

《vSphere 资源管理》

VMware vSphere 8.0

VMware ESXi 8.0

vCenter Server 8.0

您可以从 VMware 网站下载最新的技术文档:

<https://docs.vmware.com/cn/>。

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

**威睿信息技术（中国）有
限公司**
北京办公室
北京市
朝阳区新源南路 8 号
启皓北京东塔 8 层 801
www.vmware.com/cn

上海办公室
上海市
淮海中路 333 号
瑞安大厦 804-809 室
www.vmware.com/cn

广州办公室
广州市
天河路 385 号
太古汇一座 3502 室
www.vmware.com/cn

版权所有 © 2006-2022 VMware, Inc. 保留所有权利。 [版权和商标信息](#)

目录

关于 vSphere 资源管理	10
1 资源管理入门	11
资源类型	11
资源提供方	11
资源用户	12
资源管理的目标	12
2 配置资源分配设置	13
资源分配份额	13
资源分配预留	14
资源分配限制	14
资源分配设置建议	15
编辑设置	15
更改资源分配设置 — 示例	16
准入控制	17
3 CPU 虚拟化基本知识	18
基于软件的 CPU 虚拟化	18
硬件辅助的 CPU 虚拟化	18
虚拟化和特定于处理器的行为	19
CPU 虚拟化的性能影响	19
4 管理 CPU 资源	20
查看处理器信息	20
指定 CPU 配置	20
多核处理器	21
5 超线程	22
超线程和 ESXi 主机	22
启用超线程	23
6 使用 CPU 关联性	24
向特定处理器分配虚拟机	24
CPU 关联性的潜在问题	25
7 主机电源管理策略	26

- 选择 CPU 电源管理策略 27
- 为主机电源管理配置自定义策略参数 27

8 内存虚拟化基本知识 29

- 虚拟机内存 29
- 内存过载 30
- 内存共享 30
- 内存虚拟化 31
- 硬件辅助的内存虚拟化 31
- 支持大内存页 32

9 管理内存资源 33

- 了解内存开销 34
- 虚拟机上的开销内存 34
- ESXi 主机如何分配内存 35
- 闲置虚拟机的内存消耗 35
- VMX 交换文件 35
- 内存回收 36
- 内存气球驱动程序 36
- 在虚拟机之间共享内存 37
- 内存压缩 37
- 激活或停用内存压缩缓存 37
- 设置内存压缩缓存的最大大小 38
- 衡量和区分各种内存使用情况 38
- 内存可靠性 39
- 更正错误隔离通知 40
- 关于系统交换 40
- 配置系统交换 40

10 使用交换文件 42

- 交换文件位置 42
- 为 DRS 集群启用主机-本地交换 43
- 为独立主机启用主机-本地交换 43
- 交换空间和内存过载 44
- 配置主机的虚拟机交换文件属性 44
- 配置集群的虚拟机交换文件位置 45
- 删除交换文件 46

11 永久内存 47

- 配置 PMem 虚拟机的 vSphere HA 49
- vSphere HA 准入控制 PMem 预留 50

vSphere 内存监控和修复 50

12 配置虚拟图形 52

查看 GPU 统计信息 52

将 NVIDIA GRID vGPU 添加到虚拟机 53

配置主机图形 53

配置图形设备 54

13 管理存储 I/O 资源 55

关于虚拟机存储策略 56

关于 I/O 筛选器 56

Storage I/O Control 要求 56

Storage I/O Control 资源份额和限制 57

查看 Storage I/O Control 份额和限制 57

监控 Storage I/O Control 份额 57

设置 Storage I/O Control 资源份额和限制 58

启用 Storage I/O Control 58

设置 Storage I/O Control 阈值 59

Storage DRS 与存储配置文件集成 60

14 管理资源池 62

为什么使用资源池？ 63

创建资源池 65

编辑资源池 66

将虚拟机添加到资源池 66

从资源池移除虚拟机 67

移除资源池 68

资源池接入控制 68

可扩展预留示例 1 68

可扩展预留示例 2 69

15 vSphere 集群服务 71

vSphere DRS 和 vCLS 虚拟机 72

为 vCLS 虚拟机选择数据存储 72

vCLS 数据存储放置 73

监控 vSphere 集群服务 73

维护 vSphere 集群服务的运行状况 74

将集群置于撤回模式 75

检索 vCLS 虚拟机的密码 76

vCLS 虚拟机反关联性策略 77

创建或删除 vCLS 虚拟机反关联性策略 77

16 创建 DRS 集群 78

- 准入控制和初始放置 79
- 单个虚拟机打开电源 79
- 组启动 79
- 虚拟机迁移 80
- DRS 迁移阈值 81
- 迁移建议 81
- DRS 集群要求 82
- 共享存储器要求 82
- 共享的 VMFS 卷要求 82
- 处理器兼容性要求 82
- DRS 集群的 vMotion 要求 83
- 配置带有虚拟闪存的 DRS 83
- 创建集群 84
- 编辑集群设置 85
- 设置虚拟机的自定义自动化级别 87
- 停用 DRS 87
- 还原资源池树 88
- vSAN 延伸集群的 DRS 感知 88

17 具有 ROBO 企业许可证的 DRS 维护模式功能 90

- 具有 ROBO 企业许可证的 DRS 维护模式存在的限制 90
- 使用具有 ROBO 企业许可证的 DRS 维护模式 90
- 对具有 ROBO 企业许可证的 DRS 维护模式进行故障排除 91

18 使用 DRS 集群管理资源 92

- 将主机添加到集群 92
 - 将受管主机添加到集群 93
 - 将非受管主机添加到集群 93
- 将虚拟机添加到集群 94
 - 将虚拟机移到集群 94
- 从集群内移除虚拟机 95
 - 将虚拟机移出集群 95
- 从集群中移除主机 95
 - 将主机置于维护模式 96
 - 从集群中移除主机 96
- 使用待机模式 97
- DRS 集群有效性 97
 - 有效 DRS 集群 98
 - 过载的 DRS 集群 99

无效 DRS 集群	100
管理电源资源	101
为 vSphere DPM 配置 IPMI 或 iLO 设置	102
测试 vSphere DPM 的 LAN 唤醒	103
为 DRS 集群激活 vSphere DPM	104
监控 vSphere DPM	105
使用 DRS 关联性规则	106
创建主机 DRS 组	106
创建虚拟机 DRS 组	107
虚拟机-虚拟机关联性规则	107
虚拟机-主机关联性规则	108
19 创建数据存储集群	111
初始放置位置和后续平衡	112
存储迁移建议	112
创建数据存储集群	112
激活和停用 Storage DRS	113
为数据存储集群设置自动化级别	113
设置 Storage DRS 的激进级别	114
设置 Storage DRS 运行时规则	115
Datastore Cluster 要求	116
在数据存储集群中添加和移除数据存储	116
20 使用数据存储集群管理存储资源	117
使用存储 DRS 维护模式	117
将数据存储置于维护模式	117
对于维护模式忽略 Storage DRS 关联性规则	118
应用存储 DRS 建议	119
刷新存储 DRS 建议	119
更改虚拟机的存储 DRS 自动化级别	120
设置 Storage DRS 的非工作时间调度	120
Storage DRS 反关联性规则	121
创建虚拟机反关联性规则	122
创建 VMDK 反关联性规则	123
替代 VMDK 关联性规则	123
清除 Storage DRS 统计信息	124
Storage vMotion 与数据存储集群的兼容性	125
21 配合使用 NUMA 系统和 ESXi	126
什么是 NUMA?	126
对操作系统的挑战	127

- ESXi NUMA 调度的工作方式 127
- VMware NUMA 优化算法和设置 128
- 主节点和初始放置位置 128
- 动态负载平衡和页面迁移 128
- 针对 NUMA 优化的透明页共享 129
- NUMA 架构中的资源管理 129
- 使用虚拟 NUMA 130
- ESXi 8.0 中的虚拟拓扑 130
- 虚拟 NUMA 控制 131
- 指定 NUMA 控制 132
- 将虚拟机与特定处理器关联 133
- 使用内存关联性将内存分配与特定 NUMA 节点相关联 133
- 将虚拟机与指定的 NUMA 节点关联 134

22 高级属性 135

- 设置高级主机属性 135
 - 高级内存属性 136
 - 高级 NUMA 属性 137
- 设置高级虚拟机属性 138
 - 高级虚拟机属性 138
 - 高级虚拟 NUMA 属性 139
- 延迟时间敏感度 139
 - 调整延迟时间敏感度 140
- 虚拟机的虚拟超线程支持 140
- vHT 完整 CPU 预留 140
- 为虚拟机激活 vHT 141
- 关于可靠内存 141
 - 查看可靠内存 142
- 使用 1GB 页面备份客户机 vRAM 142

23 故障定义 143

- 虚拟机已固定 144
- 虚拟机与任何主机均不兼容 144
- 移动到另一台主机时违反了虚拟机/虚拟机 DRS 规则 144
- 主机与虚拟机不兼容 144
- 主机有违反虚拟机/虚拟机 DRS 规则的虚拟机 144
- 主机用于虚拟机的容量不足 145
- 主机处于错误的状态 145
- 主机用于虚拟机的物理 CPU 的数量不足 145
- 主机用于每个虚拟机 CPU 的容量不足 145
- 虚拟机正在执行 vMotion 操作 145

- 集群中没有活动主机 145
- 资源不足 145
- 资源不足以满足配置的 HA 故障切换级别 145
- 无兼容的硬关联性主机 146
- 无兼容的软关联性主机 146
- 不允许违反软规则更改 146
- 影响软规则更改 146

24 DRS 故障排除信息 147

集群问题 147

- 集群负载不均衡 147
- 集群为黄色 148
- 集群为红色，因为资源池不一致 148
- 集群为红色，因为与故障切换容量发生冲突 148
- 集群总负载低时主机电源不关闭 149
- 集群总负载高时关闭主机电源 149
- DRS 很少或从不执行 vMotion 迁移 150

主机问题 151

- DRS 建议在集群总负载低时打开主机电源以增加容量 151
- 集群总负载高 151
- 集群总负载低 152
- DRS 没有撤出请求进入维护或待机模式的主机 152
- DRS 没有将任何虚拟机移动到主机上 152
- DRS 没有从主机移动任何虚拟机 153

虚拟机问题 154

- CPU 或内存资源不足 154
- 违反了虚拟机/虚拟机 DRS 规则或者虚拟机/主机 DRS 规则 155
- 打开虚拟机电源操作失败 155
- DRS 没有移动虚拟机 156

关于 vSphere 资源管理

《vSphere 资源管理》介绍了 VMware® ESXi 和 vCenter® Server 环境中的资源管理。

本文档重点介绍了以下主题。

- 资源分配和资源管理概念
- 虚拟机属性和准入控制
- 资源池及其管理方式
- 集群、vSphere® Distributed Resource Scheduler (DRS)、vSphere Distributed Power Management (DPM) 及其使用方式
- 数据存储集群、Storage DRS、Storage I/O Control 及其使用方式
- 高级资源管理选项
- 性能注意事项

VMware 非常重视包容性。为了在客户、合作伙伴和内部社区中促进这一原则，我们采用包容性语言创建内容。

目标读者

本信息的目标读者为想要了解系统如何管理资源以及资源如何自定义默认行为的系统管理员。这些信息对想要了解并使用资源池、集群、DRS、数据存储集群、Storage DRS、Storage I/O Control 或 vSphere DPM 的所有用户同样必不可少。

本文档假定您掌握了 VMware ESXi 和 vCenter Server 的相关应用知识。

注 在本文档中，“内存”可以指物理内存或永久内存。

资源管理入门

1

要了解资源管理，必须清楚其组件、目标以及如何以最佳方式在集群设置中将其实现。

本节将讨论虚拟机的资源分配设置（份额、预留和限制），包括如何设置和查看这些设置。另外，本节还将介绍准入控制过程，系统通过该过程对照现有资源对资源分配设置进行验证。

资源管理是将资源从资源提供方分配到资源用户的一个过程。

我们之所以需要资源管理，原因就是资源会过度分配（即需求大于容量）以及需求与容量会随着时间的推移而发生变化。通过资源管理，可以动态重新分配资源，以便更高效地使用可用容量。

注 本章中“内存”是指物理内存。

本章讨论了以下主题：

- 资源类型
- 资源提供方
- 资源用户
- 资源管理的目标

资源类型

资源包括 CPU、内存、电源、存储器和网络资源。

注 ESXi 分别使用网络流量调整和按比例分配份额机制来管理每台主机上的网络带宽和磁盘资源。

资源提供方

主机和集群（包括数据存储集群）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化软件所用的资源。

集群是一组主机。可以使用 vSphere Client 创建集群，并将多个主机添加到集群。vCenter Server 一起管理这些主机的资源：集群拥有所有主机的全部 CPU 和内存。可以针对联合负载平衡或故障切换来启用集群。有关详细信息，请参见第 16 章 [创建 DRS 集群](#)。

数据存储集群是一组数据存储。与 DRS 集群一样，您可以使用 vSphere Client 创建一个数据存储集群，并将多个数据存储添加到集群中。vCenter Server 共同管理数据存储资源。可以启用 Storage DRS 来平衡 I/O 负载和空间使用情况。请参见第 19 章 [创建数据存储集群](#)。

资源用户

虚拟机是资源用户。

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，以便基于份额分配占资源提供方的总 CPU、内存以及存储 I/O 的百分比，或者分配所保证的 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。

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，用于对可用的 CPU 和内存资源按层次结构进行分区。相应地，资源池既可以被视为资源提供方，也可以被视为资源用户。它们向子资源池和虚拟机提供资源，但是，由于它们也消耗其父资源池和虚拟机的资源，因此它们同时也是资源用户。请参见第 14 章 [管理资源池](#)。

ESXi 主机根据以下因素为每个虚拟机分配一部分基础硬件资源：

- 由用户定义的资源限制。
- ESXi 主机（或集群）的可用资源总量。
- 启动的虚拟机数目和这些虚拟机的资源使用情况。
- 管理虚拟化所需的开销。

资源管理的目标

管理资源时，必须清楚自己的目标。

除了解决资源过量置备问题，资源管理还可以帮助您实现以下目标：

- 性能隔离：防止虚拟机独占资源并保证服务率的可预测性。
- 高效使用：利用分配不足的资源并在过量置备时让性能正常降低。
- 易于管理：控制虚拟机的相对重要性，提供灵活的动态分区并且符合绝对服务级别协议。

配置资源分配设置

2

当可用资源容量无法满足资源用户（和虚拟化开销）的需求时，管理员可能需要对分配给虚拟机或它们所驻留的资源池的资源量进行自定义。

资源分配设置（份额、预留和限制）用于确定为虚拟机提供的 CPU、内存和存储资源量。特别是，管理员有多个用于分配资源的选项。

- 预留主机或集群的物理资源。
- 为可以分配给虚拟机的资源量设置上限。
- 保证为特定虚拟机分配的物理资源百分比始终高于其他虚拟机。

注 本章中“内存”是指物理内存。

本章讨论了以下主题：

- 资源分配份额
- 资源分配预留
- 资源分配限制
- 资源分配设置建议
- 编辑设置
- 更改资源分配设置 — 示例
- 准入控制

资源分配份额

份额指定虚拟机（或资源池）的相对重要性。如果某个虚拟机的资源份额是另一个虚拟机的两倍，则在这两个虚拟机争用资源时，第一个虚拟机有权消耗两倍于第二个虚拟机的资源。

份额通常指定为**高**、**正常**或**低**，这些值将分别按 4:2:1 的比例指定份额值。还可以选择**自定义**为各虚拟机分配特定的份额值（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。为虚拟机分配份额时，始终会相对于其他已打开电源的虚拟机来为该虚拟机指定优先级。

下表显示了虚拟机的默认 CPU 和内存份额值。对于资源池，默认的 CPU 份额值和内存份额值是相同的，但是必须将二者相乘，就好像是资源池是具有四个虚拟 CPU 和 16 GB 内存的虚拟机一样。

表 2-1. 份额值

设置	CPU 份额值	内存份额值
高	每个虚拟 CPU 具有 2000 个份额	所配置的虚拟机内存的每兆字节具有 20 个份额。
正常	每个虚拟 CPU 具有 1000 个份额	所配置的虚拟机内存的每兆字节具有 10 个份额。
低	每个虚拟 CPU 具有 500 个份额	所配置的虚拟机内存的每兆字节具有 5 个份额。

例如，一台具有两个虚拟 CPU 和 1GB 内存且 CPU 和内存份额设置为**正常**的 SMP 虚拟机具有 $2 \times 1000 = 2000$ 个 CPU 份额和 $10 \times 1024 = 10240$ 个内存份额。

具有一个以上虚拟 CPU 的虚拟机称为 SMP（对称多处理）虚拟机。

打开新虚拟机电源时，每个份额所代表的相对优先级会改变。这将影响同一资源池内的所有虚拟机。所有虚拟机都具有相同数量的虚拟 CPU。请考虑以下示例。

- 一台聚合 CPU 容量为 8 GHz 的主机上运行着两个受 CPU 约束的虚拟机。它们的 CPU 份额设置为**正常**，因此各得 4GHz。
- 现在打开了第三个受 CPU 约束的虚拟机的电源。它的 CPU 份额设置为**高**，这意味着它拥有的份额值应该是设置为**正常**的虚拟机的两倍。新的虚拟机获得 4GHz，其他两个虚拟机各自仅获得 2GHz。如果用户为第三个虚拟机指定的自定义份额值为 2000，也会出现相同的结果。

资源分配预留

预留指定保证为虚拟机分配的最少资源量。

仅在有足够的未预留资源满足虚拟机的预留时，vCenter Server 或 ESXi 才允许您打开虚拟机电源。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。

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

预留默认为 0。可以指定预留以保证虚拟机始终可使用最少的必要 CPU 或内存量。

资源分配限制

限制功能为可以分配到虚拟机的 CPU、内存或存储 I/O 资源指定上限。

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，即使系统上有未使用的资源也是如此。限制用具体单位（兆赫兹 (GHz) 或兆字节 (MB) 或每秒 I/O 操作数）表示。

CPU、内存和存储 I/O 资源限制默认为无限制。如果内存无限制，则在创建虚拟机时为该虚拟机配置的内存量将成为其有效限制因素。

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，并且您想对用户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数量增加，性能将会降低。因此，您可以通过指定限制来模拟减少可用资源。
- 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，即使系统未充分利用并且有闲置资源可用时也是如此。请仅在充分理由的情况下指定限制。

资源分配设置建议

选择适合 ESXi 环境的资源分配设置（预留、限制和份额）。

遵循以下准则有助于使虚拟机获得更好性能。

- 使用**预留**来指定可接受的最低 CPU 量或内存量，而不是想要使用的量。预留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。主机可以根据虚拟机的限制、份额的数量和估计需求将额外的资源指定为可用资源。
- 请不要将所有资源全部指定为虚拟机的预留（请计划将至少 10% 的资源保留为未预留）。系统容量越接近于被全部预留，想要在不违反接入控制的情况下更改预留和资源池层次结构就越困难。在支持 DRS 的集群内，如果预留完全占用集群或集群内各台主机的容量，则会阻止 DRS 在主机之间迁移虚拟机。
- 如需频繁更改总可用资源，可使用**份额**在虚拟机之间合理分配资源。例如，如果使用**份额**，并且升级主机，那么，即使每个份额代表较大的内存量、CPU 量或存储 I/O 资源量，每个虚拟机也保持相同的优先级（保持相同数量的份额）。

编辑设置

可以使用“编辑设置”对话框更改内存和 CPU 资源的分配。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
- 2 右键单击并选择**编辑设置**。
- 3 编辑 CPU 资源。

选项	描述
份额	此资源池拥有的、相对于父级的总 CPU 份额值。同级资源池根据其预留和限制限定的相对份额值共享资源。选择 低 、 正常 或 高 ，这三种级别分别按 1:2:4 这个比率指定份额值。选择 自定义 可为每个虚拟机提供表示比例权重的特定份额数。
预留	保证为该资源池分配的 CPU 量。
限制	该资源池的 CPU 分配上限。选择 无限 可指定无上限。

4 编辑内存资源。

选项	描述
份额	此资源池拥有的、相对于父级的总内存份额值。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。选择 低 、 正常 或 高 ，这三种级别分别按 1:2:4 这个比率指定份额值。选择 自定义 可为每个虚拟机提供表示比例权重的特定份额数。
预留	保证为该资源池分配的内存量。
限制	该资源池的内存分配上限。选择 无限 可指定无上限。

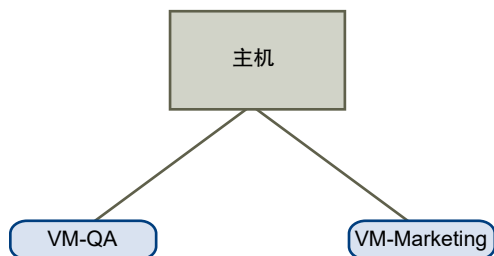
5 单击**确定**。

更改资源分配设置 — 示例

以下示例说明了如何更改资源分配设置以提高虚拟机性能。

假定在某个 ESXi 主机上，您创建了两个新的虚拟机，一台用于 QA (VM-QA) 部门，另一台用于市场 (VM-Marketing) 部门。

图 2-1. 具有两个虚拟机的单台主机



在接下来的示例中，假定 VM-QA 占用大量内存，因此，您需要将这两个虚拟机的资源分配设置相应地更改为以下内容：

- 指定当系统内存过载时，VM-QA 可使用的 CPU 和内存资源是市场部虚拟机的两倍。将 VM-QA 的 CPU 份额和内存份额设置为**高**，并将 VM-Marketing 设置为**正常**。
- 保证市场部虚拟机具有一定量的 CPU 资源。您可以使用预留设置来达到此目的。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
- 2 在要更改其份额的虚拟机上，右键单击 **VM-QA**，然后选择**编辑设置**。
- 3 在**虚拟硬件**下，展开“CPU”，然后从**共享**下拉菜单中选择**高**。
- 4 在**虚拟硬件**下，展开“内存”，然后从**共享**下拉菜单中选择**高**。
- 5 单击**确定**。
- 6 右键单击市场部虚拟机 (**VM-Marketing**)，然后选择**编辑设置**。
- 7 在**虚拟硬件**下，展开“CPU”，然后将**预留**值更改为所需值。

8 单击**确定**。

准入控制

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确定是否可保证为虚拟机所配置的预留（如果有）。此过程称为准入控制。

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。否则将显示一条资源不足警告。

注 除用户指定的内存预留外，各虚拟机还有一个开销内存量。此额外内存使用量包含在准入控制计算中。

启用了 vSphere DPM 功能时，可能会将主机置于待机模式（即关闭电源）以降低功耗。这些主机所提供的未预留资源将被视为可用于准入控制的资源。如果某个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。有关详细信息，请参见[管理电源资源](#)。

CPU 虚拟化基本知识

3

CPU 虚拟化着重于性能，只要有可能就会直接在处理器上运行。只要有可能就会使用基础物理资源，且虚拟化层仅在需要时才运行指令，使得虚拟机就像直接在物理机上运行一样。

CPU 虚拟化与仿真不同。ESXi 不使用仿真来运行虚拟 CPU。采用仿真时，所有操作均由仿真器在软件中运行。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，并在几个不同平台上运行针对一个平台而设计的软件。

CPU 资源超额分配时，ESXi 主机将在所有虚拟机之间对物理处理器进行时间划分，以便每个虚拟机在运行时就如同具有指定数目的虚拟处理器一样。运行多个虚拟机的 ESXi 主机会为各虚拟机分配一定份额的物理资源。如果使用默认资源分配设置，与同一主机关联的所有虚拟机都将在每个虚拟 CPU 上收到相同份额的 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。

本章讨论了以下主题：

- 基于软件的 CPU 虚拟化
- 硬件辅助的 CPU 虚拟化
- 虚拟化和特定于处理器的行为
- CPU 虚拟化的性能影响

基于软件的 CPU 虚拟化

采用基于软件的 CPU 虚拟化后，客户机应用程序代码直接在处理器上运行，同时转换客户机特权代码并在处理器上运行转换后的代码。

转换后的代码有点大，通常比本机版本的运行速度慢。因此，具有少量特权代码组件的客户机应用程序的运行速度与本机应用程序非常接近。而具有大量特权代码组件（如系统调用、陷阱或页面表更新）的应用程序在虚拟环境中的运行速度可能较慢。

硬件辅助的 CPU 虚拟化

某些处理器为 CPU 虚拟化提供硬件辅助。

使用此辅助时，客户机可以使用独立的执行模式（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而进入 root 模式。管理程序将在 root 模式中执行，确定退出的原因，采取任何必需的措施，并在客户机模式中重新启动客户机。

将硬件辅助用于虚拟化时，不需要再转换代码。因此，系统调用或陷阱密集型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而进入 root 模式。根据退出的次数和退出所用的总时间，硬件辅助的 CPU 虚拟化可明显提高执行的速度。

虚拟化和特定于处理器的行为

尽管 VMware 软件会虚拟化 CPU，虚拟机仍然能检测出它在其上运行的处理器的具体型号。

处理器型号可能在其提供的 CPU 功能方面不同，在虚拟机中运行的应用程序可以利用这些功能。因此，无法使用 vMotion® 在具有不同功能集的处理器上运行的系统之间迁移虚拟机。在某些情况下，通过将增强型 vMotion 兼容性 (EVC) 用于支持此功能的处理器，可以避免此限制。有关更多信息，请参见《vCenter Server 和主机管理》文档。

CPU 虚拟化的性能影响

根据工作负载和使用的虚拟化类型，CPU 虚拟化会增加不同的开销量。

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。对于此类应用程序，CPU 虚拟化开销包括必须执行的额外指令。此开销消耗应用程序本身可以使用的 CPU 处理时间。CPU 虚拟化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 虚拟化可能会提高 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。

在每台虚拟机上，ESXi 最多支持 128 个虚拟处理器 (CPU)。

注 在单处理器虚拟机（而不是带有多个 CPU 的 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 使用本该由其他虚拟机以其他方式使用的物理资源。

管理 CPU 资源

4

可以为虚拟机配置一个或多个虚拟处理器，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。

注 在本章中，“内存”可以指物理内存或永久内存。

本章讨论了以下主题：

- [查看处理器信息](#)
- [指定 CPU 配置](#)
- [多核处理器](#)

查看处理器信息

您可在 vSphere Client 中访问关于当前 CPU 配置的信息。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 在**硬件**下，展开 **CPU** 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。

注 在超线程系统中，每个硬件线程都是一个逻辑处理器。例如，激活了超线程的双核处理器具有两个内核和四个逻辑处理器。

指定 CPU 配置

可以通过指定 CPU 配置来改进资源管理。但是，如果未自定义 CPU 配置，则 ESXi 主机会使用适合大多数情况的默认值。

可以按以下方式指定 CPU 配置：

- 使用可通过 vSphere Client 访问的属性和特殊功能。使用 vSphere Client 可连接 ESXi 主机或 vCenter Server 系统。
- 在某些情况下使用高级设置。
- 将 vSphere SDK 用于脚本式 CPU 分配。

- 使用超线程。

多核处理器

多核处理器为执行虚拟机多任务的主机提供了很多优势。

注 在本主题中，“内存”可以指物理内存或永久内存。

Intel 和 AMD 均已开发将两个或更多处理器内核组合到单个集成电路（通常称为封装件或插槽）的处理器。VMware 使用“插槽”一词来描述单个封装件，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。

例如，双核处理器通过允许同时运行两个虚拟 CPU，可以提供几乎是单核处理器两倍的性能。同一处理器中的内核通常配备由所有内核使用的最低级别的共享缓存，这有可能会减少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争用相同内存总线资源且占用大量内存的工作负载，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。

ESXiCPU 调度程序独立使用每个处理器内核的每个逻辑处理器来运行虚拟机，从而提供与 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，或运行在不同物理内核的逻辑处理器上。

ESXiCPU 调度程序可以检测处理器拓扑，以及处理器内核与它上面的逻辑处理器之间的关系。它使用此信息来调度虚拟机和优化性能。

ESXiCPU 调度程序可以解释处理器拓扑（包括插槽、内核和逻辑处理器之间的关系）。调度程序使用拓扑信息来优化将虚拟 CPU 放置到不同插槽的过程。此优化可以最大程度地提高整体缓存使用率，并通过最大程度减少虚拟 CPU 迁移来提高缓存关联性。

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。为了避免将逻辑处理器和物理处理器混淆，Intel 将物理处理器称为插槽，本章的讨论也使用这一术语。

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。超线程技术允许单个处理器内核同时执行两个独立的线程。

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，使得某些重要的工作负载类型产生更大的吞吐量。如果应用程序运行在忙碌内核的一个逻辑处理器上，则与单独运行在非超线程处理器上相比，预期获得的吞吐量会稍高于一半。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）。

注 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（如内存缓存和功能单元）。此类逻辑处理器通常称为线程。

许多处理器都不支持超线程，因此每个内核仅具有一个线程。对于此类处理器，内核数目还与逻辑处理器的数目相匹配。以下处理器支持超线程，并且每个内核具有两个线程。

- 基于 Intel Xeon 5500 处理器微架构的处理器。
- Intel Pentium 4（支持 HT）
- Intel Pentium EE 840（支持 HT）

本章讨论了以下主题：

- [超线程和 ESXi 主机](#)
- [启用超线程](#)

超线程和 ESXi 主机

支持超线程的主机应具有与没有超线程的主机类似的行为。但是，如果启用超线程，则可能需要考虑某些因素。

ESXi 主机以智能方式管理处理器时间，保证负载均匀分布在系统的多个处理器内核上。相同内核上的逻辑处理器具有连续的 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。优先在两个不同的内核上调度虚拟机，然后才选择在同一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，从而释放其执行资源并允许在同一内核的另一个逻辑处理器上运行的虚拟机使用该内核的全部执行资源。VMware 调度程序会正确地考虑此暂停时间，因此使用全部内核资源运行的虚拟机的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 ESXi 资源分配规则。

在使用超线程的主机上启用 CPU 关联性之前，请考虑资源管理需求。例如，如果将高优先级虚拟机绑定到 CPU 0，并将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。请确保所有的自定义关联性设置对超线程系统都有意义。

启用超线程

要启用超线程，必须首先在系统的 BIOS 设置中将其启用，然后在 vSphere Client 中打开它。超线程在默认情况下处于启用状态。

请查阅系统文档，确定您的 CPU 是否支持超线程。

步骤

- 1 请确保您的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记为**逻辑处理器**，而有些制造商则称之为**启用超线程**。
- 3 确保为 ESXi 主机启用超线程。
 - a 在 vSphere Client 中，浏览到主机。
 - b 单击**配置**。
 - c 在**系统**下，单击**高级系统设置**，然后选择 **VMkernel.Boot.hyperthreading**。
必须重新启动主机，才能使设置生效。如果值为**有效**，将启用超线程。
- 4 在**硬件**下，单击**处理器**以查看逻辑处理器的数量。

结果

超线程已启用。

使用 CPU 关联性

6

通过为每个虚拟机指定 CPU 关联性设置，可以仅将虚拟机只分配给多处理器系统中的某个可用处理器子集。通过使用此功能，可以将每个虚拟机分配到指定关联性集合中的处理器。

CPU 关联性指定虚拟机到处理器的放置位置的限制，与由虚拟机-虚拟机或虚拟机-主机关联性规则创建的关系不同，后一关联性规则指定虚拟机到虚拟机主机的放置位置的限制。

在这个上下文中，术语“CPU”指的是超线程系统上的逻辑处理器，同时也指非超线程系统上的内核。

某一虚拟机的 CPU 关联性设置适用于与该虚拟机相关联的所有虚拟 CPU 及其他所有线程（也叫做“环境”）。这些虚拟机线程可执行仿真鼠标、键盘、屏幕、CD-ROM 及其他旧设备时所需进行的处理工作。

在某些情况下（例如，占用大量显示资源的工作负载），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了这些额外的线程与虚拟机的虚拟 CPU 同时进行调度，则性能可能会降低。例如，单处理器虚拟机与单个 CPU 关联，或双路 SMP 虚拟机仅与两个 CPU 关联。

为了获得最佳性能，在应用手动关联性设置时，VMware 建议您在关联性设置中至少要包含一个额外的物理 CPU，以便允许至少有一个虚拟机线程与其虚拟 CPU 同时调度。例如，单处理器虚拟机至少与两个 CPU 关联，或双路 SMP 虚拟机至少与三个 CPU 关联。

本章讨论了以下主题：

- [向特定处理器分配虚拟机](#)
- [CPU 关联性的潜在问题](#)

向特定处理器分配虚拟机

使用 CPU 关联性，可以向特定处理器分配虚拟机。通过此操作，可以将虚拟机只分配给多处理器系统中特定的可用处理器。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 选择**虚拟机**。
- 2 右键单击虚拟机，然后单击**编辑设置**。
- 3 在“虚拟硬件”下，展开 **CPU**。

- 4 在“调度关联性”下，选择虚拟机的物理处理器关联性。

使用“-”表示范围，使用“,”分隔值。

例如，“0, 2, 4-7”表示处理器 0、2、4、5、6 和 7。

- 5 选择要运行虚拟机的处理器，然后单击**确定**。

CPU 关联性的潜在问题

使用 CPU 关联性之前，可能需要考虑某些问题。

CPU 关联性的潜在问题包括：

- 对于多处理器系统，ESXi 系统执行自动负载均衡。避免手动指定虚拟机关联性，以改进调度程序跨处理器均衡负载的能力。
- 关联性可能会干扰 ESXi 主机满足为虚拟机指定的预留和份额的能力。
- 因为 CPU 准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 调度程序可能无法管理已经借助于关联性分配到某些处理器的虚拟机。有关详细信息，请参见[第 21 章 配合使用 NUMA 系统和 ESXi](#)。
- 关联性可能会影响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。

主机电源管理策略

7

可以在 ESXi 中应用主机硬件提供的多个电源管理功能来调整性能与电源之间的均衡。可以通过选择电源管理策略来控制 ESXi 使用这些功能的方式。

选择高性能策略可提供更多绝对性能，但每瓦特的效率和性能较低。低功耗策略提供的绝对性能较少，但效率较高。

可以使用 VMware Host Client 为管理的主机选择策略。如果未选择策略，则 ESXi 默认使用“均衡”策略。

表 7-1. CPU 电源管理策略

电源管理策略	描述
高性能	不使用任何电源管理功能。
均衡（默认值）	在对性能影响最小的情况下，减少能量消耗
低功耗	在可能降低性能的情况下，减少能量消耗
自定义	用户定义的电源管理策略。高级配置将变得可用。

当 CPU 以较低频率运行时，其运行电压也较低，这样便可省电。这种类型的电源管理通常叫做动态电压和频率缩放 (DVFS)。ESXi 会尝试调整 CPU 频率，以便不影响虚拟机性能。

当 CPU 空闲时，ESXi 可以应用深层级暂停状况（称为 C 状况）。C 状况层级越深，CPU 使用的电源就越少，但 CPU 重新开始运行的用时越长。当 CPU 变为空闲时，ESXi 会应用算法，以便预测空闲状况的持续时间并选择要进入的相应 C 状况。在不使用深层级 C 状况的电源管理策略中，ESXi 对空闲 CPU 仅使用最浅层级的暂停状况 (C1)。

本章讨论了以下主题：

- [选择 CPU 电源管理策略](#)
- [为主机电源管理配置自定义策略参数](#)

选择 CPU 电源管理策略

您可以使用 vSphere Client 为主机设置 CPU 电源管理策略。

前提条件

请确认主机系统上的 BIOS 设置允许操作系统控制电源管理（如 **OS Controlled**）。如果主机硬件不允许操作系统管理电源，则只有“不受支持”策略可用。（在某些系统上，仅“高性能”策略可用。）

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在“硬件”下，选择**电源管理**，然后单击**编辑**按钮。
- 4 为主机选择一种电源管理策略，然后单击**确定**。

所选策略保存在主机配置中，可以在引导时再次使用。您可以随时更改该策略，而不需要重新引导服务器。

为主机电源管理配置自定义策略参数

当为主机电源管理使用自定义策略时，ESXi 将其电源管理策略建立在若干高级配置参数值的基础之上。

前提条件

如[选择 CPU 电源管理策略](#)中所述，为电源管理策略选择**自定义**。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**系统**下，选择**高级系统设置**。
- 4 在右侧窗格中，可以编辑影响自定义策略的电源管理参数。

影响自定义策略的电源管理参数的描述以在“**自定义**”策略中开始。所有其他电源参数影响所有电源管理策略。

- 5 选择参数，然后单击**编辑**按钮。

注 电源管理参数的默认值与“平衡”策略匹配。

参数	描述
Power.UsePStates	处理器忙时，请使用 ACPI P 状态来节省电源。
Power.MaxCpuLoad	仅当 CPU 忙碌时间少于实际时间的给定百分比时，才使用 P 状态来节省 CPU 电源。
Power.MinFreqPct	不要使用任何低于 CPU 全速的给定百分比的 P 状态。

参数	描述
Power.UseStallCtr	当处理器频繁停止以等待缓存未命中事件时，请使用更深的 P 状态。
Power.TimerHz	控制 ESXi 重新评估每个 CPU 要处于哪种 P 状态的频率（次数/秒）。
Power.UseCStates	当处理器处于空闲状态时，请使用深 ACPI C 状态（C2 或更低）。
Power.CStateMaxLatency	不要使用其延迟时间大于此值的 C 状态。
Power.CStateResidencyCoef	当 CPU 变为空闲时，选择其延迟时间与此值的乘积小于主机的 CPU 预计空闲时间的最深的 C 状态。值越大，ESXi 愈加保守地使用深 C 状态；值越小，ESXi 愈加主动地使用深 C 状态。
Power.CStatePredictionCoef	ESXi 算法中的一个参数，用于预测变为空闲的 CPU 保持空闲状态的时间。不建议更改此值。
Power.PerfBias	性能能量偏差提示（仅适用于 Intel）。将 Intel 处理器的 MSR 设置为 Intel 建议的值。Intel 建议高性能使用 0，平衡配置使用 6，低功耗使用 15。其他值均未定义。

6 单击确定。

内存虚拟化基本知识

8

在管理内存资源之前，应当了解 ESXi 是如何虚拟化和使用这些内存资源的。

VMkernel 管理主机上所有物理内存。VMkernel 会将这种受管物理内存的一部分拿来自己使用。剩余的内存可供虚拟机使用。

虚拟和物理内存空间划分为块，块也称为页。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。根据处理器架构的不同，页通常为 4 KB 或 2 MB。请参见[高级内存属性](#)。

本章讨论了以下主题：

- [虚拟机内存](#)
- [内存过载](#)
- [内存共享](#)
- [内存虚拟化](#)
- [硬件辅助的内存虚拟化](#)
- [支持大内存页](#)

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化。

配置大小是提供给客户机操作系统的内存量。这与分配给虚拟机的物理内存量不同。后者取决于主机上的资源设置（份额、预留和限制）和内存压力级别。

例如，请考虑配置大小为 1GB 的虚拟机。当客户机操作系统引导时，系统会检测到它正运行在具有 1 GB 物理内存的专用计算机上。有些情况下，可能向虚拟机分配全部内容（即 1GB）。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，客户机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，则会为虚拟机指定相对优先级。

预留

主机保证为虚拟机预留的物理内存量下限，即使内存过载也是如此。将预留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。

在虚拟机消耗其预留的全部内存后，会允许其保留该内存量，并且不会将该内存回收，即使该虚拟机闲置也是如此。某些客户机操作系统（例如 Linux）在引导之后可能不会立即访问所配置的全部内存。在虚拟机消耗其预留的全部内存之前，VMkernel 可以将其预留的任何未使用部分分配给其他虚拟机。但是，在客户机的工作负载增加并且虚拟机消耗其全部预留之后，允许其保留此内存。

限制

主机可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的隐式限制。

内存过载

对于每个正在运行的虚拟机，系统会为虚拟机的预留（如果有）和虚拟化开销预留物理内存。

所有虚拟机的已配置内存大小总量可能超过了主机上的可用物理内存量。但是，这并不一定意味着内存已过载。当所有虚拟机的组合工作内存占用超过主机内存大小的组合工作内存占用时，内存已过载。

由于 ESXi 主机使用内存管理技术，虚拟机可以使用的虚拟内存大于主机上可用的物理内存。例如，您有一个内存为 2 GB 的主机，其上运行四个虚拟机，每个虚拟机的内存为 1 GB。这种情况下，内存会过载。例如，如果所有 4 个虚拟机均闲置，则组合消耗内存可能远低于 2GB。但是，如果所有 4GB 虚拟机均主动消耗内存，则其内存占用可能超过 2GB，并且 ESXi 主机将过载。

过载有一定的意义，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，ESXi 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”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。ESXi 实施了多种机制（如膨胀、内存共享、内存压缩和交换）来提供合理性能，即使主机尚未严重内存过载。

如果虚拟机在内存过载环境下消耗了所有可预留内存，ESXi 主机可能会内存不足。虽然已打开电源的虚拟机不受影响，但新虚拟机可能会由于内存不足而无法打开电源。

注 所有虚拟机内存开销也被视为预留。

此外，ESXi 主机上默认还会启用内存压缩，以在内存过载时提高虚拟机性能，如[内存压缩](#)中所述。

内存共享

内存共享是一项专用的 ESXi 技术，有助于增加主机上的内存密度。

内存共享取决于以下观察结果：几个虚拟机可能正在运行同一客户机操作系统的多个实例。这些虚拟机可能已加载相同的应用程序或组件，或者包含公用数据。这些情况下，主机使用专用的透明页面共享 (TPS) 技术消除内存页的冗余副本。采用内存共享后，在虚拟机中运行的工作负载消耗的内存通常要少于其在物理机上运行时可能需要的内存。因此，可以高效地支持更高级别的超额分配。通过内存共享节省的内存量取决于工作负载是否由几乎相同的虚拟机组成，这些虚拟机可能会释放更多内存。如果工作负载差异较大，则可能会导致节省的内存百分比明显降低。

注 出于安全考虑，默认情况下，虚拟机间透明页面共享处于停用状态，并且页面共享限于虚拟机内部内存共享。页面共享不能在多个虚拟机间进行，只能在虚拟机内部进行。有关详细信息，请参见[在虚拟机之间共享内存](#)。

内存虚拟化

因为虚拟化引入了额外级别的内存映射，所以 ESXi 可以跨所有虚拟机来管理内存。

虚拟机的一些物理内存可能映射到共享页面或未映射或换出的页面。

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

每个虚拟机的 VMM 保持了从客户机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。

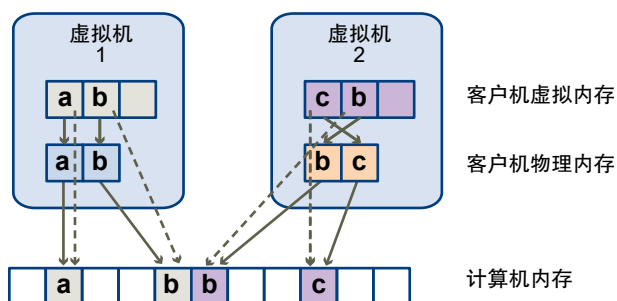
（VMware 将基础主机物理页称为“计算机”页，将客户机操作系统的物理页称为“物理”页。）

每个虚拟机均有连续的可寻址物理内存空间，该空间从零开始。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。

客户机虚拟地址到客户机物理地址的转换由客户机操作系统管理。管理程序仅负责将客户机物理地址转换为计算机地址。硬件辅助的内存虚拟化将利用硬件设施生成具有由管理程序维护的客户机页表和嵌套页表的组合映射。

该图说明了 ESXi 如何实施内存虚拟化。

图 8-1. ESXi 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机操作系统中的页表所保持的映射。（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由 VMM 保持的映射。
- 虚线箭头表示从客户机虚拟内存到计算机内存的映射，该映射也由 VMM 保持。运行虚拟机的基础处理器使用卷影页表映射。

硬件辅助的内存虚拟化

类似于 AMD SVM-V 和 Intel Xeon 5500 系列之类的部分 CPU 通过使用两层页表来提供对内存虚拟化的硬件支持。

注 在本主题中，“内存”可以指物理内存或永久内存。

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。TLB（translation look-aside buffer，转换旁视缓冲区）是由处理器的内存管理单元 (MMU) 硬件维护的转换缓存。TLB 缺失是此缓存中的缺失，而且硬件需要访问内存（可能是多次）来查找所需转换。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成计算机地址。第一层页表由客户机操作系统维护。VMM 仅维护第二层页表。

性能注意事项

使用硬件辅助时，会消除软件内存虚拟化的开销。特别是，硬件辅助消除了使卷影页表与客户机页表保持同步所需的开销。但是，使用硬件辅助时 TLB 缺失延迟时间明显较长。默认情况下，管理程序在硬件辅助模式下使用大页以减少 TLB 缺失的成本。因此，工作负载是否受益于硬件辅助主要取决于在使用软件内存虚拟化时由内存虚拟化引起的开销。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），则软件虚拟化不会引起显著开销。相反，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。

默认情况下，管理程序在硬件辅助模式下使用大页以减少 TLB 缺失的成本。通过在客户机虚拟到客户机物理以及客户机物理到计算机地址转换中使用大页，可以实现最佳性能。

LPage.LPageAlwaysTryForNPT 选项可以更改在客户机物理到计算机地址转换中使用大内存页的策略。有关详细信息，请参见[高级内存属性](#)。

支持大内存页

ESXi 为大内存页提供有限的支持。

x86 架构允许系统软件使用 4KB、2MB 和 1GB 页面。我们将 4KB 页面称为小内存页，而将 2MB 和 1GB 页面称为大内存页。大内存页可缓解旁路转换缓冲 (TLB) 压力，降低页表遍历开销，从而提高工作负载的性能。

在虚拟化环境中，Hypervisor 和客户机操作系统可独立使用大内存页。虽然客户机和 Hypervisor 都使用大内存页时可实现最大的性能影响，但在大多数情况下，即使仅在 Hypervisor 级别使用大内存页也可以观察到性能影响。

默认情况下 ESXi Hypervisor 使用 2MB 页面来备份客户机 vRAM。vSphere ESXi 支持使用 1GB 页面备份客户机 vRAM，但提供的支持有限。有关详细信息，请参见[使用 1GB 页面备份客户机 vRAM](#)。

管理内存资源

9

使用 vSphere Client，可以查看有关内存分配设置的信息并对其进行更改。为了有效管理内存资源，还必须熟悉内存开销、闲置内存消耗以及 ESXi 主机回收内存的方式。

当管理内存资源时，可以指定内存分配。如果未自定义内存分配，则 ESXi 主机使用适合大多数情况下的默认值。

可以通过几种方式指定内存分配。

- 使用可通过 vSphere Client 访问的属性和特殊功能。通过 vSphere Client，可以连接到 ESXi 主机或 vCenter Server 系统。
- 使用高级设置。
- 将 vSphere SDK 用于脚本式内存分配。

注 在本章中，“内存”可以指物理内存或永久内存。

本章讨论了以下主题：

- 了解内存开销
- 虚拟机上的开销内存
- ESXi 主机如何分配内存
- 闲置虚拟机的内存消耗
- VMX 交换文件
- 内存回收
- 内存气球驱动程序
- 在虚拟机之间共享内存
- 内存压缩
- 激活或停用内存压缩缓存
- 设置内存压缩缓存的最大大小
- 衡量和区分各种内存使用情况
- 内存可靠性
- 更正错误隔离通知

- 关于系统交换
- 配置系统交换

了解内存开销

内存资源的虚拟化会涉及一些相关开销。

ESXi 虚拟机可以引起两种内存开销：

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

ESXi 内存虚拟化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页表（基于软件的卷影页表方法或硬件辅助的两级页表方法），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。

内存空间开销有两部分：

- VMkernel 系统范围内的固定开销。
- 每个虚拟机的额外开销。

开销内存包括为虚拟机框架缓冲区和各种虚拟化数据结构（如卷影页表）预留的空间。开销内存取决于虚拟 CPU 数量以及为客户机操作系统配置的内存。

虚拟机上的开销内存

要打开虚拟机电源，需要一定数量的可用开销内存。您应当了解此开销量。

下表列出了打开虚拟机电源所需的开销内存量。当虚拟机开始运行之后，该虚拟机所使用的开销内存量可能不同于表中列出的数量。通过启用虚拟机的 VMX 交换功能以及启用硬件 MMU 来收集示例值。（默认情况下启用 VMX 交换功能。）

注 下表提供的是开销内存值的示例，并不尝试提供有关所有可能的配置的信息。您可以将虚拟机配置为最多包含 64 个虚拟 CPU，具体取决于主机上许可的 CPU 数、客户机操作系统支持的 CPU 数。

表 9-1. 虚拟机上的示例开销内存

内存 (MB)	1 个 VCPU	2 个 VCPU	4 个 VCPU	8 个 VCPU
256	20.29	24.28	32.23	48.16
1024	25.90	29.91	37.86	53.82
4096	48.64	52.72	60.67	76.78
16384	139.62	143.98	151.93	168.60

ESXi 主机如何分配内存

主机将由 `Limit` 参数指定的内存分配给每个虚拟机，除非内存过载。ESXi 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，ESXi 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。

当内存过载时，向每个虚拟机分配的内存量介于**预留**和**限制**指定的内存量之间。授予虚拟机的高于预留量的内存量会因当前的内存负载而异。

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** — ESXi 主机使用经过修改的按比例份额内存分配策略。内存份额给予虚拟机一部分可用物理内存。
- **工作集大小** — ESXi 主机通过在连续的虚拟机执行时间周期监控内存活动，来估计工作集。采用快速响应工作集大小增加且慢速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。

在默认情况下将对内存活动监控 60 秒以估计工作集大小。要修改此默认值，请调整 `Mem.SamplePeriod` 高级设置。请参见[设置高级主机属性](#)。

闲置虚拟机的内存消耗

如果虚拟机未在使用当前为其分配的所有内存，则 ESXi 对闲置内存的消耗量大于对正在使用的内存的消耗量。这样有助于防止虚拟机累积闲置内存。

闲置内存消耗以渐进方式应用。随着虚拟机闲置内存与活动内存的比率的提高，有效消耗率将增加。（在不支持分层资源池的早期版本 ESXi 中，虚拟机的所有闲置内存是以同等比率消耗的。）

可以使用 `Mem.IdleTax` 选项来修改闲置内存消耗率。使用该选项以及 `Mem.SamplePeriod` 高级属性可控制系统如何确定虚拟机的目标内存分配。请参见[设置高级主机属性](#)。

注 在大多数情况下，没有必要更改 `Mem.IdleTax`，如果更改的话，反而不合适。

VMX 交换文件

使用虚拟机可执行 (VMX) 交换文件，主机可大幅减少为 VMX 进程预留的开销内存量。

注 VMX 交换文件与交换到主机交换缓存功能或常规主机级别交换文件不相关。

ESXi 出于多种原因预留每个虚拟机的内存。打开虚拟机电源时，将完全预留特定组件（如虚拟机监控程序 (VMM) 和虚拟设备）所需的内存。但可以交换为 VMX 进程预留的一些开销内存。VMX 交换功能大大减少了 VMX 内存预留（例如，从每个虚拟机大约 50 MB 或更多减少为每个虚拟机大约 10 MB）。这样可在主机内存过载时换出剩余内存，从而减少了每个虚拟机的开销内存预留。

如果打开虚拟机电源时有足够的可用磁盘空间，主机便可自动创建 VMX 交换文件。

内存回收

ESXi 主机可以从虚拟机中回收内存。

主机将预留功能指定的内存量直接分配给虚拟机。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，则使用膨胀或交换等特殊技术进行处理。主机可使用两种技术来动态增加或减少分配给虚拟机的内存量：

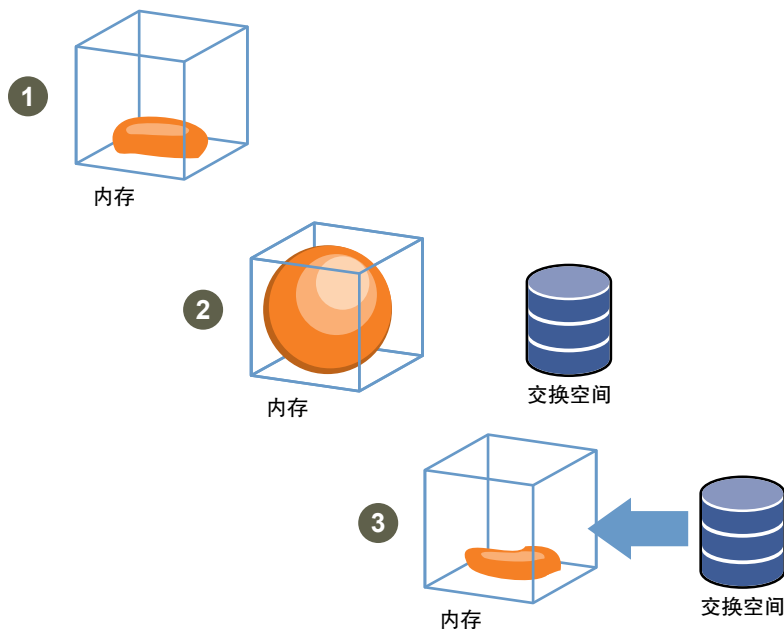
- ESXi 系统使用已加载到虚拟机中所运行的客户机操作系统的内存气球驱动程序 (vmmemctl)。请参见[内存气球驱动程序](#)。
- ESXi 系统将分页从虚拟机换出到服务器交换文件，无需客户机操作系统参与。每个虚拟机均有自己的交换文件。

内存气球驱动程序

内存气球驱动程序 (vmmemctl) 与服务器协作回收客户机操作系统认为最不重要的页面。

该驱动程序使用专用膨胀技术，提供在类似的内存限制下与本机系统的行为极为相近的可预测性能。该技术可增加或减少客户机操作系统的内存压力，使得客户机能够使用自己的本机内存管理算法。当内存很紧张时，客户机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。

图 9-1. 客户机操作系统中的内存膨胀



注 必须使用足够的交换空间来配置客户机操作系统。某些客户机操作系统具有其他限制。

如有必要，可以通过为特定虚拟机设置 **sched.mem.maxmemctl** 参数，限制由 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 (MB) 为单位。请参见[设置高级虚拟机属性](#)。

在虚拟机之间共享内存

许多 ESXi 工作负载存在跨虚拟机（以及在单个虚拟机中）共享内存的机会。

ESXi 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。此信息可在“内存”页面中交互模式的 `PSHARE` 字段中找到。

使用 **Mem.ShareScanTime** 和 **Mem.ShareScanGHz** 高级设置可控制系统扫描内存以确定内存共享机会的速率。

还可以通过设置 **sched.mem.pshare.enable** 选项为单个虚拟机配置共享。

出于安全考虑，默认情况下，虚拟机间透明页面共享处于停用状态，并且页面共享限于虚拟机内部内存共享。这意味着页面共享不会在多个虚拟机间出现，而是仅发生在虚拟机内部。为帮助解决系统管理员对透明页面共享所造成安全影响可能存在的疑问，我们引入了盐的概念。通过使用盐，可以前所未有的方式更加细化地管理参与透明页面共享的虚拟机。在新的盐设置中，仅当页面的加密盐值和内容完全相同时，虚拟机才可共享页面。新主机配置选项 **Mem.ShareForceSalting** 可配置为激活或停用撒盐加密。

有关如何设置高级选项的信息，请参见第 22 章 高级属性。

内存压缩

ESXi 提供内存压缩缓存，可在内存超额分配使用时改进虚拟机性能。内存压缩默认处于激活状态。当主机内存超额分配时，ESXi 会压缩虚拟页面并将其存储在内存中。

由于访问压缩的内存比访问交换到磁盘的内存更快，因此通过 ESXi 中的内存压缩可以使内存超额分配，但不会显著影响性能。当需要交换虚拟页面时，ESXi 会首先尝试压缩虚拟页面。可压缩至 2 KB 或更小的页面存储在虚拟机的压缩缓存中，从而增加主机的容量。

使用 vSphere Client 中的“高级设置”对话框，您可以设置压缩缓存的最大大小和停用内存压缩。

激活或停用内存压缩缓存

内存压缩默认处于激活状态。您可以使用 vSphere Client 中的“高级系统设置”来激活或停用主机的内存压缩。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**系统**下，选择**高级系统设置**。
- 4 找到“Mem.MemZipEnable”，然后单击**编辑**按钮。
- 5 输入 1 和 0 可分别激活和停用内存压缩缓存。
- 6 单击**确定**。

设置内存压缩缓存的最大大小

您可以设置主机虚拟机的内存压缩缓存的最大大小。

您可以将压缩缓存的大小设置为虚拟机的内存大小百分比。例如，如果输入 20 并且虚拟机的内存大小为 1000 MB，则 ESXi 最多可使用 200 MB 的主机内存来存储虚拟机的已压缩页面。

如果您未设置压缩缓存的大小，ESXi 会使用默认值 10%。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**系统**下，选择**高级系统设置**。
- 4 找到“Mem.MemZipMaxPct”，然后单击**编辑**按钮。

此属性的值确定虚拟机的压缩缓存的最大大小。

- 5 输入压缩缓存的最大大小。

此值是虚拟机大小的百分比并且必须介于 5% 和 100% 之间。

- 6 单击**确定**。

衡量和区分各种内存使用情况

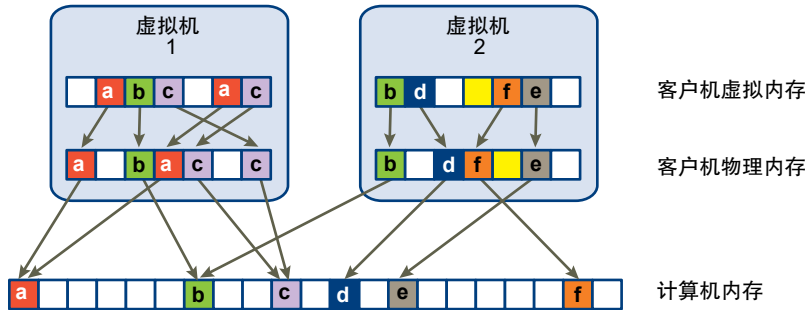
vSphere Client 的**性能**选项卡将显示可用于分析内存使用情况的多个衡量指标。

某些内存衡量指标用于衡量客户机物理内存，而另一些衡量指标用于衡量计算机内存。例如，可以使用性能衡量指标检查的两种内存使用情况是客户机物理内存和计算机内存。可以使用“已分配的内存”衡量指标（对于虚拟机）或“共享的内存”（对于主机）衡量客户机物理内存。但是，要衡量计算机内存，需要使用“消耗的内存”（对于虚拟机）或“共享的公用内存”（对于主机）。了解这些类型的内存使用情况之间的概念性差异对知道这些衡量指标的衡量对象以及如何对其进行解释十分重要。

VMkernel 会将客户机物理内存映射到计算机内存，但是它们不总是一对一映射。它可能会将客户机物理内存的多个区域映射到计算机内存的同一区域（当存在内存共享时），或者可能不会将客户机物理内存的特定区域映射到计算机内存（当 VMkernel 换出或膨胀客户机物理内存时）。在这些情况中，单个虚拟机或主机的客户机物理内存使用情况和计算机内存使用情况的计算有所不同。

请考虑下图中的示例，该图中显示了在一台主机上运行的两台虚拟机。每块代表 4 KB 内存，每个颜色/字母代表相应块上的数据集。

图 9-2. 内存使用情况示例



可以按照如下方式确定虚拟机的性能衡量指标：

- 要确定虚拟机 1 的“已分配的内存”（映射到计算机内存的客户机物理内存量），请计算虚拟机 1 的客户机物理内存中的块（含有指向计算机内存的箭头）的数量并乘以 4 KB。由于有 5 个块含有箭头，因此“已分配的内存”是 20 KB。
- “消耗的内存”是分配给虚拟机的计算机内存量，包括从共享的内存中节省的内存量。首先，计算计算机内存中的块（含有从虚拟机 1 的客户机物理内存指出的箭头）的数量。这样的块有三个，但有一个块与虚拟机 2 共享。因此，计算两个完整的块加上半个第三个块并乘以 4 KB，得到总计 10 KB 的“消耗的内存”。

这两个衡量指标之间的重要差异是：“已分配的内存”计算带箭头的客户机物理内存级块的数量，“消耗的内存”计算带箭头的计算机内存级块的数量。由于内存共享，这两个级别的块的数量不同，因此“已分配的内存”和“消耗的内存”也不同。通过共享或其他回收技术可以节省内存。

在确定主机的“共享的内存”和“共享的公用内存”时，会获得类似的结果。

- 主机的“共享的内存”是每个虚拟机“共享的内存”的总和。通过查看每个虚拟机的客户机物理内存并计算含有指向计算机内存块（计算机内存块本身也含有多个指向自己的箭头）的箭头的块数量，可计算共享的内存。在本示例中，这样的块有六个，因此主机的“共享的内存”是 24 KB。
- “共享的公用内存”是由虚拟机共享的计算机内存量。要确定公用内存，请查看计算机内存，并计算有多个箭头指向自身的块数量。这样的块有三个，因此“共享的公用内存”是 12 KB。

“共享的内存”涉及到客户机物理内存，即作为箭头起始点。而“共享的公用内存”涉及到计算机内存，即作为箭头的目标点。

用于衡量客户机物理内存和计算机内存的内存衡量指标可能会出现矛盾。事实上，它们衡量的是虚拟机内存使用情况的不同方面。通过了解这些衡量指标之间的差异，可以更好地利用它们来诊断性能问题。

内存可靠性

通过内存可靠性（也称为错误隔离），ESXi 可在其确定故障可能出现时以及已出现故障时停止使用部分内存。

在特定地址报告了足够的已更正错误时，ESXi 会停止使用该地址阻止已更正错误成为未更正的错误。

内存可靠性提高了 VMkernel 可靠性，与 RAM 中更正的和未更正的错误无关。通过内存可靠性，系统还可避免使用可能包含错误的内存页。

更正错误隔离通知

借助内存的可靠性，VMkernel 可停止使用接收错误隔离通知的页面。

当 VMkernel 从不可更正的内存错误中恢复，VMkernel 因大量可更正错误而注销显著比例的系统内存，或者存在大量无法注销的页面时，用户会在 vSphere Client 中收到事件。

步骤

- 1 腾出主机。
- 2 迁移虚拟机。
- 3 运行内存相关的硬件测试。

关于系统交换

系统交换是一种内存回收过程，可以利用整个系统内未使用的内存资源。

系统交换允许系统从内存使用者处（非虚拟机）回收内存。启用系统交换后，需要在回收其他进程内存的影响与将内存分配给可使用它的虚拟机的能力之间进行权衡。系统交换所需的空间量为 1 GB。

内存的回收通过将数据移出内存并写入后台存储实现。从后台存储访问数据的速度比从内存访问数据的速度慢，因此一定要仔细选择存储交换数据的位置。

ESXi 可自动确定系统交换应存储到的位置，这是**首选交换文件位置**。选择某一组选项可帮助确定存储位置。系统会选择最可行的选项。如果任何选项都不可行，则不会激活系统交换。

可用选项包括：

- 数据存储 - 允许使用指定的数据存储。请注意，无法为系统交换文件指定 vSAN 数据存储或 VMware vSphere® Virtual Volumes™ 数据存储。
- 主机交换缓存 - 允许使用部分主机交换缓存。
- 首选交换文件位置 - 允许使用为主机配置的首选交换文件位置。

配置系统交换

您可自定义用于确定系统交换位置的选项。

前提条件

在**编辑系统交换设置**对话框中选中**已启用**复选框。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**系统**下，选择**系统交换**。
- 4 单击**编辑**。

- 5 选中要启用的每个选项对应的复选框。
- 6 如果选择**数据存储**选项，请从下拉菜单中选择一个数据存储。
- 7 单击**确定**。

使用交换文件

10

可以指定客户机交换文件的位置、当内存超额分配时预留交换空间以及删除交换文件。

当 `vmmemctl` 驱动程序不可用或未响应时，ESXi 主机会使用交换从虚拟机中强制回收内存。

- 从未安装。
- 它已明确被停用。
- 未运行（例如，客户机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，但是已经达到最大膨胀大小。

当虚拟机需要页面时，标准需求分页技术会重新换入页面。

本章讨论了以下主题：

- 交换文件位置
- 为 DRS 集群启用主机-本地交换
- 为独立主机启用主机-本地交换
- 交换空间和内存过载
- 配置主机的虚拟机交换文件属性
- 配置集群的虚拟机交换文件位置
- 删除交换文件

交换文件位置

默认情况下，会在与虚拟机配置文件相同的位置中创建交换文件，该位置可以位于 VMFS 数据存储、vSAN 数据存储或 VMware vSphere® Virtual Volumes™ 数据存储中。在 vSAN 数据存储或 vVols 数据存储中，交换文件可作为独立的 vSAN 或 vVols 对象进行创建。

打开虚拟机电源时，ESXi 主机会创建交换文件。如果无法创建该文件，则无法打开虚拟机电源。除了接受默认值以外，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。

- 使用主机-本地交换在主机上指定存储在本地数据库存储。这样就可以在每个主机级别上进行交换，从而节省 SAN 上的空间。但是，对于 vSphere vMotion，可能会导致性能稍有下降，因为交换到源主机上的本地交换文件的页面必须通过网络传输到目标主机。当前无法为主机-本地交换指定 vSAN 和 vVols 数据存储。

为 DRS 集群启用主机-本地交换

主机-本地交换允许将存储在主机本地的数据存储指定为交换文件位置。可以为 DRS 集群启用主机-本地交换。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**。
- 3 在**配置**下，选择**常规**以查看交换文件位置，然后单击**编辑**对其进行更改。
- 4 选择**由主机指定的数据存储**选项，然后单击**确定**。
- 5 在 vSphere Client 中，浏览到集群中的主机之一。
- 6 单击**配置**。
- 7 在“虚拟机”下，选择**交换文件位置**。
- 8 单击“编辑”，选择要使用的本地数据存储，然后单击**确定**。
- 9 对集群中的每台主机重复 [步骤 5](#) 到 [步骤 8](#)。

结果

现在已为 DRS 集群启用主机-本地交换。

为独立主机启用主机-本地交换

主机-本地交换允许将存储在主机本地的数据存储指定为交换文件位置。可以为独立主机启用主机-本地交换。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**虚拟机**下，选择**交换文件位置**。
- 4 单击**编辑**，然后选择**所选数据存储**。
- 5 从列表中选择本地数据存储，然后单击**确定**。

结果

现在已为独立主机启用主机-本地交换。

交换空间和内存过载

必须在每个虚拟机交换文件中为任何未预留的虚拟机内存预留交换空间（预留和配置内存大小之间的差值）。

需要该交换预留来确保 ESXi 主机在任何情况下均能预留虚拟机内存。实际上，只有一小部分主机级别的交换空间可能会用到。

如果正在通过 ESXi 使内存过载以支持由膨胀导致的客户机内部交换，请确保客户机操作系统还有足够的交换空间。该客户机级别交换空间必须大于或等于虚拟机配置内存大小与其“预留”之间的差值。

小心 如果内存过载且客户机操作系统配置的交换空间不足，则虚拟机中的客户机操作系统可能会出现故障。

为了避免虚拟机出现故障，请增加虚拟机中交换空间的大小。

- **Windows 客户机操作系统** — 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 操作系统会尝试自动增加分页文件的大小。

请查看 [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 客户机操作系统** — 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` — 针对分页和交换启用设备和文件。

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

注 不要将交换文件存储在精简置备的 LUN 上。运行交换文件存储在精简置备的 LUN 上的虚拟机会导致交换文件增长失败，从而可能会导致虚拟机终止。

创建大型交换文件（例如，大于 100 GB 的文件）时，打开虚拟机电源所花的时间会显著增加。为避免出现这种情况，请为大型虚拟机设置较大的预留。

还可使用主机-本地交换文件将交换文件置于开销较小的存储器中。

配置主机的虚拟机交换文件属性

可通过配置主机的交换文件位置来确定虚拟机交换文件在 vSphere Client 中的默认位置。

默认情况下，虚拟机的交换文件位于数据存储上包含其他虚拟机文件的文件夹中。但是，可将主机配置为将虚拟机交换文件置于备用数据存储上。

可以使用该选项将虚拟机交换文件放到成本较低或性能较高的存储上。也可替换单一虚拟机的此主机级设置。

设置备用交换文件位置可能会导致使用 vMotion 迁移速度缓慢。为获得最佳 vMotion 性能，请将虚拟机存储在本地数据存储中，而不是与虚拟机交换文件存储在同一个目录中。如果虚拟机存储在本地数据存储中，存储交换文件和其他虚拟机文件将无法提高 vMotion 的性能。

前提条件

所需特权：**主机.配置.存储分区配置**

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**。
- 3 在**虚拟机**下，单击**交换文件位置**。

此时会显示选定的交换文件位置。如果选定主机不支持交换文件位置的配置，则此选项卡将指示该功能不受支持。

如果主机属于集群的一部分，且集群设置指定交换文件与虚拟机存储在同一个目录中，则无法从**配置**下的主机编辑交换文件位置。要更改此类主机的交换文件位置，请编辑集群设置。

- 4 单击**编辑**。
- 5 选择存储交换文件的位置。

选项	描述
虚拟机目录	将交换文件存储在与虚拟机配置文件相同的目录中。
使用特定数据存储	在您指定的位置存储交换文件。 如果无法将交换文件存储到主机指定的数据存储中，则交换文件必须与虚拟机存储在同一个文件夹中。

- 6 （可选） 如果选择**使用特定数据存储**，请从列表中选择数据存储。
- 7 单击**确定**。

结果

虚拟机交换文件将存储在选定位置。

配置集群的虚拟机交换文件位置

默认情况下，虚拟机的交换文件位于数据存储上包含其他虚拟机文件的文件夹中。但是，您可以配置集群内的主机，将虚拟机交换文件置于自己选择的替代数据存储上。

根据您的需求，可以配置备用交换文件位置，以将虚拟机交换文件置于成本较低或性能较高的存储上。

前提条件

在配置集群的虚拟机交换文件位置之前，必须按照配置**主机的虚拟机交换文件属性**中的说明配置集群内主机的虚拟机交换文件位置。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**。
- 3 选择**配置 > 常规**。
- 4 在“交换文件位置”旁，单击**编辑**。
- 5 选择存储交换文件的位置。

选项	描述
虚拟机目录	将交换文件存储在虚拟机配置文件相同的目录中。
由主机指定的数据存储	将交换文件存储在主机配置中指定的位置。 如果无法将交换文件存储到主机指定的数据存储中，则交换文件必须与虚拟机存储在 同一文件夹中。

- 6 单击**确定**。

删除交换文件

如果主机失败，并且该主机所具有的正在运行的虚拟机使用交换文件，则这些交换文件会继续存在并消耗数 GB 的磁盘空间。您可以删除这些交换文件，从而消除此问题。

步骤

- 1 重新启动故障主机上的虚拟机。
- 2 停止该虚拟机。

结果

该虚拟机的交换文件即会删除。

永久内存 (PMem) 也称为非易失性内存 (NVM)，它即使在断电之后也能够保持数据不丢失。对停机时间敏感并要求高性能的应用程序可以使用 PMem。

可将虚拟机配置为在独立主机或集群中使用 PMem。PMem 会被视为本地数据存储。永久内存可显著降低存储延迟时间。在 ESXi 中，您可以创建配置了 PMem 的虚拟机，由此带来的速度提升也可以让这些虚拟机内的应用程序受益。虚拟机首次打开电源后，无论虚拟机电源打开还是关闭，都将为其预留 PMem。此 PMem 将一直预留，直到虚拟机被迁移或移除为止。

虚拟机可在两种不同模式下消耗永久内存。旧版客户机操作系统仍可以利用虚拟永久内存磁盘功能。

■ 虚拟永久内存 (vPMem)

使用 vPMem 时，内存将作为虚拟 NVDIMM 提供给客户机操作系统使用。这使客户机操作系统能够在字节可寻址随机模式下使用 PMem。

注 您必须使用虚拟机硬件版本 14 和支持 NVM 技术的客户机操作系统。

注 为 PMem 虚拟机配置 vSphere HA 时，必须使用虚拟机硬件版本 19。有关详细信息，请参见[配置 PMem 虚拟机的 vSphere HA](#)。

■ 虚拟永久内存磁盘 (vPMemDisk)

使用 vPMemDisk 时，内存将以虚拟 SCSI 设备的形式供客户机操作系统访问，但虚拟磁盘存储在 PMem 数据存储中。

如创建带有 PMem 的虚拟机，系统会在硬盘创建时为虚拟机预留内存。在硬盘创建时还将执行准入控制。有关详细信息，请参见[vSphere HA 准入控制 PMem 预留](#)。

在集群中，每个虚拟机都具有一部分 PMem 容量。PMem 的总量不能大于集群中的可用总量。已打开电源和已关闭电源的虚拟机均会消耗 PMem。如果虚拟机配置为使用 PMem，但您未使用 DRS，则必须手动选择具有足够 PMem 可放置该虚拟机的主机。

NVDIMM 和传统存储

NVDIMM 可作为内存访问。使用传统存储时，应用程序和存储设备之间存在软件，这可能会导致处理时间出现延迟。使用 PMem 时，应用程序将直接使用存储。这意味着 PMem 性能优于传统存储。存储位于主机本地。但是，由于系统软件无法跟踪所做的更改，因此备份等解决方案目前无法与 PMem 结合使用。

如果在不完全写入非 PMem 数据存储的模式下使用 vPMem，则 vSphere HA 等解决方案的范围有限。为启用了故障切换的 vPMem 虚拟机激活 vSphere HA 后，虚拟机可以故障切换到其他主机。发生这种情况时，虚拟机将使用新主机上的 PMem 资源。为了释放旧主机上的资源，垃圾数据收集器会定期识别和释放这些资源，以供其他虚拟机使用。

命名空间

PMem 的命名空间在 ESXi 启动之前配置。命名空间与系统中的磁盘相似。ESXi 读取命名空间，并通过写入 GPT 头将多个命名空间合并成一个逻辑卷。如果以前未配置命名空间，则默认情况下命名空间会自动格式化。如果已经格式化，ESXi 会尝试挂载 PMem。

GPT 表

PMem 存储中的数据损坏可能会导致 ESXi 发生故障。为避免这种情况，ESXi 会在 PMem 挂载期间检查元数据是否有误。

PMem 区域

PMem 区域是表示单个 vNVDIMM 或 vPMemDisk 的连续字节流。每个 PMem 卷仅属于一台主机。如果管理员必须管理具有大量主机的集群中的每台主机，则可能会面临管理难题。但是，您无需管理每个单独的数据存储。相反，您可以将集群中的整个 PMem 容量视为一个数据存储。

VC 和 DRS 将自动执行 PMem 数据存储的初始放置操作。请在创建虚拟机时或向虚拟机添加设备时，选择本地的 PMem 存储配置文件。系统将自动运行其余的配置操作。这么做有一个限制，那就是 ESXi 不允许将虚拟机主目录放在 PMem 数据存储上。因为这会占用宝贵的空间来存储虚拟机日志文件和统计信息文件。这些区域用于表示虚拟机数据，并且可以作为字节可寻址 NVDIMM 或 vPMem 磁盘被访问。

迁移

因为 PMem 是本地数据存储，因此想要移动虚拟机就必须使用 Storage vMotion。具有 vPMem 的虚拟机只能迁移到具有 PMem 资源的 ESX 主机，而具有 vPMemDisk 的虚拟机可迁移到没有 PMem 资源的 ESX 主机。

错误处理和 NVDIMM 管理

主机故障可能会导致未处于完全写入模式的 vPMem 虚拟机丧失可用性。如果出现灾难性错误，您可能会丢失所有数据且必须采取手动步骤重新格式化 PMem。

vSphere Client 的 vSphere 永久内存

有关永久内存的概念简介，请参阅：



(vSphere Client 的 vSphere 永久内存)

在 vSphere Client 中使用 PMEM 增强功能

有关使用 PMem 时基于 HTML5 的 vSphere Client 中的增强功能的简要概述，请参见：



(在 vSphere Client 中使用 PMEM 增强功能)

在 vSphere Client 中迁移和克隆使用 PMEM 的虚拟机

有关迁移和克隆使用 PMem 的虚拟机的简要概述，请参见：



(在 vSphere Client 中迁移和克隆使用 PMEM 的虚拟机)

本章讨论了以下主题：

- 配置 PMem 虚拟机的 vSphere HA
- vSphere HA 准入控制 PMem 预留
- vSphere 内存监控和修复

配置 PMem 虚拟机的 vSphere HA

可以在完全写入模式下配置 PMem 虚拟机的 vSphere HA，以便在主机出现故障时，可以在另一台正常运行的主机上还原虚拟机。

前提条件

- 必须选择硬件版本 19。
- 不支持具有 vPMemDisk 的 PMem 虚拟机。

步骤

1 在**新建虚拟机**向导中创建新的虚拟机时，选择**自定义硬件**。

- 单击**添加新设备**，然后从下拉菜单中选择**添加 NVDIMM**。
- 单击**允许在另一个主机上对所有 NVDIMM 设备进行故障切换**复选框。
- 单击**下一步**并完成**新建虚拟机**向导。

在主机出现故障时，NVDIMM PMem 数据无法恢复。默认情况下，HA 不会尝试在另一台主机上重新启动此虚拟机。如果允许 HA 在主机发生故障时对虚拟机进行故障切换，将在具有新的空 NVDIMM 的另一台主机上重新启动虚拟机。

2 要在现有虚拟机上启用 HA，请浏览到该虚拟机。

- 在**虚拟机硬件**下，单击**编辑**。
- 选择 NVDIMM。
- 单击**允许在另一个主机上对所有 NVDIMM 设备进行故障切换**复选框。
- 单击**确定**。

主机出现故障时，HA 将在另一个主机上使用新的空 NVDIMM 重新启动此虚拟机。

vSphere HA 准入控制 PMem 预留

准入控制是 vSphere HA 用于确保集群内的故障切换容量的一种策略。

增加允许的潜在主机故障数将增加可用性限制和预留容量。可以预留一定比例的永久内存作为主机故障切换容量。这是被阻止且在主机关闭电源时必须考虑在内的实际存储容量。

在**编辑集群设置**下，可以选择**准入控制**以指定主机允许的故障数。

如果选择通过以下项定义的 CPU/内存预留：

- **集群资源百分比**，即使集群中的虚拟机当前未使用永久内存，集群中的部分永久内存容量也会专用于故障切换用途。此百分比可以通过替代指定，也可以根据**允许的主机故障数**设置自动计算得出。启用 PMem 准入控制后，即使有虚拟机使用 PMem 作为磁盘，也会在集群中预留 PMem 容量。
- **插槽策略 (已打开电源的虚拟机)**，永久内存准入控制将“插槽策略”替代为“集群资源百分比”策略，仅适用于永久内存资源。百分比值根据**集群允许的主机故障数目**设置自动计算得出，无法替代。
- **专用故障切换主机**，专用故障切换主机的永久内存专用于故障切换用途，且无法在这些主机上置备具有永久内存的虚拟机。

注 选择准入控制策略后，还必须单击**预留永久内存故障切换容量**复选框才能启用 PMem 准入控制。

vSphere 内存监控和修复

vMMR 收集数据并显示性能统计信息，以便您可以确定应用程序工作负载是否因内存模式而出现性能下降问题。

可以在应用直接访问模式或内存模式下，在 BIOS 设置中配置 Intel Optane 永久内存。在应用直接访问模式下，永久内存可以作为字节可寻址永久内存和 DRAM 一起访问。在内存模式下，DRAM 将成为硬件缓存，较大的 PMem 将变为易失性内存并显示为系统内存。

内存模式对虚拟机不可见且透明。在内存模式下配置系统后，系统将显示为具有 DRAM 的传统系统。一个集群可以包含一组具有不同配置的主机。vSphere 显示有关处于内存模式的系统的其他信息。ESXi 可对用于收集主机级别和虚拟机级别统计信息的相关信息的性能计数器进行编程。这些性能统计信息用于创建警报。还可以在性能图表中跟踪统计信息。

可以通过主机**摘要**选项卡下的**内存分层: 硬件**和一些其他详细信息了解系统是否处于内存模式。

Summary	Monitor	Configure	Permissions	VMs	Datastores	Networks	Updates
	Logical Processors:	96					
	NICs:	3					
	Virtual Machines:	1					
	Memory Tiering:	Hardware					
	State:	Connected					
	Uptime:	6 hours					

Intel Optane™ Persistent Memory configured in Memory Mode.

还可以在**配置 > 硬件 > 概览 > 内存**下查看 DRAM 和 PMEM 的大小。

Summary

Monitor

Configure

Permissions

VMs

Datastores

Networks

Updates

System Resource Reservation

Firewall

Services

Security Profile

System Swap

Packages

Hardware

Overview

Graphics


Memory

Total	503.68 GB
System	385.17 MB
Virtual machines	503.3 GB
Memory Tiering	Hardware ⓘ
Tier 0	256 GB DRAM (Cache)
Tier 1	503.67 GB PMem (Memory)

ESXi 收集并公开两种内存统计信息：

- **主机级别统计信息：**内存子组件通过对性能计数器进行编程衡量 DRAM 和 PMem 性能。主机级别统计信息包括不同内存类型（DRAM、PMem）的总计、读取/写入带宽、读取/写入延迟和丢失率。
- **虚拟机级别统计信息：**vSphere 监控性能计数器，获取有关虚拟机的 DRAM 和 PMEM 读取带宽数据。

主机和虚拟机的性能图表下都有新的“内存”窗格。该窗格将显示内存详细信息，如“内存利用率”、“内存回收”以及新的统计信息。在 ESXi 主机级别上，可以监控内存带宽和内存丢失率图表。在虚拟机级别，可以查看 PMem 读取带宽和 DRAM 读取带宽。高级性能图表可用于有选择地绘制任何新的统计信息。例如，您可以监控读取/写入延迟和丢失率。

从 ESXi 主机的**虚拟机**选项卡中，可以查看包含驻留在该主机上的所有虚拟机的性能信息的列表。要显示有关内存模式对虚拟机的影响的信息，请单击视图列  图标，然后选择“活动内存”、“DRAM 读取带宽”和“PMem 读取带宽”衡量指标。

有两个预配置的默认警报，一个在主机级别（主机内存模式活动 DRAM 使用情况较高），另一个在虚拟机级别（虚拟机 PMem 带宽使用情况较高）。如果满足警报条件，将发布事件以触发相应的警报。您还可以根据性能衡量指标创建自定义警报。vMMR 警报仅适用于配置了内存模式的主机。

在集群中启用并完全自动化 DRS 时，如果主机的活动内存利用率高于 DRAM 缓存大小的一定百分比，则 DRS 可能会将某些虚拟机移出主机以均衡负载。

有关详细信息，请参见《vSphere 监控和性能》。

注 Intel Broadwell、Skylake、Cascade Lake 和 Ice Lake 平台支持 vMMR。主机级别的 DRAM 统计信息在这些平台上可用。主机和虚拟机级别的 PMem 统计信息仅在内存模式下配置的 Cascade Lake 和 Ice Lake 主机中可用。

配置虚拟图形

12

您可以为受支持的图形实现编辑图形设置。

vSphere 支持多种图形实现。

- VMware 支持 AMD、Intel 和 NVIDIA 提供的三维图形解决方案。
- 支持 NVIDIA GRID。
- 允许单个 NVIDIA VIB 同时支持 vSGA 和 vGPU 实现。
- 为 Intel 和 NVIDIA 提供 vCenter GPU 性能图表。
- 为 Horizon View VDI 桌面启用图形。

您可以针对每个虚拟机配置主机图形设置，并自定义 vGPU 图形设置。

注 本章中“内存”是指物理内存。

本章讨论了以下主题：

- [查看 GPU 统计信息](#)
- [将 NVIDIA GRID vGPU 添加到虚拟机](#)
- [配置主机图形](#)
- [配置图形设备](#)

查看 GPU 统计信息

您可以查看主机显卡的详细信息。

您可以查看 GPU 温度、利用率和内存使用情况。

注 这些统计信息只有在主机上安装 GPU 驱动程序时才会显示。

步骤

- 1 在 vSphere Client 中，导航到主机。
- 2 单击[监控](#)选项卡，然后单击[性能](#)。
- 3 单击[高级](#)，然后从下拉菜单中选择 **GPU**。

将 NVIDIA GRID vGPU 添加到虚拟机

如果 ESXi 主机具有 NVIDIA GRID GPU 图形设备，则可以将虚拟机配置为使用 NVIDIA GRID 虚拟 GPU (vGPU) 技术。

NVIDIA GRID GPU 图形设备旨在优化复杂的图形操作，使这些操作能够以高性能运行且不会出现 CPU 过载。

前提条件

- 验证主机上是否安装了具有相应驱动程序的 NVIDIA GRID GPU 图形设备。请参见《vSphere 升级》文档。
- 验证虚拟机是否与 ESXi6.0 及更高版本兼容。

步骤

- 1 右键单击虚拟机，然后选择**编辑设置**。
- 2 在**虚拟硬件**选项卡上，选择**添加新设备**，然后从下拉菜单中选择**新 PCI 设备**。
- 3 展开**新 PCI 设备**，然后选择要连接虚拟机的 NVIDIA GRID vGPU 直通设备。

注 将自动应用全部内存预留，这是 PCI 设备所必需的。

- 4 选择 GPU 配置文件。
GPU 配置文件表示 vGPU 类型。
- 5 单击**确定**。

结果

虚拟机即可访问该设备。

配置主机图形

您可以针对每个主机自定义图形选项。

前提条件

应关闭虚拟机电源。

步骤

- 1 选择一个主机，然后选择**配置 > 图形**。
- 2 在**主机图形**下，选择**编辑**。

- 3 在**编辑主机图形设置**窗口中，选择：

选项	描述
共享	VMware 共享虚拟图形
直接共享	供应商共享直通图形

- 4 选择一个共享直通 GPU 分配策略。
- a 将虚拟机分散在多个 GPU 中 (最佳性能)
 - b 将虚拟机组合到 GPU 中直到已满为止 (GPU 整合)

- 5 单击**确定**。

后续步骤

单击**确定**后，您必须重新启动主机上的 Xorg。

配置图形设备

您可以编辑显卡的图形类型。

前提条件

必须关闭虚拟机电源。

步骤

- 1 在**图形设备**下，选择一个显卡并单击**编辑**。
- a 为 VMware 共享虚拟图形选择**共享**。
 - b 为供应商共享直通图形选择**直接共享**。
- 2 单击**确定**。

结果

如果选择某个设备，则会显示哪些虚拟机（如果处于活动状态）正在使用该设备。

后续步骤

单击**确定**后，您必须重新启动主机上的 Xorg。

vSphere Storage I/O Control 允许在整个集群范围内划分存储 I/O 的优先级，从而更好地整合工作负载，并有助于减少与过度置备相关的额外成本。

Storage I/O Control 通过扩展份额和限制的构成来处理存储 I/O 资源。您可以控制在 I/O 拥堵期间分配给虚拟机的存储 I/O 量，从而确保在进行 I/O 资源分配时重要性较高的虚拟机与重要性较低的虚拟机相比具有更高的优先级。

当对数据存储启用 Storage I/O Control 时，ESXi 会开始监控主机与该数据存储通信时出现的设备延迟时间。当设备延迟时间超出阈值时，该数据存储会被视为出现拥堵，此时将按访问该数据存储的每个虚拟机的份额比例向其分配 I/O 资源。您可以设置每个虚拟机的份额，并根据需要调整每个虚拟机的份额数量。

I/O 筛选器框架 (VAIO) 允许 VMware 及其合作伙伴开发用于拦截每个 VMDK 的 I/O 的筛选器，并在 VMDK 粒度级别提供所需的功能。VAIO 可与基于存储策略的管理 (SPBM) 配合使用，从而允许您通过附加到 VMDK 的存储策略设置筛选器首选项。

配置 Storage I/O Control 分为两个步骤：

- 1 为数据存储启用 Storage I/O Control。
- 2 设置每个虚拟机所允许的存储 I/O 份额数量以及每秒 I/O 操作数 (IOPS) 的上限。

默认情况下，所有虚拟机的份额均设置为“正常 (1000)”且 IOPS 无限制。

注 在已启用 Storage DRS 的数据存储集群上，Storage I/O Control 默认情况下处于启用状态。

注 本章中“内存”是指物理内存。

本章讨论了以下主题：

- [关于虚拟机存储策略](#)
- [关于 I/O 筛选器](#)
- [Storage I/O Control 要求](#)
- [Storage I/O Control 资源份额和限制](#)
- [查看 Storage I/O Control 份额和限制](#)
- [监控 Storage I/O Control 份额](#)
- [设置 Storage I/O Control 资源份额和限制](#)

- 启用 Storage I/O Control
- 设置 Storage I/O Control 阈值
- Storage DRS 与存储配置文件集成

关于虚拟机存储策略

虚拟机存储策略对虚拟机置备至关重要。这些策略将控制为虚拟机提供的存储类型、虚拟机在存储中的放置方式，以及为虚拟机提供的数据服务。

vSphere 包括默认存储策略。但是，您可以定义和分配新策略。

使用“虚拟机存储策略”界面创建存储策略。定义该策略时，可为虚拟机上运行的应用程序指定各种存储要求。您也可以使用存储策略来为虚拟磁盘请求缓存或复制等特定的数据服务。

您可以在创建、克隆或迁移虚拟机时应用该存储策略。应用存储策略之后，基于存储策略的管理 (SPBM) 机制会将虚拟机放置到匹配的数据存储中，在某些存储环境中还会确定如何在存储资源内置备和分配虚拟机存储对象，以确保所需的服务级别。SPBM 还将对虚拟机启用请求的数据服务。vCenter Server 会监控策略合规性，并在虚拟机违反所分配的存储策略时发送警示。

请参见《vSphere 存储》了解详细信息。

关于 I/O 筛选器

无论基础存储拓扑如何，与虚拟磁盘关联的 I/O 筛选器均可直接访问虚拟机 I/O 路径。

VMware 提供某些类别的 I/O 筛选器。此外，第三方供应商也可创建 I/O 筛选器。通常，它们以软件包形式分发，可提供安装程序以在 vCenter Server 和 ESXi 主机集群上部署筛选器组件。

在 ESXi 集群上部署 I/O 筛选器时，vCenter Server 会自动为集群中的每个主机配置并注册 I/O 筛选器存储提供程序（也称为“VASA 提供程序”）。这些存储提供程序与 vCenter Server 进行通信，并使 I/O 筛选器提供的数据服务在“虚拟机存储策略”界面中可见。在定义虚拟机策略的常用规则时，可以引用这些数据服务。将虚拟磁盘与此策略关联后，会在虚拟磁盘上启用 I/O 筛选器。

请参见《vSphere 存储》了解详细信息。

Storage I/O Control 要求

Storage I/O Control 有一些要求和限制。

- 启用了 Storage I/O Control 的数据存储必须由单个 vCenter Server 系统管理。
- 光纤通道连接、iSCSI 连接和 NFS 连接的存储上都可支持 Storage I/O Control。裸设备映射 (RDM) 不受支持。
- Storage I/O Control 不支持具有多个数据区的数据存储。
- 在具有自动化存储分层功能的阵列所支持的数据存储上使用 Storage I/O Control 之前，请查看《VMware 存储/SAN 兼容性指南》，以确认自动化分层存储阵列已通过认证，与 Storage I/O Control 兼容。

自动化存储分层是阵列（或阵列组）的功能，可根据用户设置的策略和当前 I/O 模式，将 LUN/卷或 LUN/卷的某些部分迁移到其他类型的存储介质（SSD、FC、SAS 和 SATA）。对于不具有这些自动迁移/分层功能的阵列（其中包括提供不同类型存储介质之间手动迁移数据功能的阵列）无需特殊认证。

Storage I/O Control 资源份额和限制

您可以分配每个虚拟机所允许的存储 I/O 份额数量以及每秒 I/O 操作数 (IOPS) 的上限。当检测到数据存储出现存储 I/O 拥堵时，会根据每个虚拟机具有的虚拟机份额比例调整访问该数据存储的虚拟机的 I/O 工作负载。

如[资源分配份额](#)中所介绍，存储 I/O 份额与用于内存和 CPU 资源分配的份额相似。这些份额表示虚拟机在存储 I/O 资源分配方面的相对重要性。在发生资源争用时，份额值较高的虚拟机访问存储阵列的机会更大。当分配存储 I/O 资源时，您可以限制虚拟机所允许的 IOPS。在默认情况下，IOPS 无限制。

[资源分配限制](#)中介绍了设置资源限制的优缺点。如果要为虚拟机设置的限制单位为 MB/秒而非 IOPS，则可根据虚拟机的典型 I/O 大小将 MB/秒转换为 IOPS。例如，要将具有 64 KB IO 的备份应用程序限定为 10 MB/秒，请将限制设置为 160 IOPS。

查看 Storage I/O Control 份额和限制

您可以查看数据存储上运行的所有虚拟机的份额和限制。通过查看此信息，可以比较访问该数据存储的所有虚拟机的设置，而不管这些虚拟机在哪个集群中运行。

步骤

- 1 在 vSphere Client 中，浏览到数据存储。
- 2 单击**虚拟机**选项卡。

该选项卡显示了数据存储上运行的每个虚拟机以及关联的份额值和数据存储份额百分比。

监控 Storage I/O Control 份额

使用数据存储**性能**选项卡，可以监控 Storage I/O Control 如何基于其份额处理访问数据存储的虚拟机的 I/O 工作负载。

使用数据存储性能图表可以监控以下信息：

- 数据存储的平均延迟时间和汇总 IOPS
- 主机之间的延迟时间
- 主机之间的队列深度
- 主机之间的读/写 IOPS
- 虚拟机磁盘之间的读/写延迟时间
- 虚拟机磁盘之间的读/写 IOPS

步骤

- 1 在 vSphere Client 中，浏览到数据存储。
- 2 在**监控**选项卡下，单击**性能**。
- 3 选择**高级**。

设置 Storage I/O Control 资源份额和限制

通过向虚拟机分配相对数量的份额，可根据重要性将存储 I/O 资源分配给虚拟机。

除非虚拟机工作负载非常相似，否则份额不必以 I/O 操作数或 MB/秒来规定分配。较高的份额可以使虚拟机在存储设备或数据存储中保持更多挂起的并行 I/O 操作（与份额较低的虚拟机相比）。根据其工作负载，两个虚拟机可能有不同的吞吐量。

前提条件

请参见《vSphere 存储》了解有关创建虚拟机存储策略和定义虚拟机存储策略常用规则的信息。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后单击**编辑设置**。
- 3 单击**虚拟硬件**选项卡，然后从列表中选择虚拟硬盘。展开**硬盘**。
- 4 从下拉菜单中选择**虚拟机存储策略**。

如果选择了某个存储策略，请勿手动配置**份额**和**限制 - IOPS**。

- 5 在**份额**下，单击下拉菜单，然后选择要分配给虚拟机的相对数量的份额（低、正常或高）。

您可以选择**自定义**以输入用户定义的份额值。

- 6 在**限制 - IOPS**下，单击下拉菜单，然后输入要分配给虚拟机的存储资源的上限。

IOPS 是每秒 I/O 操作数。在默认情况下，IOPS 无限制。选择低 (500)、正常 (1000) 或高 (2000)，或者可以选择“自定义”输入用户定义的份额量。

- 7 单击**确定**。

启用 Storage I/O Control

启用 Storage I/O Control 后，如果数据存储平均延迟时间超过阈值，则 ESXi 会监控数据存储延迟时间并限制 I/O 负载。

步骤

- 1 在 vSphere Client 中，浏览到数据存储。

- 2 单击**配置**选项卡。
- 3 依次单击**设置**和**常规**。
- 4 针对**数据存储功能**单击**编辑**。
- 5 选中**启用 Storage I/O Control** 对话框。
- 6 单击**确定**。

结果

在**数据存储功能**下，即会针对数据存储启用 Storage I/O Control。

设置 Storage I/O Control 阈值

数据存储的拥堵阈值是数据存储所允许的延迟时间上限，超过该值后，Storage I/O Control 将开始根据份额将重要性分配给虚拟机工作负载。

在大部分环境中，都不需要调整阈值设置。

小心 如果在两个不同的数据存储上共享相同的心轴，Storage I/O Control 可能无法正常运行。

如果要更改拥堵阈值设置，请根据以下事项来设置该值。

- 值越大，通常会导致总吞吐量越大，隔离越弱。除非整体平均延迟时间高于阈值，否则不会出现限制。
- 如果吞吐量比延迟时间更重要，请不要将该值设置得过低。例如，对于光纤通道磁盘，低于 20 ms 的值可降低磁盘吞吐量峰值。当该值非常大（超过 50 毫秒）时，可能会出现延迟时间长而总吞吐量却并未显著增加的情况。
- 值越小，则设备的延迟时间就越短，且虚拟机 I/O 性能隔离将越强。隔离增强意味着份额控制的实施更加频繁。设备延迟时间越短，则拥有最高份额的虚拟机的 I/O 延迟时间越短，但同时会导致份额较低的虚拟机的 I/O 延迟时间更长。
- 非常低的值（小于 20 毫秒）会使设备的延迟时间更短，I/O 之间的隔离更强，但有可能会降低数据存储总吞吐量。
- 将值设置极高或极低会导致隔离性变差。

前提条件

验证是否启用了 Storage I/O Control。

步骤

- 1 在 vSphere Client 中，浏览到数据存储。
- 2 依次单击**配置**选项卡和**设置**。
- 3 单击**常规**。
- 4 针对**数据存储功能**单击**编辑**。

5 选中启用 **Storage I/O Control** 对话框。

当数据存储以其吞吐量峰值的 90% 运行时，Storage I/O Control 会自动将延迟时间阈值设置为与估计延迟时间相对应。

6 （可选）调整**拥堵阈值**。

- ◆ 从**吞吐量峰值百分比**下拉菜单中选择一个值。

吞吐量峰值百分比指示当数据存储使用其估计吞吐量峰值百分比时的估计延迟时间阈值。

- ◆ 从**手动**下拉菜单中选择一个值。

该值必须介于 5 毫秒到 100 毫秒之间。如果设置了错误的拥堵阈值，则可能会对数据存储上虚拟机的性能不利。

7 （可选）单击**重置为默认值**，将拥堵阈值设置还原为默认值（30 毫秒）。

8 单击**确定**。

Storage DRS 与存储配置文件集成

基于存储策略的管理 (SPBM) 允许您为虚拟机指定由 Storage DRS 执行的策略。一个数据存储集群可以包含一组具备不同功能配置文件的数据存储。如果虚拟机具有关联的存储配置文件，则 Storage DRS 可以根据基础数据存储功能执行放置。

在 Storage DRS 与存储配置文件集成过程中，引入了 Storage DRS 集群级别高级选项 `EnforceStorageProfiles`。高级选项 `EnforceStorageProfiles` 使用以下整数值之一：0、1 或 2。默认值为 0。当选项设置为 0 时，表示 Storage DRS 集群中不会执行任何存储配置文件或策略。当选项设置为 1 时，表示 Storage DRS 集群中将软性执行存储配置文件或策略。这类似于 DRS 软规则。Storage DRS 将在最佳级别遵守存储配置文件或策略。但如果要求违反存储配置文件合规性，Storage DRS 也会按要求执行操作。只有在存储配置文件执行被设置为 1 时，Storage DRS 关联性规则的优先级才会高于存储配置文件。当选项设置为 2 时，表示 Storage DRS 集群中将硬性执行存储配置文件或策略。这类似于 DRS 硬规则。Storage DRS 不会违反存储配置文件或策略合规性。存储配置文件的优先级高于关联性规则。Storage DRS 将生成错误：无法解决反关联性规则的违反问题 (could not fix anti-affinity rule violation)

前提条件

默认情况下，Storage DRS 不会执行与虚拟机关联的存储策略。请根据您的需求，配置 `EnforceStorageProfiles` 选项。选项包括默认 (0)、软性要求 (1) 或硬性要求 (2)。

步骤

- 1 以管理员身份登录到 vSphere Client。
- 2 在 vSphere Client 中，单击 Storage DRS 集群，然后选择**管理 > 设置 > Storage DRS**。
- 3 单击**编辑 > 高级选项 > 配置参数**，然后选择**添加**。
- 4 单击“选项”标题下方的区域，并键入 `EnforceStorageProfiles`。
- 5 单击之前输入的高级选项名称右侧的“值”标题下方的区域，并键入值 0、1 或 2。

6 单击**确定**。

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，用于对可用的 CPU 和内存资源按层次结构进行分区。

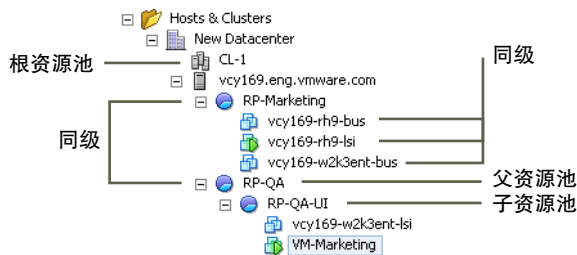
每个独立主机和每个 DRS 集群都具有一个（不可见的）根资源池，此资源池对该主机或集群的资源进行分组。根资源池之所以不显示，是因为主机（或集群）与根资源池的资源总是相同的。

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部分父级资源，然而子资源池也可以具有各自的子资源池层次结构，每个层次结构代表更小部分的计算容量。

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。集群本身表示 root 资源池。如果不创建子资源池，则只存在根资源池。

在以下示例中，RP-QA 是 RP-QA-UI 的父资源池。RP-Marketing 与 RP-QA 是同级。紧靠 RP-Marketing 下面的三个虚拟机也是同级。

图 14-1. 资源池层次结构中的父级、子级和同级



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。

注 本章中“内存”是指物理内存。

本章讨论了以下主题：

- 为什么使用资源池？
- 创建资源池
- 编辑资源池
- 将虚拟机添加到资源池

- 从资源池移除虚拟机
- 移除资源池
- 资源池接入控制
- 可扩展预留示例 1
- 可扩展预留示例 2

为什么使用资源池？

通过资源池可以委派对主机（或集群）资源的控制权，在使用资源池划分集群内的所有资源时，其优势非常明显。可以创建多个资源池作为主机或集群的直接子级，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。某部门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 — 顶级管理员使资源池可供部门级管理员使用后，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的虚拟机创建和管理操作。委派通常结合权限设置一起执行。
- 资源与硬件的分离 — 如果使用的是已启用 DRS 的集群，则所有主机的资源始终会分配给集群。这意味着管理员可以独立于提供资源的实际主机来进行资源管理。如果将三台 2GB 主机替换为两台 3GB 主机，您无需对资源分配进行更改。

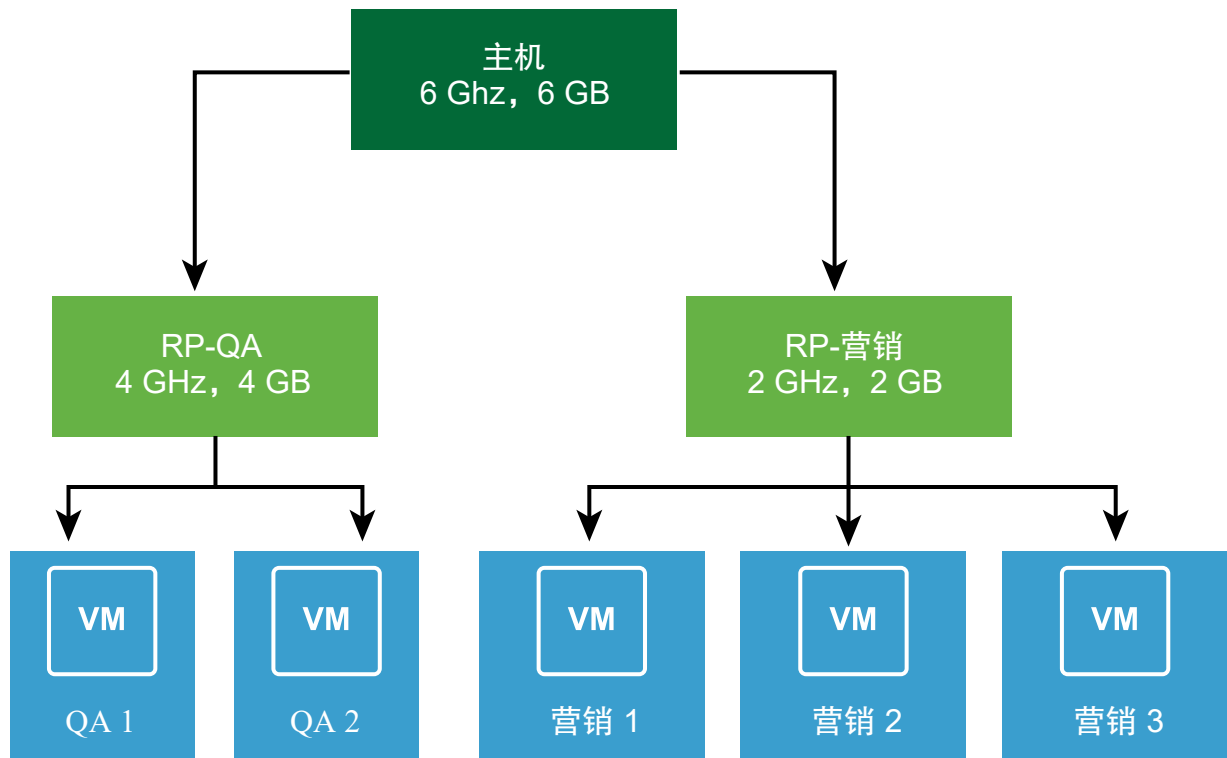
这一分离可使管理员更多地考虑聚合计算能力而非各个主机。

- 管理运行多层服务的各组虚拟机 — 为资源池中的多层服务进行虚拟机分组。您无需对每个虚拟机进行资源设置，相反，通过更改所属资源池上的设置，您可以控制对虚拟机集合的聚合资源分配。

例如，假定一台主机拥有多个虚拟机。营销部门使用其中的三个虚拟机，QA 部门使用两个虚拟机。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** 分别设置为**高**和**正常**，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到的资源，营销部门就可以使用这些可用资源。

下图中的数字显示了向资源池的有效分配。

图 14-2. 向资源池分配资源



创建资源池

可以创建任何 ESXi 主机、资源池或 DRS 集群的子资源池。

注 如果已将某台主机添加到集群，将无法创建该主机的子资源池。如果已为 DRS 启用集群，则可以创建集群的子资源池。

创建子资源池时，系统将提示您输入资源池属性信息。系统使用准入控制确保您不能分配不可用的资源。如果您希望份额在添加或移除虚拟机时动态伸缩，则可以选择可扩展份额。

注 份额在父级别进行扩展。默认情况下，从具有可扩展份额的父资源池创建的所有后代资源池都具有可扩展份额。

前提条件

将 vSphere Client 连接到 vCenter Server 系统。

步骤

- 1 在 vSphere Client 中，选择资源池的父对象（主机、其他资源池或 DRS 集群）。
- 2 右键单击对象，然后选择**新建资源池**。
- 3 键入用来标识资源池的名称。
- 4 如果要启用可扩展份额，请选中该复选框。
- 5 指定 CPU 和内存资源的分配方式。

资源池的 CPU 资源是主机为资源池预留的保证物理资源。通常，您接受默认值，并让主机处理资源分配。

选项	描述
份额	指定此资源池相对于父级的总资源的份额值。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。 <ul style="list-style-type: none">■ 选择低、正常或高，这三个级别分别按 1:2:4 这个比率指定份额值。■ 选择自定义可为每个虚拟机提供表示比例权重的特定份额数。
预留	为此资源池指定保证的 CPU 或内存分配量。默认值为“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也是如此。
可扩展预留	选中此复选框（默认设置）后，会在接入控制过程中考虑可扩展预留。如果在该资源池中打开一台虚拟机的电源，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父项的资源。
限制	指定此资源池的 CPU 或内存分配量的上限。您通常可以接受默认值（ 无限 ）。要指定限制，请取消选中 无限 复选框。

- 6 单击**确定**。

结果

创建资源池后，即可向其添加虚拟机。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。

示例：创建资源池

假定有一个主机，提供 6 GHz 的 CPU 和 3 GB 的内存，这些 CPU 和内存必须在营销部门和 QA 部门间进行共享。还需要不均等地共享资源，并授予一个部门 (QA) 更高的优先级。通过为每个部门创建一个资源池并使用**份额**属性区分资源分配优先级，可完成此任务。

本示例展示了如何使用 ESXi 主机作为父资源来创建资源池。

- 1 在**新建资源池**对话框中，键入 QA 部门的资源池的名称（例如，RP-QA）。
- 2 将 RP-QA 的 CPU 和内存资源**份额**指定为**高**。
- 3 创建第二个资源池 RP-Marketing。
将 CPU 和内存的“份额”保留为**正常**。
- 4 单击**确定**。

如果存在资源冲突，则 RP-QA 接收 4GHz 和 2GB 的内存，RP-Marketing 接收 2GHz 和 1GB 的内存。否则，它们可以接收超过此分配的量。这些资源随后即可供各自资源池内的虚拟机使用。

编辑资源池

创建资源池后，可以编辑其 CPU 和内存资源设置。

步骤

- 1 在 vSphere Client 中，浏览到资源池。
- 2 从**操作**下拉菜单中选择**编辑资源设置**。
- 3 （可选）您可以更改选定资源池的所有属性，如**创建资源池**中所述。
 - 如果要启用可扩展份额，请选中该复选框。

注 份额在父级别进行扩展。默认情况下，从具有可扩展份额的父资源池创建的所有后代资源池都具有可扩展份额。

- 在 **CPU** 下，选择 CPU 资源设置。
 - ◆ 在 **内存** 下，选择内存资源设置。
- 4 单击**确定**保存更改。

将虚拟机添加到资源池

创建虚拟机时，可以在创建过程中指定资源池位置。也可以将现有的虚拟机添加到资源池。

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、中或低，份额百分比会有所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 由于份额分配是相对于资源池的，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，以便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），将显示警告。

- 在**监控**下，**资源预留**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 如果虚拟机已关闭电源或挂起，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。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后单击**迁移**。
 - 可以将虚拟机移到另一主机。
 - 可以将虚拟机的存储移到另一数据存储。
 - 可以将虚拟机移到另一主机，并将其存储移到另一数据存储。
- 3 选择要在其中运行该虚拟机的资源池。
- 4 检查选择内容，然后单击**完成**。

结果

如果某个虚拟机已打开电源，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，因为准入控制不允许该操作。一个错误对话框将显示可用资源与请求的资源，以便您可以考虑是否能够通过调整来解决此问题。

从资源池移除虚拟机

通过将虚拟机移动到另一个资源池或将其删除，可以从资源池中移除虚拟机。

从某个资源池中移除虚拟机时，与该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**的虚拟机。假定虚拟机受 CPU 限制，每个虚拟机获得 2 GHz 的相等分配额。如果将其中一个虚拟机移至其他资源池，剩余的两个虚拟机将各获得 3GHz 的相等分配额。

步骤

- 1 在 vSphere Client 中，浏览到资源池。

- 2 选择下列方法之一将虚拟机从资源池移除。
- 右键单击虚拟机，然后选择**移至...**，将虚拟机移到其他资源池。
在移动虚拟机之前，无需关闭其电源。
 - 右键单击虚拟机，然后选择**从磁盘删除**。
必须关闭虚拟机电源才能将其完全移除。

移除资源池

您可以从清单中移除资源池。

步骤

- 1 在 vSphere Client 中，右键单击资源池，然后选择**删除**。
此时将显示确认对话框。
- 2 单击**是**以移除资源池。

资源池接入控制

在资源池内打开虚拟机电源时，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。

在打开虚拟机的电源或创建资源池之前，请使用 vSphere Client 中的**资源预留**选项卡来确保有足够的资源可用。CPU 和内存的**可用预留**值显示了未预留的资源。

如何计算可用的 CPU 和内存资源以及是否执行操作取决于**预留类型**。

表 14-1. 预留类型

预留类型	描述
固定	系统检查所选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 (默认)	系统考虑所选资源池及其直接父资源池中的可用资源。如果对于父资源池也选中了 可扩展预留 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 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，但提供的保护将会同时减少。子资源池所有者预留的资源可能大于您的预期值。

系统不允许违反预先配置的**预留**或**限制**设置。每次重新配置资源池或打开虚拟机电源时，系统都会验证所有参数以确保仍能实现各服务级别保证。

可扩展预留示例 1

此示例显示了具有可扩展预留的资源池的工作方式。

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打开具有预留的虚拟机的电源，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。

可扩展预留会造成缺少严格的隔离。S1 可使用 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。

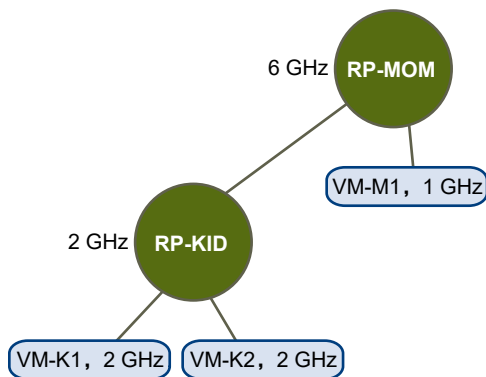
可扩展预留示例 2

此示例显示了具有可扩展预留的资源池的工作方式。

假定以下情形，如图所示。

- 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。
- 您创建了一个具有 2 GHz 预留的子资源池 RP-KID，并选中**可扩展预留**。
- 您向子资源池添加两个各具有 2 GHz 预留的虚拟机（即 VM-K1 和 VM-K2），并尝试打开其电源。
- VM-K1 可直接从 RP-KID（具有 2 GHz）预留资源。
- 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，您可以打开这个 2 GHz 虚拟机的电源。

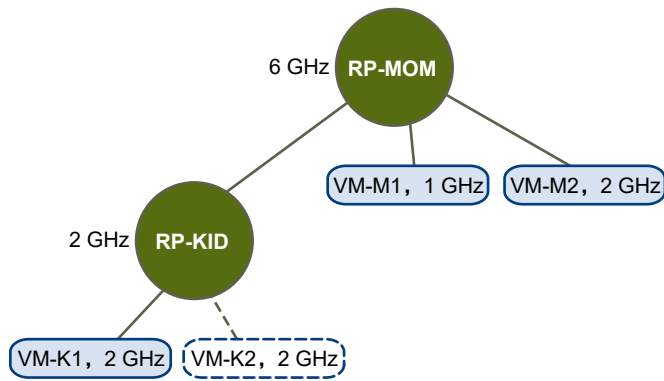
图 14-3. 可扩展资源池的接入控制：成功打开电源



现在假设另一个包含 VM-M1 和 VM-M2 的应用场景。

- 打开 RP-MOM 中总预留为 3 GHz 的两个虚拟机的电源。
- 您依然可打开 RP-KID 中的 VM-K1 的电源，因为本地有 2 GHz 可用。
- 当您尝试打开 VM-K2 的电源时，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，您无法打开需要 2 GHz 预留的 VM-K2 的电源。

图 14-4. 可扩展资源池的接入控制：无法打开电源



vSphere 集群服务 (vCLS) 默认处于激活状态，并在所有 vSphere 集群中运行。vCLS 可确保在 vCenter Server 变得不可用时，集群服务仍可用于维护在集群中运行的工作负载的资源 and 运行状况。仍需要 vCenter Server 才能运行 DRS 和 HA。

升级到 vSphere 7.0 Update 3 或新的 vSphere 7.0 Update 3 或更高版本部署时，会激活 vCLS。vCLS 会在升级 vCenter Server 过程中进行升级。

vCLS 使用代理虚拟机维护集群服务的运行状况。将主机添加到集群时，将创建 vCLS 代理虚拟机（vCLS 虚拟机）。每个 vSphere 集群中最多需要运行 3 个 vCLS 虚拟机，并在集群内进行分发。此外，也可在仅包含一个或两个主机的集群上激活 vCLS。在这些集群中，vCLS 虚拟机数量分别是 1 和 2。

将自动应用新的反关联性规则。每三分钟执行一次检查，如果多个 vCLS 虚拟机位于一个主机上，则这些虚拟机将自动重新分配到不同的主机。

表 15-1. 集群中的 vCLS 代理虚拟机数

集群中的主机数	vCLS 代理虚拟机数
1	1
2	2
3 个或更多	3

vCLS 虚拟机会在每个集群中运行，即使在集群上未激活 vSphere DRS 或 vSphere HA 等集群服务也无妨。vCLS 虚拟机的生命周期操作由 ESX Agent Manager 和工作负载控制平面等 vCenter Server 服务进行管理。vCLS 虚拟机不支持网卡。

如果 ESXi 版本与 vCenter Server 兼容，则激活了 vCLS 的集群可以包含不同版本的 ESXi 主机。vCLS 可与 vSphere Lifecycle Manager 集群配合使用。

本章讨论了以下主题：

- vSphere DRS 和 vCLS 虚拟机
- 为 vCLS 虚拟机选择数据存储
- vCLS 数据存储放置
- 监控 vSphere 集群服务
- 维护 vSphere 集群服务的运行状况

- 将集群置于撤回模式
- 检索 vCLS 虚拟机的密码
- vCLS 虚拟机反关联性策略
- 创建或删除 vCLS 虚拟机反关联性策略

vSphere DRS 和 vCLS 虚拟机

vSphere DRS 是 vSphere 的一项重要功能，要保证 vSphere 集群内运行的工作负载正常运行，必须使用此功能。DRS 取决于 vCLS 虚拟机的可用性。

注 如果尝试在 vCLS 虚拟机出现问题的集群上激活 DRS，则会在**集群摘要**页面上显示一条警告消息。

注 如果 DRS 已启用但 vCLS 虚拟机出现问题，必须解决这些问题，DRS 才能正常运行。将在**集群摘要**页面上显示一条警告消息。

如果 DRS 不起作用，这并不意味着 DRS 已停用。现有的 DRS 设置和资源池在 vCLS 虚拟机仲裁丢失后仍有效。当 vCLS 虚拟机未运行并因此跳过第一个 DRS 实例时，vCLS 运行状况仅在已激活 DRS 的集群中变为**不正常**。当至少一个 vCLS 虚拟机未运行时，vCLS 运行状况将在未激活 DRS 的集群上保持**已降级**状态。

为 vCLS 虚拟机选择数据存储

将根据连接到集群内主机的所有数据存储的排名，自动为 vCLS 虚拟机选择数据存储。

如果集群中的主机具有连接到数据存储的闲置预留 DRS 插槽，则选择该数据存储的几率更大。在选择本地数据存储之前，如有可能，算法会尝试将 vCLS 虚拟机置于共享数据存储中。首选具有更多可用空间的数据存储，并且算法尽量不将多个 vCLS 虚拟机放置在同一个数据存储上。只有在部署和打开 vCLS 虚拟机电源后，才能更改这些虚拟机的数据存储。

如果要将 vCLS 虚拟机的 VMDK 移至其他数据存储或附加不同的存储策略，可以重新配置 vCLS 虚拟机。执行此操作时，将显示一条警告消息。

可以通过执行 Storage vMotion 将 vCLS 虚拟机迁移到其他数据存储。如果要将 vCLS 虚拟机与工作负载虚拟机分开进行分组，例如，如果对在数据中心中运行的所有虚拟机使用特定的元数据策略，则可以标记 vCLS 虚拟机或附加自定义属性。

注 将数据存储置于维护模式时，如果数据存储托管 vCLS 虚拟机，则必须手动对 vCLS 虚拟机应用 Storage vMotion 以将其移至新位置或将集群置于撤回模式。将显示一条警告消息。

进入维护模式任务将开始但无法完成，因为有 1 个虚拟机驻留在数据存储上。如果决定继续，可以始终在“近期任务”中取消该任务 (The enter maintenance mode task will start but cannot finish because there is 1 virtual machine residing on the datastore. You can always cancel the task in your Recent Tasks if you decide to continue)。

所选数据存储可能存储无法关闭电源的 vSphere 集群服务虚拟机。为确保 vSphere 集群服务正常运行，必须在将此数据存储关闭进行维护之前，手动将这些虚拟机通过 vMotion 迁移到集群中的其他数据存储。请参阅以下知识库文章：KB 79892 (The selected datastore might be storing vSphere Cluster Services VMs which cannot be powered off. To ensure the health of vSphere Cluster Services, these VMs have to be manually vMotioned to a different datastore within the cluster prior to taking this datastore down for maintenance. Refer to this KB article: KB 79892)。

选中复选框**让我迁移所有虚拟机的存储，并在迁移之后继续进入维护模式。**以继续。

vCLS 数据存储放置

可以替代默认 vCLS 虚拟机数据存储放置。

vSphere 集群服务 (vCLS) 虚拟机数据存储位置是按默认数据存储选择逻辑选择的。要替代集群的默认 vCLS 虚拟机数据存储放置，可以通过浏览到集群并单击**配置 > vSphere 集群服务 > 数据存储**下的**添加**来指定一组允许的数据存储。无法为 vCLS 选择某些数据存储，因为 SRM 或 vSAN 等无法配置 vCLS 的解决方案会阻止这些数据存储。用户无法为 vCLS 虚拟机添加或移除解决方案阻止的数据存储。

监控 vSphere 集群服务

可以监控 vCLS 虚拟机消耗的资源及其运行状况。

vCLS 虚拟机不会显示在**主机和集群**选项卡的清单树中。数据中心内所有集群中的 vCLS 虚拟机都放在一个名为 **vCLS** 的单独虚拟机和模板文件夹中。仅可在 vSphere Client 的**虚拟机和模板**选项卡中查看此文件夹和 vCLS 虚拟机。这些虚拟机通过与常规工作负载虚拟机不同的图标进行标识。可以在 vCLS 虚拟机的**摘要**选项卡中查看有关 vCLS 虚拟机用途的信息。

可以在**监控**选项卡中监控 vCLS 虚拟机消耗的资源。

表 15-2. vCLS 虚拟机资源分配

属性	大小
VMDK 大小	245 MB（精简磁盘）
内存	128 MB
CPU	1 vCPU

表 15-2. vCLS 虚拟机资源分配（续）

属性	大小
硬盘	2 GB
数据存储上的存储	480 MB（精简磁盘）

注 每个 vCLS 虚拟机在集群中都预留了 100 MHz 和 100 MB 的容量。根据集群中运行的 vCLS 虚拟机数量，可以为这些虚拟机预留最多 400 MHz 和 400 MB 容量。

可以在集群的**摘要**选项卡中显示的**集群服务** portlet 中监控 vCLS 的运行状况。

表 15-3. vCLS 的运行状况

状态	颜色编码	摘要
正常	绿色	如果至少有一个 vCLS 虚拟机正在运行，则无论集群包含多少个主机，状态仍保持为正常。
已降级	黄色	如果 vCLS 虚拟机的运行时间长于 3 分钟（180 秒），则状态为已降级。
不正常	红色	如果在已启用 DRS 的集群中，vCLS 虚拟机的运行时间为 3 分钟或更短时间，则状态为不正常。

维护 vSphere 集群服务的运行状况

vCLS 虚拟机始终打开电源，因为 vSphere DRS 取决于这些虚拟机的可用性。应将这些虚拟机视为系统虚拟机。只有管理员才能对 vCLS 虚拟机执行选择性操作。为避免集群服务失败，请不要在 vCLS 虚拟机上执行任何配置或操作。

vCLS 虚拟机受到保护，以防被意外删除。集群虚拟机和文件夹受到保护，以防被用户（包括管理员）修改。

只有管理员 SSO 组的成员用户才能执行以下操作：

- 对 vCLS 虚拟机进行只读访问
- 对 vCLS 虚拟机进行控制台访问
- 使用冷迁移或热迁移将 vCLS 虚拟机重新放置到新存储和/或计算资源上
- 为 vCLS 虚拟机使用标记和自定义属性

可能会破坏 vCLS 虚拟机正常运行的操作：

- 更改 vCLS 虚拟机的电源状况
- 对 vCLS 虚拟机进行资源重新配置，例如更改 CPU、内存、磁盘大小、磁盘放置
- 虚拟机加密
- 对 vCLS 虚拟机触发 vMotion

- 更改 BIOS
- 从清单中移除 vCLS 虚拟机
- 从磁盘中删除 vCLS 虚拟机
- 启用 vCLS 虚拟机的 FT
- 克隆 vCLS 虚拟机
- 配置 PMem
- 将 vCLS 虚拟机移至其他文件夹
- 重命名 vCLS 虚拟机
- 重命名 vCLS 文件夹
- 在 vCLS 虚拟机上启用 DRS 规则和替代
- 在 vCLS 虚拟机上启用 HA 准入控制策略
- 在 vCLS 虚拟机上启用 HA 替代
- 将 vCLS 虚拟机移至资源池
- 从快照恢复 vCLS 虚拟机

在 vCLS 虚拟机上执行任何破坏性操作时，都会显示警告对话框。

故障排除：

vCLS 虚拟机的运行状况（包括电源状况）由 EAM 和 WCP 服务管理。如果 vCLS 虚拟机打开电源失败，或者由于缺少 vCLS 虚拟机仲裁而跳过集群的第一个 DRS 实例，则会在集群摘要页面中显示一个横幅，以及一个指向知识库文章的链接，以帮助对错误状态进行故障排除。

由于 vCLS 虚拟机视为系统虚拟机，因此无需对这些虚拟机执行备份或生成快照。这些虚拟机的运行状况由 vCenter 服务进行管理。

将集群置于撤回模式

将数据存储置于维护模式时，如果数据存储托管 vCLS 虚拟机，则必须手动通过 Storage vMotion 将 vCLS 虚拟机迁移到新位置或将集群置于撤回模式。

此任务说明了如何将集群置于撤回模式。

步骤

- 1 登录到 vSphere Client。
- 2 导航到必须停用 vCLS 的集群。
- 3 从浏览器的 URL 复制集群域 ID。该 ID 应类似于 **domain-c(number)**。
- 4 导航到 vCenter Server 配置选项卡。
- 5 在高级设置下，单击编辑设置按钮。

- 6 添加新条目 `config.vcls.clusters.domain-c(number).enabled`。使用在步骤 3 中复制的域 ID。
- 7 将值设置为 `False`。
- 8 单击保存。

结果

vCLS 监控服务每 30 秒运行一次。在 1 分钟内，将清理集群中的所有 vCLS 虚拟机，且集群服务运行状况将设置为已降级。如果集群激活了 DRS，则 DRS 将停止运行，并在集群摘要中额外显示一条警告。DRS 即使激活也不起作用，直到通过将 vCLS 从撤回模式移除而对其进行重新配置。

在主机出现故障的情况下，vSphere HA 不会执行最佳位置放置。HA 依赖于 DRS 来提供放置建议。HA 仍会打开虚拟机的电源，但这些虚拟机可能不会在最佳主机上打开电源。

要使集群退出撤回模式，请将步骤 7 中的值更改为 `True`。

检索 vCLS 虚拟机的密码

可以检索用于登录到 vCLS 虚拟机的密码。

要确保集群服务运行状况，请避免访问 vCLS 虚拟机。本文适用于 vCLS 虚拟机上的明确诊断。

步骤

- 1 使用 SSH 登录到 vCenter Server Appliance。
- 2 运行以下 python 脚本。

```
/usr/lib/vmware-wcp/decrypt_clustervm_pw.py
```

- 3 读取密码的输出。

```
pwd-script-output  
  
Read key from file  
  
Connected to PSQL  
  
PWD: (password displayed here)
```

结果

使用检索到的密码，可以登录到 vCLS 虚拟机。

vCLS 虚拟机反关联性策略

vSphere 支持 vCLS 虚拟机与另一组工作负载虚拟机之间的反关联性。

计算策略提供了一种方法来指定 vSphere Distributed Resource Scheduler (DRS) 应如何将虚拟机放置在资源池中的主机上。使用 vSphere 计算策略编辑器创建和删除计算策略。您可以创建或删除计算策略，但不能修改该策略。如果删除策略定义中使用的类别标记，也会删除该策略。在 vSphere 中打开**虚拟机摘要**页面，以查看应用于虚拟机的计算策略及其与每个策略的合规性状态。您可以为一组反关联到 vCLS 虚拟机组的工作负载虚拟机创建计算策略。vCLS 反关联性策略可以包含用于一组工作负载虚拟机的单个用户可见标记，另一组 vCLS 虚拟机则通过内部进行识别。

创建或删除 vCLS 虚拟机反关联性策略

vCLS 虚拟机反关联性策略描述了某类虚拟机与 vCLS 系统虚拟机之间的关系。

vCLS 虚拟机反关联性策略不支持将 vCLS 虚拟机和应用程序虚拟机放置在同一主机上。如果您不希望 vCLS 虚拟机和运行关键工作负载的虚拟机在同一主机上运行，则这种策略会非常有用。运行关键工作负载（如 SAP HANA）的一些最佳做法需要专用主机。创建策略后，放置引擎会尝试将 vCLS 虚拟机放置在未运行策略虚拟机的主机上。

vCLS 虚拟机反关联性策略的实施可能会在多个方面受到影响：

- 如果策略应用于不同主机上的多个虚拟机，并且无法提供足够的主机来分发 vCLS 虚拟机，则 vCLS 虚拟机将整合到没有策略虚拟机的主机中。
- 如果置备操作指定目标主机，则即使违反策略，也会始终遵守该指定。DRS 会在后续修复周期中尝试将 vCLS 虚拟机移至合规主机。

步骤

- 1 为要包含在 vCLS 虚拟机反关联性策略中的每个虚拟机组创建类别和标记。
- 2 标记要包含的虚拟机。
- 3 创建 vCLS 虚拟机反关联性策略。
 - a 从 vSphere 中，单击**策略和配置文件 > 计算策略**。
 - b 单击**添加**以打开**新建计算策略**向导。
 - c 填写策略**名称**，然后从**策略类型**下拉控件中选择 **vCLS 虚拟机反关联性**。
策略**名称**必须唯一。
 - d 提供策略的**描述**，然后使用**虚拟机标记**选择应用策略的**类别**和**标记**。
除非有多个虚拟机标记与某个类别相关联，否则向导会在选择标记**类别**后填入虚拟机标记。
 - e 单击**创建**以创建策略。
- 4 （可选）要删除计算策略，请打开 vSphere，单击**策略和配置文件 > 计算策略**，将每个策略显示为一个卡视图。单击“删除”以删除策略。

创建 DRS 集群

16

集群是一组具有共享资源和共享管理界面的 ESXi 主机和相关虚拟机。必须首先创建集群并激活 DRS，然后才能从集群级别资源管理中获益。

在集群中使用 vSphere Fault Tolerance (vSphere FT) 虚拟机时，DRS 的行为有所不同，具体取决于是否激活增强型 vMotion 兼容性 (EVC)。

表 16-1. 在使用 vSphere FT 虚拟机和 EVC 情况下的 DRS 行为

EVC	DRS (负载均衡)	DRS (初始放置)
已启用	已启用 (主虚拟机和辅助虚拟机)	已启用 (主虚拟机和辅助虚拟机)
已禁用	已禁用 (主虚拟机和辅助虚拟机)	已禁用 (主虚拟机) 全自动 (辅助虚拟机)

本章讨论了以下主题：

- 准入控制和初始放置
- 单个虚拟机打开电源
- 组启动
- 虚拟机迁移
- DRS 迁移阈值
- 迁移建议
- DRS 集群要求
- 共享存储器要求
- 共享的 VMFS 卷要求
- 处理器兼容性要求
- DRS 集群的 vMotion 要求
- 配置带有虚拟闪存的 DRS
- 创建集群
- 编辑集群设置
- 设置虚拟机的自定义自动化级别

- 停用 DRS
- 还原资源池树
- vSAN 延伸集群的 DRS 感知

准入控制和初始放置

尝试在已启用 DRS 的集群内打开一个或一组虚拟机的电源时，vCenter Server 会执行准入控制。它会检查集群内是否有足够的资源来支持虚拟机。

如果集群没有足够的资源，无法打开单个虚拟机的电源，或无法打开组启动尝试中任何虚拟机的电源，将会显示一条消息。否则，对于每台虚拟机，DRS 将生成要在其上运行虚拟机的主机的建议，并执行以下操作之一

- 自动执行放置建议。
- 显示用户随后可以选择接受或覆盖的放置建议。

注 对于独立主机或非 DRS 集群内的虚拟机，不提出任何初始放置建议。这些虚拟机将在打开电源时被置于当前所在的主机上。

- DRS 会考虑网络带宽。通过计算主机网络饱和度，DRS 可以更好地做出放置决策。更全面地了解环境，有助于避免虚拟机的性能下降。

单个虚拟机打开电源

在 DRS 集群中，可以打开单个虚拟机的电源，并接受初始放置位置建议。

打开单个虚拟机电源时，有两种类型的初始放置位置建议：

- 打开单个虚拟机电源，不需要任何必备条件步骤。
用户将拥有虚拟机的初始放置位置建议列表，这些建议是互斥的。您只能选择一种建议。
- 打开单个虚拟机的电源，但需要执行必备条件操作。

这些操作包括在待机模式下打开主机电源或在主机间迁移其他虚拟机。在这种情况下，提供的建议具有多行，显示每个必备条件操作。用户可以接受整个建议，也可以取消打开虚拟机电源。

组启动

可以尝试同时打开多个虚拟机的电源（组启动）。

选定进行组启动尝试的虚拟机不必位于同一个 DRS 集群内。可以在集群间选择虚拟机，但它们必须属于同一数据中心。也可以包括位于非 DRS 集群或独立主机上的虚拟机。这些虚拟机会自动开机并且不包括在任何初始放置建议中。

每个集群均提供组启动尝试的初始放置建议。如果组启动尝试的所有放置相关操作都处于自动模式，虚拟机将开机且不会提供任何初始放置建议。如果所有虚拟机的放置相关操作均处于手动模式，则会手动打开所有虚拟机的电源（包括处于自动模式的虚拟机）。这些操作包含在初始放置建议中。

对于已打开电源的虚拟机所属的每个 DRS 集群，均会有一个建议，其中包含所有必备条件（或没有建议）。所有特定于此类集群的建议都显示在**启动建议**选项卡下。

如果进行了非自动组启动尝试，并且包括不受初始放置建议限制的虚拟机（即独立主机上的虚拟机或非 DRS 集群中的虚拟机），vCenter Server 会尝试自动打开这些虚拟机的电源。如果这些虚拟机开机成功，则会显示在**已开始启动**选项卡下。所有无法开机的虚拟机将显示在**失败的启动**选项卡下。

示例：组启动

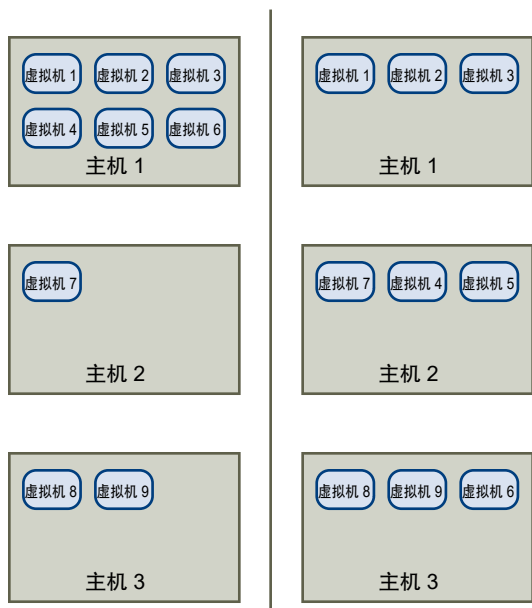
用户选择同一数据中心中的三个虚拟机进行组启动尝试。前两个虚拟机（VM1 和 VM2）在同一 DRS 集群（Cluster1）中，而第三个虚拟机（VM3）则在一台独立主机上。VM1 处于自动模式，而 VM2 处于手动模式。在此方案中，用户将获得 Cluster1 的初始放置建议（位于**启动建议**选项卡下），其中包含打开 VM1 和 VM2 电源的操作。将尝试自动打开 VM3 的电源，如果成功，将会显示在**已开始启动**选项卡下。如果此尝试失败，将会显示在**失败的启动**选项卡下。

虚拟机迁移

尽管 DRS 执行初始放置位置以便跨集群平衡负载，但是虚拟机负载和资源可用性中的更改可能会导致集群失衡。要更正此失衡情况，DRS 将生成迁移建议。

如果在集群上启用了 DRS，则可以更均匀地分配负载，从而降低不平衡程度。例如，下图中左侧的三台主机不平衡。假定主机 1、主机 2 和主机 3 具有相同的容量，且所有虚拟机的配置和负载（包括预留，如果已设置）均相同。但是，由于主机 1 有六个虚拟机，其资源可能被过度利用，而主机 2 和主机 3 上有丰富的可用资源，因此，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）。该图右侧显示了正确平衡负载之后所呈现的主机配置。

图 16-1. 负载平衡



当集群不平衡时，DRS 将根据默认的自动化级别，提出建议或迁移虚拟机：

- 如果所涉及的集群或任何虚拟机为手动或半自动，则 vCenter Server 不执行自动操作来平衡资源。“摘要”页面会指示有迁移建议，“DRS 建议”页面会显示最有效地利用集群内资源的更改建议。
- 如果所涉及的集群或虚拟机均为全自动，则 vCenter Server 将根据需要在主机间迁移正在运行的虚拟机，以确保高效利用集群资源。

注 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 vCenter Server 可能会将这些虚拟机迁移到其他主机，以优化集群资源。

默认情况下，自动化级别是为整个集群指定的。也可以为单个虚拟机指定自定义的自动化级别。

DRS 迁移阈值

DRS 迁移阈值允许您指定要生成并应用的建议（如果建议中所涉及的虚拟机处于全自动模式）或要显示的（如果处于手动模式）。此阈值是衡量 DRS 在建议迁移中提高虚拟机良好状态的激进程度的指标。

可以移动阈值滑块以使用从“保守”到“激进”这五个设置中的一个。激进程度设置越高，DRS 建议迁移以提高虚拟机良好状态的频率就越高。“保守”设置仅生成优先级 1 的建议（强制性建议）。

在建议收到优先级后，会将该级别与您所设置的迁移阈值进行比较。如果优先级低于或等于阈值设置，则会应用该建议（如果相关虚拟机均处于全自动模式），或向用户显示该建议以进行确认（如果处于手动或半自动模式）。

DRS 评分

每个迁移建议都使用虚拟机幸福感衡量指标进行计算，该衡量指标用于衡量执行效率。该衡量指标在 vSphere Client 中的集群“摘要”选项卡中显示为“DRS 评分”。DRS 负载平衡建议尝试改进虚拟机的 DRS 评分。“集群 DRS 评分”是集群中所有已打开电源的虚拟机的“虚拟机 DRS 评分”的加权平均值。“集群 DRS 评分”显示在计量器组件中。填充部分的颜色将随值而变化，以与“虚拟机 DRS 评分”直方图中的相应直条匹配。直方图中的直条显示在该范围内具有 DRS 评分的虚拟机所占百分比。可以利用以下方法通过服务器端排序和筛选查看列表：选择集群的“监控”选项卡，然后选择“vSphere DRS”，将显示集群中按其 DRS 评分升序排序的虚拟机列表。

迁移建议

如果创建带有默认模式（手动或半自动）的集群，则 vCenter Server 将在“DRS 建议”页面上显示迁移建议。

系统将提供足够的建议，以强制实施规则并平衡集群的资源。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 负载或预留。
- 平衡平均内存负载或预留。
- 满足资源池预留。
- 满足关联性规则。

- 主机正在进入维护模式或待机模式。

注 如果使用 vSphere Distributed Power Management (DPM) 功能，那么，除了迁移建议外，DRS 还会提供主机电源状况建议。

DRS 集群要求

添加到 DRS 集群的主机必须满足某些要求才能成功使用集群功能。

注 vSphere DRS 是 vSphere 的一项重要功能，要维持在 vSphere 集群内运行的工作负载正常运行，必须使用此功能。从 vSphere 7.0 Update 1 开始，DRS 依赖于 vCLS 虚拟机的可用性。有关详细信息，请参见第 15 章 [vSphere 集群服务](#)。

共享存储器要求

DRS 集群具有特定的共享存储器要求。

确保受管主机使用共享存储器。共享存储器通常位于 SAN 上，但也可以通过使用 NAS 共享存储器来实现。

有关其他共享存储器的信息，请参见《vSphere 存储》文档。

共享的 VMFS 卷要求

DRS 集群具有某些共享的 VMFS 卷要求。

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。
- 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 虚拟机交换文件还需要放在源主机和目标主机均可以访问的 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，并且使用主机-本地交换，则此要求将不适用。这种情况下，支持将带有交换文件的 vMotion 置于非共享存储器上。默认情况下，交换文件置于 VMFS 上，但管理员可以使用高级虚拟机配置选项替代此文件位置。

处理器兼容性要求

DRS 集群具有特定的处理器兼容性要求。

为了避免限制 DRS 的功能，应当将集群内源和目标主机的处理器兼容性最大化。

vMotion 在基础 ESXi 主机之间传输虚拟机的运行架构状况。vMotion 兼容性是指目标主机的处理器必须能够使用等效指令，从源主机的处理器在挂起时的状态继续执行。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。

处理器系列由处理器供应商定义。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。

有时，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）。如果不能保证通过 vMotion 成功迁移，VMware 会识别这些异常情况。

vCenter Server 提供了一些有助于确保通过 vMotion 迁移的虚拟机满足处理器兼容性要求的功能。这些功能包括：

- **增强型 vMotion 兼容性 (EVC)** - 可以使用 EVC 帮助确保集群内主机的 vMotion 兼容性。EVC 可以确保集群内的所有主机向虚拟机提供相同的 CPU 功能集，即使这些主机上的实际 CPU 不同也是如此。这样可以避免因 CPU 不兼容而导致通过 vMotion 迁移失败。

在“集群设置”对话框中配置 EVC。为了使集群能够使用 EVC，集群内的主机必须满足某些要求。有关 EVC 和 EVC 要求的信息，请参见《vCenter Server 和主机管理》文档。

- **CPU 兼容性掩码** - vCenter Server 会将虚拟机可用的 CPU 功能与目标主机的 CPU 功能进行比较，以确定是允许还是禁止通过 vMotion 迁移。通过将 CPU 兼容性掩码应用到单个虚拟机，可以向虚拟机隐藏某些 CPU 功能，从而防止由于 CPU 不兼容而造成的 vMotion 迁移失败。

DRS 集群的 vMotion 要求

DRS 集群具有特定的 vMotion 要求。

要启用 DRS 迁移建议的使用，集群内的主机必须是 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。

要为 vMotion 进行配置，集群内的每台主机必须满足下列要求：

- vMotion 不支持裸磁盘，也不支持对借助于 Microsoft 集群服务 (MSCS) 集群的应用程序进行迁移。
- vMotion 要求在所有启用了 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，需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。

配置带有虚拟闪存的 DRS

DRS 可以管理具有虚拟闪存预留的虚拟机。

虚拟闪存容量会显示为主机定期向 vSphere Client 报告的统计数据。DRS 每次运行时，都使用最新报告的容量值。

可以在每个主机上配置一个虚拟闪存资源。这表示在虚拟机打开电源期间，DRS 不需要在给定主机上的不同虚拟闪存资源之间进行选择。

DRS 选择具有足够可用虚拟闪存容量的主机以启动虚拟机。如果 DRS 无法满足虚拟机的虚拟闪存预留，则无法打开其电源。DRS 将具有虚拟闪存预留且打开电源的虚拟机视为与其当前主机之间具有软关联性。DRS 建议不要使用此类虚拟机执行 vMotion 操作，除非有必须使用的理由，例如将主机置于维护模式或者降低使用过度的主机上的负载。

创建集群

集群是一组主机。将主机添加到集群时，主机的资源将成为集群资源的一部分。集群管理其中所有主机的资源。

集群启用 vSphere High Availability (HA) 和 vSphere Distributed Resource Scheduler (DRS) 解决方案。

注 vSphere DRS 是 vSphere 的一项重要功能，要维持在 vSphere 集群内运行的工作负载正常运行，必须使用此功能。从 vSphere 7.0 Update 1 开始，DRS 依赖于 vCLS 虚拟机的可用性。有关详细信息，请参见第 15 章 vSphere 集群服务。

前提条件

- 确认您拥有足够的权限，可以创建集群对象。
- 确认清单中存在数据中心。
- 如果想要使用 vSAN，必须在配置 vSphere HA 之前启用它。

步骤

- 1 在 vSphere Client 中，浏览到数据中心。
- 2 右键单击该数据中心并选择**新建集群**。
- 3 输入集群名称。
- 4 选择 DRS 和 vSphere HA 集群功能。

选项	描述
将 DRS 用于此集群的步骤	a 选中 DRS 打开复选框。
	b 选择一个自动化级别和迁移阈值。
将 HA 用于此集群的步骤	a 选中 vSphere HA 打开复选框。
	b 选择是否启用主机监控和准入控制。
	c 如果启用准入控制，请指定策略。
	d 选择一个虚拟机监控选项。
	e 指定虚拟机监控敏感度。

- 5 选择增强型 vMotion 兼容性 (EVC) 设置。

EVC 可以确保集群内的所有主机向虚拟机提供相同的 CPU 功能集，即使这些主机上的实际 CPU 不同也是如此。这样可以避免因 CPU 不兼容而导致通过 vMotion 迁移失败。

- 6 单击**确定**。

结果

已将集群添加到清单中。

后续步骤

将主机和资源池添加到集群。

注 在**集群摘要**页面下，可以看到**集群服务**，其中显示了 vSphere 集群服务运行状况。

编辑集群设置

将主机添加到 DRS 集群时，主机的资源将成为集群资源的一部分。除了这种资源聚合外，您还可以使用 DRS 集群支持集群范围内的资源池并强制执行集群级别的资源分配策略。

还提供下面的集群级别的资源管理功能。

负载均衡

将持续监控集群内所有主机和虚拟机的 CPU 和内存资源的分布情况和使用情况。在给出集群内资源池和虚拟机的属性、当前需求以及不平衡目标的情况下，DRS 会将这些衡量指标与理想状态下的资源使用情况进行比较。然后，DRS 提供建议或相应地执行虚拟机迁移。请参见[虚拟机迁移](#)。当您在集群中打开虚拟机电源时，DRS 将尝试通过在相应主机上放置该虚拟机或提出建议来保持适当的负载均衡。请参见[准入控制和初始放置](#)。

电源管理

vSphere Distributed Power Management (DPM) 功能启用后，DRS 会将集群级别和主机级别容量与集群的虚拟机需求（包括近期历史需求）进行比较。然后，在找到足够的额外容量时，DRS 建议您将主机置于待机状态，或将主机置于待机电源模式。如果需要容量，DRS 会打开主机电源。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[管理电源资源](#)。

关联性规则

可以通过分配关联性规则控制集群内主机上的虚拟机的放置。请参见[使用 DRS 关联性规则](#)。

前提条件

可以在没有特殊许可证的情况下创建集群，但必须要有许可证才能为 vSphere DRS 或 vSphere HA 启用集群。

注 vSphere DRS 是 vSphere 的一项重要功能，要维持在 vSphere 集群内运行的工作负载正常运行，必须使用此功能。从 vSphere 7.0 Update 1 开始，DRS 依赖于 vCLS 虚拟机的可用性。有关详细信息，请参见[第 15 章 vSphere 集群服务](#)。

步骤

- 1 在 vSphere Client 中浏览到某个集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 在 **vSphere DRS** 下，单击**编辑**。

4 在 DRS 自动化下，为 DRS 选择默认的自动化级别。

自动化级别	操作
手动	<ul style="list-style-type: none"> ■ 初始放置：显示建议的主机。 ■ 迁移：显示迁移建议。
半自动	<ul style="list-style-type: none"> ■ 初始放置：自动。 ■ 迁移：显示迁移建议。
全自动	<ul style="list-style-type: none"> ■ 初始放置：自动。 ■ 迁移：自动运行建议。

5 设置 DRS 的迁移阈值。

6 选中 **Predictive DRS** 复选框。除了实时衡量指标以外，DRS 还响应 vRealize Operations 服务器提供的预测衡量指标。您还必须在支持该功能的 vRealize Operations 版本中配置 **Predictive DRS**。

7 选中**虚拟机自动化**复选框以启用个别虚拟机自动化级别。

可在“虚拟机替代项”页面中设置个别虚拟机的替代项。

8 在**其他选项**下，选中一个复选框以执行某项默认策略。

选项	描述
虚拟机分布	出于可用性目的，在各主机间分布偶数数量的虚拟机。这是 DRS 负载均衡的辅助方式。
用于负载均衡的内存衡量指标	负载均衡基于虚拟机的已消耗内存而不是活动内存。仅建议将此设置用于主机内存未超额分配的集群。 注 此设置不再受支持，将不会显示在 vCenter 7.0 中。
CPU 超额分配	控制集群中的 CPU 超额分配。
可扩展份额	为此集群上的资源池启用可扩展份额。

9 在**电源管理**下，选择“自动化级别”。

10 如果已启用 DPM，请设置 **DPM 阈值**。

11 单击**确定**。

后续步骤

注 在**集群摘要**页面下，可以看到**集群服务**，其中显示了 vSphere 集群服务运行状况。

您可以在 vSphere Client 中查看 DRS 的内存利用率。要了解详细信息，请参见：



(查看 Distributed Resource Scheduler 内存利用率)

设置虚拟机的自定义自动化级别

创建 DRS 集群后，可以为各个虚拟机自定义自动化级别，以替代集群的默认自动化级别。

例如，可以为全自动的集群内的特定虚拟机选择**手动**，或为手动集群内的特定虚拟机选择**半自动**。

如果虚拟机已设置为**已禁用**，则 vCenter Server 将不会迁移该虚拟机或为其提供迁移建议。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 依次单击**配置**选项卡和服务。
- 3 在“服务”下，选择 **vSphere DRS**，然后单击**编辑**。展开“DRS 自动化”。
- 4 选中**启用单个虚拟机自动化级别**复选框。
- 5 要临时停用任何单个虚拟机替代项，请取消选中**启用个别虚拟机自动化级别**复选框。
再次选中此复选框时，将还原虚拟机设置。
- 6 要临时挂起集群中的所有 vMotion 活动，请将集群置于手动模式，并取消选中**启用个别虚拟机自动化级别**复选框。
- 7 选择一个或多个虚拟机。
- 8 单击**自动化级别**列，然后从下拉菜单选择自动化级别。

选项	描述
手动	将显示放置和迁移建议，但在手动应用建议之前，不会运行这些建议。
全自动	放置和迁移建议会自动运行。
半自动	初始放置会自动执行。将显示迁移建议，但不运行。
已禁用	vCenter Server 将不会迁移虚拟机或为其提供迁移建议。

- 9 单击**确定**。

结果

注 其他 VMware 产品或功能（如 vSphere vApp 和 vSphere Fault Tolerance）可能会替代 DRS 集群内虚拟机的自动化级别。有关详细信息，请参见特定于产品的文档。

停用 DRS

可以关闭集群的 DRS。

停用 DRS 后，集群的资源池层次结构和关联性规则不会在您再次打开 DRS 时重新建立。如果停用 DRS，将从集群中移除资源池。为了避免丢失资源池，请在本地计算机上保存资源池树快照。激活 DRS 时，可以使用快照还原资源池。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 在 **vSphere DRS** 的下面，单击**编辑**。
- 4 取消选中**打开 vSphere DRS** 复选框。
- 5 单击**确定**，关闭 DRS。
- 6 （可选）选择用于保存资源池的选项。
 - 单击**是**以在本地计算机上保存资源池树快照。
 - 单击**否**以关闭 DRS，而不保存资源池树快照。

结果

DRS 已关闭。

注 vSphere DRS 是 vSphere 的一项重要功能，要维持在 vSphere 集群内运行的工作负载正常运行，必须使用此功能。从 vSphere 7.0 Update 1 开始，DRS 依赖于 vCLS 虚拟机的可用性。有关详细信息，请参见第 15 章 **vSphere 集群服务**。

还原资源池树

您可以还原以前保存的资源池树快照。

前提条件

- vSphere DRS 必须打开。
- 您只能在创建快照的同一集群中还原快照。
- 在该集群中不存在其他资源池。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 右键单击该集群，然后选择**还原资源池树**。
- 3 单击**浏览**，然后在本地计算机上查找快照文件。
- 4 单击**打开**。
- 5 单击**确定**以还原该资源池树。

vSAN 延伸集群的 DRS 感知

使用 vSphere 7.0 U2 启用 DRS 的延伸集群上提供 vSAN 延伸集群的 DRS 感知。vSAN 延伸集群具有读取局部性，其中虚拟机从本地站点读取数据。从远程站点获取读取可能会影响虚拟机性能。在 vSphere 7.0 U2 之前的版本中，DRS 无法感知 vSAN 延伸集群的读取局部性，并且可能会无意中将虚拟机放在没

有读取局部性的远程站点上。借助 vSAN 延伸集群的 DRS 感知，DRS 现在可以完全感知虚拟机读取局部性，并且将虚拟机放在完全满足读取局部性的站点上。这是自动操作，没有可配置选项。vSAN 延伸集群的 DRS 感知使用现有关联性规则。其适用于 vSphere 7.0 U2 和 VMware Cloud on AWS。

具有 vSphere HA 和 vSphere DRS 的 vSAN 延伸集群通过将两个数据副本分散到两个故障域以及第三个故障域中的一个见证节点提供灵活性，以防出现故障。两个活动故障域提供数据复制，以便两个故障域具有数据的当前副本。

vSAN 延伸集群提供了在两个故障域中自动移动工作负载的方法。如果整个站点发生故障，vSphere HA 将在辅助站点上重新启动虚拟机。这可确保关键生产工作负载不会停机。在主站点重新联机后，DRS 会立即使用软性关联性主机将虚拟机重新平衡回主站点。此过程会导致在虚拟机数据组件仍在重建时从辅助站点读取和写入虚拟机，并且可能会降低虚拟机性能。

在 vSphere 7.0 U2 之前的版本中，我们建议将 DRS 从全自动模式更改为半自动模式，以避免在进行重新同步时将虚拟机迁移到主站点。仅在重新同步完成后，才能将 DRS 重新设置为全自动。

在 vSphere 7.0 U2 中，vSAN 延伸集群的 DRS 感知引入全自动读取局部性解决方案，用于从 vSAN 延伸集群上的故障中恢复。读取局部性信息指出虚拟机具有完全访问权限的主机，且在将虚拟机放在 vSAN 延伸集群上的主机上时，DRS 会使用此信息。DRS 可防止在站点恢复阶段 vSAN 重新同步仍在进行时虚拟机回退到主站点。当虚拟机的数据组件达到完全读取局部性时，DRS 会自动将虚拟机迁移回主关联站点。这样，您可以在整个站点发生故障时以全自动模式运行 DRS。

在部分站点发生故障时，如果虚拟机由于丢失的数据组件超过或等于其允许的故障数而丢失读取局部性，vSphere DRS 将确定消耗极高读取带宽的虚拟机，并尝试将其重新平衡到辅助站点。这可确保在部分站点故障期间，具有包含大量读取操作的工作负载的虚拟机不会受到影响。在主站点重新联机并且数据组件已完成重新同步后，虚拟机将移回其关联站点。

具有 ROBO 企业许可证的 DRS 维护模式功能

17

主机进入维护模式时，VMware 的大型 Remote Office Branch Office (ROBO) 企业许可证支持自动撤出虚拟机。

在 ROBO 企业集群中，DRS 默认处于停用状态，您无法更改 DRS 配置。当 ROBO 企业集群中的主机进入维护模式时，DRS 会自动将虚拟机从主机撤出。从主机撤出虚拟机之前，DRS 会创建虚拟机-主机关联性映射以跟踪虚拟机的放置位置。主机退出维护模式时，之前在主机上运行的虚拟机将迁移回主机。迁移后，虚拟机-主机关联性映射将被清除。

本章讨论了以下主题：

- 具有 ROBO 企业许可证的 DRS 维护模式存在的限制
- 使用具有 ROBO 企业许可证的 DRS 维护模式
- 对具有 ROBO 企业许可证的 DRS 维护模式进行故障排除

具有 ROBO 企业许可证的 DRS 维护模式存在的限制

具有 ROBO 企业许可证的 DRS 功能不是完整的 DRS 功能。

在 ROBO 企业集群上启动维护模式之前，应当了解一些存在的限制。在 ROBO 企业集群上，DRS 默认为停用。如果已从 DRS 支持的许可证迁移到 ROBO 企业许可证，系统中可能存在具有关联性或反关联性规则的虚拟机。必须停用或删除具有关联性或反关联性规则的虚拟机，否则将停用 ROBO 企业维护模式操作。如果 DRS 未设置为全自动模式，将停用 ROBO 企业维护模式操作。DRS 自动化级别必须设置为全自动模式，以通过主机维护工作流自动撤出虚拟机。如果虚拟机替代了 DRS 全自动模式，您必须手动撤出虚拟机。

使用具有 ROBO 企业许可证的 DRS 维护模式

vSphere 支持具有 ROBO 企业许可证的有限 DRS 维护模式功能。

前提条件

- 检查集群中的所有主机是否都已安装 ROBO Enterprise 许可证。如果没有，您必须安装许可证。
- 检查是否已配置并激活任何 DRS 规则。如果是，必须停用或删除它们以使用 ROBO 企业维护模式操作。

步骤

1 为使 DRS 维护模式使用 ROBO 企业许可证，请确保集群上的每个主机已安装 ROBO 企业许可证。

- 如果未安装许可证，请转到步骤 2。
- 如果已安装许可证，请转到步骤 3。

2 安装 ROBO 企业许可证

- a 在 vSphere Client 中，浏览到主机。
- b 在配置选项卡下，选择许可证。
- c 单击分配许可证。
- d 输入您的 ROBO 企业许可证密钥，然后单击确定。

您必须为集群中的所有主机重复这些步骤。

3 选择集群中的主机，右键单击并选择进入维护模式，然后单击确定。

将自动撤出主机上的虚拟机。

结果

主机退出维护模式后，虚拟机将自动迁移回主机。主机将还原到原始状态。但是，如果主机过载，DRS 则无法将虚拟机迁回原始主机。DRS 会尝试将主机还原到原始状态，但不能使主机过载。

后续步骤

如果您要停用具有 ROBO 企业许可证的 DRS 维护模式，可以编辑 vpxd.cfg 文件。打开 vpxd.cfg 文件。在 `<cluster>` 选项下，将 `<roboMMEEnabled>true</roboMMEEnabled>` 改为 `<roboMMEEnabled>>false</roboMMEEnabled>`。这是运行时配置，因此更新配置后无需重新启动 vpxd。

对具有 ROBO 企业许可证的 DRS 维护模式进行故障排除

如果在 ROBO 企业集群中使用维护模式时遇到问题，请考虑以下注意事项。

为使维护模式在 ROBO 企业集群中正常工作：

- 检查集群中的所有主机是否都已安装 ROBO Enterprise 许可证。如果没有，您必须安装许可证。
- 检查是否已配置并激活任何 DRS 规则。如果是，必须停用或删除它们以使用 ROBO 企业维护模式操作。
- 如果兼容性检查失败，请确保其他主机与虚拟机兼容。

使用 DRS 集群管理资源

18

创建 DRS 集群后，可以对其进行自定义，并使用它来管理资源。

要自定义 DRS 集群及其包含的资源，可以配置关联性规则，并添加和移除主机和虚拟机。定义集群的设置和资源后，应当确保它是且保持为有效集群。还可以使用有效 DRS 集群管理电源资源，并与 vSphere HA 互操作。

注 在本章中，“内存”可以指物理内存或永久内存。

本章讨论了以下主题：

- 将主机添加到集群
- 将虚拟机添加到集群
- 从集群内移除虚拟机
- 从集群中移除主机
- DRS 集群有效性
- 管理电源资源
- 使用 DRS 关联性规则

将主机添加到集群

对于由同一 vCenter Server 管理的主机（受管主机）和未由该服务器管理的主机，将主机添加到集群的步骤有所不同。

添加某个主机之后，部署到该主机的虚拟机将变为集群的一部分，而且 DRS 会建议将某些虚拟机迁移到集群内的其他主机。

注 vSphere DRS 是 vSphere 的一项重要功能，要维持在 vSphere 集群内运行的工作负载正常运行，必须使用此功能。从 vSphere 7.0 Update 1 开始，DRS 依赖于 vCLS 虚拟机的可用性。有关详细信息，请参见第 15 章 vSphere 集群服务。

将受管主机添加到集群

当将 vCenter Server 正在管理的独立主机添加到 DRS 集群时，该主机的资源将与集群相关联。

可以决定是要将现有的虚拟机和资源池与集群的根资源池相关联，还是移植资源池层次结构。

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

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 右键单击主机，然后选择**移至...**。
- 3 选择一个集群。
- 4 单击**确定**应用更改。
- 5 选择要对主机的虚拟机和资源池执行的操作。

- **将此主机的虚拟机置于集群的根资源池中**

vCenter Server 会移除主机上所有现有的资源池，而该主机层次结构中的虚拟机都将被附加到根。因为份额分配是相对于资源池的，而上述操作破坏了资源池层次结构，所以在选择此选项后可能必须手动更改虚拟机的份额。

- **为此主机的虚拟机和资源池创建资源池**

vCenter Server 创建将成为集群的直接子级的顶层资源池并将主机的所有子级添加到新资源池。您可以命名这个新的顶层资源池。默认名称是**移植自 <host_name>**。

结果

此时主机即会添加到集群。

将非受管主机添加到集群

可将非受管主机添加到集群。此类主机当前并未由集群所在的 vCenter Server 系统管理，而且在 vSphere Client 中不可见。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 右键单击该集群并选择**添加主机**。
- 3 输入主机名、用户名和密码，然后单击**下一步**。
- 4 查看摘要信息并单击**下一步**。
- 5 分配现有的或新的许可证密钥，然后单击**下一步**。
- 6 （可选） 您可以启用锁定模式以避免远程用户直接登录到主机。

如果不启用锁定模式，则可以稍后通过编辑主机设置中的“安全配置文件”来配置该选项。

7 选择要对主机的虚拟机和资源池执行的操作。

■ 将此主机的虚拟机置于集群的根资源池中

vCenter Server 会移除主机上所有现有的资源池，而该主机层次结构中的虚拟机都将被附加到根。因为份额分配是相对于资源池的，而上述操作破坏了资源池层次结构，所以在选择此选项后可能必须手动更改虚拟机的份额。

■ 为此主机的虚拟机和资源池创建资源池

vCenter Server 创建将成为集群的直接子级的顶层资源池并将主机的所有子级添加到新资源池。您可以命名这个新的顶层资源池。默认名称是**移植自 <host_name>**。

8 查看设置，然后单击**完成**。

结果

此时主机即会添加到集群。

将虚拟机添加到集群

可通过以下几种方式将虚拟机添加到集群。

- 如果将某个主机添加到一个集群，则该主机上的所有虚拟机均会添加到此集群。
- 当创建虚拟机时，**创建新的虚拟机**向导会提示您选择放置虚拟机的位置。可以选择独立主机或集群并选择主机或集群内的任意资源池。
- 可以使用**迁移虚拟机**向导将虚拟机从一台独立主机迁移到一个集群或者从一个集群迁移到另一个集群。要开始该向导，请右键单击虚拟机名称，然后选择**迁移**。

将虚拟机移到集群

可以将虚拟机移到集群中。

步骤

- 1 在 vSphere Client 中找到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后选择**移至...**。
- 3 选择一个集群。
- 4 单击**确定**。

从集群内移除虚拟机

可以从集群内移除虚拟机。

可通过两种方式从集群内移除虚拟机。

- 当从集群内移除主机时，所有未迁移到其他主机的已关闭电源的虚拟机也会被移除。主机只有在维护模式或断开的情况下才可以被移除。如果从 DRS 集群内移除主机，集群可能会因集群过载而变成黄色。
- 可以使用**迁移**向导将虚拟机从集群迁移到独立主机，或者从一个集群迁移到另一个集群。要启动此向导，请右键单击虚拟机名称，然后选择**迁移**。

将虚拟机移出集群

可以将虚拟机移出集群。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后选择**迁移**。
- 3 选择**更改数据存储**，然后单击**下一步**。
- 4 选择一个数据存储，然后单击**下一步**。
- 5 单击**完成**。

如果虚拟机属于 DRS 集群规则组，则 vCenter Server 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。必须在执行迁移操作之前确认该警告。

从集群中移除主机

从 DRS 集群中移除主机时，会影响资源池层次结构、虚拟机，而且可能会创建无效集群。在移除主机之前，请先考虑受影响的对象。

- 资源池层次结构 - 即使在将某个主机添加到集群时使用了 DRS 集群并决定移植主机资源池，在将该主机从集群内移除后，其上也只保留根资源池。在这种情况下，层次结构将随集群保留。可以创建一个特定于主机的资源池层次结构。

注 必须先将主机置于维护模式，才能将其从集群内移除。相反，如果先断开主机的连接，然后再将其从集群内移除，则主机会保留反映集群层次结构的资源池。

- 虚拟机 - 主机必须处于维护模式才能从集群中移除，而且对于要进入维护模式的主机，必须将所有已打开电源的虚拟机迁移出该主机。当请求主机进入维护模式时，会询问您是否要将该主机上所有已关闭电源的虚拟机迁移到集群内的其他主机上。

- 无效集群 - 当从集群内移除主机时，可供集群使用的资源会减少。如果集群有足够的资源用于满足集群内所有虚拟机和资源池的预留需要，则集群会调整资源的分配以反映减少的资源量。如果集群没有足够的资源满足所有资源池的预留需要，但是有足够的资源满足所有虚拟机的预留需要，就会出现警报，而且该集群会被标记为黄色。DRS 继续运行。

将主机置于维护模式

当需要维护主机时（例如，要安装更多内存），请将主机置于维护模式。主机仅会因用户请求而进入或离开维护模式。

如果主机将进入维护模式，则需将其上正在运行的虚拟机迁移到其他主机。此时主机将处于**进入维护模式**这一状况，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为止。如果主机正在进入维护模式，则无法打开其上的虚拟机电源，也无法将虚拟机迁移到该主机。

当主机上不再有正在运行的虚拟机时，该主机的图标将发生变化，并新增显示**维护模式**，并且该主机的“摘要”面板会指示新的状况。在维护模式下，主机不允许您部署虚拟机，也不允许您打开虚拟机电源。

注 如果主机进入所请求的模式后会违反 vSphere HA 故障切换级别，则 DRS 不会建议将任何虚拟机从进入维护或待机模式的主机中迁出（在全自动模式下，则不执行这样的迁移）。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 右键单击主机，然后选择**维护模式 > 进入维护模式**。
 - 如果主机是部分自动化或手动 DRS 集群的一部分，请浏览到**集群 > 监控 > DRS > 建议**，然后单击**应用建议**。
 - 如果主机属于自动模式下的 DRS 集群，则虚拟机将在主机进入维护模式时迁移到其他主机。
- 3 如果适用，请单击**是**。

结果

在选择**维护模式 > 退出维护模式**之前，主机一直处于维护模式。

从集群中移除主机

可以从集群内移除主机。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 右键单击主机，然后选择**维护模式 > 进入维护模式**。
主机处于维护模式时，可以将其移到其他清单位置，该位置可以是顶层数据中心或者其他集群。
- 3 右键单击主机，然后选择**移至...**。
- 4 为主机选择一个新位置，然后单击**确定**。

结果

移动主机时，主机的资源会从集群内移除。如果将主机的资源池层次结构移植到集群上，则该层次结构将随集群保留。

后续步骤

从集群中移除主机后，可以执行以下任务。

- 将主机从 vCenter Server 中移除。
- 在 vCenter Server 下将主机作为独立主机运行。
- 将主机移至另一个集群。

使用待机模式

将主机置于待机模式时，会将其关闭电源。

通常，主机由 vSphere DPM 功能置于待机模式以优化电源使用情况。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电源状态，请将其置于维护模式并将其关闭电源。

DRS 集群有效性

vSphere Client 会指示 DRS 集群是有效、过载（黄色）还是无效（红色）。

DRS 集群由于多个原因而变得过载或无效。

- 集群可能由于一台主机发生故障而过载。
- 如果 vCenter Server 不可用，并且使用 vSphere Client 打开虚拟机电源，则 DRS 集群将变为无效。
- 如果用户在虚拟机进行故障切换时减少父资源池上的预留，则集群将变为无效。
- 如果在 vCenter Server 不可用时使用 vSphere Client 对主机或虚拟机进行更改，则这些更改将生效。但是，当 vCenter Server 再次可用时，您可能会发现集群由于不再满足集群要求而变为红色或黄色。

当考虑集群有效性情况时，应当了解以下术语。

预留

保证分配给资源池的固定量，由用户输入。

使用的预留

预留总量或每个子级资源池所使用的预留量（以较大者为准），以递归方式相加。

未预留

这个非负数会根据资源池类型不同而有所不同。

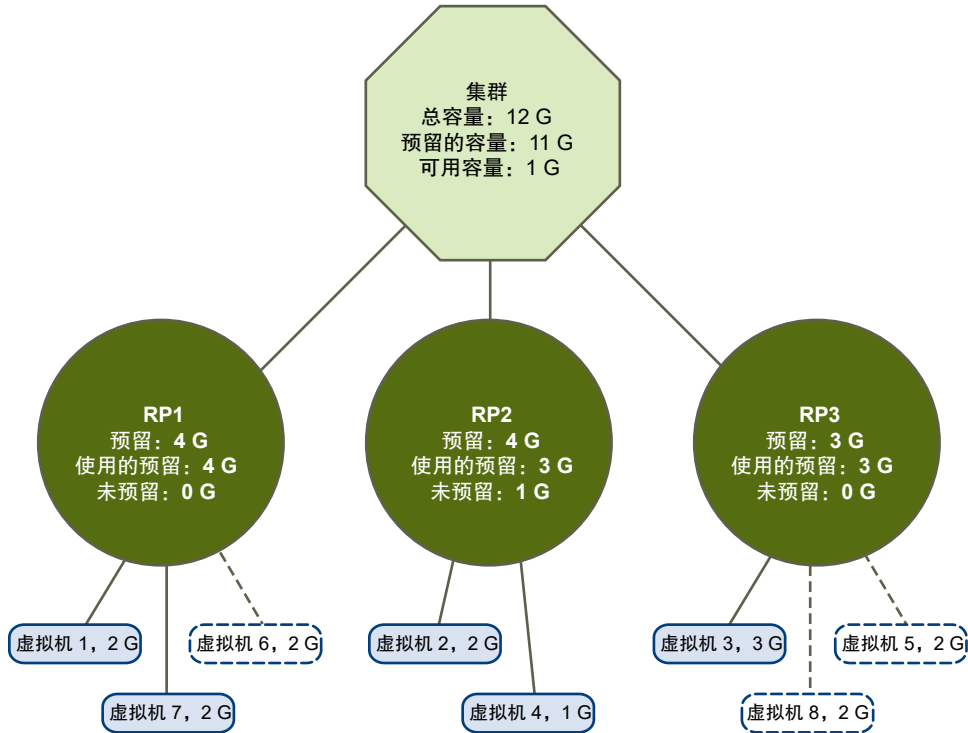
- 不可扩展的资源池：预留减去已使用的预留。
- 可扩展的资源池：（预留减去已使用的预留）加上任何可从祖先资源池借来的未预留资源。

有效 DRS 集群

有效集群拥有足够资源来满足所有预留以及支持所有正在运行的虚拟机。

下图显示具有固定资源池的有效集群的示例以及如何计算其 CPU 和内存资源。

图 18-1. 具有固定资源池的有效集群

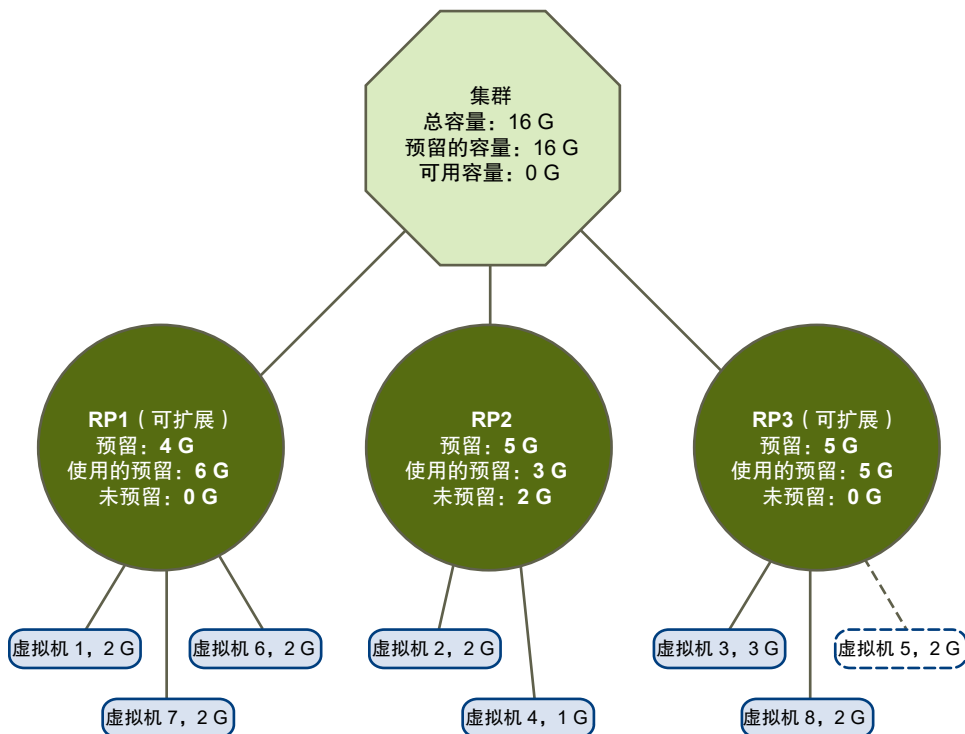


该集群具有以下特性：

- 总资源为 12 GHz 的集群。
- 三个类型均为**固定**（未选择**可扩展预留**）的资源池。
- 三个资源池合起来的总预留为 11 GHz (4+4+3 GHz)。总数显示在集群的**预留的容量**字段中。
- RP1 是使用 4 GHz 预留量创建的。两个虚拟机。打开了 VM1 和 VM7 的电源，分别各占 2 GHz（**使用的预留：4 GHz**）。未剩下资源用于打开额外的虚拟机的电源。VM6 显示为未打开电源。它不消耗任何预留。
- RP2 是使用 4 GHz 预留创建的。打开了两个虚拟机的电源，分别各占 1 GHz 和 2 GHz（**使用的预留：3 GHz**）。还剩 1 GHz 未预留。
- RP3 是使用 3 GHz 预留量创建的。打开了一个占用 3 GHz 的虚拟机的电源。没有资源可于打开额外的虚拟机的电源。

下图举例说明具有某些资源池（RP1 和 RP3）的有效集群，这些资源池的预留类型为**可扩展**。

图 18-2. 具有可扩展资源池的有效集群



可按如下方式配置有效集群:

- 总资源为 16 GHz 的集群。
- RP1 和 RP3 的类型为**可扩展**, RP2 的类型为“**固定**”。
- 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 显示为顶层集群的**预留的容量**。
- RP1 是使用 4 GHz 预留量创建的。打开了三个虚拟机的电源, 分别各占用 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) 可以使用集群资源池中的预留。(如果此资源池的类型为**固定**, 则无法打开额外的虚拟机的电源。)
- RP2 是使用 5 GHz 预留创建的。打开了两个虚拟机的电源, 分别各占 1 GHz 和 2 GHz (**使用的预留: 3 GHz**)。还剩 2 GHz 未预留。

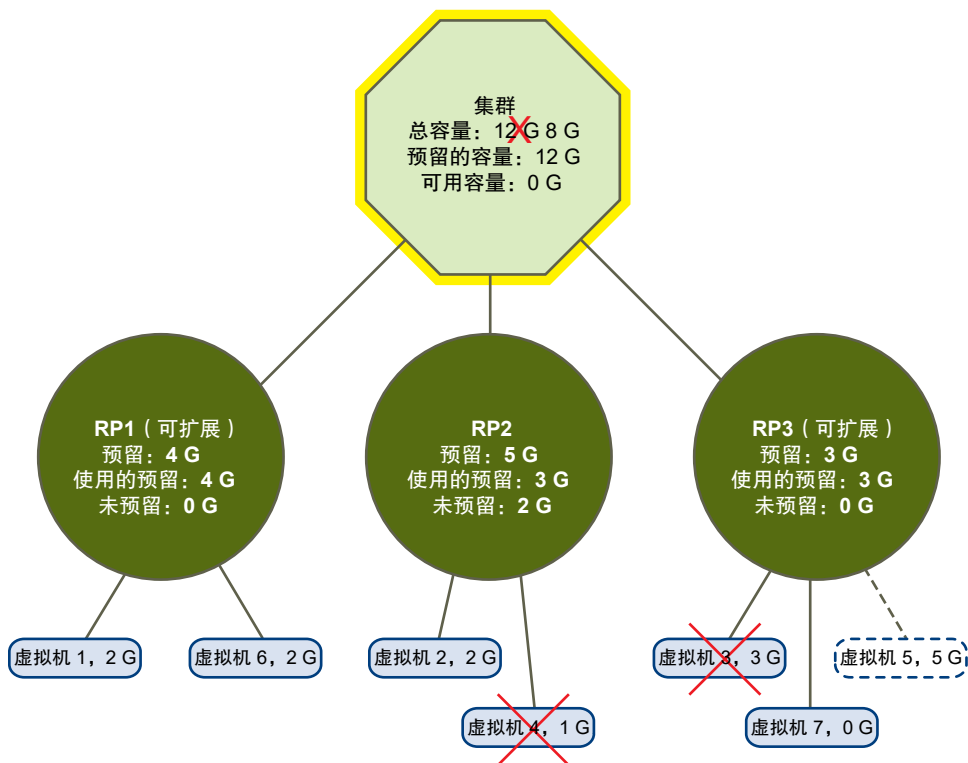
RP3 是使用 5 GHz 预留量创建的。打开了两个虚拟机电源, 分别各占 3 GHz 和 2 GHz。即使此资源池的类型为**可扩展**, 也无法打开额外的 2 GHz 虚拟机的电源, 因为父资源池的额外资源已被 RP1 占用。

过载的 DRS 集群

当资源池和虚拟机的树在内部是一致的, 但集群内没有足够容量来支持子资源池所预留的所有资源时, 集群将会变为过载 (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。当集群容量突然减少时 (例如, 集群内的一台主机不可用时), 集群通常会变为黄色。VMware 建议留足额外的集群资源, 以避免集群变为黄色。

图 18-3. 黄色集群



在此示例中：

- 总资源为 12GHz（分别来自三台各有 4GHz 资源的主机）的集群。
- 预留了总共 12GHz 资源的三个资源池。
- 三个资源池合起来所使用的总预留为 12GHz (4+5+3GHz)。该数值显示为集群内**预留的容量**。
- 由于其中一个 4GHz 主机不可用，因此总资源减少至 8GHz。
- 同时，故障主机上运行的 VM4 (1GHz) 和 VM3 (3GHz) 都不再运行。
- 该集群现在正在运行的虚拟机总共需要 6GHz 资源。该集群仍有 8GHz 的资源可用，足够满足虚拟机需求。

由于不再能达到 12GHz 的资源池预留，因此集群会被标记成黄色。

无效 DRS 集群

当树内部不再一致，即未遵守资源限制时，已启用 DRS 的集群会变为无效（红色）。

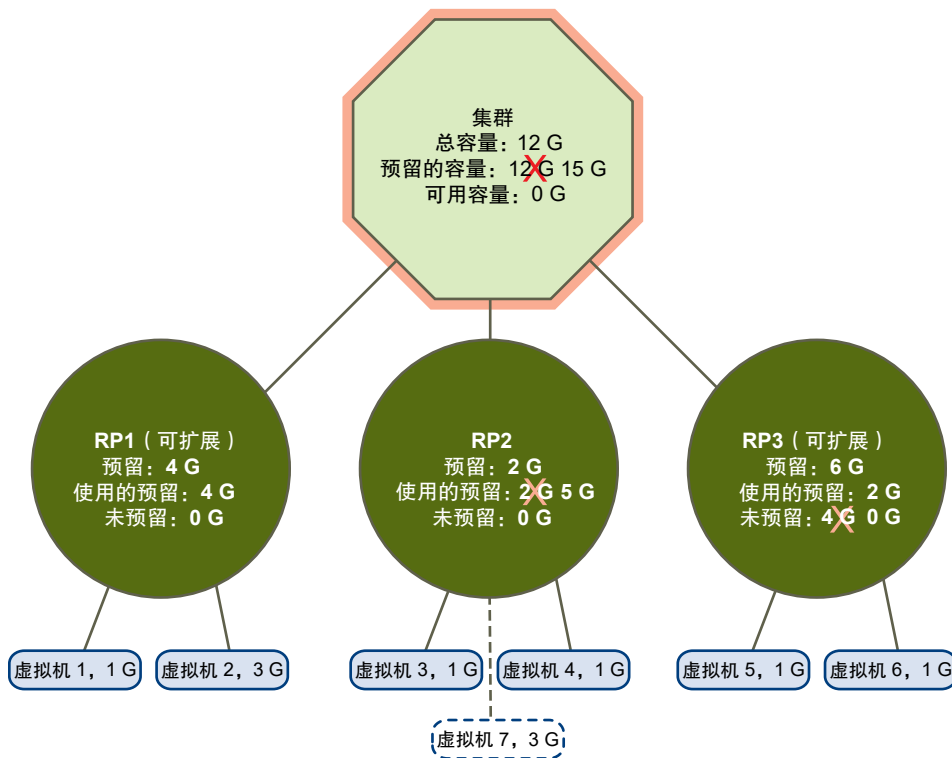
集群内资源的总数量与该集群是否为红色并无直接关联。如果在子级别中存在不一致，即使在根级别中存在足够的资源，集群也可能为红色。

可通过关闭一个或多个虚拟机的电源、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 集群问题。添加资源通常仅在处于黄色状况时才有用。

如果在虚拟机正在进行故障切换时重新配置资源池，则集群也可能会变为红色。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。在故障切换完成前，可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中。如果池的使用量大于新的预留，则该集群将变为红色。

如下图所示，如果用户能够（以不支持的方式）启动一个使用资源池 2 下 3 GHz 预留的虚拟机，则该集群会变为红色。

图 18-4. 红色集群



管理电源资源

通过 vSphere Distributed Power Management (DPM) 功能，DRS 集群可以根据集群资源利用率来打开和关闭主机电源，从而减少其功耗。

vSphere DPM 监控集群中所有虚拟机对内存和 CPU 资源的累积需求，并将其与集群中所有主机的可用资源总量进行比较。如果找到足够的额外容量，则 vSphere DPM 会将一台或多台主机置于待机模式，并将其虚拟机迁移到其他主机，然后关闭其电源。相反，当认为容量不够时，DRS 会使这些主机退出待机模式（打开它们的电源），并使用 vMotion 将虚拟机迁移到这些主机上。当进行这些计算时，vSphere DPM 不仅考虑当前需求，而且还会考虑用户指定的所有虚拟机资源预留。

如果您在创建 DRS 集群时启用**预测衡量指标**，DPM 将根据您选择的滚动预测窗口提前给出建议。

注 ESXi 主机不能自动退出待机模式，除非它们在 vCenter Server 管理的集群中运行。

vSphere DPM 可以使用三个电源管理协议之一使主机退出待机模式：智能平台管理界面 (IPMI)、Hewlett-Packard Integrated Lights-Out (iLO) 或 LAN 唤醒 (WOL)。每个协议均需要其各自的硬件支持和配置。如果主机不支持以上任何协议，则无法通过 vSphere DPM 将其置于待机模式。如果主机支持多个协议，则按以下顺序使用它们：IPMI、iLO、WOL。

注 不要在待机模式中断开主机，也不要未打开电源的情况下将其从 DRS 集群中移出，否则 vCenter Server 将无法再次打开该主机的电源。

为 vSphere DPM 配置 IPMI 或 iLO 设置

IPMI 是硬件级别规范，而 Hewlett-Packard iLO 是嵌入式服务器管理技术。它们均介绍并提供用于远程监控和控制计算机的接口。

必须在每台主机上执行以下过程。

前提条件

IPMI 和 iLO 需要硬件底板管理控制器 (BMC) 提供用于访问硬件控制功能的网关，并允许使用串行或 LAN 连接从远程系统访问该接口。即使主机自身已关闭电源，BMC 仍是打开电源的。如果已正确启用，则 BMC 可响应远程打开电源命令。

如果计划将 IPMI 或 iLO 用作唤醒协议，则必须配置 BMC。BMC 配置步骤根据型号而异。有关详细信息，请参见供应商的文档。使用 IPMI，还必须确保 BMC LAN 通道已配置为始终可用且允许操作员特权命令。在某些 IPMI 系统上，当启用“LAN 上的 IPMI”时，必须在 BIOS 中对其进行配置并指定特定的 IPMI 帐户。

仅使用 IPMI 的 vSphere DPM 支持基于 MD5 和纯文本的身份验证，但不支持基于 MD2 的身份验证。如果主机的 BMC 报告操作员角色支持并启用了 MD2 的身份验证，则 vCenter Server 使用 MD5。否则，如果 BMC 报告支持和启用了基于纯文本的身份验证，则使用基于纯文本的身份验证。如果既未启用 MD5 身份验证，也未启用纯文本身份验证，则 IPMI 无法与主机配合使用，并且 vCenter Server 将尝试使用 LAN 唤醒。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**选项卡。
- 3 在**系统**下，单击**电源管理**。
- 4 单击**编辑**。
- 5 输入以下信息。
 - BMC 帐户的用户名和密码。（该用户名必须能够远程打开主机电源。）
 - 与 BMC 关联的网卡的 IP 地址，不同于主机的 IP 地址。该 IP 地址应是具有无限租期的静态或 DHCP 地址。
 - 与 BMC 关联的网卡的 MAC 地址。
- 6 单击**确定**。

测试 vSphere DPM 的 LAN 唤醒

如果根据 VMware 准则配置用于 vSphere DPM 功能的 LAN 唤醒并成功对其进行测试，系统将完全支持对 WOL 的使用。为集群首次激活 vSphere DPM 之前，或在要添加到正在使用 vSphere DPM 的集群的任何主机上，必须执行这些步骤。

前提条件

在测试 WOL 之前，请确保集群满足先决条件。

- 集群必须至少包含两个 ESX 3.5（或 ESX 3i 版本 3.5）或更高版本的主机。
- 每台主机的 vMotion 网络链路必须工作正常。vMotion 网络还应当是单个 IP 子网，而不是由路由器分隔的多个子网。
- 每台主机上的 vMotion 网卡都必须支持 WOL。要检查 WOL 支持，请首先通过在 vSphere Client 的“清单”面板中选择主机，再选择**配置**选项卡，然后单击**网络**，以确定对应于 VMkernel 端口的物理网络适配器的名称。获取此信息后，单击**网络适配器**，并查找对应于网络适配器的条目。相关适配器应在**支持 LAN 唤醒**列中显示“是”。
- 要显示主机上每个网卡的 WOL 兼容状态，请在 vSphere Client 的“清单”面板中选择主机，再选择**配置**选项卡，然后单击**网络适配器**。网卡应在**支持 LAN 唤醒**列中显示“是”。
- 每个支持 WOL 的 vMotion 网卡所插入到的交换机端口应设置为自动协商链路速度，而不是设置为固定速度（例如，1000 Mb/s）。当主机关闭电源时，许多网卡仅在可切换到 100 Mb/s 或更慢速度时，才支持 WOL。

在验证这些必备条件之后，测试每个将使用 WOL 支持 vSphere DPM 的 ESXi 主机。测试这些主机时，请确保为集群停用 vSphere DPM 功能。

小心 确保对添加到使用 WOL 作为唤醒协议的 vSphere DPM 集群的任何主机进行测试，如果测试失败，则禁止其使用电源管理。如果未完成此操作，则 vSphere DPM 可能会关闭随后无法再次打开电源的主机的电源。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 右键单击主机，然后选择**电源 > 进入待机模式**
此操作将关闭主机。
- 3 右键单击主机，然后选择**电源 > 打开电源**，尝试使其退出待机模式。
- 4 观察主机是否再次成功打开电源。
- 5 对于未能成功退出待机模式的主机，请执行以下步骤：
 - a 在 vSphere Client 中选择主机，然后选择**配置**选项卡。
 - b 在**硬件 > 电源管理**下，单击**编辑**，调整电源管理策略。执行此操作后，vSphere DPM 不会考虑将该主机作为要关闭电源的候选主机。

为 DRS 集群激活 vSphere DPM

在执行了每台主机上正使用的唤醒协议所需的配置或测试步骤后，可以启用 vSphere DPM。

请配置电源管理自动化级别、阈值和主机级替代项。这些设置在集群的“设置”对话框中的**电源管理**下进行配置。

您也可以使用**调度任务: 更改集群电源设置**向导创建调度任务来为集群启用和停用 DPM。

注 如果 DRS 集群中的主机已连接 USB 设备，请对该主机停用 DPM。否则，DPM 可能会关闭主机，并断开设备与正在使用该设备的虚拟机之间的连接。

自动化级别

是否自动运行由 vSphere DPM 生成的主机电源状况和迁移建议取决于为该功能选择的电源管理自动化级别。

自动化级别在集群的“设置”对话框中的**电源管理**下进行配置。

注 电源管理自动化级别与 DRS 自动化级别不同。

表 18-1. 电源管理自动化级别

选项	描述
关闭	停用该功能且不提供建议。
手动	提供主机电源操作和相关虚拟机迁移建议，但不自动运行。
自动	如果可以自动运行相关虚拟机迁移，则将自动运行主机电源操作。

vSphere DPM 阈值

由 vSphere DPM 功能生成的电源状况（主机打开电源或关闭电源）建议按优先级进行分配，建议范围为 - 从优先级 1 到优先级 5。

这些优先级分类的基础为：DRS 集群内资源的过度利用率或利用率不足，以及预期对主机电源状况的改善。。优先级 1 的建议是强制性的，而优先级 5 的建议仅带来轻微改善。

该阈值在集群的“设置”对话框中的**电源管理**下进行配置。每次将 vSphere DPM 阈值滑块向右移动一个级别后，都会使自动执行的一组建议的优先级或显示为要手动执行的建议的优先级下降一个级别。在“保守”设置中，vSphere DPM 仅生成优先级 1 的建议，向右的下一级别则生成优先级 2 以及更高级别的建议，然后依次类推，直至“激进”级别，该级别生成优先级 5 的建议以及更高级别的建议（即，生成所有建议）。

注 DRS 阈值和 vSphere DPM 阈值本质上是相互独立的。您可以区分它们分别提供的迁移和主机电源状况建议的激进程度。

主机级替代项

在 DRS 集群内激活 vSphere DPM 时，默认情况下，集群内的所有主机都将继承其 vSphere DPM 自动化级别。

通过选择集群的“设置”对话框的“主机选项”页面并单击其**电源管理**设置，可以替代单个主机的此默认值。可以将此设置更改为以下选项：

- 已禁用
- 手动
- 自动

注 如果由于退出待机模式测试失败而将主机的“电源管理”设置为“已禁用”，请不要更改其设置。

在激活和运行 vSphere DPM 之后，通过查看每台主机的**上次退出待机模式的时间**信息（在集群“设置”对话框中的“主机选项”页面和每个集群的**主机选项卡**上显示），可以验证其是否能够正常工作。此字段会显示一个时间戳，以及 vCenter Server 上次尝试使主机退出待机模式的结果是“成功”还是“失败”。如果未曾进行此类尝试，则字段将显示“从不”。

注 **上次退出待机模式的时间**文本框的时间派生自 vCenter Server 事件日志。如果清除此日志，则时间将重置为“从不”。

监控 vSphere DPM

可以在 vCenter Server 中使用基于事件的警报来监控 vSphere DPM。

使用 vSphere DPM 时可能出现的最严重的错误是：主机在 DRS 集群需要其容量时无法退出待机模式。可以使用在 vCenter Server 中预先配置的**退出待机错误**警报来监控出现此错误时的情况。如果 vSphere DPM 无法使主机退出待机模式（vCenter Server 事件 DrsExitStandbyModeFailedEvent），可以将此警报配置为向管理员发送警示电子邮件或者使用 SNMP 陷阱发送通知。默认情况下，在 vCenter Server 能够成功连接到该主机之后，将清除此警报。

要监控 vSphere DPM 活动，还可以为以下 vCenter Server 事件创建警报。

表 18-2. vCenter Server 事件

事件类型	事件名称
正在进入待机模式（即将关闭主机电源）	DrsEnteringStandbyModeEvent
已成功进入待机模式（主机已成功关闭电源）	DrsEnteredStandbyModeEvent
正在退出待机模式（即将打开主机电源）	DrsExitingStandbyModeEvent
已成功退出待机模式（已成功打开电源）	DrsExitedStandbyModeEvent

有关创建和编辑警报的详细信息，请参见《vSphere 监控和性能》文档。

如果是使用监控软件而不是 vCenter Server，并且该软件会在意外关闭物理主机的电源时触发警报，那么，当 vSphere DPM 使主机进入待机模式时可能会出现生成“无效”警报的情况。如果不希望接收这些警报，请配合供应商部署一个与 vCenter Server 集成的监控软件版本。还可以使用 vCenter Server 本身作为监控解决方案，因为从 vSphere 4.x 开始，它本身能够识别 vSphere DPM 且不会触发这些无效警报。

使用 DRS 关联性规则

您可以使用关联性规则，控制集群内主机上的虚拟机的放置位置。

可以创建两种类型的规则。

- 用于指定虚拟机组和主机组之间的关联性或反关联性。关联性规则规定，所选虚拟机 DRS 组的成员可以或必须在特定的主机 DRS 组成员上运行。反关联性规则规定，所选虚拟机 DRS 组的成员不能在特定的主机 DRS 组成员上运行。

有关创建和使用此类型规则的信息，请参见[虚拟机-主机关联性规则](#)。

- 用于指定各个虚拟机之间的关联性或反关联性。指定关联性的规则会使 DRS 尝试将指定的虚拟机一起保留在同一台主机上（例如，出于性能考虑）。根据反关联性规则，DRS 尝试将指定的虚拟机分开，例如，当一台主机出现问题时，将不会同时丢失两台虚拟机。

有关创建和使用此类型规则的信息，请参见[虚拟机-虚拟机关联性规则](#)。

添加或编辑关联性规则时，如果集群的当前状态违反规则，系统将运行并尝试更正冲突。对于处于手动和半自动模式的 DRS 集群，将以规则实现和负载平衡为依据给出迁移建议，以待审批。您不一定要遵循这些规则，但在规则实现之前，相应的建议将一直保留。

要检查是否违反了任何已启用的关联性规则，且是否无法由 DRS 进行更正时，可以选择集群的 **DRS** 选项卡，然后单击**故障**。如果违反了某规则，则在此页面中将会显示与之相对应的错误。请阅读该错误以确定为什么 DRS 不能满足特定规则。规则违反也会生成日志事件。

注 虚拟机-虚拟机关联性规则与虚拟机-主机关联性规则与各个主机的 CPU 关联性规则不同。

创建主机 DRS 组

虚拟机与主机间的关联性规则将建立虚拟机 DRS 组与主机 DRS 组之间的关联性（或反关联性）关系。必须创建全部这两个组才能创建链接它们的规则。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**选项卡。
- 3 在**配置**下，选择**虚拟机/主机组**，然后单击**添加**。
- 4 在**创建虚拟机/主机组**对话框中，键入组的名称。
- 5 从**类型**下拉框中选择**主机组**，然后单击**添加**。
- 6 单击主机旁边的复选框以添加该主机。继续此过程，直到已添加所有需要的主机。

7 单击**确定**。

后续步骤

通过使用此主机 DRS 组，可以创建虚拟机与主机间的关联性规则，从而与适当的虚拟机 DRS 组建立关联性（或反关联性）关系。

[创建虚拟机 DRS 组](#)

[创建虚拟机-主机关联性规则](#)

创建虚拟机 DRS 组

关联性规则建立 DRS 组之间的关联性（或反关联性）关系。必须先创建 DRS 组，然后才能创建链接它们的规则。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**选项卡。
- 3 在**配置**下，选择**虚拟机/主机组**，然后单击**添加**。
- 4 在**创建虚拟机/主机组**对话框中，键入组的名称。
- 5 从**类型**下拉框中选择**虚拟机组**，然后单击**添加**。
- 6 单击虚拟机旁边的复选框以添加该虚拟机。继续此过程，直到已添加所有需要的虚拟机。
- 7 单击**确定**。

后续步骤

[创建主机 DRS 组](#)

[创建虚拟机-主机关联性规则](#)

[创建虚拟机-虚拟机关联性规则](#)

虚拟机-虚拟机关联性规则

虚拟机-虚拟机关联性规则指定选定的单个虚拟机是应在同一主机上运行还是应保留在其他主机上。此类型规则用于创建所选单个虚拟机之间的关联性或反关联性。

当创建关联性规则时，DRS 会尝试将指定的虚拟机都保留在同一主机上。例如，可能需要出于性能原因而这样做。

使用反关联性规则时，DRS 会尝试将指定的虚拟机分开。如果为了保证某些虚拟机始终在不同物理主机上，则可以使用此类规则。在该情况下，如果一个主机出现问题，不会将所有虚拟机都置于风险中。

创建虚拟机-虚拟机关联性规则

您可以创建虚拟机-虚拟机关联性规则，以指定选定的个别虚拟机是应在同一主机上运行还是应在不同主机上运行。

注 如果使用“vSphere HA 指定故障切换主机”准入控制策略，并指定多个故障切换主机，则不支持虚拟机-虚拟机关联性规则。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**选项卡。
- 3 在**配置**下，单击**虚拟机/主机规则**。
- 4 单击**添加**。
- 5 在**创建虚拟机/主机规则**对话框中，键入规则的名称。
- 6 从**类型**下拉菜单中，选择**聚集虚拟机**或**分开虚拟机**。
- 7 单击**添加**。
- 8 至少选择两个要应用该规则的虚拟机，然后单击**确定**。
- 9 单击**确定**。

虚拟机-虚拟机关联性规则冲突

您可以创建并使用多个虚拟机-虚拟机关联性规则，但是，这可能会导致规则相互冲突的情况发生。

如果两条虚拟机-虚拟机关联性规则存在冲突，则无法同时激活这两条规则。例如，如果一条规则要求两个虚拟机始终在一起，而另一条规则要求这两个虚拟机始终分开，则无法同时激活这两条规则。选择应用其中一条规则，并停用或删除与之冲突的规则。

当两条虚拟机-虚拟机关联性规则发生冲突时，旧规则优先，新规则将停用。DRS 仅尝试满足激活的规则，而忽略停用的规则。与关联性规则的冲突相比，DRS 将优先阻止反关联性规则的冲突。

虚拟机-主机关联性规则

虚拟机-主机关联性规则指定选定的虚拟机 DRS 组的成员是否可在特定主机 DRS 组的成员上运行。

与指定各个虚拟机之间的关联性（或反关联性）的虚拟机-虚拟机关联性规则不同，虚拟机-主机关联性规则会指定一组虚拟机与一组主机之间的关联性关系。存在“必要”规则（由“必须”指定）和“首选”规则（由“应该”指定）。

虚拟机-主机关联性规则包括以下组件。

- 一个虚拟机 DRS 组。
- 一个主机 DRS 组。
- 指定规则是必要（“必须”）还是首选项（“应该”），以及规则是关联性（“运行于”）还是反关联性（“不运行于”）。

由于虚拟机-主机关联性规则是基于集群的，因此规则中包含的虚拟机和主机必须全部位于同一集群中。如果从集群移除虚拟机，则虚拟机会丢失其 DRS 组关联性（即使稍后恢复到集群也是如此）。

创建虚拟机-主机关联性规则

您可以创建虚拟机-主机关联性规则，以指定选定的虚拟机 DRS 组的成员是否可在特定主机 DRS 组的成员上运行。

前提条件

创建应用虚拟机-主机关联性规则的虚拟机和主机 DRS 组。

步骤

- 1 在 vSphere Client 中，浏览到集群。
- 2 单击**配置**选项卡。
- 3 在**配置**下，单击**虚拟机/主机规则**。
- 4 单击**添加**。
- 5 在**创建虚拟机/主机规则**对话框中，键入规则的名称。
- 6 从**类型**下拉菜单中，选择**虚拟机到主机**。
- 7 选择该规则所应用到的虚拟机 DRS 组和主机 DRS 组。
- 8 为该规则选择规范。
 - **必须在组中的主机上运行。**虚拟机组 1 中的虚拟机必须在主机组 A 中的主机上运行。
 - **应在组中的主机上运行。**虚拟机组 1 中的虚拟机应当（但不是必须）在主机组 A 中的主机上运行。
 - **不得在组中的主机上运行。**虚拟机组 1 中的虚拟机绝对不能在主机组 A 中的主机上运行。
 - **不应在组中的主机上运行。**虚拟机组 1 中的虚拟机不应当（但可以）在主机组 A 的主机上运行。
- 9 单击**确定**。

使用虚拟机-主机关联性规则

您可以使用虚拟机-主机关联性规则，以指定虚拟机组和主机组之间的关联性关系。使用虚拟机-主机关联性规则时，您应该了解这些规则何时最有用，如何解决规则之间的冲突以及小心设置所需关联性规则的重要性。

如果创建多个虚拟机-主机关联性规则，这些规则不会进行排序，将平等应用。请注意，这会对规则的交互方式有影响。例如，属于两个 DRS 组（每个组都属于不同的必要规则）的虚拟机只能在同时属于这两个主机 DRS 组（如规则中所述）的主机上运行。

创建虚拟机-主机关联性规则时，不会检查该规则是否能在与其他规则相关的情况下运行。因此，您创建的规则可能会与正在使用的其他规则相冲突。当两条虚拟机-主机关联性规则发生冲突时，旧规则优先，新规则将停用。DRS 仅尝试满足激活的规则，而忽略停用的规则。

DRS、vSphere HA 和 vSphere DPM 不会采取任何会导致违反必要关联性规则（虚拟机 DRS 组“必须运行于”或“不得运行于”主机 DRS 组上）的操作。相应地，您应该小心使用此类型的规则，因为可能会对集群运行造成负面影响。如果未正确使用，必要虚拟机-主机关联性规则可能会将集群分为多个段，并阻碍 DRS、vSphere HA 和 vSphere DPM 正确运行。

如果这样做会违反必要关联性规则，则不会执行许多集群功能。

- DRS 不会撤出虚拟机，以将主机置于维护模式。
- DRS 不会将虚拟机置于打开电源状态，也不会对虚拟机进行负载均衡。
- vSphere HA 不会执行故障切换。
- vSphere DPM 不会通过将主机置于待机模式来优化电源管理。

要避免这些情况，在创建多个必要关联性规则时或考虑仅使用首选的虚拟机-主机关联性规则（虚拟机 DRS 组“应运行于”或“不应运行于”主机 DRS 组上）时，请倍加小心。请确保与每个虚拟机关联的集群中的主机数目足够大，即使丢失一个主机也不会导致缺少可运行虚拟机的主机。可以违反首选规则，以使 DRS、vSphere HA 和 vSphere DPM 可以正确运行。

注 您可以创建基于事件的警报，当虚拟机违反虚拟机-主机关联性规则时触发该警报。为虚拟机添加新警报，并选择**虚拟机正在违反虚拟机-主机关联性规则**作为事件触发器。有关创建和编辑警报的详细信息，请参见《vSphere 监控和性能》文档。

创建数据存储集群

19

数据存储集群是具有共享资源和共享管理接口的数据存储的集合。数据存储集群之于数据存储，如同集群之于主机。创建数据存储集群时，可以使用 vSphere Storage DRS 管理存储资源。

注 在 vSphere API 中，数据存储集群称为存储单元。

将数据存储添加到数据存储集群时，数据存储的资源将成为数据存储集群资源的一部分。和主机集群一样，您可以使用数据存储集群聚合存储资源，以便在数据存储集群级别上支持资源分配策略。还提供了以下数据存储集群级别的资源管理功能。

空间使用负载均衡

可以设置空间利用率阈值。当数据存储中的空间利用率超出阈值时，存储 DRS 会生成建议或执行 Storage vMotion 迁移来在整个数据存储集群中均衡空间使用情况。

I/O 滞后时间负载均衡

为避免出现瓶颈，可以设置 I/O 滞后时间阈值。当数据存储中的 I/O 滞后时间超出阈值时，存储 DRS 会生成建议或执行 Storage vMotion 迁移来帮助缓解高 I/O 负载。

反关联性规则

可以为虚拟机磁盘创建反关联性规则。例如，某个虚拟机的虚拟磁盘必须保存在不同的数据存储中。默认情况下，一个虚拟机的所有虚拟磁盘都放在同一数据存储中。

本章讨论了以下主题：

- 初始放置位置和后续平衡
- 存储迁移建议
- 创建数据存储集群
- 激活和停用 Storage DRS
- 为数据存储集群设置自动化级别
- 设置 Storage DRS 的激进级别
- Datastore Cluster 要求
- 在数据存储集群中添加和移除数据存储

初始放置位置和后续平衡

存储 DRS 为启用了存储 DRS 的数据存储集群中的数据提供初始放置位置建议和后续平衡建议。

当存储 DRS 选择数据存储集群内的数据存储以在其上放置虚拟机磁盘时，会发生初始放置位置。在以下情形下会发生此情况：创建或克隆虚拟机时、将虚拟机磁盘迁移至其他数据存储集群时或将磁盘添加到现有虚拟机时。

将根据空间限制并相对于空间目标和 I/O 负载平衡生成初始放置位置建议。这些目标旨在将超置备一个数据存储的风险、存储 I/O 瓶颈和对虚拟机的性能影响降至最低。

将以配置的频率（默认情况下，为每八小时）调用存储 DRS，或当数据存储集群中的一个或多个数据存储超出用户可配置的空间利用率阈值时，调用存储 DRS。调用存储 DRS 后，它将根据阈值检查每个数据存储的空间利用率和 I/O 滞后时间值。对于 I/O 滞后时间，存储 DRS 使用一天当中测量的 I/O 滞后时间的 90 个百分点来与阈值进行比较。

存储迁移建议

vCenter Server 在具有手动自动模式的数据存储集群的“存储 DRS 建议”页面上显示迁移建议。

系统将提供足够的建议，以强制实施存储 DRS 规则并平衡数据存储集群的空间和 I/O 资源。每条建议均包含虚拟机名称、虚拟磁盘名称、数据存储集群名称、源数据存储、目标数据存储以及提出建议的原因。

- 平衡数据存储空间使用情况
- 平衡数据存储 I/O 负载

存储 DRS 在以下情况下提供强制性迁移建议：

- 数据存储空间不足。
- 违反了反关联性或关联性规则。
- 数据存储正进入维护模式且必须撤出。

此外，当数据存储空间接近不足或者应对空间和 I/O 负载平衡进行调整时，还会提供可选建议。

存储 DRS 应考虑移动已关闭电源或打开电源的虚拟机来平衡空间。在这些注意事项中，存储 DRS 考虑到了已关闭电源的具有快照的虚拟机。

创建数据存储集群

可使用 Storage DRS 管理数据存储集群资源。

步骤

- 1 在 vSphere Client 中，浏览到数据中心。
- 2 右键单击数据中心对象并选择**新建数据存储集群**。
- 3 要完成**新建数据存储集群**向导，请按照提示进行操作。
- 4 单击**完成**。

激活和停用 Storage DRS

通过 Storage DRS 可以管理数据存储集群的聚合资源。激活 Storage DRS 后，它会对虚拟机磁盘放置和迁移提供建议，以均衡数据存储集群内所有数据存储之间的空间和 I/O 资源。

激活 Storage DRS 时，将激活以下功能。

- 数据存储集群中数据存储之间的空间负载均衡。
- 数据存储集群中数据存储之间的 I/O 负载均衡。
- 最初基于空间和 I/O 工作负载放置虚拟磁盘。

“数据存储集群设置”对话框中的“启用 Storage DRS”复选框用于一次激活或停用所有这些组件。如果需要，您可以独立于空间均衡功能，停用 Storage DRS 的 I/O 相关功能。

在数据存储集群上停用 Storage DRS 时，会保留 Storage DRS 设置。激活 Storage DRS 时，会将数据存储集群的设置还原至 Storage DRS 被停用的位置。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 选择 **Storage DRS**，然后单击**编辑**。
- 4 选择**打开 vSphere DRS**，然后单击**确定**。
- 5 （可选）若只停用 Storage DRS 的 I/O 相关功能而让空间相关控件保持激活状态，请执行以下步骤。
 - a 在 **Storage DRS** 下，选择**编辑**。
 - b 取消选中**为 Storage DRS 启用 I/O 衡量指标**复选框，然后单击**确定**。

为数据存储集群设置自动化级别

数据存储集群的自动化级别指定是否自动应用来自 Storage DRS 的放置建议和迁移建议。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 选择 **DRS**，然后单击**编辑**。

4 展开“DRS 自动化”，然后选择自动化级别。

手动是默认的自动化级别。

选项	描述
非自动 (手动模式)	将显示放置和迁移建议，但在手动应用建议之前，不会运行这些建议。
半自动	将自动运行放置建议，并显示迁移建议，但在手动应用迁移建议之前，不会运行这些建议。
全自动	放置和迁移建议会自动运行。

5 单击确定。

设置 Storage DRS 的激进级别

通过指定已用空间和 I/O 延迟时间的阈值来确定 Storage DRS 的激进程度。

Storage DRS 收集数据存储集群中数据存储的资源使用情况信息。vCenter Server 使用此信息生成虚拟磁盘在数据存储上的放置位置的建议。

为数据存储集群设置低激进级别时，Storage DRS 建议仅在绝对有必要时（例如，当 I/O 负载、空间利用率或其不平衡高时）进行 Storage vMotion 迁移。为数据存储集群设置高激进级别时，Storage DRS 建议只要数据存储集群可从空间或 I/O 负载平衡中受益便进行迁移。

在 vSphere Client 中，可以使用以下阈值设置 Storage DRS 的激进级别：

空间利用率

当数据存储上的空间利用率超过在 vSphere Client 中设置的阈值时，Storage DRS 将生成建议或执行迁移。

I/O 延迟时间

当一天中为数据存储测量的 I/O 延迟时间的第 90 个百分点超过阈值时，Storage DRS 将生成建议或执行迁移。

还可以设置高级选项以进一步配置 Storage DRS 的激进级别。

空间利用率差异

该阈值确保源的空间利用率与目标的空间利用率之间存在一些最小差异。例如，如果数据存储 A 上的空间利用率为 82%，数据存储 B 为 79%，差异为 3。如果阈值为 5，则 Storage DRS 不会建议从数据存储 A 迁移到数据存储 B。

I/O 负载平衡调用时间间隔

此时间间隔后，Storage DRS 将运行以平衡 I/O 负载。

I/O 不平衡阈值

降低该值可减少 I/O 负载平衡的激进程度。Storage DRS 计算 0 与 1 之间的 I/O 公平性衡量指标，其中 1 是最公平的分发。仅当计算的衡量指标小于 1 时，I/O 负载平衡才运行 - (I/O 不平衡阈值 / 100)。

设置 Storage DRS 运行时规则

设置 Storage DRS 触发器，并为数据存储集群配置高级选项。

步骤

- 1 (可选) 选中或取消选中为 **SDRS 建议启用 I/O 衡量指标** 复选框，以激活或停用包含 I/O 衡量指标。

如果停用该选项，则在提出 Storage DRS 建议时，vCenter Server 不会考虑 I/O 衡量指标。停用此选项时，将停用 Storage DRS 的以下元素：

- 数据存储集群中数据存储之间的 I/O 负载均衡。
- 基于 I/O 工作负载的虚拟磁盘的初始放置位置。初始放置位置仅基于空间。

- 2 (可选) 设置 Storage DRS 阈值。

通过指定已用空间和 I/O 延迟时间的阈值，来设置 Storage DRS 的激进级别。

- 使用“已利用空间”滑块指示存储 DRS 触发前允许的已占用空间的最大百分比。数据存储上的空间利用率超出阈值时，Storage DRS 将提出建议并执行迁移。
- 使用“I/O 滞后时间”滑块指示存储 DRS 触发前允许的最大 I/O 滞后时间。延迟时间超出阈值时，Storage DRS 将提出建议并执行迁移。

注 数据存储集群的 Storage DRS I/O 延迟时间阈值应低于或等于 Storage I/O Control 拥堵阈值。

- 3 (可选) 配置高级选项。

- 无建议，直到源与目标之间的利用率差异为：使用滑块指定空间利用率差异阈值。利用率等于使用情况 * 100/容量。

该阈值确保源的空间利用率与目标的空间利用率之间存在一些最小差异。例如，如果数据存储 A 上的空间利用率为 82%，数据存储 B 为 79%，差异为 3。如果阈值为 5，则 Storage DRS 不会建议从数据存储 A 迁移到数据存储 B。

- 检查失衡情况时间间隔：指定 Storage DRS 评估空间和 I/O 负载均衡的频率。
- I/O 不均衡阈值：使用滑块指示 I/O 负载均衡的激进程度。降低该值可减少 I/O 负载均衡的激进程度。Storage DRS 计算 0 与 1 之间的 I/O 公平性衡量指标，其中 1 是最公平的分发。仅当计算的衡量指标小于 1 时，I/O 负载均衡才运行 - (I/O 不均衡阈值 / 100)。

- 4 单击**确定**。

Datastore Cluster 要求

数据存储以及与数据存储集群关联的主机必须符合特定要求，才能成功使用数据存储集群功能。

创建数据存储集群时，请遵循下列准则。

- 数据存储集群必须包含类似的或可互换的数据存储。
一个数据存储集群中可以混用不同大小和 I/O 能力的数据存储，还可以混用来自不同阵列和供应商的数据存储。但是，下列类型的数据存储不能共存于一个数据存储集群中。
 - 在同一个数据存储集群中，不能组合使用 NFS 和 VMFS 数据存储。
 - 在同一个启用了 Storage DRS 的数据存储集群中，不能结合使用复制的数据存储和非复制的数据存储。
- 连接到数据存储集群中的数据存储的所有主机必须是 ESXi 5.0 及更高版本。如果数据存储集群中的数据存储连接到 ESX/ESXi 4.x 及更早版本的主机，则 Storage DRS 不会运行。
- 数据存储集群中不能包含跨多个数据中心共享的数据存储。
- 最佳做法是，启用了硬件加速的数据存储不能与未启用硬件加速的数据存储放在同一个数据存储集群中。数据存储集群中的数据存储必须属于同类，才能保证实现硬件加速支持的行为。

在数据存储集群中添加和移除数据存储

可在现有数据存储集群中添加和移除数据存储。

在 vSphere Client 清单中，可以将主机上挂载的任何数据存储添加到数据存储集群，以下情况除外：

- 连接到数据存储的所有主机必须是 ESXi 5.0 及更高版本。
- 数据存储不能位于 vSphere Client 同一实例中的多个数据中心内。

从数据存储集群中移除数据存储后，该数据存储仍在 vSphere Client 清单中且未从主机上卸载。

使用数据存储集群管理存储资源

20

创建数据存储集群后，可以对其进行自定义，并使用它来管理存储 I/O 和空间利用率资源。

本章讨论了以下主题：

- 使用存储 DRS 维护模式
- 应用存储 DRS 建议
- 更改虚拟机的存储 DRS 自动化级别
- 设置 Storage DRS 的非工作时间调度
- Storage DRS 反关联性规则
- 清除 Storage DRS 统计信息
- Storage vMotion 与数据存储集群的兼容性

使用存储 DRS 维护模式

当您需要暂停使用数据存储以对其进行维护时，请将其置于维护模式。数据存储仅会因用户要求而进入或离开维护模式。

维护模式适用于启用了存储 DRS 的数据存储集群中的数据存储。独立数据存储不能置于维护模式。

必须手动或使用存储 DRS 将即将进入维护模式的数据存储中的虚拟磁盘迁移到其他数据存储。当您尝试将数据存储置于维护模式时，**放置位置建议**选项卡将显示迁移建议列表，以及同一数据存储集群中可以迁移虚拟磁盘的数据存储。在**故障**选项卡上，vCenter Server 将显示无法迁移的磁盘列表及其原因。如果存储 DRS 关联性或反关联性规则阻止迁移磁盘，则可以选择为“维护”选项启用“忽略关联性规则”。

虚拟磁盘全部迁移之前，数据存储会处于“正在进入维护模式”状态。

将数据存储置于维护模式

如果需要中止数据存储，可将该数据存储置于 Storage DRS 维护模式。

前提条件

包含要进入维护模式的数据存储的数据存储集群上已启用 Storage DRS。

没有 CD-ROM 映像文件存储在数据存储中。

数据存储集群中至少有两个数据存储。

步骤

- 1 在 vSphere Client 中，浏览到数据存储。
- 2 右键单击数据存储，然后选择**维护模式 > 进入维护模式**。

此时会显示数据存储维护模式迁移的建议列表。

- 3 （可选）在“放置位置建议”选项卡中，取消选择您不想应用的建议。

注 如果未清空所有磁盘，数据存储将无法进入维护模式。如果取消选中建议，则必须手动移动受影响的虚拟机。

- 4 如有必要，请单击**应用建议**。

vCenter Server 使用 Storage vMotion 将虚拟磁盘从源数据存储迁移到目标数据存储，然后数据存储进入维护模式。

结果

可能无法立即更新数据存储图标，以反映数据存储的当前状况。要立即更新图标，请单击**刷新**。

对于维护模式忽略 Storage DRS 关联性规则

Storage DRS 关联性或反关联性规则可能会阻止数据存储进入维护模式。当将数据存储置于维护模式时，可以忽略这些规则。

为数据存储集群激活“对维护忽略关联性规则”选项时，vCenter Server 将忽略会阻止数据存储进入维护模式的 Storage DRS 关联性和反关联性规则。

仅针对疏散建议忽略 Storage DRS 规则。vCenter Server 在生成空间和负载均衡建议或初始放置建议时不违反规则。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 选择 **DRS**，然后单击**编辑**。
- 4 展开**高级选项**，然后单击**添加**。
- 5 在“选项”列中，键入 **IgnoreAffinityRulesForMaintenance**。
- 6 在“值”列中，键入 **1** 可激活该选项。

键入 **0** 可停用该选项。

- 7 单击**确定**。

结果

针对“维护模式”选项的“忽略关联性规则”将应用于数据存储集群。

应用存储 DRS 建议

Storage DRS 收集数据存储集群中所有数据存储的资源使用情况信息。Storage DRS 使用此信息生成虚拟机磁盘在数据存储集群中的数据存储上的放置建议。

Storage DRS 建议显示在 vSphere Client 的“数据存储”视图中的 **Storage DRS** 选项卡上。建议也会在尝试将数据存储置于 Storage DRS 维护模式时显示。应用 Storage DRS 建议时，vCenter Server 会使用 Storage vMotion 将虚拟机磁盘迁移到数据存储集群中的其他数据存储，以平衡资源。

您可以通过选中“替代给出的 DRS 建议”复选框并选择要应用的每个建议来应用一部分建议。

表 20-1. 存储 DRS 建议

标签	描述
优先级	建议的优先级级别 (1-5)。(默认情况下隐藏。)
建议	Storage DRS 建议的操作。
原因	需要执行操作的原因。
(源) 和 (目标) 之前的空间利用率 %	迁移之前在源和目标数据存储上使用的空间百分比。
(源) 和 (目标) 之后的空间利用率 %	迁移之后在源和目标数据存储上使用的空间百分比。
(源) 之前的 I/O 滞后时间	迁移之前源数据存储上的 I/O 延迟时间值。
(目标) 之前的 I/O 滞后时间	迁移之前目标数据存储上的 I/O 延迟时间值。

刷新存储 DRS 建议

Storage DRS 迁移建议显示在 vSphere Client 中的 **Storage DRS** 选项卡上。可以通过运行 Storage DRS 刷新这些建议。

前提条件

vSphere Client 清单中必须至少存在一个数据存储集群。

为数据存储集群启用 Storage DRS。**Storage DRS** 选项卡仅在启用了 Storage DRS 时才会出现。

步骤

- 1 在 vSphere Client 的“数据存储”视图中，选择数据存储集群，然后单击 **Storage DRS** 选项卡。
- 2 选择 **建议** 视图，然后单击右上角的 **运行 Storage DRS** 链接。

结果

即会更新建议。“上次更新”时间戳将显示刷新 Storage DRS 建议的时间。

更改虚拟机的存储 DRS 自动化级别

可以替代各个虚拟机的数据存储集群范围的自动化级别。也可以替代默认虚拟磁盘关联性规则。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**配置**。
- 3 在**虚拟机替代**项下，选择**添加**。
- 4 选择虚拟机。
- 5 单击“自动化级别”下拉菜单，然后为虚拟机选择自动化级别。

选项	描述
默认（手动）	将显示放置和迁移建议，但在手动应用建议之前，不会运行这些建议。
全自动	放置和迁移建议会自动运行。
已禁用	vCenter Server 将不会迁移虚拟机或为其提供迁移建议。

- 6 单击**聚集 VMDK** 下拉菜单以替代默认 VMDK 关联性。

请参见替代 [VMDK 关联性规则](#)。

- 7 单击**确定**。

设置 Storage DRS 的非工作时间调度

可以创建一个已调度任务来更改数据存储集群的 Storage DRS 设置，从而使完全自动化的数据存储集群的迁移更可能发生在非高峰期间。

可以创建已调度任务来更改数据存储集群的自动化级别和激进级别。例如，可以在性能优先时将 Storage DRS 配置为在高峰期间降低运行的激进程度，以尽量避免发生存储迁移。在非高峰期间，Storage DRS 可以在更为激进的模式下运行，并且可以对其更频繁地调用。

前提条件

启用 Storage DRS。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**服务**。
- 3 在 **vSphere DRS** 下，单击**调度 DRS** 按钮。
- 4 在“编辑数据存储集群”对话框中，单击 **SDRS 调度**。

5 展开 DRS 自动化。

- a 选择自动化级别。
- b 设置迁移阈值。

使用“迁移”滑块选择 vCenter Server 建议的优先级别，以调整集群的负载平衡。

- c 选择是否启用“虚拟机自动化”。

可在“虚拟机替代项”页面中设置个别虚拟机的替代项。

6 展开电源管理。

- a 选择自动化级别。
- b 设置 DPM 阈值。

使用“DPM”滑块，选择 vCenter Server 将应用的电源建议。

7 键入任务名称。

8 为已经创建的任务键入描述。

9 在“配置的调度程序”下，单击**更改**，然后选择运行任务的时间，最后单击**确定**。

10 键入电子邮件地址，以便完成任务时将通知电子邮件发送到该地址。

11 单击**确定**。

结果

已调度任务会在特定的时间运行。

Storage DRS 反关联性规则

可以创建 Storage DRS 反关联性规则，以控制哪些虚拟磁盘不应置于数据存储集群中的同一数据存储上。默认情况下，虚拟机的虚拟磁盘聚集在同一数据存储上。

如果您创建反关联性规则，则该规则将适用于数据存储集群中的相关虚拟磁盘。在初始放置和 Storage DRS 建议迁移期间强制实施反关联性规则，但在用户启动迁移时不强制实施反关联性规则。

注 反关联性规则不适用于存储在数据存储集群中的数据存储上的 CD-ROM ISO 映像文件，也不适用于存储在用户定义的位置中的交换文件。

虚拟机反关联性规则

指定哪些虚拟机不应保存在相同的数据存储上。请参见[创建虚拟机反关联性规则](#)。

VMDK 反关联性规则

指定与特定虚拟机关联的哪些虚拟磁盘必须保存在不同的数据存储上。请参见[创建 VMDK 反关联性规则](#)。

如果您将某个虚拟磁盘移出数据存储集群，关联性规则或反关联性规则将不再适用于该磁盘。

当将虚拟磁盘文件移入具有现有关联性规则和反关联性规则的数据存储集群时，将具有以下行为：

- 数据存储集群 B 具有虚拟机内部关联性规则。当将虚拟磁盘移出数据存储集群 A 并移入数据存储集群 B 时，适用于数据存储集群 A 中给定虚拟机的虚拟磁盘的任何规则都将不再适用。虚拟磁盘现遵循数据存储集群 B 中的虚拟机内部关联性规则。
- 数据存储集群 B 具有虚拟机反关联性规则。当将虚拟磁盘移出数据存储集群 A 并移入数据存储集群 B 时，适用于数据存储集群 A 中给定虚拟机的虚拟磁盘的任何规则都将不再适用。虚拟磁盘现遵循数据存储集群 B 中的虚拟机反关联性规则。
- 数据存储集群 B 具有 VMDK 反关联性规则。当将虚拟磁盘移出数据存储集群 A 并移入数据存储集群 B 时，VMDK 反关联性规则不适用于给定虚拟机的虚拟磁盘，因为该规则仅限于数据存储集群 B 中的指定虚拟磁盘。

注 Storage DRS 规则可能会阻止数据存储进入维护模式。可以通过为“维护”选项启用“忽略关联性规则”，选择对维护模式忽略 Storage DRS 规则。

创建虚拟机反关联性规则

您可以创建反关联性规则，以指示某些虚拟机的所有虚拟磁盘都必须保留在不同的数据存储上。此规则将应用到各数据存储集群。

数据存储集群中应用虚拟机反关联性规则的虚拟机，都必须与此数据存储集群中的虚拟机内部关联性规则相关联。这些虚拟机也必须符合虚拟机内部关联性规则。

当虚拟机受虚拟机反关联性规则限制时，将具有以下行为：

- Storage DRS 将根据规则放置虚拟机的虚拟磁盘。
- 即使是强制进行迁移（如将数据存储置于维护模式），Storage DRS 也会根据规则使用 vMotion 迁移虚拟磁盘。
- 如果虚拟机的虚拟磁盘违反了规则，则 Storage DRS 将提出迁移建议来更正这一错误，或者在无法提出更正错误的建议时将此违反报告为故障。

默认情况下，未定义任何虚拟机反关联性规则。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**配置**。
- 3 选择**虚拟机/主机规则**。
- 4 单击**添加**。
- 5 键入规则的名称。
- 6 从“类型”菜单中，选择**虚拟机反关联性**。
- 7 单击**添加**。
- 8 单击**选择虚拟机**。

9 至少选择两台虚拟机，然后单击**确定**。

10 单击**确定**以保存该规则。

创建 VMDK 反关联性规则

您可以为虚拟机创建 VMDK 反关联性规则，以指示虚拟机的哪些虚拟磁盘必须保留在不同的数据存储上。

VMDK 反关联性规则适用于已定义此规则的虚拟机，并非适用于所有虚拟机。此规则是指一个要相互分离的虚拟磁盘的列表。

如果尝试为虚拟机同时设置 VMDK 反关联性规则和虚拟机内部关联性规则，则 vCenter Server 将拒绝最近定义的规则。

当虚拟机受 VMDK 反关联性规则限制时，将具有以下行为：

- Storage DRS 将根据规则放置虚拟机的虚拟磁盘。
- 即使是强制进行迁移（如将数据存储置于维护模式），Storage DRS 也会根据规则使用 vMotion 迁移虚拟磁盘。
- 如果虚拟机的虚拟磁盘违反了规则，则 Storage DRS 将提出迁移建议来更正这一错误，或者在无法提出更正错误的建议时将此违反报告为故障。

默认情况下，未定义任何 VMDK 反关联性规则。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**配置**。
- 3 选择**虚拟机/主机规则**。
- 4 单击**添加**。
- 5 键入规则的名称。
- 6 从“类型”菜单中，选择 **VMDK 反关联性**。
- 7 单击**添加**。
- 8 单击**选择虚拟机**。
- 9 选择虚拟机，然后单击**确定**。
- 10 至少选择两个要应用该规则的虚拟磁盘，然后单击**确定**。
- 11 单击**确定**以保存该规则。

替代 VMDK 关联性规则

VMDK 关联性规则指示数据存储集群中所有与特定虚拟机关联的虚拟磁盘是否都位于此数据存储集群中的同一数据存储上。这些规则将应用到各数据存储集群。

默认情况下，数据存储集群中的所有虚拟机均已激活 VMDK 关联性规则。可以替代数据存储集群或各虚拟机的默认设置。

受 VMDK 关联性规则限制的虚拟机将具有以下行为：

- Storage DRS 将根据规则放置虚拟机的虚拟磁盘。
- 即使是强制进行迁移（如将数据存储置于维护模式），Storage DRS 也会根据规则使用 vMotion 迁移虚拟磁盘。
- 如果虚拟机的虚拟磁盘违反了规则，则 Storage DRS 将提出迁移建议来更正这一错误，或者在无法提出更正错误的建议时将此违反报告为故障。

将数据存储添加到已为 Storage DRS 激活的数据存储集群时，如果在该数据存储上具有虚拟磁盘的虚拟机还在其他数据存储上具有虚拟磁盘，则为此虚拟机停用 VMDK 关联性规则。

步骤

- 1 在 vSphere Client 中，浏览到数据存储集群。
- 2 依次单击**配置**选项卡和**配置**。
- 3 选择**虚拟机替代项**。
- 4 单击**添加**。
- 5 使用 **+** 按钮选择虚拟机。
- 6 单击**聚集 VMDK** 下拉列表，然后选择**否**。
- 7 单击**确定**。

清除 Storage DRS 统计信息

要诊断 Storage DRS 问题，可以先清除 Storage DRS 统计信息，然后再手动运行 Storage DRS。

重要说明 激活清除 Storage DRS 统计信息的选项后，会在每次运行 Storage DRS 时清除统计信息，直到停用该选项。诊断 Storage DRS 问题后，始终停用该选项。

前提条件

为数据存储集群激活 Storage DRS。

步骤

- 1 激活 **ClearIoStatsOnSdrsRun** 选项。
 - a 在 vSphere Client 中，浏览到数据存储集群。
 - b 依次单击**配置**选项卡和**服务**。
 - c 选择 **vSphere DRS**，然后单击**编辑**。
 - d 展开**高级选项**，然后单击**添加**。
 - e 在“选项”列中，键入 **ClearIoStatsOnSdrsRun**。

f 在相应的“值”文本框中，键入 **1**。

g 单击**确定**。

2 对数据存储集群运行 Storage DRS。

vSphere Client 清单中所有数据存储集群中的所有数据存储和虚拟磁盘当前的 Storage DRS 统计信息已清除，但未收集新的统计信息。

3 将 **ClearIoStatsOnSdrsRun** 标记值更改为 **0** 以将其停用。

4 再次运行 Storage DRS。

Storage DRS 将正常运行。允许新设置几个小时后生效。

Storage vMotion 与数据存储集群的兼容性

数据存储集群具有特定的 vSphere Storage vMotion[®] 要求。

- 主机必须运行支持 Storage vMotion 的 ESXi 版本。
- 主机必须对源数据存储和目标数据存储同时具有写入访问权限。
- 主机必须具有足够的可用内存资源来容纳 Storage vMotion。
- 目标数据存储必须具有足够的磁盘空间。
- 目标数据存储不得处于维护模式或进入维护模式。

ESXi 在支持 NUMA（非一致性内存访问）的服务器架构中，支持 Intel 和 AMD Opteron 处理器的内存访问优化。

在了解如何执行 ESXi NUMA 调度以及 VMware NUMA 算法如何工作之后，可以指定 NUMA 控件以优化虚拟机的性能。

本章讨论了以下主题：

- 什么是 NUMA？
- 对操作系统的挑战
- ESXi NUMA 调度的工作方式
- VMware NUMA 优化算法和设置
- 主节点和初始放置位置
- 动态负载平衡和页面迁移
- 针对 NUMA 优化的透明页共享
- NUMA 架构中的资源管理
- 使用虚拟 NUMA
- ESXi 8.0 中的虚拟拓扑
- 虚拟 NUMA 控制
- 指定 NUMA 控制
- 将虚拟机与特定处理器关联
- 使用内存关联性将内存分配与特定 NUMA 节点相关联
- 将虚拟机与指定的 NUMA 节点关联

什么是 NUMA？

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。

几 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, 它使用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。顾名思义, 这种技术的内存访问时间是不一致的, 而且取决于内存的位置和通过其访问内存的节点。

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。

远程内存访问的延迟时间较长, 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。

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。

ESXi NUMA 调度的工作方式

ESXi 使用复杂的 NUMA 调度程序来动态均衡处理器负载、内存局部性或处理器负载。

- 1 由 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 (SRAT) 所示。
- 2 将内存分配给虚拟机时, ESXi 主机会优先从主节点分配内存。虚拟机的虚拟 CPU 被限制在主节点上运行以使内存局部性最大化。
- 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。该调度程序可能会将虚拟机迁移到新的主节点, 以减少处理器负载的不均衡。因为这可能会导致使用更多远程内存, 所以调度程序可能会将虚拟机的内存动态迁移到新的主节点, 以改善内存局部性。在改善总体内存局部性的同时, 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 ESXi NUMA 调度程序管理。例如, 如果为虚拟机手动设置了处理器或内存关联性, NUMA 调度程序可能无法管理该虚拟机。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 ESXi 的 NUMA 优化中受益。

ESXi 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均衡虚拟机的复杂事情。

无论客户机操作系统的类型如何，优化措施都可以顺利发挥作用。ESXi 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，即使是使用旧版操作系统，也可以利用新的硬件。

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，则可以自动管理该虚拟机。NUMA 调度程序调控此类虚拟机的方法是使其跨越各 NUMA 节点。即，虚拟机分为多个 NUMA 客户端，每个客户端都分配到一个节点，然后由调度程序将其作为正常、非跨越客户端进行管理。这可提高某些具有较高局部性且占用大量内存的工作负载的性能。有关配置此功能的行为的信息，请参见[高级虚拟机属性](#)。

ESXi 包括对向客户机操作系统公开虚拟 NUMA 拓扑的支持。有关虚拟 NUMA 控制的详细信息，请参见[使用虚拟 NUMA](#)。

VMware NUMA 优化算法和设置

本节介绍了 ESXi 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和设置。

主节点和初始放置位置

当打开虚拟机电源时，ESXi 会向其分配主节点。虚拟机仅在其主节点内的处理器上运行，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相关联的性能损失。

当打开虚拟机的电源时，会为其分配初始主节点以使 NUMA 节点间的整体 CPU 和内存负载保持均衡。由于在大的 NUMA 系统中节点间的滞后时间变化很大，ESXi 会在引导时确定这些节点间滞后时间，并在初始放置虚拟机（比单个 NUMA 节点更宽）时使用此信息。这些宽的虚拟机放置在彼此靠近的 NUMA 节点上，以实现最低的内存访问滞后。

对于仅运行单个工作负载（例如基准配置，它不会在系统运行过程中发生变化）的系统，仅初始放置方法通常已足够。但是，此方法无法保证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。因此，除了初始放置之外，ESXi 还支持在 NUMA 节点之间动态迁移虚拟 CPU 和内存，以促进 CPU 均衡和扩大内存覆盖区域。

动态负载平衡和页面迁移

ESXi 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再平衡负载。

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可用量的情况下改善性能。

再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，再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），虚拟机将在新的主节点上分配内存，并且仅在新主节点内的处理器上运行。

再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。再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相关联的性能损失。ESXi 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 系统使用虚拟机监控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，从而消除了远程访问内存所带来的损失。

当虚拟机移至新的节点时，ESXi 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 ESXi 主机不会无用地移动内存。

当初始放置位置、动态再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。

针对 NUMA 优化的透明页共享

许多 ESXi 工作负载存在跨虚拟机共享内存的机会。

您可能有多个虚拟机运行同一客户机操作系统的实例，加载了相同的应用程序或组件，或者包含公用数据。在这些情况下，ESXi 系统使用专用的透明页共享技术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消耗的内存通常要少于其在物理机上运行时所需的内存。因此，可以高效地支持更高级别的超额分配。

ESXi 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，因此对于频繁共享的页面，每个 NUMA 节点都有自己的本地副本。当虚拟机使用共享页面时，它们无需访问远程内存。

注 此默认行为在 ESX 和 ESXi 的所有先前版本中亦然如此。

NUMA 架构中的资源管理

可以使用不同类型的 NUMA 架构进行资源管理。

通过安装高度多核系统，NUMA 架构会越来越受欢迎，因为这些架构可改善占用大量内存的工作负载的性能。所有现代的 Intel 和 AMD 系统都具有内置于处理器的 NUMA 支持。此外，还具有传统的 NUMA 系统（例如 IBM 企业 X 型架构），这些系统使用具有专用芯片集支持的 NUMA 行为扩展 Intel 和 AMD 处理器。

通常，您可以使用 BIOS 设置激活和停用 NUMA 行为。例如，在基于 AMD Opteron 的 HP ProLiant 服务器中，可以通过在 BIOS 中激活节点交叉来停用 NUMA。如果激活 NUMA，BIOS 将生成系统资源分配表 (SRAT)，ESXi 使用该表生成用于优化的 NUMA 信息。为了确保调度的公平性，将不会为每个 NUMA 节点内核数太少或总内核数太少的系统激活 NUMA 优化。可修改 `numa.rebalancecorestotal` 和 `numa.rebalancecoresnode` 选项以更改此行为。

使用虚拟 NUMA

vSphere 包括支持向客户机操作系统公开虚拟 NUMA 拓扑，这样便于客户机操作系统和应用程序 NUMA 优化，从而可提高性能。

虚拟 NUMA 拓扑可用于虚拟机，且默认情况下在虚拟 CPU 的数目大于 8 时激活。也可以使用高级配置选项手动影响虚拟 NUMA 拓扑。

首次打开虚拟 NUMA 激活的虚拟机的电源时，其虚拟 NUMA 拓扑基于底层物理主机的 NUMA 拓扑。初始化虚拟机的虚拟 NUMA 拓扑后，除非该虚拟机中的 vCPU 数量已更改，否则该拓扑不会发生变化。

虚拟 NUMA 拓扑不考虑配置到虚拟机的内存。虚拟 NUMA 拓扑不受虚拟机的虚拟插槽数和每个插槽的内核数影响。

如果需要替代虚拟 NUMA 拓扑，请参见 [虚拟 NUMA 控制](#)。

注 启用 CPU HotAdd 将停用虚拟 NUMA。请参见 <https://kb.vmware.com/kb/2040375>。

ESXi 8.0 中的虚拟拓扑

ESXi 8.0 包含增强的虚拟拓扑功能。

虚拟机的虚拟拓扑支持在 GOS 中优化放置和负载均衡。选择与运行虚拟机的宿主机的底层物理拓扑一致的准确虚拟拓扑，对应用程序的性能至关重要。

ESXi 8.0 自动为虚拟机选择最佳 `coresPerSocket`，并选择最佳虚拟 L3 大小。它还包含新的虚拟主板布局，用于在启用 CPU 热插拔时公开虚拟设备的 NUMA 和 vNUMA 拓扑。

注 增强型虚拟拓扑仅在 ESXi 8.0 上可用。虚拟机必须具有硬件版本 20 或更高版本才能使用此功能。

步骤

- 1 要手动配置虚拟机拓扑，请先浏览到虚拟机。
- 2 选择**虚拟机选项**。在 **CPU 拓扑**下，可以调整**每个插槽内核数**和 **NUMA 节点**。

为了在新 NUMA 节点中启用热添加 CPU，请在高级配置选项下添加 `numa.allowHotadd`。然后，您可以手动添加 NUMA 配置。

注 默认情况下，启用 CPU 热插拔的虚拟机将实施单个 NUMA 节点拓扑。任何热添加的 CPU 都将转到单个 NUMA 节点。

- 3 在**设备分配**下，还可以将设备分配给虚拟 NUMA 节点，也可以将其保留为未分配状态。

结果

此新配置的拓扑将在现有虚拟机拓扑部分中显示为**手动**。如果不进行手动配置，则此选项卡将显示为**打开电源时分配**。

虚拟 NUMA 控制

对于内存消耗量大得不成比例的虚拟机，可以使用高级选项来替代默认虚拟 CPU 设置。

可以将以下高级选项添加到虚拟机配置文件中。

表 21-1. 虚拟 NUMA 控制的高级选项

选项	描述	默认值
<code>cpuid.coresPerSocket</code>	确定每个虚拟 CPU 插槽的虚拟内核数。除非配置了 <code>numa.vcpu.followcorespersocket</code> ，否则该选项并不会影响虚拟 NUMA 拓扑。 注 ESXi 8.0 自动为虚拟机选择最佳 <code>coresPerSocket</code> ，默认值显示为 0。	1
<code>numa.vcpu.maxPerVirtualNode</code>	通过以该值作为除数均匀拆分 vCPU 总数来确定虚拟 NUMA 节点数。	8
<code>numa.autosize.once</code>	使用这些设置创建虚拟机模板时，如果使用默认值 <code>TRUE</code> ，则每当您随后打开虚拟机电源时，设置都将保持不变。如果该值设置为 <code>FALSE</code> ，则每次打开电源时，虚拟 NUMA 拓扑都会进行更新。无论何时修改虚拟机中已配置的虚拟 CPU 数，都会对虚拟 NUMA 拓扑重新评估。	<code>FALSE</code>
<code>numa.vcpu.min</code>	虚拟机中生成虚拟 NUMA 拓扑所需的虚拟 CPU 的最小数量。当该值小于 <code>numa.vcpu.min</code> 时，虚拟机将始终为 UMA	9
<code>numa.vcpu.followcorespersocket</code>	设置为 1 时，将恢复为根据 <code>cpuid.coresPerSocket</code> 调整虚拟 NUMA 节点大小这一旧行为。	0
<code>numa.allowHotadd</code>	为了在新 NUMA 节点中激活热添加 CPU 的容量，请在高级配置选项下添加 <code>numa.allowHotadd</code> 。然后，您可以在激活 CPU 热添加时手动添加 NUMA 配置。	<code>FALSE</code>
<code>numa.vcpu.coresPerNode</code>	用于配置虚拟 NUMA 节点大小的 VMX 参数，从 UI 重新配置中解释。此参数仅对 HWv20 有效。默认为 0，表示 ESXi 自动选择 vNUMA 大小。 注 如果此选项与 <code>numa.vcpu.maxPerVirtualNode</code> 冲突，将无法打开虚拟机电源。	0
<code>vcpu.hotadd</code>	当此选项为 <code>TRUE</code> 时，会停用虚拟 NUMA。当虚拟机能够进行 CPU 热插拔时，虚拟机始终会看到一个虚拟 NUMA 节点。	
<code>llc.multiLLCPerSocket</code>	当此选项为 <code>TRUE</code> 时，虚拟机会在 AMD Epyc 上公开真实的 vLLC。公开的 vLLC 大小可以不同于虚拟套接字大小。	<code>FALSE</code>
<code>llc.size.vcpu</code>	为 AMD Epyc 上的 vLLC 手动配置的 vCPU 数。如果该值与虚拟机的其他设置不兼容，则会忽略该值。	

表 21-1. 虚拟 NUMA 控制的高级选项（续）

选项	描述	默认值
chipset.motherboardLayout	此虚拟机使用的虚拟主板的类型。它只能具有以下两个值之一： acpi : 从 HWv 20 开始的新主板布局。 i440bx : 旧版主板布局	
cpuid.coresPerSocket.cookie	这是由 ESXi 生成的 vmx 条目，用于存储自动生成的 coresPerSocket 值。这样做可确保 vMotion 的一致性。请勿手动更改或删除它。	

指定 NUMA 控制

如果您有一些占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（例如内存中的数据库或具有大型数据集的科学计算应用程序），指定控制将非常有用。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 8 个虚拟机而且具有类似工作负载的 8 个处理器组成的系统，很容易进行明确地优化。

注 大多数情况下，ESXi 主机的自动 NUMA 优化会产生良好的性能。

ESXi 为 NUMA 放置位置提供了三组控制，因此管理员可以控制虚拟机的内存和处理器位置。

可以指定以下选项。

NUMA 节点关联性

设置该选项时，NUMA 仅可以在关联性中指定的节点上调度虚拟机。

CPU 关联性

如果设置了此选项，则虚拟机仅使用关联性中指定的处理器。

内存关联性

如果设置了此选项，则服务器仅在指定的节点上分配内存。

即使指定了 NUMA 节点关联性，虚拟机仍由 NUMA 管理，但其虚拟 CPU 仅可以在 NUMA 节点关联性中指定的节点上进行调度。同样，仅可以从 NUMA 节点关联性中指定的节点上获取内存。如果指定了 CPU 或内存关联性，则虚拟机不再受 NUMA 管理。这些虚拟机的 NUMA 管理在移除 CPU 和内存关联性限制后有效。

手动 NUMA 放置位置可能会干扰 ESXi 资源管理算法，这种算法在系统之间公平地分发处理器资源。例如，如果将具有占用大量处理器的工作负载的 10 个虚拟机手动置于一个节点上，并且仅将 2 个虚拟机手动置于另一个节点上，则系统不可能为所有的 12 个虚拟机赋予相等份额的系统资源。

将虚拟机与特定处理器关联

通过将虚拟机的虚拟 CPU 固定到固定处理器，可能会改善虚拟机上应用程序的性能。这样可以防止虚拟 CPU 在 NUMA 节点之间进行迁移。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后单击**编辑设置**。
- 3 选择**虚拟硬件**选项卡，然后展开 **CPU**。
- 4 在“调度关联性”下，为偏好处理器设置 CPU 关联性。

注 必须手动选择 NUMA 节点中的所有处理器。CPU 关联性是按照处理器指定的，而不是按照节点指定的。

使用内存关联性将内存分配与特定 NUMA 节点相关联

可以指定虚拟机上所有的后续内存分配使用与特定 NUMA 节点关联的页（也称为手动内存关联性）。

注 只有在指定了 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 再平衡功能将无法正常工作。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
- 2 单击**配置**选项卡。
- 3 单击**设置**，然后单击**虚拟机硬件**。
- 4 单击**编辑**。
- 5 选择**虚拟机**选项卡，然后展开**内存**。
- 6 在“NUMA 内存关联性”下，设置内存关联性。

示例：将虚拟机绑定到单个 NUMA 节点

以下示例说明了将最后四个物理 CPU 手动绑定到 8 路服务器上双路虚拟机的单个 NUMA 节点。

CPU（例如 4、5、6 和 7）是物理 CPU 编号。

- 1 在 vSphere Client 中，右键单击虚拟机，然后选择**编辑设置**。
- 2 选择**选项**并单击**高级**。
- 3 单击**配置参数**按钮。

4 在 vSphere Client 中，为处理器 4、5、6 和 7 打开 CPU 关联性。

接着，您希望此虚拟机仅在节点 1 上运行。

1 在 vSphere Client “清单” 面板中，选择该虚拟机并选择**编辑设置**。

2 选择**选项**并单击**高级**。

3 单击**配置参数**按钮。

4 在 vSphere 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在 NUMA 节点 1 上运行，并在可能的情况下从同一个节点分配内存。

将虚拟机与指定的 NUMA 节点关联

将 NUMA 节点与虚拟机关联以指定 NUMA 节点关联性时，限制 ESXi 可在上面调度虚拟机的虚拟 CPU 和内存的一组 NUMA 节点。

注 如果限制 NUMA 节点关联性，则可能会影响 ESXi NUMA 调度程序公平地在 NUMA 节点之间重新均衡虚拟机的功能。仅在考虑重新均衡问题后指定 NUMA 节点关联性。

步骤

1 在 vSphere Client 中，浏览到集群。

2 依次单击**配置**选项卡和**设置**。

3 在**虚拟机选项**下，单击**编辑**按钮。

4 选择**虚拟机选项**选项卡，然后展开**高级**。

5 在**配置参数**下，单击**编辑配置**按钮。

6 单击**添加行**添加新选项。

7

- 要为虚拟机指定 NUMA 节点，请在“名称”列中输入 **numa.nodeAffinity**。
- 要为虚拟机上的特定虚拟 NUMA 节点指定 NUMA 节点，请在“名称”列中输入 **sched.nodeX.affinity**，其中 X 是虚拟 NUMA 节点编号。例如，**sched.node0.affinity** 指定虚拟机上的虚拟 NUMA 节点 0。

8 在“值”列中，输入可在上面调度虚拟机或虚拟 NUMA 节点的 NUMA 节点。

如果有多个节点，则以逗号分隔。例如，输入 **0,1** 以将虚拟机资源调度限制为 NUMA 节点 0 和 1。

9 单击**确定**。

10 单击**确定**关闭“编辑虚拟机”对话框。

可以为主机或单个虚拟机设置高级属性以帮助自定义资源管理。

大多数情况下，调整基本资源分配设置（预留、限制和份额）或接受默认设置可以获得适当的资源分配结果。但是，可以使用高级属性为主机或特定虚拟机自定义资源管理。

本章讨论了以下主题：

- 设置高级主机属性
- 设置高级虚拟机属性
- 延迟时间敏感度
- 虚拟机的虚拟超线程支持
- vHT 完整 CPU 预留
- 为虚拟机激活 vHT
- 关于可靠内存
- 使用 1GB 页面备份客户机 vRAM

设置高级主机属性

可以为主机设置高级属性。

小心 更改高级选项将被视为不受支持。通常，使用默认设置即可获得最佳结果。仅当 VMware 技术支持或知识库文章提供了具体指示时，才能更改高级选项。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 单击**配置**选项卡。
- 3 在**系统**下，单击**高级系统设置**。
- 4 单击**编辑**按钮。
- 5 找到相应的项目并更改其值。
- 6 单击**确定**。

高级内存属性

可以使用高级内存属性自定义内存资源使用情况。

表 22-1. 高级内存属性

属性	描述	默认
Mem.ShareForceSalting	<p>Mem.ShareForceSalting 0: 仍会保留虚拟机间透明页面共享 (TPS) 行为。VMX 选项 sched.mem.pshare.salt 的值即使存在也会被忽略。</p> <p>Mem.ShareForceSalting 1: 默认情况下, 加密盐值来自 sched.mem.pshare.salt。如果未指定, 则将虚拟机的加密盐值视为 0, 从而回退到旧版 TPS (虚拟机间) 行为。</p> <p>Mem.ShareForceSalting 2: 默认情况下, 加密盐值来自 sched.mem.pshare.salt (如果存在) 或 vc.uuid。如果不存在, 则页面共享算法会生成一个用于为每台虚拟机设置盐的唯一随机值, 用户不可对该值进行配置。</p>	2
Mem.SamplePeriod	指定虚拟机执行时间的周期性时间间隔 (以秒为度量单位), 在该执行时间内监控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性时间间隔, 以秒为单位。可用内存量的重大更改也会触发重新分配。	15
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使用的内存的消耗量。0 % 的消耗率定义的分配策略将忽略工作集并严格按照份额分配内存。较高的消耗率产生的分配策略允许要重新分配的闲置内存远离以非生产性方式累积闲置内存的虚拟机。	75
Mem.ShareScanGHz	指定每 1 GHz 可用主机 CPU 资源为寻找页面共享机会, 每秒内可用于扫描的最大内存页面量。例如, 默认值为每 GHz 的速率为 4 MB/秒。	4
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。默认值为 60 分钟。	60
Mem.CtlMaxPercent	根据所配置内存大小的百分比, 使用内存气球驱动程序 (vmmemctl) 限制从任何虚拟机回收的最大内存量。指定 0 将停止回收所有虚拟机。	65
Mem.AllocGuestLargePage	激活使用主机大页支持客户机大页。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。0 = 停用。	1
Mem.AllocUsePSharePool 和 Mem.AllocUseGuestPool	通过提高让主机大页作为客户机备用大页的可能性来减少内存碎片。如果主机内存有碎片, 则主机大页的可用性会降低。0 = 停用。	15
Mem.MemZipEnable	激活主机的内存压缩。0 = 停用。	1
Mem.MemZipMaxPct	根据每个虚拟机的可存储为压缩内存的内存最大百分比, 指定压缩缓存的最大大小。	10
LPage.LPageDefragEnable	激活大页碎片整理。0 = 停用。	1
LPage.LPageDefragRateVM	每个虚拟机上每秒内最多可尝试的大页碎片整理次数。可接受的值在 1 到 1024 之间。	32

表 22-1. 高级内存属性（续）

属性	描述	默认
LPage.LPageDefragRateTotal	每秒内最多可尝试的大页碎片整理次数。可接受的值在 1 到 10240 之间。	256
LPage.LPageAlwaysTryForNPT	尝试为嵌套页表（AMD 称为“RVI”，Intel 称为“EPT”）分配大页。如果激活此选项，则所有客户机内存都受到使用嵌套页表的计算机（例如，AMD Barcelona）中的大页支持。如果 NPT 不可用，则只有部分客户机内存受到大页支持。0 = 停用。	1

高级 NUMA 属性

可以使用高级 NUMA 属性自定义 NUMA 使用情况。

表 22-2. 高级 NUMA 属性

属性	描述	默认
Numa.RebalancePeriod	控制重新均衡周期的频率，以毫秒为单位指定。重新均衡的频率越大，CPU 开销也越大，是在运行大量虚拟机的计算机上尤其如此。频繁的重重新均衡还可以提高公平性。	2000
Numa.MigImbalanceThreshold	NUMA 重新均衡器计算节点之间 CPU 的不均衡，考虑每个虚拟机的 CPU 时间可用量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均衡，以百分比为单位。	10
Numa.RebalanceEnable	激活 NUMA 重新均衡和调度。将此选项设置为 0 可针对虚拟机停用所有的 NUMA 重新均衡和初始放置，从而有效地停用 NUMA 调度系统。	1
Numa.RebalanceCoresTotal	指定主机上激活 NUMA 重新均衡器所需的处理器内核的最小总数。	4
Numa.RebalanceCoresNode	指定每个节点上激活 NUMA 重新均衡器所需的处理器内核的最小数量。 在小型 NUMA 配置（例如，2 路 Opteron 主机）中停用 NUMA 重新均衡时，此选项和 Numa.RebalanceCoresTotal 会非常有用，在这样的配置中，如果激活了 NUMA 重新均衡功能，而且处理器总数或每个节点上的处理器较少，则会影响调度的公平性。	2
Numa.AutoMemAffinity	自动设置具有 CPU 关联性集合的虚拟机的内存关联性。	1
Numa.PageMigEnable	在 NUMA 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

设置高级虚拟机属性

可以为虚拟机设置高级属性。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后选择**编辑设置**。
- 3 单击**虚拟机选项**。
- 4 展开**高级**。
- 5 在“配置参数”下，单击**编辑配置**按钮。
- 6 在显示的对话框中，单击**添加行**以输入新参数及其值。
- 7 单击**确定**。

高级虚拟机属性

可以使用高级虚拟机属性自定义虚拟机配置。

表 22-3. 高级虚拟机属性

属性	描述	默认
sched.mem.maxmemctl	通过膨胀而从选定虚拟机中回收的最大内存量，以兆字节 (MB) 为单位。如果 ESXi 主机需要回收更多内存，则强制进行交换。交换的优先级低于膨胀。	-1（无限制）
sched.mem.pshare.enabled	为选定的虚拟机启用内存共享。 此布尔值默认为“有效”。如果将虚拟机的该属性设置为“无效”，则将关闭内存共享。	有效
sched.mem.pshare.salt	加密盐值是每个虚拟机的可配置 VMX 选项。如果虚拟机的 VMX 文件中不存在此选项，则 vc.uuid vmx 选项的值将作为默认值。由于每个虚拟机的 vc.uuid 各不相同，因此默认情况下，透明页面共享仅发生在属于特定虚拟机的页面之间（虚拟机内部）。如果一组虚拟机被视为可信，则可以通过为所有这些虚拟机设置公用加密盐值来在它们之间共享页面（虚拟机间）。	用户可配置
sched.swap.persist	指定关闭虚拟机电源时应保留还是删除虚拟机的交换文件。默认情况下，当虚拟机打开电源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。	无效
sched.swap.dir	虚拟机交换文件的目录位置。默认为虚拟机的工作目录，即包含其配置文件的目录。此目录必须保留在虚拟机可访问的主机上。如果移动虚拟机（或从虚拟机创建的任何克隆），则可能需要重置此属性。	等于 workingDir

高级虚拟 NUMA 属性

可以使用高级虚拟 NUMA 属性自定义虚拟 NUMA 使用情况。

表 22-4. 高级 NUMA 属性

属性	描述	默认
<code>cpuid.coresPerSocket</code>	确定每个虚拟 CPU 插槽的虚拟内核数。如果该值大于 1 且虚拟机具备虚拟 NUMA 拓扑，则还可确定虚拟 NUMA 节点的大小。如果知道每个物理主机精确的虚拟 NUMA 拓扑，则可以设置此选项。	1
<code>numa.autosize</code>	设置此选项时，虚拟 NUMA 拓扑中每个虚拟节点的虚拟 CPU 数等于每个物理节点的内核数。	FALSE
<code>numa.autosize.once</code>	当使用这些设置创建虚拟机模板时，请确保这些设置在您以后每次打开虚拟机电源时保持不变。如果修改了虚拟机上配置的虚拟 CPU 数，则需要重新评估虚拟 NUMA 拓扑。	TRUE
<code>numa.vcpu.maxPerVirtualNode</code>	如果 <code>cpuid.coresPerSocket</code> 严格限定为 2 的幂，则可以直接设置 <code>numa.vcpu.maxPerVirtualNode</code> 。在这种情况下，请不要设置 <code>cpuid.coresPerSocket</code> 。	8
<code>numa.vcpu.min</code>	虚拟机中生成虚拟 NUMA 拓扑所需的虚拟 CPU 的最小数目。	9
<code>numa.vcpu.maxPerMachineNode</code>	同一虚拟机中可同时调度到某个 NUMA 节点上的虚拟 CPU 的最大数目。使用该属性将不同的 NUMA 客户端强制分配到不同的 NUMA 节点可确保最大带宽。	运行虚拟机的物理主机上每个节点的内核数。
<code>numa.vcpu.maxPerClient</code>	NUMA 客户端中的虚拟 CPU 的最大数目。客户端是由 NUMA 作为单个实体进行管理的一组虚拟 CPU。默认情况下，每个虚拟 NUMA 节点为一个 NUMA 客户端。但是，如果虚拟 NUMA 节点大于物理 NUMA 节点，则单个虚拟 NUMA 节点可由多个 NUMA 客户端支持。	等于 <code>numa.vcpu.maxPerMachineNode</code>
<code>numa.nodeAffinity</code>	限制可在其上调度虚拟机的虚拟 CPU 和内存的一组 NUMA 节点。 注 如果限制 NUMA 节点关联性，可能会影响 NUMA 调度程序为确保公平性在 NUMA 节点之间再平衡虚拟机的功能。仅在考虑再平衡问题后指定 NUMA 节点关联性。	
<code>numa.mem.interleave</code>	指定分配给虚拟机的内存是否在所有 NUMA 节点（其上运行作为其组成部分的 NUMA 客户端）之间静态交叉，且未显示虚拟 NUMA 拓扑。	有效

延迟时间敏感度

您可以调整虚拟机的延迟时间敏感度，以优化延迟时间敏感度应用程序的调度延迟。

ESXi 已经过优化，可提供高吞吐量。您可以优化虚拟机，以满足延迟时间敏感度应用程序的短延迟时间要求。延迟时间敏感度应用程序示例包括 VOIP 或媒体播放器应用程序，或者需要频繁访问鼠标或键盘设备的应用程序。

调整延迟时间敏感度

可以调整虚拟机的延迟时间敏感度。

前提条件

延迟敏感度设置为**高**时，ESXi 需要完全 CPU 预留才能打开具有硬件版本 14 的虚拟机的电源。

步骤

- 1 在 vSphere Client 中，浏览到虚拟机。
 - a 要查找虚拟机，请选择数据中心、文件夹、集群、资源池或主机。
 - b 单击**虚拟机**选项卡。
- 2 右键单击虚拟机，然后单击**编辑设置**。
- 3 单击**虚拟机选项**，然后单击**高级**。
- 4 从**延迟时间敏感度**下拉菜单中选择一个设置。
- 5 单击**确定**。

虚拟机的虚拟超线程支持

虚拟机支持虚拟超线程 (vHT)。

在 vSphere 8.0 中，支持虚拟机 vHT。vHT 默认处于停用状态，可以在每个虚拟机的延迟敏感度设置下激活。vHT 支持的最大 HT 大小为 2。

vHT 是延迟敏感度高功能的扩展。受益于超线程感知的应用程序将因延迟敏感度高和激活的 vHT 而获得性能提升。性能提升可能来自于足够的资源预留，以及虚拟机具有独占的物理 CPU。

如果未在 ESXi 上激活 vHT，则每个虚拟 CPU (vCPU) 相当于客户机操作系统可用的单个非超线程内核。激活 vHT 后，将每个客户机 vCPU 视为虚拟内核 (vCore) 的单个超线程。

同一 vCore 的虚拟超线程占用同一物理内核。因此，虚拟机的 vCPU 可以共享同一内核，而不是在已停用 vHT 的高延迟敏感度的虚拟机上使用多个内核。

运行较旧硬件版本的 ESXi 主机和虚拟机无法使用此功能。

vHT 完整 CPU 预留

您可以使用公式计算 vHT 的完整 CPU 预留。

对于没有 vHT 的低延迟虚拟机，虚拟机的每个 vCPU 都具有与物理内核的线程的独占关联性。对于已激活超线程的主机，合作伙伴超线程与空闲环境具有独占关联性。为低延迟虚拟机的每个 vCPU 分配一个专用物理内核。

低延迟虚拟机的 CPU 预留计算如下：

```
低延迟虚拟机（无 vHT）CPU 最小预留 = numVcpus * cpuFrequency
```

但是，当为虚拟机激活 vHT 时，物理内核的每个 hptwin 将在虚拟机的多个 vCPU 之间共享，其中每个超 hypertwin 与虚拟机的一个 vCPU 具有独占关联性。这意味着具有 numSMT 个物理超线程的内核由多个 numSMT 虚拟线程共享。在这种情况下，CPU 预留要求计算如下：

$$\text{低延迟虚拟机（具有 vHT）CPU 最小预留} = (\text{numVcpus} / \text{numSMT}) * \text{cpuFrequency}$$

表 22-5. 在 CPU 频率为 2 GHz 的主机上引导具有 20 个 vCPU 的低延迟虚拟机的示例

	numSMT = 1（不含 vHT）	numSMT = 2（使用 vHT）
numVcpus	20	20
物理内核数	20	10（每个内核由 2 个 vCPU 共享）
所需的最小 CPU 预留	20 * 2.0 GHz = 40 GHz	(20/2) * 2.0 GHz = 20 GHz

为虚拟机激活 vHT

ESXi 8.0 支持 vHT，但 vHT 默认处于停用状态。您可以在每个虚拟机的延迟敏感度设置下激活 vHT。

前提条件

激活 vHT 后，CPU 和内存必须设置为完全预留。如果将预留设置得较低，则会显示一条警告。

步骤

- 1 在 vSphere 中，选择虚拟机。
- 2 选择**操作**，然后单击**编辑设置**。
- 3 在“延迟敏感度”下，单击下拉菜单，然后选择**高 (超线程)**。
- 4 单击**确定**。

结果

此时，vHT 处于激活状态。

关于可靠内存

ESXi 支持可靠内存。

一些系统具有可靠内存，可靠内存是指相较于系统中其他部分的内存，不太会发生硬件内存错误的那部分内存。如果硬件公开有关不同级别的可靠性的信息，则 ESXi 可能能够实现更高的系统可靠性。

查看可靠内存

您可以查看许可证是否允许可靠内存。

步骤

- 1 在 vSphere Client 中，浏览到主机。
- 2 依次单击**配置**选项卡和**系统**。
- 3 选择**许可**。
- 4 在**已获许可的功能**下，验证是否已显示可靠内存。

后续步骤

可以使用 `ESXCLI hardware memory get` 命令查找被视为可靠的内存量。

使用 1GB 页面备份客户机 vRAM

vSphere ESXi 支持使用 1 GB 页面备份客户机 vRAM，但提供的支持有限。

要使用 1 GB 页面备份客户机内存，必须对虚拟机应用 `sched.mem.lpage.enable1GPage = "TRUE"` 选项。选择**编辑设置**后，可以在“高级”选项下进行此项设置。只能在关闭电源的虚拟机上启用 1 GB 页面。

启用了 1 GB 页面的虚拟机必须预留全部内存，否则，虚拟机将无法打开电源。启用了 1 GB 页面的虚拟机的所有 vRAM 在打开电源时即进行预先分配。由于这些虚拟机会预留全部内存，因此它们不受内存回收的影响，而且它们的内存使用量在虚拟机的整个生命周期内都保持在最高水平。

能否使用 1 GB 页面备份 vRAM 须视情况而定，而系统会尽最大努力分配 1 GB 页面。其中包括主机 CPU 不支持 1 GB 页面功能的情况。要最大限度提高使用 1 GB 页面备份客户机 vRAM 的几率，我们建议在刚刚引导的主机上启动需要使用 1 GB 页面的虚拟机，因为随着时间的推移主机 RAM 会出现碎片。

启用了 1 GB 页面的虚拟机可迁移至其他主机。但是，目标主机可能不会像源主机那样分配 1 GB 页面大小。您还可能会发现，在源主机上使用 1 GB 页面进行备份的部分 vRAM 在目标主机上不再使用 1 GB 页面进行备份。

1 GB 页面这种视情况而定的性质也适用于 HA 和 DRS 之类的 vSphere 服务，也就是说这些服务可能不会保留使用 1 GB 页面备份 vRAM 的功能。原因是，这些服务并不知晓目标主机是否支持 1 GB 页面功能，在做出放置决策时也不会考虑 1 GB 内存备份。

DRS 故障会指示阻止生成 DRS 操作（或阻止在手动模式下提出 DRS 操作建议）的原因。

本节定义了 DRS 故障。

注 在本章中，“内存”可以指物理内存或永久内存。

本章讨论了以下主题：

- 虚拟机已固定
- 虚拟机与任何主机均不兼容
- 移动到另一台主机时违反了虚拟机/虚拟机 DRS 规则
- 主机与虚拟机不兼容
- 主机有违反虚拟机/虚拟机 DRS 规则的虚拟机
- 主机用于虚拟机的容量不足
- 主机处于错误的状态
- 主机用于虚拟机的物理 CPU 的数量不足
- 主机用于每个虚拟机 CPU 的容量不足
- 虚拟机正在执行 vMotion 操作
- 集群中没有活动主机
- 资源不足
- 资源不足以满足配置的 HA 故障切换级别
- 无兼容的硬关联性主机
- 无兼容的软关联性主机
- 不允许违反软规则更改
- 影响软规则更改

虚拟机已固定

当因为虚拟机上已停用 DRS 而导致 DRS 不能移动虚拟机时，会发生此故障。即，虚拟机在其注册的主机上“固定”了。

虚拟机与任何主机均不兼容

当 DRS 找不到可以运行虚拟机的主机时，会出现此错误。

例如，如果没有主机可以满足虚拟机的 CPU 或内存资源需求，或者目前没有主机拥有虚拟机所需的网络或存储访问权限，可能会出现此错误。

要解决此问题，请提供能够满足虚拟机要求的主机。

移动到另一台主机时违反了虚拟机/虚拟机 DRS 规则

如果在同一主机上运行多个相互共享关联性规则的虚拟机，那么，当无法将这些虚拟机移动到另一个主机时，会发生此错误。

由于只有部分虚拟机可以通过 vMotion 从当前主机移出，因此有可能发生此错误。例如，组中的一个虚拟机停用了 DRS。

要防止发生此错误，请检查该组中某些虚拟机无法通过 vMotion 移动的原因。

主机与虚拟机不兼容

当 DRS 考虑将虚拟机迁移到主机，但发现主机与给定虚拟机不兼容时，此错误会出现。

当目标主机无权访问虚拟机所需的网络或存储连接时，可能发生此错误。发生此故障的另一个原因是目标主机的 CPU 与当前主机相差太大，以致于无法支持在主机间使用 vMotion。

为避免此错误，请在创建集群时使所有主机的配置一致而且主机间的 vMotion 兼容。

主机与虚拟机不兼容的另一个原因是，必须存在一个虚拟机/主机 DRS 规则，该规则要求 DRS 绝不该将此虚拟机放置在此主机上。

主机有违反虚拟机/虚拟机 DRS 规则的虚拟机

当通过启动 vMotion 打开虚拟机电源或移动虚拟机会违反虚拟机/虚拟机 DRS 规则时，会出现此故障。

仍可以手动打开虚拟机电源或手动使用 vMotion 移动虚拟机，但 vCenter Server 无法自动执行这些操作。

主机用于虚拟机的容量不足

当主机没有足够的 CPU 或内存容量用于运行虚拟机时，此错误会出现。

主机处于错误的状态

若主机进入维护或待机模式时需要进行 DRS 操作，则会出现此故障。

要更正此错误，请取消有关主机进入待机或维护模式的请求。

主机用于虚拟机的物理 CPU 的数量不足

当主机硬件没有足够的 CPU（超线程）来支持虚拟机中的虚拟 CPU 数目时，就会出现此故障。

主机用于每个虚拟机 CPU 的容量不足

当主机没有足够的 CPU 容量用于运行虚拟机时，此故障会出现。

虚拟机正在执行 vMotion 操作

当 DRS 因为虚拟机正在执行 vMotion 操作而不能移动它时，会发生此故障。

集群中没有活动主机

如果集群内的虚拟机正在被移动，且该集群不包含任何处于连接状态和非维护状态的主机，则会发生此错误。

例如，如果所有主机均断开或处于维护模式，便可能出现此情况。

资源不足

当所尝试的操作与资源配置策略相冲突时，会出现此错误。

例如，如果打开电源操作预留的内存多于分配到资源池的内存时，此错误可能出现。

调整资源以允许更多内存后，重试该操作。

资源不足以满足配置的 HA 故障切换级别

当违反为故障切换保留的 CPU 或内存资源的 HA 配置，或 HA 配置不足以运行 DRS 操作时，就会出现此故障。

在以下情况下将报告此故障：

- 请求主机进入维护或待机模式。
- 尝试打开虚拟机电源时与故障切换发生冲突。

无兼容的硬关联性主机

没有主机可用于满足其强制性虚拟机/主机 DRS 关联性或反关联性规则的虚拟机。

无兼容的软关联性主机

没有主机可用于满足其首选虚拟机/主机 DRS 关联性或反关联性规则的虚拟机。

不允许违反软规则更改

DRS 迁移阈值被设置为仅强制性。

这会禁止生成更正非强制性虚拟机/主机 DRS 关联性规则的 DRS 操作。

影响软规则更改

因为会影响性能，所以不对非强制性虚拟机/主机 DRS 关联性规则进行更正。

此信息描述了特定类别的 vSphere® Distributed Resource Scheduler (DRS) 问题：集群、主机和虚拟机问题。

注 在本章中，“内存”可以指物理内存或永久内存。

本章讨论了以下主题：

- 集群问题
- 主机问题
- 虚拟机问题

集群问题

集群问题可导致 DRS 无法以最佳状态执行或报告故障。

集群负载不均衡

集群资源负载不均衡。

问题

由于虚拟机的资源需求不平均并且主机容量也不相同，因此集群可能会不均衡。

原因

以下是集群负载不均衡的可能原因：

- 迁移阈值过高。
阈值越高，集群越容易出现负载不均衡。
- 虚拟机/虚拟机或虚拟机/主机 DRS 规则可阻止移动虚拟机。
- 为一个或多个虚拟机停用了 DRS。
- 某个设备挂载到了一个或多个虚拟机上，使 DRS 无法移动虚拟机，从而均衡负载。
- 虚拟机与 DRS 要将它们移动到的主机不兼容。这就是说，集群中至少有一个主机与将要迁移的虚拟机不兼容。例如，如果主机 A 的 CPU 不与主机 B 的 CPU vMotion 兼容，则主机 A 将与在主机 B 上运行的已打开电源虚拟机变得不兼容。

- 与虚拟机保留在当前位置继续运行相比，移动虚拟机可能会对其性能更加不利。当负载不稳定或者迁移成本比移动虚拟机所带来的收益要高时，就会出现上述情况。
- 没有为集群中的主机激活或设置 vMotion。

解决方案

解决导致负载不均衡的问题。

集群为黄色

该集群由于资源短缺而变为黄色。

问题

如果集群没有足够的资源满足所有资源池和虚拟机的预留，但有足够的资源来满足所有正在运行中的虚拟机的预留，则 DRS 将继续运行，同时集群显示为黄色。

原因

如果从集群中移除了主机资源（例如，主机出现故障），则集群可能会变为黄色。

解决方案

将主机资源添加到集群，或减少资源池预留量。

集群为红色，因为资源池不一致

DRS 集群无效时显示为红色。其可能因为资源池树内部不一致而变为红色。

问题

如果集群资源池树内部不一致（例如，子资源池预留总数大于父资源池不可扩展的预留），集群就没有足够的资源来满足所有正在运行的虚拟机的预留，从而导致集群显示为红色。

原因

如果 vCenter Server 不可用，或者如果资源池设置在虚拟机处于故障切换状态时发生了更改，则可能出现此情况。

解决方案

恢复关联的更改，或者修改资源池设置。

集群为红色，因为与故障切换容量发生冲突

DRS 集群无效时显示为红色。其可能因为与故障切换容量发生冲突而变为红色。

问题

集群会在主机发生故障时尝试对虚拟机进行故障切换，但不能保证有足够的可用资源对故障切换要求所涵盖的所有虚拟机进行故障切换。

原因

如果启用了 HA 的集群失去的资源过多以致无法再满足故障切换要求，会显示一条消息，而且集群状态将变成红色。

解决方案

查看集群摘要页面顶部黄色框中的配置问题列表，并解决导致该状况的问题。

集群总负载低时主机电源不关闭

集群总负载低时主机电源不关闭。

问题

因为 HA 故障切换预留需要额外容量，因此当集群总负载低时，不会关闭主机电源。

原因

主机可能会因为以下原因而无法关闭电源：

- 需要满足 MinPoweredOn{Cpu|Memory}Capacity 高级选项设置。
- 由于资源预留、虚拟机/主机 DRS 规则、虚拟机/虚拟机 DRS 规则、未激活 DRS 或者与具有可用容量的主机不兼容，无法将虚拟机整合到较少数量的主机上。
- 负载不稳定。
- DRS 迁移阈值处于最高设置，仅允许强制移动。
- vMotion 无法运行，因为未进行配置。
- 在可能已关闭电源的主机上停用了 DPM。
- 主机与将要移动到另一主机上的虚拟机不兼容。
- 主机不具备唤醒 LAN、IPMI 或 iLO 技术。必须满足其中任一条件，DPM 才能进入处于待机状态的主机。

解决方案

解决导致集群总负载低时无法关闭主机电源的问题。

集群总负载高时关闭主机电源

集群总负载高时主机电源关闭。

问题

DRS 确保虚拟机可以在较少的主机上运行，同时不降低主机或虚拟机性能。另外，还限制 DRS 将高利用率主机上运行的虚拟机移动到计划关闭电源的主机上。

原因

集群总负载过高。

解决方案

降低集群负载。

DRS 很少或从不执行 vMotion 迁移

DRS 很少或从不执行 vMotion 迁移。

问题

DRS 不执行 vMotion 迁移。

原因

集群中出现以下一个或多个问题时，DRS 从不执行 vMotion 迁移。

- 在集群上停用了 DRS。
- 主机没有共享存储。
- 集群内的主机不包含 vMotion 网络。
- DRS 需手动操作而无人批准迁移。

集群中出现以下一个或多个问题时，DRS 很少执行 vMotion：

- 负载不稳定，或者 vMotion 耗时过长，抑或二者兼有。移动不适当。
- DRS 很少或从不迁移虚拟机。
- DRS 迁移阈值设置得过高。

DRS 移动虚拟机的原因如下：

- 用户请求其进入维护或待机模式的主机撤出。
- 虚拟机/主机 DRS 规则或虚拟机/虚拟机 DRS 规则。
- 预留冲突。
- 负载不均衡。
- 电源管理。

解决方案

请解决导致 DRS 避免执行 vMotion 迁移的问题。

主机问题

主机问题可能会导致 DRS 无法按预期方式执行。

DRS 建议在集群总负载低时打开主机电源以增加容量

必须打开主机电源，这样才有助于为集群提供更多容量或者对过量分配的主机提供帮助。

问题

DRS 建议当集群总负载低时，打开主机电源来增加容量。

原因

可能会进行此建议，其原因是：

- 集群是 DRS-HA 集群。需要其他已打开电源的主机来提供更多的故障切换功能。
- 部分主机过载，而且可以将目前已打开电源的主机上的虚拟机移动到待机模式下的主机以平衡负载。
- 需要容量以满足 `MinPoweredOn{Cpu|Memory}Capacity` 高级选项的要求。

解决方案

打开该主机电源。

集群总负载高

集群总负载较高。

问题

当集群总负载高时，DRS 不会打开主机电源。

原因

以下是 DRS 无法打开主机电源的可能原因：

- 虚拟机/虚拟机 DRS 规则或虚拟机/主机 DRS 规则阻止将虚拟机移动到该主机。
- 虚拟机已固定到其当前主机，因此 DRS 无法将这些虚拟机移动到待机模式下的主机以达到负载均衡。
- DRS 或 DPM 处于手动模式中，且未采用建议。
- 没有将高利用率主机上的任何虚拟机移动到该主机。
- 因用户设置或主机之前退出待机模式失败，主机上已停用 DPM。

解决方案

解决阻止 DRS 打开主机电源的问题。

集群总负载低

集群总负载低。

问题

当集群总负载低时，DRS 不会打开主机电源。

原因

以下是 DRS 无法打开主机电源的可能原因：

- 分布式电源管理 (DPM) 检测到更好的电源关闭候选对象。
- vSphere HA 需要额外的容量进行故障切换。
- 负载不够低，不足以触发主机关闭电源操作。
- DPM 预测负载将增加。
- 没有为主机启用 DPM。
- DPM 阈值设置得过高。
- 为主机启用 DPM 期间，没有适合主机的打开电源机制存在。
- DRS 不能撤出主机。
- DRS 迁移阈值处于最高设置，仅可执行强制移动。

解决方案

解决阻止 DRS 关闭主机电源的问题。

DRS 没有撤出请求进入维护或待机模式的主机

DRS 没有撤出请求进入维护模式或待机模式的主机。

问题

尝试将主机置于维护模式或待机模式时，DRS 没有正常撤出主机。

原因

激活 vSphere HA 后，撤出该主机可能会与 HA 故障切换容量发生冲突。

解决方案

无解决方案。如果适用，请先停用 vSphere HA，然后再尝试将主机置于维护模式或待机模式。

DRS 没有将任何虚拟机移动到主机上

DRS 没有将任何虚拟机移动到主机上。

问题

如果主机已添加到激活了 DRS 的集群，DRS 不建议将虚拟机迁移到这样的主机上。

原因

将主机添加到已激活 DRS 的集群后，部署到该主机的虚拟机将变为集群的一部分。DRS 可能会建议将部分虚拟机迁移到刚添加到集群的主机。如果没有发生上述行为，则可能是 vMotion、主机兼容性或关联性规则存在问题。以下是可能的原因：

- 此主机未配置或未激活 vMotion。
- 其他主机上的虚拟机与此主机不兼容。
- 主机没有足够的资源用于任何虚拟机。
- 向该主机中移动任何虚拟机都会违反虚拟机/虚拟机 DRS 规则或虚拟机/主机 DRS 规则。
- 此主机为 HA 故障切换容量预留。
- 设备挂载到虚拟机。
- vMotion 阈值过高。
- 已为虚拟机停用 DRS，因此无法将该虚拟机移动到目标主机。

解决方案

请解决阻止 DRS 将虚拟机移动到主机上的问题。

DRS 没有从主机移动任何虚拟机

DRS 未从主机移动任何虚拟机。

问题

没有从该主机中移动任何虚拟机。

原因

这可能是由 vMotion、DRS 或主机兼容性问题导致的。以下是可能的原因：

- 此主机未配置或未激活 vMotion。
- 为此主机上的虚拟机停用了 DRS。
- 此主机上的虚拟机与任何其他主机不兼容。
- 其他主机都没有足够的资源供该主机上的任何虚拟机使用。
- 从该主机中移动任何虚拟机都会违反虚拟机/虚拟机 DRS 规则或虚拟机/主机 DRS 规则。
- 为该主机上的一个或多个虚拟机停用了 DRS。
- 设备挂载到虚拟机。

解决方案

解决使 DRS 无法从主机移动虚拟机的问题。

虚拟机问题

虚拟机问题可能会导致 DRS 无法按预期方式执行。

CPU 或内存资源不足

虚拟机未收到足够的 CPU 或内存资源。

问题

在某些情况中，虚拟机的需求大于其资源授权。发生这种情况时，虚拟机不会收到足够的 CPU 或内存资源。

原因

以下各节介绍影响虚拟机授权的因素。

集群为黄色或红色

如果集群为黄色或红色，则容量不足以满足为集群中所有虚拟机和资源池配置的资源预留。特殊的虚拟机可能就是没有收到预留的那个虚拟机。检查（红色或黄色）集群的状态，并解决该情况。

资源限制过于严格

虚拟机、其父资源池或其资源池祖先可能有过于严格的配置资源限制。检查需求是否等于或高于所有配置的限制。

集群过载

正在运行虚拟机的集群可能资源不足。此外，相比该虚拟机的共享值，其他虚拟机被成比例地授予了更多资源。要确定需求是否大于容量，请检查集群统计信息。

主机过载

要确定是否超额预订了主机的资源，请检查主机统计信息。如果超额预订了它们，则考虑为什么 DRS 没有将主机上现在正在运行的虚拟机移动到其他主机。以下是可能存在这种状况的原因：

- 虚拟机/虚拟机 DRS 规则和虚拟机/主机 DRS 规则需要当前的“虚拟机到主机”映射。如果在集群中配置了这样的规则，则考虑停用其中的一个或多个。然后运行 DRS 并检查情况是否已更正。
- DRS 不能将此虚拟机或足够的其他虚拟机移动到其他主机以释放容量。DRS 不会移动虚拟机的原因包括以下几种：
 - 已为虚拟机禁用 DRS。
 - 主机设备已挂载到虚拟机。
 - 虚拟机的资源预留很大，以致虚拟机不能在集群中的任何其他主机上运行。
 - 虚拟机与集群中的任何其他主机不兼容。

检查是否存在上述虚拟机的问题。如果都不存在，则集群中的其他虚拟机可能存在问题。如果是这样，则 DRS 将无法均衡集群以满足虚拟机的需要。

- 请减小 DRS 迁移阈值设置并检查问题是否已解决。

- 增加虚拟机预留。

解决方案

解决导致虚拟机未收到足够 CPU 或内存资源的问题。

违反了虚拟机/虚拟机 DRS 规则或者虚拟机/主机 DRS 规则

DRS 规则指定虚拟机必须位于或不能位于哪台主机，或哪些虚拟机必须位于或不能位于同一主机。

问题

违反虚拟机/虚拟机 DRS 规则或虚拟机/主机 DRS 规则。

原因

虚拟机/虚拟机 DRS 规则指定选定的虚拟机应当放置在相同主机上（关联性），或指定虚拟机应当放置在不同主机上（反关联性）。虚拟机/主机 DRS 规则指定选定的虚拟机应当放置在特定的主机上（关联性），或者选定的虚拟机不应当放置在特定的主机上（反关联性）。

当违反了虚拟机/虚拟机 DRS 规则或虚拟机/主机 DRS 规则时，则可能是因为 DRS 而无法移动规则中的部分或全部虚拟机。关联性规则中的虚拟机或其他虚拟机的预留，或其父资源池可能阻止 DRS 查找同一主机上的所有虚拟机。

解决方案

- 检查 DRS 故障面板，查找与关联性规则相关的故障。
- 计算关联性规则中所有虚拟机的预留总数。如果该值大于任何主机的可用容量，则该规则将无法实现。
- 计算其父资源池的预留总数。如果该值大于任何主机的可用容量，并且资源是从单个主机获取的，则该规则将无法实现。

打开虚拟机电源操作失败

显示一条错误消息指出未能打开虚拟机电源。

问题

未能打开虚拟机电源。

原因

由于资源不足或虚拟机没有兼容的主机，未能打开虚拟机电源。

解决方案

如果集群没有足够的资源来打开单个虚拟机的电源，或组中的任何虚拟机尝试打开电源，请对照集群或其父资源池中的可用资源来检查虚拟机所需的资源。如有必要，减少要打开电源的虚拟机及其同级虚拟机的预留，或者增加集群或其父资源池中的可用资源。

DRS 没有移动虚拟机

最初打开虚拟机电源时，尽管主机上资源不足，DRS 也不会移动虚拟机。

问题

打开虚拟机电源时，如果虚拟机注册到的主机上资源不足，DRS 不会正常迁移虚拟机。

原因

DRS 不移动虚拟机的可能原因如下。

- 在此虚拟机上停用了 DRS。
- 虚拟机已挂载设备。
- 虚拟机与其他任何主机都不兼容。
- 其他主机都没有足够数量的物理 CPU 或容量供此虚拟机的每个 CPU 使用。
- 其他主机都没有足够的 CPU 或内存资源可满足此虚拟机的预留和所需的内存。
- 移动此虚拟机将违反关联性或反关联性规则。
- 此虚拟机的 DRS 自动化级别为手动，并且用户没有批准迁移建议。
- DRS 将不移动已激活容错的虚拟机。

解决方案

请解决阻止 DRS 移动虚拟机的问题。