资源管理指南

ESX Server 3.5、 ESX Server 3i 版本 3.5

VirtualCenter 2.5

资源管理指南
修订时间：20080410
项目：VI-CHS-Q208-534

我们的网站将提供最新技术文档，网址为：

http://www.vmware.com/cn/support/

此外， VMware 网站还提供最新的产品更新。

如果对本文档有任何意见或建议，请将反馈信息提交至以下地址：

docfeedback@vmware.com

VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
www.vmware.com

VMware Global, Inc.
北京办公室 北京市东城区长安街一号东方广场 W2 办公楼 6 层 601 室
邮编：100738 电话：+86-10-8520-0148
上海办公室 上海市浦东新区浦东南路 999 号新梅联合广场 23 楼
邮编：200120 电话 : +86-21-6160 -1168
广州办公室 广州市天河北路 233 号中信广场 7401 室
邮编：510613 电话：+86-20-3877-1938
http://www.vmware.com/cn

目录

关于本书

本手册 （《*资源管理指南*》）介绍 VMware® Virtual Infrastructure 环境的资源管理。重点讲述以下几大主题：

- 资源分配和资源管理概念
- 虚拟机属性和接入控制
- 资源池及其管理方式
- 群集、VMware Distributed Resource Scheduler (DRS)、VMware High Availability (HA) 及其使用方式
- 高级资源管理选项
- 性能注意事项

《资源管理指南》 涵盖了 ESX Server 3.5 和 ESX Server 3i 版本 3.5。为方便讲解，本书使用以下产品命名约定：

- 对于特定于 ESX Server 3.5 的主题，本书使用术语 “ESX Server 3”。
- 对于特定于 ESX Server 3i 版本 3.5 的主题，本书使用术语 “ESX Server 3i”。
- 对于上述两款产品的通用主题，本书使用术语 “ESX Server”。
- 如果讲解内容需要明确辨识某特定版本，本书将使用带版本号的完整名称指代该产品。
- 如果讲解内容适用于 VMware Infrastructure 3 的所有 ESX Server 版本，则本书使用术语 “ESX Server 3.x”。

目标读者

本手册专供要了解系统如何管理资源以及用户如何定制默认行为的系统管理员使用。此外，对于要了解和使用资源池、群集、 DRS 或 HA 的用户，本手册亦是必不可少的。

本手册假定您具有 ESX Server 和 VirtualCenter Server 的相关应用知识。

文档反馈

VMware 欢迎您提出宝贵建议，以便改进我们的文档。如有意见，请将反馈发送到：

docfeedback@vmware.com

VMware Infrastructure 文档

VMware Infrastructure 文档包括 VMware VirtualCenter 和 ESX Server 文档集。

图中使用的缩写

本手册中的图片使用表 1 中列出的缩写形式。

表 1. 缩写

缩写	描述
数据库	VirtualCenter 数据库
数据存储	受管主机的存储
dsk#	受管主机的存储磁盘
host*n*	VirtualCenter 管理的主机
RP	资源池
SAN	受管主机之间共享的存储区域网络类型数据存储
tmplt	模板
user#	具有访问权限的用户
VC	VirtualCenter
VI	VMware Infrastructure Client
VM#	受管主机上的虚拟机

技术支持和教育资源

下面几节介绍为您提供的技术支持资源。可以通过下列网址访问本手册及其他书籍的最新版本：

http://www.vmware.com/support/pubs

联机支持和电话支持

通过在线支持可提交技术支持请求、查看产品和合同信息，以及注册您的产品。网址为：http://www.vmware.com/cn/support。

客户只要拥有相应的支持合同，就可以通过电话支持，尽快获得对优先级高的问题的答复。网址为：http://www.vmware.com/cn/support/phone_support.html。

支持服务项目

了解 VMware 支持服务项目如何帮助您满足业务需求。网址为：http://www.vmware.com/cn/support/services。

VMware 教育服务

VMware 课程提供了大量实践操作环境、案例研究示例，以及设计作为作业参考工具的课程材料。有关 VMware 教育服务的详细信息，请访问 http://mylearn1.vmware.com/mgrreg/index.cfm。

1 资源管理入门

本章将使用一个简单示例介绍资源管理的基本概念。本章将使您逐步了解两种环境中的资源分配，首先是单个主机环境，然后是较复杂的多主机环境。

本章将讨论以下主题：

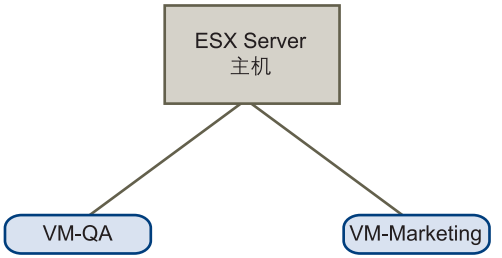
查看主机资源信息

本节将引导您了解一台主机的资源，并了解如何决定资源的使用者。

注意 您还可以使用 ESX Server 系统所连接的 VI Client 或服务器所连接的 VI Web Access Client 执行本章中的多项任务。

假设一家小公司的系统管理员在一台 ESX Server 主机上设置了两台虚拟机，即 VM-QA 和 VM-Marketing。请参见图 1-1。

图 1-1. 具有两台虚拟机的单个主机

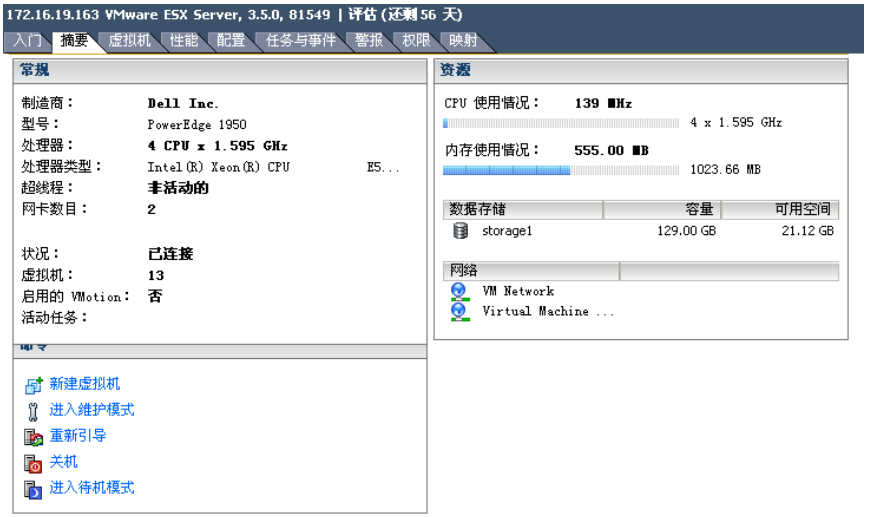


查看关于主机的信息

1 启动 VMware Infrastructure Client (VI Client) 并连接 VirtualCenter Server。

2 在左侧清单面板中选择主机。
选择 **[摘要 (Summary)]** 选项卡后，面板会显示有关该主机的以下信息。

[摘要 (Summary)] 面板	**显示的信息**
[常规 (General)] 面板	显示有关处理器、处理器类型等方面的信息。
[命令 (Commands)] 面板	可让您选择要为选定主机执行的命令。
[资源 (Resources)] 面板	显示有关选定主机的总资源的信息。此面板包含有关该主机所连接的数据存储的信息。



3 有关可用内存的详细信息，请单击 **[配置 (Configuration)]** 选项卡，然后选择 **[内存 (Memory)]**。

该面板会列出总资源、虚拟机所使用的资源量及服务控制台 （仅限 ESX Server 3）所使用的资源量。

可供虚拟机使用的物理内存量总是低于物理主机的内存量，因为虚拟层会占用一些资源。例如，具有双 3.2 GHz CPU 和 2 GB 内存的主机可能只有 6 GHz 的 CPU 资源和 1.5 GB 的内存可供虚拟机使用。

4 有关这两台虚拟机如何使用主机资源的详细信息，请单击 **[分配资源 (Resource Allocation)]** 选项卡。

此时您会看到 **[CPU 预留 (CPU Reservation)]** 和 **[内存预留 (Memory Reservation)]** 、已使用的预留量及可用的预留量。

注意 在显示的 [分配资源 (Resource Allocation)] 选项卡中，没有任何虚拟机正在运行，因此未使用任何 CPU 或内存。您可以在启动虚拟机后再次访问该选项卡。

各字段将会显示以下信息。

字段	描述
CPU 预留	该主机可用的总 CPU 资源。
使用的 CPU 预留	该主机中由正在运行的虚拟机预留的总 CPU 资源。 **注意**：未启动的虚拟机不消耗 CPU 资源。对于已启动的虚拟机，系统将根据每台虚拟机的 **[预留 (Reservation)]** 设置预留 CPU 资源。
未使用的 CPU 预留	该主机中当前未预留的总 CPU 资源。 假设有一台预留 =2 GHz 且完全闲置的虚拟机。该虚拟机预留了 2 GHz，但未使用任何预留。 ■ 其他虚拟机*不能预留* 2 GHz。 ■ 其他虚拟机*可使用* 2 GHz，也就是说闲置的 CPU 预留不会浪费。
内存预留	该主机可用的总内存资源。 如果某台虚拟机具有内存预留但尚未使用全部预留，则未使用的内存可重新分配给其他虚拟机。

字段	描述
使用的内存预留	该主机中由正在运行的虚拟机和虚拟化开销预留的总内存资源。 **注意**：未启动的虚拟机不消耗内存资源。对于已启动的虚拟机，系统将根据每台虚拟机的 **[预留 (Reservation)]** 设置和开销预留内存资源。 虚拟机使用了其全部预留后， ESX Server 将允许该虚拟机保留这些内存，并且不会将其回收，即使该虚拟机闲置并停止使用内存也是如此。
未使用的内存预留	该主机中当前未预留的总内存资源。

5 单击 **[内存 (Memory)]** 或 **[CPU]** 按钮查看所需的信息。

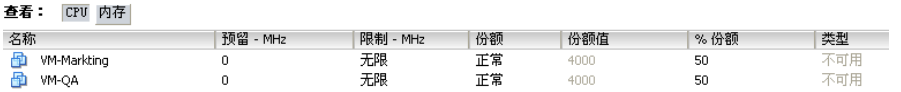


表 1-1. 虚拟机属性

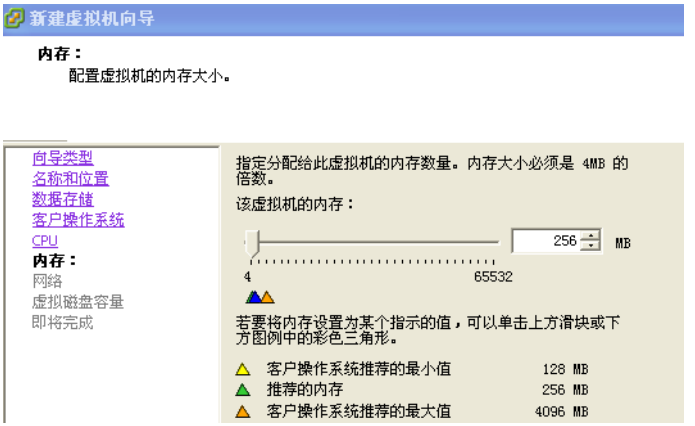
字段	描述
名称	虚拟机的名称。
预留 - MHz/MB	为该虚拟机预留的 CPU 或内存量。 默认情况下未指定任何预留，并显示 **[0]**。请参见 “预留”（第 21 页）。
限制	指定为该虚拟机的上限的 CPU 或内存量。 默认情况下未指定任何限制，并显示 **[无限 (Unlimited)]** 。请参见 “限制”（第 21 页）。
份额	为该虚拟机指定的份额。每台虚拟机都有权使用与其指定份额成比例的资源，该份额受其预留和限制的约束。所得份额是其他虚拟机两倍的虚拟机有权使用两倍的资源。 份额默认为 **[正常 (Normal)]**。请参见 “份额”（第 20 页）。
份额值	分配给该虚拟机的份额数。
% 份额	分配给该虚拟机的份额百分比。
类型	对于资源池，可为 **[可扩展的 (Expandable)]** 或 **[固定的 (Fixed)]**。请参见 “了解可扩展预留”（第 27 页）。

了解虚拟机资源分配

创建虚拟机时，新建虚拟机向导会提示您为该虚拟机指定内存大小。此内存量与物理机中所安装的内存量相同。

注意 ESX Server 主机将向虚拟机提供此内存。主机会将预留中指定的兆字节数直接分配给该虚拟机。超出预留的任何部分都使用主机的物理资源进行分配，如果物理资源不可用，则使用伸缩或交换等特殊技术进行处理。请参见 “ESX Server 主机如何回收内存”（第 129 页）。

图 1-2. 虚拟机内存配置



如果所选择的操作系统支持一个以上的虚拟处理器 (CPU)，系统还会提示您指定虚拟处理器的数量。

图 1-3. 虚拟 CPU 配置

CPU 资源过量使用时， ESX Server 主机将在所有虚拟机之间对物理处理器进行时间划分，以便每台虚拟机在运行时就如同具有指定数目的处理器一样。

运行多台虚拟机的 ESX Server 主机会为各虚拟机分配一定份额的物理资源。在默认的资源分配设置中，与同一主机相关联的所有虚拟机都将按如下方式获得资源：

- 每个虚拟 CPU 获得相等份额的 CPU。这意味着单处理器虚拟机分配到的资源只有双处理器虚拟机的一半。
- 每 MB 虚拟内存大小获得相等的份额。这意味着 8 GB 虚拟机有权使用的内存是 1 GB 虚拟机的 8 倍。

预留主机资源

在某些情况下，系统管理员需要了解某虚拟机的特定内存量是否直接来自 ESX Server 计算机的物理资源。同样，管理员可能还想确保特定虚拟机获得的物理资源百分比总是高于其他虚拟机。

您可以使用各个虚拟机的属性预留主机的物理资源，下一节将对此进行论述。

注意 在大多数情况下，请使用默认设置。有关如何以最佳方式使用自定义资源分配的信息，请参见第 11 章，“最佳做法”（第 151 页）。

虚拟机属性：份额、预留和限制

可为每台虚拟机指定份额、预留 （最小值）和限制 （最大值）。本节介绍指定这些属性的意义。

份额

份额将指定虚拟机的相对优先级或重要性。如果某一虚拟机的资源份额是另一虚拟机的两倍，它将有权消耗两倍的资源。份额通常指定为 **[高 (High)]**、**[正常 (Normal)]** 或 **[低 (Low)]**，这些值将分别按 4:2:1 的比例指定份额值。您还可以选择 **[自定义 (Custom)]** 为各虚拟机分配特定份额数 （表示比例权重）。

指定份额仅对同级虚拟机或资源池 （即在资源池层次结构中具有相同父级的虚拟机或资源池）有意义。同级将根据其相对份额值共享资源，该份额值受预留和限制的约束。有关层次结构和同级概念的说明，请参见 “什么是资源池？”（第 42 页）。

为虚拟机分配份额时，总会指定该虚拟机的相对优先级。

CPU 和内存份额值分别默认为：

- **[高 (High)]** - 每个虚拟 CPU 2000 个份额及每 MB 虚拟机内存 20 个份额
- **[正常 (Normal)]** - 每个虚拟 CPU 1000 个份额及每 MB 虚拟机内存 10 个份额
- **[低 (Low)] -** 每个虚拟 CPU 500 个份额及每 MB 虚拟机内存 5 个份额

您还可以指定 **[自定义 (Custom)]** 份额值。

例如，一台具有两个虚拟 CPU 和 1 GB RAM 且 CPU 和内存份额设置为 **[正常 (Normal)]** 的 SMP 虚拟机具有 2x1000=2000 个 CPU 份额和 10x1024=10240 个内存份额。

注意 具有一个以上虚拟 CPU 的虚拟机称为 SMP（Symmetric Multiprocessing, 对称多处理）虚拟机。

启动新的虚拟机时，每个份额所代表的资源量会改变。这将影响同一资源池内的所有虚拟机。例如：

- 一台 8 GHz 主机上运行着两台虚拟机。它们的 CPU 份额设置为 **[正常 (Normal)]**，因此各得 4 GHz。
- 现在启动了第三台虚拟机。它的 CPU 份额设置为 **[高 (High)]**，这意味着它拥有的份额应该是设置为 **[正常 (Normal)]** 的虚拟机的两倍。新的虚拟机获得 4 GHz，其他两台虚拟机各自仅获得 2 GHz。注意，如果用户为第三台虚拟机指定了 2000 的自定义份额值，也会出现相同的结果。

预留

预留指定虚拟机的保证预留。只有在 CPU 和内存预留可用的情况下，服务器才会允许启动虚拟机。即使物理服务器负载较重，服务器也会确保该资源量。预留用具体单位（兆赫兹 (GHz) 或兆字节 (MB)）表示。资源未使用时， ESX Server 主机可将其提供给其他虚拟机。

例如，假定有 2 GHz 可用，并且为 VM1 和 VM2 各指定了 1 GHz 的预留。现在每台虚拟机都能保证在需要时获得 1 GHz。但是，如果 VM1 只用了 500 MHz，则 VM2 可使用 1.5 GHz。

预留默认为 0。最好指定预留以确保虚拟机始终可使用必要的 CPU 或内存。

限制

限制指定虚拟机的 CPU 或内存上限。服务器分配给虚拟机的资源可大于预留，但决不可大于限制，即使系统上有尚未利用的 CPU 或内存也是如此。限制用具体单位 （兆赫兹 (GHz) 或兆字节 (MB)）表示。

CPU 和内存限制默认为无限。内存限制为无限时，创建虚拟机时为其配置的内存量一般会成为其隐式限制。

多数情况下无需指定限制。指定限制的优缺点如下：

- **优点** - 如果开始时虚拟机的数量较少，并且您想对期望数量的虚拟机进行管理，则分配一个限制将非常有效。但随着用户添加的虚拟机的数量的增加，性能将会降低。因此，您可以指定一个限制，减少可用的资源。
- **缺点** - 如果指定限制，可能会浪费闲置资源。系统不允许虚拟机使用的资源超过限制，即使系统利用不足并有闲置资源可用时也是如此。请仅在有充分理由的情况下指定限制。

接入控制

启动虚拟机时，系统会检查尚未预留的 CPU 和内存资源量。系统将根据可用的未预留资源决定是否可保证为虚拟机所配置的预留 （如果有）。此过程称为*接入控制*。

如果有足够的未预留 CPU 和内存可用，或者没有预留，虚拟机将启动。否则将会出现一条 `Insufficient Resources` 的警告。

注意 除用户指定的内存预留外，各虚拟机还有一个开销内存量。此额外内存承诺包含在接入控制计算中。请参见 "了解内存开销"（第 126 页）。

启用了处于实验阶段的分布式电源管理功能时，可能会将主机置于待机模式 （即关闭）以降低功耗。出于接入控制的目的，这些主机提供的未预留资源被认为可用。如果某虚拟机没有这些资源就无法启动，系统会建议启动足够的待机主机。请参见 "分布式电源管理"（第 66 页）。

更改虚拟机属性

在本章的前面部分，您了解了主机、虚拟机以及其资源分配，但尚未为虚拟机指定份额、预留和限制。在本示例中，假设：

- QA 虚拟机需占用大量内存。您想指定，当系统内存过量使用时， VM-QA 可使用的内存和 CPU 量是市场部虚拟机的两倍。将内存份额和 CPU 份额设置为 **[高 (High)]**。
- 确保市场部虚拟机有一定保证的 CPU 资源量。您可以使用 [预留 (Reservation)] 设置来达到此目的。

编辑虚拟机的资源分配

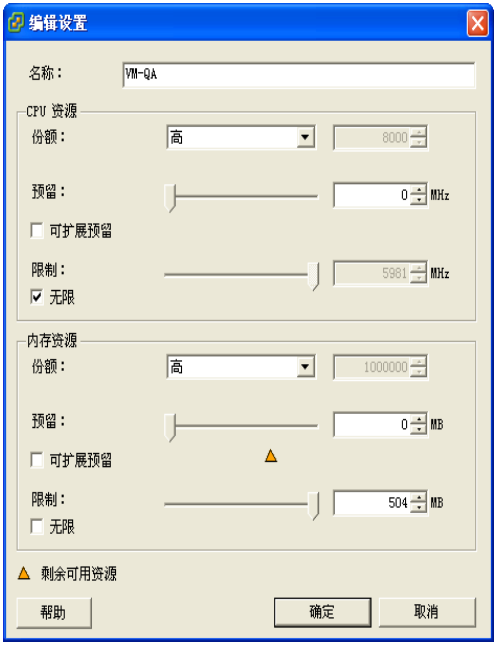
1 启动 VI Client 并连接 VirtualCenter Server。

2 在清单面板中选择主机，然后单击 **[分配资源 (Resource Allocation)]** 选项卡。

3 右键单击要为其更改份额的虚拟机 **[VM-QA]**，然后选择 **[编辑资源设置 (Edit Resource Settings)]**。

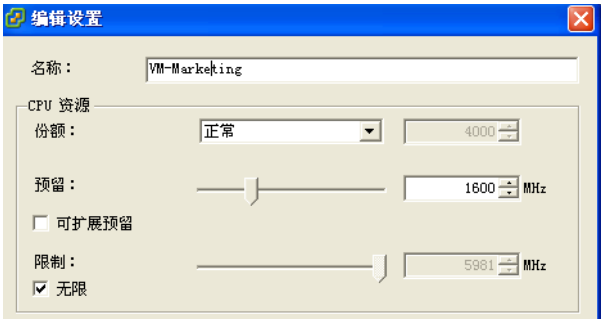
4 在 [CPU 资源 (CPU Resources)] 面板中，从 [份额 (Shares)] 下拉菜单中选择 **[高 (High)]**。

5 在 **[内存资源 (Memory Resources)]** 面板中重复上述步骤，然后单击 **[确定 (OK)]**。



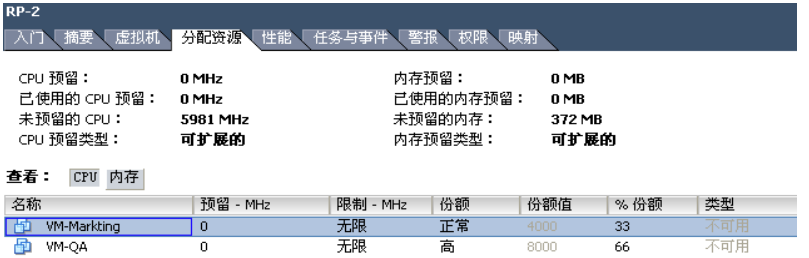
6 右键单击市场部虚拟机 **([VM-Marketing])**。

7 将 **[预留 (Reservation)]** 字段中的值更改为适当数值，然后单击 **[确定 (OK)]**。



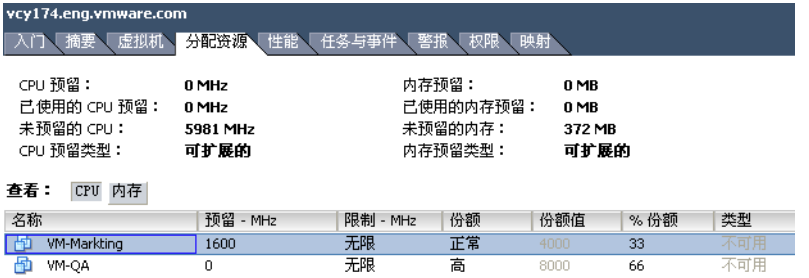
8 完成后，单击 **[确定 (OK)]**。

9 选择主机的 **[分配资源 (Resource Allocation)]** 选项卡，然后单击 **[CPU]**，此时会看到 **[VM-QA]** 的份额是另一虚拟机的两倍。



由于虚拟机尚未启动，因此 **[使用的预留 (Reservation Used)]** 字段尚未改变。

10 启动 **[VM-Marketing]**，然后查看 **[已使用的 CPU 预留 (CPU Reservation Used)]** 和 **[未预留的 CPU (CPU Unreserved)]** 字段的变化情况。



创建并定制资源池

随着组织不断壮大，各组织可提供更快更好的系统，并向不同部门分配更多资源。本节将带您了解如何使用资源池划分主机资源。您还可以将资源池与 VMware 群集配合使用，将群集中所有主机的资源作为一个资源池来管理。

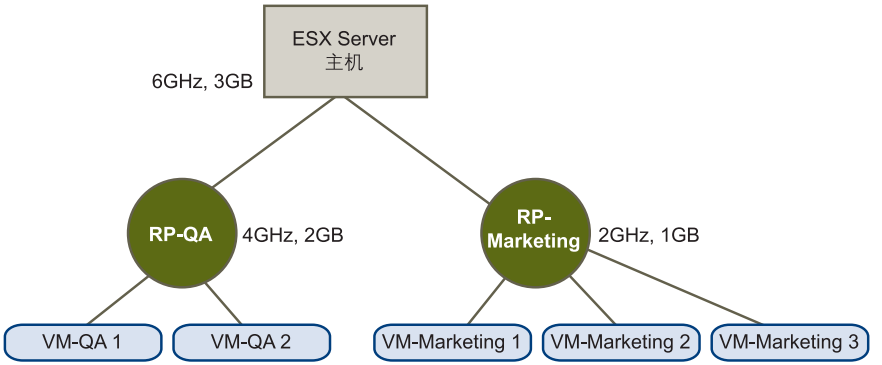
创建资源池时，请指定以下属性：

- 预留、限制和份额的行为方式与虚拟机的情况相同。请参见 “更改虚拟机属性”（第 22 页）。
- **[预留类型 (Reservation Type)]** 属性可让您设置资源池，从而使资源池在其本地可用资源不足时可预留其父级中的可用资源。请参见 “了解可扩展预留”（第 27 页）。

继续上述例子：假设您不再想为 QA 和营销部门各分配一台虚拟机，而是想为每个部门提供预定义数量的资源，部门管理员可根据需要为部门创建虚拟机。

例如，如果开始时主机可提供 6 GHz 的 CPU 和 3 GB 的内存，您可以为 RP-QA 选择 **[高 (High)]** 的份额分配，而为 RP-Marketing 选择 **[正常 (Normal)]** 的份额分配。这将使 RP-QA 获得大约 4 GHz 和 2 GB 的内存，而 RP-Marketing 则获得 2 GHz 和 1 GB。这些资源随后即可供各自资源池内的虚拟机使用。请参见图 1-4。

图 1-4. 具有两个资源池的 ESX Server 主机



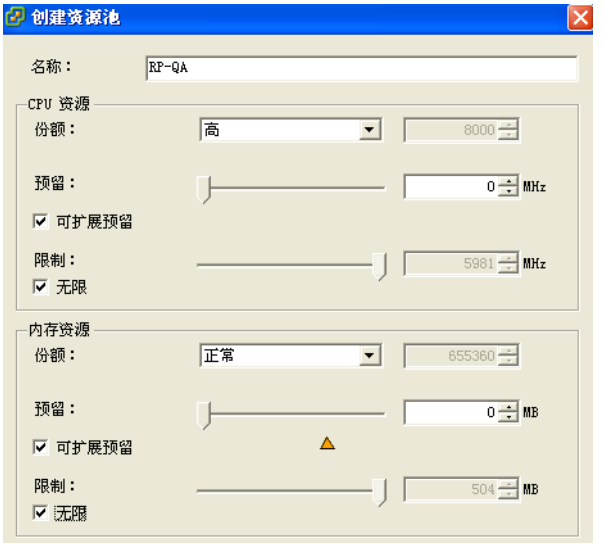
创建并定制资源池

1 启动 VI Client 并连接 VirtualCenter Server。

2 在左侧清单面板中选择主机，并在右侧 **[命令 (Commands)]** 面板中选择 **[创建资源池 (New Resource Pool)]**。

3 在 [创建资源池 (Create Resource Pool)] 对话框中，键入资源池的名称 （例如 RP-QA）。

4 将 RP-QA 的 CPU 和内存资源 **[份额 (Shares)]** 指定为 **[高 (High)]**。



5 创建第二个资源池 RP-Marketing：

a 将 CPU 和内存的 [份额 (Shares)] 保留为 **[正常 (Normal)]**。

b 为 CPU 和内存指定 **[预留 (Reservation)]**。

c 单击 **[确定 (OK)]** 退出。

6 在清单面板中选择主机，然后单击 **[分配资源 (Resource Allocation)]** 选项卡。

资源池已添加到了显示屏幕中。在顶部面板中，已从未预留资源中减去了第二个资源池的 [预留 (Reservation)]。在第二个面板中，可查看资源池信息，其中包括资源池类型。

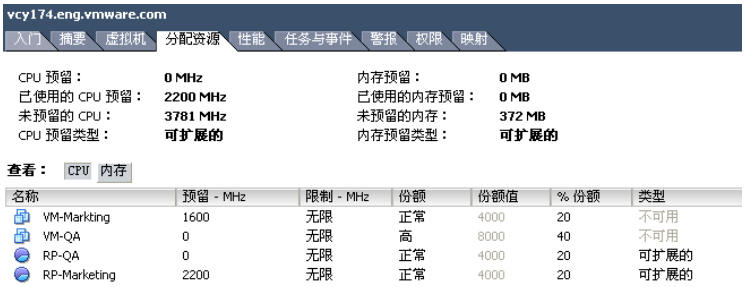


表 1-2 概述了可为资源池指定的值。

表 1-2. 资源池属性

字段	描述
CPU 份额 **内存份额**	可让您为该资源池指定份额。基本原则与虚拟机的情况相同，如 “份额”（第 20 页）中所述。
预留	显示主机为该资源池预留的 CPU 或内存量。默认为 **[0]**。 非零预留将从父级 （主机或资源池）的未预留资源中减去。这些资源被认为是预留资源，无论虚拟机是否与该资源池相关联。
可扩展预留	选中 （默认）此复选框时，如果资源池需要做出的预留高于其自身预留（例如，要启动一台虚拟机），资源池可使用父资源池中的资源并预留这些资源。 请参见 “了解可扩展预留”（第 27 页）。
限制	显示主机分配给选定资源池的 CPU 或内存上限。默认为无限。此默认设置可避免浪费闲置资源。 取消选中 **[无限 (Unlimited)]** 复选框可指定其他限制。 在某些情况下 （例如，如果您想向组管理员分配一定量的资源），资源池限制将很有用。组管理员可根据需要为该组创建虚拟机，但使用的资源决不可超过由限制指定的量。

创建资源池后，可向各资源池添加虚拟机。虚拟机的份额与同一父资源池内的其他虚拟机 （或资源池）相关。

注意 向资源池添加虚拟机之后，选择该资源池的 **[分配资源 (Resource Allocation)]** 选项卡，了解有关预留和未预留资源的信息。

了解可扩展预留

要了解可扩展预留的工作原理，最简单的方法是举例说明。

假设以下应用场景 （如图 1-5 中所示）：

1 父资源池 RP-MOM 具有 6 GHz 的预留及一台预留了 1 GHz 的运行中的虚拟机。

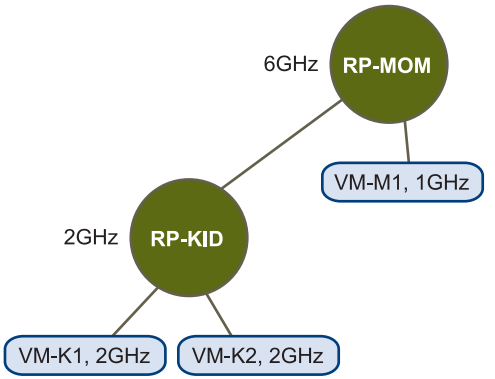
2 您创建一个具有 2 GHz 预留的子资源池 RP-KID，并选中 **[可扩展预留 (Expandable Reservation)]**。

3 您向子资源池添加两台各具有 2 GHz 预留的虚拟机，即 VM-K1 和 VM-K2，并尝试启动它们。

4 VM-K1 可直接从 RP-KID （具有 2 GHz）预留资源。

5 VM-K2 没有本地资源可用，因此它将从父资源池 RP-MOM 中借用资源。RP-MOM 现有资源为 6 GHz 减去 1 GHz （由虚拟机预留）再减去 2 GHz （由 RP-KID 预留），剩下 3 GHz 的未预留资源。利用 3 GHz 的可用资源，您可以启动此 2 GHz 虚拟机。

图 1-5. 可扩展资源池的接入控制，示例 1



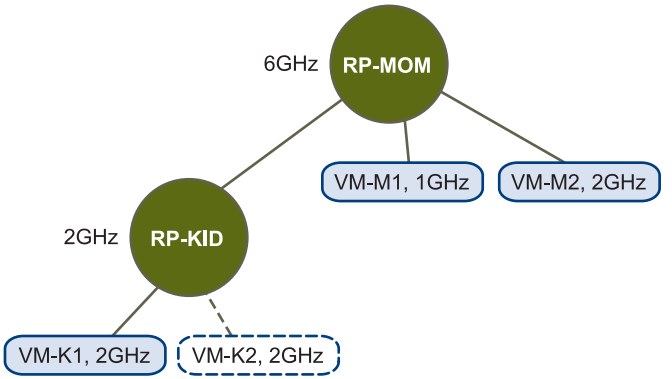
现在假设另一个包含 VM-M1 和 VM-M2 的应用场景 （如图 1-6 中所示）：

1 启动 RP-MOM 中总预留为 3 GHz 的两台虚拟机。

2 您依然可启动 RP-KID 中的 VM-K1，因为本地有 2 GHz 可用。

3 当您尝试启动 VM-K2 时， RP-KID 已无未预留的 CPU 容量，因此会检查其父级。RP-MOM 只有 1 GHz 的未预留容量可用 （RP-MOM 的 5 GHz 已被占用 - 3 GHz 由本地虚拟机预留，2 GHz 由 RP-KID 预留）。因此，您无法启动需要 2 GHz 预留的 VM-K2。

图 1-6. 可扩展资源池的接入控制，示例 2



创建并定制群集

在上一节中，您设置了两个共享单台主机资源的资源池而群集由一组主机组成。如果启用了 VMware DRS （Distributed Resource Scheduling），则群集将支持共享的资源池并为群集内的虚拟机执行位置放置和动态负载平衡。处于实验阶段的分布式电源管理功能也可与 DRS 一同启用。此功能可在有足够的额外容量时建议将主机置于待机电源模式，从而降低群集的功耗。如果启用了 VMware HA (High Availability)，群集将支持故障切换。某一主机出现故障时，会在其他主机上重新启动所有相关联的虚拟机。

注意 您必须获得许可才能使用群集功能。

本节将逐步指导您创建群集，并介绍群集的基本功能。重点内容是基本群集的默认行为。

假设您有一个由三台物理主机组成的群集。每台主机提供 3 GHz 和 1.5 GB，共有 9 GHz 和 4.5 GB 可用。如果为群集启用了 DRS，则可以创建具有不同预留或份额的资源池，从而按照一定标准 （例如按照部门、项目或用户）控制对虚拟机组的总体分配。

对于已启用 DRS 的群集，启动虚拟机时，系统会将虚拟机放置在最合适的物理主机上（或给出放置位置的建议）。具体行为取决于群集的默认自动化级别或特定虚拟机的自动化模式。

创建并定制群集

1 启动 VI Client 并连接 VirtualCenter Server。

2 在左侧清单面板中，右键单击数据中心并选择 **[新建群集 (New Cluster)]**。

3 为该群集命名，并为其启用 HA 和 DRS。

4 保留 DRS 的默认值，即全自动。

5 保留 HA 的主机故障和接入控制的默认设置。

6 为虚拟机的交换文件策略选择合适的选项。

7 单击 **[完成 (Finish)]**。
VirtualCenter Server 将创建具有指定属性的新群集。

有关 DRS、 HA 和可用属性的信息，请参见第 5 章，“创建 VMware 群集”（第 83 页）。

下一项任务是向群集中添加多台主机。使用已启用 DRS 的群集很有意义，即使群集内只有两台主机也是如此。

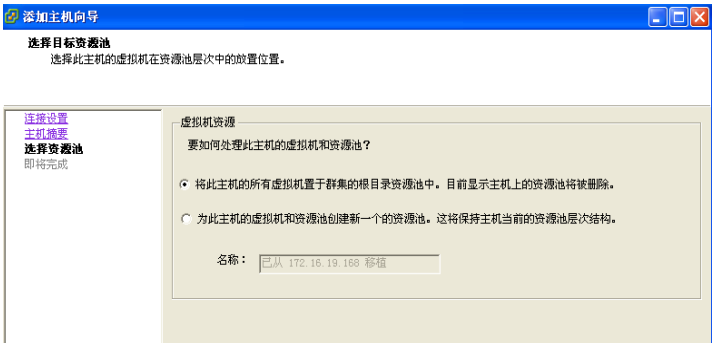
已启用 HA 的群集最多可支持四起并发的主机故障。在以下步骤中，您将向受同一 VirtualCenter Server 管理的群集中添加主机。

向群集中添加主机

1 在 VI Client 的左面板中，选择主机并将其拖至群集图标之上。

如果该群集已启用 DRS，系统将提示您是将该主机的虚拟机直接添加到群集的（不可见的）根资源池中，还是创建新的资源池来代表该主机。根资源池位于顶层且不会显示，因为这些资源为群集所有。

如果该群集未启用 DRS，则会移除所有资源池。



2 选择合适的选项。
如果选择第一个选项，原来位于要添加到群集中的主机上的资源池层级结构将折叠，并且所有资源都将受群集管理。如果为该主机创建了资源池，请选择第二个选项。

注意 如果使用已启用 HA 的群集，则该群集可能会标有红色警告图标，直到添加了足够的主机，可以满足指定的故障切换容量为止。请参见 “有效群集、黄色群集和红色群集”（第 76 页）。

---

3 选择群集，然后选择其 **[分配资源 (Resource Allocation)]** 选项卡以添加更多主机并查看该群集的资源分配信息。

2 资源管理概念

本章将讨论以下主题：

资源的定义

资源包括 CPU、内存、电源、磁盘和网络资源。本手册着重说明 CPU 和内存资源。电源资源可以用分布式电源管理实验功能进行管理。请参见 "分布式电源管理"（第 66 页）。有关磁盘和网络资源的信息，请参见 *《ESX Server 配置指南》*。

资源提供方和使用方

在虚拟基础架构环境中，考虑资源提供方和使用方很有用。

主机 和 *群集* 是物理资源的提供方。

对于主机，可用的资源是主机的硬件规格减去虚拟软件所用的资源。

群集是一组主机。可以使用 VMware VirtualCenter 创建群集，还可以将多个主机添加到群集。VirtualCenter 以联合方式管理这些主机的资源：群集拥有所有主机的所有 CPU 和内存。可以启用群集用于联合负载平衡或故障切换。有关群集的简介，请参见第 4 章，"了解群集"（第 57 页）。

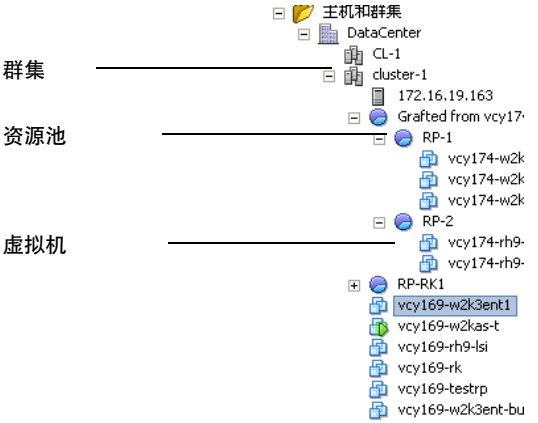
资源池 是灵活管理资源的逻辑抽象。资源池可以分组为层次结构。资源池既可以视为资源提供方，也可以视为资源使用方。资源池向子资源池和虚拟机提供资源。资源池也

是资源使用方，因为资源池会消耗其父级的资源。请参见第 3 章，“了解和管理资源池”（第 41 页）。

虚拟机 是资源使用方。创建期间分配的默认资源设置适用于大多数计算机。可以在以后编辑虚拟机设置，以便分配资源提供方的总 CPU 和内存的共享百分比，还可以分配保证的 CPU 和内存预留量。启动虚拟机时，服务器检查是否有足够的未预留资源可用，并仅在有足够的资源时才允许启动虚拟机。（此过程称为*接入控制*。）

若要查看群集、资源池和虚拟机在 VI Client 中的显示方式，请参见图 2-1。

图 2-1. VI Client 中的群集、资源池和虚拟机



ESX Server 如何管理资源

每个虚拟机均会消耗 ESX Server 主机的一部分 CPU、内存、网络带宽和存储资源。主机将根据以下因素来保证每个虚拟机占用一份基础硬件资源：

- ESX Server 主机 （或群集）可用的资源。
- 预留量、限制和虚拟机份额。虚拟机的这些属性具有默认值，您可以更改这些值来定制资源分配。请参见 “了解虚拟机资源分配”（第 18 页）。
- 启动的虚拟机数和这些虚拟机的资源利用率。
- 管理员在资源池层次结构中分配给资源池的预留量、限制和份额。
- 管理虚拟所需的开销。

服务器按不同的方式管理不同的资源。服务器根据可用的总资源和上述列出的因素来管理 CPU 和内存资源。

服务器基于每台主机来管理网络和磁盘资源。 VMware Server：

- 使用比例分配机制来管理磁盘资源。
- 用网络流量调整来控制网络带宽。

注意 有关磁盘和网络资源的信息，建议参见 *《ESX Server 配置指南》*。*《光纤通道 SAN 配置指南》* 和 *《iSCSI SAN 配置指南》* 提供了将 ESX Server 与 SAN 存储器搭配使用的背景信息和设置信息。

管理员如何配置资源

在很多情况下，系统在创建虚拟机时使用的默认值都是很合适的。而在某些情况下，您可能会发现，对虚拟机进行定制可能非常有效，因为这样可以增加或减少系统分配的资源量。

本指南讨论了虚拟机和资源池属性以及如何定制虚拟机和资源池属性。有关介绍，请参见 “管理员如何影响 CPU 管理”（第 35 页）和 “管理员可以如何影响内存管理”（第 36 页）。

资源利用率和性能

资源利用率是性能的关键。要从虚拟基础架构组件获得最高的性能，最好的方式是确保没有资源瓶颈。请参见第 11 章，“最佳做法”（第 151 页）。有关 `resxtop` 和 `esxtop` 性能测量工具的信息，请参见附录 A，“性能监视实用程序：resxtop 和 esxtop”（第 159 页）。

了解 ESX Server 架构

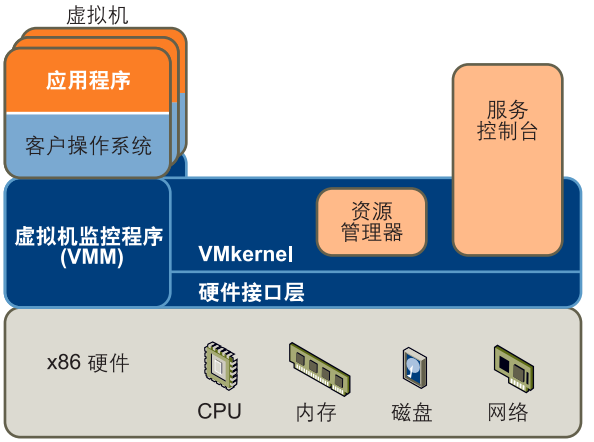
ESX Server 系统的不同组件协同工作，以运行虚拟机并赋予其访问资源的权限。本节简要说明了 ESX Server 架构。

注意 如果您的重点是资源管理的实际应用，请跳过本节。

图 2-2 显示了 ESX Server 主机的主要组件。

注意 只有使用 ESX Server 3 时，图 2-2 中所示的服务控制台组件才适用。ESX Server 3i 未提供服务控制台。

图 2-2. ESX Server 主机组件



VMkernel

VMkernel 是 VMware 开发的高性能操作系统，直接运行在 ESX Server 主机上。VMkernel 控制和管理硬件的大部分物理资源，包括：

- 内存
- 物理处理器
- 存储器和网络控制器

VMkernel 包括 CPU、内存和磁盘访问的调度程序，并具有完全合格的存储器和网络堆栈。它还包括 Virtual Machine File System (VMFS)。 VMFS 是为虚拟机磁盘和交换文件等大型文件优化的分布式文件系统。

VMkernel 资源管理器

资源管理器对基础服务器的物理资源进行分区。它采用资源预留量和比例分配调度等机制向启动的虚拟机分配 CPU、内存和磁盘资源。有关资源分配的信息，请参见第 9 章，“高级资源管理”（第 115 页）。

用户可以为每个虚拟机指定份额、预留量和限制。资源管理器在向每个虚拟机分配 CPU 和内存时会考虑此信息。请参见 “ESX Server 如何管理资源”（第 32 页）。

VMkernel 硬件接口层

硬件接口向 ESX Server （和虚拟机）用户隐藏硬件差异。它启用了特定硬件的服务，并包括设备驱动程序。

虚拟机监视器

虚拟机监视器 (Virtual Machine Monitor, VMM) 负责虚拟化 x86 硬件，包括处理器和内存。当虚拟机开始运行时，控制会转到 VMM，此时 VMM 开始执行来自虚拟机的指令。将控制转到 VMM 涉及设置系统状况，以便 VMM 直接在硬件上运行。

服务控制台

服务控制台是基于 Red Hat Enterprise Linux 3 Update 8 (RHEL 3 U8) 的 Linux 的有限分发版本。服务控制台提供了监视和管理 ESX Server 3 系统的执行环境。ESX Server 3i 未提供服务控制台。

注意 大多数情况下，管理员使用连接 ESX Server 系统或 VirtualCenter Server 的 VI Client 来监视和管理 ESX Server 系统。

管理员如何影响 CPU 管理

通过 VI Client 或使用 Virtual Infrastructure SDK 可以访问有关当前 CPU 分配的信息。

按以下方式指定 CPU 分配：

- 通过 VI Client 使用可用的属性和特殊功能。
  VI Client 图形用户界面 (Graphical User Interface, GUI) 允许用户连接 ESX Server 主机或 VirtualCenter Server。有关介绍，请参见第 1 章，“资源管理入门”（第 13 页）。
- 在某些情况下使用高级设置。请参见第 9 章，“高级资源管理”（第 115 页）。
- 将 Virtual Infrastructure SDK 用于通过脚本进行的 CPU 分配。
- 按照 “超线程”（第 120 页）的讨论来使用超线程。

  注意 通常不建议使用 CPU 关联性。有关 CPU 关联性及其潜在问题的信息，请参见 “使用 CPU 关联性向特定处理器分配虚拟机”（第 117 页）。

如果未定制 CPU 分配，则 ESX Server 主机使用适合大多数情况的默认值。

管理员可以如何影响内存管理

通过 VI Client 或使用 Virtual Infrastructure SDK 可以访问有关当前内存分配的信息和其他状态信息。

按以下方式指定内存分配：

- 通过 VI Client 使用可用的属性和特殊功能。
  VI Client GUI 允许用户连接 ESX Server 主机或 VirtualCenter Server。有关介绍，请参见第 1 章，“资源管理入门”（第 13 页）。
- 在某些情况下使用高级设置。请参见第 9 章，“高级资源管理”（第 115 页）。
- 将 Virtual Infrastructure SDK 用于通过脚本进行的内存分配。

如果未定制内存分配，则 ESX Server 主机使用适合大多数情况的默认值。

有关采用 NUMA 架构的服务器，请参见第 10 章，“配合使用 NUMA 系统和 ESX Server”（第 141 页）。

了解 CPU 和内存虚拟化

本节讨论了虚拟化及其对虚拟机可用资源的意义。

CPU 虚拟化基本知识

可以用一个或多个虚拟处理器配置虚拟机，每个处理器均具有自己的寄存器和控制结构集合。当调度虚拟机时，会调度其虚拟处理器在物理处理器上运行。 VMkernel 资源管理器在物理 CPU 上调度虚拟 CPU，从而管理虚拟机对物理 CPU 资源的访问。 ESX Server 支持最多具有四个虚拟处理器的虚拟机。请参见 “多核处理器”（第 119 页）。

注意 当客户操作系统是 Windows Vista 时，每个虚拟机仅支持两个虚拟 CPU。

查看有关物理和逻辑处理器的信息

1 在 VI Client 中，选择主机，然后单击 **[配置 (Configuration)]** 选项卡。

2 选择 **[处理器 (Processors)]**。

可以查看有关物理处理器数量和类型以及逻辑处理器数量的信息。还可以通过单击 **[属性 (Properties)]** 禁用或启用超线程。

注意 在超线程系统中，每个硬件线程均是逻辑处理器。启用超线程的双核处理器具有两个内核和四个逻辑处理器。

内存虚拟基本知识

VMkernel 管理所有计算机内存。（一种例外情况是在 ESX Server 3 中分配到服务控制台的内存。） VMkernel 会将这种受管计算机内存的一部分拿来自己使用。剩余的内存可供虚拟机使用。虚拟机将计算机内存用于两个用途：每个虚拟机均需要有自己的内存，且 VMM 需要一些内存用于其代码和数据。

查看如何使用主机内存的信息

1 在 VI Client 中，选择主机。

2 单击 **[配置 (Configuration)]** 选项卡。

3 选择 **[内存 (Memory)]**。

可以查看有关总内存的信息和可用于虚拟机的内存信息。在 ESX Server 3 中，还可以查看分配到服务控制台的内存。

虚拟机内存

每个虚拟机均会根据其配置大小消耗内存，还会消耗额外开销内存以用于虚拟。

配置大小。 配置大小是一种由虚拟机的虚拟层来维持的构造。它是提供给客户操作系统的内存量，但独立于分配给虚拟机的物理 RAM 量，取决于下文所述的资源设置（份额、预留量、限制）。

以配置大小为 1 GB 的虚拟机为例。当客户操作系统引导时，系统会认为正运行在具有 1 GB 物理内存的专用计算机上。分配给虚拟机的物理主机内存实际数量取决于其内存资源设置和 ESX Server 主机的内存争用情况。有些情况下，可能向虚拟机分配全部 1 GB。在其他情况下，可能会得到较小的分配量。无论实际分配如何，客户操作系统都会继续运行，就好像正运行在具有 1 GB 物理内存的专用计算机上一样。

份额。 如果可用量超过预留量，则为虚拟机指定相对优先级。请参见 “份额”（第 20 页）。

预留 。 是主机为虚拟机预留的、有保证的物理内存量下限，即使内存过量使用也如此。将预留量设置为确保虚拟机高效运行的足够内存水平，这样就不会有过多的内存分页。

限制。 是主机将分配给虚拟机的物理内存量的上限。虚拟机的内存分配还受其配置大小隐式限制。

开销内存 包括为虚拟机框架缓冲区和各种虚拟数据结构预留的空间。请参见 “了解内存开销”（第 126 页）。

内存过量使用

对于每个运行的虚拟机，系统会为虚拟机的预留量 （若有）和虚拟开销预留物理内存。由于 ESX Server 主机使用内存管理技术，虚拟机可以使用的内存大于物理机 （主机）可用的内存。例如，您可以有一个内存为 2 GB 的主机，并且运行四个虚拟机，每个虚拟机的内存为 1 GB。这种情况下，内存被过量使用。

过量使用有一定的意义，因为通常情况下有些虚拟机负载较轻，而有些虚拟机负载较重，相对活动水平会随着时间的推移而有所差异。

为了改善内存利用率， ESX Server 主机将闲置虚拟机的内存转移给需要更多内存的虚拟机。使用预留量或份额参数可优先向重要的虚拟机分配内存。如果这部分内存未使用，可以用于其他虚拟机。

内存共享

许多工作负载存在跨虚拟机共享内存的机会。例如，几个虚拟机可能正运行同一客户操作系统的实例，加载了相同的应用程序或组件，或包含公共数据。 ESX Server 系统使用专用的分页共享技术安全地消除了内存页的冗余副本。

采用内存共享，由多个虚拟机组成的工作负载通常消耗的内存要少于其在物理机上运行时所需的内存。因此，系统可以有效地支持更高级别的过量使用。

内存共享保存的内存量取决于工作负载特性。许多几乎相同的虚拟机的工作负载可能释放 30% 以上的内存，而有较大差异的工作负载可能节省的内存少于 5%。

3 了解和管理资源池

本章介绍资源池并说明如何使用 Virtual Infrastructure 查看和操作资源池。

本章将讨论以下主题：

所有任务均假定您有相应的操作权限。有关权限及如何设置权限的信息，请参见联机帮助。

什么是资源池？

资源池用于对可用的 CPU 和内存资源进行层次结构式的分区。

每台独立主机和每个 DRS 群集都具有一个 （不可见的）根资源池，该资源池对该主机或群集的资源进行分组。根资源池之所以不显示，是因为主机 （或群集）与根资源池的资源总是相同的。

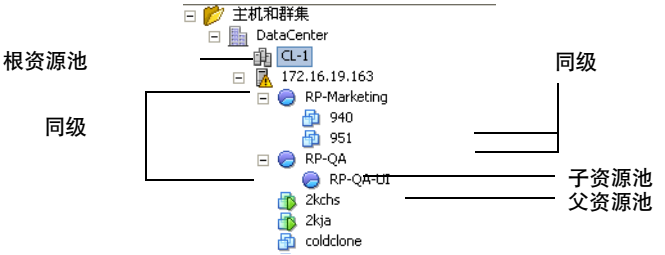
> **注意** VMware DRS 可帮助您平衡虚拟机之间的资源。"了解 VMware DRS"（第 60 页）中将对此进行论述。

如果不创建子资源池，则只存在根资源池。

用户可以创建根资源池的子资源池，也可以创建用户创建的任何子资源池的子资源池。每个子资源池都拥有一部分父级资源，它们又可能具有相继表示更小计算容量单元的子资源池层次结构。

一个资源池可包含多个子资源池和 / 或虚拟机。您可以创建共享资源的层次结构。处于较高级别的资源池称为*父资源池*。处于同一级别的资源池和虚拟机称为*同级*。群集本身表示根资源池。

图 3-1. 资源池层次结构中的父级、子级和同级



在图 3-1 中， RP-QA 是 RP-QA-UI 的父资源池。 RP-Marketing 与 RP-QA 是同级。紧靠 RP-Marketing 下面的两个虚拟机也是同级。

对于每个资源池，均可指定预留、限制、份额以及预留是否应为可扩展。随后该资源池的资源将可用于子资源池和虚拟机。

为什么使用资源池？

通过资源池可以委派对主机 （或群集）资源的控制权，在使用资源池划分群集中的所有资源时，其优势尤为明显。可以创建多个资源池作为主机或群集的直接子级，并对它们进行配置。然后便可向其他个人或组织委派对资源池的控制权。

使用资源池具有下列优点：

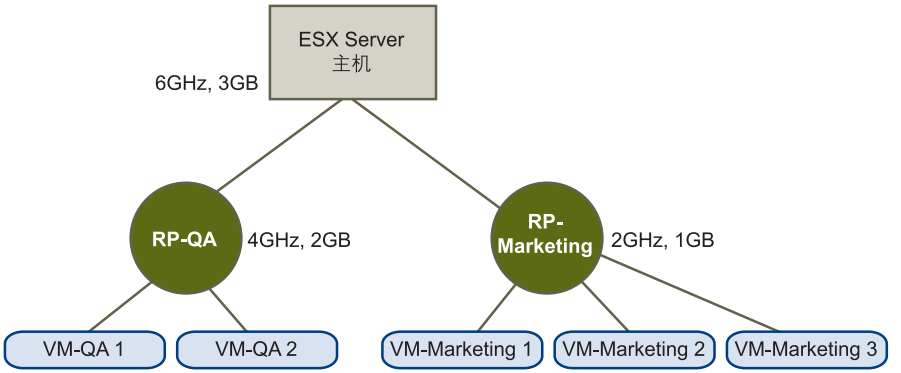
- **灵活的层次结构组织** - 根据需要添加、移除或重组资源池，或者更改资源分配。
- **资源池之间相互隔离，资源池内部相互共享** - 顶级管理员可向部门级管理员提供一个资源池。某一部门级资源池内部的资源分配变化不会对其他不相关的资源池造成不公平的影响。
- **访问控制和委派** - 顶级管理员向部门级管理员提供资源池后，该管理员可以在当前的份额、预留和限制设置向该资源池授予的资源范围内进行所有虚拟机创建和管理操作。委派通常与权限设置共同完成，《*Virtual Infrastructure 简介*》中对权限设置进行了论述。
- **资源与硬件的分离** - 使用已启用 DRS 的群集时，所有主机的资源始终会分配给群集。这意味着管理员可以脱离提供资源的实际主机独立地进行资源管理。如果将三台 2 GB 主机替换为两台 3 GB 主机，您无需对资源分配进行更改。

  这一分离可使管理员更多地考虑聚合计算容量而非单个主机。
- **管理运行多层服务的各组虚拟机** - 您无需对每台虚拟机进行资源设置，而是可以通过更改该组虚拟机所属资源池的设置来控制对这些虚拟机的总体资源分配。

例如，假定一台主机拥有多台虚拟机。营销部门使用其中的三台虚拟机， QA 部门使用两台虚拟机。由于 QA 部门需要更多的 CPU 和内存，管理员为每组创建了一个资源池。管理员将 QA 部门资源池和营销部门资源池的 **[CPU 份额 (CPU Shares)]** 分别设置为 **[高 (High)]** 和 **[正常 (Normal)]**，以便 QA 部门的用户可以运行自动测试。 CPU 和内存资源较少的第二个资源池足以满足营销工作人员的较低负载要求。只要 QA 部门未完全利用所分配的资源，营销部门就可以使用这些可用资源。

图 3-2 中演示了此应用场景。这些数字显示了向资源池的有效分配。

图 3-2. 向资源池分配资源



主资源池和群集资源池

您可以创建独立 ESX Server 主机或 DRS 群集的子资源池。

- 对于独立 ESX Server 主机，可以将资源池作为主机的子级进行创建和管理。每台主机均支持自身的资源池层次结构。
- 如果对未启用 DRS 的群集添加一台主机，主机的资源池层次结构将会丢弃，并且无法创建任何资源池层次结构。
- 对于已启用 DRS 的群集，所有主机的资源将分配给该群集。

  向 DRS 群集添加一台带有资源池的主机时，系统会提示您确定资源池的放置位置。默认情况下将丢弃资源池层次结构，并在虚拟机所处的同一级别添加主机。您可以选择将主机的资源池移植到群集的资源池层次结构上，并为顶层资源池选择一个名称。请参见 “资源池和群集”（第 53 页）。

  由于所有资源均已合并，因此您无需对单个主机进行资源管理，而只需在群集环境中管理所有资源。您可将虚拟机分配给具有预定义特性的资源池。如果随后因添加、移除或升级主机而改变了容量，则可能必须更改资源池的资源分配。

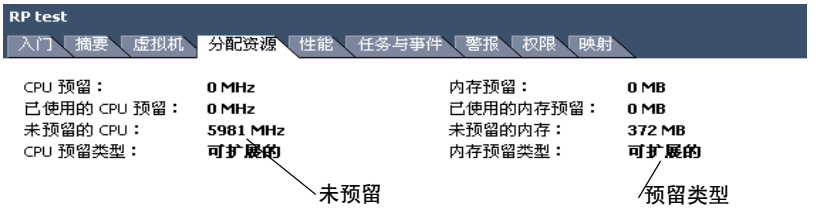
  如果 VirtualCenter Server 不可用，可以通过 ESX Server 主机所连接的 VI Client 进行更改。但是当 VirtualCenter Server 再次可用时，群集可能会变黄 （过量使用）或变红 （无效）。请参见 “有效群集、黄色群集和红色群集”（第 76 页）。如果群集处于自动模式，则当 VirtualCenter Server 再次可用时，VirtualCenter 会重新申请上次已知的群集配置 （并可能撤消您的更改）。

资源池接入控制

在 ESX Server 主机上启动虚拟机时，主机首先会执行基本接入控制，如 “接入控制”（第 22 页）中所述。在资源池内启动虚拟机时，或尝试创建子资源池时，系统会执行其他接入控制以确保不违反资源池的限制。

启动虚拟机或创建资源池之前，请在资源池的 **[分配资源 (Resource Allocation)]** 选项卡中检查 [未预留的 CPU (CPU Unreserved)] 和 [未预留的内存 (Memory Unreserved)] 字段，以确定 （参见图 3-3）是否有足够的资源可用。

图 3-3. 资源池预留信息



如何计算未预留的 CPU 和内存以及是否执行操作取决于预留类型：

- **[固定的 (Fixed)]** 预留类型。系统检查资源池是否有足够的未预留资源。如果有，则可以执行操作。否则将显示一条消息，而且无法执行操作。
- **[可扩展的 (Expandable)]** 预留类型。系统检查资源池是否有足够的资源来满足要求。
  - 如果有足够的资源，则将执行操作。
  - 否则管理服务器将检查父资源池 （直接父级或祖先）中是否有可用资源。如果有，则将执行操作，并预留父资源池资源。否则将显示一条消息，而且不会执行操作。请参见 “了解可扩展预留”（第 27 页）。

系统不允许违反预先配置的 **[预留 (Reservation)]** 或 **[限制 (Limit)]** 设置。每次重新配置资源池或启动虚拟机时，系统都会验证所有参数以确保仍能实现各服务级别保证。

创建资源池

您可以创建任何 ESX Server 3.x 主机、资源池或 DRS 群集的子资源池。

注意 如果已将某台主机添加到群集中，将无法创建该主机的子资源池。如果群集已启用 DRS，则可以创建该群集的子资源池。

创建子资源池时，系统将提示您输入资源池属性信息。系统使用接入控制确保您不能分配不可用的资源。例如，如果您的资源池预留为 10 GB，则在创建了一个预留为 6 GB

的子资源池后，便无法再创建第二个预留为 6 GB 且 **[类型 (Type)]** 设置为 **[固定的 (Fixed)]** 的子资源池。

创建资源池

1 选择所需的父级，然后选择 **[文件 (File)] > [新建 (New)] > [创建资源池 (New Resource Pool)]**（或在 **[摘要 (Summary)]** 选项卡的 [命令 (Commands)] 面板中单击 **[创建资源池 (New Resource Pool)]**）。

2 在 [创建资源池 (New Resource Pool)] 对话框中，为资源池提供下列信息。

字段	描述
名称	新资源池的名称。
CPU 资源	
份额	资源池拥有的、相对于父级总数的 CPU 份额值。同级资源池根据由其预留和限制限定的相对份额值共享资源。您可以选择 **[低 (Low)]**、**[正常 (Normal)]** 或 **[高 (High)]**，也可以选择 **[自定义 (Custom)]** 来指定表示份额值的数字。
预留	保证为该资源池分配的 CPU 量。
可扩展预留	表示在接入控制期间是否考虑可扩展预留。选中 （默认）该复选框时，如果在该资源池中启动一台虚拟机，并且虚拟机的总预留大于该资源池的预留，则该资源池可以使用父级或祖先的资源。
限制	主机为该资源池提供的 CPU 量的上限。默认设置为 [无限 (Unlimited)]。要指定限制量，请取消选中 [无限 (Unlimited)]，然后键入数值。
内存资源	
份额	资源池拥有的、相对于父级总数的内存份额值。同级资源池根据由其预留和限制限定的相对份额值共享资源。您可以选择 **[低 (Low)]**、**[正常 (Normal)]** 或 **[高 (High)]**，也可以选择 **[自定义 (Custom)]** 来指定表示份额值的数字。
预留	保证为该资源池分配的内存量。
可扩展预留	表示在接入控制期间是否考虑可扩展预留。选中 （默认）该复选框时，如果在该资源池中启动一台虚拟机，并且虚拟机的总预留大于该资源池的预留，则该资源池可以使用父级或祖先的资源。
限制	该资源池的内存分配的上限。默认设置为 **[无限 (Unlimited)]**。要指定其他限制量，请取消选中 **[无限 (Unlimited)]** 复选框。

3 完成所有选择操作后，请单击 **[确定 (OK)]**。
此时 VirtualCenter 将创建该资源池，并将其显示在清单面板中。

如有任何选定值因可用 CPU 和内存总量限制而无效，将显示黄色三角形。

了解可扩展预留

启动虚拟机或创建资源池时，系统将检查 CPU 和内存预留是否可用于执行该操作。

如果未选中 **[可扩展预留 (Expandable Reservation)]**，则系统仅考虑选定资源池中的可用资源。

如果已选中 （默认）**[可扩展预留 (Expandable Reservation)]**，系统将考虑选定资源池及其直接父资源池中的可用资源。如果父资源池也选中了 **[可扩展预留 (Expandable Reservation)]** 选项，它还可以从其父资源池中借用资源。只要选中了 **[可扩展预留 (Expandable Reservation)]** 选项，就会以递归方式向当前资源池的祖先借用资源。将该选项保持选中状态可提供更高的灵活性，但同时提供的保护将会减少。子资源池所有者预留的资源可能大于您的预期值。

注意 只有当您相信子资源池的管理员不会过多预留资源时，才应将此选项保持选中状态。

可扩展预留示例 假定某个管理员负责管理资源池 P，并定义了两个子资源池 S1 和 S2，分别用于两个不同的用户 （或组）。

该管理员知道用户将要启动具有预留的虚拟机，但不知道每个用户需要预留多少资源。为 S1 和 S2 设置可扩展预留可使管理员更加灵活地共享和继承资源池 P 的公用预留。

如果不使用可扩展预留，管理员需要向 S1 和 S2 明确分配具体的资源量。这种具体的分配可能欠缺灵活性，特别是在较深的资源池层次结构中，并且可能使资源池层次结构中的预留设置操作复杂化。

可扩展预留会造成资源池缺少严格的隔离，也就是说 S1 可使用资源池 P 的全部预留启动，致使 S2 无法直接使用任何 CPU 或内存资源。

查看资源池信息

在 VI Client 中选择某个资源池时，**[摘要 (Summary)]** 选项卡将显示该资源池的相关信息。下一节将列出有关 “资源池摘要选项卡”（第 48 页）和 “资源池资源分配选项卡”（第 49 页）的信息。

注意 所有其他选项卡将在联机帮助中详细介绍。

资源池摘要选项卡

资源池的 **[摘要 (Summary)]** 选项卡显示有关资源池的高级别统计信息。

表 3-1. 资源池摘要选项卡的各栏

栏	描述
常规	[常规 (General)] 面板显示资源池的统计信息。 ■ **[虚拟机数目 (Number of Virtual Machines)]** - 该资源池中的虚拟机数目。不包括子资源池中的虚拟机数目。 ■ **[正在运行的虚拟机数 (Number of Running Virtual Machines)]** - 该资源池中正在运行的虚拟机数。不包括子资源池中正在运行的虚拟机数。 ■ **[子资源池数 (Number of Child Resource Pools)]** - 不包括层次机构中的所有资源池，而是仅包括直接子级。
CPU	显示为该资源池指定的 CPU **[份额 (Shares)]**、**[预留 (Reservation)]**、**[预留类型 (Reservation Type)]** 和 **[限制 (Limit)]**。同时还显示当前未预留的 CPU 量。
命令	允许您调用常用命令。 ■ **[新建虚拟机 (New Virtual Machine)]** - 启动新建虚拟机向导，在该资源池中创建一台新虚拟机。 ■ **[新建资源池 (New Resource Pool)]** - 显示 [创建资源池 (New Resource Pool)] 对话框，该对话框允许您为选定资源池创建一个子资源池。 ■ **[编辑设置 (Edit Settings)]** - 允许您更改选定资源池的 CPU 和内存属性。

表 3-1. 资源池摘要选项卡的各栏 （续）

栏	描述
资源	显示选定资源池内的虚拟机的运行时 **[CPU 使用情况 (CPU Usage)]** 和 **[内存使用情况 (Memory Usage)]**。
内存	显示为该资源池指定的 **[份额 (Shares)]**、**[预留 (Reservation)]**、**[预留类型 (Reservation Type)]** 和 **[限制 (Limit)]**。同时还显示当前未预留的内存量。

资源池资源分配选项卡

资源池的 **[分配资源 (Resource Allocation)]** 选项卡显示有关当前为资源池预留的资源及资源池可用的资源的详细信息，并列出资源的用户，如表 3-2 和表 3-3 中所述。

图 3-4. 资源池资源分配选项卡

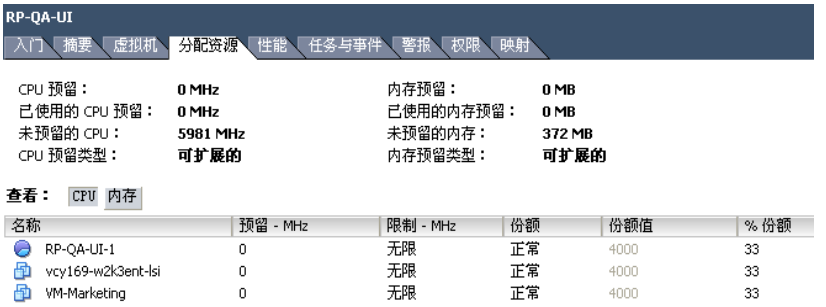


图 3-4 的顶部指定了表 3-2 中关于资源池本身的信息。

表 3-2. 资源分配选项卡字段

字段	描述
CPU 预留 / 内存预留	该资源池的预留中指定的 CPU 或内存量。可在创建资源池的过程中或随后通过编辑资源池指定预留。
已使用的 CPU 预留 / 已使用的内存预留	已使用的 CPU 或内存预留。预留将由正在运行的虚拟机或具有预留的子资源池使用。
未预留的 CPU/ 未预留的内存	当前未预留并可供虚拟机和资源池预留的 CPU 或内存。 **注意**：尝试确定是否可创建一定大小的子资源池或者是否可启动一台具有一定预留的虚拟机时请查看该数值。
CPU 预留类型 / 内存预留类型	**[可扩展的 (Expandable)]** 或 **[固定的 (Fixed)]**。请参见 “了解可扩展预留”（第 47 页）。

在图 3-4 中，特定于资源池的信息下方是该资源池的虚拟机和子资源池列表。该列表不包含分配给该资源池的子资源池的虚拟机。

单击 **[CPU]** 或 **[内存 (Memory)]** 选项卡显示表 3-3 中所述的信息。

表 3-3. 资源分配 CPU 和内存字段

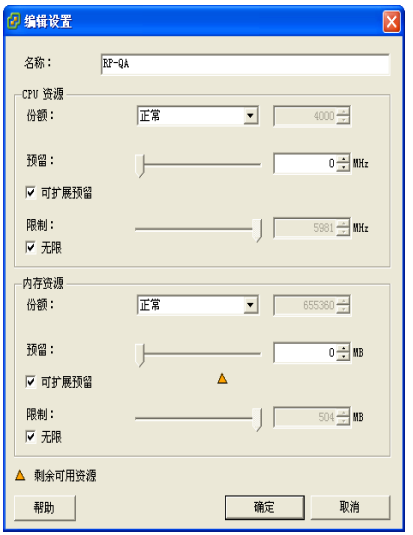
字段	描述
名称	资源池或虚拟机的名称。
预留	该虚拟机或资源池的指定预留。默认设置为 0，即系统不为该资源池预留任何资源。
限制	该虚拟机或资源池的指定限制。默认设置为 **[无限 (Unlimited)]**，即系统向该虚拟机分配尽可能多的资源。
份额	该虚拟机或资源池的指定份额。如果选择了某个默认设置，则为 **[高 (High)]**、**[正常 (Normal)]**、**[低 (Low)]** 三者之一。如果选择自定义设置，则为 **[自定义 (Custom)]**。
份额值	分配给该虚拟机或资源池的份额数。该数值取决于份额设置 （**[高 (High)]**、**[正常 (Normal)]**、**[低 (Low)]** 或 **[自定义 (Custom)]**）。 请参见 “份额”（第 20 页）。
% 份额	该资源池或虚拟机的份额值除以分配给父资源池中所有子资源池的份额总数。该值与父资源池的本地份额分配无关。
类型	预留类型。可以是 **[固定的 (Fixed)]** 或 **[可扩展的 (Expandable)]**。请参见 “了解可扩展预留”（第 27 页）。

更改资源池属性

对资源池进行更改

1 在 VI Client 清单面板中选择资源池。

2 在 **[摘要 (Summary)]** 选项卡 [命令 (Command)] 面板中，选择 **[编辑设置 (Edit Settings)]**。



3 在 [编辑设置 (Edit Settings)] 对话框中，您可以更改选定资源池的全部属性。各选项在 “创建资源池”（第 45 页）中进行了论述。

监视资源池性能

监视资源池的性能有助于了解资源池分配的有效性。

监视资源池的性能

1 在清单面板中选择资源池。

2 单击 **[性能 (Performance)]** 选项卡。

此时即可看到有关资源池性能的信息。单击 **[更改图表选项 (Change Chart Options)]** 定制性能图表。有关性能图表及其配置方法的论述，请参见联机帮助。

将虚拟机添加到资源池

创建新的虚拟机时，可以通过新建虚拟机向导在创建过程中将该虚拟机添加到资源池。您也可以将现有虚拟机添加到资源池。本节将论述这两种任务。

创建一台虚拟机并将其添加到资源池中

1 选择一台主机，然后选择 **[文件 (File)] > [新建 (New)] > [虚拟机 (Virtual Machine)]**（或按 Ctrl+n）。

2 提供虚拟机的信息，同时在出现向导提示时选择一个资源池作为位置。

向导将把虚拟机添加到您选择的资源池中。

将现有虚拟机添加到资源池

1 从清单中的任意位置选择虚拟机。

该虚拟机可以与独立主机、群集或另一个资源池关联。

2 将该虚拟机 （或这些虚拟机）拖至所需的资源池对象。

将虚拟机移至新的资源池时：

- 该虚拟机的预留和限制不会发生变化。
- 如果该虚拟机的份额为高、中等或低，**[% 份额 (%Shares)]** 会进行调整以反映新资源池中使用的份额总数。
- 如果已为该虚拟机指定了自定义份额，该份额值将保持不变。

  注意 由于份额分配与资源池有关，因此当您将虚拟机移入资源池时可能必须手动更改虚拟机的份额，以使虚拟机的份额与新资源池中的相对值保持一致。如果虚拟机所占总份额的比例过大 （或过小），将出现警告。

- **[分配资源 (Resource Allocation)]** 选项卡中显示的有关资源池的预留和未预留 CPU 和内存资源的信息将发生变化，以反映与该虚拟机关联的预留 （如果有）。

  注意 如果虚拟机已关闭或挂起，可以移动该虚拟机，但资源池的可用资源总量 （例如预留和未预留的 CPU 和内存资源）不受影响。

如果某台虚拟机已启动，且目标资源池的 CPU 或内存不足以保证该虚拟机的预留，移动操作将会失败，因为接入控制不允许该操作。此时将显示一个错误对话框，解释这种情况。该错误对话框会将可用资源与所需资源进行比较，以便您考虑可否通过调整来解决此问题。请参见 “资源池接入控制”（第 45 页）。

从资源池中移除虚拟机

可通过多种方法将虚拟机从资源池中移除，具体取决于您对虚拟机的打算。

将虚拟机移至其他资源池。 请参见 “将现有虚拟机添加到资源池”（第 52 页）。如果仅需移动虚拟机，则无需将其关闭。

从资源池中移除虚拟机时，与资源池相关联的份额总数将减少，从而使每个剩余的份额代表更多资源。例如，假定您有一个有权使用 6 GHz 的资源池，其中包含三台份额设置为 **[正常 (Normal)]** 的虚拟机。假定虚拟机受 CPU 限制，每台各获得 2 GHz 的相等分配额。如果将其中一台虚拟机移至其他资源池，剩余的两台虚拟机将各获得 3 GHz 的相等分配额。

从清单中移除虚拟机或将其从磁盘中删除。 右键单击虚拟机 （或按 [Delete]）。

您需要关闭虚拟机才能将其完全移除。请参见 *《Virtual Infrastructure 用户指南》*。

资源池和群集

将带有现存资源池层次结构的主机添加到群集时，随后将发生的情况取决于群集。您有两种选择：

- “已启用 DRS 的群集”（第 53 页）
- “未启用 DRS 的群集”（第 54 页）

已启用 DRS 的群集

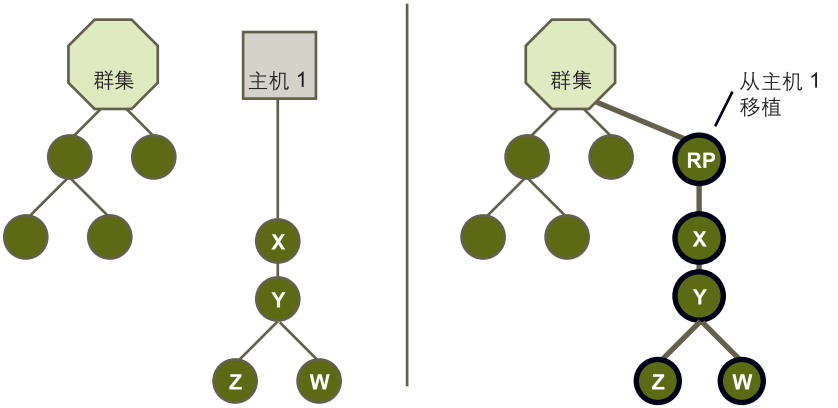
如果群集已启用 DRS，那么当您将一台或多台主机移入群集时，向导将允许您选择对主机的资源池采取何种操作。

将此主机的虚拟机置于群集的根资源中。 折叠主机的资源池层次结构，使所有虚拟机成为群集的直接子级。该行为与 “未启用 DRS 的群集”（第 54 页）中所表现的行为相同。

注意 由于原始主机层次结构中的份额与主机上的资源池有关，因此您可能必须手动调整与单个虚拟机相关联的份额值。

为此主机的虚拟机和资源池创建一个新的资源池。 创建一个与该主机的根资源池对应的资源池。默认情况下，该资源池命名为 **[从 <host_name> 移植 (Grafted from <host_name>)]**，但您也可选择其他名称。选择使用 “移植”一词是因为主机树的分支已添加到群集树的分支。

图 3-5. 资源池层次结构移植到群集



在图 3-5 中的示例中，群集 **[群集 (Cluster)]** 和主机 **[主机 1 (Host1)]** 各有一个资源池层次结构。将主机添加到群集后，主机中不可见的顶层资源池将移植到群集的资源池层次结构，并默认命名为 **[从主机 1 移植 (grafted from Host1)]**。

注意 份额的分配情况与主机移入群集前相同。百分比将根据需要进行调整。

资源池层次结构目前已完全独立于主机。如果随后将主机从群集中移除，群集会保留资源池层次结构，而主机将丢失资源池层次机构 （虽然虚拟机将与主机一同移除）。请参见 〝从群集移除主机〞（第 95 页）。

注意 主机必须处于维护模式才能将其从群集中移除。请参见 〝主机维护模式和待机模式〞（第 67 页）。

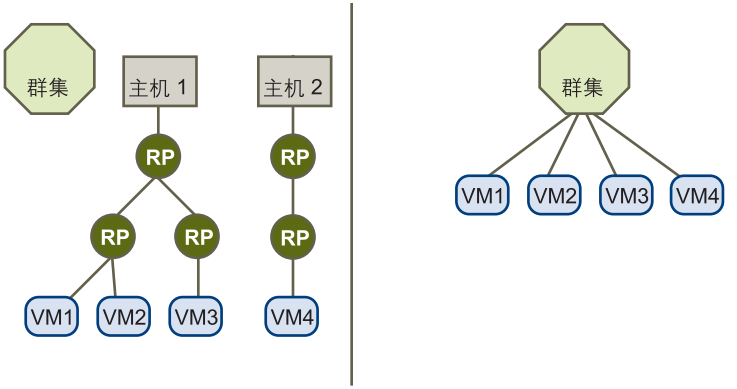
未启用 DRS 的群集

如果群集仅启用了 HA （或者 HA 和 DRS 均未启用），那么当您将一台或多台主机移入群集时，群集将占有资源的所有权。主机和虚拟机将与群集产生关联。资源池层次结构将展平。

注意 在非 DRS 群集中，不会根据份额进行群集范围的资源管理。虚拟机份额与各主机保持相关。

在图 3-6 中，主机 H1 和 H2 各有一个资源池和虚拟机层次结构。将两台主机添加到群集 C 后，资源池层次结构将展平，所有虚拟机都将成为群集的直接子级。

图 3-6. 已展平的资源池层次结构



4 了解群集

本章提供群集以及 VMware Distributed Resource Scheduler (DRS) 和 High Availability (HA) 功能的概念简介。

本章将讨论以下主题：

注意 所述的所有任务均假定您有执行这些任务的权限。有关权限及如何设置权限的信息，请参见联机帮助。

群集简介

群集是一组具有共享资源和共享管理界面的 ESX Server 主机和相关虚拟机。将主机添加到群集时，主机的资源将成为群集资源的一部分。创建群集时，可以选择为 DRS 和/或 HA 启用群集。

有关群集中虚拟机以及如何配置这些虚拟机的信息，请参见 "群集先决条件"（第 83 页）。

注意 可以在没有特殊许可证的情况下创建群集，但必须要有许可证才能为 DRS 或 HA 启用群集。

VMware DRS

DRS 功能可改善所有主机和资源池之间的资源分配。 DRS 会收集群集中所有主机和虚拟机的资源使用信息，并为虚拟机放置位置和主机电源状况生成建议。这些建议可以自动应用。根据配置的 DRS 自动控制级别， DRS 将显示或应用下列类型的建议：

- **[初始放置位置 (Initial placement)]** - 当您在群集中首次启动某个虚拟机时， DRS 将在适当的主机上放置该虚拟机或提出建议。请参见 “初始放置位置”（第 60 页）。
- **[迁移 (Migration)]** - 在运行时，DRS 会通过执行虚拟机的迁移或提供虚拟机迁移建议，尝试解决规则冲突并提高群集中的资源利用率。请参见 “负载平衡和虚拟机迁移”（第 64 页）。
- **[电源管理 (Power management)]** - 启用分布式电源管理功能时， DRS 会将群集和主机级容量与运行群集的虚拟机的需求 （包括近期历史需求）进行比较。如果找到足够的额外容量，它会建议将主机置于待机电源模式，或者如果需要容量，则建议启动主机。根据提出的主机电源状况建议，可能需要将虚拟机迁移到主机并从主机迁移虚拟机。请参见 “分布式电源管理”（第 66 页）。

VMware HA

为主机故障启用 HA 监控的群集。如果主机不可用，将立即在其他主机上重新启动该主机上的所有虚拟机。

为 HA 启用群集时，可以指定允许的主机故障数。如果将主机故障数设置为 **[1]**， HA 将使整个群集具备足够的容量来处理一台主机的故障，从而使该主机上所有正在运行的虚拟机都能在其余主机上重新启动。默认情况下，如果启动虚拟机会与故障切换所需的容量发生冲突，则无法启动此虚拟机 （严格的接入控制）。请参见 “了解 VMware HA”（第 69 页）。

图 4-1. VMware HA

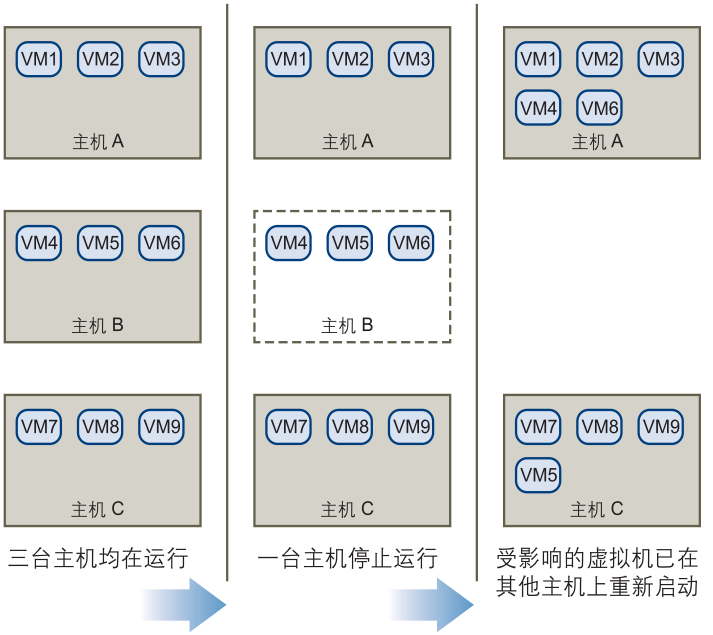


图 4-1 中，三台主机各有三个虚拟机，并且为相应 HA 群集配置了一台主机的故障切换。当主机 B 不可用时， HA 会将虚拟机从主机 B 迁移到主机 A 和主机 C。

群集和 VirtualCenter 故障

VirtualCenter Server 在每台主机上放置一个代理。如果 VirtualCenter Server 不可用，则 HA 和 DRS 功能会发生如下变化：

- **HA** - 即使 VirtualCenter Server 不可用， HA 群集仍会继续运作，并且仍然可以在故障切换时在其他主机上重新启动虚拟机。但是，有关虚拟机特定群集属性 （如优先级和隔离响应）的信息则取决于 VirtualCenter Server 发生故障前该群集的状况。
- **DRS** - DRS 群集中的主机将使用可用资源继续运行。但是，不会有资源优化建议。

VirtualCenter Server 不可用时，如果必须使用连接 ESX Server 主机的 VI Client 更改主机或虚拟机，这些更改会生效。当 VirtualCenter 再次可用时，您可能会发现群集由于不再满足群集要求而变为红色或黄色。

了解 VMware DRS

为 DRS 启用群集时， VirtualCenter 将持续监视群集中所有主机和虚拟机的 CPU 和内存资源分布。 DRS 会将这些衡量指标与理想状态下群集中资源池和虚拟机的属性和当前需求应给出的资源利用率进行比较，并提出相应的迁移建议。此外，启用分布式电源管理功能时， DRS 将监视群集层以及主机层可供运行虚拟机的容量，并会根据容量是多余还是不足，提出有关关闭或启动主机的建议。

将主机添加到 DRS 群集时，该主机的资源会与群集相关联。系统将提示是要将任何现有的虚拟机和资源池与群集的根资源池相关联，还是移植资源池层次结构。请参见 "资源池和群集"（第 53 页）。然后 VirtualCenter 就可以进行虚拟机的初始放置、为了负载平衡或规则进行虚拟机迁移，以及执行分布式电源管理 （如果启用了的话）。

初始放置位置

尝试在已启用 DRS 的群集中启动一个虚拟机或一组虚拟机时， VirtualCenter 会检查群集中是否有足够的资源来支持虚拟机。然后，对于每个虚拟机， VirtualCenter 将标识可在其上运行的主机并执行以下操作之一：

自动放置。 如果所有与放置相关的操作 （启动或迁移虚拟机，或启动主机）都处于自动模式，则不必对用户提出建议而自动执行这些步骤。

提出初始放置建议。 如果任何与放置相关的操作的自动化级别都处于手动模式，则将提出初始放置建议。根据启动一个还是多个虚拟机，用户在非自动放置情形收到的初始放置位置建议会有所不同。

注意 对于独立主机或非 DRS 群集中的虚拟机，不提出任何初始放置位置建议。这些虚拟机将会在启动时被置于当前所驻留的主机上。

单一虚拟机启动

启动一个虚拟机时，有两种类型的初始放置位置建议：

- 启动一个虚拟机，不需要任何先决条件步骤。

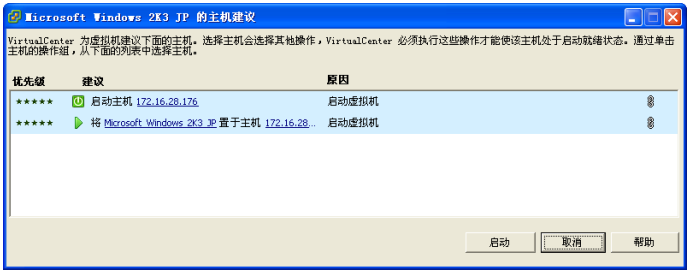
  用户将拥有虚拟机互斥初始放置位置建议列表 (图 4-2)。只能选择一种建议。

图 4-2. 单一虚拟机，无先决条件

- 启动一个虚拟机，但需要先决条件操作。

  这些操作包括在待机模式下启动主机或在主机间迁移其他虚拟机。在这种情况下，仅提供一个建议，该建议具有多行，显示每个先决条件操作。用户可以接受整个建议，也可以取消启动虚拟机 (图 4-3)。

图 4-3. 单一虚拟机，有先决条件



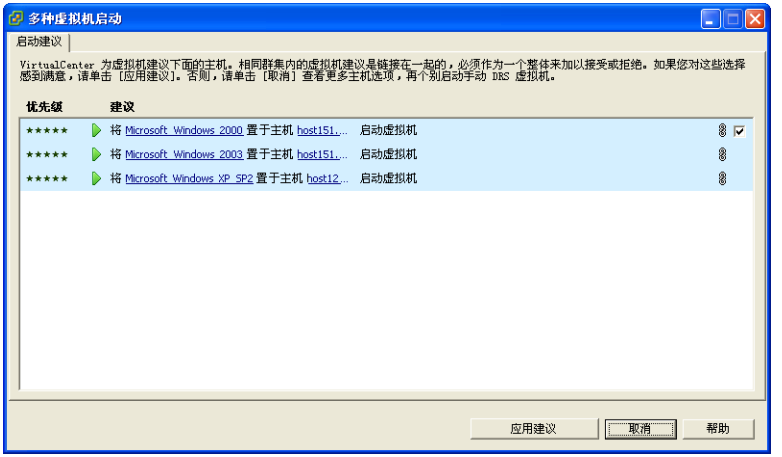
组启动

可以尝试同时启动多个虚拟机 （*组启动*）。选择组启动尝试的虚拟机不需要在同一 DRS 群集中。可以在群集间选择虚拟机，但它们必须属于同一数据中心。也可以包括位于非 DRS 群集或独立主机上的虚拟机，但这些虚拟机自动启动并且不包括在任何初始放置位置建议中。

每个群集均提供组启动尝试的初始放置位置建议。如果组启动尝试的所有与放置相关的操作都处于自动模式，虚拟机将启动，不提出任何初始放置位置建议。如果任何虚拟机与放置相关的操作处于手动模式，则所有虚拟机 （包括处于自动模式的虚拟机）都将手动启动，并且包括在初始放置位置建议中。

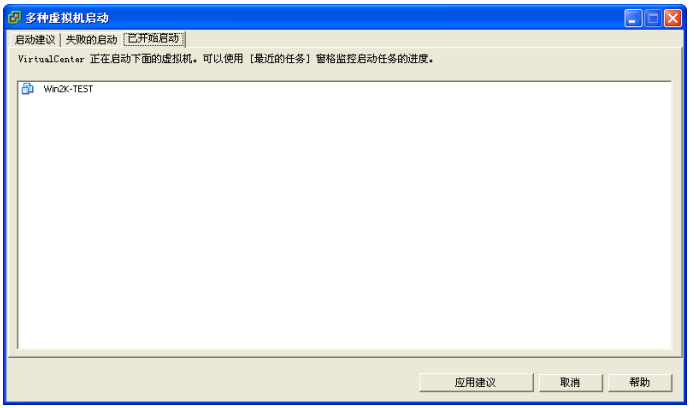
对于已启动的虚拟机所属的每个 DRS 群集，均会有一个包含所有所需先决条件的建议（或没有建议）。所有特定于此类群集的建议都显示在 **[启动建议 (Power On Recommendations)]** 选项卡下 (图 4-4)。

图 4-4. 组启动建议



如果进行了非自动组启动尝试，且包括了不受限于初始放置位置建议的虚拟机 （即独立主机或非 DRS 群集上的虚拟机）， VirtualCenter 会尝试自动启动这些虚拟机。如果这些虚拟机自动启动成功，则会在 **[已开始启动 (Started Power-Ons)]** 选项卡下列出。那些无法以此方式启动的虚拟机则在 **[失败的启动 (Failed Power-Ons)]** 选项卡下列出。请参见图 4-5。

图 4-5. 自动组启动

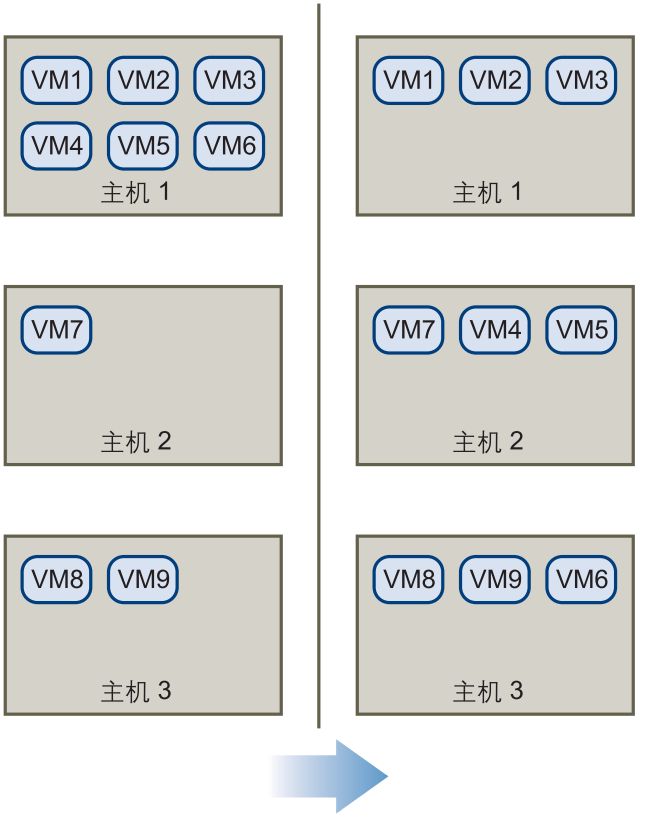


组启动示例：用户选择同一数据中心中的三个虚拟机进行组启动尝试。前两个虚拟机（VM1 和 VM2）在同一 DRS 群集 (Cluster1) 中，而第三个虚拟机 (VM3) 则在一台独立主机上。 VM1 处于自动模式，而 VM2 处于手动模式。在此方案中，用户将获得 Cluster1 的初始放置位置建议 （位于 **[启动建议 (Power On Recommendations)]** 选项卡下），其中包含启动 VM1 和 VM2 的操作。将尝试自动启动 VM3，如果成功，则会在 **[已开始启动 (Started Power Ons)]** 选项卡下列出 VM3。如果此尝试失败，则会在 **[失败的启动 (Failed Power Ons)]** 选项卡下列出 VM3。

负载平衡和虚拟机迁移

为 DRS 启用的群集可能会变得不平衡。例如，请参见图 4-6。该图左侧的三台主机不平衡。假定主机 1、主机 2 和主机 3 均有相同的容量，且所有虚拟机的配置和负载均相同。但是，由于主机 1 有六个虚拟机，其资源被过度利用，而主机 2 和主机 3 资源充足。DRS 会将虚拟机从主机 1 迁移到主机 2 和主机 3 （或提出迁移建议）。该图右侧显示了适当平衡负载之后所得到的主机配置。

图 4-6. VMware DRS



当群集不平衡时，DRS 将根据默认自动化级别，提出建议或迁移虚拟机：

- 如果涉及的群集或任何虚拟机为*手动*或*半*自动，则 VirtualCenter 不执行自动操作来平衡资源。[摘要 (Summary)] 页面会指示有迁移建议，并且 [DRS 建议 (DRS Recommendations)] 页面会显示最有效地利用群集中资源的更改建议。

- 如果涉及的群集或虚拟机均为全自动，则 VirtualCenter 将根据需要在主机间迁移正在运行的虚拟机，以确保有效地利用群集资源。

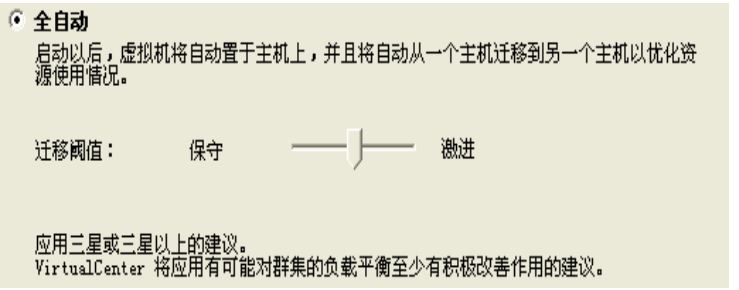
注意 即使是在自动迁移设置中，用户也可以显式迁移单个虚拟机，但 VirtualCenter 可能会将这些虚拟机迁移到其他主机，以优化群集资源。

默认情况下，自动化级别是为整个群集指定的。也可以为单个虚拟机指定自定义自动化级别。

迁移阈值

迁移阈值可让您指定群集处于全自动模式时应用的建议。请参见图 4-7。可以移动滑块，以使用从 “保守”（进行最小数目的迁移）到 “激进”（进行最大数目的迁移）的五种级别之一。这五种迁移级别将根据其所分配的星级应用建议。

图 4-7. 迁移阈值选项



根据所带来的群集负载平衡改善情况，对迁移建议进行了星级分配 - 从五星建议 （强制性建议）到一星建议 （仅带来轻微改善的建议）。每次将滑块向右移动一个级别将会允许应用下一较低星级的建议。保守设置将仅应用五星建议，向右的下一级别则将应用四星建议以及较高级别的建议，然后依次类推，直至激进级别，该级别将应用一星建议和较高级别的建议 （即，应用所有建议）。

迁移建议

如果创建带有默认手动或半自动模式的群集，则 VirtualCenter 将在 [DRS 建议 (DRS Recommendations)] 页面上显示迁移建议。系统将提供足够的建议，以强制实施规则和平衡群集的资源。每条建议均包含要移动的虚拟机、当前 （源）主机和目标主机，以及提出建议的原因。原因可能为以下之一：

- 平衡平均 CPU 负载。
- 平衡平均内存负载。
- 满足资源池预留量。

- 满足关联性 （或反关联性）规则。请参见 “使用 DRS 关联性规则”（第 99 页）。
- 主机正在进入维护模式。请参见 “主机维护模式和待机模式”（第 67 页）。

分布式电源管理

启用此实验性功能时， DRS 将根据比较群集级容量和需求的结果，提出建议来减少其功耗。如果认为容量不够， DRS 将建议启动主机，并 （使用 VMware VMotion）将虚拟机迁移到这些主机。相反，找到额外容量时， DRS 会建议将某些主机置于*待机*模式，并将这些主机上正在运行的任何虚拟机撤出至其他主机。请参见 “待机模式”（第 68 页）。是否自动执行这些主机电源状况和迁移建议，取决于所选的分布式电源管理功能的自动化级别。

在为 DRS 群集启用分布式电源管理前，必须先确保 ESX Server 主机有合适的硬件支持和配置。特别是， VMkernel 网络使用的网卡必须具备 LAN 唤醒 (Wake-on-LAN, WOL) 功能，该功能用于将 ESX Server 主机从关闭状况唤醒。应该在每台要部署分布式电源管理的 ESX Server 3.5 （或 ESX Server 3i 版本 3.5）主机上测试唤醒功能操作。为此，请确保群集中至少启动有另外一台 ESX Server 主机 （用于发送唤醒数据包），并将要测试的主机显式置于待机状况，而一旦该主机进入待机状况后，请确保成功显式请求重新启动待机主机。否则， VMware 会建议您不要配置该主机由分布式电源管理来管理电源。

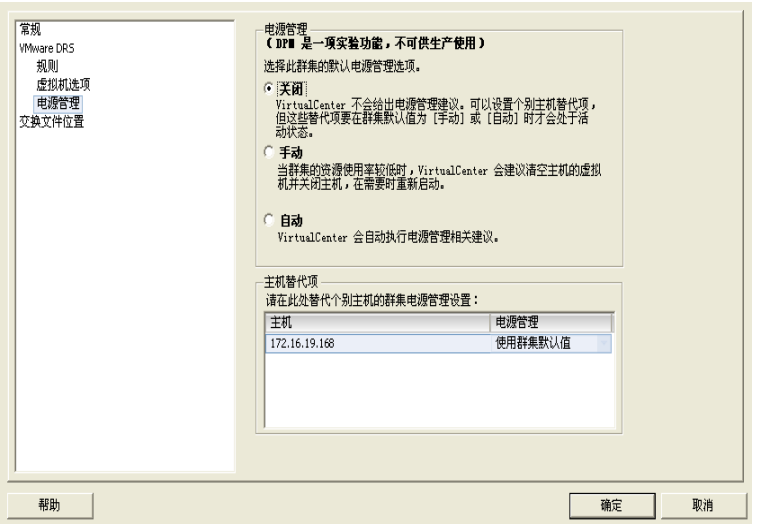
小心 实施分布式电源管理前，请测试主机的 LAN 唤醒功能。如果 WOL 功能失败，电源管理功能可能会关闭主机，而稍后并不能重新启动这些主机。

DRS 群集的默认电源管理自动化级别是从群集的 [设置 (Settings)] 对话框的 **[电源管理 (Power Management)]** 选项卡中选择的。请参见图 4-8。只要此设置不是 “关闭”，就会启用该功能。可用选项包括：

- **[关闭 (Off)]** - 禁用该功能且不提供建议。
- **[手动 (Manual)]** - 提供主机电源操作和相关虚拟机迁移建议，但不自动执行。
- **[自动 (Automatic)]** - 如果可以自动执行所有相关虚拟机迁移，则将自动执行主机电源操作。

除了这些群集级设置之外，还可以为单个主机设置替代项，以便这些主机与群集的自动化级别不同。只有在为群集启用该功能 （未设置为 “关闭”）时，才会应用这些替代项。

图 4-8. 电源管理自动化级别和主机替代项



注意 电源管理自动化级别与之前所介绍的 DRS 自动化级别 （用于负载平衡）不同。此外，由分布式电源管理生成的建议会被分配以星级来显示其相对重要性，但却并不受 DRS 迁移阈值控制。

DRS 群集、资源池和 ESX Server

对于已启用 DRS 的群集，所有主机的资源将分配给该群集。

DRS 在内部使用每台主机资源池层次结构来实现群集范围的资源池层次结构。使用连接 VirtualCenter Server 的 VI Client 查看群集时，会看到由 DRS 实现的资源池层次结构。

使用连接 ESX Server 主机的 VI Client 查看单个主机时，会看到资源池的基础层次结构。由于 DRS 实现了其所能达到的最佳平衡资源池层次结构，因此请不要修改单个 ESX Server 主机上可见的层次结构。如果进行了更改， DRS 将会立即撤消更改。

主机维护模式和待机模式

主机的维护模式和待机模式有一些相似之处。尤其是，两者都禁止虚拟机运行。但是，这两种模式分别有不同的用途。当需要维护主机时 （例如，安装更多内存或升级其上运行的 ESX Server 版本），您可以将主机置于维护模式，而后该主机会一直处于维护模式，直到您使其退出维护模式。相比之下， DRS 会自动使主机进入和退出待机模式，以优化电源使用情况。

维护模式

独立主机和群集中的主机都支持维护模式。维护模式将限制主机上的虚拟机操作，使用户可以关闭正在运行的虚拟机，以便为关闭主机做好准备。仅 ESX Server 3.0 及更高版本才支持独立主机的维护模式。

主机仅会因用户要求而进入或离开维护模式。如果主机位于群集中，则在其进入维护模式时，用户可以选择撤出已关闭的虚拟机。如果选择了此选项，除非群集中没有可用于虚拟机的兼容主机，否则会将每台已关闭的虚拟机迁移到其他主机。在维护模式下，主机不允许您部署虚拟机，也不允许您启动虚拟机。需要将正在进入维护模式的主机上正在运行的虚拟机迁移到其他主机或将其关闭 （可以手动操作或由 DRS 自动操作）。

注意 如果 DRS 无法提供虚拟机的迁移建议，则会生成一个事件 （请检查该虚拟机的 [任务和事件 (Tasks & Events)] 选项卡）。必须手动迁移或关闭该虚拟机，主机才能进入维护模式。

当主机上不再有正在运行的虚拟机时，该主机的图标将变成包含 **[维护中 (under maintenance)]**，并且该主机的 [摘要 (Summary)] 面板会指示新的状况。

虚拟机的默认自动化模式将决定该虚拟机在其上运行的主机 （在 DRS 群集中）进入维护模式时的行为：

- 自动迁移任何全自动虚拟机。

  注意 如果没有合适的可用主机，DRS 将在 **[任务和事件 (Tasks & Events)]** 选项卡中显示相关信息。

- 对于半自动的或手动的虚拟机，则将生成并显示有关其他用户操作的建议。

待机模式

将主机置于待机模式时，会关闭主机。通常，由分布式电源管理功能将主机置于待机模式。还可以手动将主机置于待机模式；但是， DRS 可能会在其下次运行时撤消 （或建议撤消）更改。要强制主机保持关闭状态，请将其置于维护模式并关闭。还可以禁用主机上的分布式电源管理 （或将其设置为手动模式），以防止自动启动该主机。

如果分布式电源管理功能决定需要将该主机从待机模式唤醒 （即重新启动），可以使用 LAN 唤醒技术来实现。

要提供此功能，必须确保执行了以下步骤：

- VMkernel 网络堆栈连接的网卡 （选为 VMotion 网卡）必须与 WOL 兼容。要显示主机上每个网卡的 WOL 兼容状态，请在 VI Client 的清单面板中选择主机，再选择 **[配置 (Configuration)]** 选项卡，然后单击 **[网络适配器 (Network Adapters)]**。

- 每个群集的 VMotion 网络都必须位于单个 IP 子网上。

注意 从不选择将没有 WOL 兼容网卡的主机置于待机模式。

了解 VMware HA

HA 群集功能允许运行在 ESX Server 系统上的虚拟机可自动从主机故障中恢复。当主机不可用时，将自动在系统内其他主机上重新启动所有相关的虚拟机。本节首先考虑 VMware HA 群集和传统群集解决方案的差异，然后介绍 HA 群集概念。

传统解决方案和 HA 故障切换解决方案

VMware HA 和传统群集解决方案都支持自动从主机故障中恢复。由于以下方面的差异，两者互为补充：

- 硬件和软件要求
- 恢复时间
- 应用程序依赖度。

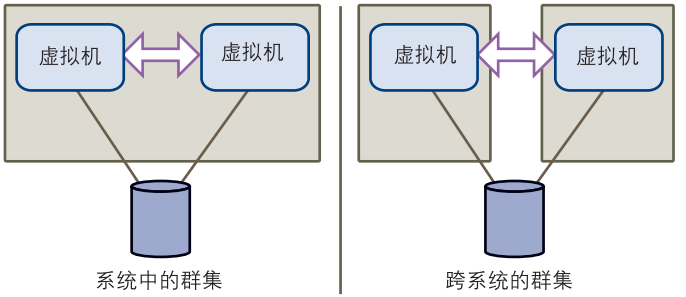
传统群集解决方案

传统群集解决方案 （如 Microsoft 群集服务 (Microsoft Cluster Service, MSCS) 或 Veritas 群集服务）可以在主机或虚拟机发生故障时，以最少的停机时间立即恢复应用程序。要做到这一点，必须按以下说明设置 IT 基础架构：

- 每台计算机 （或虚拟机）必须有一个镜像虚拟机 （可能在其他主机上）。
- 使用群集软件设置计算机 （或虚拟机及其主机）以相互镜像。通常，主虚拟机会向镜像发送检测信号。万一发生故障，镜像会无缝取代对方。

图 4-9 显示用于设置虚拟机的传统群集的不同选项。

图 4-9. VMware 群集设置



设置和维护此类群集解决方案需占用大量资源。每次添加新的虚拟机，都需要对应的故障切换虚拟机，而且还可能需要额外的主机。必须设置、连接和配置所有的新计算机，并更新群集应用程序的配置。

传统解决方案可保证快速恢复，但会占用大量资源，而且比较费时。有关不同群集类型以及如何配置这些群集的详细信息，请参见 VMware 文档 *《Microsoft 群集服务的设置》*。

VMware HA 解决方案

在 VMware HA 解决方案中，会将一组 ESX Server 主机与一个带有一个共享资源池的群集结合在一起。 VirtualCenter 会监视该群集中的所有主机。如果其中一个主机发生故障， VirtualCenter 将立即通过在其他主机上重新启动每个关联的虚拟机来做出响应。

使用 VMware HA 有许多好处：

- **最少的设置** - 将使用新建群集向导进行初始设置。可以使用 VI Client 添加主机和新的虚拟机。群集中的所有虚拟机无需另外配置即可获得故障切换支持。
- **减少了硬件成本和设置** - 在传统群集解决方案中，必须正确连接和配置重复的硬件和软件。虚拟机可充当应用程序的移动容器，可随意移动。可以避免在多台计算机上进行重复的配置。使用 VMware HA 时，必须拥有足够的资源来故障切换要保证的主机数。但是， VirtualCenter Server 会负责资源管理和群集配置。
- **提高了应用程序的可用性** - 虚拟机内部运行的任何应用程序的可用性变得更高。由于虚拟机可以从硬件故障中恢复，因此提高了设置在引导周期内启动的所有应用程序的可用性，而且没有额外的成本，即使该应用程序本身不是群集应用程序也一样。

使用 HA 群集的可能限制包括丢失运行时间状况，以及与带热待机的传统群集环境相比，应用程序的停机时间更长。如果这些限制造成问题，请考虑结合使用这两种方法。

VMware HA 功能

已启用 HA 的群集：

- 支持使用 VI Client 进行的易用配置。
- 在硬件发生故障时，为故障切换容量范围内所有正在运行的虚拟机提供故障切换（请参见 “故障切换容量”（第 71 页））。
- 可与传统应用程序级故障切换结合使用，并增强其功能。
- 与 DRS 完全集成。如果主机发生了故障，并且在其他主机上重新启动了虚拟机，则 DRS 会提出迁移建议或迁移虚拟以平衡资源分配。如果迁移的源主机和 / 或目标主机发生故障，则 HA 会帮助从该故障中恢复。

故障切换容量

为 HA 启用群集时，新建群集向导将提示您输入要故障切换容量的主机数。此数目显示为 VI Client 中的 **[配置的故障切换容量 (Configured Failover Capacity)]**。HA 使用此数目来确定群集内是否有足够资源来启动虚拟机。

只需指定要故障切换容量的主机数。HA 会使用保守估计来计算故障切换这些主机的虚拟机所需的资源，并且在不能再保证故障切换容量时，禁止启动虚拟机。

注意 即使虚拟机违反可用性限制，也可以允许群集启动这些虚拟机。如果这样做，群集会变成红色，表示可能不再保证故障切换。请参见 “有效群集、黄色群集和红色群集”（第 76 页）。

创建群集后，可以向其中添加主机。将主机添加到未启用 DRS 的 HA 群集时，会立即从该主机移除所有资源池，所有虚拟机会直接与群集相关联。

注意 如果群集还启用了 DRS，则可以选择保留资源池层次结构。请参见 “资源池和群集”（第 53 页）。

规划 HA 群集

规划 HA 群集时，请考虑以下几点：

- 每台主机要有一些内存和 CPU，以启动虚拟机。
- 必须保证每个虚拟机的 CPU 和内存预留要求。

一般而言，建议使用统一设置。HA 将针对最坏故障情况来进行规划。计算所需故障切换容量时，HA 将计算任何当前已启动虚拟机所需的最大内存和 CPU 预留量，并将此

称为*槽*。槽是足够当前已启动的所有虚拟机 （计算当前故障切换级别时，不考虑关闭或挂起的虚拟机）使用的 CPU 和内存资源量。

例如，如果有两个虚拟机，一个带有 1 GHz CPU 预留量和 1 GB 内存预留量，而另一个带有 2 GHz CPU 预留量和 512 MB 内存预留量，则槽定义为 2 GHz CPU 预留量和 1 GB 内存预留量。 HA 将根据主机的 CPU 和内存容量确定 ″适合″ 每台主机的槽数。然后， HA 会确定在保证群集仍至少有足够数量的槽来启动虚拟机的前提下，允许发生故障的主机数。此数字即为当前故障切换级别。

规划期间，请决定要保证故障切换的主机数。 HA 会尝试通过限制启动的虚拟机数（虚拟机将消耗资源），来至少为此数量的主机故障预留出资源。

如果不选中 **[允许启动虚拟机，即使这些虚拟机违反可用性限制 (Allow virtual machine to be started even if they violate availability constraints)]** 选项 （严格的接入控制），并且如果虚拟机可能导致当前故障切换级别低于所配置的故障切换级别，则 VMware HA 不允许启动虚拟机。如果虚拟机可能导致当前故障切换级别超过所配置的故障切换级别，则 VMware HA 也不允许以下操作：

- 将关闭的虚拟机恢复到已启动快照。
- 将正在运行的虚拟机迁移至群集。
- 重新配置正在运行的虚拟机，以增加其 CPU 或内存预留量。

如果启用 HA 时选择了 **[允许启动虚拟机，即使这些虚拟机违反可用性限制 (Allow virtual machine to be started even if they violate availability constraints)]** 选项，则可以启动比 HA 建议数量更多的虚拟机。由于配置了系统允许此操作，因此群集并不会变成红色。如果发生故障的主机数超过所配置的数目，则当前 （可用的）故障切换级别也会降至所配置的故障切换级别之下。例如，如果已配置群集允许一台主机发生故障，且已达容量 （当前故障切换级别等于所配置的故障切换级别），而这时两台主机发生故障，则群集会变成红色。

低于所配置故障切换级别的群集仍然可以在主机发生故障时执行虚拟机故障切换，此时它会使用虚拟机优先级来决定要首先启动的虚拟机。请参见 ″对虚拟机进行 HA 定制″（第 106 页）。

小心 不建议使用红色群集。如果使用，则不保证指定级别的故障切换。

VMware HA 和特殊情况

VMware HA 知道如何处理特殊情况来保留您的数据：

关闭主机。 如果关闭主机，则 HA 会在其他主机上重新启动正在该主机上运行的任何虚拟机。

使用 VMotion 迁移虚拟机。 如果正使用 VMotion 将虚拟机迁移到其他主机，而源主机或目标主机不可用，则根据所在的迁移阶段，可能会将该虚拟机留在故障 （关闭）状况中。 HA 会处理此故障，并在合适主机上启动该虚拟机：

- 如果源主机不可用，则 HA 会在目标主机上启动该虚拟机。
- 如果目标主机不可用，则 HA 会在源主机上启动该虚拟机。
- 如果两个主机都不可用，则 HA 会在群集中的第三台主机 （如果存在的话）上启动该虚拟机。

当前故障切换容量与所配置的故障切换容量不一致。 如果当前故障切换容量小于所配置的故障切换容量，则群集将变为红色。因为发生故障的主机数多于配置群集可处理的故障主机数，所以会发生这种情况。如果关闭了严格的接入控制，则即使启动的虚拟机数超过可容纳的数目，群集也不会变为红色。

容量不足时， HA 将先故障切换优先级较高的虚拟机，然后再尝试故障切换其他虚拟机。在此情况下，请给对恢复环境最为重要的虚拟机赋予高优先级。请参见 “对虚拟机进行 HA 定制”（第 106 页）。

主机网络隔离。 HA 群集中的主机可能会失去其控制台网络 （或 ESX Server 3i 中的 VMkernel 网络）连接。此主机会与群集中的其他主机隔离。群集中的其他主机会认为该主机发生了故障，并会尝试故障切换其上运行的虚拟机。如果某个虚拟机继续在隔离的主机上运行， VMFS 磁盘锁定将防止其在其他地方启动。如果虚拟机共享同一个网络适配器，则它们将无法访问网络。可能要在其他主机上启动该虚拟机。

默认情况下，如果出现主机网络隔离的意外情况，隔离主机上的虚拟机将关闭，以便可以在群集中其他未隔离的主机上重新启动虚拟机。可以更改群集的默认行为，使隔离主机上的虚拟机保持启动状态或将其关闭。也可以更改每个虚拟机的这种行为。请参见 “对虚拟机进行 HA 定制”（第 106 页）。

主要主机和次要主机

HA 群集中的某些主机会被指定为*主要*主机，这些主机会保存元数据和故障切换信息。群集中的前五台主机为主要主机，而所有其他主机则为*次要*主机。将主机添加到 HA 群集时，该主机必须与同一群集中一台现有主要主机进行通信，以完成其配置 （除非它是群集中的第一台主机，这样它就会是主要主机）。当一台主要主机不可用或被移除

时，HA 会将另一台主机提升至主要状态。主要主机可帮助提供冗余，并用来启动故障切换操作。

如果群集中的所有主机都没有响应，而这时将新的主机添加到该群集，则 HA 配置会失败，因为新主机无法与任何主要主机进行通信。在此情况下，必须先断开所有不响应的主机，然后才能添加新的主机。新主机将成为第一台主要主机。当其他主机再次可用时，将重新配置其 HA 服务，而后这些主机将成为主要主机或次要主机，具体取决于现有主要主机数目。

HA 群集与维护模式

将主机置于维护模式时，表示您正准备关闭该主机或对其进行维护。不能启动维护模式下主机上的虚拟机。主机发生故障时，HA 不会故障切换任何虚拟机至维护模式下的主机。HA 计算当前故障切换级别时，并不考虑这类主机。

主机退出维护模式后，会重新启用该主机上的 HA 服务，因此可再次用于故障切换。

主机正在进入维护模式时，如果其上没有任何已启动的虚拟机，则不能取消转换到维护模式的这一操作。

HA 群集和断开的主机

主机断开连接后，会存在于 VirtualCenter 清单中，但 VirtualCenter 不会从该主机获得任何更新，也不会监视该主机，并且不知道该主机的健康状况。由于不知道该主机的状态，又由于 VirtualCenter 没有与该主机进行通信，因此 HA 无法将其用作有保证的故障切换目标。计算当前故障切换级别时，HA 并不考虑断开的主机。

重新连接该主机后，该主机可再次用于故障切换。

以下列表介绍断开的主机和不响应的主机的差异。

- *断开的主机* 已被用户显式断开。断开主机过程中，VirtualCenter 会禁用该主机上的 HA。该主机上的虚拟机不会进行故障切换，并且 VirtualCenter 在计算当前故障切换级别时也不会考虑这些虚拟机。
- 如果*主机不响应*，则 VirtualCenter Server 不再接收来自该主机的检测信号。由于某些原因，可能会发生这种情况，例如，网络有问题、该主机发生故障或 VirtualCenter 代理发生故障。

  VirtualCenter 在计算当前故障切换级别时并不会包含此类主机，但假设如果该主机发生故障，将会故障切换不响应的主机上运行的任何虚拟机。不响应的主机上的虚拟机会影响接入控制检查。

HA 群集和主机网络隔离

主机上 HA 服务停止发送检测信号至群集中其他主机后 15 秒将进行主机故障检测。（可以更改该故障检测时间间隔的默认值。请参见 “设置高级 HA 选项”（第 113 页）。）如果发生故障或与网络隔离，则主机会停止发送检测信号。此时，群集中的其他主机将认为该主机发生了故障，而该主机本身则会在其失去网络连接超过 12 秒后，声明已与网络隔离。

如果隔离主机可访问 SAN，它将保留虚拟机文件上的磁盘锁，因而任何尝试故障切换虚拟机至其他主机的操作均会失败。虚拟机将继续在隔离主机上运行。VMFS 磁盘锁定可防止对虚拟机磁盘文件同时进行写操作，并且可以防止可能出现的损坏。默认情况下，隔离主机将关闭其虚拟机。然后，这些虚拟机就可以成功故障切换到群集中的其他主机。

如果网络连接在 12 秒内得以恢复，则群集中的其他主机不会将其视为主机故障。而会将其视为瞬间现象。此外，出现瞬间网络连接问题的主机并不会声明自己已与网络隔离，而是会继续运行。

如果 15 秒或更长时间内网络连接仍未恢复，则群集中的其他主机将认为该主机发生了故障，并会尝试故障切换该主机上的虚拟机。隔离主机会关闭其虚拟机，以便在其他网络连接正常的主机上重新启动这些虚拟机。

在 12 到 14 秒之间，隔离主机上的群集服务会声明自己已隔离，并开始使用默认隔离响应设置关闭虚拟机。如果这段时间内网络连接得以恢复，已关闭的虚拟机并不会在其他主机上重新启动，因为其他主机上的 HA 服务尚未认为该主机发生了故障。

因此，如果网络连接在主机失去连接后的 12 到 14 秒之间得以恢复，则虚拟机会关闭，但并不会进行故障切换。

结合使用 HA 和 DRS

HA 执行故障切换并在其他主机上重新启动虚拟机时，其首要的优先级是所有虚拟机的立即可用性。重新启动虚拟机后，启动这些虚拟机的主机可能会负载过重，而其他主机则相对负载较轻。 HA 将使用 CPU 和内存预留量来确定故障切换，而实际使用情况可能会更高。

结合使用 HA 和 DRS 可将自动故障切换与负载平衡结合起来。这种结合可在 HA 将虚拟机移至其他主机后快速重新平衡虚拟机。可以设置关联性和反关联性规则，以优先在相同主机 （关联性）或不同主机 （反关联性）上启动两个或更多个虚拟机。例如，可

以使用反关联性规则来确保运行重要应用程序的两个虚拟机永远不会运行在相同主机上。请参见 “使用 DRS 关联性规则”（第 99 页）。

注意 在使用 DRS 和 HA 并且启用了 HA 接入控制的群集中，可能不会从正在进入维护模式的主机上撤出虚拟机。这是由于预留用于维护故障切换级别的资源造成的。在此情况下，必须使用 VMotion 手动将虚拟机迁移出主机。

有效群集、黄色群集和红色群集

VI Client 会指示群集是有效、过量使用 （黄色），还是无效 （红色）。群集可能由于 DRS 冲突而过量使用。群集可能由于 DRS 冲突或 HA 冲突而变为无效。 [摘要 (Summary)] 页面显示的一条消息会指示该问题。

有效群集

有效群集拥有足够资源来满足所有预留量以及支持所有正在运行的虚拟机。群集会一直可用，除非发生某些情况而导致其过量使用或无效。

- DRS 群集可能由于一台主机发生故障而过量使用。
- 如果 VirtualCenter 不可用，并且使用直接连接 ESX Server 主机的 VI Client 启动虚拟机，则 DRS 群集将变为无效。
- 如果当前故障切换容量低于所配置的故障切换容量，或者如果群集内所有主要主机均不响应，则 HA 群集将变为无效。请参见 “主要主机和次要主机”（第 73 页）。
- 如果虚拟机正在进行故障切换时，用户减少了父资源池上的预留量，则 DRS 或 HA 群集将变为无效。

考虑下面的示例前，请注意以下术语的定义：

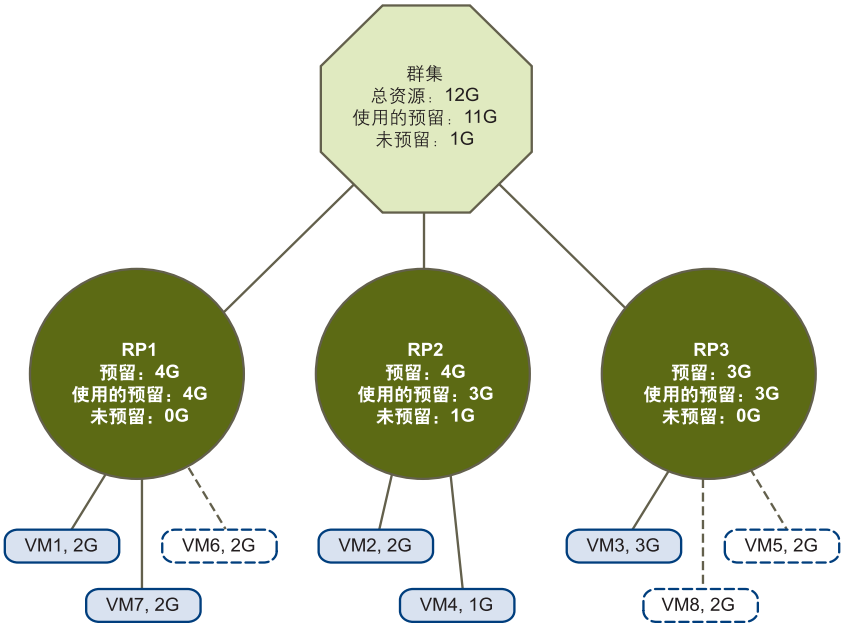
- **预留** （用于资源池） - 用户输入的资源池的固定、保证分配。
- **使用的预留** （用于群集或资源池） - 预留总量或已用于各子级的预留总量 （不管哪个较大），可重复添加。
- **未预留** （用于群集或资源池） - 一个非负数，该数字也会根据资源池类型的不同而有所不同：
  - 对于群集，它等于总容量 - 使用的预留。
  - 对于不可扩展的资源池，它等于预留 - 使用的预留。
  - 对于可扩展的资源池，它等于 （预留 - 使用的预留） + 任何可从祖先资源池借来的未预留资源。

示例 1：有效群集，固定的类型的所有资源池

图 4-10 显示有效群集以及如何计算其 CPU 资源。该群集具有以下特性：

- 总资源为 12 GHz 的群集。
- 三个资源池，每个类型均为 **[固定的 (Fixed)]**（未选择 **[可扩展预留 (Expandable Reservation)]**）。
- 三个资源池合起来的总预留为 11 GHz (4+4+3 GHz)。总数显示在群集的 **[使用的预留 (Reservation Used)]** 字段中。
- RP1 是使用 4 GHz 预留创建的。两个虚拟机。启动了 （VM1 和 VM7），每个各占 2 GHz （**[使用的预留 (Reservation Used)]**：4 GHz）。未剩下资源用于启动额外的虚拟机。 VM6 显示为未启动。它不消耗任何预留。
- RP2 是使用 4 GHz 预留创建的。启动了两个虚拟机，分别各占 1 GHz 和 2 GHz （**[使用的预留 (Reservation Used)]**：3 GHz）。还剩 1 GHz 未预留。
- RP3 是使用 3 GHz 预留创建的 （预留）。启动了一个占用 3 GHz 的虚拟机。没有资源可用于启动额外的虚拟机。

图 4-10. 有效群集 （固定的资源池）

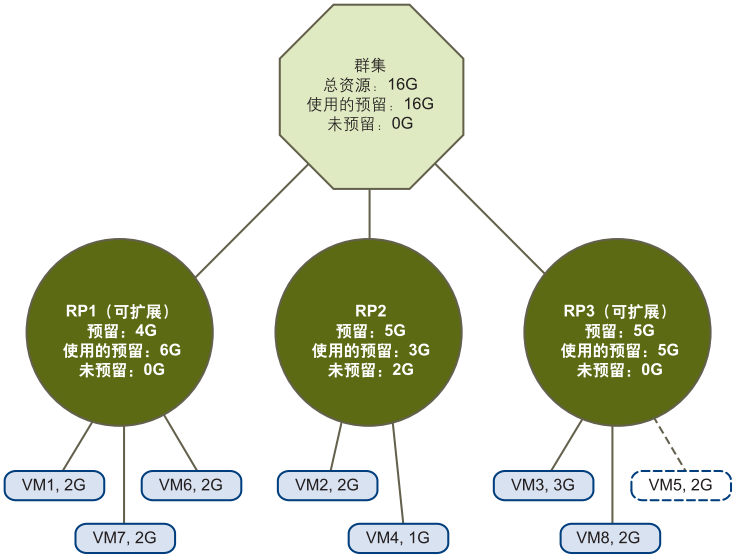


示例 2：有效群集，可扩展的类型的一些资源池

示例 2 (图 4-11) 使用类似示例 1 的设置。但是，RP1 和 RP3 使用的预留类型为 **[可扩展的 (Expandable)]**。可按如下配置有效群集：

- 总资源为 16 GHz 的群集。
- RP1 和 RP3 的类型为 **[可扩展的 (Expandable)]**，RP2 的类型为 [固定的 (Fixed)]。
- 这三个资源池合起来所使用的总预留是 16 GHz （其中 RP1 占 6 GHz， RP2 占 5 GHz， RP3 占 5 GHz）。 16 GHz 显示为顶层群集的 **[使用的预留 (Reservation Used)]**。
- RP1 是使用 4 GHz 预留创建的。启动了三个虚拟机，每个各占用 2 GHz。这些虚拟机中的两个 （例如， VM1 和 VM7）可以使用 RP1 的预留，第三个虚拟机 (VM6) 可以使用群集资源池中的预留。（如果此资源池的类型为 **[固定的 (Fixed)]**，则无法启动额外的虚拟机。）
- RP2 是使用 5 GHz 预留创建的。启动了两个虚拟机，分别各占 1 GHz 和 2 GHz （**[使用的预留 (Reservation Used)]**：3 GHz）。还剩 2 GHz 未预留。
- RP3 是使用 5 GHz 预留创建的。启动了两个虚拟机，分别各占 3 GHz 和 2 GHz。即使此资源池的类型为 **[可扩展的 (Expandable)]**，也无法启动额外的 2 GHz 虚拟机，因为父资源池的额外资源已被 RP1 占用。

图 4-11. 有效群集 （可扩展的资源池）



黄色群集

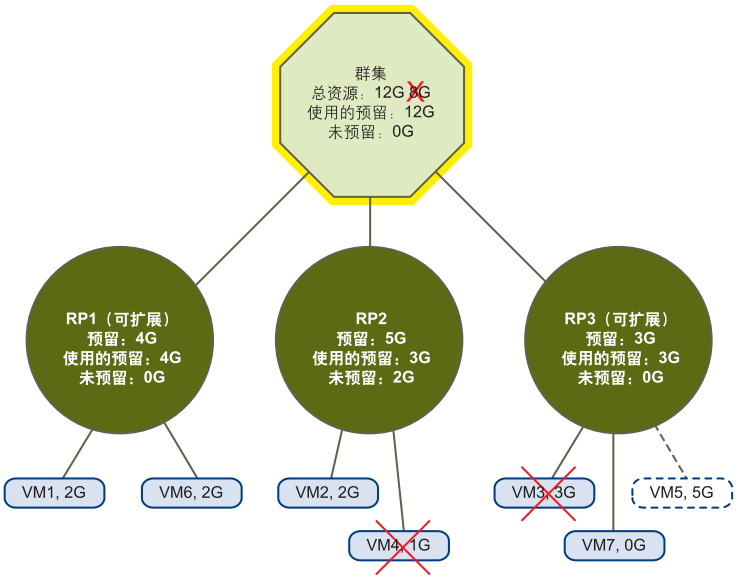
当资源池和虚拟机的树在内部是一致的，但群集中没有足够容量来支持子资源池所预留的所有资源时，群集将会变为黄色。始终会有足够的资源来支持所有正在运行的虚拟机，因为当主机不可用时，其所有的虚拟机也不可用。

当群集容量突然减少 （例如，群集中一台主机不可用时），群集通常会变为黄色。VMware 建议留足额外的群集资源，以避免群集变为黄色。

请考虑以下示例，如图 4-12 所示：

- 总资源为 12 GHz （分别来自三台各有 4 GHz 资源的主机）的群集。
- 预留了总共 12 GHz 资源的三个资源池。
- 三个资源池合起来所使用的总预留为 12 GHz (4+5+3 GHz)。该数值显示为群集中的 **[使用的预留 (Reservation Used)]**。
- 由于其中一个 4 GHz 主机不可用，因此总资源减少至 8 GHz。
- 同时，故障主机上运行的 VM4 (1 GHz) 和 VM3 (3 GHz) 都不再运行。
- 该群集现在正在运行的虚拟机需要总共 6 GHz 的资源。该群集仍有 8 GHz 的资源可用，足够满足虚拟机需求。
- 由于不再能达到 12 GHz 的资源池预留，因此群集会被标记成黄色。

图 4-12. 黄色群集



红色群集

群集可能由于 DRS 冲突或 HA 冲突而变为红色。群集的行为取决于冲突的类型，如本节所述。

红色 DRS 群集

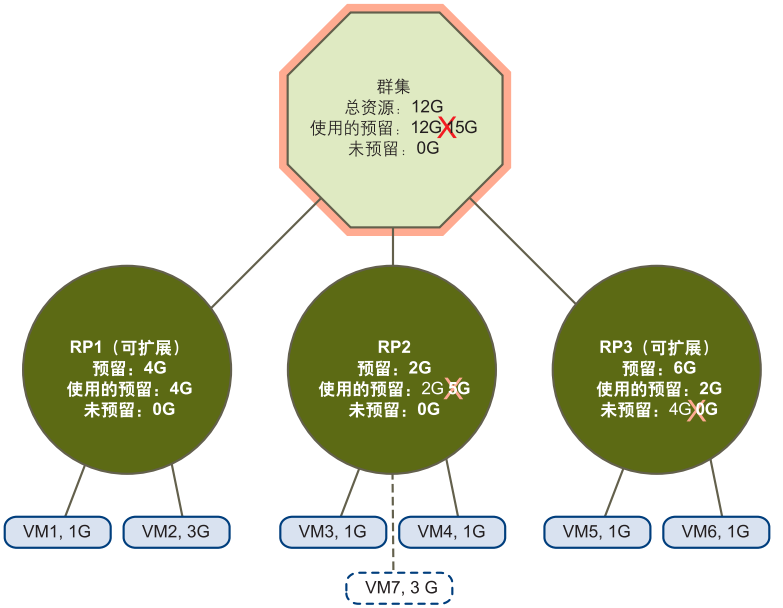
当树内部不再一致，即未遵守资源限制时，已启用 DRS 的群集会变为红色。群集中资源的总数量并不会对该群集是否为红色造成影响。即使根级别中有足够资源，但如果子级别存在不一致性，该群集也有可能会变为 DRS 红色。

可通过关闭一个或多个虚拟机、将虚拟机移至树中有足够资源的部分，或者编辑红色部分的资源池设置，来解决红色 DRS 群集问题。添加资源通常仅在处于黄色状况时才有用。

如果虚拟机正在进行故障切换时，重新配置资源池，则群集也可能会变为红色。正在进行故障切换的虚拟机会断开连接，并且不会算入父资源池所使用的预留。可在故障切换完成前，减少父资源池的预留。故障切换完成后，会再次将虚拟机资源纳入父资源池计算。如果池的使用情况大于新的预留，则该群集将变为红色。

请考虑图 4-13 中的示例。

图 4-13. 红色群集



如图 4-13 中的示例所示，如果用户 （以不支持的方式）能够启动一个使用资源池 2 下 3 GHz 预留的虚拟机，则该群集会变为红色。

红色 HA 群集

当已启动的虚拟机数超过了故障切换需求，即当前故障切换容量小于所配置的故障切换容量，则已启用 HA 的群集将会变为红色。如果禁用了严格的接入控制，则无论主机能否保证故障切换，群集都不会变为红色。

有时可能会出现故障切换容量不足的情况，例如，如果启动了许多虚拟机，以致该群集不再有足够的资源来保证指定数目主机的故障切换。

如果在有四台主机的群集中为两台主机故障设置了 HA，而一台主机发生了故障，也会出现这种情况。剩余的三台主机可能不再能满足两台主机故障。

如果已启用 HA 的群集变为红色，或者如果当前故障切换容量降至所配置的故障切换容量之下，则将不能再保证指定数目主机的故障切换，但会继续执行故障切换。万一主机发生故障，HA 将首先按优先级顺序故障切换一台主机上的虚拟机，然后再按优先级顺序故障切换第二台主机上的虚拟机，依此类推。请参见 “对虚拟机进行 HA 定制”（第 106 页）。

[摘要 (Summary)] 页面显示红色和黄色群集的配置问题列表。该列表说明造成群集过量使用或无效的原因。

注意 如果群集是由于 HA 问题而变为红色的，则 DRS 行为不会受到影响。

创建 VMware 群集 5

本章将讨论以下主题：

注意 所有任务均假定您有相应的操作权限。有关权限的信息，请参见联机帮助。

群集先决条件

要成功使用 VMware 群集功能，系统必须满足某些先决条件。

- 一般而言，如果虚拟机满足 VMotion 要求，则 DRS 和 HA 可以达到最佳工作状态，如下一节所述。
- 如果要使用 DRS 进行负载平衡，则群集中的主机必须是 VMotion 网络的一部分。如果主机不在 VMotion 网络中， DRS 仍可提供初始放置位置建议。

启用 HA 的群集

因为必须确保能启动群集中任意主机上的虚拟机，所以在启用 HA 的群集中，所有虚拟机及其配置文件必须驻留在共享存储器 （例如 SAN）上。而且还必须配置主机，使其能够访问相同的虚拟机网络以及其他资源。

HA 群集中的每一台主机都必须能够解析主机名称以及群集中所有其他主机的 IP 地址。为此，可以在每个主机上设置 DNS （首选）或者手动填写 `/etc/hosts` 条目 （容易出错，不建议采用这种方法）。要使用 DNS 解析名称，必须确保启用 ESX Server 主机防火墙上的 NIS 客户端服务。（如果使用 ESX Server 3i，则不必这样做。）

启用 NIS 客户端服务

1 在 VI Client 中，选择主机，然后单击 **[配置 (Configuration)]** 选项卡。

2 选择 **[安全配置文件 (Security Profile)]**。

3 如果 **[防火墙 (Firewall)]** 的 **[出站连接 (Outgoing Connections)]** 下面没有列出 **[NIS 客户端 (NIS Client)]**，请单击 **[属性 (Properties)]**。

4 在 [防火墙属性 (Firewall Properties)] 对话框中选择 **[NIS 客户端 (NIS Client)]**，然后单击 **[确定 (OK)]**。

注意 HA 群集中的所有主机都必须配置 DNS 以便群集中任何主机的短主机名 （不带域后缀）都可以从群集中任意其他主机解析为正确的 IP 地址。否则，配置 HA (`Configuring HA`) 任务会失败。如果使用 IP 地址添加主机，则另外还要启用反向 DNS 查找 （IP 地址应可解析为短主机名）。

当 VMware HA 在 ESX Server 3 中使用时，建议使用冗余控制台网络 （尽管不是必须的）。类似地，对于 ESX Server 3i，也建议使用冗余 VMkernel 网络。如果未提供冗余，则故障切换设置中只有一个故障点。当一个主机的网络连接出现故障时，第二个连接可以向其他主机发送检测信号。

要设置冗余，每个主机上需要有两个物理网络适配器。使用两个服务控制台界面 （ESX Server 3i 中的 VMkernel 网络界面），或者使用单个界面 （使用网卡成组），将它们连接到相应的服务控制台 （或 ESX Server 3i 中的 VMkernel 网络）。

注意 在为 HA 群集中的一台主机添加网卡以后，必须在该主机上重新配置 HA。

VirtualCenter VMotion 要求

要配置 VMotion，群集中的每台主机必须满足下列要求。

共享存储器

确保受管主机使用共享存储器。共享存储器一般位于存储区域网络 (Storage Area Network, SAN) 上。请参见 *《iSCSI SAN 配置指南》*以及 *《光纤通道 SAN 配置指南》*，了解有关 SAN 的详细信息；请参见 *《ESX Server 配置指南》*，了解有关其他共享存储器的信息。

共享 VMFS 卷

配置所有受管主机以使用共享 VMFS 卷。

- 将所有虚拟机的磁盘置于可通过源主机和目标主机访问的 VMFS 卷上。
- 将共享 VMFS 的访问模式设置为公用。
- 确保 VMFS 卷足够大，可以存储虚拟机的所有虚拟磁盘。
- 确保源主机及目标主机上的所有 VMFS 卷都使用了卷名称，并且所有虚拟机都使用这些卷名称来指定虚拟磁盘。

注意 虚拟机交换文件同样需要置于源主机和目标主机可以访问的 VMFS 上 （就像 `.vmdk` 虚拟磁盘文件一样）。如果所有的源主机及目标主机都是 ESX Server 3.5 或更高版本，则此要求将不再适用。这种情况下，支持交换文件位于非共享存储器上的 VMotion。默认情况下，交换文件置于 VMFS 上，但管理员可能会使用高级虚拟机配置选项替代此文件位置。

处理器兼容性

确保源主机和目标主机具有一组兼容的处理器。

VMotion 在基础 VMware ESX Server 系统之间传输虚拟机的运行架构状态。VMotion 兼容性的意思是目标主机的处理器必须能够使用源主机的处理器在挂起时使用的等效指令来恢复执行。处理器时钟速度和缓存大小可能不同，但处理器必须属于相同的供应商类别 （Intel 与 AMD）和相同的处理器系列，以便达到通过 VMotion 迁移所需的兼容性。

处理器系列 （如 Xeon MP 和 Opteron）是由处理器供应商定义的。可以通过比较处理器的型号、步进级别和扩展功能来区分同一系列中的不同处理器版本。

- 在大多数情况下，同一系列中的不同处理器版本都很类似，足以保持兼容性。

- 在某些情况下，处理器供应商在同一处理器系列中引入了重大的架构更改 （例如 64 位扩展及 SSE3）。如果不能保证使用 VMotion 进行成功迁移，VMware 会识别这些异常情况。

注意 VMware 正与处理器和硬件供应商共同合作，致力于在最大范围的处理器之间实现 VMotion 兼容性。有关当前的信息，请联系 VMware 代表。

其他要求

必须遵守的其他 VMotion 要求：

- 对于 ESX Server 3.x， ESX Server 主机的虚拟机配置文件必须驻留在 VMFS 上。
- VMotion 不支持裸磁盘或使用 Microsoft 群集服务 (Microsoft Cluster Service, MSCS) 群集的应用程序迁移。
- VMotion 要求在所有启用 VMotion 的受管主机之间设置专用的千兆以太网迁移网络。在受管主机上启用 VMotion 后，要为受管主机配置唯一的网络标识对象并将其连接到专用迁移网络。

群集创建概述

创建群集时，请确保系统满足群集先决条件。（请参见 “群集先决条件”（第 83 页）。）启动新建群集向导。

启动群集向导

1 右键单击数据中心或文件夹，然后选择 **[新建群集 (New Cluster)]**。（键盘快捷键是 Ctrl+I）。

2 按向导提示和本章说明选择群集设置。

在第一个面板中，选择创建一个支持 VMware DRS、VMware HA，还是同时支持这两者的群集。该选择影响到随后显示的页面，而且也隐式决定了显示在向导左面板中的任务列表。如果选择 [DRS] 和 [HA]，向导会提示您为这些选项提供配置信息。

创建群集时，群集一开始并不包含任何主机或虚拟机：

- “向 DRS 群集添加主机”（第 94 页）和“将主机添加至 HA 群集”（第 110 页）讨论了添加主机的方法。
- 第 7 章，“群集和虚拟机”（第 103 页）讨论了添加虚拟机的方法。

创建群集

本节讨论新建群集向导的每个页面。

选择群集功能

新建群集向导中的第一个面板允许指定以下信息：

- **[名称 (Name)]** - 指定群集的名称。该名称显示在 VI Client 清单面板中。必须指定一个名称，然后继续创建群集。
- **[启用 VMware HA (Enable VMware HA)]** - 如果选中了此框，则当源主机发生故障时， VirtualCenter 将在另一个主机上重新开始运行虚拟机。请参见 “了解 VMware HA”（第 69 页）。
- **[启用 VMware DRS (Enable VMware DRS)]** - 如果选中了此框，则 DRS 使用负载分布信息来提出初始放置建议和负载平衡建议，或者自动放置和迁移虚拟机。请参见 “了解 VMware DRS”（第 60 页）。

指定名称并选择一个或两个群集功能。单击 **[下一步 (Next)]** 继续。

注意 可以更改已选择的群集功能。请参见第 6 章，“管理 VMware DRS”（第 93 页）和第 8 章，“管理 VMware HA”（第 109 页）。

选择自动化级别

如果在向导的第二个面板中选择了 **[启用 VMware DRS (Enable VMware DRS)]** 选项，则 [VMware DRS] 面板可让您选择默认自动化级别。有关不同选择的详细讨论，请参见 “了解 VMware DRS”（第 60 页）。

注意 可以更改整个群集或单个虚拟机的自动化级别。请参见 “重新配置 DRS”（第 98 页）和 “对虚拟机进行 DRS 定制”（第 105 页）。

表 5-1 概括了向导提供的选项。

表 5-1. DRS 自动化级别

	初始放置位置	迁移
[手动 (Manual)]	显示建议的主机。	显示迁移建议。
[半自动 (Partially Automated)]	自动放置。	显示迁移建议。
[全自动 (Fully Automated)]	自动放置。	自动执行迁移建议。

注意 默认群集自动化级别及特定虚拟机的自动化模式均不影响分布式电源管理功能产生的建议和操作。通过选择*电源管理*自动化级别可以实现这一点。请参见 "分布式电源管理"（第 66 页）。

选择 HA 选项

如果启用 HA，新建群集向导允许设置表 5-2 中列出的选项。请参见 "处理 VMware HA"（第 112 页）。

表 5-2. VMware HA 选项

选项	描述
[主机故障 (Host Failures)]	指定要保证故障切换的主机故障数。
[重新启动优先级 (Restart Priority)]	确定在主机发生故障时重新启动虚拟机的顺序。可选值为：**[禁用 (Disabled)]**、**[低 (Low)]**、**[中等 (Medium)]**、**[高 (High)]**。默认为 **[中等 (Medium)]**。如果选择了 **[禁用 (Disabled)]**，则虚拟机禁用 HA。 请参见 "重新启动优先级"（第 106 页）。
[隔离响应 (Isolation Response)]	确定当 HA 群集中的某个主机失去其控制台网络 （或 ESX Server 3i 中的 VMkernel 网络）连接但仍在运行时发生的情况。可选值为：**[关闭 (Power off)]** （默认）及 **[保持启动 (Leave powered on)]**。请参见 "隔离响应"（第 107 页）。
[接入控制 (Admission Control)]	提供两个选项： ■ **[如果虚拟机违反可用性限制，则不启动虚拟机 (Do not power on virtual machines if they violate availability constraints)]** 强制执行上面设置的故障切换容量。 ■ **[即使虚拟机违反可用性限制也允许启动虚拟机 (Allow virtual machines to be powered on even if they violate availability constraints)]** 允许在故障切换指定的主机数量无法保证的情况下启动虚拟机。如果没有选择此选项 （默认），则当下面的操作会违反可用性限制时，这些操作也不会被允许执行： 　■ 将关闭的虚拟机恢复到已启动快照 　■ 将虚拟机迁移至群集 　■ 重新配置虚拟机，以增加其 CPU 或内存预留量

选择虚拟机交换文件位置

此向导页面允许为虚拟机的交换文件选择位置。可以将交换文件与虚拟机本身存储在同一目录中，或者将交换文件存储在主机指定的数据存储中 （主机 - 本地交换）。请参见 "交换"（第 131 页）。

完成群集创建

完成群集的所有选择后，向导会显示一个 [摘要 (Summary)] 页面，列出选择的选项。单击 **[完成 (Finish)]** 以完成群集的创建，或单击 **[上一步 (Back)]** 返回并对群集设置进行修改。

可以查看群集信息 （请参见 “查看群集信息”（第 89 页））或者向群集添加主机及虚拟机 （请参见 “向 DRS 群集添加主机”（第 94 页）及 “将主机添加至 HA 群集”（第 110 页））。

还可以定制群集选项，如第 6 章，“管理 VMware DRS”（第 93 页）及第 8 章，“管理 VMware HA”（第 109 页）中所述。

查看群集信息

本节讨论当您在清单面板中选择群集时显示的信息页面。

注意 有关所有其他页面的信息，请参见联机帮助。

[摘要 (Summary)] 页面

群集 [摘要 (Summary)] 页面显示群集的摘要信息。请参见图 5-1。

图 5-1. 群集 [摘要 (Summary)] 选项卡

表 5-3. 群集摘要信息

面板	描述
[常规 (General)]	包含群集的信息： **[VMware DRS]** - [已启用 (Enabled)] 或 [已禁用 (Disabled)]。 **[VMware HA]** - [已启用 (Enabled)] 或 [已禁用 (Disabled)]。 **[总的 CPU 资源 (Total CPU Resources)]** - 可用于群集的总 CPU 资源。来自主机的所有可用资源的总和。 **[总内存 (Total Memory)]** - 群集的总内存。来自主机的所有可用资源的总和。 **[主机数 (Number of Hosts)]** - 群集中主机的数目。如果添加或移除主机，该数目会更改。 **[处理器总数 (Total Processors)]** - 所有主机的所有处理器的总和。 **[虚拟机数目 (Number of Virtual Machines)]** - 群集或群集任意子资源池中所有虚拟机的总数。包括当前未启动的虚拟机。 **[迁移总数 (Total Migrations)]** - 自群集创建以来由 DRS 或用户执行的迁移总数。
[命令 (Commands)]	允许调用常用的群集命令。 **[新建虚拟机 (New Virtual Machine)]** - 显示新建虚拟机向导。该向导提示您从群集中选择一个主机。 **[添加主机 (Add Host)]** - 添加一个当前未由同一 VirtualCenter Server 管理的主机。要添加由同一 VirtualCenter Server 管理的主机，请拖放清单面板中的主机。 **[创建资源池 (New Resource Pool)]** - 创建群集的子资源池。 **[编辑设置 (Edit Settings)]** - 显示群集的 [编辑设置 (Edit Settings)] 对话框。
[VMware HA]	显示接入控制设置、当前故障切换容量，以及启用 HA 的群集的已配置故障切换容量。 无论何时在群集中添加或移除主机，或者启动或关闭虚拟机，系统都会更新当前的故障切换容量。
[VMware DRS]	显示默认自动化级别、迁移阈值，以及群集的未完成迁移建议。 如果选择 **[DRS 建议 (DRS Recommendations)]** 选项卡，会显示迁移建议。请参见 “迁移建议”（第 65 页）。 默认自动化级别及迁移阈值在群集创建时进行设置。请参见 “迁移阈值”（第 65 页）。
[DRS 资源分发 (DRS Resource Distribution)]	显示两个实时柱状图：**[利用率百分比 (Utilization Percent)]** 及 **[已递送授权资源百分比 (Percent of Entitled Resources Delivered)]**。这两个图显示了群集的平衡情况。请参见 “DRS 资源分发图”（第 91 页）。

DRS 资源分发图

这两个 [DRS 资源分发 (DRS Resource Distribution)] 图允许评估群集的健康状况。这两个图在 [摘要 (Summary)] 页面每次显示的时候都会进行更新，并且会在性能限制允许的情况下定期进行更新。

上面的 DRS 资源分发图

该图是一个柱状图，X 轴显示主机数，Y 轴显示利用率百分比。如果群集不平衡，您会看到多个竖条，这些竖条对应不同的利用率水平。例如，可能有一个 CPU 利用率为 20% 的主机，一个 CPU 利用率为 80% 的主机，每个主机的利用率都用一个蓝色竖条来表示。在具有自动默认自动化级别的群集中， DRS 可以将虚拟机从负载较重的主机迁移到资源利用率为 20% 的主机上。结果会出现一个范围在 40% 到 50% 之间的蓝色竖条，表示容量相似的主机。

对于一个平衡的群集，该图显示两个竖条：一个表示 CPU 利用率，一个表示内存利用率。但是，如果群集中的主机利用率不高，则在平衡的群集中， CPU 和内存都可能会有多个竖条。

下面的 DRS 资源分发图

该图是一个柱状图，Y 轴显示主机数，X 轴显示每个主机的已递送授权资源百分比。上面的图报告原始资源利用率值，而下面的图也包括有关虚拟机及资源池的资源设置的信息。

DRS 根据配置的份额、预留、限制设置，以及当前资源池配置和设置，计算每个虚拟机的资源权利。然后，通过将运行在该主机上的所有虚拟机的资源权利进行累加，计算每个主机的资源权利。已递送授权资源百分比等于该主机的容量除以它的权利。

对于平衡群集，主机的容量应大于或等于其权利，因此在理想情况下，该图应有一个竖条表示每种资源，该竖条位于柱状图中 90% 到 100% 的那一段。不平衡的群集有多个竖条。 X 轴值较低的竖条表示这些主机上的虚拟机未获得其有权获得的资源。

[DRS 建议 (DRS Recommendations)] 页面

[DRS 建议 (DRS Recommendations)] 页面显示为通过迁移或电源管理优化群集中的资源利用而生成的一组当前建议。根据为群集设置的值， VirtualCenter 会定期更新建议列表。

如果没有当前建议，[DRS 建议 (DRS Recommendations)] 页面会显示 **[此时没有 DRS 建议 (No DRS recommendations at this time)]**。如果有当前建议，此部分的显示内容将如同图 5-2 所示：

图 5-2. DRS 建议

[DRS 建议 (DRS Recommendations)] 部分显示有关表 5-4 中所述的列中每一项的信息。

表 5-4. [DRS 建议 (DRS Recommendations)] 信息

列	描述
[优先级 (Priority)]	此建议的优先级 （用星数表示）。五星 （最高级别）表示由于主机进入维护模式或违反关联性规则需要强制移动。其他级别表示建议能在多大程度上提高群集性能，从四星 （显著提高）到一星 （轻微提高）。
[建议 (Recommendation)]	本列显示的内容取决于建议的类型： ■ 对于虚拟机迁移：要迁移的虚拟机名称、（虚拟机当前正在其上运行的）源主机以及 （虚拟机要迁移到其上的）目标主机。 ■ 对于主机电源状况更改：要启动或关闭的主机名称。
[原因 (Reason)]	建议的原因：为何建议迁移虚拟机或建议更改主机电源状况。原因可能与以下情况有关： ■ 平衡平均 CPU 或内存负载。 ■ 满足关联性或反关联性规则。 ■ 主机正在进入维护模式。 ■ 降低功耗。 ■ 关闭特定主机。 ■ 增加群集容量。

有关使用此页面的信息，请参见 ″应用 DRS 建议″（第 97 页）。

注意 紧靠在 [DRS 建议 (DRS Recommendations)] 部分下面的 [DRS 操作历史记录 (DRS Action History)] 部分显示某段时间内此群集应用的建议。

6 管理 VMware DRS

本章说明如何向 DRS 群集中添加及移除主机，以及如何定制 DRS。

本章将讨论以下主题：

注意 所述所有任务均假定您已获得许可并拥有执行这些任务的权限。有关权限及如何设置权限的信息，请参见联机帮助。

定制 DRS

创建群集后，可以为 DRS 和 / 或 HA 启用群集。然后可以继续添加或移除主机，并用其他方法定制群集。

可以按以下说明定制 DRS：

- 在群集创建过程中指定默认的自动化级别及迁移阈值。请参见 ″选择自动化级别″（第 87 页）。
- 为该群集添加主机。请参见 ″向 DRS 群集添加主机″（第 94 页）

- 更改现有群集的默认自动化级别或迁移阈值，如 “重新配置 DRS”（第 98 页）中所述。
- 可以为群集中的单个虚拟机设置自定义自动化模式以替代群集范围的设置。例如，可以将群集的默认自动化级别设置为自动，而将其中一些计算机的模式设置为手动。请参见 “对虚拟机进行 DRS 定制”（第 105 页）。
- 使用关联性规则对虚拟机进行分组。关联性规则指定选定的虚拟机应始终放置在同一主机上。反关联性规则指定选定的虚拟机应始终放置在不同的主机上。请参见 “使用 DRS 关联性规则”（第 99 页）。

向 DRS 群集添加主机

对于当前由同一 VirtualCenter Server 管理的主机 （受管主机）和当前未由该服务器管理的主机，将主机添加至群集的步骤有所不同。

添加完主机后，部署至该主机的虚拟机将成为群集的一部分。 DRS 可能会建议将部分虚拟机迁移到群集中的其他主机上。

将受管主机添加至群集

VirtualCenter 清单面板显示由该 VirtualCenter Server 管理的所有群集和所有主机。有关向 VirtualCenter Server 添加主机的信息，请参见 *《Virtual Infrastructure 用户指南》*。

将受管主机添加至群集

1 从清单或列表视图中选择主机。

2 将主机拖至目标群集对象。

3 向导会询问您要对主机上的虚拟机及资源池进行什么操作。

- 如果选择 **[将该主机的虚拟机放到群集的根资源池中 (Put this host's virtual machines in the cluster's root resource pool)]**， VirtualCenter 将移除该主机现有的所有资源池，并将该主机层次结构中的虚拟机全部连接到根资源池。

注意 因为份额的分配与资源池有关，而上述操作破坏了资源池的层次结构，所以在执行上述操作后可能必须手动更改虚拟机的份额。

- 如果选择 **[为此主机的虚拟机和资源池创建一个新的资源池 (Create a new resource pool for this host's virtual machines and resource pools)]**，VirtualCenter 将创建一个顶层资源池作为群集的直接子级，并将主机的所有

子级添加到这个新的资源池。您可以命名这个新的顶层资源池。默认为 **[已从 <host_name> 移植 (Grafted from <host_name>)]**。

注意 如果主机没有子资源池或虚拟机，其资源将添加到群集，但不会创建带有顶层资源池的资源池层次结构。

要利用自动迁移功能，还必须设置主机的 VMotion 网络。

注意 如果以后从群集移除主机，资源池层次结构仍会是群集的一部分。主机会失去该资源池层次结构。由于资源池的目的之一就是支持独立于主机的资源分配，因此这种丢失是有意义的。例如，可以移除两个主机，用一个功能相似的主机代替而不需要经过额外的重新配置。

将非受管主机添加至群集

可以添加当前并未由群集所在同一 VirtualCenter Server 管理的主机 （该主机在 VI Client 中不可见）。

将非受管主机添加至群集

1 选择要将主机添加至的群集，并从右键菜单中选择 **[添加主机 (Add Host)]**。

2 提供主机名、用户名和密码，然后单击 **[下一步 (Next)]**。

3 查看摘要信息并单击 **[下一步 (Next)]**。

4 回答 “将受管主机添加至群集”（第 94 页）中讨论过的关于虚拟机及资源池位置的提示。

从群集移除主机

要从群集中移除主机，必须将主机置于维护模式。有关背景信息，请参见 “主机维护模式和待机模式”（第 67 页）。

将主机置于维护模式

1 选择主机，并从右键菜单中选择 **[进入维护模式 (Enter Maintenance Mode)]**。

此时主机将处于 **[正在进入维护模式 (Entering Maintenance Mode)]** 的状况，直到关闭所有正在运行的虚拟机或将虚拟机迁移到其他主机。无法启动正在进入维护模式的主机上的虚拟机，或者将虚拟机迁移到该主机。

如果没有已启动的虚拟机，主机会进入维护模式。

2 主机处于维护模式时，可以将其拖到其他清单位置，该位置可以是顶层数据中心或者其他群集。

移动主机时，主机的资源会从群集中移除。如果将主机的资源池层次结构移植到群集上，则该层次结构将保留在群集中。

3 移动该主机后，您可以：

- 从 VirtualCenter Server 移除该主机（在右键菜单中选择 **[移除 (Remove)]**）。
- 将该主机作为 VirtualCenter 下的独立主机运行 （在右键菜单中选择 **[退出维护模式 (Exit Maintenance Mode)]**）。
- 将主机移至另一个群集。

主机移除及资源池层次结构

即使在将主机添加到群集时使用了 DRS 群集并决定移植主机资源池，将该主机从群集中移除后，主机也只保留 （不可见的）根资源池。在这种情况下，层次结构将保留在群集中。

可以创建一个新的、特定于主机的资源池层次结构。

主机移除及虚拟机

因为移除主机前必须将其置于维护模式，所以必须关闭所有虚拟机。从群集移除主机时，当前与该主机关联的虚拟机也会从该群集中移除。

注意 因为 DRS 可以将虚拟机从一个主机迁移至另一个主机，所以主机上的虚拟机可能不同于最初添加主机时的虚拟机。

主机移除与无效群集

如果从群集中移除主机，群集的可用资源会减少。

如果群集启用了 DRS，移除主机会导致以下结果：

- 如果群集仍然有足够的资源用于满足群集中所有虚拟机和资源池预留的需要，群集会调整资源的分配以反映减少的资源量。
- 如果群集没有足够的资源满足所有资源池预留的需要，但是有足够的资源满足所有虚拟机预留的需要，就会出现警报，而且该群集会被标记为黄色。DRS 继续运行。

如果启用 HA 的群集失去的资源过多以致无法再满足故障切换的需要，会出现一条消息，而且群集将变成红色。主机发生故障时群集可以对虚拟机进行故障切换，但不能保证有足够的可用资源对所有虚拟机进行故障切换。

应用 DRS 建议

VirtualCenter 在 [DRS 建议 (DRS Recommendations)] 页面显示群集的迁移及电源管理建议。请参见 “[DRS 建议 (DRS Recommendations)] 页面”（第 91 页）。该页面也是应用建议的位置。请参见图 6-1。

图 6-1. DRS 建议

建议分组

建议页面组织成若干个框的形式。每个框包含共享某种相互依赖度的建议，而不同框中显示的建议可以认为是相互独立的。在同一个框内，依赖于其他建议的建议放置在其*先决条件*下面。在应用建议时，这些相互依赖关系将导致以下操作：

- 选择一个存在依赖关系的建议后（选中 **[应用 (Apply)]** 复选框），在同一个框内该建议上方作为其先决条件的所有建议也会被选中。如果没有这些建议，则无法应用存在依赖关系的建议。但并不是所有在其上方的建议都是该建议的先决条件。
- 当取消选择一个先决条件建议时，同一个框内在该建议之下并且依赖于该建议的所有建议也会被取消选择。如果不应用先决条件，则这些建议也不会被应用。但并不是所有在其下方的建议都依赖于该建议。

在这些建议之间，显示在该页面上的另一种类型的相互依赖关系是：如果操作是*原子*操作，则只能作为一个单元应用。这些类型的建议可能是满足某种关联性（或反关联性）规则所必需的，此类建议由一个链接图标及一个 **[应用 (Apply)]** 复选框表示。

使用 [DRS 建议 (DRS Recommendations)] 页面

默认情况下，[DRS 建议 (DRS Recommendations)] 页面上所有 DRS 建议的 **[应用 (Apply)]** 复选框都会被选中且不可用 （无法取消选择）。要取消选择建议，请选中 **[覆盖建议的 DRS 操作 (Override suggested DRS actions)]** 复选框。该操作会激活 [应用 (Apply)] 复选框。在决定要应用哪个 （些）建议后，单击 **[应用建议 (Apply Recommendations)]** 按钮。

还可以从 [DRS 建议 (DRS Recommendations)] 页面执行其他两项操作：

- 单击 **[生成建议 (Generate Recommendations)]** 按钮刷新整个页面。如果更改了群集的配置并想立即查看新配置的更新建议，您可能会这样做。
- 阈值链接以星级形式 （例如 *2 星或 2 星以上*）显示在建议表上方，单击此阈值链接打开 [群集设置 (Cluster Settings)] 对话框。可以调整默认群集自动化级别和电源管理自动化级别，以及其他群集级别的设置。

重新配置 DRS

可以为群集关闭 DRS 或更改配置选项。

关闭 DRS

1 选择群集。

2 从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

3 在左侧面板中选择 **[常规 (General)]** 并取消选中 **[VMware DRS]** 复选框。

此时系统会警告您关闭 DRS 将销毁群集中的所有资源池。

4 单击 **[确定 (OK)]** 关闭 DRS 并销毁所有资源池。

小心 再次启用 DRS 不会重新建立资源池。虽然将 DRS 自动化级别更改为手动 （而不是将其关闭）会阻止任何 DRS 自动操作，但这样做可以保留资源池层次结构。

重新配置 DRS

1 选择群集。

2 从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

3 在 [群集设置 (Cluster Settings)] 对话框中选择 **[VMware DRS]**。

4 设置默认自动化级别：

- 选择一个单选按钮更改自动化级别。请参见 “选择自动化级别” （第 87 页）。

- 如果选择了 **[全自动 (Fully automated)]**，可以移动 **[迁移阈值 (Migration Threshold)]** 滑块来更改迁移阈值。请参见 “迁移阈值”（第 65 页）。

注意 当您与 VMware 客户支持合作解决问题时，[高级选项 (Advanced Options)] 对话框很有用。其他情况下，不建议设置高级选项。

要在不更改资源池层次结构的情况下挂起 DRS 建议的 VMotion 迁移，请将 DRS 自动化级别设置为手动或半自动。设置完成后， DRS 仍会建议进行 VMotion 迁移，但 VirtualCenter 不经用户许可不会执行迁移操作。如果已将单个虚拟机的自动化级别设置为全自动，请将其恢复为群集的默认设置。有关说明，请参见 “对虚拟机进行 DRS 定制”（第 105 页）。

使用 DRS 关联性规则

创建 DRS 群集之后，可以编辑其属性以创建指定关联性的规则。 DRS 不会违反用户指定的规则，但如果有两条规则相互冲突，则无法同时启用这两条规则。例如，如果一个规则要求两个虚拟机始终在一起，而另一个规则要求这两个虚拟机始终分开，则无法同时启用这两个规则。您可使用这些规则确定以下事项：

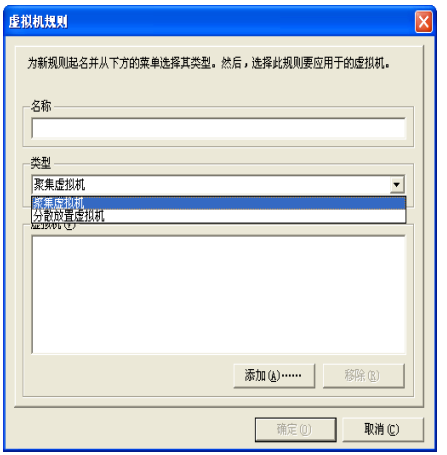
- DRS 应尝试将某些虚拟机一起保留在同一台主机上 （例如，出于性能考虑）。
- DRS 应尝试确保某些虚拟机没有放置在一起 （例如，为实现高可用性）。您可能希望保证某些虚拟机始终位于不同的物理主机上。当一台主机出现问题时，不会同时失去两台虚拟机。

注意 DRS 关联性规则与单个主机的 CPU 关联性规则完全不同。“使用 CPU 关联性向特定处理器分配虚拟机”（第 117 页）对 CPU 关联性规则进行了论述。

创建 DRS 规则

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 在 [群集设置 (Cluster Settings)] 对话框中选择 **[规则 (Rules)]**。



3 在 [虚拟机规则 (Virtual Machine Rule)] 对话框中命名该规则，以便对其进行查找并编辑。

4 从弹出菜单中选择一个选项：

- **[聚集虚拟机 (Keep Virtual Machines Together)]**

  一个虚拟机不能设置多条这样的规则。

- **[分散放置虚拟机 (Separate Virtual Machines)]**

  这种类型的规则不能包含两个以上 （不含两个）的虚拟机。

5 单击 **[添加 (Add)]** 添加虚拟机，然后在完成操作时单击 **[确定 (OK)]**。

添加规则之后，可以编辑它、查找冲突规则或删除它。

编辑现有规则

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 在左面板中选择 **[规则 (Rules)]** （在 **[VMware DRS]** 下）。

3 有关冲突规则等主题的详细信息，请单击 **[详细信息 (Details)]**。

4 在对话框中进行更改，然后在完成操作时单击 **[确定 (OK)]**。

了解规则结果

添加或编辑规则且群集立即与该规则发生冲突时，系统将继续运行并尝试更正该冲突。

对于默认自动化级别为手动或半自动的 DRS 群集，迁移建议将以规则实现和负载平衡为依据。您不一定要遵循规则，但在实现规则之前，相应的建议将一直保留。

禁用或删除规则

您可以禁用规则或将其完全删除。

禁用规则

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 在左面板中选择 **[规则 (Rules)]**（在 **[VMware DRS]** 下）。

3 取消选中规则左侧的复选框，然后单击 **[确定 (OK)]**。

稍后可以通过重新选中该复选框启用该规则。

删除规则

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 在左面板中选择 **[规则 (Rules)]**（在 **[VMware DRS]** 下）。

3 选择要移除的规则并单击 **[移除 (Remove)]**。

此时会删除该规则。

7 群集和虚拟机

本章说明如何添加、移除及定制虚拟机。

本章将讨论以下主题：

注意 所有任务均假定您有相应的操作权限。有关权限及如何设置权限的信息，请参见联机帮助。

将虚拟机添加到群集

可以通过向群集迁移虚拟机，或者向群集添加带有虚拟机的主机，在群集创建后向其中添加虚拟机。

在创建过程中添加虚拟机

创建虚拟机时，可以将其作为创建过程的一部分添加到群集中。当新建虚拟机向导提示您选择虚拟机位置时，您可选择一个独立主机或群集，并且可选择该主机或群集中的任意资源池。

有关部署虚拟机的详细信息，请参见 *《虚拟机管理指南》*。

将虚拟机迁移到群集

可以将现有的虚拟机从一个独立主机迁移到一个群集或者从一个群集迁移到另一个群集。虚拟机启动或关闭均可。用 VirtualCenter 移动虚拟机，有两种选择：

- 将虚拟机对象拖到群集对象上。
- 右键单击虚拟机名称并选择 **[迁移 (Migrate)]**。

对于 DRS 群集，系统会提示用户提供下面的信息：

- 群集本身或者群集内部资源池的位置。
- 启动及运行虚拟机的主机 （如果群集处于手动模式的话）。如果群集具有全自动或半自动的默认自动化级别，则 DRS 会选择主机。

注意 可以直接将虚拟机拖到群集内的资源池。在这种情况下，迁移向导会启动，但是资源池选择页不会出现。因为资源池控制资源，所以不允许直接向群集中的主机迁移。

向群集添加带有虚拟机的主机

如果将一个主机添加到一个群集，则该主机上的所有虚拟机均会添加到该群集。请参见“向 DRS 群集添加主机”（第 94 页）和 “将主机添加至 HA 群集”（第 110 页）。

启动群集中的虚拟机

当在属于某一群集的主机上启动虚拟机时，产生的 VirtualCenter 行为取决于群集的类型。

DRS 已启用

如果启动一个或一组虚拟机，并且启用了 DRS，则 VirtualCenter 首先执行接入控制。它会检查群集及资源池是否能为虚拟机提供足够的资源。如果群集没有足够的资源来启动单个虚拟机，或在组启动尝试中无法启动任何虚拟机，将会显示一条消息。

如果有足够资源可用，则 VirtualCenter 按如下方式进行操作：

- 如果将要进行的任何操作 （启动虚拟机、迁移虚拟机或启动主机）的自动化级别均为手动，则 VirtualCenter 会显示初始位置建议。请参见“初始放置位置”（第 60 页）。
- 如果所有的这些操作都是自动的，则 VirtualCenter 会将虚拟机放在最合适的主机上，而不给出建议。

HA 已启用

如果启动虚拟机并启用 HA，则 VirtualCenter 会执行 HA 接入控制。它会检查在启动虚拟机时是否有足够资源存在以允许执行指定数量的主机故障切换。

- 如果有足够的资源存在，则虚拟机会启动。
- 如果没有足够的资源存在，并且使用了严格接入控制 （默认），会有一条消息通知您虚拟机不能启动。如果未使用严格接入控制，则不会出现警告，虚拟机将直接启动。

从群集移除虚拟机

通过将虚拟机迁移出群集或将带有虚拟机的主机从群集移除，可以从群集移除虚拟机。

将虚拟机迁移出群集

您可以用下面两种方法之一，将虚拟机从群集迁移到一个独立主机或者从一个群集迁移到另一个群集：

- 使用标准的拖放方法。
- 从虚拟机右键菜单或 VirtualCenter 菜单栏中选择 **[迁移 (Migrate)]**。

如果虚拟机属于某一 DRS 群集关联性规则组 （请参见 “使用 DRS 关联性规则”（第 99 页）），则 VirtualCenter 会在允许迁移之前显示警告。该警告提示从属的虚拟机没有自动迁移。必须在迁移操作执行之前确认该警告。

从群集中移除带有虚拟机的主机

从群集中移除带有虚拟机的主机时，该主机中的所有虚拟机都将同时被移除。主机只有在维护模式或断开的情况下才可以被移除。请参见 “从群集移除主机”（第 95 页）。

注意 如果从 HA 群集中移除主机，群集可能会因没有足够的资源用于故障切换而变成红色。如果从 DRS 群集中移除主机，群集可能会因群集过量使用而变成黄色。请参见 “有效群集、黄色群集和红色群集”（第 76 页）。

对虚拟机进行 DRS 定制

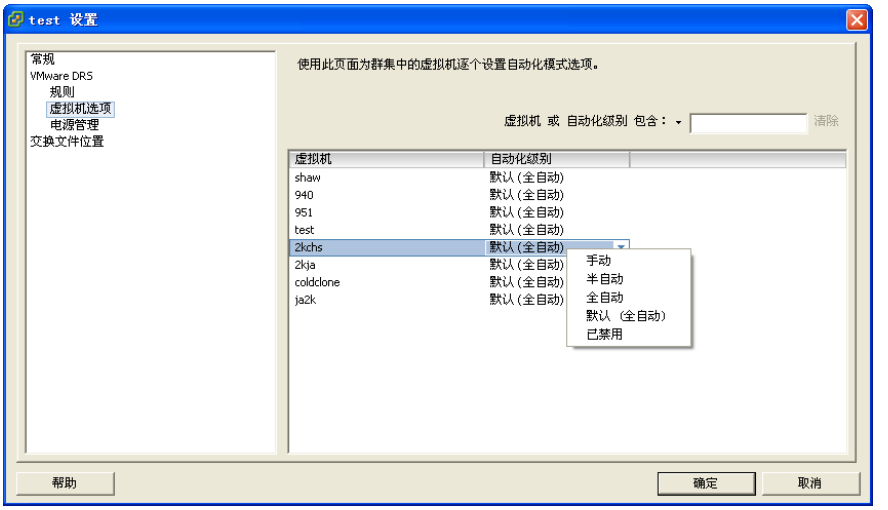
可以为 DRS 群集中的单个虚拟机定制自动模式，以重写该群集的默认自动化级别。例如，可以为完全自动的群集中的特定虚拟机选择 **[手动 (Manual)]**，或者为设置为 **[手动 (Manual)]** 的群集中的特定虚拟机选择 **[半自动 (Partially Automated)]**。

如果虚拟机已设置为 **[已禁用 (Disabled)]**，VirtualCenter 不会迁移该虚拟机或为其提供迁移建议。

为一个或多个虚拟机设置自定义自动模式

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 在 [群集设置 (Cluster Settings)] 对话框中，选择左边一列中的 **[虚拟机选项 (Virtual Machine Options)]**。



3 选择单个虚拟机，或者在按住 Shift 或 Ctrl 的同时选择多个虚拟机。

4 在右键菜单中，选择一种自动模式，然后单击 **[确定 (OK)]**。

对虚拟机进行 HA 定制

可以定制 HA 的重新启动优先级和隔离响应：

重新启动优先级。 确定在主机发生故障时重新启动虚拟机的顺序。重新启动优先级始终是需要考虑的因素，但在以下情况下尤为重要：

- 如果已将主机故障设置为特定的主机数 （例如 3 个），而发生故障的主机数大于该数 （例如 4 个）。
- 如果已关闭严格的接入控制，而且已启动的虚拟机数大于 HA 设置的可支持数量。

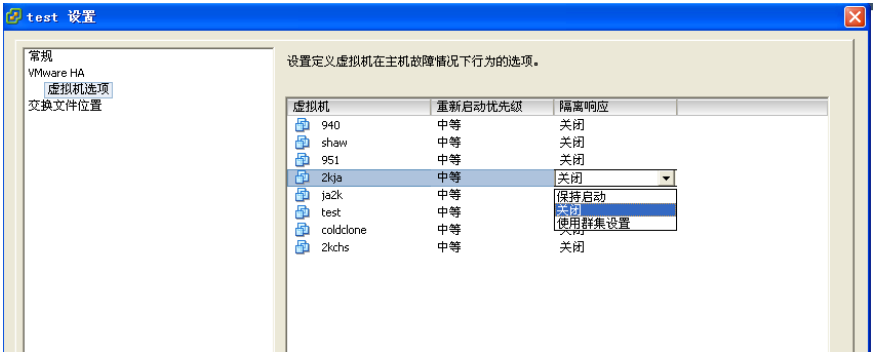
隔离响应。 确定当 HA 群集中的某个主机失去其控制台网络 （或 ESX Server 3i 中的 VMkernel 网络）连接但仍在运行时发生的情况。群集中的其他主机不再从此主机获得检测信号，声明该主机已停止运行，并尝试重新启动其虚拟机。磁盘锁定可以防止虚拟机的两个实例在两个不同的主机上运行。默认情况下，如果出现主机隔离的意外情况，隔离主机上的虚拟机将关闭，以便可以在另一个主机上重新启动虚拟机。您可更改各个虚拟机的这种行为。

如果使用 NAS 或 iSCSI 存储器，则当主机失去控制台网络连接（或 ESX Server 3i 中的 VMkernel 网络连接）时，虚拟机可能也无法访问其磁盘。在这种情况下，磁盘锁可能被破坏，虚拟机可以在其他主机上成功启动。由于失去了磁盘锁，因此隔离主机上的虚拟机可以继续运行，但是无法访问其磁盘 （即使它再次获得网络连接也是如此）。该虚拟机可能会创建并占用网络 I/O。对于 NAS 或 iSCSI 存储器上的虚拟机，VMware 建议将 **[隔离响应 (Isolation Response)]** 保持为 **[关闭 (Power off)]** （默认设置）。

为各个虚拟机定制 HA 行为

1 选择群集并从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 选择 **[VMware HA]** 下面的 **[虚拟机选项 (Virtual Machine Options)]**。



3 对于每个虚拟机，可以在 **[重新启动优先级 (Restart Priority)]** 或 **[隔离响应 (Isolation Response)]** 菜单中进行选择，以定制其设置。

- **[重新启动优先级 (Restart Priority)]** - 指定发生主机故障时重新启动虚拟机的相对优先级。优先级较高的虚拟机将首先启动。

  注意 仅按照主机来应用此优先级。如果多个主机发生故障， VirtualCenter 将首先按优先级顺序迁移第一个主机上的所有虚拟机，然后按优先级顺序迁移第二个主机上的所有虚拟机，依此类推。

- **[隔离响应 (Isolation Response)]** - 指定已与其群集失去连接的 ESX Server 主机应对正在运行的虚拟机执行的操作。默认情况下，每个虚拟机都被设置为在主机隔离意外情况发生时关机。

  可以选择 **[保持启动 (Leave powered on)]**，指示即使主机无法再与群集中的其他主机通信，隔离主机上的虚拟机仍应继续运行。如果虚拟机网络是在另一个稳定并有冗余的网络上，或者如果要让虚拟机一直运行，您就可以这样做。

注意 将主机添加到群集时，群集中的所有虚拟机默认为群集的默认重新启动优先级（未指定情况下为 **[中等 (Medium)]**）以及默认隔离响应 （未指定情况下为 **[关闭 (Power off)]**）。

8 管理 VMware HA

本章说明如何将主机添加至 HA 群集、如何移除主机以及如何定制 HA 群集。

本章将讨论以下主题：

注意 所述所有任务均假定您已获得许可并拥有执行这些任务的权限。有关权限的信息，请参见联机帮助。

定制 HA

创建群集后，可以为 DRS 和 / 或 HA 启用群集。然后就可以添加或移除主机，以及定制该群集。

可以按以下说明定制 HA：

- 创建群集时，接受或更改群集的默认重新启动优先级和隔离响应、选择群集的主机故障数并指明是否要执行严格的接入控制。请参见 ″选择 HA 选项″（第 88 页）。
- 按 ″将主机添加至 HA 群集″（第 110 页）中所述添加主机。
- 按 ″处理 VMware HA″（第 112 页）中所述更改现有群集的主机故障数或接入控制。

- 设置单个虚拟机的优先级。HA 使用虚拟机优先级来决定重新启动顺序，因此，在资源不足时，同一主机中优先级较高的虚拟机将优先获得资源。请参见 “对虚拟机进行 HA 定制”（第 106 页）。
- 设置单个虚拟机的隔离响应。默认情况下，如果主机与网络隔离，则关闭所有虚拟机。请参见 “对虚拟机进行 HA 定制”（第 106 页）。

将主机添加至 HA 群集

对于当前由同一 VirtualCenter Server 管理的主机 （受管主机）和当前未由该服务器管理的主机，将主机添加至群集的步骤有所不同。添加完主机后，部署至该主机的虚拟机将成为群集的一部分。

将受管主机添加至群集

VirtualCenter 清单面板显示由该 VirtualCenter Server 管理的所有群集和所有主机。有关将主机添加至 VirtualCenter Server 的信息，请参见 *《ESX Server 配置指南》*。

将受管主机添加至群集

1 从清单或列表视图中选择主机。

2 将主机拖至目标群集对象。

将主机添加至群集会在该主机上生成一个配置 HA (Configuring HA) 系统任务。完成此任务后，该主机就会包含在 HA 服务中。

将非受管主机添加至群集

可以添加当前并未由群集所在同一 VirtualCenter Server 管理的主机 （因此不可见）。

将非受管主机添加至群集

1 选择要添加主机的群集，并从右键菜单中选择 **[添加主机 (Add Host)]**。

2 提供主机名、用户名和密码，然后单击 **[下一步 (Next)]**。

此时主机即会添加至群集。将主机添加至群集会在该主机上生成一个配置 HA (Configuring HA) 系统任务。完成此任务后，该主机就会包含在 HA 服务中。

将主机添加至群集的结果

将主机添加至 HA 群集后：

- 该主机的资源会立即出现在群集的根资源池中，供群集使用。
- 除非群集还启用了 DRS，否则所有资源池均会折叠至群集的顶层 （不可见）资源池。

注意 资源池层次结构会丢失。之后将主机从群集移除时，该结构将不可用。

- 该主机上正在运行的每台虚拟机所需的容量或保证为其提供的容量之外的任何容量将用作该群集池中的备用容量。万一主机发生故障，可以使用此备用容量来启动其他主机上的虚拟机。
- 如果添加了一台带有多台正在运行的虚拟机的主机，群集由于添加了这台主机而不再能满足其故障切换要求，则会出现一个警告，并且该群集会被标成红色。
- 默认情况下，所添加主机上的所有虚拟机会获得群集默认重新启动优先级 （**[中等 (Medium)]**，如果未另行指定的话）和群集默认隔离响应 （**[关闭 (Power off)]**，如果未另行指定的话）。有关这些选项的信息，请参见 “选择 HA 选项”（第 88 页）。
- 系统还将监视每台主机上 HA 服务的状态，并在 [摘要 (Summary)] 页面显示有关配置问题的信息。
- 将主机从该群集中移除 （或者断开主机或将其置于维护模式）时，将取消 HA 服务的配置。您可能会在该主机上看到系统生成的取消配置 HA (Unconfiguring HA) 系统任务，该任务必须完成。

在主机上配置 HA 和取消 HA 的配置

将主机添加至 HA 群集时，会生成配置 HA (Configuring HA) 系统任务。必须成功完成该任务，才可将主机用于 HA。正在进行 HA 配置或取消 HA 的配置时，主机状况为黄色，而 [摘要 (Summary)] 页面会显示可能挂起的操作。

如果进行以下操作，则会为 HA 配置主机：

- 启用群集的 HA
- 连接 HA 群集中的主机
- 退出主机上的维护模式

如果进行以下操作，则会取消为 HA 配置主机：

- 禁用群集上的 HA
- 断开主机
- 进入主机上的维护模式

小心 将主机与 HA 群集断开时，将减少故障切换操作的可用资源。如果群集的故障切换容量小于或等于配置的故障切换容量，并且已开始断开主机的过程，则会收到一个群集故障切换警告。如果已完成断开，则该群集可能无法维持配置的故障切换级别。

可能会生成取消配置 HA (Unconfiguring HA) 系统任务。如果断开连接或进入维护模式，则取消配置操作将作为相应任务的一部分来完成，并且不会生成单独的系统任务。另外还将监视每台主机上的 HA 服务，并且主机的 [摘要 (Summary)] 页面会指示任何错误。该主机会被标记成红色。

配置任务或取消配置任务失败时，可在该任务的相关事件中获得详细信息。可能还需要检查该主机上的日志。如果修复了该错误，该主机将有一个重新配置 HA (Unconfiguring HA) 任务来重新配置发生故障的主机上的 HA。

注意 配置 HA 时，需要一台 DNS 服务器来解析主机名称。但是，配置以后，HA 会将名称解析缓存起来，不需要 DNS 查找来执行故障切换操作。

处理 VMware HA

重新配置 HA 可以是关闭 HA 或重新配置 HA 的选项。

关闭 HA

1 选择群集。

2 从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

3 在左面板中选择 **[常规 (General)]**，并取消选择 **[启用 VMware HA (Enable VMware HA)]** 复选框。

重新配置 HA

1 选择群集。

2 从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

3 在 [群集设置 (Cluster Settings)] 对话框中，选择 **[启用 VMware HA (Enable VMware HA)]**。

4 更改主机故障切换数或接入控制行为。请参见 “选择 HA 选项”（第 88 页）。

设置高级 HA 选项

本节指导您设置 HA 的高级属性，并列出了可能要设置的一些属性。因为这些属性会影响 HA 的运行，所以更改时请小心谨慎。

设置 HA 的高级选项

1 在群集的 [设置 (Settings)] 对话框中，选择 **[VMware HA]**。

2 单击 **[高级选项 (Advanced Options)]** 按钮以打开 [高级选项 (HA) (Advanced Options (HA))] 对话框。

3 输入要在 [选项 (Option)] 列中的文本框中更改的每个高级属性和应该在 [值 (Value)] 列中设置的值。

4 单击 **[确定 (OK)]**。

表 8-1. 高级 HA 属性

属性	描述
das.isolationaddress	设置需要 ping 的地址，以确定主机是否与网络隔离。如果未指定此选项，则使用控制台网络的默认网关。此默认网关必须是某个可用的可靠地址，以便主机可以确定它是否与网络隔离。可以为群集指定多个隔离地址 （最多 10 个）：`das.isolationaddress`*X*，其中 *X* = 1-10。
das.usedefaultisolationaddress	默认情况下， HA 使用控制台网络的默认网关作为隔离地址。该属性指定是否应该使用它 （`true|false`）。
das.defaultfailoverhost	如果设置了此属性， HA 会尝试将主机故障切换到此选项指定的主机。如果将某个主机用作备用故障切换主机，此属性非常有用，但是不建议这样做，因为 HA 会尝试利用群集内所有主机之间的所有可用备用容量。 如果指定的主机没有足够的备用容量， HA 会尝试将虚拟机故障切换到群集内有足够容量的任何其他主机上。
das.failuredetectiontime	更改默认故障检测时间 （默认为 15000 毫秒）。这是主机未从其他主机接收到任何检测信号时在声明其他主机停止运行前等待的时间段。
das.failuredetectioninverval	更改 HA 主机间的检测信号时间间隔。默认情况下，每秒钟发生一次。

表 8-1. 高级 HA 属性 （续）

属性	描述
das.vmMemoryMinMB	指定群集中任何虚拟机足够使用的最低内存量 （兆字节）。仅在没有为虚拟机指定内存保留量以及内存保留量用于 HA 接入控制和计算当前故障切换级别时，才使用此值。如果未指定任何值，则默认值为 256 MB。
das.vmCpuMinMHz	指定群集中任何虚拟机足够使用的最低 CPU 量 （兆赫兹）。仅在没有为虚拟机指定 CPU 保留量以及 CPU 保留量用于 HA 接入控制和计算当前故障切换级别时，才使用此值。如果未指定任何值，则默认值为 256 MHz。

9 高级资源管理

本章将讨论一些高级资源管理主题。其中包括概念信息和对可以设置的高级参数的讨论。大多数情况下，您不必使用高级设置，不正确地使用高级设置可能对系统性能不利。但是，经验丰富的管理员可能会发现这些高级配置选项有助于微调 ESX Server 环境的性能。

注意 对于本章讨论的任何主题和任务，均不需要 DRS 和 HA 的许可证。

本章将讨论以下主题：

CPU 虚拟

了解 CPU 相关问题以及仿真和虚拟之间的差异。

采用仿真时，所有操作均由仿真器在软件中执行。软件仿真器允许程序在不同于最初编写时所针对的计算机系统上运行。仿真器通过接受相同的数据或输入并获得相同的结果，来模拟或再现原始计算机的行为，从而实现仿真。仿真提供了可移植能力，通常用于在几个不同平台上运行针对一个平台设计的软件。

采用虚拟后，只要有可能就会使用基础物理资源，且虚拟层仅在需要时才执行指令，使得虚拟机就像直接运行在物理机上一样。虚拟着重于性能，只要有可能就会直接运行在处理器上。

软件 CPU 虚拟

采用软件 CPU 虚拟后，客户应用程序代码直接运行在处理器上，同时转换客户机特权代码且转换的代码在处理器上执行。转换后的代码大小稍微比转换前的代码大，这导致客户机执行速度降低。因此，具有少量特权代码组件的客户机程序运行速度非常接近本机，而具有大量特权代码组件的程序 （例如系统调用、陷阱或页面表更新）可能在虚拟环境中运行速度较慢。

硬件辅助的 CPU 虚拟

某些处理器 （例如 Intel VT 和 AMD SVM）为 CPU 虚拟提供了硬件辅助。使用此辅助时，客户机会获得独立的执行模式，称为客户机模式。应用程序代码或特权代码等客户机代码均在客户机模式中运行。出现某些事件时，处理器退出客户机模式，进入根模式。然后管理程序在根模式中执行，确定退出的原因，采取任何需要的措施，并在客户机模式中重新启动客户机。

将硬件辅助用于虚拟时，不需要再转换代码。因此系统调用或陷阱密集的工作负载在运行时的速度非常接近本机速度。但是，诸如涉及更新页面表之类的一些工作负载会导致多次退出客户机模式进入根模式。根据退出的次数和退出所用的总时间，这可能会明显降低执行的速度。

虚拟和特定处理器行为

因为 VMware 软件虚拟化 CPU，所以虚拟机能识别出它在其上运行的处理器的具体型号。一些操作系统安装有为特定处理器型号调整的不同内核版本，且这些内核也会安装到虚拟机中。由于内核版本不同，因此不可以将安装在运行一种处理器型号 （例如 AMD）的系统上的虚拟机迁移到运行在不同处理器 （例如 Intel）上的系统。

性能影响

根据工作负载和使用的虚拟类型， CPU 虚拟会添加不同的开销量。

如果应用程序的大多数时间用于执行指令而不是等待用户交互、设备输入或数据检索等外部事件，则应用程序是受 CPU 约束的。对于这些应用程序， CPU 虚拟开销需要执行其他指令，这会占用可能由应用程序本身使用的 CPU 处理时间。 CPU 虚拟开销通常会导致整体性能下降。

对于不受 CPU 约束的应用程序，CPU 虚拟可能会提高 CPU 利用率。如果备用 CPU 容量可用于吸收开销，则仍然可以在整体吞吐量方面提供不错的性能。

ESX Server 3 支持每个虚拟机最多可以有四个虚拟处理器 (CPU)。

注意 在单处理器虚拟机 （而不是 SMP 虚拟机）上部署单线程应用程序可获得最佳的性能和资源利用率。

单线程应用程序只能利用单个 CPU。在双处理器虚拟机中部署这些应用程序不会加快应用程序的速度。相反，这样会使得第二个虚拟 CPU 使用本该在其他方式下由其他虚拟机使用的物理资源。

使用 CPU 关联性向特定处理器分配虚拟机

关联性意味着可以在多处理器系统中限制将虚拟机分配给可用处理器的子集。通过为每个虚拟机指定关联性设置可以做到这点。

小心 使用关联性可能不合适。请参见 “使用关联性的潜在问题”（第 118 页）。

虚拟机的 CPU 关联性设置不仅应用到与虚拟机关联的所有虚拟 CPU，还会应用到与虚拟机关联的所有其他线程 （也称为 “环境”）。这些虚拟机线程执行仿真鼠标、键盘、屏幕、 CD-ROM 和其他老设备所需的处理。

一些情况下，像显示密集的工作负载，可能会在虚拟 CPU 和其他虚拟机线程之间出现大量通信。如果虚拟机的关联性设置阻止这些额外的线程同时用虚拟机的虚拟 CPU 调度 （例如，单处理器虚拟机与单个 CPU 关联，或 2 路 SMP 虚拟机仅与两个 CPU 关联），则性能可能会降低。

为了获得最佳性能，在使用手动关联性设置时， VMware 建议在关联性设置中至少包括一个额外的物理 CPU，以便允许至少有虚拟机的其中一个线程与其虚拟 CPU 同时调度 （例如，单处理器虚拟机至少与两个 CPU 关联或 2 路 SMP 虚拟机至少与三个 CPU 关联）。

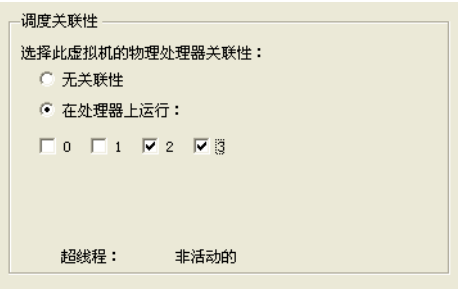
注意 CPU 关联性不同于 DRS 关联性，如 “对虚拟机进行 DRS 定制”（第 105 页）中所述。

向特定处理器分配虚拟机

1 在 VI Client 清单面板中，选择一个虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 选择 **[资源 (Resources)]** 选项卡，然后选择 **[CPU]**。

3 单击 **[在处理器上运行 (Run on processor(s))]** 按钮。



4 选择要在其上运行虚拟机的处理器，然后单击 **[确定 (OK)]**。

使用关联性的潜在问题 虚拟机关联性将每个虚拟机分配到指定关联性集合中的处理器。使用关联性之前，要考虑以下问题：

- 对于多处理器系统， ESX Server 系统执行自动负载平衡。避免手动指定虚拟机关联性，以改进调度程序跨处理器平衡负载的能力。
- 关联性可能会干扰 ESX Server 主机满足为虚拟机指定的预留量和份额的能力。
- 因为 CPU 接入控制不考虑关联性，所以具有手动关联性设置的虚拟机可能不会始终得到其完整的预留量。

  没有手动关联性设置的虚拟机不会受到具有手动关联性设置的虚拟机的负面影响。
- 将虚拟机从一个主机移动到另一个主机时，因为新的主机可能具有不同的处理器数，所以关联性可能不再适用。
- NUMA 调度程序可能不能管理已经使用关联性分配到某些处理器的虚拟机。有关搭配使用 NUMA 和 ESX Server 主机的更多信息，请参见第 10 章，“配合使用 NUMA 系统和 ESX Server”（第 141 页）。
- 关联性可能会影响 ESX Server 主机在多核或超线程处理器上调度虚拟机以充分利用在这些处理器上共享资源的能力。

多核处理器

Intel 和 AMD 均已开发了将两个或两个以上处理器内核组合到单个集成电路的处理器（通常称为*封装件* 或*插件*）。VMware 使用术语*物理处理器* 或*插件* 来描述单个封装件，该封装件可以具有一个或多个处理器内核且每个内核具有一个或多个*逻辑处理器*。多核处理器为执行虚拟机多任务的 ESX Server 主机提供了很多优势。

例如，与单核处理器相比，双核处理器通过允许同时执行两个虚拟机，可以提供几乎两倍的性能。每个内核均有自己的内存缓存，或可以与其他内核共享其部分缓存，潜在减少了缓存未中率和访问较慢主内存的必要性。如果运行在逻辑处理器上的虚拟机正运行竞争相同内存总线资源且占用大量内存的工作负载，则将物理处理器连接到主内存的共享内存总线可能会限制其逻辑处理器的性能。

ESX Server CPU 调度程序可以独立将每个处理器内核的每个逻辑处理器用于执行虚拟机，从而提供了与传统对称多处理 (Symmetric MultiProcessing, SMP) 系统类似的功能。例如， 2 路虚拟机可以让虚拟处理器运行在属于相同内核的逻辑处理器上，或运行在不同物理处理器的逻辑处理器上。

表 9-1 提供了处理器及其属性的列表。

注意 在具有 Intel 超线程技术的处理器上，每个内核可以具有两个逻辑处理器，这两个逻辑处理器共享大多数内核资源，例如内存缓存和执行管线。这些逻辑处理器通常称为*线程*。

表 9-1. 处理器和内核属性

处理器	内核	线程 / 内核	逻辑处理器
Intel Pentium III	1	1	1
Intel Pentium 4（禁用 HT）	1	1	1
Intel Pentium 4（启用 HT）	1	2	2
Intel Pentium D 940	2	1	2
Intel Pentium EE 840 （启用 HT）	2	2	4
Intel Core 2 Duo	2	1	2
Intel Core 2 Quad	4	1	4
AMD Athlon64	1	1	1
AMD Athlon64 X2	2	1	2

表 9-1. 处理器和内核属性 （续）

处理器	内核	线程 / 内核	逻辑处理器
AMD Opteron	1	1	1
AMD Opteron Dual Core	2	1	2

ESX Server CPU 调度程序知道处理器内核和其上逻辑处理器之间的处理器拓扑和关系。它使用此知识来调度虚拟机和优化性能。

超线程

Intel Corporation 开发了超线程技术来增强 Pentium IV 和 Xeon 处理器系列的性能。该技术允许单个处理器内核同时执行两个独立的线程。虽然此功能未提供真实双处理器系统的性能，但它可以改进芯片资源的利用率，使得某些重要的工作负载类型产生更大的吞吐量。

请参见 Intel 网站，了解有关超线程技术的深入讨论。

有关更多信息，请参见 VMware 网站上提供的白皮书 《ESX Server 2 对超线程的支持》。

超线程技术允许单个物理处理器内核像两个逻辑处理器一样工作。处理器可以同时运行两个独立的应用程序。为了避免混淆逻辑处理器和物理处理器， Intel 将物理处理器称为*插件*，本章的讨论也使用这一术语。

虽然超线程不会使系统的性能加倍，但是它可以通过更好地利用空闲资源来提高性能。如果应用程序运行在忙碌内核的一个逻辑处理器上，则与其单独运行在非超线程处理器上相比，预期获得的吞吐量会高出一半。但是，超线程性能改进情况与应用程序有很大关系，有些应用程序使用超线程可能会出现性能下降的情况，因为两个逻辑处理器之间会共享许多处理器资源 （例如缓存）。

启用超线程

默认情况下，启用超线程。如果禁用，可以启用。

具有 512K 二级缓存的所有 Intel Xeon MP 处理器和所有 Intel Xeon DP 处理器均支持超线程；但是，并非每个 Intel Xeon 系统均配有支持超线程的 BIOS。请查询系统文档，以查看 BIOS 是否包括对超线程的支持。 VMware ESX Server 无法在具有 16 个以上物理 CPU 的系统上启用超线程，因为 ESX Server 具有 32 个 CPU 的逻辑限制。

启用超线程

1 确保系统支持超线程技术。

2 在系统 BIOS 中启用超线程。
有些制造商将该选项标为 **[逻辑处理器 (Logical Processor)]**，而有些制造商则称之为 **[启用超线程 (Enable Hyperthreading)]**。

3 确保为 ESX Server 主机开启了超线程。

a 在 VI Client 中，选择主机，然后单击 **[配置 (Configuration)]** 选项卡。

b 选择 **[处理器 (Processors)]** 并单击 **[属性 (Properties)]**。

c 在该对话框中，可以查看超线程状态，还可以开启 （默认）或关闭超线程。

超线程和 ESX Server

启用超线程的 ESX Server 系统应像标准系统一样具有几乎相同的行为。相同内核上的逻辑处理器具有邻近的 CPU 编号，因此 CPU 0 和 1 一起在第一个内核上，而 CPU 2 和 3 在第二个内核上，依此类推。

VMware ESX Server 系统智能管理处理器时间，保证负荷均匀分布在系统的多个处理器内核上。优先在两个不同的内核上调度虚拟机，然后才选择在相同内核的两个逻辑处理器上调度虚拟机。

如果逻辑处理器没有工作，则将其置于*暂停*状况，释放其执行资源并允许运行在相同内核的另一个逻辑处理器上的虚拟机使用该内核的全部执行资源。 VMware 调度程序完全占用此暂停时间，因此使用内核全部资源运行的虚拟机效率要高于在半个内核上运行的虚拟机。按这种方法管理处理器可确保服务器不会违反任何标准的 ESX Server 资源分配规则。

超线程的高级服务器配置

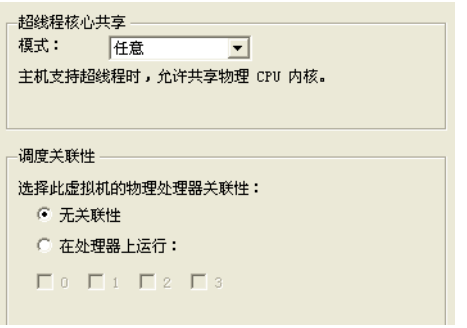
可以指定虚拟机的虚拟 CPU 如何在超线程系统上共享物理内核。如果两个虚拟 CPU 同时运行在内核的逻辑 CPU 上，则这两个虚拟 CPU 共享内核。可以为各个虚拟机设置此选项。

为虚拟机设置超线程共享选项

1 在 VI Client 清单面板中，右键单击虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 单击 **[资源 (Resources)]** 选项卡，然后单击 **[高级 CPU (Advanced CPU)]**。

3 从 **[模式 (Mode)]** 下拉菜单中进行选择，为该虚拟机指定超线程。



这时您有以下选择：

选项	描述
[任意 (Any)]	超线程系统上所有虚拟机的默认值。具有该设置的虚拟机的虚拟 CPU 可与该虚拟机或任何其他虚拟机的其他虚拟 CPU 随意共享内核。
[无 (None)]	虚拟机的虚拟 CPU 不应彼此共享内核，或不应与其他虚拟机的虚拟 CPU 共享内核。即，该虚拟机的每个虚拟 CPU 本身始终应获得完整的内核，而该内核上的另一个逻辑 CPU 则置于暂停状况。
[内部 (Internal)]	该选项类似于 **[无 (none)]**。不允许该虚拟机的虚拟 CPU 与其他虚拟机的虚拟 CPU 共享内核。这些虚拟 CPU 可以与同一虚拟机的其他虚拟 CPU 共享内核。 此选项仅可用于 SMP 虚拟机。如果应用于单处理器虚拟机，则系统将该选项更改为 **[无 (none)]**。

这些选项不会影响公平性或 CPU 时间分配。无论虚拟机的超线程设置如何，它仍然会得到与 CPU 份额成比例的 CPU 时间，且会受到 CPU 预留和 CPU 限制值的约束。

对于典型的工作负载，自定义超线程设置并非必要设置。对于与超线程交互不良的不常见工作负载，该选项很有用。例如，具有缓存颠簸问题的应用程序可能会让共享其物理

内核的应用程序降低速度。可以将运行该应用程序的虚拟机置于 **[无 (none)]** 或 **[内部 (internal)]** 超线程状态，将其与其他虚拟机隔离开。

如果虚拟 CPU 具有超线程限制，不允许该虚拟 CPU 与其他虚拟 CPU 共享内核，则当其他虚拟 CPU 有资格消耗处理器时间时，该系统可能取消对该虚拟 CPU 的调度。如果没有超线程限制，则两个虚拟 CPU 可能已经调度到了相同的内核上。

对于 （每个虚拟机）内核数有限的系统，问题会变得更糟。这些情况下，可能没有内核来让取消调度的虚拟机进行迁移。因此，超线程设置为 **[无 (none)]** 或 **[内部 (internal)]** 的虚拟机可能会降低性能，这一点对于内核数有限的系统而言尤其明显。

隔离

在某些很少的情况下， ESX Server 系统可能检测到应用程序与超线程技术交互不良。例如，对于与问题代码共享一个内核的应用程序，某些类型的自修改代码可能中断 Pentium IV 跟踪缓存的正常行为，导致速度大幅度降低 （最多 90%）。在这些情况下，ESX Server 主机隔离运行该代码的虚拟 CPU，并适当将其虚拟机置于 **[无 (none)]** 或 **[内部 (internal)]** 模式。必须隔离的情形很少，而且隔离对用户而言是透明的。

将主机的 **[Cpu.MachineClearThreshold]** 高级设置设置为 **[0]** 以禁用隔离。请参见 “设置高级主机属性”（第 134 页）。

超线程和 CPU 关联性

在使用超线程的系统上设置 CPU 关联性之前考虑您的情况。例如，如果高优先级虚拟机绑定到 CPU 0，而另一个高优先级虚拟机绑定到 CPU 1，则这两个虚拟机必须共享相同的物理内核。这种情况下，可能无法满足这些虚拟机的资源需求。确保超线程系统的任何自定义关联性设置有意义。在此示例中，将虚拟机绑定到 CPU 0 和 CPU 2 时，不应使用关联性设置。请参见 “使用 CPU 关联性向特定处理器分配虚拟机”（第 117 页）。

内存虚拟化

所有现代的操作系统均提供了对虚拟内存的支持，允许软件使用的内存要多于计算机实际拥有的内存。虚拟内存空间划分为块，通常每个块 4 KB，块也称为页。物理内存也划分为块，通常每个块也是 4 KB。当物理内存占满时，不在物理内存中的虚拟页的数据将存储到磁盘上。

软件内存虚拟化

ESX Server 通过添加附加级别的地址转换来虚拟化客户机物理内存。

- 每个虚拟机的 VMM 保持了从客户操作系统的物理内存页到基础计算机上物理内存页的映射。（VMware 将基础物理页称为计算机页，将客户操作系统的物理页称为物理页。）

  每个虚拟机均有连续的、基于零的、可寻址的物理内存空间。每个虚拟机使用的服务器上的基础计算机内存不一定是连续的。

- VMM 截取操作客户操作系统内存管理结构的虚拟机指令，因此虚拟机不会直接更新处理器上的实际内存管理单元 (Memory Management Unit, MMU)。

- ESX Server 主机在保持最新物理 - 计算机映射（由 VMM 保持，参见上文）的阴影页表格中保持虚拟 - 计算机页的映射。

- 阴影页表格直接由处理器的分页硬件使用。

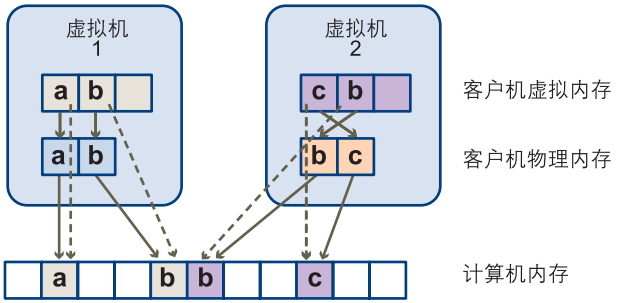
这种地址转换方法在设置阴影页表格之后，允许执行虚拟机中的正常内存访问，而不会添加地址转换开销。因为处理器上的转换后备缓冲区 (Translation Look-aside Buffer, TLB) 缓存从阴影页表格中读取的直接虚拟 - 计算机映射，所以 VMM 访问内存时不会添加额外开销。

硬件辅助的内存虚拟化

像 AMD SVM-V 等一些 CPU 通过使用两层页表，提供了对内存虚拟的硬件支持。第一层页表存储客户机虚拟 - 物理转换，而第二层页表存储客户机物理 - 计算机转换。如果 TLB 中没有某个客户机虚拟地址，则硬件会查看这两个页表，将客户机虚拟地址转换为主机物理地址。

图 9-1 中的插图说明了 ESX Server 如何实施内存虚拟。

图 9-1. ESX Server 内存映射



- 方框表示页，而箭头表示不同的内存映射。
- 从客户机虚拟内存到客户机物理内存的箭头表示客户操作系统中页表保持的映射。（插图未显示 x86 架构处理器从虚拟内存到线性内存的映射。）
- 从客户机物理内存到计算机内存的箭头表示 VMM 保持的映射。
- 虚线箭头表示也由 VMM 保持的阴影页表中从客户机虚拟内存到计算机内存的映射。运行虚拟机的基础处理器使用阴影页表映射。

因为虚拟引入了额外级别的内存映射，所以 ESX Server 可以高效管理所有虚拟机的内存。虚拟机的一些物理内存可能映射到共享页或未映射或换出的页面。

ESX Server 主机执行虚拟内存管理时无需了解客户操作系统，也不会干涉客户操作系统自身的内存管理子系统。

性能影响

本节讨论了基于软件和硬件辅助的内存虚拟的性能影响。

对于软件内存虚拟

使用两个页面整理的页表具有以下性能影响：

- 对于常规客户机内存访问不会产生开销。
- 在虚拟机内映射内存需要额外时间，这可能意味着：
  - 虚拟机操作系统正设置或更新虚拟地址到物理地址的映射。
  - 虚拟机操作系统从一个地址空间切换到另一个地址空间 （上下文切换）。
- 类似于 CPU 虚拟，内存虚拟开销取决于工作负载。

对于硬件辅助的内存虚拟

使用硬件辅助时，会消除软件内存虚拟的开销。特别是，硬件辅助消除了保持阴影页表与客户机页表同步所需的开销。但是，使用硬件辅助时 TLB 缺失等待时间明显较长。因此，工作负载使用硬件辅助受益与否主要取决于使用软件内存虚拟时内存虚拟引起的开销。如果工作负载涉及少量页表活动 （例如进程创建、映射内存或上下文切换），则软件虚拟不会引起大量开销。另一方面，具有大量页表活动的工作负载可能会因使用硬件辅助而受益。

了解内存开销

ESX Server 虚拟机可以引起两种内存开销：

- 在虚拟机内访问内存的额外时间。
- 超出向每个虚拟机分配的内存后， ESX Server 主机自身代码和数据结构所需的额外空间。

ESX Server 内存虚拟向内存访问添加很少的时间开销。因为处理器分页硬件直接使用阴影页表，所以虚拟机中的大多数内存访问在执行时没有地址转换开销。

例如，如果虚拟机中页面出错，则控制会切换到 VMM，以便 VMM 可以更新其数据结构。

内存空间开销有两部分：

- VMkernel 和 （仅用于 ESX Server 3）服务控制台的固定系统范围开销。
- 每个虚拟机的额外开销。

对于 ESX Server 3，服务控制台通常使用 272 MB，而 VMkernel 则使用更少的内存。使用的内存量取决于正使用的设备驱动程序的数量和大小。有关如何确定主机可用内存的信息，请参见 “查看主机资源信息”（第 14 页）。

开销内存包括为虚拟机框架缓冲区和各种虚拟数据结构预留的空间。开销内存取决于虚拟 CPU 数量、为客户操作系统配置的内存，以及使用的客户操作系统是 32 位还是 64 位。表 9-2 列出了每种情况的开销。

表 9-2. 虚拟机上的开销内存

虚拟 CPU	内存 (MB)	32 位虚拟机的开销 (MB)	64 位虚拟机的开销 (MB)
1	256	87.56	107.54
1	512	90.82	110.81
1	1,024	97.35	117.35
1	2,048	110.40	130.42
1	4,096	136.50	156.57
1	8,192	188.69	208.85
1	16,384	293.07	313.42
1	32,768	501.84	522.56
1	65,536	919.37	940.84
2	256	108.73	146.41

表 9-2. 虚拟机上的开销内存 （续）

虚拟 CPU	内存 (MB)	32 位虚拟机的开销 (MB)	64 位虚拟机的开销 (MB)
2	512	114.49	152.20
2	1,024	126.04	163.79
2	2,048	149.11	186.96
2	4,096	195.27	233.30
2	8,192	287.57	325.98
2	16,384	472.18	511.34
2	32,768	841.40	882.06
2	65,536	1,579.84	1,623.50
4	256	146.75	219.82
4	512	153.52	226.64
4	1,024	167.09	240.30
4	2,048	194.20	267.61
4	4,096	248.45	322.22
4	8,192	356.91	431.44
4	16,384	573.85	649.88
4	32,768	1,007.73	1,086.75
4	65,536	1,875.48	1,960.52

ESX Server 还提供了内存共享 （请参见 “在虚拟机之间共享内存”（第 133 页））等优化措施来减少基础服务器上使用的物理内存量。这些优化措施可以节省的内存多于开销占用的内存。

内存分配和闲置内存消耗

本节讨论了 ESX Server 主机如何分配内存以及您可以如何使用 **[Mem.IdleTax]** 配置参数来更改 ESX Server 主机回收闲置内存的方式。

ESX Server 主机如何分配内存

ESX Server 主机将 **[限制 (Limit)]** 参数指定的内存分配给每个虚拟机，除非内存过量使用。 ESX Server 主机向虚拟机分配的内存决不会超过指定的物理内存大小。例如，1 GB 虚拟机可能具有默认的限制 （无限）或用户指定的限制 （例如 2 GB）。在这两种

情况下， ESX Server 主机分配的内存决不会超过 1 GB，即不会超过为其指定的物理内存大小。

当内存过量使用时，向每个虚拟机分配的内存量介于 **[预留 (Reservation)]** 和 **[限制 (Limit)]** 指定的内存量之间 （请参见 “内存过量使用”（第 39 页））。在预留基础上分配给虚拟机的内存量通常会随着当前内存负载的变化而变化。

ESX Server 主机根据分配给虚拟机的份额数和对最近工作集大小的估计，确定每个虚拟机的分配量。

- **份额 -** ESX Server 主机使用改良的按比例分配内存策略。内存份额给予虚拟机一部分可用物理内存。请参见 “份额”（第 20 页）。
- **工作集大小 -** ESX Server 主机通过在连续的虚拟机执行时间周期监视内存活动，来估计工作集。采用快速响应工作集大小增加、较慢响应工作集大小减小的技术，在几个时间周期内进行平稳估计。

  该方法确保虚拟机开始更活跃地使用其内存时，已经回收闲置内存的虚拟机可以快速达到基于完整份额的分配量。

  通过调整 **[Mem.SamplePeriod]** 高级设置可以修改 60 秒的默认监视周期。**[Mem.SamplePeriod]** 指定虚拟机执行时间的周期时间间隔 （以秒为单位），在该执行时间内监视内存活动来估计工作集大小。请参见 “设置高级主机属性”（第 134 页）。

如何使用主机内存

可以使用 VI Client 查看如何使用主机内存。

查看有关物理内存使用情况的信息

1 在 VI Client 中，选择一个主机，然后单击 **[配置 (Configuration)]** 选项卡。

2 单击 **[内存 (Memory)]**。

此时会出现如表 9-3 中所述的以下信息。

表 9-3. 主机内存信息

字段	描述
总计	该主机的总物理内存。
系统	ESX Server 系统使用的内存。 ESX Server 3.x 至少使用 50 MB 系统内存用于 VMkernel，并使用额外内存用于设备驱动程序。 ESX Server 已加载且无法配置时，分配该内存。 虚拟层实际所需的内存取决于主机上外围组件互连 (Peripheral Component Interconnect, PCI) 设备的数量和类型。有些驱动程序需要 40 MB，几乎是基本系统内存的两倍。 ESX Server 主机还尝试一直保持一些内存可用，以便高效处理动态分配请求。 ESX Server 设置大约 6% 的可用内存来运行虚拟机。
虚拟机	在选定主机上运行的虚拟机使用的内存。 大多数主机内存用于运行虚拟机。 ESX Server 主机根据管理参数和系统负荷，管理向虚拟机分配此内存。
服务控制台	为服务控制台预留的内存。 单击 **[属性 (Properties)]** 以更改可用于服务控制台的内存量。该字段仅出现在 ESX Server 3 中。 ESX Server 3i 不提供服务控制台。

闲置虚拟机的内存消耗

如果虚拟机未活跃使用目前分配的内存，则 ESX Server 对闲置内存的消耗量大于对正在使用的内存的消耗量。（ESX Server 决不会改变用户指定的份额分配，但内存消耗却有类似的效果。）

内存消耗帮助虚拟机避免累积闲置内存。默认消耗率为 75%，即闲置页面的开销与四个活动页面的开销一样多。

[Mem.IdleTax] 高级设置允许您控制回收闲置内存的策略。使用该选项以及 **[Mem.SamplePeriod]** 高级属性来控制系统如何回收内存。请参见 “设置高级主机属性”（第 134 页）。

注意 大多数情况下，没有必要更改 **[Mem.IdleTax]**，而且在某些情况下不应进行更改。

ESX Server 主机如何回收内存

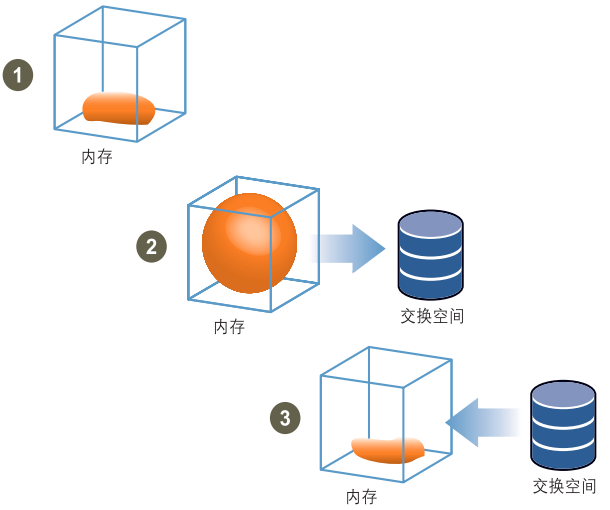
本节提供了 ESX Server 主机如何从虚拟机中回收内存的背景信息。主机使用两种技术来动态增加或减少分配给虚拟机的内存量：

- ESX Server 系统使用已加载到虚拟机中运行的客户操作系统的内存伸缩驱动程序 (vmmemctl)。请参见 “内存伸缩 (vmmemctl) 驱动程序”
- ESX Server 系统从虚拟机分页到服务器交换文件，无需客户操作系统参与。每个虚拟机均有自己的交换文件。请参见 “交换”（第 131 页）。

内存伸缩 (vmmemctl) 驱动程序

vmmemctl 驱动程序与服务器协作回收客户操作系统认为最不重要的页面。该驱动程序使用专用伸缩技术，提供了可预测的性能，在类似的内存约束下与本机系统的行为很相似。该技术可增加或减少客户操作系统的内存压力，使得客户机能够调用自己的本机内存管理算法。当内存很紧张时，客户操作系统决定要回收哪些页面，并在必要时将这些页面换到自己的虚拟磁盘上。请参见图 9-2。

图 9-2. 客户操作系统中的内存伸缩



注意 必须用足够的交换空间配置客户操作系统。一些客户操作系统具有其他限制。请参见 “交换空间和客户操作系统”（第 131 页）。

如有必要，通过为特定虚拟机设置 **[sched.mem.maxmemctl]** 参数，可以限制 vmmemctl 回收的内存量。该选项指定了可以从虚拟机中回收的最大内存量，以兆字节 (MegaByte, MB) 为单位。请参见 “设置高级虚拟机属性”（第 137 页）。

交换空间和客户操作系统

如果选择用 ESX Server 过量使用内存，需要确保客户操作系统具有足够的交换空间。该交换空间必须大于或等于虚拟机配置内存大小与其 **[预留 (Reservation)]** 之间的差值。

小心 如果内存过量使用且客户操作系统配置的交换空间不足，则虚拟机中的客户操作系统可能会出现故障。

要避免虚拟机故障，请增加虚拟机中交换空间的大小：

- **Windows 客户操作系统** - Windows 操作系统将其交换空间称为分页文件。如果有足够的可用磁盘空间，一些 Windows 操作系统会尝试增加分页文件的大小。

  请参见 Microsoft Windows 文档或搜索 Windows 帮助文件来了解 “分页文件”。按照说明更改虚拟内存分页文件的大小。

- **Linux 客户操作系统** - Linux 操作系统将其交换空间称为交换文件。有关增加交换文件的信息，请参见以下 Linux 手册页：

  - `mkswap` - 设置 Linux 交换区。
  - `swapon` - 启用设备和文件用于分页和交换。

具有大量内存和较小虚拟磁盘的客户操作系统 （例如，具有 8 GB RAM 和 2 GB 虚拟磁盘的虚拟机）更容易出现交换空间不足的情况。

交换

当虚拟机启动时， ESX Server 主机会创建交换文件。如果无法创建该文件，则无法启动虚拟机。默认情况下，在与虚拟机配置文件相同的位置中创建交换文件。除了接受此默认值以外，您还可以：

- 使用每个虚拟机配置选项将数据存储更改为另一个共享的存储位置。请参见 “设置高级虚拟机属性”（第 137 页）和表 9-7。
- 使用主机 - 本地交换，允许您在主机上指定本地存储的数据存储。这样就可以在每个主机级别进行交换，节省了 SAN 上的空间。但是对于 VMotion，可能会导致性能下降。

为群集启用主机 - 本地交换

1 右键单击 VI Client 清单面板中的群集，并单击 **[编辑设置 (Edit Settings)]**。

2 在出现的 [群集设置 (cluster Settings)] 对话框的左窗格中，单击 **[交换文件位置 (Swapfile Location)]**。

3 选择 **[将交换文件存储在主机指定的数据存储中 (Store the swapfile in the datastore specified by the host)]** 选项并单击 **[确定 (OK)]**。

4 在 VI Client 清单面板中选择其中一个群集的主机，然后单击 **[配置 (Configuration)]** 选项卡。

5 选择 **[虚拟机交换文件位置 (Virtual Machine Swapfile Location)]**。

6 单击 **[交换文件数据存储 (Swapfile Datastore)]** 选项卡，从提供的列表中选择要使用的本地数据存储，然后单击 **[确定 (OK)]**。

7 对群集中的每个主机重复步骤 4 到步骤 6。

为独立主机启用主机 - 本地交换

1 在 VI Client 清单面板中选择主机，然后单击 **[配置 (Configuration)]** 选项卡。

2 选择 **[虚拟机交换文件位置 (Virtual Machine Swapfile Location)]**。

3 在出现的 [虚拟机交换文件位置 (Virtual Machine Swapfile Location)] 对话框的 **[交换文件位置 (Swapfile location)]** 选项卡下，选择 **[将交换文件存储到交换文件数据存储中 (Store the swapfile in the swapfile datastore)]** 选项。

4 单击 **[交换文件数据存储 (Swapfile Datastore)]** 选项卡，从提供的列表中选择要使用的本地数据存储，然后单击 **[确定 (OK)]**。

当 `vmmemctl` 驱动程序由于以下原因而不可用时，ESX Server 主机会使用交换从虚拟机中强制回收内存：

- 未安装
- 已被明确禁用
- 未运行 （例如，客户操作系统正在引导时）
- 暂时无法以足够快的速度回收内存来满足当前系统需求
- 正常工作，但是已经达到最大伸缩大小。

当虚拟机需要页面时，标准需求分页技术会重新换入页面。

注意 为了获得最佳性能，只要有可能， ESX Server 主机就会使用伸缩方法 （通过 vmmemctl 驱动程序实施）。交换是只有必须回收内存时主机才会使用的最后可靠机制。

交换空间和内存过量使用

必须在 ESX Server 主机上为任何未预留的虚拟机内存预留交换空间，该交换空间是预留和配置内存大小之间的差值。需要该交换预留来确保系统在任何情况下均能保留虚拟机内存。实际上，只有一小部分交换空间可能会用到。

类似地，当内存预留用于接入控制时，实际内存分配会动态变化，不会浪费未用的预留。

交换文件和 ESX Server 故障

如果 ESX Server 系统发生故障，并且该系统有正在运行的、正在使用交换文件的虚拟机，则这些交换文件将继续存在并占用磁盘空间，即使 ESX Server 系统重新启动以后也是如此。

删除交换文件

1 再次启动虚拟机。

2 明确停止虚拟机。

注意 这些交换文件可能消耗数千兆字节的磁盘空间，因此请确保正确删除这些交换文件。

在虚拟机之间共享内存

许多 ESX Server 工作负载存在跨虚拟机共享内存的机会。例如，几个虚拟机可能正运行同一客户操作系统的实例，加载了相同的应用程序或组件，或包含公共数据。这些情况下， ESX Server 主机使用专用的透明页共享技术安全地消除内存页的冗余副本。采用内存共享，在虚拟机中运行的工作负载通常消耗的内存要少于其在物理机上运行时所需的内存。因此，可以有效地支持更高级别的过量使用。

使用 **[Mem.ShareScanTime]** 和 **[Mem.ShareScanGHz]** 高级设置来控制系统扫描内存以识别共享内存机会的速率。

通过将 **[sched.mem.pshare.enable]** 选项设置为 **[假 (FALSE)]**（该选项默认为 **[真 (TRUE)]**），还可以为个别虚拟机禁用共享。请参见 “设置高级虚拟机属性”（第 137 页）。

ESX Server 内存共享作为后台活动运行，随着时间的推移而扫描共享机会。节省的内存量随着时间而变化。对于相当固定的工作负载，在使用所有共享机会之前，内存量一般会缓慢增加。

要确定给定工作负载内存共享的有效性，请尝试运行工作负载，并使用 resxtop 或 esxtop 观察实际节省的内存量。在 [内存 (Memory)] 页面中交互模式的 PSHARE 字段中查找此信息。请参见 “以交互模式使用实用程序”（第 160 页）。

高级属性及其作用

本节列出了可用于定制内存管理的高级属性。

小心 只有在特殊情况下才适合使用这些高级属性。大多数情况下，更改基本设置（**[预留 (Reservation)]**、**[限制 (Limit)]**、**[份额 (Shares)]**）或使用默认设置可以获得适当的分配结果。

设置高级主机属性

本节指导您为主机设置高级属性，并列出了可能要在某些情况下设置的一些属性。

为主机设置高级属性

1 在 VI Client 清单面板中，选择要定制的虚拟机。

2 在 [命令 (Commands)] 面板中选择 **[编辑设置 (Edit Settings)]**，然后选择 **[选项 (Options)]** 选项卡。

3 选择 **[高级 (Advanced)]**，然后单击 **[配置参数 (Configuration Parameters)]** 按钮。

4 单击 **[高级设置 (Advanced Settings)]**。

5 在 [高级设置 (Advanced Settings)] 对话框中，选择合适的项 （例如 **[CPU]** 或 **[内存 (Memory)]**），并在右侧面板中滚动以查找和更改属性。

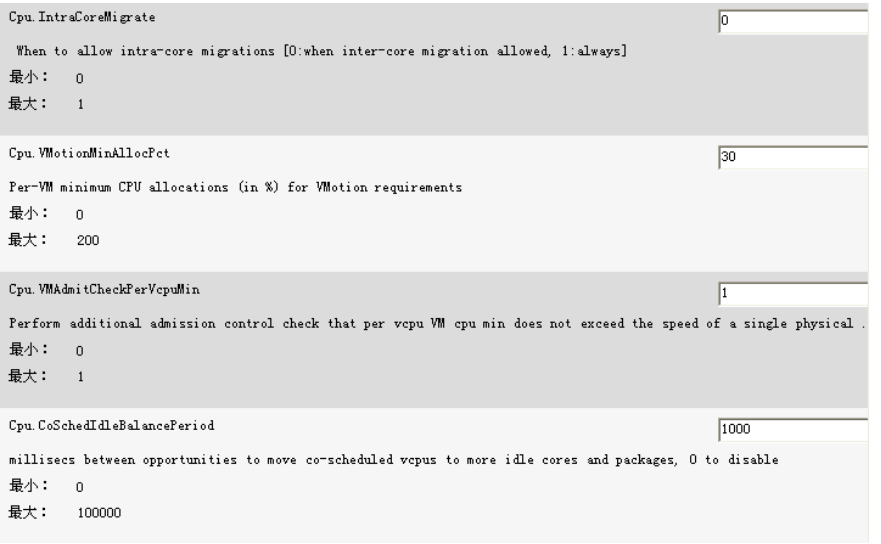


表 9-5、表 9-6 和表 9-7 列出了本文档中讨论的高级资源管理属性。

小心 建议只有具备 ESX Server 主机使用经验的高级用户才能设置这些属性。大多数情况下，使用默认设置即可获得最佳结果。

表 9-4. 高级 CPU 属性

属性	描述
CPU.MachineClearThreshold	如果使用启用了超线程的主机且将该属性设置为 **[0]**，则禁用隔离。请参见 “隔离”（第 123 页）。

表 9-5. 高级内存属性

属性	描述	默认值
Mem.CtlMaxPercent	根据配置内存大小的百分比，使用 `vmmemctl` 限制可以从任何虚拟机回收的最大内存量。为所有虚拟机指定 **[0]** 会通过 `vmmemctl` 禁用回收。	65
Mem.ShareScanTime	指定要扫描整个虚拟机以寻找页面共享机会所用的时间，以分钟为单位。默认为 60 分钟。	60

表 9-5. 高级内存属性 （续）

属性	描述	默认值
Mem.ShareScanGHz	为每个 GHz 的可用主机 CPU 资源指定要扫描 （每秒）页面共享机会的最大内存页面量。 默认为每 1 GHz 4 MB/s	4
Mem.IdleTax	指定闲置内存消耗率，以百分比为单位。虚拟机对闲置内存的消耗量大于对正在积极使用的内存的消耗量。 0% 的消耗率定义了忽略工作集并严格按照份额分配内存的分配策略。较高消耗率产生的分配策略允许要重新分配的闲置内存远离非生产性累积闲置内存的虚拟机。	75
Mem.SamplePeriod	指定虚拟机执行时间的周期时间间隔 （以秒为单位），在该执行时间内监视内存活动来估计工作集大小。	60
Mem.BalancePeriod	指定自动内存重新分配的周期时间间隔，以秒为单位。重新分配也可由可用内存量中的显著变化触发。	15
Mem.AllocGuestLargePage	将该选项设置为 1，让主机大页作为客户机的备用大页。在使用客户机大页的服务器工作负载中减少 TLB 缺失并改善性能。	1
Mem.AllocUsePSharePool 和 Mem.AllocUseGuestPool	将这些选项设置为 1 以减少内存碎片。如果主机内存有碎片，则主机大页的可用性会降低。这些选项可以提高让主机大页作为客户机备用大页的可能性。	1

表 9-6. 高级 NUMA 属性

属性	描述	默认值
Numa.RebalanceEnable	将该选项设置为 **[0]** 以禁用所有 NUMA 重新平衡和虚拟机的初始放置，从而有效禁用 NUMA 调度系统。	1
Numa.PageMigEnable	如果该选项设置为 **[0]**，则系统不会在节点间自动迁移页面以改善内存局域性。手动设置的页面迁移率仍然有效。	1
Numa.AutoMemAffinity	如果该选项设置为 **[0]**，则系统不会自动用 CPU 关联性集合来设置虚拟机的内存关联性。	1
Numa.MigImbalanceThreshold	NUMA 重新平衡器计算节点之间 CPU 的不平衡，考虑每个虚拟机的 CPU 时间权利与其实际消耗量之间的差值。该选项控制节点之间触发虚拟机迁移所需的最小负载不平衡，以百分比为单位。	10

表 9-6. 高级 NUMA 属性 （续）

属性	描述	默认值
Numa.RebalancePeriod	控制重新平衡周期的频率，以毫秒为单位指定。频繁的重新平衡会增加 CPU 开销，特别是在运行大量虚拟机的计算机上。频繁的重新平衡还可以提高公平性。	2000
Numa.RebalanceCoresTotal	在需要启用 NUMA 重新平衡器的主机上指定处理器内核的最小总数。	4
Numa.RebalanceCoresNode	在每个需要启用 NUMA 重新平衡器的节点上指定处理器内核的最小数量。 由于启用 NUMA 重新平衡时，少量处理器总数或每个节点的处理器可能会影响调度公平性，因此在小型 NUMA 配置 （例如， 2 路 Opteron 主机）上禁用 NUMA 重新平衡时，可以使用该选项和 **[Numa.RebalanceCoresTotal]**。	2

有关更多信息，请参见第 10 章，“配合使用 NUMA 系统和 ESX Server”（第 141 页）。

设置高级虚拟机属性

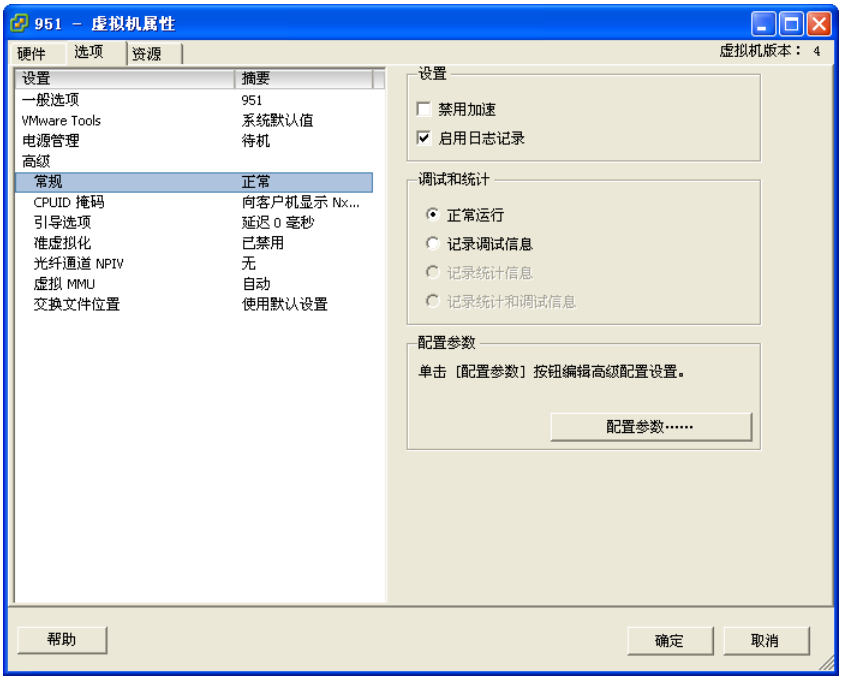
本节指导您为虚拟机设置高级属性，并列出了可能要设置的属性。

为虚拟机设置高级属性

1 在 VI Client 清单面板中选择虚拟机，然后从右键菜单中选择 **[编辑设置 (Edit Settings)]**。

2 单击 **[选项 (Options)]**，然后单击 **[高级 (Advanced)] > [常规 (General)]**。

3 单击 **[配置参数 (Configuration Parameters)]** 按钮。



4 在出现的对话框中，单击 **[添加行 (Add Row)]** 以输入新的参数及其值。

为虚拟机设置以下高级属性。

表 9-7. 高级虚拟机属性

属性	描述
sched.mem.maxmemctl	通过伸缩可以从选定虚拟机中回收的最大内存量，以兆字节 (MegaByte, MB) 为单位。如果 ESX Server 主机需要回收更多内存，则强制进行交换。交换的优先级低于伸缩。
sched.mem.pshare.enable	为选定的虚拟机启用内存共享。 该布尔值默认为 **True**。如果为虚拟机将其设置为 **False**，则关闭内存共享。
sched.swap.persist	指定关闭虚拟机时应保留还是删除虚拟机的交换文件。默认情况下，当虚拟机启动时系统为虚拟机创建交换文件，当虚拟机关闭时删除该交换文件。

表 9-7. 高级虚拟机属性 （续）

属性	描述
sched.swap.dir	虚拟机交换文件所在的 VMFS 目录。默认为虚拟机的工作目录，即包含其配置文件的 VMFS 目录。
sched.swap.file	虚拟机交换文件的文件名。默认情况下，系统在创建交换文件时会生成唯一的名称。

小心 如果修改 DRS 群集中某个虚拟机的 `sched.swap.dir` 属性，请确保群集中的每个主机均能访问您指定的交换文件位置，否则必须禁用该虚拟机的 DRS。

配合使用 NUMA 系统和 ESX Server

10

ESX Server 在支持非一致性内存访问 (Non-Uniform Memory Access, NUMA) 的服务器架构中，支持 Intel 和 AMD Opteron 处理器的内存访问优化。本章介绍了有关 NUMA 技术的背景信息，以及可用于 ESX Server 的优化。

本章将讨论以下主题：

NUMA 简介

NUMA 系统是具有多个系统总线的高级服务器平台。可以在单个系统映像中利用大量处理器，具有极高的性价比。提供 NUMA 平台以支持业界标准操作系统的系统包括基于 AMD CPU 或 IBM 企业 X 架构 (Enterprise X-Architecture) 的系统。

NUMA 的定义

在过去的十年中，处理器时钟速度获得了巨大的提升。但是，多 GHz 的 CPU 需要具备大量的内存带宽，才能有效利用其处理能力。即使是运行占用大量内存的工作负载（例如科学计算应用程序）的单个 CPU，也会受到内存带宽的限制。

在对称多处理 (Symmetric MultiProcessing, SMP) 系统上，这个问题会变得更加严重，因为许多处理器必须竞争同一系统总线上的带宽。一些高端系统通常通过构建高速数据总线来尝试解决这个问题。但是这种解决方案价格昂贵而且可扩展性也受到限制。

NUMA 是一种替代方法，它通过高性能连接将多个具有成本效益的小型节点连接起来。每个节点均包含处理器和内存，很像一个小型 SMP 系统。但是，高级内存控制器允许节点使用所有其他节点上的内存，从而创建了单个系统映像。当处理器访问不在自己节点内的内存 （远程内存）时，数据必须通过 NUMA 连接来传输，这种传输的速度比访问本地内存的速度慢。如这种技术的名称所示，内存访问的时间是不一致的，而且取决于内存的位置和通过其访问内存的节点。

NUMA 对操作系统的挑战

因为 NUMA 架构提供单个系统映像，所以通常可以运行没有经过专门优化的操作系统。例如， IBM x440 完全支持 Windows 2000，尽管 Windows 2000 并未针对与 NUMA 配合使用而设计。

在 NUMA 平台上使用这种操作系统有许多缺点。远程内存访问的滞后时间较长，会使处理器得不到充分利用，经常要等待数据传输到本地节点，而且 NUMA 连接会成为具有高内存带宽需求的应用程序的瓶颈。

而且，这种系统上的性能会有很大变化。例如，如果应用程序在一次基准运行时将内存放置在本地，但后来的一次运行碰巧将所有的这些内存放在远程节点上，此时性能就会发生变化。此现象会让容量规划变得困难。最后，多个节点之间的处理器时钟可能会不同步，因此直接读取时钟的应用程序可能会出现错误的行为。

一些高端 UNIX 系统支持在编译器和编程库中进行 NUMA 优化。此支持需要软件开发人员调整和重新编译他们的程序才能获得最佳的性能。针对一个系统进行的优化不能保证在下一代相同的系统上也能正常发挥作用。其他系统允许管理员明确决定运行应用程序的节点。对于要求其所有内存均必须是本地内存的某些应用程序，可能接受这种做法，不过当工作负载变化时会造成管理负担并且会导致节点之间不平衡。

理想情况下，系统软件提供了透明的 NUMA 支持，因此应用程序可以立即受益，无需进行修改。该系统应充分利用本地内存并且智能调度程序，不需要管理员经常干预。最后，该系统必须在不影响公平性或性能的情况下，对不断变化的状况作出良好的响应。

ESX Server NUMA 调度

ESX Server 使用复杂的 NUMA 调度程序来动态平衡处理器负载和内存局域性或处理器负载平衡，如下所示：

1 NUMA 调度程序管理的每个虚拟机均分配有主节点；主节点是系统的 NUMA 节点之一，其中包含处理器和本地内存，如系统资源分配表 (System Resource Allocation Table, SRAT) 所示。

2 将内存分配给虚拟机时， ESX Server 主机优先从主节点分配内存。

3 NUMA 调度程序可以动态更改虚拟机的主节点以响应系统负载的变化。该调度程序可能将虚拟机迁移到新的主节点，以减少处理器负载的不平衡。因为这可能会导致使用更多远程内存，所以调度程序可能将虚拟机的内存动态迁移到新的主节点，以改善内存局域性。在改善总体内存局域性的同时， NUMA 调度程序还可能在节点之间交换虚拟机。

一些虚拟机不受 ESX Server NUMA 调度程序管理。例如，如果为虚拟机手动设置处理器关联性， NUMA 调度程序可能无法管理该虚拟机。如果虚拟机上的虚拟处理器数量超过单个硬件节点上可用的物理处理器内核数，则无法自动管理该虚拟机。未受 NUMA 调度程序管理的虚拟机仍然可以正确运行。但是，这些虚拟机不能从 ESX Server 的 NUMA 优化中受益。

VMware ESX Server 中的 NUMA 调度和内存放置策略可以透明地管理所有虚拟机，因此管理员不需要明确处理在节点之间平衡虚拟机的复杂事情。

无论客户操作系统的类型如何，优化措施都可以顺利发挥作用。 ESX Server 甚至为不支持 NUMA 硬件的虚拟机 （例如 Windows NT 4.0）也提供了 NUMA 支持。因此，即使是使用老的操作系统，也可以利用新的硬件。

VMware NUMA 优化算法

本节介绍了 VMware ESX Server 在维持资源保证量的同时，用来充分提高应用程序性能的算法。

主节点和初始放置位置

当启动虚拟机时， ESX Server 会向其分配主节点。虚拟机仅运行在其主节点的处理器上，而且新分配的内存也来自该主节点。除非虚拟机的主节点更改，否则虚拟机仅使用本地内存，从而避免了与其他 NUMA 节点的远程内存访问关联的性能损失。

新的虚拟机最初以循环方式分配到主节点，第一个虚拟机分配到第一个节点，第二个虚拟机分配到第二个节点，以此类推。该策略确保在系统的所有节点上均匀地使用内存。

诸如 Windows Server 2003 之类的一些操作系统提供了这一级别的 NUMA 支持（称为初始放置位置）。对于仅运行单个工作负载 （例如基准配置，它不会在系统的正常运行时间过程中发生变化）的系统，这可能够用了。但是，初始放置位置还不够完善，不能保证预期支持工作负载变化的数据中心级系统的良好性能和公平性。

要了解仅采用初始放置位置的系统的缺点，请考虑以下示例：管理员启动四个虚拟机，系统将其中两个虚拟机置于第一个节点上。剩下的两个虚拟机置于第二个节点上。如果第二个节点上的两个虚拟机均停止，或如果这两个虚拟机均闲置，则系统将完全不平衡，全部负载都会置于第一个节点上。即使系统允许剩余的虚拟机中可以有一个虚拟机远程运行在第二个节点上，它也会因为所有的内存都保留在原始节点上而遭受严重的性能损失。

动态负载平衡和页迁移

ESX Server 结合了传统的初始放置位置方法和动态重新平衡算法。系统定期 （默认情况下每两秒一次）检查各个节点的负载，并且确定是否应通过将虚拟机从一个节点移至另一个节点来重新平衡负载。此计算考虑了虚拟机和资源池的资源设置，以便在不违反公平性或资源权利的情况下改善性能。

该重新平衡器选择合适的虚拟机，并将其主节点更改为负载最少的节点。如果可以的话，重新平衡器会移动目标节点上已经有一些内存的虚拟机。从此之后 （除非再次移动），虚拟机在新的主节点上分配内存并且仅运行在新的主节点内的处理器上。

重新平衡是维持公平性和确保完全使用所有节点的有效解决方案。重新平衡器可能需要将虚拟机移至已经分配少量内存或没有分配内存的节点上。这种情况下，虚拟机会遭受与大量远程内存访问关联的性能损失。 ESX Server 通过将内存从虚拟机的原始节点以透明的方式迁移到新的主节点，可以消除该损失：

1 系统选择原始节点上的页 （4 KB 连续内存），并将其数据复制到目标节点中的页上。

2 系统使用虚拟机监视层和处理器的内存管理硬件来无缝地重新映射虚拟机的内存视图，因此系统将目标节点上的页用于后续的所有引用，从而消除了远程内存访问的损失。

当虚拟机移至新的节点时， ESX Server 主机立即开始按此方式迁移其内存。主机会管理迁移速率，以避免让系统负担过重，特别是在虚拟机剩下很少的远程内存或目标节点

的可用内存很少时。如果虚拟机只是短时间内移至新的节点，则内存迁移算法还可以确保 ESX Server 主机不会无用地移动内存。

当初始放置位置、动态重新平衡和智能内存迁移配合使用时，即使工作负载出现变化，也能确保 NUMA 系统的良好内存性能。当主要工作负载出现变化时 （例如启动新的虚拟机时），系统会需要一些时间来重新调整，将虚拟机和内存迁移到新的位置。经过很短的时间之后 （通常是几秒钟或几分钟），系统就可以完成重新调整并达到稳定状况。

为 NUMA 优化的透明页共享

许多 ESX Server 工作负载存在跨虚拟机共享内存的机会。例如，几个虚拟机可能正运行同一客户操作系统的实例，加载了相同的应用程序或组件，或包含公共数据。这些情况下， ESX Server 系统使用专用的透明页共享技术安全地消除了内存页的冗余副本。采用内存共享，在虚拟机中运行的工作负载通常消耗的内存要少于其在物理机上运行时所需的内存。因此，可以有效地支持更高级别的过量使用。

ESX Server 系统的透明页共享也针对在 NUMA 系统上的使用而经过了优化。在 NUMA 系统上，页按照节点进行共享，因此每个 NUMA 节点都有自己的频繁共享页本地副本。当虚拟机使用共享页时，不需要访问远程内存。

手动 NUMA 控制

如果有一些使用大量内存的应用程序或者有少量的虚拟机，可能要通过明确指定虚拟机 CPU 和内存放置位置来优化性能。如果虚拟机运行占用大量内存的工作负载 （例如内存数据库或具有大型数据集的科学计算应用程序），这样做很有用。如果已知系统工作负载很简单而且不会变化，您可能还想手动优化 NUMA 放置位置。例如，对于一个由运行八个虚拟机而且具有类似工作负载的八个处理器组成的系统，很容易进行明确地优化。

注意 大多数情况下， ESX Server 主机的自动 NUMA 优化会产生良好的性能。

ESX Server 为 NUMA 放置位置提供了两组控制，因此管理员可以控制虚拟机的内存和处理器放置位置。

VI Client 允许您指定：

- **[CPU 关联性 (CPU Affinity)]** - 虚拟机应仅使用给定节点上的处理器。请参见 “将虚拟机与单个 NUMA 节点关联的 CPU 关联性”（第 147 页）。
- **[内存关联性 (Memory Affinity)]** - 服务器应仅在指定的节点上分配内存。请参见 “将内存分配与 NUMA 节点关联的内存关联性”（第 148 页）。

如果在虚拟机启动前设置了这两个选项，则虚拟机仅运行在选定的节点上并且在本地分配它的所有内存。

虚拟机已经开始运行后，管理员还可以手动将虚拟机移至另一个节点。这种情况下，还应手动设置虚拟机的页迁移速率，以便虚拟机前一个节点中的内存可以移至新的节点。

手动 NUMA 放置位置可能会干扰 ESX Server 资源管理算法，这种算法尝试向每个虚拟机赋予公平份额的系统处理器资源。例如，如果将具有占用大量处理器的工作负载的十个虚拟机手动置于一个节点，并且仅将两个虚拟机手动置于另一个节点，则系统不可能为所有的十二个虚拟机赋予相等份额的系统资源。在作出手动 NUMA 放置位置决定时，必须考虑这些问题。

IBM 企业 X 架构概述

IBM 企业 X 架构支持最多具有四个节点的服务器 （在 IBM 术语中也称为 CEC 或 SMP 扩展联合）。每个节点最多可以包含四个 Intel Xeon MP 处理器，共 16 个 CPU。下一代 IBM eServer x445 使用增强版本的企业 X 架构，并扩展为八个节点，每个节点最多四个 Xeon MP 处理器，共 32 个 CPU。第三代 IBM eServer x460 提供了类似的可扩展性，但另外还支持 64 位 Xeon MP 处理器。所有这些系统的高可扩展性均源于企业 X 架构的 NUMA 设计；基于 POWER4 的 IBM 高端 pSeries 服务器也采用了该设计。有关企业 X 架构的详细说明，请参见 IBM 红皮书 *《IBM eServer xSeries 440 计划和安装向导》*。

基于 AMD Opteron 的系统概述

诸如 HP ProLiant DL585 Server 之类基于 AMD Opteron 的系统也提供了 NUMA 支持。节点交叉的 BIOS 设置决定了系统行为更像 NUMA 系统还是更像统一内存架构 (Uniform Memory Architecture, UMA) 系统。请参见 《HP ProLiant DL585 Server 技术》技术摘要。另请参见惠普网站上的 *《基于 HP ROM 的安装实用程序用户向导》*。

默认情况下，禁用节点交叉，因此每个处理器都有自己的内存。 BIOS 生成系统资源分配表 (System Resource Allocation Table, SRAT)，因此 ESX Server 主机将系统作为 NUMA 来进行检测并应用 NUMA 优化。如果启用节点交叉 （也称为交叉内存），则 BIOS 不生成 SRAT，因此 ESX Server 主机不会将系统作为 NUMA 来进行检测。

当前随附的 Opteron 处理器的每个插件最多有四个内核。当启用节点内存时，会划分 Opteron 处理器上的内存，以便每个插件有一些本地内存，但其他插件的内存则是远程的。单内核 Opteron 系统的每个 NUMA 节点有单个处理器，而双内核 Opteron 系统的每个 NUMA 节点有两个处理器。

SMP 虚拟机 （有两个虚拟处理器）无法驻留在具有单个内核的 NUMA 节点内，例如单内核 Opteron 处理器。这也意味着 ESX Server NUMA 调度程序无法管理这些虚拟机。未受 NUMA 调度程序管理的虚拟机仍然可以正确运行。但是，这些虚拟机不会从

ESX Server NUMA 优化中受益。单处理器虚拟机 （具有单个虚拟处理器）可以驻留在单个 NUMA 节点内，并且由 ESX Server NUMA 调度程序进行管理。

注意 对于小型 Opteron 系统，现在默认禁用 NUMA 重新平衡，以确保调度的公平性。可以使用 **[Numa.RebalanceCoresTotal]** 和 **[Numa.RebalanceCoresNode]** 选项更改此行为。请参见 “设置高级虚拟机属性”（第 137 页）。

获得 NUMA 配置信息和统计信息

可以在 `resxtop` （或 `esxtop`）实用程序的 [内存 (Memory)] 面板中查看 NUMA 配置信息和统计信息。请参见 “内存面板”（第 166 页）。

将虚拟机与单个 NUMA 节点关联的 CPU 关联性

通过将虚拟机关联到单个 NUMA 节点上的 CPU 编号 （手动 CPU 关联性），可能会改善虚拟机上应用程序的性能。

小心 如果使用 CPU 关联性，会有许多潜在的问题。请参见 “使用关联性的潜在问题”（第 118 页）。

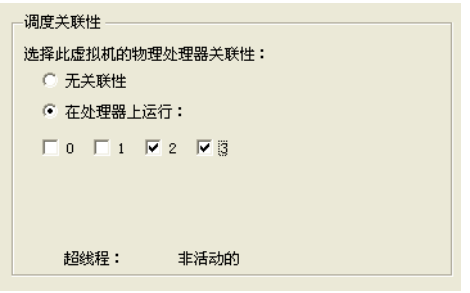
为单个 NUMA 节点设置 CPU 关联性

1 使用 VI Client，右键单击虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 在 [虚拟机属性 (Virtual Machine Properties)] 对话框中，选择 **[资源 (Resources)]** 选项卡并选择 **[高级 CPU (Advanced CPU)]**。

3 在 [调度关联性 (Scheduling Affinity)] 面板中，为不同的 NUMA 节点设置 CPU 关联性。

注意 必须为 NUMA 节点中的所有处理器手动选择这些框。 CPU 关联性是按照处理器指定的，而不是按照节点指定的。



将内存分配与 NUMA 节点关联的内存关联性

可以指定虚拟机上的所有后续内存分配使用与单个 NUMA 节点关联的页 （也称为手动内存关联性）。当虚拟机使用本地内存时，该虚拟机上的性能会得到改善。

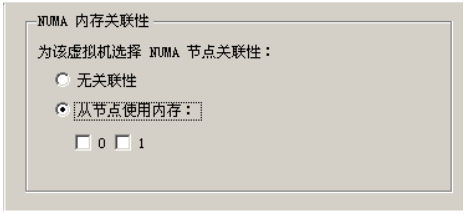
注意 只有在同时指定了 CPU 关联性时，才能指定要用于以后内存分配的节点。如果仅对内存关联性设置进行了手动更改，则自动 NUMA 重新平衡功能将无法正常工作。

将内存分配与某个 NUMA 节点相关联

1 使用 VI Client，右键单击虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 在 [虚拟机属性 (Virtual Machine Properties)] 对话框中，选择 **[资源 (Resources)]** 选项卡并选择 **[内存 (Memory)]**。

3 在 [NUMA 内存关联性 (NUMA Memory Affinity)] 面板中，设置内存关联性。



示例：将虚拟机绑定到单个 NUMA 节点 以下示例说明了将四个 CPU 绑定到 8 路服务器上的虚拟机的单个 NUMA 节点上。您想让该虚拟机仅运行在节点 1 上。

CPU （例如 4、 5、 6 和 7）是物理 CPU 编号。

绑定 2 路虚拟机以使用八处理器计算机的最后四个物理 CPU

1 在 VI Client 清单面板中，选择该虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 选择 **[选项 (Options)]** 并单击 **[高级 (Advanced)]**。

3 单击 **[配置参数 (Configuration Parameters)]** 按钮。

4 在 VI Client 中，为处理器 4、 5 和 6 启用 CPU 关联性。

设置虚拟机的内存以指定虚拟机的所有内存应分配在节点 1 上

1 在 VI Client 清单面板中，选择该虚拟机并选择 **[编辑设置 (Edit Settings)]**。

2 选择 **[选项 (Options)]** 并单击 **[高级 (Advanced)]**。

3 单击 **[配置参数 (Configuration Parameters)]** 按钮。

4 在 VI Client 中，将 NUMA 节点的内存关联性设置为 1。

完成这两个任务可以确保虚拟机仅运行在 NUMA 节点 1 上，并在可能的情况下从同一个节点分配内存。

11 最佳做法

本章将讨论一些适合 ESX Server 和 VirtualCenter 用户的最佳做法。

本章将讨论以下主题：

资源管理最佳做法

遵循以下准则有助于使虚拟机获得最佳性能：

- 如需频繁更改总可用资源，可使用 **[份额 (Shares)]** 合理分配虚拟机资源。例如，如果使用 **[份额 (Shares)]**，并且升级主机，则即使每个份额代表较大的内存量或 CPU 量，每个虚拟机也保持相同的优先级 （保持相同数量的份额）。
- 使用 **[预留 (Reservation)]** 来指定可接受的最低 CPU 量或内存量而不是想要使用的量。主机可以根据份额的数量和虚拟机的限制将额外的资源指定为可用资源。预留值表示的具体资源量不会随环境改变 （例如添加或移除虚拟机）而变化。
- 不要将 **[预留 (Reservation)]** 设置得太高。预留值设置得太高会限制资源池中的虚拟机数量。
- 不要将所有资源全部指定为虚拟机的预留值。系统容量越接近于被全部预留，想要在不违反接入控制的条件下更改预留值和资源池层次结构就越困难。在一个启用 DRS 的群集中，预留值占用群集或群集中单个主机的全部容量会阻止 DRS 在主机之间迁移虚拟机。

- 使用资源池进行委派资源管理。要完全隔离一个资源池，需将资源池类型设置为 **[固定的 (Fixed)]** 并使用 **[预留 (Reservation)]** 及 **[限制 (Limit)]**。
- 为资源池中的多层服务进行虚拟机分组。资源池允许 ESX Server 主机将服务资源作为一个整体来进行分配。

创建和部署虚拟机

本节提供规划及创建虚拟机最佳做法的信息。

规划

在部署虚拟机之前，需要：

- 规划负载的构成。
- 了解目标及预期效果。
- 了解要求及其对成功实现目标的重要性。
- 避免混合使用争夺资源要求的虚拟机。
- 如果有特定的性能期望，应在部署之前进行测试。

虚拟化可以使许多虚拟机共享主机资源。它并不会创建新的资源。虚拟化会产生开销。

创建虚拟机

创建虚拟机时，请务必像物理机一样根据实际需要确定其大小。虚拟机配置过高会浪费可共享资源。

为优化性能，应禁用不用的虚拟设备，例如 COM 端口、 LPT 端口、软盘驱动器、CD-ROM、 USB 适配器等等。即使在不使用的情况下，客户操作系统仍会定期轮询这些设备。这种无用的轮询会浪费可共享资源。

安装 VMware Tools，该工具有助于获得较高的性能，提高 CPU 使用效率，而且包括磁盘、网络和内存回收驱动程序。

部署客户操作系统

使用注册表、交换空间等等，像调整物理机操作系统一样调整虚拟机操作系统。禁用不必要的程序和服务，例如屏幕保护程序。不必要的程序和服务会浪费可共享资源。

确保客户操作系统及时更新最新的修补程序。如果使用 Microsoft Windows 作为客户操作系统，可在 Microsoft 知识库文章中查找任何已知的操作系统问题。

注意 必须用足够的交换空间配置客户操作系统。一些客户操作系统具有其他限制。请参见 “交换空间和客户操作系统”（第 131 页）。

部署客户应用程序

请按照在物理机上调整应用程序的相同方法在虚拟机上调整应用程序。

不要在 SMP 虚拟机上运行单线程应用程序。单线程工作负载无法利用额外的虚拟 CPU，未使用的虚拟 CPU 会浪费可共享资源。但是，如果一个工作负载包含若干同时运行的单线程应用程序，也许可以利用额外的虚拟 CPU。

配置 VMkernel 内存

VMkernel 通过伸缩和交换回收内存。请参见第 9 章，“高级资源管理”（第 115 页）。要以最佳方式使用内存资源，请通过正确设置虚拟机大小和避免内存过量使用，来避免频繁的回收活动。请参见 “内存过量使用”（第 39 页）。

VMkernel 会执行 NUMA 调度程序，该调度程序支持 IBM 及 AMD NUMA 架构。该调度程序可以在同一个 NUMA 节点上设置虚拟机内存和虚拟 CPU。这样可以防止因远程访问内存而导致的性能下降。主机硬件应配置为主机物理内存在各 NUMA 节点上均衡分布。请参见第 10 章，“配合使用 NUMA 系统和 ESX Server”（第 141 页）。

VMware HA 最佳做法

使用以下适用于 ESX Server 实施以及网络架构的 VMware HA 最佳做法。

网络最佳做法

如何配置 ESX Server 主机网络和名称解析，以及 ESX Server 主机外部的网络基础架构（交换机、路由器和防火墙）对优化 VMware HA 设置至关重要。在配置这些组件时，请使用以下最佳做法来提高 VMware HA 性能。

- 如果您的交换机支持 *PortFast*（或等效）设置，请在连接服务器的物理网络交换机上启用该设置。这样有助于避免发生跨树隔离事件。有关该选项的详细信息，请参见您的网络交换机供应商提供的文档。
- 请务必通过服务控制台对所有 ESX Server 3 主机开放以下防火墙端口来进行通信：
  入站端口 TCP/UDP 8042-8045
  出站端口 TCP/UDP 2050-2250

- 在服务器间配置端到端双重网络路径用于服务控制台网络，以获得更好的检测信号可靠性。请在群集中服务器之间配置较短的网络路径。跃点过多的路由会导致检测信号的网络数据包延迟。
- 在执行任何网络维护操作 （该操作可能会禁用主机之间的所有检测信号路径）时，请禁用 VMware HA （使用 VirtualCenter，在群集 [设置 (Settings)] 对话框中清除 **[启用 VMware HA (Enable VMware HA)]** 复选框）。
- 使用 DNS 进行名称解析，而不使用在 ESX Server 主机上手动编辑本地 /etc/hosts 文件的方法 （这种方法容易出错）。如果确实要编辑 /etc/hosts，必须同时包含长名称和短名称。
- 在 VLAN 上使用一致的公用网络端口名称。端口名称用于重新配置虚拟机对网络的访问。如果在原始服务器和故障切换服务器间使用的名称不一致，虚拟机将在故障切换后断开网络连接。
- 在一个 VMware HA 群集的所有服务器上使用有效的虚拟机网络标签。虚拟机使用这些标签在重新启动后重新建立网络连接。

设置网络冗余

群集节点之间的网络冗余对 VMware HA 可靠性非常重要。ESX Server 3 上的冗余服务控制台网络 （或 ESX Server 3i 上的 VMkernel 网络）可以可靠地检测故障并防止发生隔离的情况，因为检测信号可以通过多个网络发送。

可以在网卡级别或服务控制台 /VMkernel 端口级别实现网络冗余。在大多数实现中，网卡成组可以提供足够的冗余，但如果需要额外的冗余，可以使用或增加服务控制台 / 端口冗余。

网卡成组

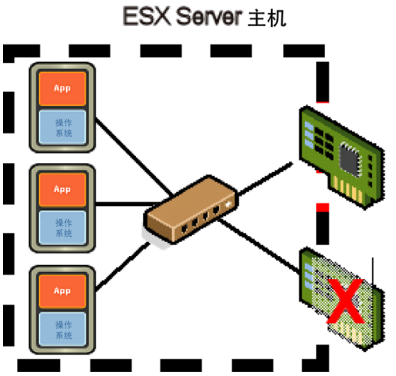
如图 11-1 所示，使用两个网卡构成网卡组隔离物理交换机可以提高服务控制台 （或 ESX Server 3i 中的 VMkernel）网络的可靠性。因为通过两个网卡 （并且通过单独的交换机）连接的服务器具有两条独立的路径来发送和接收检测信号，所以群集具有更好的弹性。

要为服务控制台配置网卡组，请在活动 / 待机配置的 vSwitch 配置中配置 vNIC。建议的 vNIC 参数设置如下：

- [滚动故障切换 (Rolling Failover)] = [是 (Yes)]
- 默认的 [负载平衡 (Load balancing)] = [基于源端口 ID 的路由 (route based on originating port ID)]

注意 在为 VMware HA 群集中的一个主机添加网卡以后，必须在该主机上重新配置 HA。

图 11-1. 使用网卡成组的服务控制台冗余



网卡成组方案 下面的方案说明使用带有网卡成组的单个服务控制台网络实现网络冗余的情况：

- 在只有一个服务控制台网络 （子网 10.20.XX.XX）的群集中配置主机存在一定风险。可以使用两个成组网卡以防网卡发生故障。
- 将默认超时增加到 60 秒 (`[das.failuredetectiontime]` = 60000)。

次要服务控制台网络

除了使用网卡成组提供检测信号冗余之外，还可以创建一个次要服务控制台 （或 ESX Server 3i 的 VMkernel 端口），并将其连接到一个单独的虚拟交换机上。主服务控制台仍用于网络和管理。次要服务控制台创建之后，VMware HA 会同时通过主要和次要服务控制台发送检测信号。如果一条路径发生故障，VMware HA 仍可通过另一条路径发送和接收检测信号。

默认情况下，每个 ESX Server 主机的服务控制台网络配置中指定的网关 IP 地址用作隔离地址。每个服务控制台网络都必须有一个可以到达的隔离地址。设置服务控制台冗余时，必须为次要服务控制台网络指定额外的隔离响应地址 (`[das.isolationaddress2]`)。此隔离地址的网络跃点数应尽量少。指定次要隔离地

址时，VMware 建议将 [das.failuredetectiontime] 设置增加到 20000 毫秒或更长时间。请参见 “设置高级 HA 选项”（第 113 页）。

通过为 VMotion vswitch 添加次要服务控制台网络，可以进一步优化网络 （如果您已经配置了 VMotion 网络）。如图 11-2 所示， VMotion 网络和次要服务控制台网络可以共享虚拟交换机。

图 11-2. 带有次要服务控制台的网络冗余

冗余服务控制台网络方案 下面的方案说明使用冗余服务控制台网络的情况：

- 利用一个现有的 VMotion 网络 （子网 10.20.YY.YY 和 192.168.ZZ.ZZ），为群集中的每个主机配置两个服务控制台网络。
- 将默认网关用于第一个网络并且指定 [das.isolationaddress2] = 192.168.1.103 作为额外的隔离地址用于第二个网络。
- 将默认超时增加到 20 秒 ([das.failuredetectiontime] = 20000)。

其他 VMware HA 群集注意事项

优化 VMware HA 群集性能的其他注意事项包括：

- 在启用 VMware HA 的群集中，使用较大规模的同类服务器组以提高使用水平（平均情况）。
  - 增加每个群集中的节点数可以在保证故障切换能力的情况下提高主机故障的容错能力。

- 由于适当利用了接入控制试探法，因此带有许多虚拟机的大型服务器可以在发生故障时切换到较小的服务器。

- 要定义用于接入控制的大小估计值，请为所需的最小资源设置合理的预留值。
  - 如果不设置预留值，则接入控制将超过故障切换能力；否则，VMware HA 将使用指定为 "slot" 大小的最大预留值 （请参见 "规划 HA 群集"（第 71 页））。
  - 将一些虚拟机的平均预留值设置为最小值。
- 通过选择 **[即使虚拟机违反可用性限制也允许启动虚拟机 (Allow virtual machines to be powered on even if they violate availability constraints)]**，可以执行您自己的容量规划。在主机和虚拟机大小相差悬殊的情况下，接入控制可能过于保守。VMware HA 仍然会尝试重新启动尽可能多的虚拟机。

性能监视实用程序：resxtop 和 esxtop

A

通过 resxtop 和 esxtop 命令行实用程序，您可以实时详细查看 ESX Server 使用资源的情况。可以按以下三种模式之一启动任一实用程序：交互 （默认）、批处理或重放。本附录说明了如何在这些模式中使用 resxtop 和 esxtop，并对可用的命令给出了参考信息，另外还显示了统计信息。除非另有指定，否则这两个实用程序的命令和统计信息均相同。本附录讨论了以下主题：

决定使用 resxtop 或 esxtop

resxtop 和 esxtop 的基本区别是可以远程 （或本地）使用 resxtop，而 esxtop 只能通过本地 ESX Server 主机的服务控制台来启动。

使用 resxtop 实用程序

resxtop 实用程序是远程命令行界面 (Remote Command Line Interface, Remote CLI) 命令，在使用任何 Remote CLI 命令前，必须下载、安装和配置 Remote CLI 虚拟设备。请参见 *《ESX Server 3i 版本 3.5 配置指南》*。

设置虚拟设备后，从远程 Linux 客户端的命令行启动 resxtop，它具有与 esxtop 相同的命令行选项。因此您正在使用的客户端可以连接远程服务器并且由远程服务器进行身份验证，您可以使用以下其他选项：

[server] - 要连接的远程服务器主机的名称 （必需）。

[portnumber] - 要连接的远程服务器上的端口号。默认端口为 443，除非在服务器上更改了这一端口，否则不需要此选项。

[username] - 连接远程主机时要进行身份验证的用户名。远程服务器还会提示您输入密码。

注意 resxtop 不使用与其他 Remote CLI 命令共享的所有选项。

您还可以在本地 ESX Server 主机上使用 resxtop。为此，在命令行上省略 server 选项，该命令将会默认为 localhost。

使用 esxtop 实用程序

esxtop 实用程序仅运行在 ESX Server 主机的服务控制台上。

启动 esxtop

1 确保您具有超级用户特权。

2 使用所需的选项，键入命令：

```
esxtop [–] [h] [v] [b] [s] [a] [c filename] [R  vm–support_dir_path]
               [d delay] [n iter]
```

esxtop 实用程序从 .esxtop310rc 读取其默认配置。该配置文件由七行组成。

前六行包含小写字母和大写字母，指定在 CPU、内存、存储适配器、存储设备、虚拟机存储器和网络面板上以什么顺序显示哪些字段。这些字母对应于各个 esxtop 面板的 [字段 (Fields)] 或 [顺序 (Order)] 面板中的字母。

第七行包含了有关其他选项的信息。最重要的是，如果以安全模式保存了配置，则不从 .esxtop310rc 文件的第七行移除 s，也不会获得不安全的 esxtop。用一个数字指定更新之间的延迟时间。与交互模式相同，键入 c、m、d、u、v 或 n 决定 esxtop 启动时的面板。

注意不建议编辑该文件。请在运行的 esxtop 进程中选择这些字段和顺序，进行更改，并使用 **W** 交互命令保存该文件。

以交互模式使用实用程序

默认情况下， resxtop 和 esxtop 以交互模式运行。交互模式在不同的面板中显示统计信息。

每个面板均可以使用帮助菜单。

交互模式命令行选项

在交互模式中可以使用表 A-1 中列出的命令行选项。

表 A-1. 交互模式命令行选项

选项	描述
h	显示 resxtop （或 esxtop）命令行选项的帮助。
v	显示 resxtop （或 esxtop）版本号。
s	以安全模式调用 resxtop （或 esxtop）。在安全模式中，禁用了指定更新之间延迟的 –d 命令。
d	指定更新之间的延迟。默认为 5 秒。最小值为 2 秒。可以使用交互命令 s 更改此命令。如果指定的延迟少于 2 秒，延迟将设置为 2 秒。
n	迭代次数。更新显示 n 次，然后退出。
服务器	要连接的远程服务器主机的名称 （仅 resxtop 需要）。
portnumber	要连接的远程服务器上的端口号。默认端口为 443，除非在服务器上更改了这一端口，否则不需要此选项。（仅 resxtop）
username	连接远程主机时要进行身份验证的用户名。远程服务器还会提示您输入密码 （仅 resxtop）。
a	显示所有统计信息。该选项会替代配置文件设置并显示所有统计信息。配置文件可以是默认的 ~/.esxtop310rc 配置文件或用户定义的配置文件。
c <filename>	加载用户定义的配置文件。如果未使用 -c 选项，则默认配置文件名为 ~/.esxtop310rc。使用 W 单键交互命令创建自己的配置文件，同时指定不同的文件名。有关 W 的信息，请参见 “交互模式单键命令”（第 162 页）。

公共统计信息描述

当 resxtop （或 esxtop）以交互模式运行时，不同的面板上会显示一些统计信息。以下统计信息是所有四个面板的公共信息。

四个 resxtop （或 esxtop）面板的顶部显示的 **[正常运行时间 (Uptime)]** 行显示了当前时间、自上一次重新引导以来所经过的时间、当前运行的环境数量和平均负载。环境是 ESX Server VMkernel 可调度的实体，类似于其他操作系统中的进程或线程。

其下显示的是过去 1 分钟、5 分钟和 15 分钟内的平均负载。平均负载同时考虑了正在运行和准备运行的环境。 1.00 的平均负载表示完全利用了所有物理 CPU。 2.00 的平均负载表示 ESX Server 系统可能需要当前可用数目两倍的物理 CPU。类似地， 0.50 的平均负载表示 ESX Server 系统上的物理 CPU 有一半得到了利用。

交互模式单键命令

以交互模式运行时， resxtop （或 esxtop）可识别几个单键命令。所有四个面板都可以识别表 A-2 中列出的命令。如果已经在命令行上给出 s 选项，则禁用指定更新之间延迟的命令 （请参见 “交互模式命令行选项”（第 161 页））。所有排序的交互命令按降序排序。

表 A-2. 交互模式单键命令

键	描述
h 或 ?	显示当前面板的帮助菜单，给出命令的简短摘要和安全模式的状态。
space	立即更新当前面板。
^L	擦除和重绘当前面板。
f 或 F	显示将统计信息列 （字段）添加到当前面板或从当前面板移除统计信息列 （字段）的面板。
o 或 O	显示在当前面板上更改统计信息列顺序的面板。
#	提示您输入要显示的统计信息的行数。大于 0 的任意值均会替代根据窗口大小测量自动确定的显示行数。如果在一个 resxtop （或 esxtop）面板中更改该数值，此更改会影响所有四个面板。
s	提示您输入更新之间的延迟，以秒为单位。小数值可以识别到微秒。默认值为 5 秒。最小值为 2 秒。安全模式中不可以使用该命令。
W	将当前设置写入 esxtop （或 resxtop）配置文件。这是写入配置文件的推荐方式。默认文件名是通过 -c 选项指定的文件名，如果不使用 –c 选项，则为 ~/.esxtop310rc。还可以在该 W 命令生成的提示中指定不同的文件名。
q	退出交互模式。
c	切换到 CPU 资源利用率面板。
m	切换到内存资源利用率面板。
d	切换到存储 （磁盘）适配器资源利用率面板。
u	切换到存储 （磁盘）设备资源利用率屏幕。请参见 “存储设备面板”（第 173 页）。
v	切换到存储 （磁盘）虚拟机资源利用率屏幕。请参见 “虚拟机存储面板”（第 175 页）
n	切换到网络资源利用率面板。

统计信息列和顺序页

如果按下 f、F、o 或 O，系统会显示一个页面，该页面在最上面的一行指定字段顺序和字段内容的简短描述。如果对应于字段的字段字符串中的字母为大写，则显示该字段。字段描述前面的星号表示是否显示字段。

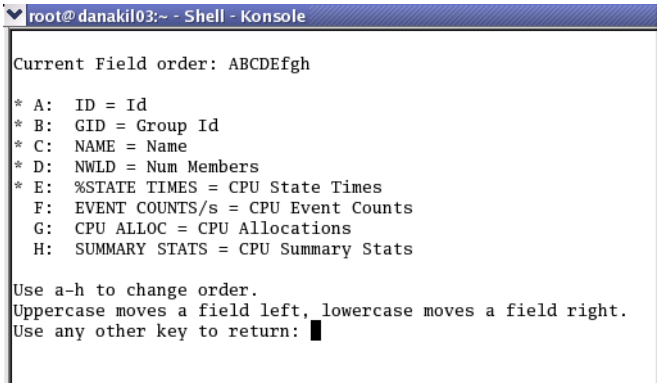
这些字段的顺序对应于字符串中字母的顺序。

从 [字段选择 (Field Select)] 面板中，您可以：

- 通过按下对应的字母，切换字段的显示
- 通过按下对应的大写字母，向左移动字段。
- 通过按下对应的小写字母，向右移动字段。

图 A-1 显示了字段顺序变化。

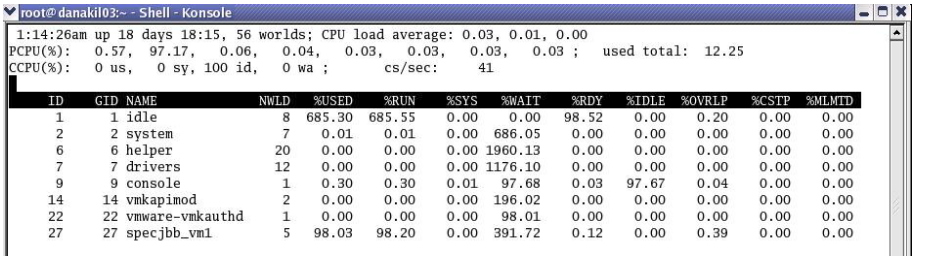
图 A-1. 字段顺序变化



CPU 面板

CPU 面板显示了服务器范围的统计信息以及单个环境、资源池和虚拟机 CPU 利用率的统计信息。资源池、正在运行的虚拟机或其他环境有时作为组引用。对于属于虚拟机的环境，显示正在运行的虚拟机的统计信息。所有其他环境按逻辑方式聚合到包含这些环境的资源池中。

图 A-2. CPU 面板



可以使用单键命令来更改该显示。表 A-3 和表 A-4 讨论了统计信息和单键命令。

表 A-3. CPU 面板统计信息

行	描述
PCPU(%)	每个物理 CPU 的 CPU 利用率百分比和平均物理 CPU 总利用率百分比。
LCPU(%)	每个逻辑 CPU 的 CPU 利用率百分比。属于封装件的逻辑 CPU 的百分比总计为 100%。只有存在并启用超线程时，才会出现该行。请参见 "超线程和 ESX Server"（第 121 页）。
CCPU(%)	ESX Server 服务控制台报告的 CPU 总时间的百分比。 ■ `us` - 用户时间百分比。 ■ `sy` - 系统时间百分比。 ■ `id` - 闲置时间百分比。 ■ `wa` - 等待时间百分比。 ■ `cs/sec` - 服务控制台记录的每秒上下文切换次数。
ID	正在运行环境的资源池或虚拟机的资源池 ID 或虚拟机 ID，或正在运行环境的环境 ID。
GID	正在运行环境的资源池或虚拟机的资源池 ID。
NAME	正在运行环境的资源池或虚拟机的名称，或正在运行环境的名称。
NWLD	正在运行环境的资源池或虚拟机中的成员数量。如果使用交互命令 `e`（请参见交互命令）扩展组，则所有生成环境的 NWLD 为 1（类似控制台资源池的一些资源池只有一个成员）。
%STATE TIMES	由以下百分比构成的 CPU 统计信息集合。对于环境，百分比是一个物理 CPU 的百分比。
%USED	由资源池、虚拟机或环境使用的物理 CPU 百分比。
%SYS	代表资源池、虚拟机或环境在 ESX Server VMkernel 中处理中断和执行其他系统活动所用的时间百分比。该时间是用于计算上述 **%USED** 的时间的一部分。
%WAIT	资源池、虚拟机或环境在阻止或遇忙等待状况所占的时间百分比。该百分比包括资源池、虚拟机或环境闲置的时间百分比。
%IDLE	资源池、虚拟机或环境闲置的时间百分比。从上述 **%WAIT** 中减去该百分比，可得出资源池、虚拟机或环境等待某个事件所用的时间百分比。
%RDY	资源池、虚拟机或环境准备运行的时间百分比。
%MLMTD	ESX Server VMkernel 故意未运行资源池、虚拟机或环境的时间百分比，因为如果运行的话，会违反资源池、虚拟机或环境的限制设置。即使资源池、虚拟机或环境被阻止以此方式运行时准备运行，**[%MLMTD]** 时间也不会包括在 **[%RDY]** 时间内。
EVENT COUNTS/s	由每秒事件速率构成的 CPU 统计信息集合。这些统计信息仅供 VMware 内部使用。

表 A-3. CPU 面板统计信息 （续）

行	描述
CPU ALLOC	由以下 CPU 分配配置参数构成的 CPU 统计信息集合。
AMIN	资源池、虚拟机或环境属性 **[预留 (Reservation)]**。请参见 “创建并定制资源池”（第 24 页）。
AMAX	资源池、虚拟机或环境属性 **[限制 (Limit)]**。 -1 值表示无限制。请参见 “创建并定制资源池”（第 24 页）。
ASHRS	资源池、虚拟机或环境属性 **[份额 (Shares)]**。请参见 “创建并定制资源池”（第 24 页）。
SUMMARY STATS	由以下 CPU 配置参数和统计信息构成的 CPU 统计信息集合。这些统计信息仅适用于环境，不适用于虚拟机或资源池。
AFFINITY BIT MASK	显示环境的当前调度关联性的位掩码。请参见 “使用 CPU 关联性向特定处理器分配虚拟机”（第 117 页）。
HTSHARING	当前超线程配置。请参见 “超线程的高级服务器配置”（第 122 页）。
CPU	当 `resxtop` （或 `esxtop`）获得该信息时，正在运行环境的物理或逻辑处理器。
HTQ	表示环境当前是否已隔离。 **[N]** 表示否， **[Y]** 表示是。请参见 “隔离”（第 123 页）。
TIMER/s	该环境的定时器速率。
%OVRLP	调度资源池、虚拟机或环境时，代表不同资源池、虚拟机或环境在调度资源池、虚拟机或环境期间所用系统时间的百分比。该时间不包括在 **[%SYS]** 中。例如，如果当前正在调度虚拟机 A 而且虚拟机 B 的网络包由 ESX Server VMkernel 处理，则虚拟机 A 所用的时间显示为 **[%OVRLP]**，而虚拟机 B 所用的时间显示为 **[%SYS]**。
%RUN	调度的总时间百分比。该时间不算超线程和系统时间。在启用超线程的服务器上， %RUN 可以是 **[%USED]** 大小的两倍。
%CSTP	资源池在就绪、共同取消调度状况中所用的时间百分比。 （注意：您可能会看到该统计信息显示出来，但其仅供 VMware 使用。）

表 A-4. CPU 面板单键命令

命令	描述
e	切换 CPU 统计信息扩展显示还是不扩展显示。 扩展显示包括按照属于资源池或虚拟机的单个环境细分的 CPU 资源利用率统计信息。单个环境的所有百分比是单个物理 CPU 的百分比。 考虑以下示例： ■ 如果在 2 路服务器上按资源池细分的 **[%Used]** 为 30%，则该资源池正在利用 30% 的两个物理 CPU 资源。 ■ 如果在 2 路服务器上按属于资源池的环境细分的 **[%Used]** 为 30%，则该环境正在利用 30% 的一个物理 CPU 资源。
U	按资源池或虚拟机的 **[%Used]** 列对资源池、虚拟机和环境进行排序。
R	按资源池或虚拟机的 **[%RDY]** 列对资源池、虚拟机和环境进行排序。
N	按 [GID] 列对资源池、虚拟机和环境进行排序。这是默认的排序顺序。
V	仅显示虚拟机实例。

内存面板

内存面板显示了服务器范围和组的内存利用率统计信息。与 CPU 面板类似，组对应于资源池、正在运行的虚拟机或正在消耗内存的其他环境。有关计算机内存和物理内存之间的区别，请参见 “内存虚拟化”（第 123 页）。

内存面板顶部第一行 （请参见图 A-3）显示了当前时间、自上一次重新引导以来所经过的时间、当前运行的环境数量和内存过量使用平均值。显示过去 1 分钟、5 分钟和 15 分钟内内存过量使用的平均值。1.00 的内存过量使用表示 100% 的内存过量使用。请参见 “内存过量使用”（第 39 页）。

图 A-3. 内存面板

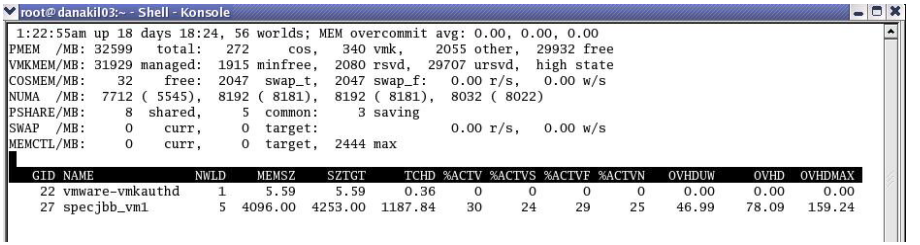


表 A-5. 内存面板统计信息

字段	描述
PMEM (MB)	显示服务器的计算机内存统计信息。所有数字都以兆字节为单位。 ■ total - 服务器中计算机内存总量。 ■ cos - 分配给 ESX Server 服务控制台 （仅 ESX Server 3）的计算机内存量。 ■ vmk - 正由 ESX Server Vmkernel 使用的计算机内存量。 ■ other - 除 ESX 服务控制台 （仅 ESX Server 3）和 ESX Server Vmkernel 之外其他各项正在使用的计算机内存量。 ■ free - 可用的计算机内存量。
VMKMEM (MB)	显示 ESX Server Vmkernel 的计算机内存统计信息。所有数字都以兆字节为单位。 ■ managed - 由 ESX Server Vmkernel 管理的计算机内存总量。 ■ min free - ESX Server VMkernel 旨在保持可用的计算机内存最小量。 ■ rsvd - 当前由资源池预留的计算机内存总量。 ■ ursvd - 当前未预留的计算机内存总量。 ■ state - 当前计算机内存可用性状况。可能的值为 high、 soft、 hard 和 low。 high 表示计算机内存没有任何压力， low 表示有压力。
COSMEM (MB)	显示 ESX Server 服务控制台 （仅 ESX Server 3）报告的内存统计信息。所有数字都以兆字节为单位。 ■ free - 闲置的内存量。 ■ swap_t - 配置的总交换量。 ■ swap_f - 可用的交换量。 ■ r/s is - 从磁盘换入内存的速率。 ■ w/s - 内存交换到磁盘的速率。
NUMA (MB)	显示 ESX Server NUMA 统计信息。只有当 ESX Server 主机正运行在 NUMA 服务器上时，才会显示该行。所有数字都以兆字节为单位。 对于服务器中的每个 NUMA 节点，显示两个统计信息： ■ 在由 ESX Server 管理的 NUMA 节点中的计算机内存总量。 ■ 在当前可用节点中的计算机内存量 （在圆括号中）。
PSHARE (MB)	显示 ESX Server 页共享统计信息。所有数字都以兆字节为单位。 ■ shared - 正共享的物理内存量。 ■ common - 环境之间共用的计算机内存量。 ■ saving - 由于页共享而节省的计算机内存量。

表 A-5. 内存面板统计信息 （续）

字段	描述
SWAP (MB)	显示 ESX Server 交换使用量统计信息。所有数字都以兆字节为单位。 ■ curr - 当前的交换使用量 ■ target - ESX Server 系统预期交换使用量的目标。 ■ r/s - 由 ESX Server 系统从磁盘换入内存的速率。 ■ w/s - 由 ESX Server 系统将内存交换到磁盘的速率。 有关背景信息，请参见 “交换”（第 131 页）。
MEMCTL (MB)	显示内存伸缩统计信息。所有数字都以兆字节为单位。 ■ curr - 使用 `vmmemctl` 模块回收的物理内存总量。 ■ target - ESX Server 主机尝试使用 `vmmemctl` 模块回收的物理内存总量。 ■ max - ESX Server 主机可以使用 `vmmemctl` 模块回收的物理内存最大量。 请参见 “内存伸缩 (vmmemctl) 驱动程序”（第 130 页）。
AMIN	该资源池或虚拟机的内存预留。请参见 “预留”（第 21 页）。
AMAX	该资源池或虚拟机的内存限制。 -1 值表示无限制。请参见 “限制”（第 21 页）。
ASHRS	该资源池或虚拟机的内存份额。请参见 “份额 (Shares)”（第 20 页）。
NHN	资源池或虚拟机的当前主节点。该统计信息仅适用于 NUMA 系统。如果虚拟机没有主节点，则显示短划线 (-)。
NRMEM (MB)	分配到虚拟机或资源池的当前远程内存量。该统计信息仅适用于 NUMA 系统。请参见 “VMware NUMA 优化算法”（第 143 页）。
N%L	分配到虚拟机或资源池的当前本地内存百分比。
MEMSZ (MB)	分配到资源池或虚拟机的物理内存量。
SZTGT (MB)	ESX Server VMkernel 想要分配到资源池或虚拟机的计算机内存量。
TCHD (MB)	资源池或虚拟机的工作集估计。请参见 “内存分配和闲置内存消耗”（第 127 页）。
%ACTV	正由客户机引用的客户机物理内存的百分比。这是瞬时值。
%ACTVS	正由客户机引用的客户机物理内存的百分比。这是慢速移动平均值。
%ACTVF	正由客户机引用的客户机物理内存的百分比。这是快速移动平均值。
%ACTVN	正由客户机引用的客户机物理内存的百分比。这是估计值。（您可能会看到该统计信息显示出来，但其仅供 VMware 使用。）
MCTL?	是否已安装内存伸缩驱动程序。 **[N]** 表示否， **[Y]** 表示是。
MCTLSZ (MB)	通过伸缩从资源池回收的物理内存量。

表 A-5. 内存面板统计信息 （续）

字段	描述
MCTLTGT (MB)	ESX Server 系统可以通过伸缩从资源池或虚拟机回收的物理内存量。
MCTLMAX (MB)	ESX Server 系统可以通过伸缩从资源池或虚拟机回收的最大物理内存量。该最大值取决于客户操作系统类型。
SWCUR (MB)	该资源池或虚拟机使用的当前交换量。
SWTGT (MB)	ESX Server 主机预期资源池或虚拟机交换使用量的目标。
SWR/s (MB)	ESX Server 主机为资源池或虚拟机从磁盘换入内存的速率。
SWW/s (MB)	ESX Server 主机将资源池或虚拟机内存交换到磁盘的速率。
CPTRD (MB)	从检查点文件中读取的数据量。
CPTTGT (MB)	检查点文件大小。
ZERO (MB)	置零的资源池或虚拟机物理页。
SHRD (MB)	共享的资源池或虚拟机物理页。
SHRDSVD (MB)	因为资源池或虚拟机共享页而节省的计算机页。
OVHD (MB)	资源池的当前空间开销。请参见 “了解内存开销”（第 126 页）。
OVHDMAX (MB)	可能由资源池或虚拟机造成的最大空间开销。请参见 “了解内存开销”（第 126 页）。
OVHDUW (MB)	用户环境的当前空间开销。（您可能会看到该统计信息显示出来，但其仅供 VMware 使用。）
GST_NDx (MB)	为 NUMA 节点 *x* 上的资源池分配的客户机内存。该统计信息仅适用于 NUMA 系统。
OVD_NDx (MB)	为 NUMA 节点 *x* 上的资源池分配的 VMM 开销内存。该统计信息仅适用于 NUMA 系统。

表 A-6. 内存面板交互命令

命令	描述
M	按 **[映射的组 (Group Mapped)]** 列对资源池或虚拟机排序。
B	按 **[组 Memctl (Group Memctl)]** 列对资源池或虚拟机排序。
N	按 **[GID]** 列对资源池或虚拟机排序。这是默认的排序顺序。
V	仅显示虚拟机实例。

存储面板

三个存储面板显示了服务器范围的存储利用率统计信息。

本节介绍了三个存储面板：

- “存储适配器面板”（第 170 页）
- “存储设备面板”（第 173 页）
- “虚拟机存储面板”（第 175 页）

存储适配器面板

存储适配器面板显示了图 A-4 中所示的信息。默认情况下，按照存储适配器来聚集统计信息。统计信息还可以按照存储通道、目标、 LUN 或使用 LUN 的环境来查看。

图 A-4. 存储适配器面板

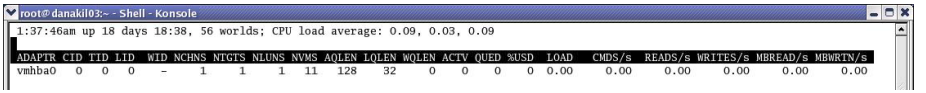


表 A-7. 存储适配器面板统计信息

列	描述
ADAPTR	存储适配器的名称。
CID	存储适配器通道 ID。只有扩展对应的适配器，该 ID 才可见。请参见下面的交互命令 e。
TID	存储适配器通道目标 ID。只有扩展对应的适配器和通道，该 ID 才可见。请参见下面的交互命令 e 和 a。
LID	存储适配器通道目标 LUN ID。只有扩展对应的适配器、通道和目标，该 ID 才可见。请参见下面的交互命令 e、 a 和 t。
WID	存储适配器通道目标 LUN 环境 ID。只有扩展对应的适配器、通道、目标和 LUN，该 ID 才可见。请参见下面的交互命令 e、 a、 t 和 l。
NCHNS	通道数量。
NTGTS	目标数量。
NLUNS	LUN 数量。
NVMS	环境数量。
SHARES	份额数量。
BLKSZ	以字节为单位的块大小。该统计信息仅适用于 LUN。

表 A-7. 存储适配器面板统计信息 （续）

列	描述
AQLEN	存储适配器队列深度。配置适配器驱动程序支持的 ESX Server VMkernel 活动命令的最大数目。
LQLEN	LUN 队列深度。允许 LUN 具有的 ESX Server VMkernel 活动命令的最大数目。
WQLEN	环境队列深度。允许环境具有的 ESX Server VMkernel 活动命令的最大数目。这是对于环境而言每个 LUN 的最大值。
%USD	ESX Server VMkernel 活动命令使用的队列深度 （适配器、 LUN 或环境）百分比。
LOAD	ESX Server VMkernel 活动命令加上 ESX Server VMkernel 排队命令与队列深度 （适配器、 LUN 或环境）的比率。
ACTV	当前活动的 ESX Server VMkernel 中的命令数目。
QUED	当前排队的 ESX Server VMkernel 中的命令数目。
CMDS/s	每秒发出的命令数目。
READS/s	每秒发出的读取命令数目。
WRITES/s	每秒发出的写入命令数目。
MBREAD/s	每秒读取的兆字节数。
MBWRTN/s	每秒写入的兆字节数。
DAVG/cmd	每条命令的平均设备滞后时间，以毫秒为单位。
KAVG/cmd	每条命令的平均 ESX Server Vmkernel 滞后时间，以毫秒为单位。
GAVG/cmd	每条命令的平均虚拟机操作系统滞后时间，以毫秒为单位。
DAVG/rd	每个读取操作的平均设备读取滞后时间，以毫秒为单位。
KAVG/rd	每个读取操作的平均 ESX Server Vmkernel 读取滞后时间，以毫秒为单位。
GAVG/rd	每个读取操作的平均客户操作系统读取滞后时间，以毫秒为单位。
DAVG/wr	每个写入操作的平均设备写入滞后时间，以毫秒为单位。
KAVG/wr	每个写入操作的平均 ESX Server Vmkernel 写入滞后时间，以毫秒为单位。
GAVG/wr	每个写入操作的平均客户操作系统写入滞后时间，以毫秒为单位。
QAVG/cmd	每条命令的平均队列滞后时间，以毫秒为单位。
QAVG/rd	每个读取操作的平均队列滞后时间，以毫秒为单位。
QAVG/wr	每个写入操作的平均队列滞后时间，以毫秒为单位。
ABRTS/s	每秒中止的命令数目。

表 A-7. 存储适配器面板统计信息 （续）

列	描述
RESETS/s	每秒重置的命令数目。
PAECMD/s	每秒的物理地址扩展 (Physical Address Extension, PAE) 命令数目。
PAECP/s	每秒的 PAE 副本数。
SPLTCMD/s	每秒的拆分命令数目。
SPLTCP/s	每秒的拆分副本数。

表 A-8. 存储适配器面板交互命令

命令	描述
e	切换存储适配器统计信息扩展显示还是不扩展显示。允许查看按照属于扩展存储适配器的单个通道细分的存储资源利用率统计信息。提示您输入适配器名称。
E	切换存储适配器统计信息扩展显示还是不扩展显示。允许查看按照属于扩展存储适配器的环境细分的存储资源利用率统计信息。请勿汇总到适配器统计信息。提示您输入适配器名称。
P	切换存储适配器统计信息扩展显示还是不扩展显示。允许查看按照属于扩展存储适配器的路径细分的存储资源利用率统计信息。请勿汇总到适配器统计信息。提示您输入适配器名称。
a	切换存储通道统计信息扩展显示还是不扩展显示。允许查看按照属于扩展存储通道的单个目标细分的存储资源利用率统计信息。提示您输入适配器名称和通道 ID。扩展通道本身之前，需要先扩展通道适配器。
t	切换存储目标统计信息以扩展模式显示还是以不扩展模式显示。允许查看按照属于扩展存储目标的单个路径细分的存储资源利用率统计信息。提示您输入适配器名称、通道 ID 和目标 ID。扩展目标本身之前，需要先扩展目标通道和适配器。
l	切换路径以扩展模式显示还是以不扩展模式显示。允许查看利用扩展存储路径的单个环境细分的存储资源利用率统计信息。提示您输入适配器名称、通道 ID、目标 ID 和 LUN ID。扩展路径本身之前，必须先扩展路径目标、通道和适配器。
r	按 **[读取 (Reads)]** 列排序。
w	按 **[写入 (Writes)]** 列排序。
R	按 **[读取的 MB (MB read)]** 列排序。

表 A-8. 存储适配器面板交互命令 （续）

命令	描述
T	按 **[写入的 MB (MB written)]** 列排序。
N	先按 **[ADAPTR]** 列排序，再按每个 **[ADAPTR]** 内的 **[CID]** 列排序，再按每个 **[CID]** 内的 **[TID]** 列排序，然后按每个 **[TID]** 内的 **[LID]** 列排序，最后按每个 **[LID]** 内的 **[WID]** 列排序。这是默认的排序顺序。

存储设备面板

存储设备面板显示了服务器范围的存储利用率统计信息。默认情况下，该信息按存储设备分组。还可以按照路径、环境或分区对统计信息分组。

图 A-5. 存储设备面板

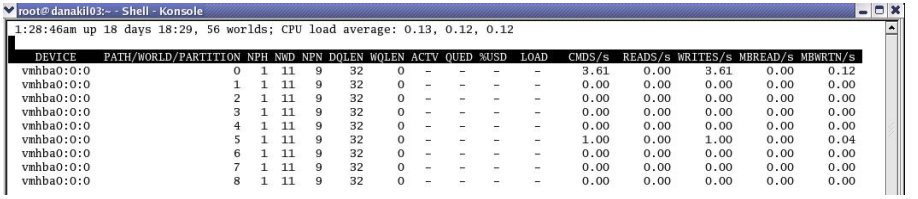


表 A-9. 存储设备面板统计信息

列	描述
DEVICE	存储设备的名称。
PATH	路径名称。只有对应的设备扩展到路径，该名称才可见。请参见下面的交互命令 **p**。
WORLD	环境 ID。只有对应的设备扩展到环境，该 ID 才可见。请参见下面的交互命令 **e**。环境统计信息按环境和设备显示。
PARTITION	分区 ID。只有对应的设备扩展到分区，该 ID 才可见。请参见下面的交互命令 **t**。
NPH	路径数量。
NWD	环境数量。
NPN	分区数量。
SHARES	份额数量。该统计信息仅适用于环境。
BLKSZ	以字节为单位的块大小。
NUMBLKS	设备的块数。
DQLEN	存储设备队列深度。这是配置设备支持的 ESX Server VMkernel 活动命令的最大数目。

表 A-9. 存储设备面板统计信息 （续）

列	描述
WQLEN	环境队列深度。这是允许环境具有的 ESX Server VMkernel 活动命令的最大数目。这是对于环境而言每个设备的最大值。只有对应的设备扩展到环境，此列才有效。
ACTV	当前活动的 ESX Server VMkernel 中的命令数目。该统计信息仅适用于环境和设备。
QUED	当前排队的 ESX Server VMkernel 中的命令数目。该统计信息仅适用于环境和设备。
%USD	ESX Server VMkernel 活动命令使用的队列深度百分比。该统计信息仅适用于环境和设备。
LOAD	ESX Server VMkernel 活动命令加上 ESX Server VMkernel 排队命令与队列深度的比率。该统计信息仅适用于环境和设备。
CMDS/s	每秒发出的命令数目。
READS/s	每秒发出的读取命令数目。
WRITES/s	每秒发出的写入命令数目。
MBREAD/s	每秒读取的兆字节数。
MBWRTN/s	每秒写入的兆字节数。
DAVG/cmd	每条命令的平均设备滞后时间，以毫秒为单位。
KAVG/cmd	每条命令的平均 ESX Server Vmkernel 滞后时间，以毫秒为单位。
GAVG/cmd	每条命令的平均客户操作系统滞后时间，以毫秒为单位。
QAVG/cmd	每条命令的平均队列滞后时间，以毫秒为单位。
DAVG/rd	每个读取操作的平均设备读取滞后时间，以毫秒为单位。
KAVG/rd	每个读取操作的平均 ESX Server Vmkernel 读取滞后时间，以毫秒为单位。
GAVG/rd	每个读取操作的平均客户操作系统读取滞后时间，以毫秒为单位。
QAVG/rd	每个读取操作的平均队列读取滞后时间，以毫秒为单位。
DAVG/wr	每个写入操作的平均设备写入滞后时间，以毫秒为单位。
KAVG/wr	每个写入操作的平均 ESX Server Vmkernel 写入滞后时间，以毫秒为单位。
GAVG/wr	每个写入操作的平均客户操作系统写入滞后时间，以毫秒为单位。
QAVG/wr	每个写入操作的平均队列写入滞后时间，以毫秒为单位。
ABRTS/s	每秒中止的命令数目。
RESETS/s	每秒重置的命令数目。
PAECMD/s	每秒的 PAE 命令数目。该统计信息仅适用于路径。

表 A-9. 存储设备面板统计信息 （续）

列	描述
PAECP/s	每秒的 PAE 副本数。该统计信息仅适用于路径。
SPLTCMD/s	每秒的拆分命令数目。该统计信息仅适用于路径。
SPLTCP/s	每秒的拆分副本数。该统计信息仅适用于路径。

表 A-10. 存储设备面板交互命令

命令	描述
e	扩展或汇总存储环境统计信息。该命令允许查看属于扩展存储设备的单个环境分隔的存储资源利用率统计信息。提示您输入设备名称。统计信息按环境和设备显示。
p	扩展或汇总存储路径统计信息。该命令允许查看属于扩展存储设备的单个路径分隔的存储资源利用率统计信息。提示您输入设备名称。
t	扩展或汇总存储分区统计信息。该命令允许查看属于扩展存储设备的单个分区分隔的存储资源利用率统计信息。提示您输入设备名称。
r	按 **[READS/s]** 列排序。
w	按 **[WRITES/s]** 列排序。
R	按 **[MBREAD/s]** 列排序。
T	按 **[MBWRTN]** 列排序。
N	先按 **[设备 (DEVICE)]** 列排序，再按 **[路径 / 环境 / 分区 (PATH/WORLD/PARTITION)]** 列排序。这是默认的排序顺序。

虚拟机存储面板

该面板显示了以虚拟机为中心的存储统计信息。默认情况下，按照资源池聚集统计信息。一个虚拟机具有一个对应的资源池，因此该面板实际上按照虚拟机显示统计信息。还可以按照环境或按照环境和设备查看统计信息。

图 A-6. 虚拟机存储面板

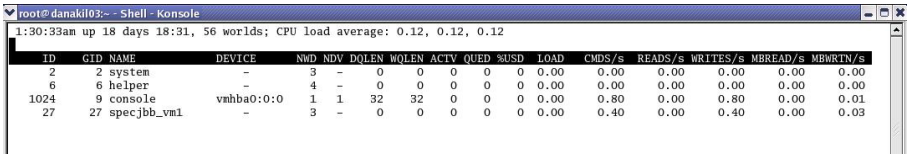


表 A-11. 虚拟机存储面板统计信息

列	描述
ID	正在运行环境的资源池的资源池 ID 或正在运行环境的环境 ID。
GID	正在运行环境的资源池的资源池 ID。
NAME	正在运行环境的资源池名称或正在运行环境的名称。
Device	存储设备名称。只有对应的环境扩展到设备，该名称才可见。请参见下面的交互命令 **i** 。
NWD	环境数量。
NDV	设备数量。只有对应的资源池扩展到环境，此数量才有效
SHARES	份额数量。该统计信息仅适用于环境。只有对应的资源池扩展到环境，此列才有效。
BLKSZ	以字节为单位的块大小。只有对应的环境扩展到设备，此列才有效。
NUMBLKS	设备的块数。只有对应的环境扩展到设备，此列才有效。
DQLEN	存储设备队列深度。这是配置设备支持的 ESX Server VMkernel 活动命令的最大数目。只有对应的环境扩展到设备，显示的数字才有效。
WQLEN	环境队列深度。该列显示了允许环境具有的 ESX Server VMkernel 活动命令的最大数目。只有对应的环境扩展到设备，该数字才有效。这是对于环境而言每个设备的最大值。
ACTV	当前活动的 ESX Server VMkernel 中的命令数目。该数字仅适用于环境和设备。
QUED	当前排队的 ESX Server VMkernel 中的命令数目。该数字仅适用于环境和设备。
%USD	ESX Server VMkernel 活动命令使用的队列深度百分比。该数字仅适用于环境和设备。
LOAD	ESX Server VMkernel 活动命令加上 ESX Server VMkernel 排队命令与队列深度的比率。该数字仅适用于环境和设备。
CMDS/s	每秒发出的命令数目。
READS/s	每秒发出的读取命令数目。
WRITES/s	每秒发出的写入命令数目。

表 A-11. 虚拟机存储面板统计信息 （续）

列	描述
MBREAD/s	每秒读取的兆字节数。
MBWRTN/s	每秒写入的兆字节数。
DAVG/cmd	每条命令的平均设备滞后时间，以毫秒为单位。
KAVG/cmd	每条命令的平均 ESX Server Vmkernel 滞后时间，以毫秒为单位。
GAVG/cmd	每条命令的平均客户操作系统滞后时间，以毫秒为单位。
QAVG/cmd	每条命令的平均队列滞后时间，以毫秒为单位。
DAVG/rd	每个读取操作的平均设备读取滞后时间，以毫秒为单位。
KAVG/rd	每个读取操作的平均 ESX Server Vmkernel 读取滞后时间，以毫秒为单位。
GAVG/rd	每个读取操作的平均客户操作系统读取滞后时间，以毫秒为单位。
QAVG/rd	每个读取操作的平均队列读取滞后时间，以毫秒为单位。
DAVG/wr	每个写入操作的平均设备写入滞后时间，以毫秒为单位。
KAVG/wr	每个写入操作的平均 ESX Server Vmkernel 写入滞后时间，以毫秒为单位。
GAVG/wr	每个写入操作的平均客户操作系统写入滞后时间，以毫秒为单位。
QAVG/wr	每个写入操作的平均队列写入滞后时间，以毫秒为单位。
ABRTS/s	每秒中止的命令数目，以毫秒为单位。
RESETS/s	每秒重置的命令数目，以毫秒为单位。

表 A-12. 虚拟机存储面板交互命令

命令	描述
e	扩展或汇总存储环境统计信息。允许查看属于组的单个环境分隔的存储资源利用率统计信息。提示您输入组 ID。该统计信息按环境显示。
l	扩展或汇总存储设备 （即 LUN）统计信息。允许查看属于扩展环境的单个设备分隔的存储资源利用率统计信息。提示您输入环境 ID。
V	仅显示虚拟机实例。
r	按 **[READS/s]** 列排序。
w	按 **[WRITES/s]** 列排序。
R	按 **[MBREAD/s]** 列排序。
T	按 **[MBWRTN/s]** 列排序。
N	先按虚拟机列排序，然后按 [环境 (WORLD)] 列排序。这是默认的排序顺序。

网络面板

图 A-7 中所示的面板显示了服务器范围的网络利用率统计信息。统计信息按照每个配置的虚拟网络设备端口进行排列。有关物理网络适配器统计信息，请参见对应于连接物理网络适配器的端口的行。有关在特定虚拟机上配置的虚拟网络适配器的统计信息，请参见对应于连接虚拟网络适配器的端口的行。

图 A-7. 网络面板

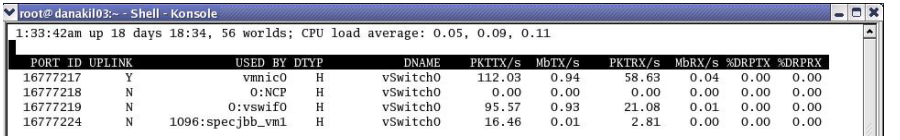


表 A-13. 网络面板统计信息

列	描述
PORT	虚拟网络设备端口 ID。
UPLINK	**[Y]** 表示对应的端口是上行链路。 **[N]** 表示不是。
UP	**[Y]** 表示对应的链路是上行链路。 **[N]** 表示不是。
SPEED	以每秒兆位为单位的链路速度。
FDUPLX	**[Y]** 表示对应的链路以全双工方式运行。 **[N]** 表示不是。
USED	虚拟网络设备端口用户。
DTYP	虚拟网络设备类型。 **[H]** 表示集线器， **[S]** 表示交换机。
DNAME	虚拟网络设备名称。
PKTTX/s	每秒传输的数据包数。
PKTRX/s	每秒接收的数据包数。
MbTX/s	每秒传输的兆位数。
MbRX/s	每秒接收的兆位数。
%DRPTX	丢弃的传输数据包百分比。
%DRPRX	丢弃的接收数据包百分比。

表 A-14. 网络面板交互命令

命令	描述
T	按 **[Mb Tx]** 列排序。
R	按 **[Mb Rx]** 列排序。

表 A-14. 网络面板交互命令 （续）

命令	描述
t	按 **[Packets Tx]** 列排序。
r	按 **[Packets Rx]** 列排序。
N	按 **[端口 ID (PORT ID)]** 列排序。这是默认的排序顺序。

以批处理模式使用实用程序

批处理模式允许您收集资源利用率统计信息并将其保存到文件中。要以批处理模式运行，必须先准备批处理模式。

准备以批处理模式运行 resxtop 或 esxtop

1 以交互模式运行 resxtop （或 esxtop）。

2 在每个面板中，选择您想要的列。

3 使用 *W* 交互命令将该配置保存到文件 （默认为 ~/.esxtop310rc）中。

以批处理模式运行 resxtop 或 esxtop

1 启动 resxtop （或 esxtop）将输出重定向到文件。例如：

```
esxtop –b > my_file.csv
```

文件名必须具有 .csv 扩展名。该实用程序不强制要求这点，但后处理工具需要该扩展名。

2 使用诸如 Microsoft Excel 和 Perfmon 之类的工具处理批处理模式中收集的统计信息。

在批处理模式中， resxtop （或 esxtop）不接受交互命令。在批处理模式中，该实用程序运行到产生请求的迭代次数为止 （有关详细信息，请参见下面的命令行选项 n），或运行到通过按 Ctrl+c 终止进程为止。

在批处理模式中可以使用表 A-15 中的命令行选项。

表 A-15. 批处理模式中的命令行选项

选项	描述
a	显示所有统计信息。该选项会替代配置文件设置并显示所有统计信息。配置文件可以是默认的 ~/.esxtop310rc 配置文件或用户定义的配置文件。
b	以批处理模式运行 resxtop （或 esxtop）。

表 A-15. 批处理模式中的命令行选项 （续）

选项	描述
c <filename>	加载用户定义的配置文件。如果未使用 –c 选项，则默认配置文件名为 ~/.esxtop310rc。使用 W 单键交互命令创建自己的配置文件，同时指定不同的文件名。有关 W 的信息，请参见 “交互模式单键命令”（第 162 页）。
d	指定统计信息快照之间的延迟。默认为 5 秒。最小值为 2 秒。如果指定的延迟少于 2 秒，延迟将设置为 2 秒。
n	迭代次数。 resxtop （或 esxtop）对统计信息迭代执行此次数的收集和保存操作，然后退出。
server	要连接的远程服务器主机的名称 （仅 resxtop 需要）。
portnumber	要连接的远程服务器上的端口号。默认端口为 443，除非在服务器上更改了这一端口，否则不需要此选项。（仅 resxtop）
username	连接远程主机时要进行身份验证的用户名。远程服务器还会提示您输入密码 （仅 resxtop）。

以重放模式使用实用程序

在重放模式中， resxtop （或 esxtop）重放使用 vm–support 收集的资源利用率统计信息。请参见 vm–support 手册页。

要以重放模式运行，必须先准备重放模式。

准备以重放模式运行 resxtop 或 esxtop

1 在 ESX Server 服务控制台 （仅 ESX Server 3）上以快照模式运行 vm–support。

使用以下命令：

```
vm–support –S –d duration –i interval
```

2 解压缩生成的 tar 文件，以便 resxtop （或 esxtop）可以在重放模式中使用该文件。

以重放模式运行 resxtop 或 esxtop

在命令行提示符下输入以下内容：

```
resxtop –R <vm–support_dir_path>
```

表 A-16 中列出了其他命令行选项。

您可以不必在 ESX Server 服务控制台上运行重放模式。

可以运行重放模式，按照与批处理模式相同的样式产生输出 （请参见下面的命令行选项 b）。

在重放模式中， resxtop （或 esxtop）接受与交互模式相同的交互命令集，并运行到不再有 vm-support 收集的快照要读取为止，或者运行到请求的迭代次数已完成为止（有关详细信息，请参见命令行选项 n）。

表 A-16 列出了可用于 resxtop （或 esxtop）重放模式的命令行选项。

表 A-16. 重放模式中的命令行选项

选项	描述
R	vm-support 收集的快照目录的路径。
a	显示所有统计信息。该选项会替代配置文件设置并显示所有统计信息。配置文件可以是默认的 ~/.esxtop310rc 配置文件或用户定义的配置文件。
b	以批处理模式运行 resxtop （或 esxtop）。
c <filename>	加载用户定义的配置文件。如果未使用 –c 选项，则默认配置文件名为 ~/.esxtop310rc。使用 W 单键交互命令创建自己的配置文件并指定不同的文件名。有关 W 的信息，请参见 “交互模式单键命令”（第 162 页）。
d	指定面板更新之间的延迟。默认为 5 秒。最小值为 2 秒。如果指定的延迟少于 2 秒，延迟将设置为 2 秒。
n	迭代次数。 resxtop （或 esxtop）对显示迭代执行此次数的更新操作，然后退出。

索引