）等工具提供完整和可重复的测试。

咨询制造商，了解是否有其他服务器元素需要配置。

7.5 安装

本节描述各种vSphere 5安装选项和规范。

7.5.1 ESXi 5服务器

vSphere 5有两种主要安装模式：交互式安装和自动部署。

1. 交互式安装

安装ESXi服务器很容易，也很快（10 ~ 15分钟）。因为ESXi没有服务控制面板，安

装采用文本模式。

需要如下信息：

- ❑ VLAN ID
- ❑ IP 地址（可以在 DHCP 下安装）
- ❑ 子网掩码
- ❑ 网关
- ❑ 主要 DNS 和辅助 DNS
- ❑ ESXi 主机名
- ❑ 安装位置（至少需要 5GB 磁盘空间）
- ❑ 迁移还是新安装（迁移保留 VMFS 版本）
- ❑ 根密码（6 ~ 64 个字符）

图 7-1 中所示的屏幕在你启动服务器并插入 DVD-ROM 时显示。

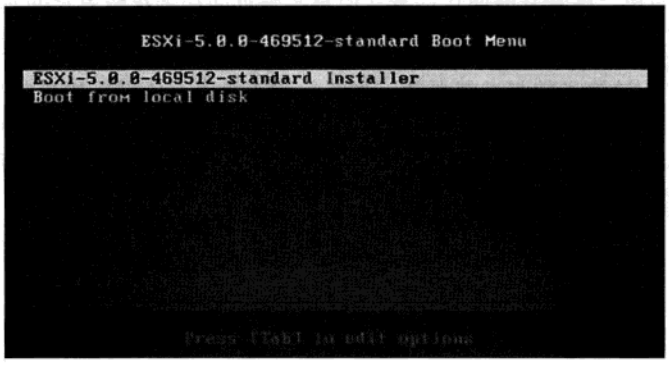


图 7-1 启动菜单

根据指令，输入需要的信息。安装 ESXi 之后，你将看到图 7-2 所示的界面。

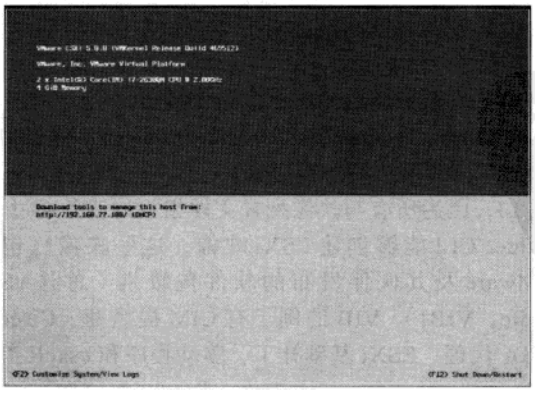


图 7-2 ESXi 安装完成屏幕

2. 用自动部署安装

自动部署是 vSphere 的新特性。它是专用于大规模生产环境的解决方案，能够从被称作映像配置文件（image profile）的映像中部署几十台 ESXi 主机服务器。这个 ESXi 映像在服务器启动时直接加载到主机服务器的 RAM 中。

和要求将系统安装到磁盘的传统安装不同，自动部署不向服务器本地硬盘安装任何东西。

这种模式的主要优点是必须升级（例如补丁或者更新）时非常高效的管理，这是因为它只需要更新参考映像，服务器启动时，它们会读取新的映像。这种模式的操作任务比重新安装服务器要少得多。

注意： 自动部署模式与使用 vSphere 更新管理器更新相比有显著的优势，因为在物理服务器应用更新之后无法稳定响应的情况下，更容易进行回溯。

这个解决方案使服务器完全可以互换（服务器可以用一个或者另一个映像启动），测试新映像变得非常简单。尽管它能简化部署，但是实施自动部署花费更长的时间进行创建和配置，并且需要通过 PowerCLI 使用命令提示符。

注意： 不要混淆从 SAN 启动和自动部署。尽管两种方法都不使用本地磁盘启动 ESXi，但是两者不同。在前一种情况下，每台服务器安装一个专用的 OS（ESXi 虚拟化管理器），写入在磁盘上完成。后者则是 OS 映像通过网络动态加载到服务器内存中，严格上说，这不是一个安装，而是一次映像广播（OS 流）。

从 SAN 启动往往用于使服务器可互换以及简化 DRP。尽管这种方法已经证明工作得很好，但是需要严格管理 SAN 和服务器级别上的实施（例如，分区、LUN 或者屏蔽）以避免处理错误。在这种框架内，自动部署能够替代 SAN 重启，最大限度地降低风险，并且更容易更新服务器的 ESXi 映像。

1）自动部署如何工作

自动部署服务器存储和管理映像。通过 Auto Deploy PowerCLI，管理员创建主机和映像配置文件的关联规则。还可以创建规则关联一个主机配置文件来配置服务器。

使用自动部署的主机服务器通过网络进行 ESXi 映像的远程启动（称作 PXE 启动），为此，需要如下条件：

- ❑ 一台 DHCP 服务器，在服务器启动时分配 IP 地址。
- ❑ 一台简单文件传输协议（Trivial File Transfer Protocol，TFTP）服务器，这是 FTP 的轻型版本，主要用于通过网络的远程启动。
- ❑ VMware PowerCLI，因为所有与映像配置文件相关的任务通过 PowerCLI 进行。
- ❑ ESXi Image Builder CLI 能够创建 ESXi 映像，这些映像打包了某些特定补丁、驱动程序或者 VMware 及其伙伴发布的软件包级别（称作 vSphere 安装包 [vSphere Installation Bundle，VIB]）。VIB 的例子有 CIM 提供商、Cisco Nexus、vSheld 插件、Lab Manager、HA 代理、ESXi 基础补丁、驱动程序和 esxcli 扩展等。

注意： 如果映像还未指定给某台主机服务器，可以为映像配置文件添加 VIB。添加 VIB 时，必须控制签名级别，以知晓在出现不稳定的时候谁确保支持。这个级别称为接受度级别，限制可多可少：VMware（最严格）、合作伙伴或者社区（较不严格，不建议用于生产环境）。

2）主机配置文件

前面已经讨论过，自动部署的作用是通过将更新集成到映像中，减轻 ESXi 部署的困难。它不能在代码已经安装到内存之后配置 ESXi。主机配置文件特性补充了自动部署，可以通过预先定义的安装和配置参数进行定制（见图 7-3）。这能确保基础架构同质，且所有指定服务器遵从规定的策略。

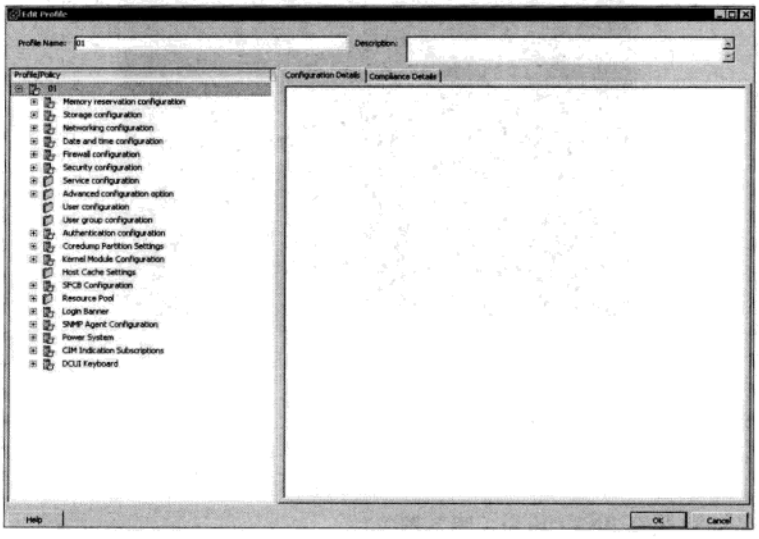


图 7-3 主机配置文件的编辑对话框

主机配置文件提供 ESXi 主机配置和设置的同质化并且允许如下配置：

- 存储：使用 VMFS 数据存储、NFS 卷、iSCSI、多路径等
- 网络：包括使用 vSwitch、端口组、物理网卡和相关的安全策略和 NIC 组合
- 许可证密钥
- DNS 和路由：DNS 服务器、默认网关
- 防火墙：网络和服务端口（sshd、snmpd、CIM 等）
- 处理器：是否使用超线程

为了创建一个主机配置文件，ESXi 必须用允许定义策略的参数来配置，这个策略作为创建所谓“金映像”的基础。定义主机配置文件之后，可以将其应用于独立主机服务器或者一个群集。

注意： 主机配置文件只在 ESXi 处于维护模式时才能使用。

如果启用相容性检查（Check Compliance），vCenter 检查服务器与已定义策略的相容性，并通知管理员有关状态。可能的状态有相容（Compliant）、未知（Unknown）和不相容（Non-compliant）。补救（Remediation）功能将服务器带回到相容状态。可以导出 VPF 格式的主机配置文件，用于其他地方。

7.5.2 vCenter Server 5 安装

图 7-4 所示的屏幕在插入 vCenter DVD 之后出现。

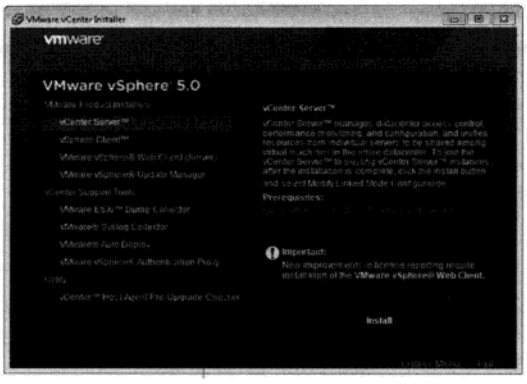


图 7-4 vCenter Server 初始安装屏幕

vCenter Server 安装 DVD 上有如下组件：

- ❑ vCenter Server
- ❑ vSphere Client
- ❑ VMware vSphere Web Client（服务器）
- ❑ VMware vSphere Upadate Manager（更新管理器）
- ❑ VMware ESXi Dump Collector
- ❑ Syslog collector
- ❑ VMware Auto Deploy（自动部署）
- ❑ VMware vSphere Authentication Proxy
- ❑ vCenter Host Agent Pre-Upgrade Checker

vCenter Server 5 需要一个数据库用于存储和组织数据。你可以依赖现有的数据库。VMware 建议为 vCenter Server 和 vCenter 更新管理器使用单独的数据库。vSphere 5 支持 Oracle、Microsoft SQL Server 和 IBM DB2。

与 vCenter Server 兼容的数据库完整列表参见 VMware 网站：http://partnerweb.vmware.com/comp_guide2/sim/interop_matrix.php。

Microsoft SQL Server 2008 Express 版本可以用于少于 5 台主机和 50 个 VM 的情况。超

过这一水平时，最好使用 Microsoft SQL Server 的其他生产版本、Oracle 或者 DB2。

7.5.3 升级到 vSphere 5

升级到 vSphere5 分几步完成，更新的顺序如下：

1）vCenter Server

2）vSphere 更新管理器

3）ESXi 服务器

4）VM——VMware Tools 和虚拟硬件

1. 更新到 vCenter Server 5

vCenter Server 必须为版本 5，以便管理 ESXi 5。升级 vCenter 是更新整个基础架构的第一步。可以作如下迁移：

- ❑ 从 virtual Center 2.5 update 6 到 vCenter Server 5
- ❑ vCenter 4.x 到 vCenter Server 5

提醒： vCenter Server 5 只能安装在 64 位环境中（64 位的 OS 和服务器）。

在升级 vCenter 之前，记得备份数据库和 vCenter Server 的证书，还要参考 VMware 兼容性矩阵，确保现有数据库兼容。安装 DVD 上可以找到 Pre-Upgrade Check（升级前检查）工具，它能显示一个报告，也可能在安装前解决一些问题。

2. 更新 vSphere 更新管理器

vSphere 更新管理器（VUM）4.x 可以升级到 5.0 版本。

备份 VUM 数据库以保留前一版本的 VUM 配置是必要的。

注意： 在安装期间，VUM 卸载过时的现有补丁（例如，ESX3.x 的更正）。

升级更新管理器服务器之后，在 vCenter Server 中更新 Update Manager 插件。

3. 升级到 ESXi 5

可以用 VUM 直接从 ESXi 4 更新到 ESXi 5。ESX4（具有服务控制台）到 ESXi 5 的迁移可以直接完成，但是必须小心进行，因为许多配置文件无法迁移（例如代理和第三方脚本）。很有必要参考更新指南，根据制造商的约束来验证架构。（参见 vsphere-esxi-vcenter-server-50-upgrade-guide.pdf）。

从 ESXi 3.5 直接更新为 ESXi 5 是不可能的。在这种情况下，需要分两步完成迁移：从 ESX3.5 迁移到 ESX4，然后迁移到 ESXi 5。然而，在后两种情况下，根据经验，我们建议逐个重新安装 ESXi 节点，以便从健康的基础架构开始，这更可取，而且快得多。

将 ESXi3.5 之前的主机升级为 ESXi5.0 也是不可能的，只能重新安装 ESXi。

提示： vSphere 5 只包含 ESXi 版本。必须验证与现有解决方案的兼容性，因为现有方案在没有服务控制台的情况下可能无法工作（例如备份和监控工具）。执行升级之后，没有回滚的可能（ESX3.5 到 ESX 4.x 的迁移与此不同，提供了回滚的可能性）。

警告： VUM 不能执行已经从 3.5 升级到 4.x 的主机的升级。

升级 ESX/ESXi 主机最简单的方式是通过 vSphere 更新管理器 5.0，如图 7-5 所示。

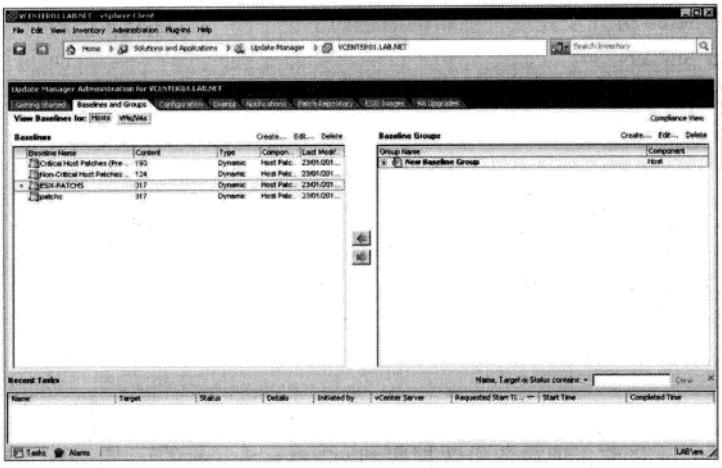


图 7-5 vSphere Client 更新管理器管理控制台

升级还可以使用其他方法。对于数量有限的主机（10 个左右），可以通过一个 ISO 映像进行交互式安装（手工）。对于大量 ESXi 主机，使用脚本更为合适。

4. 升级 VM

必须从升级 VM 中的 VMware Tools 开始。这可以使用 VUM 在每个 VM 上单独进行。注意，vSphere 5 的 VMware Tools 特性与版本 4.x 的主机兼容，所以如果在迁移阶段群集中某些主机还没有升级到版本 5，也不会有问题。

VM 的虚拟硬件也必须被升级为第 8 版本（在 VM 关闭的时候）以利用这个硬件版本的新特性。（参见 2.6.2 节）升级虚拟硬件是重要的考虑因素之一是运行虚拟硬件第 8 版的虚拟机只能运行于 ESXi 5.0 主机上。建议最好是在群集中所有主机都升级到 ESXi5.0 之后才升级虚拟机的虚拟硬件版本。

警告： 升级虚拟机的虚拟硬件版本是单向操作。升级完成之后没有逆向操作的选项。

7.6 不同的连接方法

如图 7-6 所示，连接到 vSphere 5 环境有多种方法，包括：

- 在 ESXi 服务器上通过直接控制台用户界面（Direct Console User Interface，DCUI）进行本地连接。
- 使用 vSphere Client 直接连接 vCenter Server 或者 ESXi 主机服务器
- 在 vCenter Server 上使用 Web Access 或者连接到 ESXi 服务器

- 在 vCenter Server 上使用 Windows 远程桌面
- 使用脚本工具（有经验的管理员可开发自己的脚本和任务自动化解决方案）

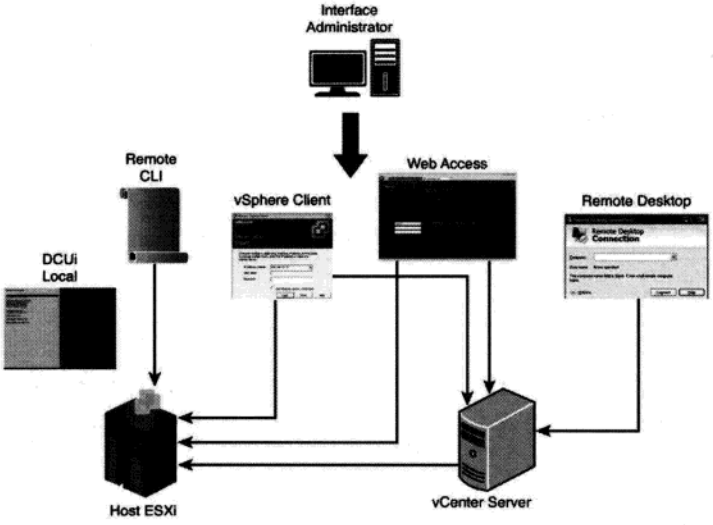


图 7-6 vSphere 5 环境连接选项

7.6.1 直接控制台用户界面

直接控制台用户界面（Direct Console User Interface，DCUI）允许 ESXi 服务器的本地连接，可以进行管理 IP 地址的配置（见图 7-7）。管理网络（只能从 DCUI 配置）在 ESXi 安

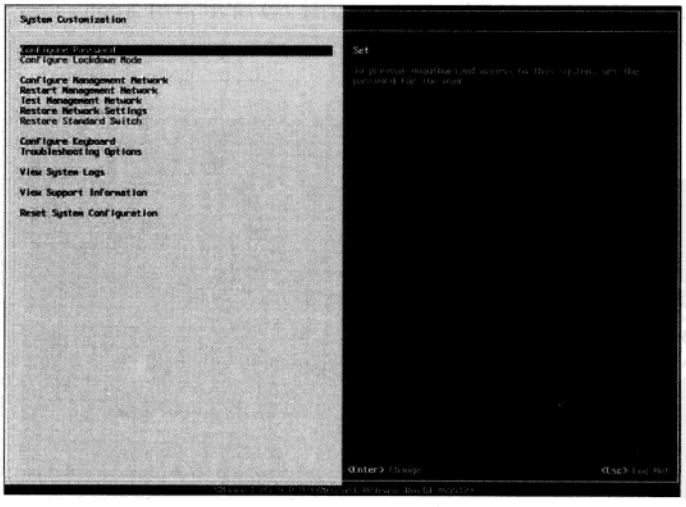


图 7-7 系统自定义屏幕

装时自动创建，使服务器可达且可以配置，与 vSphere 4 的服务控制台端口等价。DCUI 可以作如下配置：

- ❑ 主机名称
- ❑ 根密码
- ❑ 诊断
- ❑ 重启 hostd 和 vpxa 管理代理
- ❑ 复位，恢复原始配置

注意： ESXi 服务器集成到 vCenter Server 时，安装 vCenter vpxa 代理。ESXi 服务器的 hostd 管理代理负责从 ESXi 服务器接收数据以及与 vSphere Client 通信。

7.6.2 vSphere Client

vSphere Client 能直接连接到一台 ESXi 服务器或者 vCenter Server。连接 ESXi 服务器必须使用根账户（安装时配置）。用其他用户连接也是可能的。要连接到 vCenter Server，用户必须具有管理员的权限。

连接 ESX 服务器只需要输入服务器的 IP 地址（或者主机名）以及安装时定义的用户名及密码，如图 7-8 所示。vCenter Server 使用 Windows 用户账户库，要连接到它，可以使用安装服务器的本地账户或者 vCenter 加入的活动目录账户。

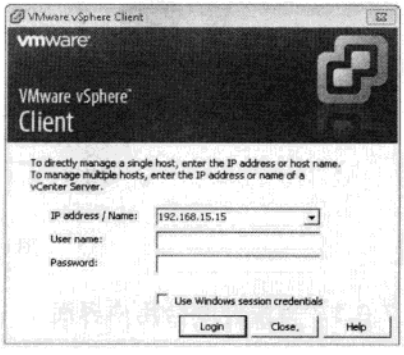


图 7-8 vSphere Client 登录屏幕

7.6.3 vSphere Web Client

vSphere Web Access 是一个 Web 界面，如图 7-9 所示，用 Adobe Flex 开发，在虚拟和 ESXi 机器上进行基本管理和配置工作。

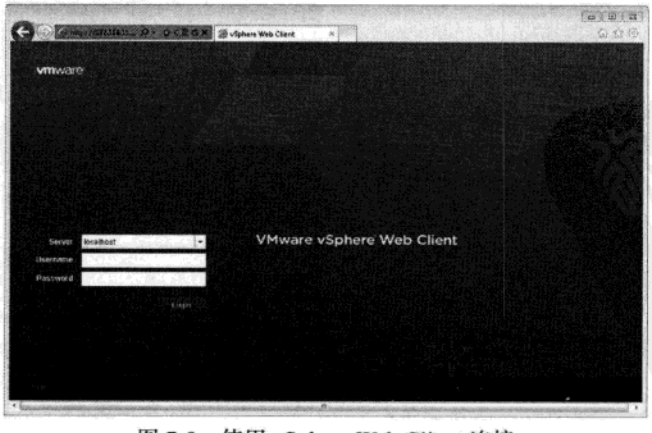


图 7-9 使用 vSphere Web Client 连接

注意： Web Access 不能在 ESXi 上直接访问，但是如果安装了 vCenter Web Administration 模块，可以通过 vCenter 访问。

7.6.4 脚本工具

为了自动化部署，VMware 开发了日常操作所用的工具，可以创建脚本，以自动化和 vSpher Client 具有相同功能的任务：

- vSphere Managerment Assistant（VMA）：由多个管理工具组成的一个虚拟用具。VMA 包含了 vSphere 命令行界面和 vSphere SDK Perl。
- vSphere PowerCLI：用于自动化主机、VM、客户操作系统等的命令行工具。PowerCLI 包含了超过 150 条命令（称为 cmdlets，是特殊的 Powershell 命令）。
- vSphere 命令行界面（vCLI）：用于部署、配置和维护 ESXi 主机的一个实用工具。vCLI 包含了许多命令，包括 VMkfs-tools、VMware-cmd 和 resxtop。

7.7 vCenter 配置

在 VMware 环境中，vCenter Server 是主要的管理工具。图 7-10 展示了连接到 vCenter Server 的 vSPhere Client。

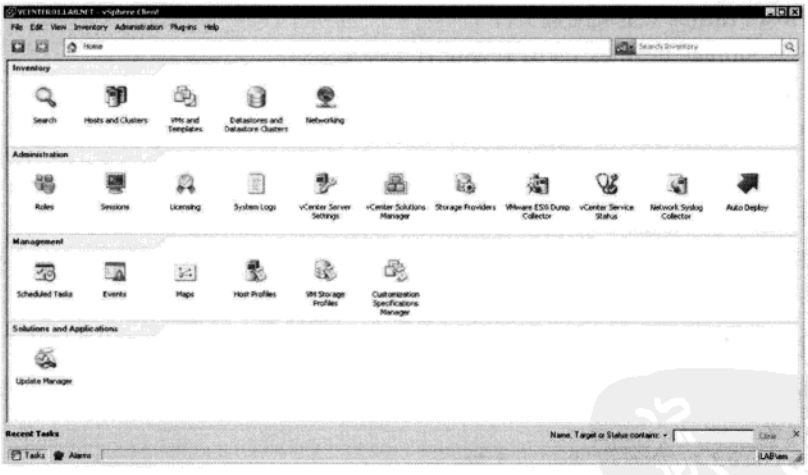


图 7-10 vCenter 常规视图

根据安装的许可证和插件，你可能发现如下对象。

- Inventory（库存）组
  - Search（搜索）：寻找 vCenter 中的对象
  - Host and Clusters（主机和群集）：管理主机服务器和群集
  - VMs and Templates（VM 与模板）：用于创建 VM 和模板

- ❑ Datastores and Datastore Clusters（数据存储和数据存储群集）：管理存储空间
- ❑ Networking（网络）：管理 vSwitch 和分布式 vSwitch 网络

❑ Administration（系统管理）组

- ❑ Roles（角色）
- ❑ Sessions（会话）
- ❑ Licensing（许可证）
- ❑ System Logs（系统日志）
- ❑ vCenter Solutions Manager（vCenter 解决方案管理器）
- ❑ Storage Provider（存储提供者）
- ❑ VMware ESXi Dump Collector
- ❑ vCenter Service Status（vCenter 服务状态）
- ❑ Network Syslog Collector
- ❑ Auto Deploy（自动部署）

❑ Management（管理）组

- ❑ Scheduled Tasks（调度任务）
- ❑ Events（事件）
- ❑ Maps（映射）
- ❑ Host Profiles（主机配置文件）
- ❑ VM Storage Profiles（VM 存储配置文件）
- ❑ Customization Specifications Manager（自定义规范管理器）

❑ Soulutions and Applications（解决方案和应用程序）组

- ❑ Update Manager（更新管理器）

7.7.1 许可证

每台 ESXi 服务器需要一个许可证密钥。vCenter Server 也需要一个许可证密钥。（见 2.2 节。）如果没有输入许可证密钥，就会开始一个 60 天的评估期。

7.7.2 常规设置

vCenter 的其他设置可以在主菜单的 vCenter Settings 中找到（见图 7-11），下面的列表将描述这些设置。

- ❑ Licensing（许可证）：参见 2.2 节
- ❑ Statistics（统计）：确定组件里记录的监控信息级别（级别越高，生成的信息越多）
- ❑ Runtime Settings（运行时设置）：用于多 vCenter Server 实例环境
- ❑ Active Directory（活动目录）：活动目录的同步设置
- ❑ Mail（邮件）：通过电子邮件发送警告
- ❑ SNMP：向管理控制台发送警告和简单网络管理协议（Simple Network Management Protocol，SNMP）陷阱

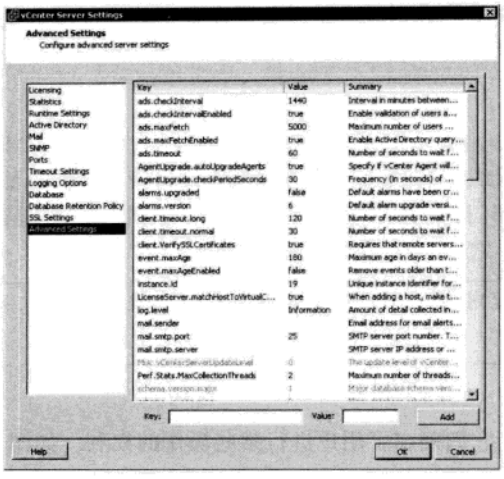


图 7-11 vCenter Server 设置，高级设置菜单

- ❑ Ports（端口）：修改 vCenter 默认使用端口
- ❑ Timeout Settings（超时设置）：超时参数
- ❑ Logging Options（日志选项）：日志文件的详细程度
- ❑ Database（数据库）：数据库最大连接数
- ❑ Database Retention Policy（数据库保留策略）：事件和任务保留设置
- ❑ SSL Settings（SSL 设置）：激活或者禁用 SSL 证书
- ❑ Advance Settings（高级设置）：所有 vCenter Server 的高级参数

7.7.3 主机与群集

你可以从 vSphere Client 管理控制台上轻松地进行主机和群集管理，如图 7-12 所示。

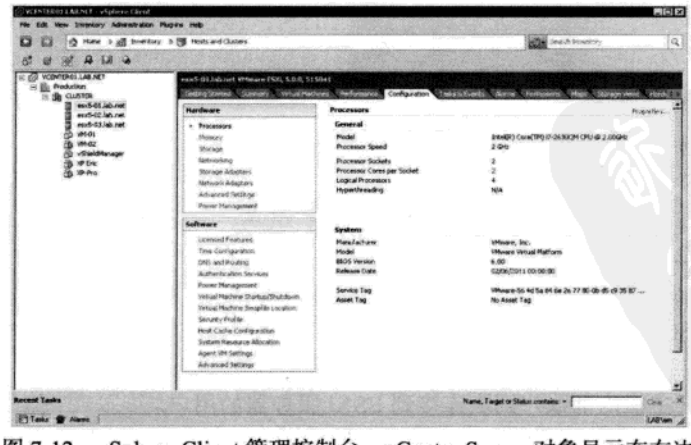


图 7-12 vSphere Client 管理控制台：vCenter Server 对象显示在右边

只要点击一个对象，就能看到对应于对象可设置参数的选项卡。每次对象修改的状态和进度都会显示在 Recent Tasks（最新任务）下方。

7.7.4 数据中心创建

必须先创建数据中心，才能添加 ESXi 主机服务器或者创建 VM。数据中心提供虚拟环境中所有对象的结构化组织：主机、群集、虚拟机和目录以及资源池。数据中心必须反映根据地理位置或者功能 / 部门组织的网络架构。数据中心的概念只有在连接到 vCenter Server 时才存在。创建数据中心之后，可以添加 ESXi 服务器。

注意： 你可以从数据中心内的一个群集向另一个群集实施 vMotion，但是不能迁移到另一个数据中心。

7.7.5 权限管理

权限非常重要，因为它们确定了用户进行某些操作的权利。如图 7-13 所示，权限的管理分为 3 个步骤：

1）创建用户；

2）创建角色；

3）将角色与用户关联以建立权限。

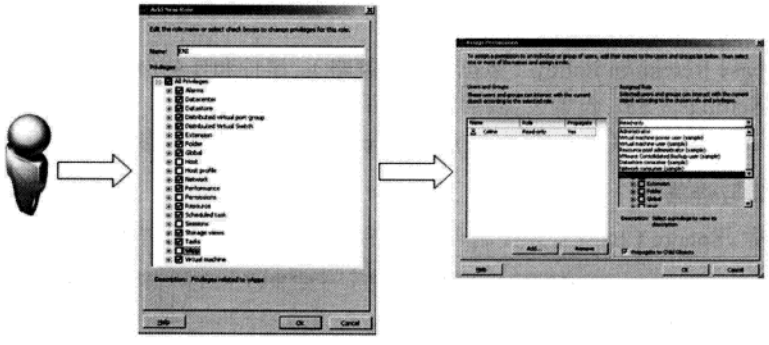


图 7-13 权限管理步骤

用户分为两类：

- 根据 Windows 账户连接到 vCenter Server 的用户。用户必须存在于活动目录（AD）或者 vCenter Server 本地。在 vCenter 中不可能创建或者删除用户，修改必须在 AD 级别进行。vCenter Server 连接到 ESXi 主机时使用 vpxuser 账户。
- 根据 ESXi 主机内部账户列表连接到 ESXi 服务器的用户。

角色（role）的定义是一组用于为环境中的对象指定特定任务或者行为的特权。对象

（object）是 vCenter Server 能够管理的任何东西，例如，数据中心、主机或者群集。在图 7-12 所示的主机和群集屏幕中，对象列在左边。

有一些预定义的标准角色，但是可以通过定义特权创建其他角色，如图 7-14 所示。

创建角色之后，可以将角色与用户关联以建立权限。

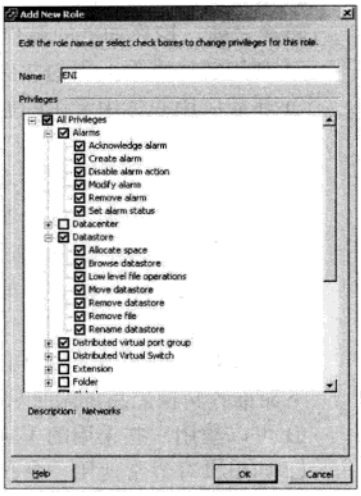


图 7-14 在 Add New Role（新建角色）对话框中分配特权

7.7.6 存储和网络

请参阅第 3 章和第 4 章，其中详细介绍了存储和网络配置。

7.7.7 P2V 转换

物理－虚拟（P2V）转换可以完成如下工作：

- 创建一个新 VM
- 用 VMware Converter 之类的工具将一台物理服务器或者 PC 转换为一个虚拟实例

注意： 其他类型的转换也有可能，如在不同 VMware 产品之间的转换（例如，ESXi 和 VMware Workstaion 或者 Fusion）和虚拟－物理（V2P）转换。

VMware 提供了一个单独的转化工具 VMware Converter。这个工具很适合于少量机器的转换。对于大规模迁移，最好采用市场上的工具，因为它们能够工业化迁移过程，提供更丰富的功能。

1. P2V 之后的变化

将物理服务器转换为虚拟服务器产生的变化必须加以考虑：

- MAC 地址变化。每个虚拟网卡（vNIC）都会分配一个范围在 00:50:56:00:00:00 到 00:50:56:3f:ff:ff 或者 00:0c:29 的新 MAC 地址。这可能造成一些后果（最明显的是许可证，它可能基于 MAC 地址）。
- 创建虚拟硬盘。迁移到虚拟环境时，物理磁盘被转换为虚拟磁盘（vmdk 文件）。在大容量磁盘（几百 GB）的时候应该考虑这一点，转换的时间将会较长。在 P2V 期间，虚拟磁盘的大小可以修改。
- 许可证模式可能改变。软件许可证是迁移时需要考虑的重要问题。有必要验证制造商在虚拟环境中对产品的支持以及许可证系统是否清晰定义。有些制造商根据 vCPU 数量颁发许可证，而其他制造商根据物理处理器（插槽）或者主机服务器的物理核心颁发许可证。
- 使用 Sysprep（为虚拟机创建新的安全标识符 [SID] 的一个 Microsoft 工具）时，缓存的凭据被删除。有些转换工具在迁移过程使用 Sysprep。
- 迁移之前，确保你可以使用本地管理员账户访问服务器。使用 Sysprep 时，所有本地存储的连接配置文件以及相关的凭据（缓存凭据）都被删除。只有原始的安装账户得

以保留，确切地说是本地管理员账户。没有这个账户就无法建立连接。而且，Sysprep复位机器的SID。这意味着在第一次连接时，该机器不再能连接到域，必须重新手工集成到域（而且必须首先从活动目录中删除前一个机器的账户）。

2. 最佳实践

转换之前，推荐如下的最佳实践：

- ❑ 利用转换删除无用的文件（例如临时文件或者几个月甚至几年未用的文件）。
- ❑ 卸载应用并禁用无用的服务。
- ❑ 整理硬盘碎片。
- ❑ 如果你不能访问机器的本地管理员账户，创建一个仅用于迁移的临时管理员账户。记得在迁移之后删除这个账户。
- ❑ 停止所有可能妨碍迁移过程的服务（例如防病毒、磁盘扫描和备份）。
- ❑ 执行热迁移的时候，最好停止应用，特别是数据库等事务性应用。
- ❑ 在服务器活动较少时进行迁移。（例如，可以在夜间进行迁移）
- ❑ 确保目标主机服务器有必要的资源（最重要的是磁盘空间和内存）以运行新的VM。

下面推荐转换之后的最佳实践：

- ❑ 可以禁用一些无用的VM设备（例如串行端口、USB端口、并行端口和磁盘）。
- ❑ 如果没有经常使用，断开VM中的CD-ROM驱动器。
- ❑ 卸载或者停止管理和监控服务以及与硬件相关且不再需要的代理。
- ❑ 禁用管理HBA卡或者网卡负载分配的软件（VMware负责这些管理）。
- ❑ 验证VMware Tools正常安装。
- ❑ 验证Windows设备管理器中正常安装了虚拟硬件（也就是说，没有表示问题的问号或者感叹号）。
- ❑ 删除在事件日志中产生问题的设备（例如，ILO卡）。

7.8 高效管理虚拟环境

本节提供信息，帮助你最大限度地提升虚拟环境管理效率，主要关注如下领域：

- ❑ 主机服务器监控
- ❑ 警告和映射
- ❑ 资源共享
- ❑ 资源池
- ❑ 整合率
- ❑ 服务器性能
- ❑ 复制和模板
- ❑ vApp

7.8.1 主机服务器监控

基础架构的监控需要组合管理员的主动行为和根据规定警示采取的对应措施。使用

vCenter Server 可以组合这两种方法。市场上的其他工具能够补充 vCenter，改进信息和对象表现的分析。

在虚拟环境中，性能特征是根本，因为 VM 共享 ESXi 主机服务器资源。因此，监控主机和 VM 的活动对于确保 VM 有运行应用的足够资源来说是必要的。经过改进的资源分析必须能够改进基础架构的整合，从而达到显著地节约，因为不需要进行新服务器的投资。

将不适合的服务器投入大负载中可能导致服务器总体性能的崩溃。相反，没有充分利用服务器的性能，VM 只使用一部分资源，会导致公司无法最大限度地利用这种解决方案的好处。

7.8.2 警告与结构图

警告（alarm）是根据规定的标准采取的相应措施。它们可以触发向管理员发送邮件等动作。警告应该谨慎使用，只突出重大的故障。

例如，如果 VM 的 CPU 使用率（%）达到 75% ~ 90% 之间，应该创建一个警告，如果超过 90%，则要发送一个警告。

你可以关联一个动作（非强制），如向管理控制台发送一封电子邮件或者 SNMP 陷阱。注意，如果没有关联动作，警告仍然可以在警告区域中查看，如图 7-15 所示。

图 7-15 在 vSphere Client 管理控制台上查看警告定义

7.8.3 资源共享

资源共享是共享环境中的基本元素。VMware 提供共享机制，确保 VM 在发生争用时有必要的资源。这些“安全保障”是必不可少的，能够在密集活动的时候带来心灵的平静。重要的 VM 有所需的资源。在 VMware 下，如图 7-16 所示，可以设置不同级别的服务质量。

如图 7-17 所示，结构图（map）提供虚拟环境的图形视图。可以用它们观察主机服务器和 VM、网络、数据存储的关系。

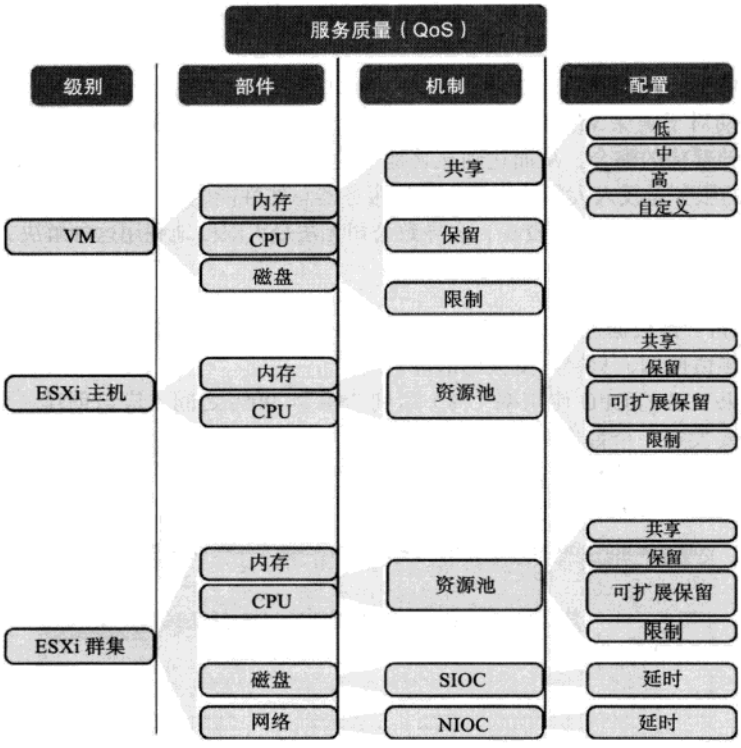


图 7-16　可以为内存、处理器、磁盘和网络设定不同级别的服务质量

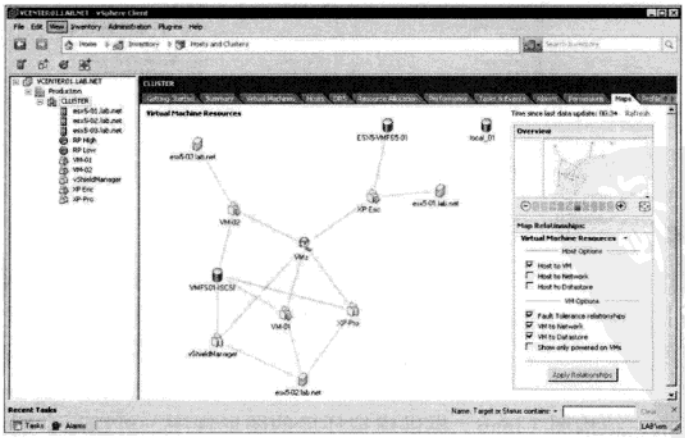


图 7-17　资源结构图示例

在VM中可以共享内存、处理器和磁盘。如图7-18所示，有三个可用设置：保留（Reservation）、共享（Share）和限制（Limit）。

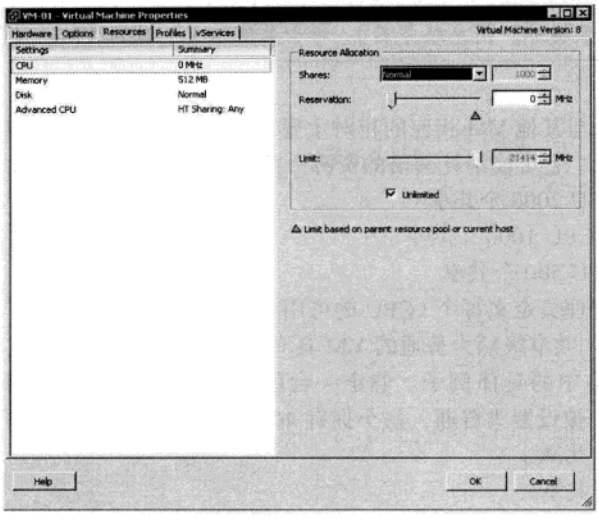


图7-18 VM资源分配设置

1. 保留

保留代表着保证虚拟机能得到的资源量。ESXi允许VM仅在指定资源量可用的情况下启动。这能确保VM在任何时候都能访问最低限度的资源量，即使同一台主机服务器上的VM正在进行密集的活动。保留分为如下类别。

- CPU保留：在正常的活动中，如果VM不能使用CPU保留资源并且不需要它们，CPU会被释放供其他VM使用。但是，如果ESXi主机服务器的活动密集，CPU时间根据配置的共享和保留动态地分配给VM。
- 内存保留：服务器物理RAM中不考虑活动而保留的内存量。指定了内存保留的VM不管其他VM的负载水平如何都能保证得到这些内存。这一机制说明，TPS、交换、气球和内存压缩都不能获得VM的保留内存。如果RAM中可用内存不足，VM不能启动。如图7-19所示，创建一个扩展名为vswp的交换文件，大小等于配置的内存减去保留内存。

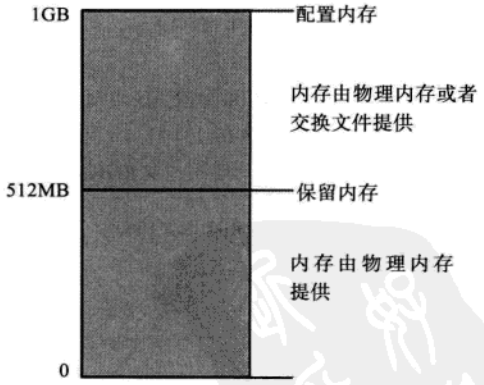


图7-19 保留内存由服务器物理RAM提供。在这个例子中，交换文件占据512MB

注意： 使用 vSphere HA 群集时，保留是插槽大小计算的重要因素（参见第 5 章）。

提示： 保留不应该过于频繁使用，它应该用于关键 VM（例如，活动目录、vCenter）或者关键应用（例如，数据库、消息服务器）。对于其他不太关键的 VM，使用共享更为合适。

2. 共享

共享（share）是与其他 VM 相比的相对重要性或者优先级概念。如果一个虚拟机拥有另一个 VM 两倍的资源，它可能消耗两倍的资源。共享有 4 个级别：低、普通、高和自定义：

- 高 = 每个 vCPU 2000 个共享
- 普通 = 每个 vCPU 1000 个共享
- 低 = 每个 vCPU 500 个共享
- 自定义 = 由管理员定义每个 vCPU 的可用共享数

具有两个 vCPU，共享级别为普通的 VM 具有 2 × 1000=2000 个共享。

我们来看图 7-20 中的具体例子。假定一台服务器上有两个 VM，该服务器有 8GHz 的 CPU 资源。两个 VM 被设置为普通，每个拥有 4GHz 资源。假设如下条件：

- VM1=1000 个共享
- VM2=1000 个共享
- VM1=4GHz
- VM2=4GHz

如果共享级别为高的第三个 VM（如图 7-21）启动，它有 2000 个共享：

- VM1=1000 个共享
- VM2=1000 个共享
- VM3=2000 个共享

资源的比例计算如下：

- VM1=1000/4000 × 8GHz=1/4 × 8GHZz=2GHz
- VM2=1000/4000=1/4=1/4 × 8GHz=2GHz
- VM3=2/4000=1/2=1/2 × 8GHz=4GHz

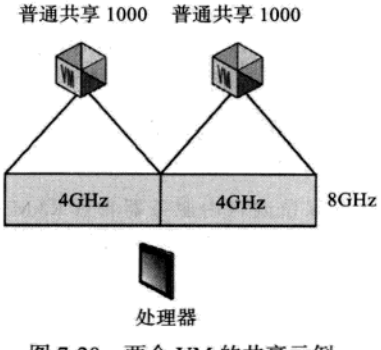


图 7-20 两个 VM 的共享示例

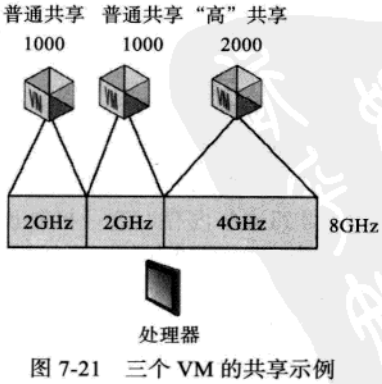


图 7-21 三个 VM 的共享示例

注意： 设置共享和创建资源池要好于保留。

3. 限制

你应该熟悉的两个限制因素是内存限制（memory limit）和 CPU 限制（CPU limit）。

内存限制是 VM 可以占用的最大 RAM 数量。限制了 RAM 的 VM 必须交换到磁盘上，获得“可用内存空间”，从而妨碍了性能。

在图 7-22 中，交换文件等于配置内存减去保留内存：1.5GB−0.5GB=1GB。如果 VM 需要超过 1GB 的内存，它不能从 RAM 中取得内存：额外的内存将由交换文件提供。512MB 内存（配置内存减去保留内存）将从磁盘交换文件中得到。

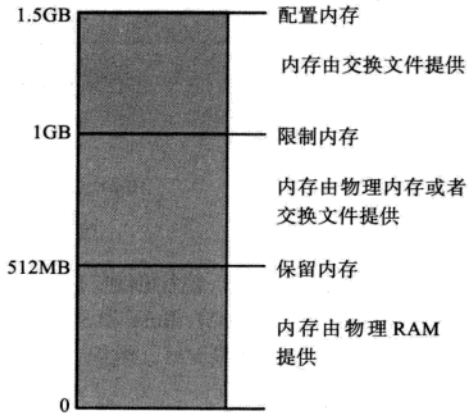


图 7-22 内存限制示例

注意： 应该避免过大的交换文件，所以 VM 不应该配置过量的内存。

提示： 即使 VM 不关键，使用限制也没有多少好处。最好使用共享而不使用限制。

CPU 限制代表 VM 可以使用的最大 CPU 时间。ESXi 不会分配多于指定限制的 CPU 时间（用 MHz 表示）。CPU 限制主要用于测试性能，以了解 VM 在最大规定值下的表现。对于生产 VM，使用限制是不明智的，应该使用共享。

7.8.4 资源池

利用资源池，CPU 和内存资源可以用图 7-23 所示的对话框进行层次化风格的分区。

数据中心的可用资源被组合为多个实体，能够精确地分配资源。可以用继承自上级资源的子资源池建立层次化资源池，这种可用资源管理使管理员不能在每个虚拟机上分配资源，而代之于整体的分配方式。

图 7-24 显示了父资源池和子资源池。

资源池的创建可以在 ESXi 主机或者群集的级别上进行。对于 CPU 和内存，可以设置精确的保留、限制和共享参数值。

当选中 Expandable Reservation（可扩展保留）选项时（见图 7-23），如果子资源池需要更高的保留值，资源可以从父资源池中取得。

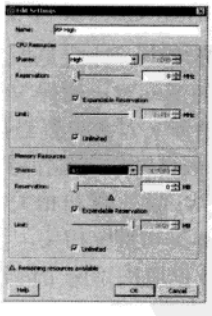


图 7-23 使用 Edit Settings（编辑设置）对话框设置资源池层次

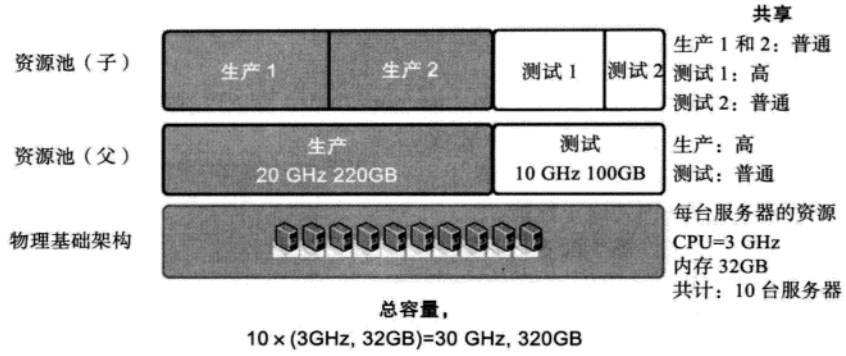


图 7-24　父资源池和子资源池

资源池非常有用，往往创建三个重要性级别——低、中、高。还为 SRM 定义了三个分组：SRM Low、SRM Medium 和 SRM High。

资源池不用来组织 VM，组织 VM 必须使用目录。

7.8.5　整合率

整合率（Consolidation Rate）很重要，因为它代表着可以在 ESXi 主机服务器上同时运行且符合必要性能和服务质量水平的 VM 数量。整合率直接与管理员对环境的掌握相关，团队对技术越精通，整合率就越高。

当然，建立理想的比率（每台服务器 VM 数）取决于应用负载。最佳的整合率使用管理和自动化工具进行观察，以获得高的整合水平。

注意： 制造商、发布者和零售商不一定鼓励各家公司优化他们的基础架构，公司必须自己进行这一工作。使用有效的工具能够得到可观的效果。2012 年，每台服务器 30 个 VM 似乎已经很多，但是几年之内在技术的帮助下，整合率可能达到每台服务器数百个 VM。

7.8.6　vCenter Server 中的性能

在虚拟环境中，性能是根本特征，因为 VM 共享 ESXi 主机服务器的资源。因此，监控主机和 VM 的活动对于确保 VM 有充足的资源来运行应用是很有必要的。这就是管理员必须为虚拟环境的所有对象设置最佳大小的原因。这只能通过管理员必须掌握的工具进行。

在 vSphere 5 中，性能得到改进，如图 7-25 所示，添加了图形指标来方便分析。这些指标提供了 ESXi 服务器和相关 VM 状态的简单视图。记录的指标与 CPU、内存、磁盘、网络、资源、群集和系统相关。

报告的数据取决于两个因素：收集级别和收集间隔。

1. 收集级别

有 4 个收集级别：

❑ 级别 1（默认）：手机 ESXi 的一般使用信息。记录的指标与 CPU、内存、磁盘和网络

以及系统信息相关：运行时间、Heartbeat 和 DRS 指标。这种级别说明了硬件资源的使用情况。

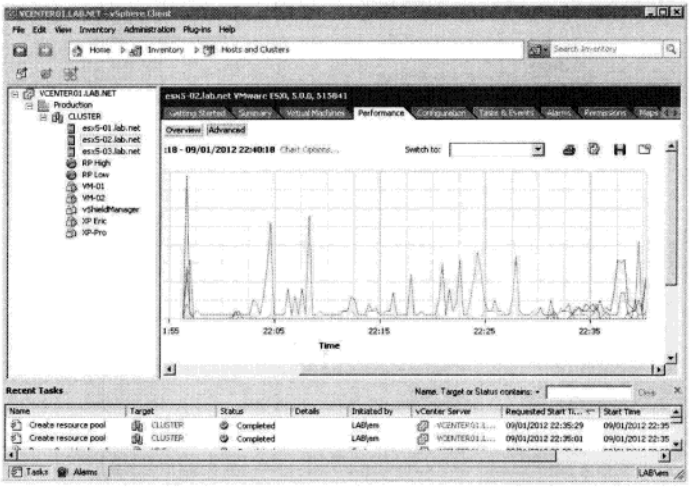


图 7-25 vSphere 5 性能视图采用图形指标

- 级别 2：收集比级别 1 更详细的信息，特别是与内存相关的指标：交换、内存共享和活动内存。这能预测是否可以在同一台主机上放置其他 VM。
- 级别 3：包括级别 1 和级别 2 的相同指标，加上每个网卡、主机总线适配器（HBA）卡或者核心的设备级信息，包括了 CPU 的使用信息。它能够确定 vSMP 的效率（通过比较 vCPU 的就绪时间和延时）和性能。
- 级别 4：包含 vCenter Server 支持的所有信息，可以确定部件是否超载。

级别 1 收集信息时消耗的主机服务器资源很少。级别 2 到级别 4 使用的资源可以忽略。级别 4 应该只用于短期内分析问题。

2. 收集间隔

默认情况下，vCenter Server 提供多种收集间隔：实时、每日、每周、每月、每年和自定义，如图 7-26 所示。每种间隔指定 vCenter 数据库记录的统计数字的持续时间。数据每隔 20 秒收集一次。

- 实时统计数字不存储在数据库中。
- “每日”收集采用实时收集（每 20 秒）方式，每隔 5 分钟整合数据：每小时 12 个数值，24 个小时有 288 个值。
- “每周”收集采用每日收集方式，每隔 30 分钟整合数据：每 30 分钟 1 个数值，每天 48 个值，每周 336 个数值。
- “每月”收集采用每周收集方式，每隔两小时整合数据：每天 12 个数值，每月 360 个数值。

❑“每年”收集采用每月收集方式，每天整合数据，每年 365 个数值。

保留数据的时间越长，数据库需要的存储空间越多。这就是数据不应该保留过长的原因。然后，可以在 vCenter 中查看选中的指标，如图 7-27 所示。

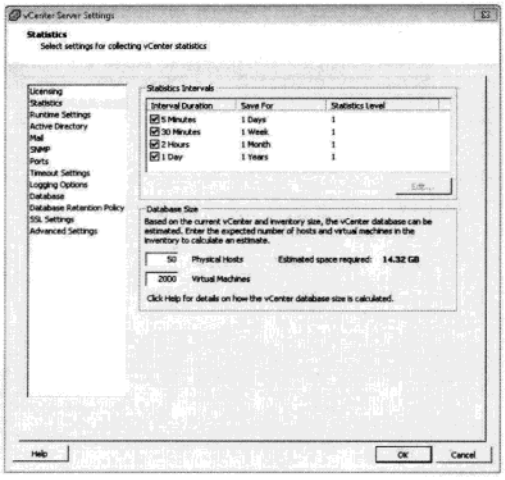


图 7-26　收集统计数字

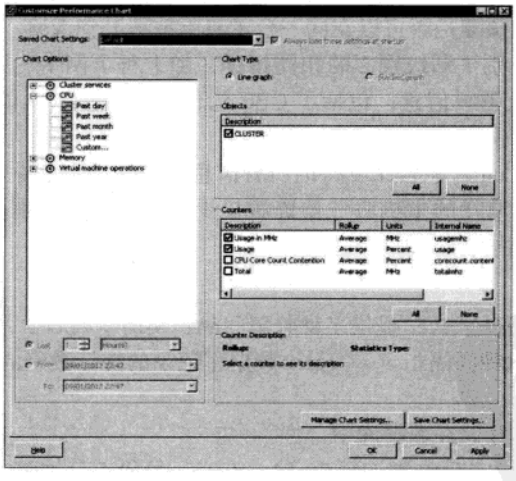


图 7-27　收集间隔设置

7.8.7　复制和模板

虚拟化提供了很大的使用灵活性，这归功于文件中封装的 VM。VM 很容易处理，提供某些可能性。为了更快地部署 VM，你可以创建模板或者复制。

复制基于已经部署的活动 VM，是原始虚拟机的一个映像。这个新机器可以与原始机器完全相同（保留相同的网络标识：SSID、MAC 地址等），或者在内容上相同而在网络上不同。在这种情况下，Sysprep 工具可以生成唯一的 ID（SID）和一个与原始机器不同的 MAC 地址。复制可以用于快速部署来自原始 VM 的特殊 VM，因为没有必要重新安装所有软件，这能够节约许多时间。

模板是一个静态映像——称作黄金母板（Golden Master），可以从共同的基础上大量部署 VM。这类似于大公司用于工业化服务器部署和使服务器同质化的母板。在 VM 安装、配置、优化并处于稳定状态之后，必须创建一个模板。

模板是一个不能启动的非活动 VM，因此不容易修改。（为了修改，必须将它重新转换为 VM）。管理员完全确定给定时刻创建的模板与原来创建的 VM 完全相同。

因为用复制来部署 VM 并不能保证复制从创建以来没有修改过，使用模板更为合适，可以减少使用错误映像的风险。

7.8.8 vApp

vApp 将多个 VM 重新组合成一个单元。vApp 的好处是可以通过定义 VM 的启动顺序，在一次点击中启动（或者停止）一组 VM。VM 可以在定义的时间间隔（默认为 120 秒）或者在 VMware 工具启动之后逐个顺序启动（或者停止）。

在需要以相同方式部署很相似的环境时，vApp 很有价值，例如，用于大企业的远程站点。测试环境如果需要相同的应用或者公共（私有）云服务以快速构建完整的架构，也可以使用 vApp。

7.8.9 最佳实践

下面的最佳实践能改进虚拟环境的管理效率：

- 禁用屏幕保护程序：屏幕保护程序消耗 CPU 事件。这种消耗相对低，但是如果同一台服务器上运行几十个 VM，累积的 CPU 时间就很可观。因此最好是禁用它们。
- 不要使用桌面背景：避免使用桌面背景，它们也会使用 CPU 时间和内存，往往没有理由。
- 限制防病毒扫描和自动更新：避免过于频繁地扫描磁盘，并计划更新。过于频繁扫描磁盘的防病毒技术对性能的影响很大。磁盘扫描应该规划在特定时间进行，避免在 ESXi 主机服务器上的活动密集期进行。
- 检查电力消耗选项：在 Windows 2008 Server 中，默认选项为 Balanced（平衡）。将其修改为 High Performance（高性能），可以完全利用 VM 性能。
- 优化 VM 大小：特别是内存。
- 不要分配附加的虚拟硬件：这些硬件可能消耗 CPU 时间。
- 断开没有必要的端口，如 USB、串行口和并行口，以及磁盘或者 CD-ROM 驱动程序：客户操作系统经常访问所有设备，包括 CD-ROM 驱动程序。如果多个 VM 与服务器物理 CD-ROM 驱动器通信，性能可能受影响，最好使用 ISO 映像文件。

7.8.10 精心规划的架构是关键

在任何环境中，精心规划的架构可以最大限度地利用技术。vSphere 也不例外，但是 vSphere 的许多核心设计也有利于灵活性。某些决策看似关键，但是你往往也可以根据事实解决问题。

正确地安装 vSphere 很容易。ESXi 占用资源很少且经过简化，可以很快安装，vCenter 还有一个向导，进一步帮助你成功地进行安装。配置 VM 时，重要的是确定合适的规模。和物理服务器不同，VM 的规模过大往往对性能有不利的影响。vSphere 提供丰富的方法以确保合适的性能，例如保留、限制和共享。大部分生产环境使用这些特性，但是使用当中应该谨慎。VM 是一个强大的实体，通过实施本章中的解决方案，你可以最大限度地利用其潜力。

第8章
管理虚拟化项目

在本章中，我们以一个决定加速信息系统中虚拟化渗透率的大企业作为例子。这家公司已经熟悉了虚拟化技术。大约30%的服务器环境已经完成了虚拟化。它希望进入第二阶段，大规模实施虚拟化，包括运行关键应用的服务器。为了确保项目的成功，建立了个项目团队。

8.1 背景

这家企业的IT系统由硬件和软件上都已经过时的服务器组成。362台服务器（x86系统）分布在两个数据中心，50%以上的服务器使用时间已经超过4年，在当年年底将到达制造商的支持期限。IT预算逐年缩减，无法应付这些服务器的替换成本。某些陈旧的应用（10年以上）仍然在用，因为兼容性原因无法迁移到当前的硬件和操作系统中。该公司甚至没有重新安装这些应用所需的文档。

IT部门面临压力。它必须应对内部客户的新需求（每年40项），公司管理层要求在缩减成本的同时提供更高的服务水平。在紧缩的经济环境下，内部IT团队承受了沉重的工作负担，因为和许多服务供应商的协议都没有更新。

数据中心是另一个忧虑的来源——它们正在接近占地空间和电力消耗的极限（只有几千瓦的余量了）。无法添加新的硬件，也就无法实施新项目。

公司有一个灾难恢复计划（DRP），但是不令人满意，只有5%的应用安全地放在远程站点上，偶尔进行测试。

备份也有问题，部署了多种解决方案，但是没有全局的管理，备份窗口也常常过窄。许多从磁带上进行的恢复因为缺乏备份测试而遭到失败。

公司决定更多依靠最适合于应对这些问题的虚拟化技术，它决定继续大规模部署所有应用的虚拟化，甚至包括最关键的应用。

8.1.1 目标

IT部门定义如下目标：

- ❑ 成本降低
  - ❑ 降低数据中心的电力消耗
  - ❑ 基础架构合理化
  - ❑ 管理员执行的某些任务的自动化
- ❑ 实施DRP，恢复点目标（RPO）设定为0，恢复时间目标（RTO）设置为整个IT系统少于4个小时，加强生产站点安全。
- ❑ IT系统必须成为公司增长的引擎。它必须有利于发明、创造和新的项目。

IT部门还确定了在直接目标之外的如下远景项目：

- ❑ 部署云类型服务
- ❑ 为最终用户和项目团队提供IT服务的简单访问
- ❑ 采用成本分摊工具
- ❑ 实施统一备份解决方案

- 实施第三个复制站点
- 标准化过程

8.1.2 选择解决方案的标准

为了评估现有虚拟化解决方案能否达到目标，该公司研究了来自 3 家主要公司的产品：Microsoft、Citrix 和 VMware。在深入研究不同解决方案之后，该公司根据如下的标准选择了 VMware：

- 产品在大型数据中心的成熟度
- 解决方案的可升级特性
- 通过 VMware 高级特性获得高服务水平的可能性
- 使用单个工具进行管理、虚拟化平台监控和 DRP，在 vCenter 中集成站点恢复管理器。
- 健全的生态系统，具有许多选项（包括迁移工具、与备份工具的广泛兼容性，以及报告工具）
- 实施全局备份策略的能力
- 原有的内部 VMware 技能

8.2 项目各阶段

如图 8-1 所示，这个虚拟化项目计划分为 4 个阶段：规划、设计、实施和管理。

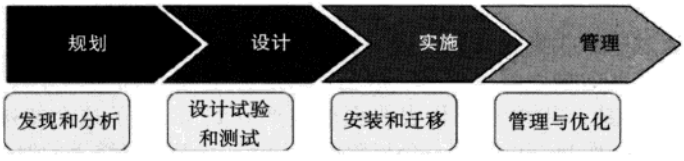


图 8-1 虚拟化项目规划的各个阶段

对任何虚拟化项目（包括关键生产服务器），规划阶段的目标都是初步的资格研究。规划阶段包括两个步骤：

1）发现，包括收集服务器、存储和生产设备的所有信息，草拟生产计划。

2）分析，包括解读发现步骤的结果，草拟优化的整合方案。

设计阶段涉及提出各种目标方案，考虑资格研究的结果，并提供目标平台所需的技术规范和性能指标。这保证了性能水平至少与虚拟化之前相同（实际上要高得多）。在这个阶段中，要草拟详细的参考架构。这个文档用于设备规范的拟定，使得不同的供应商提供符合规定需求的解决方案。

实施阶段按照 IT 部门定义的优先级（如服务器保修合同到期、紧急或者容量），确定迁移的时间表。如果项目中最重要的是降低电力消耗（千瓦时，kWh），应该定义最耗电的应用并首先迁移。目标平台的实施通过使用物理 - 虚拟（P2V）转换工具完成，遵照详细参考架构进行。

管理阶段应用最佳实践，提高环境的运营效率，必须遵循严格的规则以维持控制。可以使用 VMware 提供的辅助工具管理日常操作。

8.3 规划

前一小节已经提到过，规划阶段包括发现和分析（见图 8-2）。

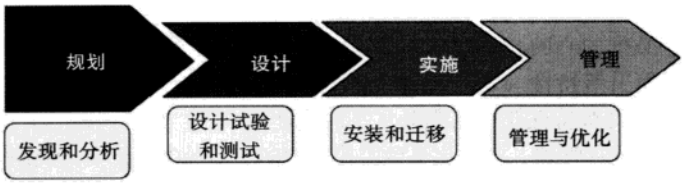


图 8-2 规划阶段

8.3.1 发现

这一阶段包括生产计划（production plan），精确的服务器库存和规定时期内各种要素使用的准确信息。收集的信息用于确定目标架构的规模，确定哪些服务器不具备虚拟化项目的资格。收集的信息还用于站点恢复管理器（SRM）的正确实施。

项目的范围包括基于 x86 处理器的服务器。其他类型服务器（如 UNIX 或者大型主机）被排除在外。

信息的收集通过一台收集服务器完成。基础架构的物理服务器发回一段时间内与生产相关的所有数据，这些数据能够表现活动情况。通常选择 30 天周期，排除 7 月和 8 月（这两个月的业务活动减速，往往不能代表典型的生产计划）。为了这个特殊项目，选择了 PlateSpin PowerRecon 作为分析工具。分析阶段可以在 30 天周期结束的时候开始，涵盖了表 8-1 中列出的技术角度。

表 8-1 发现阶段的规划目标

审计主题	目 标	备 注
OS 类型	验证 OS 与 VMware vSphere5 兼容性矩阵的兼容性	不支持的操作系统不具备资格，因为这意味着通常没有工具能够进行 P2V 或者制造商不支持该 OS
CPU 使用率	能够确定当前服务器场计算能力的平均使用率。对确定目标架构规模很有用	密集使用 CPU 的服务器通常不是虚拟化的好候选，但是如果目标是简化 DRP，即使处理器密集的服务器可以迁移到 vSphere（每台 ESXi 主机 1 个 VM）
CPU 使用模式	能够了解 CPU 活动是否可以预测（每周、每天的特定时间）或者是否随机	可以确定哪些服务器处于危险之中。验证物理机器虚拟化资格的确定因素之一
内存使用确定	能够确定内存平均使用率	内存是确定虚拟化架构规模所要考虑的第一要素

（续）

审计主题	目　标	备　注
确定存储卷大小	定义需要的存储容量和性能	目标存储架构的选择是极其重要的。存储是虚拟化环境最关键的部件，因为如果存储空间和性能（IOPS）的大小没有正确定义，项目可能会失败
网络连接	定义确保网络数据流处于最优状况的先决条件	服务器的特性（支持应用）和报告计划（批处理、备份）应该加以详细分析，以确定网络活动
应用程序分析	能够确定每个应用的关键性和性能。切换到虚拟环境只应该在服务水平和性能至少相同的情况下进行	应该特别注意关键应用的保护和迁移

注意：在此期间，数据收集操作和分析团队不能扰乱目前的生产，必须尊重公司的业务现实。

数据收集遵循逻辑过程。CPU、内存、网络和磁盘部件的活动始终被收集。每过五分钟得到一个最大值。一个小时之后，从这些最大值求出平均值。每个小时这样计算，建立一个24小时的档案。这样可以找出每台服务器一个月的平均值。可以根据每台服务器的档案进行计算，创建如图8-3所示的图形演示，说明所有机器的活动情况。这种收集能够确定服务器的活动及其生产计划。

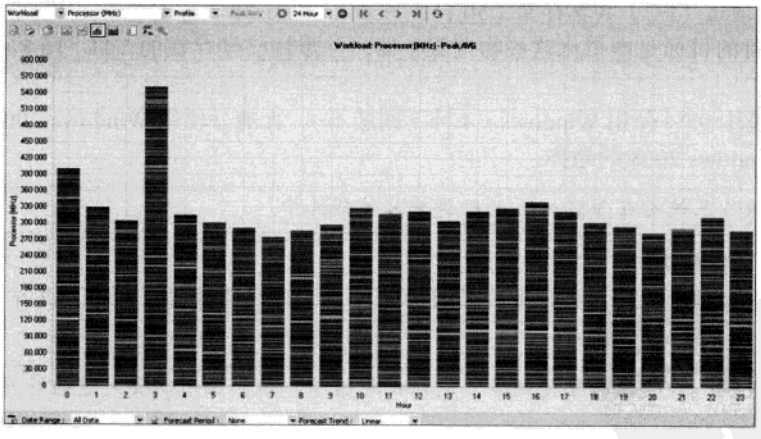


图8-3　数据中心生产计划演示

在图8-3中，堆积柱状图的每种颜色代表数据中心中一台机器的活动。收集的活动累计数代表观察到的最大负载，这是虚拟化平台就绪后必须提供的，以确保至少提供与虚拟化之前相同的性能。注意，大部分活动发生在夜间：备份、批处理、数据库索引、碎片整理等。在某些服务器上，最繁忙的时间在夜里。

如表 8-2 所示，必须设置阈值，以确定哪些服务器具备进行虚拟化的资格。这些阈值根据 VMware 的建议和我们的经验设置。每个机构都可能设置不同的阈值。

表 8-2 阈值

	阈值设置
CPU	6GHz
内存	6GB
磁盘	磁盘传输 1700IOPS
	带宽 20MBps
网络	20MBps

超出这些阈值的服务器将被隔离，没有资格参与这个虚拟化项目。

1. 操作系统

如表 8-3 所示，收集工作提供了服务器的精确描述。

表 8-3 现有物理和虚拟服务器分布

x86 服务器总数	物理服务器数量	VM 数量	连接到 SAN 的服务器数量	在 SAN 上复制的 VM 数量
362	242	120	50	35

362 台物理和虚拟服务器被包含在这个项目的范围之内。大约 30% 的服务器已经用 vSphere 4.1 实现了虚拟化。大部分服务器使用内部存储，但是有些使用了存储区域网络（SAN）光纤通道（FC）共享存储（50 台服务器）。只有 35 台服务器在远程站点上加强了安全。这个项目的目标是虚拟化这些物理服务器，并重用已经存在的 VM。图 8-4 展示了操作系统环境。

当前环境中 95% 使用 Windows（4 种不同版本）。大部分使用 Windows 2003，但是可以看到转向 Windows 2008 的趋势。

注意： 所有操作系统都在 VMware 的硬件兼容性矩阵中。

2. CPU 数据收集

在物理服务器中，50% 是单处理器服务器，44% 是双处理器服务器。只有 6% 使用四处理器。

注意： 经验说明，服务器的插槽越多，虚拟化越困难，因为应用需要大量的资源。

服务器 CPU 使用的深入视图（见图 8-5）展示了活动情况。

在图 8-5 中你可以看到，大部分（87%）服务器合理使用 CPU（低于 4GHz）。有 47 台服务器（13%）有很高的 CPU 活动率（超过 4GHz）。4%（15 台）服务器 CPU 使用率超过阈值 6GHz。

注意： 这些数字是整体平均值，不排除特定服务器上有过量的使用。平均值在能够代表活动的周期内计算。

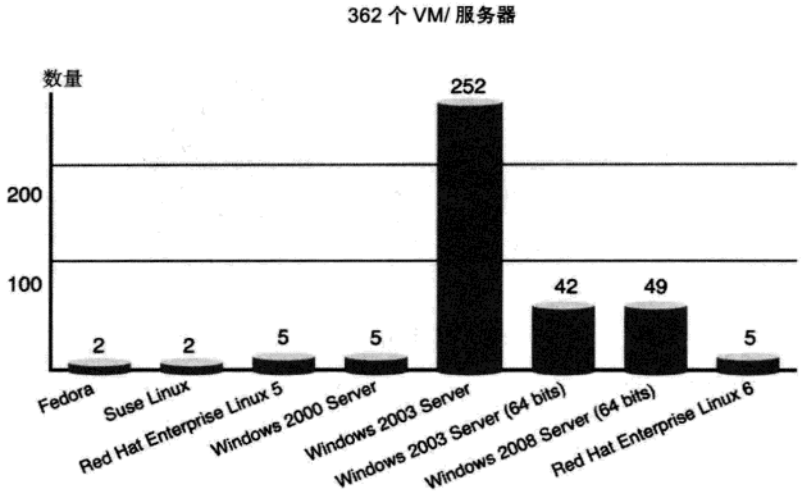


图 8-4 在用操作系统

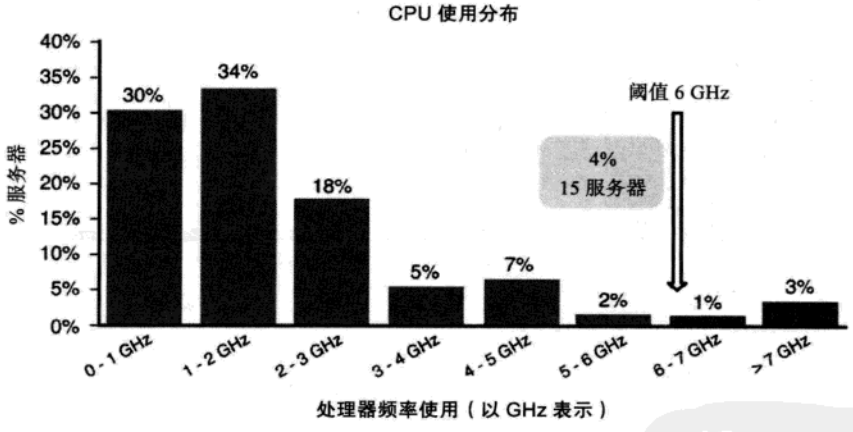


图 8-5 服务器 CPU 活动情况

3. 内存数据收集

统计所有服务器中安装的内存，你可以看到所有内存容量使用了 64%，空闲内存为 36%（见图 8-6）。

图 8-7 显示了一个深入的视图，说明了在用内存的分布。

如图 8-7 所示，90% 的服务器使用小于或者等于 8GB 的 RAM，这些服务器可以虚拟化。3% 的服务器使用大量内存，超过 8GB。这 10 台服务器主要用于 SQL 和 Oracle 数据库。它们超过了阈值，没有资格参与这个项目。

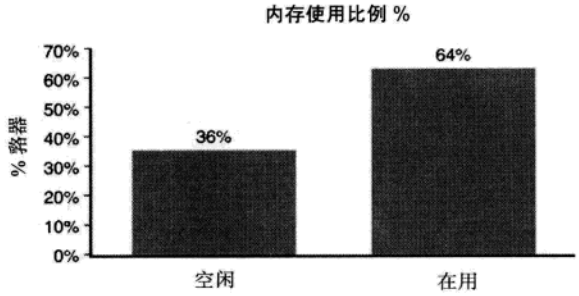


图 8-6 服务器内存总使用率

使用大量内存的服务器必须更详细地进行分析，确定其是否可以虚拟化。在 vSphere 5 中，许可证部分与 VM 配置内存相关，所以在需要大容量内存时必须考虑这一点。

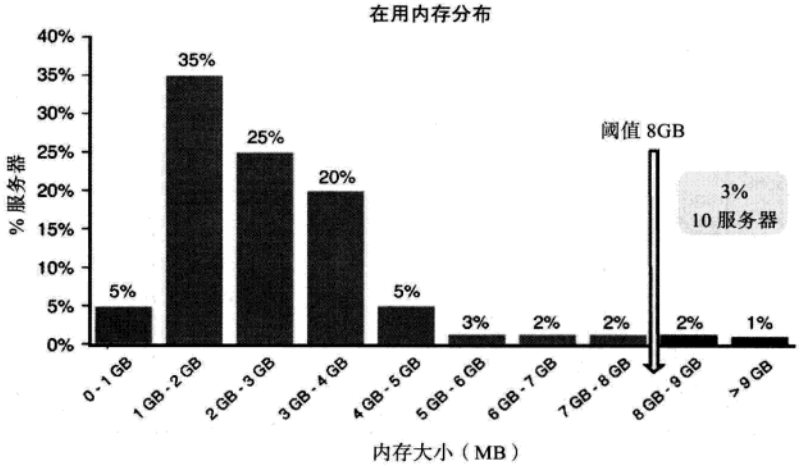


图 8-7 在用内存分布的详细演示

注意： 具有大量内存的服务器并不是在任何情况下都要排除，实际上，如果主要的目标是简单地建立一个 DRP，将这样的服务器保留在项目范围内可能很有趣。

4. 磁盘

有 312 台服务器使用本地存储，50 台连接到 SAN。362 台服务器的总存储容量（本地和 SAN）为 25TB，包括 15TB 的已用空间（55%）。平均每台服务器配置 70G 磁盘容量，其中 40GB 已用。实施集中化存储的架构能够整合和改进存储使用率。

注意： 这以信息可以作为存储规模的基础。存储容量需求是选择磁盘正确类型（vmdk、RDMp 或者 RDMv）的指标之一。

图 8-8 中的图形说明了使用不同磁盘容量的服务器数量。

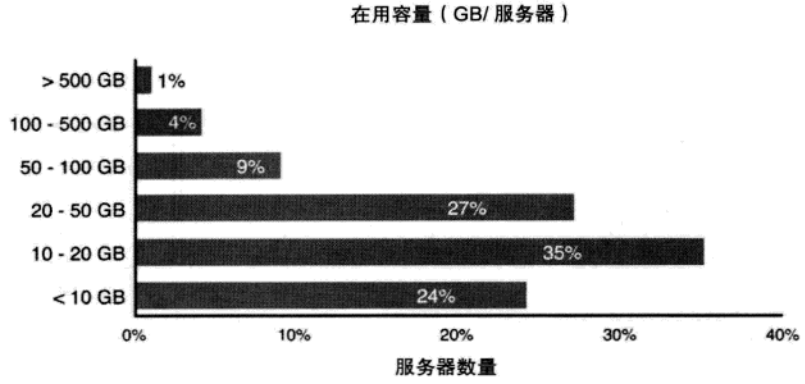


图 8-8　按照磁盘容量分类的服务器数量

86% 的服务器使用合理的容量（100GB 以下），这些服务器可以进行虚拟化。14% 的服务器具有很大的容量，其中两台超过了 1TB。大容量的服务器需要深入研究。

注意： *原始设备映射（RDM）模式磁盘适合大的容量。*

带宽是资格研究的重要方面。通常来说，虚拟化的好候选需要低于 20MBps 的 I/O 磁盘带宽。幸运的是，如图 8-9 所示，大部分服务器都在这个限值以下。

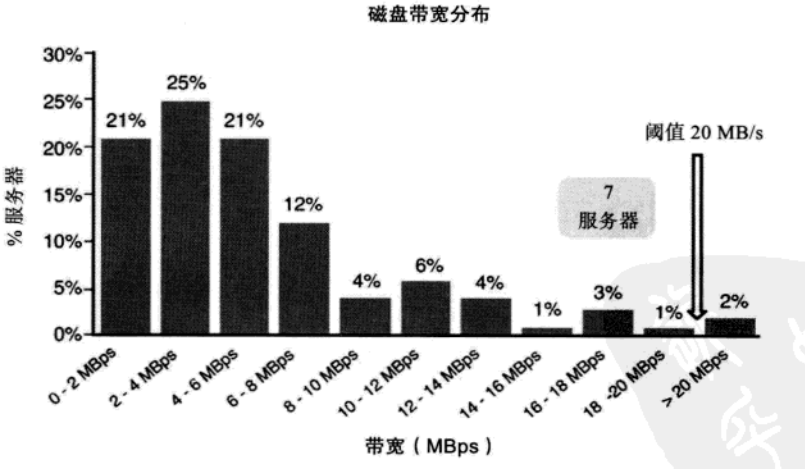


图 8-9　2% 的服务器（7 台服务器）需要超过 20MBps 带宽

需要深入的研究来确认数值和周期。图 8-10 中所示的 I/O 活动是虚拟机所在主机共享磁盘时需要监控的关键要素。

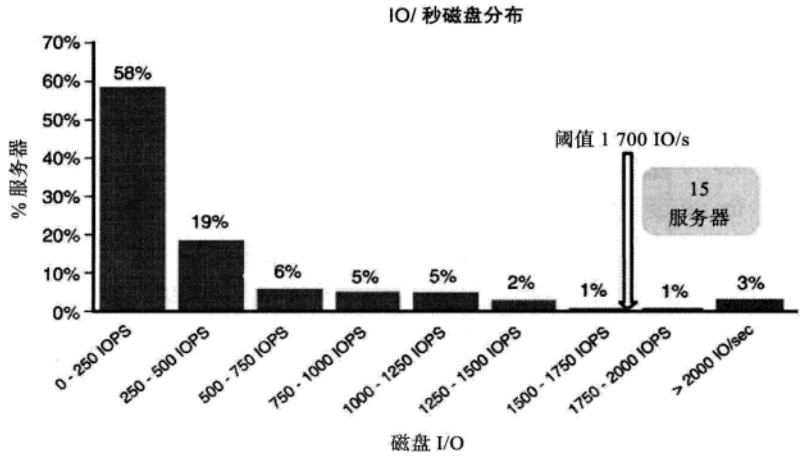


图 8-10　IOPS 活动

在图 8-10 中可以看到，服务器的总体 IOPS 活动很合理。15 台服务器有大量活动（大于 1700IOPS），需要深入研究。

峰值主要出现在夜间。这些活动不一定会妨碍虚拟化迁移，但是有一定的代价。为了让 VM 提供必要的性能，它必须有专用的 RDM 磁盘和足够数量的磁盘主轴来消化负载。

作为目标的一部分，建议为这个 VM 保留 SAN 的专用部分。阻碍虚拟化的是财务因素，而非技术原因。

5. 网络

如图 8-11 所示，264 台服务器使用小于 20MB/ 秒的带宽，10 台服务器使用大于 20MBps 的带宽，其中有用于备份、防病毒软件和扫描应用的服务器。在虚拟化环境中，20MBps 是

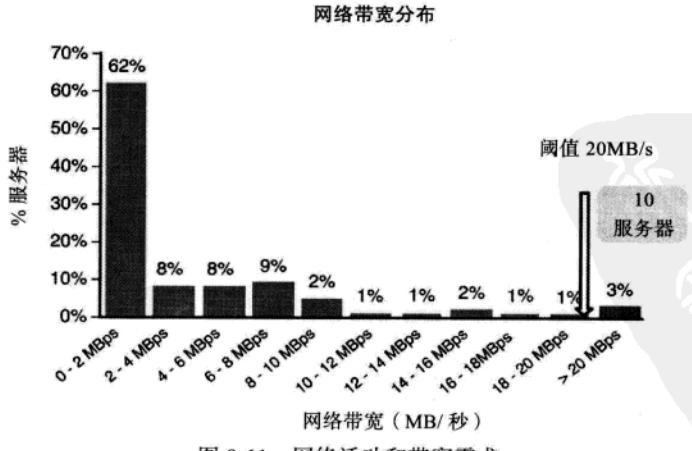


图 8-11　网络活动和带宽需求

机器网络流量的极限。一个千兆以太网适配器通常不能提供超过 70MBps 的带宽。

峰值主要出现在夜间，与备份时段对应。必须针对目标基础架构的备份问题与客户进行讨论。如果建立 SAN 架构，网络活动可能减轻。

6. 应用

收集的数据提供了应用环境的精确描述：

- SQL Server：有 124 台服务器上部署了 SQL。
- Oracle Server：50 台服务器
- 大约 100 个现有 VM 用于测试、开发、生产前等。
- 其余的服务器与基础架构相关：*活动目录*（AD）、*域名系统*（DNS）、Web 服务器、文件服务器、打印服务器、黑莓、业务应用等。

SQL 服务器：我们已经发现 124 台服务器 /VM 部署了 SQL。某些服务器在早上 5 点至下午 7 点之间有很长的队列。队列长度超过 12 的服务器上可能发生性能问题。如果在虚拟化环境中这一问题仍然存在，就有必要在构建这些服务器的*逻辑单元号*（LUN）时增加磁盘主轴数量，为每个队列长度大于 12 的单位分配 RAID 组的一个磁盘。这种增加很重要，否则目标磁盘的争用可能成为严重的问题，可能恶化其他 VM 的性能。

Oracle 服务器：有 50 台服务器运行 Oracle 数据库。必须向 Oracle 直接了解所需要的许可证数量。有些服务器的 CPU 和磁盘 I/O 消耗很大，对这些服务器进行分析能够根据定义好的阈值确定它们是不是虚拟化的候选。

注意： *队列长度在操作系统中表示等待写入的 I/O 操作。也就是在磁盘上未确认的 I/O 等待队列。导致这种等待的原因可能是磁盘上的活动量峰值或者 I/O 管理器与文件系统之间通信速度的降低（例如，主机级别上的防病毒软件或者复制驱动程序）。*

队列长度的审计能够确定存储和服务器之间队列的瓶颈。目前，有 14 台服务器的队列长度证明它们应该被排除在虚拟化项目之外。

8.3.2 分析

公司整个服务器集合中的平均 CPU 使用率是 19%　与我们的其他许多客户相比（他们的使用率在 10% ~ 15% 以下），这是个很不错的数字。大部分服务器使用单 CPU 是高 CPU 使用率的原因。还有 81% 的资源没有利用，CPU 审计发现 24 台服务器有高的活动，其中 4 台的表现没有规律。

服务器内存总使用率为 74%。这是一个不错的数字。

一般来说，理想的虚拟化候选是活动量可预测、资源需求相对低（和服务器所能提供的相比）的服务器。

数据收集和分析提供了该公司服务器的平均配置：

平均 CPU 活动量在 1.5GHz 至 2GHz 之间，在用内存 2GB，存储空间 70GB（使用率 40%），运行 Microsoft Windows 2003 Server。

平均配置只能提供参考信息。

为了确定哪些物理服务器将被虚拟化，采取如下的排除标准（exclusion criteria）：

- 在 VMware 硬件兼容性矩阵中找不到的 OS 或者硬件（例如，传真卡、软件狗、旧的操作系统）
- 和前面提到的阈值相比，过于密集的资源使用
- 服务或者性能水平低于物理环境（很少）
- 业务原因（例如，非常敏感的应用）

图 8-12 展示了这个项目中服务器资格调查的结果。

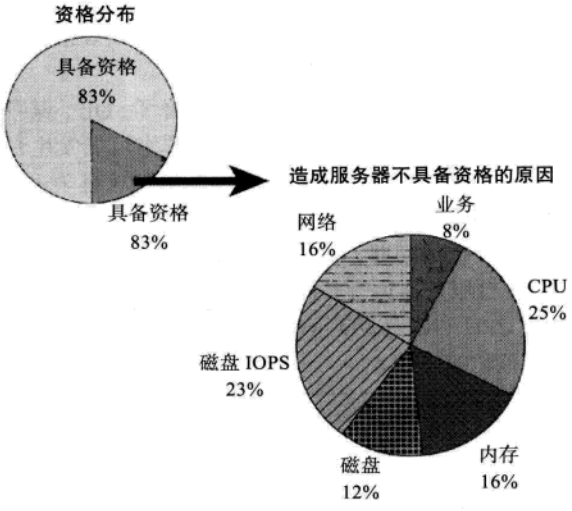


图 8-12　导致服务器不具备虚拟化资格的各种原因

大约 17%（62 台）的服务器被认为“不可虚拟化”。它们不处于这个项目的考虑范围，因为决策必须很快做出，这些服务器可能成为使整个项目处于危险的障碍。这一阶段的重点是获得用户的信任。随后，可以研究虚拟化的可能性，包括：

- 在夜间某些时段磁盘和网络带宽过量使用：是否可能缓和这些负载？
- 早上 10 点到晚上 7 点之间 CPU 使用率超高：这些活动是否可以分散到多个 VM 上？
- 早上 8 点网络带宽的过量使用：负载是否可以分散到更长的周期，将使用率降到 20MBps 以下？
- 在多种情况下的超长队列：为了降低队列长度，在 RAID 组中使用更多磁盘主轴。
- 在多数夜间有大量的磁盘 I/O：是否可能降低负载？必须修订备份策略。

8.4　设计

在设计阶段，重点是目标架构和规模的确定（见图 8-13）。

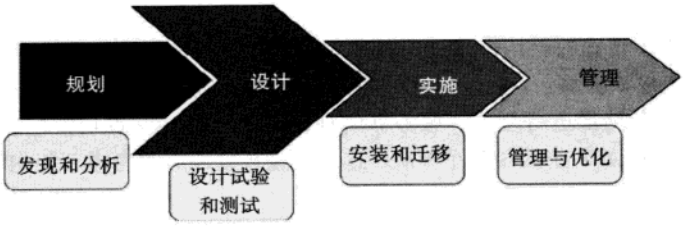


图 8-13 设计阶段

1. 目标架构

有资格参与虚拟化的物理服务器数为 180 个（242 台服务器减去 60 台）。加上现有的 VM，目标架构共有 300 个 VM。为了达到确立的目标，目标架构如图 8-14 所示。

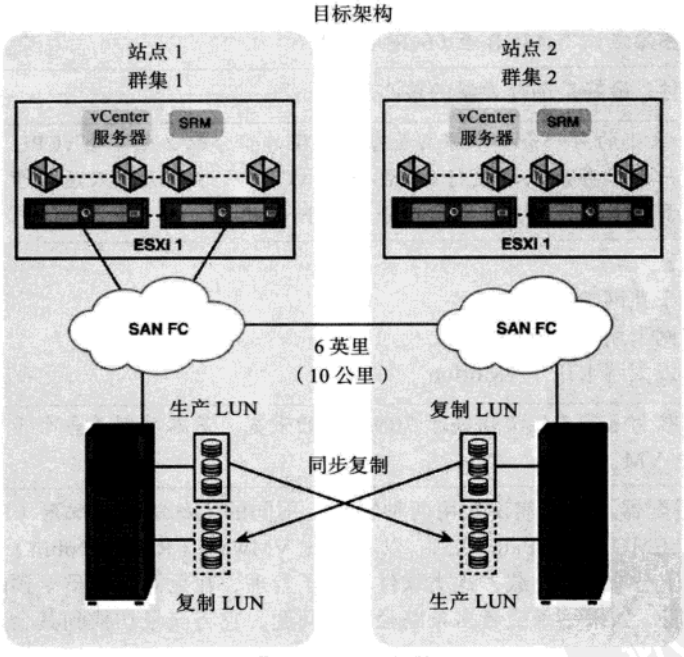


图 8-14 目标架构

目标架构包括 2 个数据中心：站点 1 和站点 2，相距 6 英里（10 公里）。每个站点的所有 ESXi 服务器组成一个群集，连接到一个 SAN FC，两个存储阵列同步复制。DRP 考虑了 SRM 5 的使用。

注意： 实施这种解决方案符合 2010 年 12 月 16 日发布的《巴塞尔第 3 版协议》，该协议要求（金融机构）有两个备份站点。

2. 确定规模

发现阶段提供了精确的已用容量，这是实现最佳架构设计需要的根本信息。公司将实施VMware的SRM5解决方案，在重大事故在另一个站点上完整恢复主用站点。该架构分布在两个站点上，所有存储都进行复制。服务器的规模必须能够在站点切换之后支持所有负载。负载能力是可以处理DRP中的300个VM。这种容量规划还必须考虑三年内的升级能力，每年升级幅度为30%。

ESXi主机服务器：为了提供最高要求情况下的计算能力，必须安装16台Intel Westmere E5670（2个处理器、6核心、2.93GHz）服务器（每个站点8台），每台服务器96GB RAM。

内存是确定目标架构规模时首先考虑的要素。平均有效容量为2.3GB/服务器，代表着总内存容量大约为700GB。高级的内存超量配置技术能够保证活动峰值时保持好的工作状况。它们使内存数量可靠，不需要投入过多的内存。

提示： 确定内存大小的基本原则是每个物理核心4GB至8GB RAM。例如，共有12个核心的一台服务器应该有48GB至96GB RAM。

选择处理器时，最好是选择大量的核心。

提示： 确定CPU大小的另一条基本原则是每个物理核心分配2至4个vCPU。例如，具有12个核心的一台服务器可以运行24至48个vCPU。这条原则只适用于第2层和第3层应用。对于第1层应用，每个核心分配1个vCPU。

网卡配置如下：

- ❑ 两个物理千兆网卡用于VM
- ❑ 两个千兆网卡用于管理
- ❑ 两个千兆以太网卡用于vMotion

提示： 千兆以太网卡通常不能提供高于70MBps的带宽。基本原则是每个千兆网卡不能服务超过10个VM。

至于SAN适配器，我们建议使用两种物理上不同的主机总线适配器（HBA）卡，加上冗余软件（例如EMC PowerPath VE），以利用比VMware（Round Robin）更高级的特性。通过在可用路径上分布I/O负载，这类软件简化了管理，在无需管理员干预的情况下得到更好的I/O磁盘性能。如果一条或者多条路径出现问题，它将流量切换到其余路径，自动检测HS链接。

存储需求的大小根据300个具备资格服务器实际使用的存储容量计算。使用如下规则：

需要提供的容量=（有效容量+30%用于升级的容量）+（RAM总容量）+（20%用于快照的容量）+10%安全容量

因此，总容量为12TB+3.6TB+1.3TB+3.3TB+2TB=22.2TB。

DRP必须保留同样的容量，所以上述结果要加倍（总计44.4TB），以得到预计容量（见表8-4）。

表 8-4 预计容量

站点	生产	复制	总计
站点 A	12.2TB	10TB	22.2TB
站点 B	10TB	12.2TB	22.2TB

预计每年有 30% 的增加（见表 8-5）。

表 8-5 后两年估计增加

	2012 年	2013 年	2014 年
ESXi 5 数量	12.2TB	14TB	16TB
VM 数量	300	350	400
存储 TB	22.2	29	38

DRP 涉及 270 台服务器。如果一个站点崩溃，所有 VM，不管其环境或者排名（有 3 个重要性级别）必须在剩下的站点运行。

提示： 对于 VMFS 数据存储，我们建议每个 800GB 的卷最多允许 32 个 vmdk，连接的 ESXi 服务器不超过 8 台。

图 8-15 显示了实施的配置。

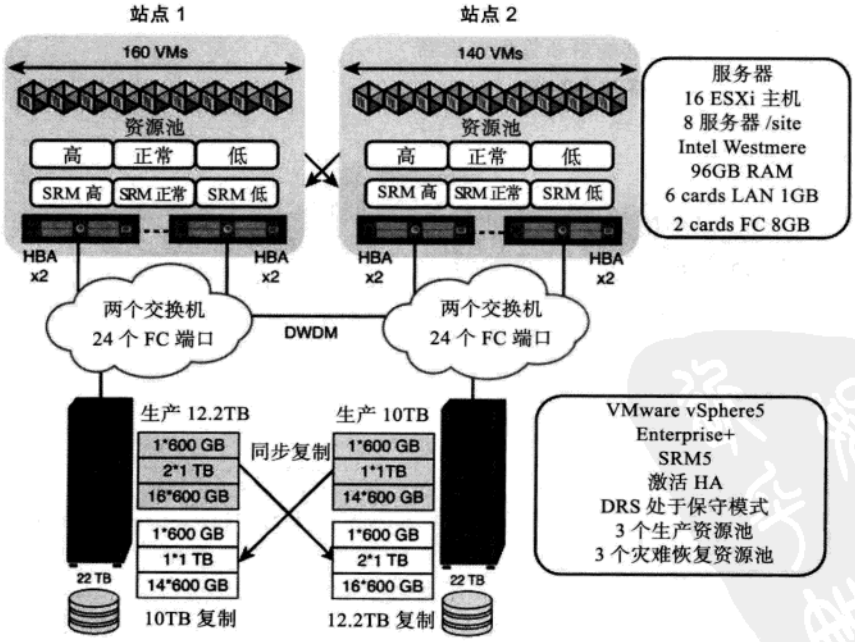


图 8-15 配置的目标架构

8.5 实施

前面定义的要素在实施阶段投入使用，如图 8-16 所示。这一阶段要求团队之间有良好的沟通，这样才能够遵照计划、控制风险。必须严格监控和重新评估先决条件和可交付成果，如果必要，限制风险区域。

1. 安装 vSphere 5 平台

安装按照如下顺序完成：

1）服务器的物理安装，以及机架中的存储阵列
2）服务器部件和 BIOS 固件更新
3）BIOS 配置
4）在服务器上安装 vSphere5
5）vSphere 服务器基本配置
6）安装 HBA 卡冗余软件
7）存储准备、RAID 创建、LUN 创建等
8）新 vSphere 服务器的分区（Zoning）和屏蔽（Masking）
9）本地 vSwitch 及其管理网络的配置
10）为 vMotion 配置 VMkernel 端口
11）在 vCenter 中添加 vSphere 服务器
12）主机配置文件的应用
13）通过 vSphere 服务器的更新管理器更新
14）验证 vSphere 服务器基本配置
15）P2V 迁移

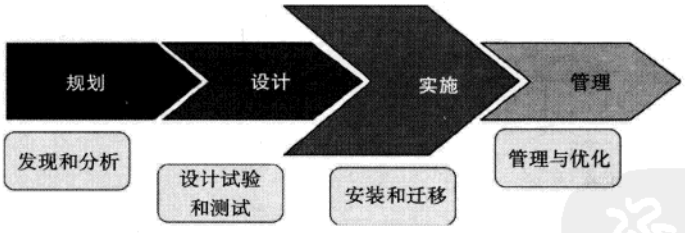


图 8-16 实施阶段

2. P2V 迁移

物理机器向 VM 的转换根据支持结束日期和电力消耗进行组织。迁移将花费 4 周时间。选择 PlateSpin Migrate 作为 P2V 迁移工具，因为它可能造成的服务中断期很短。选择这个工具来代替 VMware Converter 是因为它能使用计划任务，进行物理机器和 ESXi 服务器的差分同步。使用同一个工具，还可以启动目标 VM 并在最终切换之前进行隔离测试，从而确保了 VM 确实能用于关键应用。

迁移方案根据如下优先级标准草拟：

- 电力消耗：目标是快速释放数据中心使用的电力资源，用它来进行所有迁移，同时确保数据中心无法恢复时的服务质量。因此，消耗最多电力（以千瓦时表示）的服务器被认为应该优先进行迁移。
- 硬件陈旧：迁移较老的服务器，有些已经临近制造商保修期限，以降低故障风险。
- 迁移依赖性：物理机器组迁移时要考虑它们之间的依赖性。
- 新需求

迁移之前的周末完成完全同步。在迁移的当天，只有前一个同步点以来修改过的数据块被同步且发送到目标 VM。

进行 P2V 活动之前，必须准备服务器，关闭防病毒软件和资源密集应用。当迁移完成，与硬件相关的工具不再需要，必须卸载。

要让关键的物理机器退役，只要拔除网络电缆，保持 2 周。（最好是从物理上停止服务器，避免硬盘问题。）如果虚拟化迁移出现问题，而物理服务器已经退役，可以通过虚拟－物理（V2P）迁移切换回来。

实施 SRM 5

为了实施 SRM5，必须知道如下信息：

- DRP 包括哪些虚拟机。
- VM 的拓扑结构（特别是与存储相关的结构）。某些 VM 可能有 VM 摘要中无法看到的 RDM 卷。
- 存在哪些 VM 依赖性，这是为了以正确的顺序重启它们。

启动顺序根据业务需求、服务相关 VM、基础架构等标准确定。

示例： VMware 环境的启动顺序之一是活动目录，然后是 DNS，接着是 vCenter 数据库、vCenter Server 等。

SRM 必须涵盖整个机构。VM 必须在生产站点和备份站点上找到相同的资源层次结构。网络资源、资源池和 VM 组织文件在两个站点之间重新映射。因此，关键的是正确地组织和命名 VM，以便在以后找到路径。

1. 一致性分组

有些存储阵列制造商复制解决方案能够创建一致性分组，这很重要。

有些 VM（例如数据库）可以存储在不同的数据存储中。复制所用的一致性分组（称作数据存储分组）允许将不同 LUN 当作单个 LUN 进行逻辑管理，确保了所有 LUN 的相关性。

示例： 如果一个 VM 在数据存储 A 上有一个 vmdk 虚拟磁盘，第二个 vmdk 在数据存储 B 上，就必须创建存储复制的一致性分组。一致性分组计算数据存储 A 和数据存储 B 的相互依赖性。这些数据存储不能分开，它们组成了一个相关的整体。

VMFS 数据存储可能是一个独立的 LUN，也可以由多个 LUN 扩展组成。如果数据存

储由多个扩展组成，一致性分组能够保证组成 VMFS 的不同 LUN 的复制状态相关，即所有 LUN 处于相同的复制状态。

在两个数据中心上，必须对目标群集中的所有 ESXi 主机引入一个 2GB 的 LUN，称作占位数据存储（placeholder）。它代表每个受保护 VM（源机器的 v m x 文件）的配置信息，确保目标平台上的业务恢复。

2. 映射资源层次结构

我们强烈建议正确地标识站点 1 和站点 2 之间的资源池。例如，在生产站点，创建 3 个资源池：高、普通和低。在站点 2 则创建 SRM-High、SRM-Normal 和 SRM-Low。

前缀清晰地表明哪些资源池保存备份 VM，哪些保存生产 VM。这种专用的生产资源映射保证了生产和备份 VM 的分离。确实，在启用 VM 保护时，源 VM 将在目标环境中被标识。如果生产和备份 VM 不分离，它们在目标站点上将被混在一起，使得日常管理更加复杂。

创建保护组时，可以在数据存储中添加 VM。在这种情况下，“保护组”发现该 VM 不被认为是业务恢复计划的一部分，会发送一封电子邮件，并在 SRM 管理控制台上出现一条警告信息。

8.6 管理

在架构建立之后，它必须开始处理负载。确保 CPU、内存和 I/O 磁盘负载不会饱和，否则会降低性能。图 8-17 所示的管理阶段，是对解决方案的监控和优化。

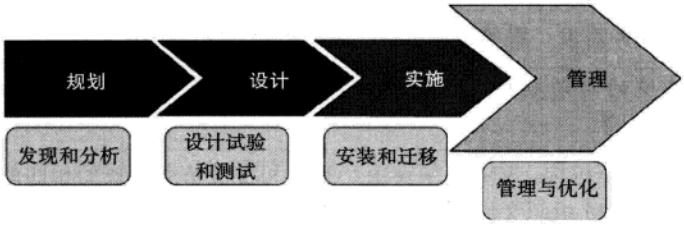


图 8-17 管理阶段

ESXi 主机服务器的物理 CPU 负载不应该超过 75%。如果 CPU 负载非常密集，为关键 VM 使用保留，保证最低需求。关键 VM 必须将共享设置为“高”。监控 VM 的 vCPU 的 CPU 就绪时间，它必须超过 2000 毫秒。

DRS 群集以常规模式激活。为 4 个非常关键的 VM 启用 FT，为 25% 的服务器启用 HA。

管理员必须监控内存气球、内存换入 / 换出和磁盘读 / 写延时（必须低于 20 毫秒）。

在争用的情况下采取一个简单的措施确定资源的优先级。创建三个资源池（高、普通、低）很容易，这对于根据 VM 的重要性来分配 CPU 和 RAM 优先级也很有效。vMotion 和 Storage vMotion 简化了管理和购置新硬件时的规划维护操作。这也减少了这些阶段的服务中断，而这些操作在传统的物理环境中非常容易出问题。

8.7 总结

这个项目已经取得了成功，客户对此完全满意，因为时间表和预算都得到了尊重，目标也已经实现。从节约电力中得到的成本降低甚至超出了预期。

在 16 台 ESXi 主机（每个数据中心 8 台服务器）上运行 300 个 VM。基础架构提供 RPO 为 0 的高服务水平。故障切换测试显示，RTO 小于 4 个小时。IT 团队经过了全面的培训，每年进行两次故障切换测试。

这个系统在电工要求数据中心完全断电进行工作时已经证明很稳定。通过 SRM5 计划迁移，团队将所有生产任务切换到站点 2。这种操作在物理环境中通常很麻烦且容易出错，但是在该系统中没有遇到任何问题，所需要的准备时间也降到最低。

新请求（队列中大约 50 个）都在基础架构部署完毕后很快就得到处理。

基础架构的合理化令人印象深刻。在项目结束时，如图 8-18 所示，只有 78 台服务器就能代替 252 台物理服务器——减少了 69%。

项目完成的时候，252 台现有的物理服务器（242 台物理服务器 +10 台部署了 120 个 VM 的老一代 ESX 服务器）被 78 台物理服务器（16 台 ESXi 服务器 +62 台未作虚拟化的物理服务器）取代。

如图 8-19 所示，虚拟化之后的电力消耗为 16 千瓦时，减少了 48 千瓦时（75%）。

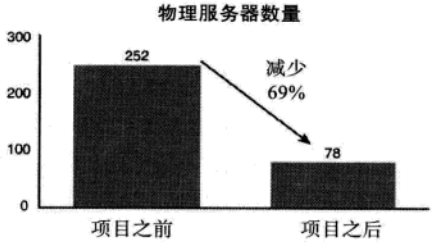


图 8-18 项目前后的物理服务器库存量

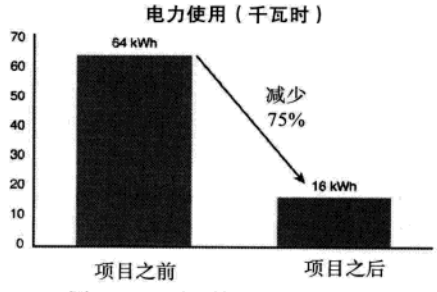


图 8-19 项目前后的电力消耗

如图 8-20 所示，空间的占用明显减少了，从大约 740U（1U=1.75 英寸 =4.445 厘米）减少到 240U（降低 68%）。整合率为 1:18，这是一个很好的平均数。电力消耗减少了 75%。

该公司现在打算自动化一些规程，为内部客户提供服务目录。公司正在研究具有重复数据消除功能的备份解决方案，用统一的解决方案替代多种现有的解决方案。

通过这个虚拟化项目示例，我们可以看到这种技术给数据中心带来的好处。

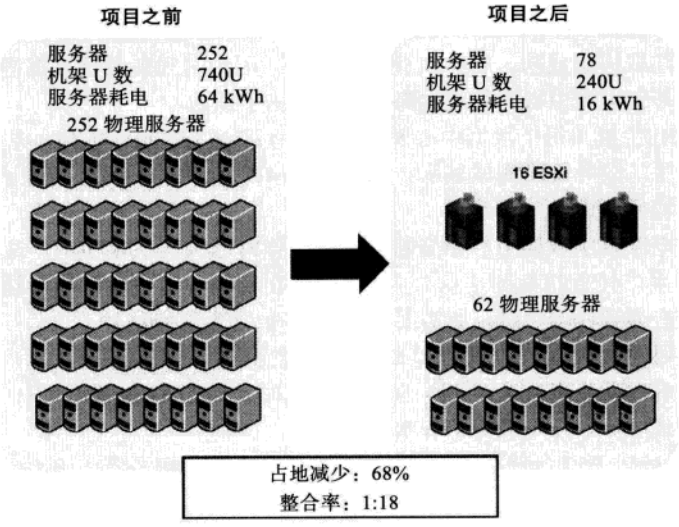


图 8-20 项目前后的空间使用

8.8 一个引人入胜的故事

在你阅读有关这个项目的内容，了解机构所面对的挑战时，你可能会觉得它很熟悉！在任何经济形势下（尤其是形势不好的时候），公司都在寻求降低成本的手段，使用 vSphere 5 的虚拟化是这方面的利器。在这个项目中，你学习了如何确定 VM 主机和 VM 的规模，然后将物理服务器转换为 VM。为此，你列出了 242 台物理服务器的需求，做出了一些关于虚拟化最佳候选的决策，并进行了这些工作。项目中设置的阈值是由项目团队确定的，但是每个机构可能有不同的阈值，vSphere 能够虚拟化大部分 x86 服务器。这个项目还实施了 VMware SRM 5，所以这家机构还能从全面实施 DRP 中获益。

项目产生的节约是惊人的，物理服务器数量减少了 69%，电力消耗减少了 75%。现在，许多行业都推行绿色 IT 倡议，节约已经超出了财务考虑的范畴。虚拟化节约金钱、时间并保护环境，在某些情况下，也保护了你的企业！

常用缩略词

AAM（Automated Availability Manager，自动化可用性管理器）
ADAM（Active Directory Application Mode，活动目录应用模式）
API（Application Program Interface，应用程序编程接口）
ASR（Automatic Server Restart，自动服务器重启）
BIA（Business Impact Analysis，业务影响分析）
BRP（Business Recovery Plan，业务恢复计划）
CBT（Changed Block Tracking，变更数据块跟踪）
CLI（Command-Line Interface，命令行接口）
CNA（Converged Network Adapter，聚合网络适配器）
DAS（Direct Attached Storage，直接连接存储）
DCUI（Direct Console User Interface，直接控制台用户界面）
DMZ（Demilitarized Zone，非军事区）
DNS（Domain Name Server，域名服务器）
DPM（Distributed Power Management，分布式电源管理）
DRP（Disaster Recovery Plan，灾难恢复计划）
DRS（Distributed Resource Scheduler，分布式资源调度器）
ERP（Enterprise Resource Planning，企业资源计划）
EVC（Enhanced vMotion Compatibility，增强型 vMotion 兼容性）

FCoE（Fibre Channel over Ethernet，以太网光纤通道）
FDM（Fault Domain Manager，故障域管理器）
FT（Fault Tolerance，容错）
GPT（GUID Partition Table，GUID 分区表）
HA（High Availability，高可用性）
IA（Information Availability，信息可用性）
ICMP（Internet Control Message Protocol，互联网控制信息协议）
iLO（Intergrated Lights-out，集成关机）
IPMI（Intelligent Power Management Interface，智能电源管理接口）
ITIL（Information Technology Infrastructure Library，信息技术基础架构库）
LUN（Logical Unit Number，逻辑单元号）
MBR（Mater Boot Record，主引导记录）
MMU（Memory Management Unit，内存管理单元）
MPIO（Multi Path I/O，多路径输入 / 输出）
MPP（Multipathing Plugin，多路径插件）
MRU（Most Recently Used，最近使用）
MSCS（Microsoft Cluster Service，微软群集服务）
MTBF（Mean Time Between Failure，平均故障间隔时间）
MTTR（Mean Time To Repair，平均修复时间）
MTU（Maximum Transmission Unit，最大传输单元）
NFS（Network File System，网络文件系统）
NIS（Network Information Service，网络信息服务）
NL-SAS（Near Line-SAS，近线 SAS）
NMP（Native Multipathing，原生多路径）
P2V（Physical-to-Virtual，物理 - 虚拟）
POD（Pool of Datastores，数据存储池）
PSA（Pluggable Storage Architecture，可插入存储架构）
PSP（Path Selection Plug-in，路径选择插件）
PXE（Preboot Execution Environment，预启动执行环境）

[General Information]
书名=VMWARE VSPHERE 5 虚拟数据中心构建指南=VMWARE VSPHERE 5：BUILDING A VIRTUAL DATACENTER
作者=（法）ERIC MAILLE RENE-FRANCOIS MENNECIER著；姚军等译
页数=200
出版社=机械工业出版社
出版日期=2013
SS号=13205460
DX号=000011718587
URL=http://book.szdnet.org.cn/bookDetail.jsp?dxNumber=000011718587&d=6ABA542E83D44B45552DC43D0AD4111A