采用ONTAP的vSphere Metro存储集群

Enterprise applications

NetApp
May 03, 2024

目录

采用ONTAP的vSphere Metro存储集群

采用ONTAP的vSphere Metro存储集群

VMware行业领先的vSphere虚拟机管理程序可部署为延伸型集群、称为vSphere Metro Storage Cluster (VMSC)。

NetApp®MetroCluster™和SnapMirror主动同步(以前称为SnapMirror业务连续性或SMBC)均支持VMSC解决方案，如果一个或多个故障域发生完全中断，则可以提供高级业务连续性。不同故障模式的故障恢复能力取决于您选择的配置选项。

适用于vSphere环境的持续可用性解决方案

ONTAP架构是一个灵活且可扩展的存储平台、可为数据存储库提供SAN (FCP、iSCSI和NVMe-oF)和NAS (NFS v3和v4.1)服务。NetApp AFF、ASA和FAS存储系统使用ONTAP操作系统为子系统存储访问提供其他协议、例如S3和SMB/CCIFS。

NetApp MetroCluster使用NetApp的HA (控制器故障转移或CFO)功能来防止控制器发生故障。它还包括本地SyncMirror技术、灾难时集群故障转移(控制器按需故障转移或CFOD)、硬件冗余和地理分隔、以实现高可用性。SyncMirror通过将数据写入两个丛(本地磁盘架上主动提供数据的本地丛和通常不提供数据的远程丛)、在MetroCluster配置的两部分之间同步镜像数据。所有MetroCluster组件(例如控制器、存储、缆线、交换机(用于光纤MetroCluster)和适配器)都具有硬件冗余。

NetApp SnapMirror主动同步可通过FCP和iSCSI SAN协议提供数据存储库粒度保护、从而使您可以有选择地仅保护高优先级工作负载。与主动-备用解决方案NetApp MetroCluster不同、它可以同时对本地和远程站点进行主动-主动访问。目前、主动同步是一种非对称解决方案、其中一方优先于另一方、可提供更好的性能。这可通过ALOA (非对称逻辑单元访问)功能来实现、ALOA功能会自动通知ESXi主机首选控制器。但是、NetApp已宣布活动同步很快将启用完全对称访问。

要在两个站点之间创建VMware HA/DRS集群、需要使用vCenter Server Appliance (VCA)来管理ESXi主机。vSphere管理、vMotion®和虚拟机网络通过两个站点之间的冗余网络进行连接。管理HA/DRS集群的vCenter Server可以连接到两个站点上的ESXi主机、并且应使用vCenter HA进行配置。

请参见 "如何在vSphere Client中创建和配置集群" 配置vCenter HA。

您还应参考 "VMware vSphere Metro Storage Cluster 建议的实践"。

什么是vSphere Metro Storage Cluster?

vSphere Metro Storage Cluster (VMSC)是一种经过认证的配置、可保护虚拟机(VM)和容器免受故障的影响。这可以通过使用延伸型存储概念以及分布在不同故障域(例如机架、建筑物、园区甚至城市)中的ESXi主机集群来实现。NetApp MetroCluster和SnapMirror主动同步存储技术用于分别为主机集群提供RPO = 0或接近RPO = 0的保护。VMSC配置旨在确保数据始终可用、即使完整的物理或逻辑"站点"发生故障也是如此。在成功完成VMSC认证过程后、属于VMSC配置的存储设备必须经过认证。可在中找到所有受支持的存储设备 "《VMware存储兼容性指南》"。

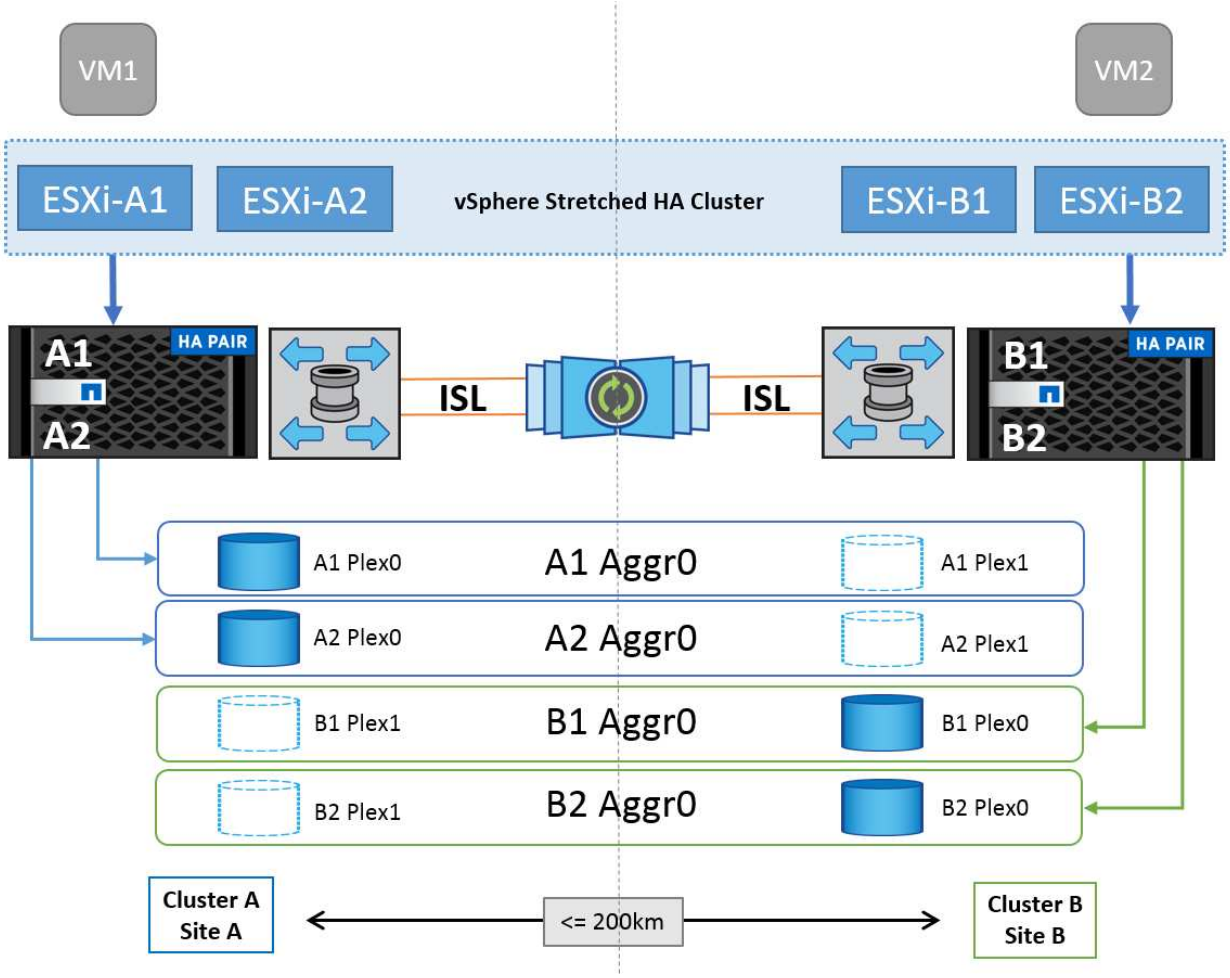
如果您需要有关vSphere Metro Storage Cluster设计准则的详细信息、请参阅以下文档：

- "NetApp MetroCluster支持VMware vSphere"
- "VMware vSphere支持NetApp SnapMirror业务连续性" (现在称为SnapMirror活动同步)

根据延迟注意事项、可以将NetApp MetroCluster部署在两种不同的配置中以用于vSphere：

- 延伸型MetroCluster
- 光纤MetroCluster

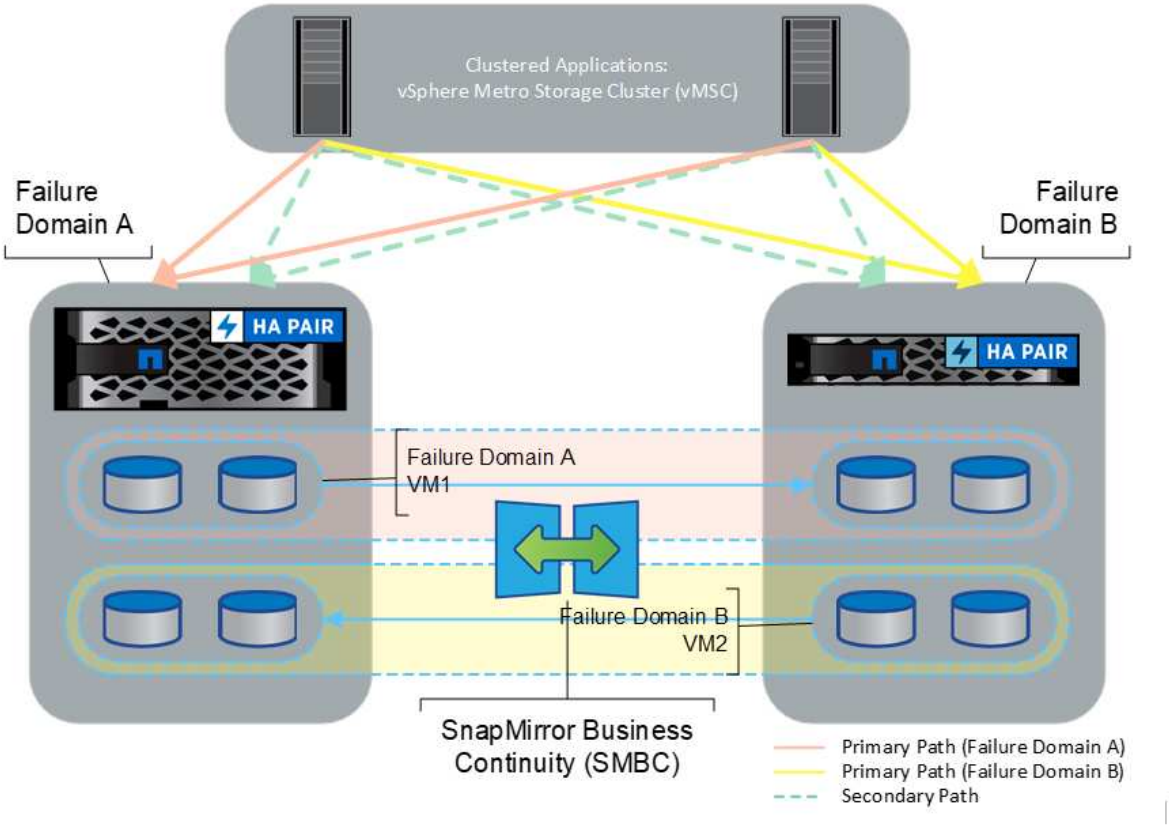
下图展示了延伸型MetroCluster的高层面拓扑图。



请参见 "MetroCluster 文档" 了解MetroCluster的特定设计和部署信息。

SnapMirror主动同步也可以通过两种不同的方式进行部署。

- 非对称
- 对称(ONTAP 9.14.1中的私有预览)



请参见 "NetApp文档" 有关SnapMirror活动同步的特定设计和部署信息、请参见。

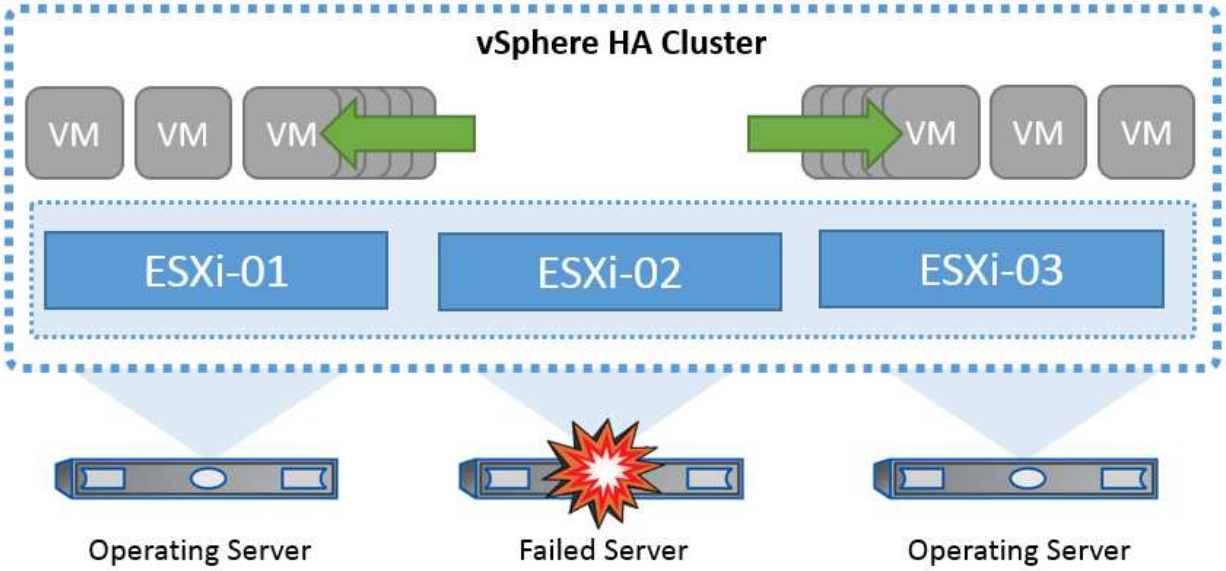
VMware vSphere解决方案概述

vCenter Server Appliance (VCCSA)是一款功能强大的集中式管理系统和适用于vSphere的单一管理平台、可使管理员有效地运行ESXi集群。它有助于实现关键功能、例如VM配置、vMotion操作、高可用性(HA)、分布式资源计划程序(DRS)、Tanzu Kubernetes Grid等。它是VMware云环境中的一个重要组件、在设计时应考虑服务可用性。

vSphere高可用性

VMware的集群技术可将ESXi服务器分组到虚拟机的共享资源池中、并提供vSphere High Availability (HA)。vSphere HA可为虚拟机中运行的应用程序提供易于使用的高可用性。在集群上启用HA功能后、每个ESXi服务器都会与其他主机保持通信、以便在任何ESXi主机无响应或隔离时、 HA集群可以在集群中的无故障主机之间协商恢复该ESXi主机上运行的虚拟机。如果子操作系统发生故障、vSphere HA会在同一物理服务器上重新启动受影响的虚拟机。借助vSphere HA、可以减少计划内停机、防止计划外停机并从中断中快速恢复。

vSphere HA集群从故障服务器中恢复VM。



请务必了解、VMware vSphere不了解NetApp MetroCluster或SnapMirror活动同步、并且会根据主机和VM组关联性配置、将vSphere集群中的所有ESXi主机视为符合HA集群操作条件的主机。

主机故障检测

创建HA集群后、集群中的所有主机都会参与选择、其中一个主机将成为主主机。每个从节点对主节点执行网络检测信号、而主节点则对所有从节点主机执行网络检测信号。vSphere HA集群的主主机负责检测从主机的故障。

根据检测到的故障类型、可能需要对主机上运行的虚拟机进行故障转移。

在vSphere HA集群中、检测到三种类型的主机故障：

- 故障—主机停止运行。
- 隔离—主机变为网络隔离。
- 分区-主机与主主机断开网络连接。

主主机监控集群中的从主机。此通信通过每秒交换一次网络检测信号来完成。当主主机停止从从从主机接收这些检测信号时、它会先检查主机活动性、然后再声明主机出现故障。主主机执行的活动性检查用于确定从主机是否正在与某个数据存储库交换检测搏。此外、主主机还会检查该主机是否对发送到其管理IP地址的ICMP ping做出响应、以检测它是仅与其主节点隔离还是与网络完全隔离。它通过对默认网关执行pinging来实现此目的。可以手动指定一个或多个隔离地址、以提高隔离验证的可靠性。

最佳实践

NetApp建议至少指定两个额外的隔离地址、并且每个地址都是站点本地地址。这将提高隔离验证的可靠性。

主机隔离响应

隔离响应是vSphere HA中的一项设置、用于确定当vSphere HA集群中的主机丢失其管理网络连接但仍继续运行时在虚拟机上触发的操作。此设置有三个选项："Disabled (禁用)"、"Shut Down and Restart VMs"(关闭并重新启动VM)和"Power Off and Restart VMs"(关闭并重新启动VM)。

"Shut down (关闭)"优于"Power Off (关闭)"、后者不会刷新最近对磁盘所做的更改或提交事务。如果虚拟机在300秒内未关闭、则会将其关闭。要更改等待时间、请使用高级选项as.isolationshutdowntimeout。

在HA启动隔离响应之前、它会首先检查vSphere HA主代理是否拥有包含VM配置文件的数据存储库。否则、主机将不会触发隔离响应、因为没有主节点可重新启动VM。主机将定期检查数据存储库状态、以确定是否由具有主角色的vSphere HA代理声明数据存储库。

最佳实践

NetApp建议将"主机隔离响应"设置为"已禁用"。

如果主机与vSphere HA主主机隔离或分区、并且主主机无法通过检测信号数据存储库或ping进行通信、则可能发生脑裂情况。主节点会声明隔离的主机已停止运行、并在集群中的其他主机上重新启动VM。现在存在脑裂情况、因为虚拟机有两个实例正在运行、其中只有一个实例可以读取或写入虚拟磁盘。现在、可以通过配置虚拟机组件保护(VM Component Protection、VMCP)来避免脑裂情况。

VM组件保护(VMCP)

vSphere 6中与HA相关的一项增强功能是VMCP。VMCP可针对块(FC、iSCSI、FCoE)和文件存储(NFS)提供增强的保护、使其免受所有路径关闭(APD)和永久设备丢失(PDL)情况的影响。

永久设备丢失(永久设备丢失)(财产和财产

如果存储设备永久出现故障或被管理员删除、并且不希望返回、则会出现上述情况。NetApp存储阵列向ESXi发出SCSI检测代码、声明设备已永久丢失。在vSphere HA的故障条件和VM响应部分中、您可以配置检测到无效条件后的响应。

最佳实践

NetApp建议将"Response for Data with PCL"(使用数据存储库的响应)设置为"关闭并重新启动**VMS**"。检测到这种情况后、VM将在vSphere HA集群中运行正常的主机上立即重新启动。

所有路径已关闭(APD)

APD是指主机无法访问存储设备且没有指向阵列的路径时发生的情况。ESXi认为此问题是设备的临时问题、并希望它能够再次可用。

检测到APD情况时、计时器将启动。140秒后、系统将正式声明APD条件、并且设备会标记为APD超时。超过140秒后、HA将开始计算VM故障转移APD延迟中指定的分钟数。指定时间过后、HA将重新启动受影响的虚拟机。您可以根据需要将VMCP配置为以不同方式响应("Disabled (已禁用)"、"VM Events (问题描述事件)"或"Power Off and Restart VM (关闭并重新启动VM)")。

最佳实践

NetApp建议将"Response for Data with APD"配置为"关闭并重新启动**VM (**保守**)**"。

保守是指HA能够重新启动VM的可能性。如果设置为保守、则只有在HA知道其他主机可以重新启动受APD影响的虚拟机时、它才会重新启动该虚拟机。如果发生主动、即使HA不知道其他主机的状态、也会尝试重新启动虚拟机。如果任何主机都无法访问虚拟机所在的数据存储库、则可能导致虚拟机无法重新启动。

如果APD状态为已解决、并且在超时之前还原了对存储的访问、则HA不会不必要地重新启动虚拟机、除非您明确对此虚拟机进行配置。如果即使环境已从APD条件中恢复、也需要响应、则应将APD超时后APD恢复的响应配置为重置VM。

NetApp建议将APD超时后APD恢复的响应配置为已禁用。

适用于NetApp MetroCluster的VMware DRS实施

VMware DRS是一项将主机资源聚合到集群中的功能、主要用于在虚拟基础架构中的集群内进行负载平衡。VMware DRS主要计算在集群中执行负载平衡所需的CPU和内存资源。由于vSphere无法识别延伸型集群、因此在进行负载平衡时、它会考虑两个站点中的所有主机。为了避免跨站点流量、NetApp建议配置DRS关联性规则、以管理VM的逻辑隔离。这样可以确保、除非完全发生站点故障、否则HA和DRS将仅使用本地主机。

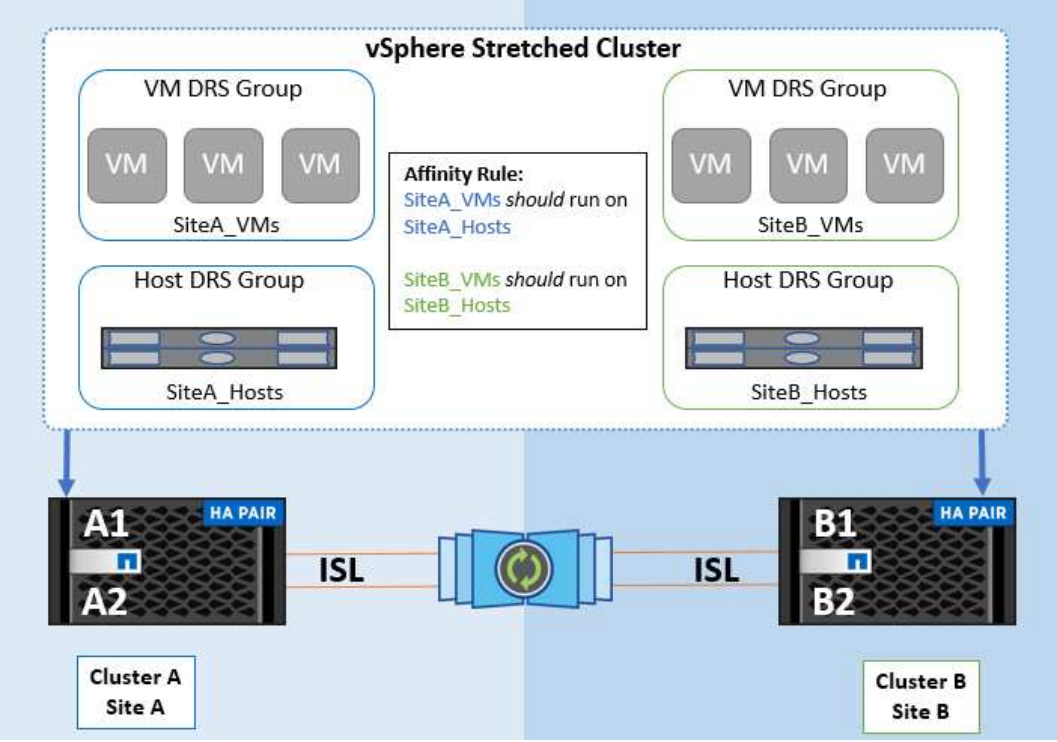
如果为集群创建DRS关联性规则、则可以指定vSphere在虚拟机故障转移期间如何应用该规则。

您可以通过两种类型的规则来指定vSphere HA故障转移行为：

- VM反关联性规则会强制指定的虚拟机在故障转移操作期间保持分离状态。
- 在故障转移操作期间、VM主机关联性规则会将指定的虚拟机放置在特定主机或已定义主机组的成员上。

使用VMware DRS中的VM主机关联性规则、可以在站点A和站点B之间进行逻辑隔离、以便VM与配置为给定数据存储库的主读/写控制器的阵列在同一站点的主机上运行。此外、VM主机关联性规则还可以使虚拟机保持在存储本地、从而确保在站点间发生网络故障时虚拟机连接。

以下是VM主机组和关联性规则的示例。



最佳实践

NetApp建议实施"应该"规则、而不是"必须"规则、因为如果发生故障、vSphere HA会违反这些规则。使用"必须"规则可能会导致服务中断。

服务的可用性应始终高于性能。如果完整数据中心发生故障、则"必须"规则必须从VM主机关联性组中选择主

机、并且当数据中心不可用时、虚拟机不会重新启动。

使用NetApp MetroCluster实施VMware存储DRS

通过VMware Storage DRS功能、可以将数据存储库聚合到一个单元中、并在超过存储I/O控制阈值时平衡虚拟机磁盘。

默认情况下、启用了存储DRS的DRS集群会启用存储I/O控制。通过存储I/O控制、管理员可以控制在I/O拥塞期间分配给虚拟机的存储I/O量、这样、在分配I/O资源时、更重要的虚拟机就可以优先于不太重要的虚拟机。

存储DRS使用Storage vMotion将虚拟机迁移到数据存储库集群中的不同数据存储库。在NetApp MetroCluster环境中、需要在该站点的数据存储库中控制虚拟机迁移。例如、在站点A的主机上运行的虚拟机A最好在站点A的SVM数据存储库中进行迁移否则、虚拟机将继续运行、但性能会下降、因为虚拟磁盘读/写操作将通过站点间链路从站点B进行。

最佳实践

NetApp建议创建与存储站点关联性相关的数据存储库集群；也就是说、与站点A具有站点关联性的数据存储库不应与与与站点B具有站点关联性的数据存储库的数据存储库集群混用

每当使用Storage vMotion新配置或迁移虚拟机时、NetApp建议相应地手动更新特定于这些虚拟机的所有VMware DRS规则。这样可以确定主机和数据存储库在站点级别的虚拟机关联性、从而降低网络和存储开销。

《VMSC设计和实施准则》

本文档概述了使用ONTAP存储系统的VMSC的设计和实施指南。

NetApp存储配置

有关NetApp MetroCluster (称为MCC配置)的设置说明、请参见 "MetroCluster 文档"。有关SnapMirror活动同步的说明、请参见 "SnapMirror 业务连续性概述"。

配置MetroCluster后、对其进行管理就像管理传统ONTAP环境一样。您可以使用命令行界面(CLI)、System Manager或Ans得 等各种工具设置Storage Virtual Machine (SVM)。配置SVM后、在集群上创建要用于正常操作的逻辑接口(Logical Interface、Li)、卷和逻辑单元号(Logical Unit Number、LUN)。这些对象将自动通过集群对等网络复制到另一个集群。

如果不使用MetroCluster、则可以使用SnapMirror主动同步、它可以在不同故障域的多个ONTAP集群之间提供数据存储库粒度保护和主动-主动访问。SnapMirror主动同步使用一致性组来确保一个或多个数据存储库之间的写入顺序一致性、您可以根据应用程序和数据存储库要求创建多个一致性组。对于需要在多个数据存储库之间同步数据的应用程序、一致性组尤其有用。SnapMirror主动同步还支持原始设备映射(Raw Device Mapping、RDM)以及子系统连接的存储(具有子系统内iSCSI启动程序)。有关一致性组的详细信息、请参见 "一致性组概述"。

与MetroCluster相比、使用SnapMirror活动同步管理VMSC配置有一些不同。首先、这是一种仅SAN配置、任何NFS数据存储库都无法通过SnapMirror主动同步进行保护。其次、您必须将两个LUN副本映射到ESXi主机、以使其能够访问这两个故障域中复制的数据存储库。

VMware vSphere HA

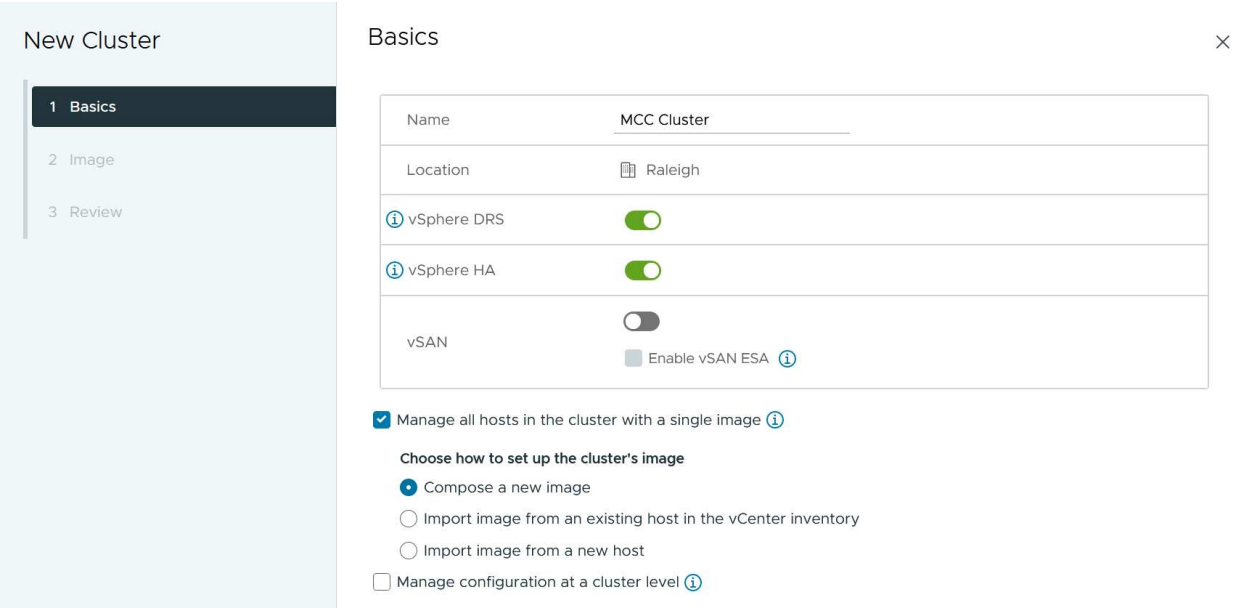
创建**vSphere HA**集群

创建vSphere HA集群是一个多步骤过程、有关详细信息、请参见 "如何在docs.vmware.com上的vSphere Client中创建和配置集群"。简而言之、您必须先创建一个空集群、然后使用vCenter添加主机并指定集群的vSphere HA和其他设置。

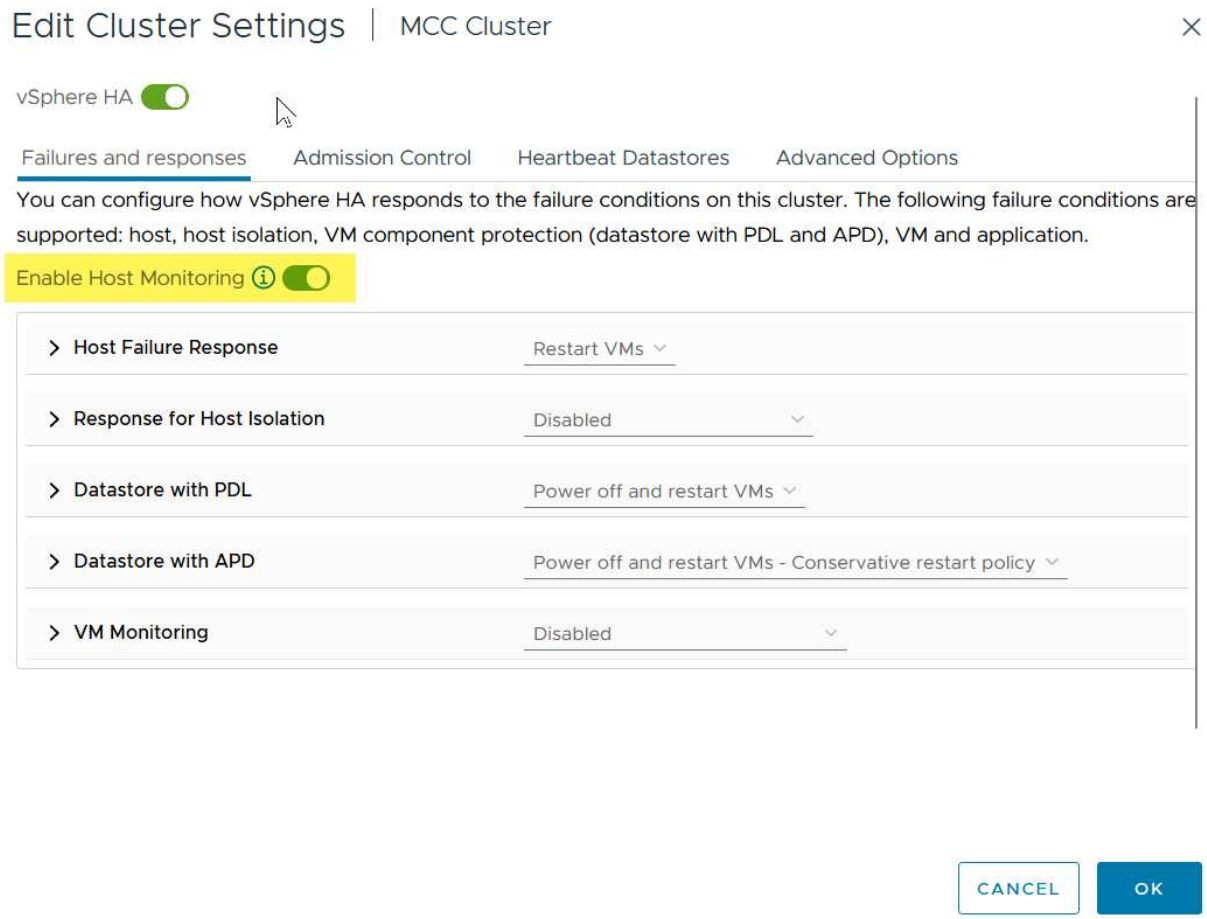
*注：*本文档中的任何内容均不可取代 "VMware vSphere Metro Storage Cluster 建议的实践"

要配置HA集群、请完成以下步骤：

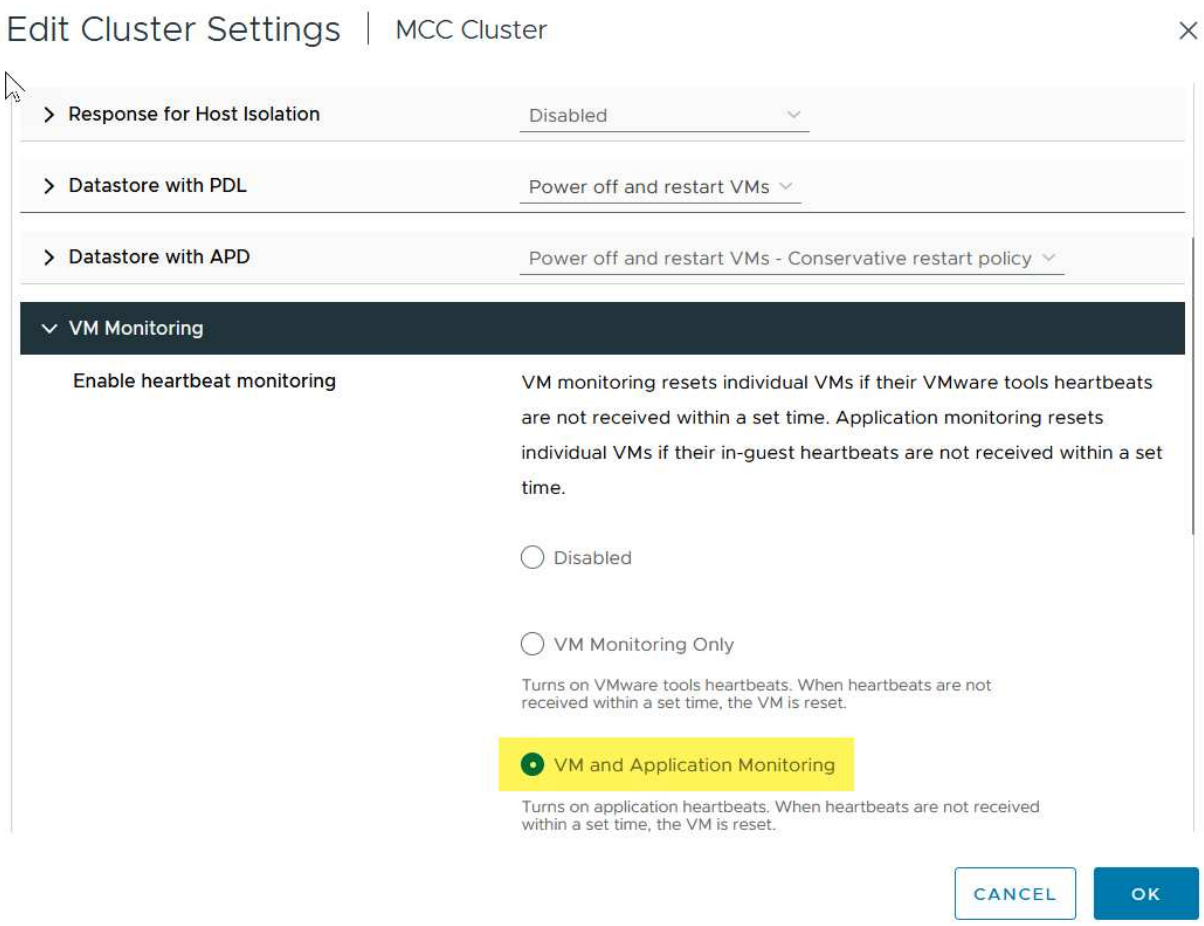
1. 连接到vCenter UI。
2. 在主机和集群中、浏览到要创建HA集群的数据中心。
3. 右键单击数据中心对象、然后选择"New Cluster"(新建集群)。在基础下、确保已启用vSphere DRS和vSphere HA。完成向导。



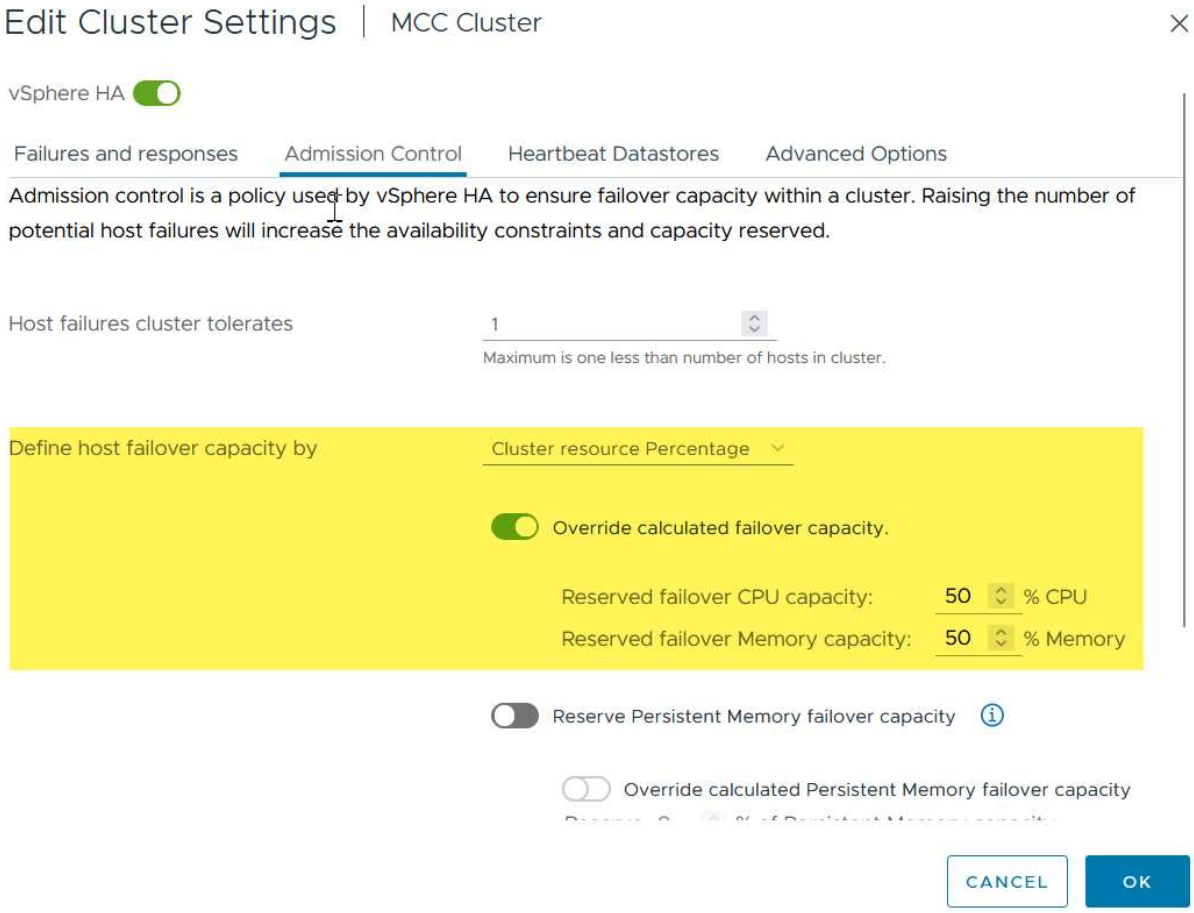
1. 选择集群并转到配置选项卡。选择vSphere HA、然后单击编辑。
2. 在"Host Monitoring"(主机监控)下、选择"Enable Host Monitoring"(启用主机监控)选项。



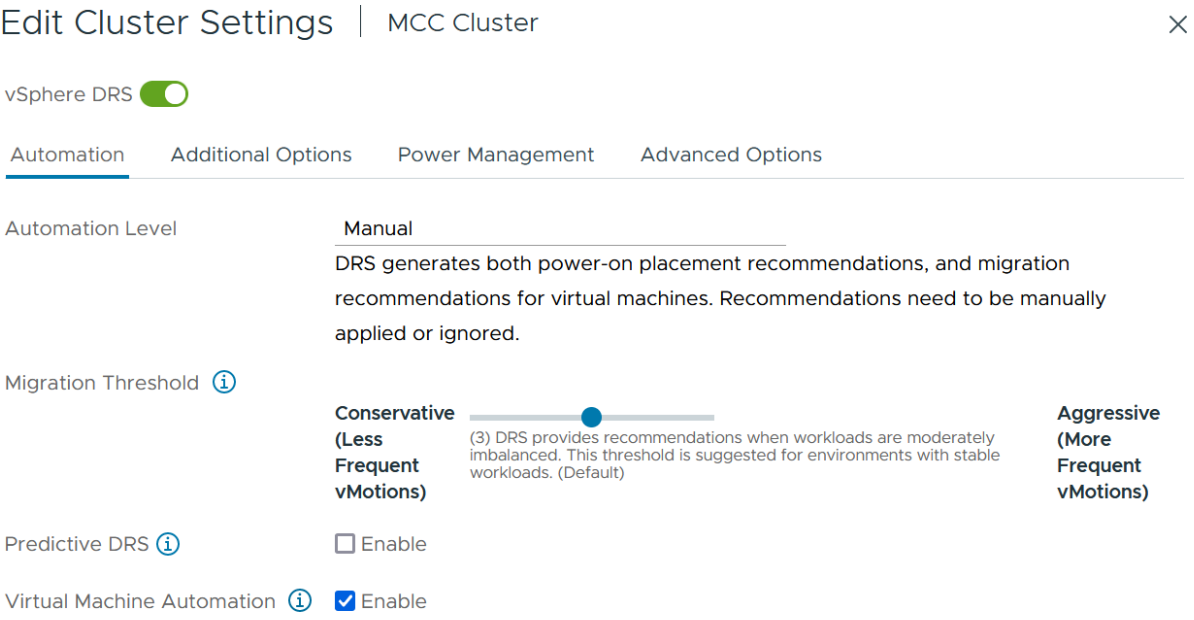
1. 仍在故障和响应选项卡上的VM监控下、选择仅VM监控选项或VM和应用程序监控选项。



1. 在"Admission Control"(准入控制)下、将HA准入控制选项设置为"Cluster Resource resource"(集群资源预留)；使用50% CPU/MEM。



1. 单击"OK"(确定)。
2. 选择DRS并单击编辑。
3. 除非您的应用程序要求、否则请将自动化级别设置为手动。



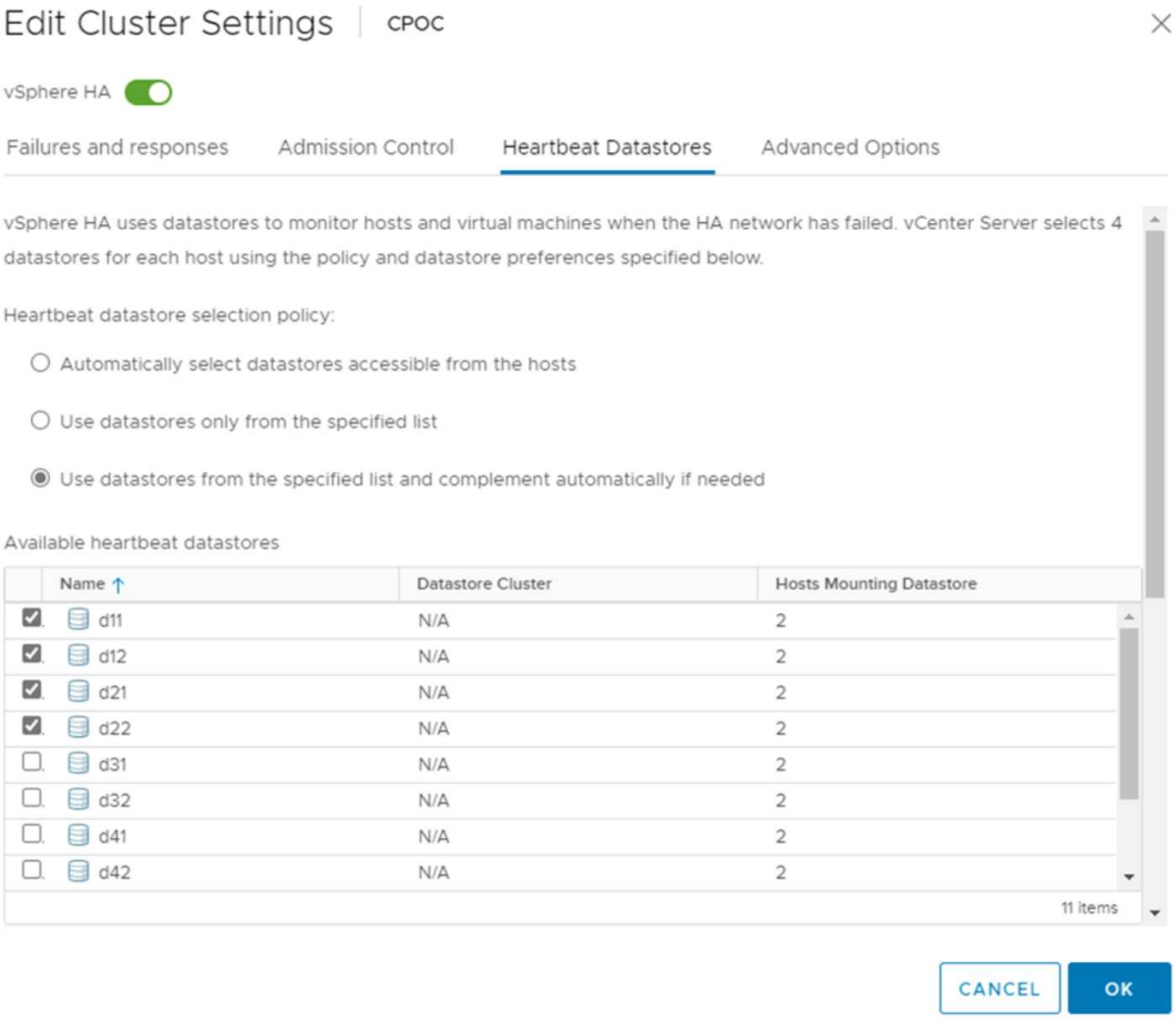
1. 启用VM组件保护、请参见 "docs.vmware.com"。
2. 对于采用MCC的VMSC、建议使用以下附加vSphere HA设置：

失败	响应
主机故障	重新启动VM
主机隔离	已禁用
具有永久设备丢失(永久设备丢失)的数据存储库	关闭并重新启动VM
所有路径均已关闭的数据存储库(APD)	关闭并重新启动VM
子系统不检测信号	重置虚拟机
VM重新启动策略	由虚拟机的重要性决定
主机隔离响应	关闭并重新启动VM
对使用了基于数据存储库的数据存储库的响应	关闭并重新启动VM
使用APD响应数据存储库	关闭并重新启动VM (保守)
APD的VM故障转移延迟	3分钟
响应APD恢复并显示APD超时	已禁用
VM监控敏感度	预设为高

配置用于检测信号的存储库

当管理网络出现故障时、vSphere HA使用数据存储库监控主机和虚拟机。您可以配置vCenter选择检测信号数据存储库的方式。要为数据存储库配置检测信号、请完成以下步骤：

1. 在数据存储库检测信号部分中、选择使用指定列表中的数据存储库并根据需要自动完成。
2. 从两个站点中选择要vCenter使用的数据存储库、然后按OK。

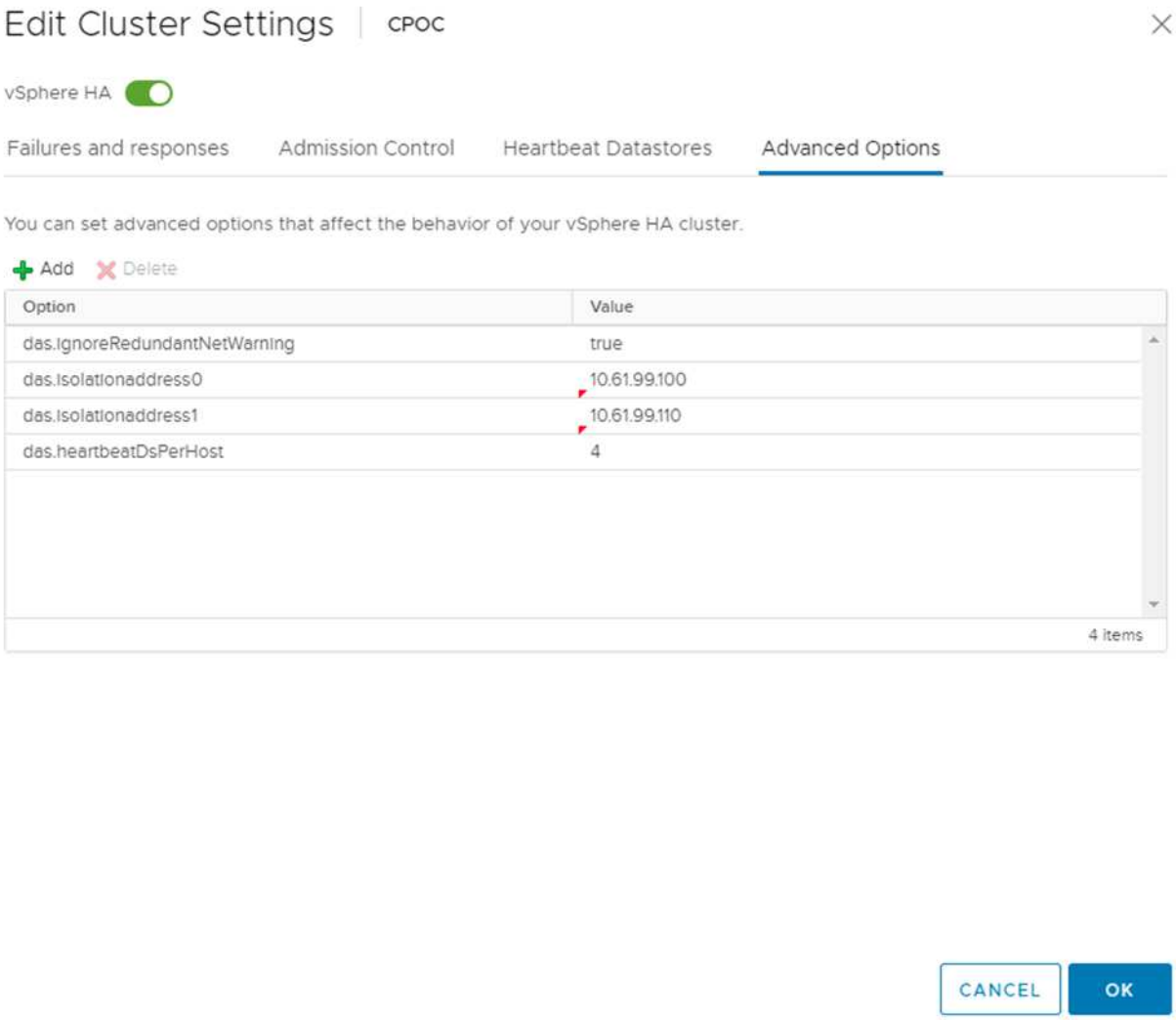


配置高级选项

主机故障检测

如果HA集群中的主机与网络或集群中的其他主机断开连接、则会发生隔离事件。默认情况下、vSphere HA将使用其管理网络的默认网关作为默认隔离地址。但是、您可以为要执行ping操作的主机指定其他隔离地址、以确定是否应触发隔离响应。添加两个可执行ping操作的隔离IP、每个站点一个。请勿使用网关IP。使用的vSphere HA高级设置为"as.isolationaddress"。为此、您可以使用ONTAP或调解器IP地址。

请参见 "core.vmware.com" 有关详细信息.



添加名为ds.heartbeatDsPerHost的高级设置可以增加检测信号数据存储库的数量。使用四个检测信号数据存储库(HB DSS)—每个站点两个。使用“从列表中选择但恭维”选项。这是必需的、因为如果一个站点发生故障、您仍需要两个HB DSS。但是、这些数据不必通过MCC或SnapMirror主动同步进行保护。

请参见 "core.vmware.com" 有关详细信息.

适用于NetApp MetroCluster的VMware DRS关联

在本节中、我们将为MetroCluster环境中每个站点\集群的VM和主机创建DRS组。然后、我们配置VM\Host规则、使VM主机与本地存储资源的关联性保持一致。例如、站点A的VM属于VM组sitea_vm、站点A的主机属于主机组sitea_hosts。接下来、在VM\Host规则中、我们说明site_vm应在sitea_hosts中的主机上运行。

最佳实践

- NetApp强烈建议使用规范“*应在组中的主机上运行”，而不是规范“必须在组中的主机上运行”。如果站点A主机发生故障、则需要通过vSphere HA在站点B的主机上重新启动站点A的VM、但后一种规范不允许HA重新启动站点B上的VM、因为这是一条硬规则。前一种规范是一种软规则、在发生HA时会违反该规范、从而实现可用性而非性能。

*注意：*您可以创建基于事件的警报，当虚拟机违反VM-主机关联性规则时触发该警报。在vSphere Client中、为虚拟机添加新警报、然后选择"VM is violating VM-Host Affinity Rule "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**DRS**主机组

要创建特定于站点A和站点B的DRS主机组、请完成以下步骤：

1. 在vSphere Web Client中、右键单击清单中的集群、然后选择设置。
2. 单击VM\Host Groups。
3. 单击添加。
4. 键入组的名称(例如、sitea_hosts)。
5. 从类型菜单中、选择主机组。
6. 单击Add、然后从站点A中选择所需主机、然后单击OK。
7. 重复上述步骤、为站点B添加另一个主机组
8. 单击确定。

创建**DRS VM**组

要创建特定于站点A和站点B的DRS VM组、请完成以下步骤：

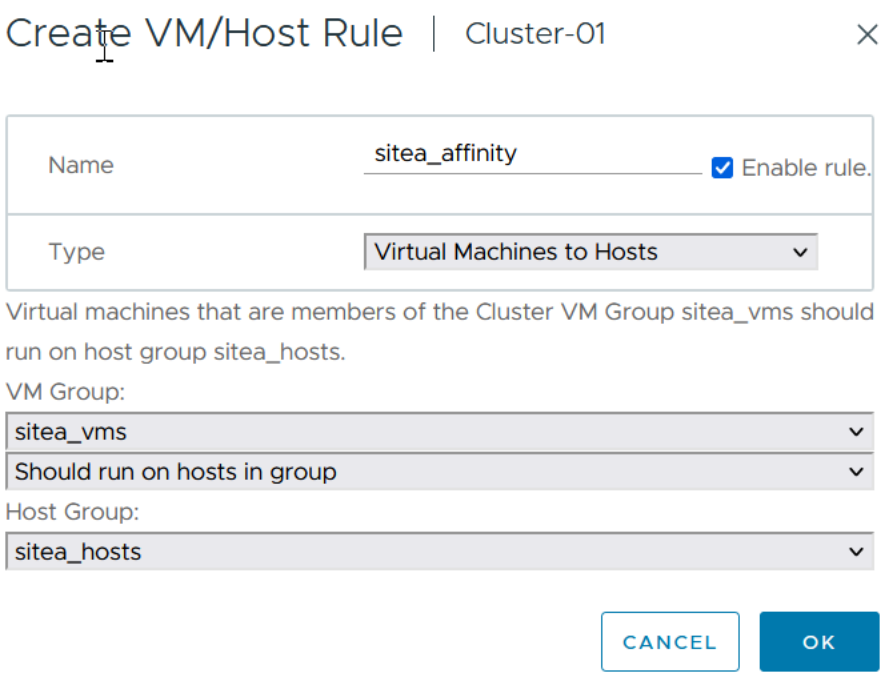
1. 在vSphere Web Client中、右键单击清单中的集群、然后选择设置。
2. 单击VM\Host Groups。
3. 单击添加。
4. 键入组的名称(例如、sitea_VMs.)。
5. 从Type菜单中、选择VM Group。
6. 单击添加并从站点A选择所需的VM、然后单击确定。
7. 重复上述步骤、为站点B添加另一个主机组
8. 单击确定。

创建**VM**主机规则

要创建特定于站点A和站点B的DRS相关性规则、请完成以下步骤：

1. 在vSphere Web Client中、右键单击清单中的集群、然后选择设置。
2. 单击VM\Host Rule。
3. 单击添加。
4. 键入规则的名称(例如、sitea_affinity)。
5. 验证是否已选中"Enable Rule (启用规则)"选项。
6. 从类型菜单中、选择虚拟机到主机。
7. 选择VM组(例如、sitea_vm)。
8. 选择主机组(例如、sitea_hosts)。

9. 重复上述步骤、为站点B添加另一个VM\Host规则
10. 单击确定。



适用于NetApp MetroCluster的VMware vSphere存储DRS

创建数据存储库集群

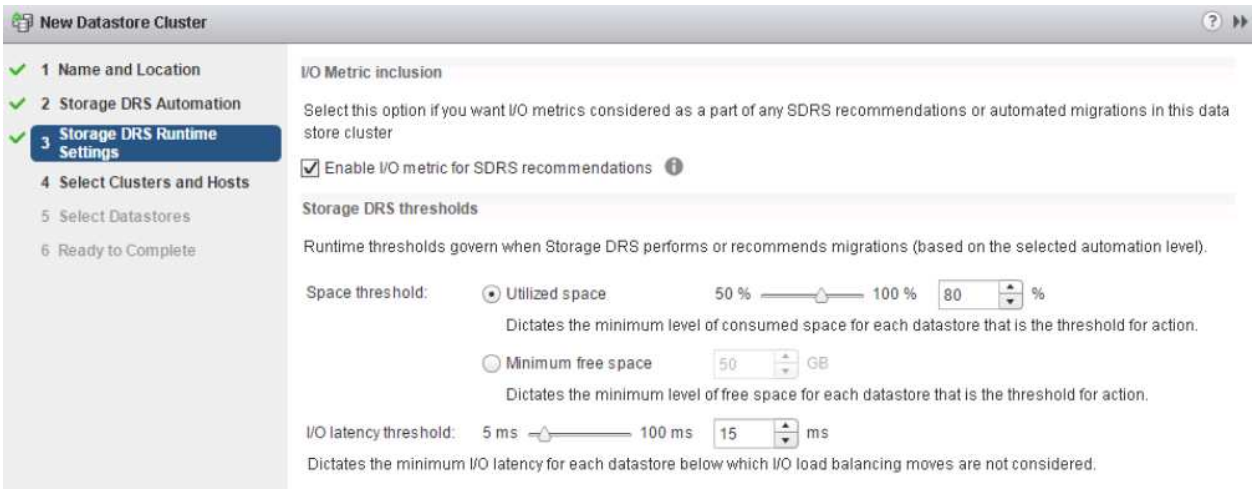
要为每个站点配置数据存储库集群、请完成以下步骤：

1. 使用vSphere Web Client、浏览到"Storage"(存储)下HA集群所在的数据中心。
2. 右键单击数据中心对象、然后选择"Storage"(存储)>"New Datastore Cluster"(新建数据存储库集群)。
3. 选择"Turn on Storage DRs"(打开存储DRS)选项、然后单击"Next"(下一步)。
4. 将所有选项设置为无自动化(手动模式)、然后单击下一步。

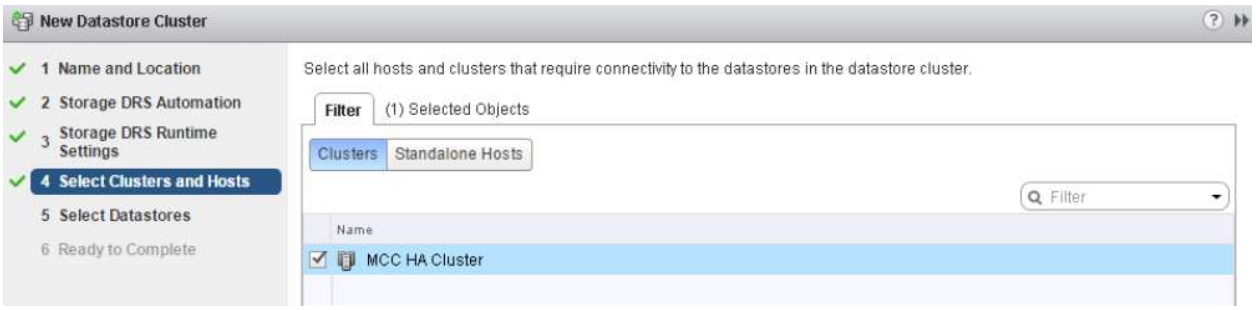
最佳实践

- NetApp建议在手动模式下配置存储DRS、以便管理员能够决定和控制何时需要进行迁移。

1. 验证是否已选中为SDRS建议启用I/O指标复选框；指标设置可以保留默认值。



1. 选择HA集群、然后单击"Next"(下一步)。



1. 选择属于站点A的数据存储库、然后单击下一步。

1. 查看选项、然后单击完成。
2. 重复上述步骤以创建站点B数据存储库集群、并验证是否仅选择了站点B的数据存储库。

vCenter Server可用性

您的vCenter Server设备(VCSA)应通过vCenter HA进行保护。通过vCenter HA、您可以在一个主动-被动HA对中部署两个VCSA。每个故障域一个。您可以在上阅读有关vCenter HA的更多信息 "docs.vmware.com"。

计划内和计划外事件的故障恢复能力

NetApp MetroCluster和SnapMirror主动同步是功能强大的工具，可提高NetApp硬件和ONTAP®软件的高可用性和无中断运行。

这些工具可为整个存储环境提供站点范围的保护、确保数据始终可用。无论您使用的是独立服务器、高可用性服

务器集群、Docker容器还是虚拟化服务器、NetApp技术都可以在因断电、散热或网络连接中断、存储阵列关闭或操作错误而导致全面中断时无缝保持存储可用性。

MetroCluster和SnapMirror主动同步提供了三种在发生计划内或计划外事件时保持数据连续性的基本方法：

- 冗余组件、用于防止出现单组件故障
- 影响单个控制器的事件的本地HA接管
- 全面的站点保护—通过将存储和客户端访问从源集群移动到目标集群、快速恢复服务

这意味着在单个组件发生故障时可以无缝地继续运行、并在更换故障组件后自动恢复为冗余操作。

除单节点集群(通常为软件定义的版本、例如ONTAP Select)之外、所有ONTAP集群都具有称为接管和交还的内置HA功能。集群中的每个控制器都会与另一个控制器配对、形成一个HA对。这些对可确保每个节点在本地连接到存储。

接管是一个自动化过程、其中一个节点接管另一个节点的存储以维护数据服务。相反、恢复过程会恢复正常操作。接管可以是计划内的(例如在执行硬件维护或ONTAP升级时)、也可以是计划外的(因节点崩溃或硬件故障而导致)。

在接管期间、MetroCluster配置中的网络连接存储逻辑接口(NAS LUN)会自动进行故障转移。但是、存储区域网络Lifs (SAN Lifs)不会进行故障转移；它们将继续使用逻辑单元号(Logical Unit Number、LUN)的直接路径。

有关HA接管和交还的详细信息、请参见 "HA对管理概述"。值得注意的是、此功能并不特定于MetroCluster或SnapMirror活动同步。

如果一个站点脱机或作为计划内活动进行站点范围维护、则会使用MetroCluster进行站点切换。另一个站点接管脱机集群的存储资源(磁盘和聚合)的所有权、故障站点上的SVM将联机并在灾难站点上重新启动、从而保留其完整身份、以供客户端和主机访问。

使用SnapMirror主动同步时、由于两个副本会同时使用、因此现有主机将继续运行。要确保正确进行站点故障转移、需要使用NetApp调解器。

使用MCC的VMSC的故障情形

以下各节概述了VMSC和NetApp MetroCluster系统的各种故障情形的预期结果。

单个存储路径故障

在这种情况下、如果组件(例如HBA端口、网络端口、前端数据交换机端口或FC或以太网缆线)发生故障、ESXi主机会将存储设备的特定路径标记为无活动。如果通过在HB/网络/交换机端口提供故障恢复能力来为存储设备配置多个路径、则ESXi最好执行路径切换。在此期间、虚拟机将保持运行状态、而不会受到影响、因为通过提供存储设备的多个路径、可以确保存储可用性。

*注意：*在这种情况下、MetroCluster行为没有变化、所有数据存储库在其各自的站点中仍保持完好。

最佳实践

在使用NFS/iSCSI卷的环境中、NetApp建议为标准vSwitch中的NFS vmkernel端口至少配置两个网络上行链路、并且在为分布式vSwitch映射NFS vmkernel接口的端口组上配置相同的网络上行链路。NIC绑定可以配置为主动-主动或主动-备用。

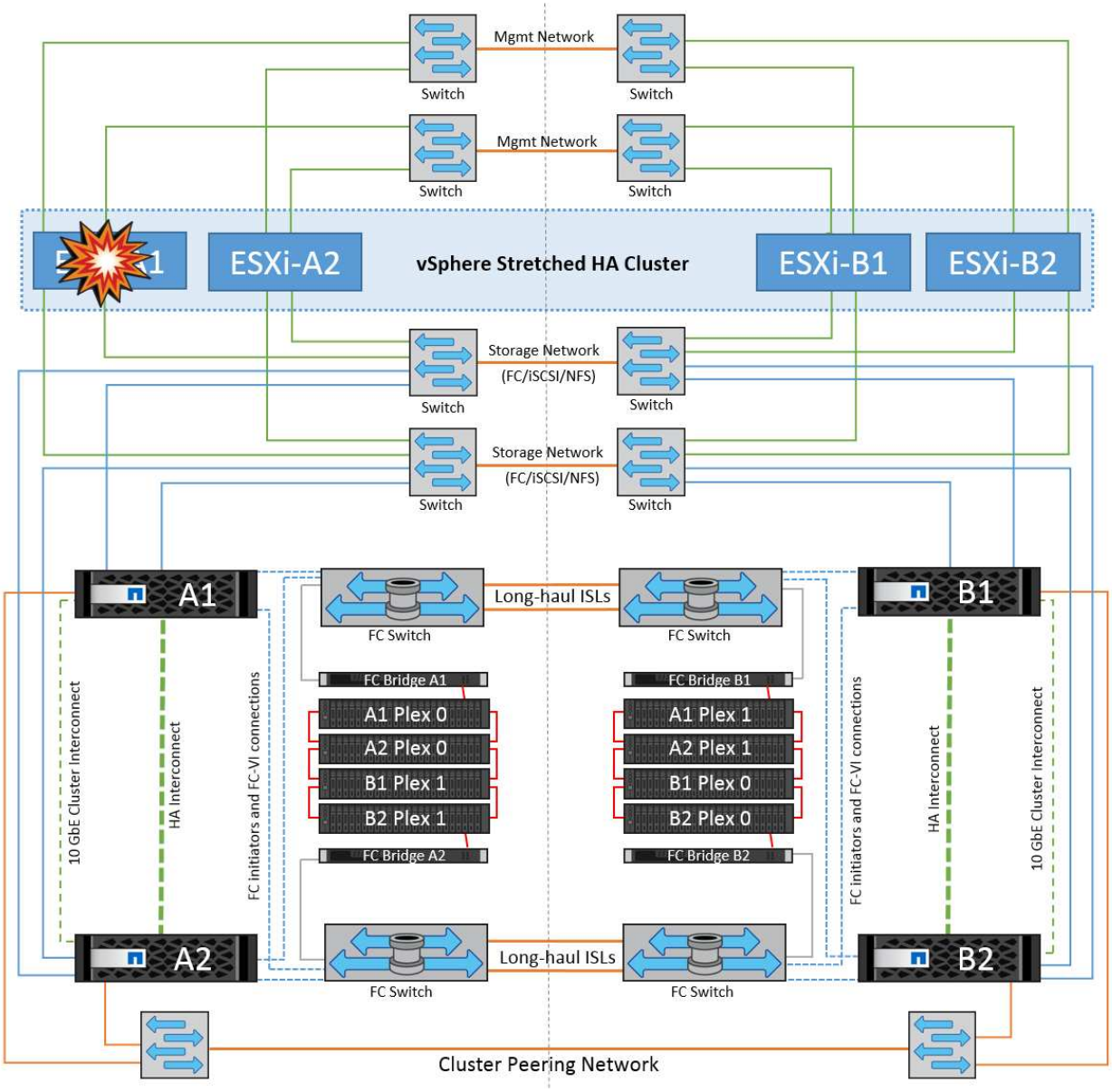
此外、对于iSCSI LUN、必须通过将vmkernel接口绑定到iSCSI网络适配器来配置多路径。有关详细信息、请参阅vSphere存储文档。

最佳实践

在使用光纤通道LUN的环境中、NetApp建议至少配置两个HBA、以确保HHBA/Port级别的故障恢复能力。NetApp还建议配置分区的最佳做法是、为单个目标分区配置单个启动程序。

应使用虚拟存储控制台(VSC)设置多路径策略、因为它会为所有新的和现有的NetApp存储设备设置策略。

单个**ESXi**主机故障



在这种情况下、如果ESXi主机发生故障、VMware HA集群中的主节点会检测到主机故障、因为它不再接收网络检测信号。为了确定主机是否确实已关闭或仅为网络分区、主节点会监控数据存储库检测点、如果没有检测点、它会对故障主机的管理IP地址执行屏显操作来执行最终检查。如果所有这些检查均为否定、则主节点会将此主机声明为故障主机、并且在此故障主机上运行的所有虚拟机都会在集群中的无故障主机上重新启动。

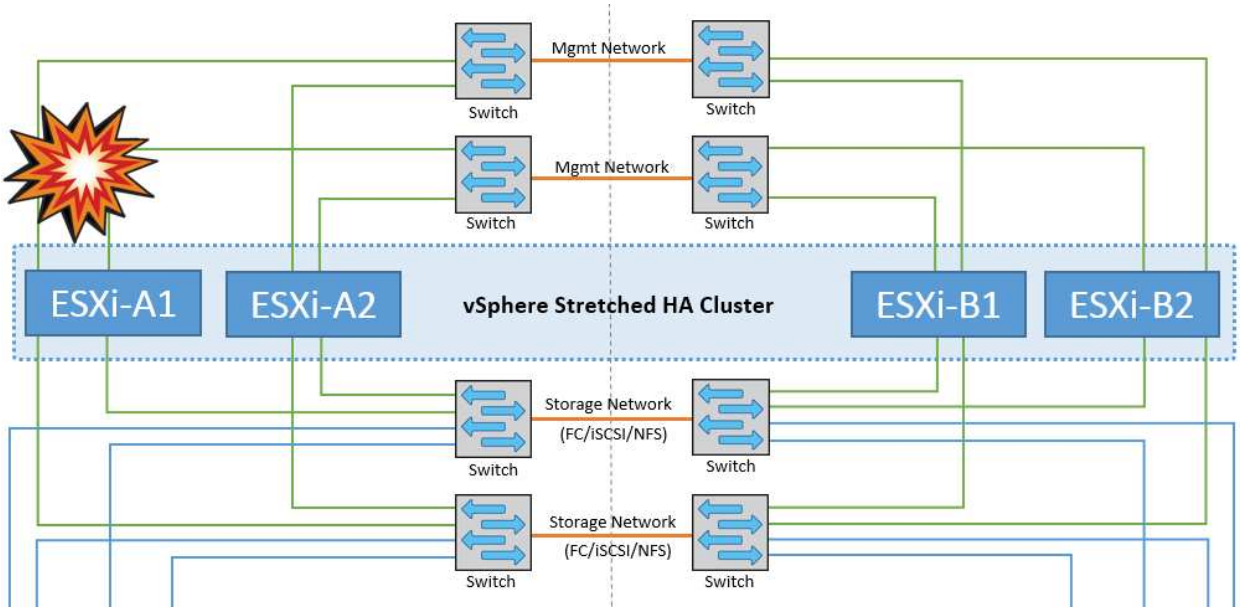
如果已配置DRS VM和主机关联性规则(VM组site_VMs中的VM应运行主机组site_hosts中的主机)、则HA主节点

会首先检查站点A上的可用资源如果站点A上没有可用主机、主节点将尝试重新启动站点B主机上的VM

如果本地站点存在资源限制、则虚拟机可能会在另一站点的ESXi主机上启动。但是、如果将虚拟机迁移回本地站点中任何无故障的ESXi主机时违反了任何规则、则定义的DRS VM和主机关联性规则将会进行更正。如果DRS设置为手动、则NetApp建议调用DRS并应用建议以更正虚拟机放置。

在这种情况下、MetroCluster的行为没有变化、所有数据存储库在其各自的站点中仍保持完好。

ESXi主机隔离

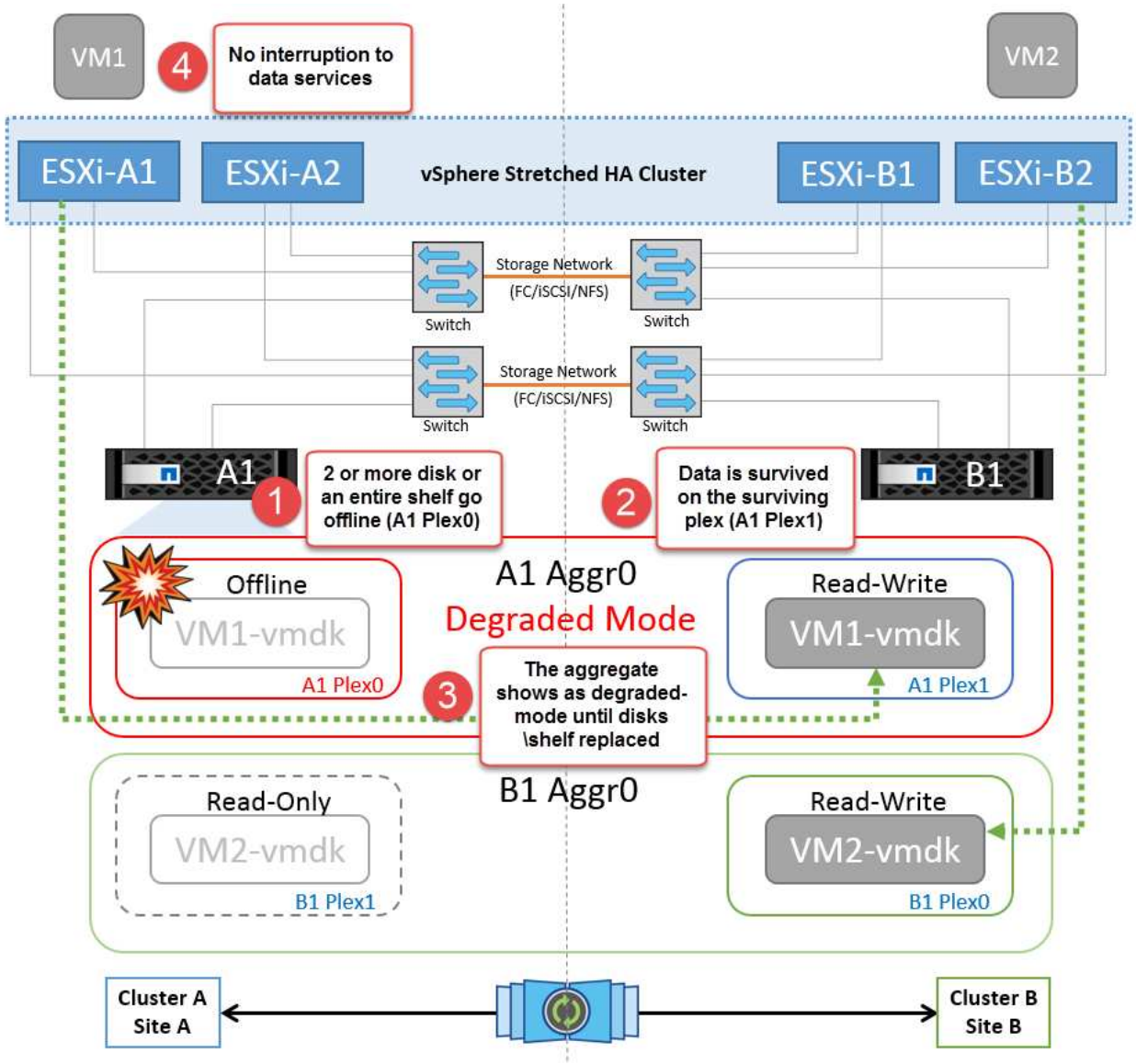


在这种情况下、如果ESXi主机的管理网络关闭、HA集群中的主节点将不会收到任何检测信号、因此此主机将在网络中隔离。要确定数据存储库是发生故障还是仅被隔离、主节点会开始监控数据存储库检测信号。如果存在、则主节点会声明主机已隔离。根据配置的隔离响应、主机可以选择关闭电源、关闭虚拟机、甚至保持虚拟机处于打开状态。隔离响应的默认间隔为30秒。

在这种情况下、MetroCluster的行为没有变化、所有数据存储库在其各自的站点中仍保持完好。

磁盘架故障

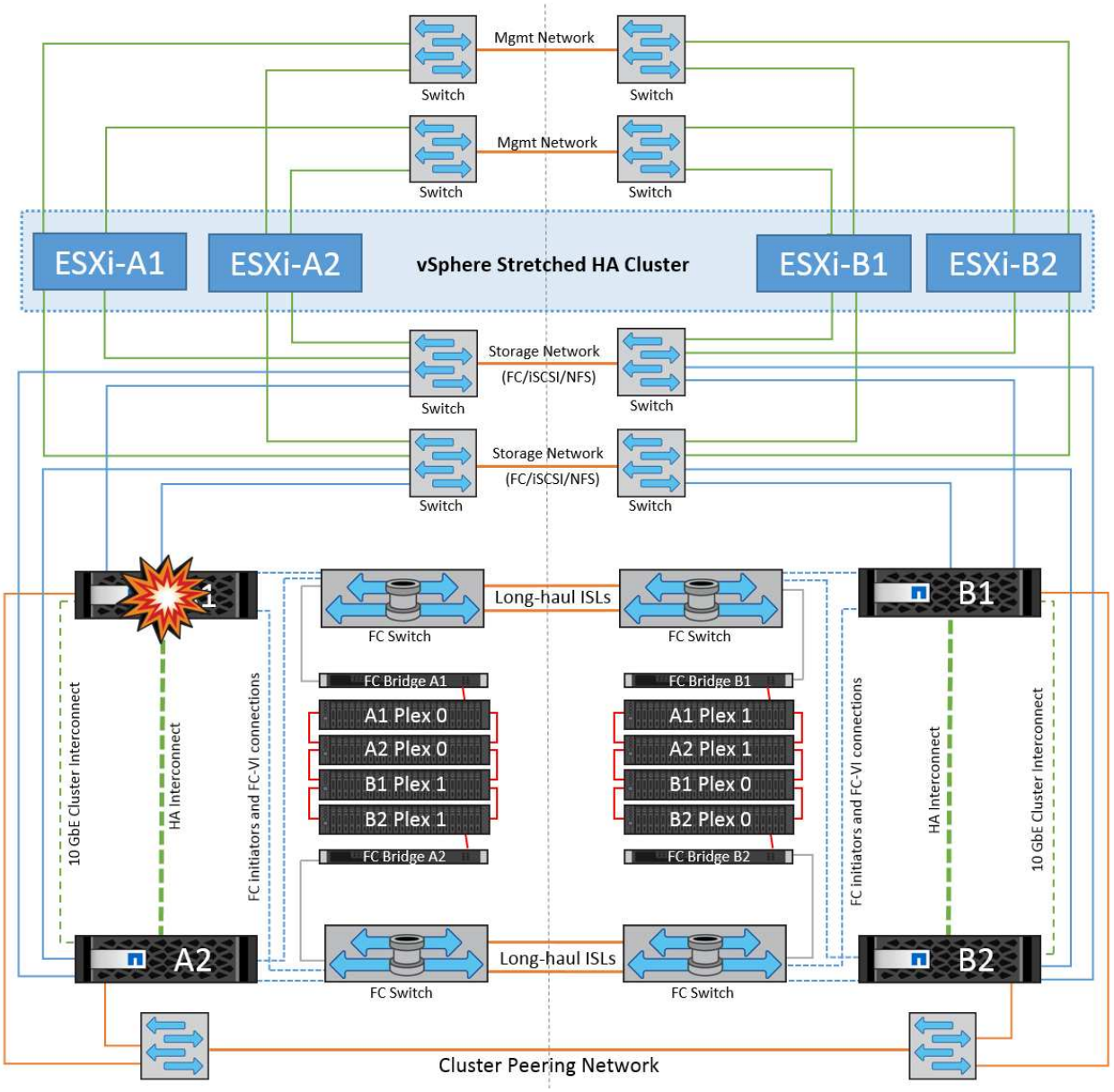
在这种情况下、出现两个以上磁盘或整个磁盘架故障。数据从无故障丛提供、而不会中断数据服务。磁盘故障可能会影响本地丛或远程丛。聚合将显示为降级模式、因为只有一个丛处于活动状态。更换故障磁盘后、受影响的聚合将自动重新同步以重建数据。重新同步后、聚合将自动恢复为正常镜像模式。如果一个RAID组中有两个以上的磁盘出现故障、则必须从头开始重建丛。



*注意：*在此期间、虚拟机I/O操作不会受到影响、但性能会下降、因为数据是通过ISL链路从远程磁盘架访问的。

单个存储控制器故障

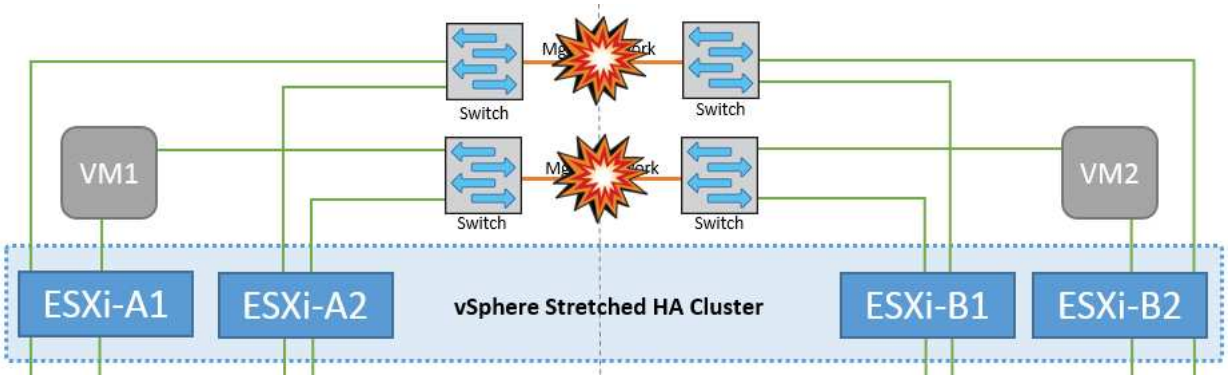
在这种情况下、一个站点上的两个存储控制器之一发生故障。由于每个站点都有一个HA对、因此一个节点发生故障会透明地自动触发故障转移到另一个节点。例如、如果节点A1发生故障、其存储和工作负载将自动传输到节点A2。虚拟机不会受到影响、因为所有的plexes都保持可用。第二个站点节点(B1和B2)不受影响。此外、vSphere HA不会执行任何操作、因为集群中的主节点仍将接收网络检测信号。



如果故障转移是滚动灾难的一部分(节点A1故障转移到A2)、则在后续发生A2故障或站点A完全故障时、站点B可能会发生灾难后切换

交换机间链路故障

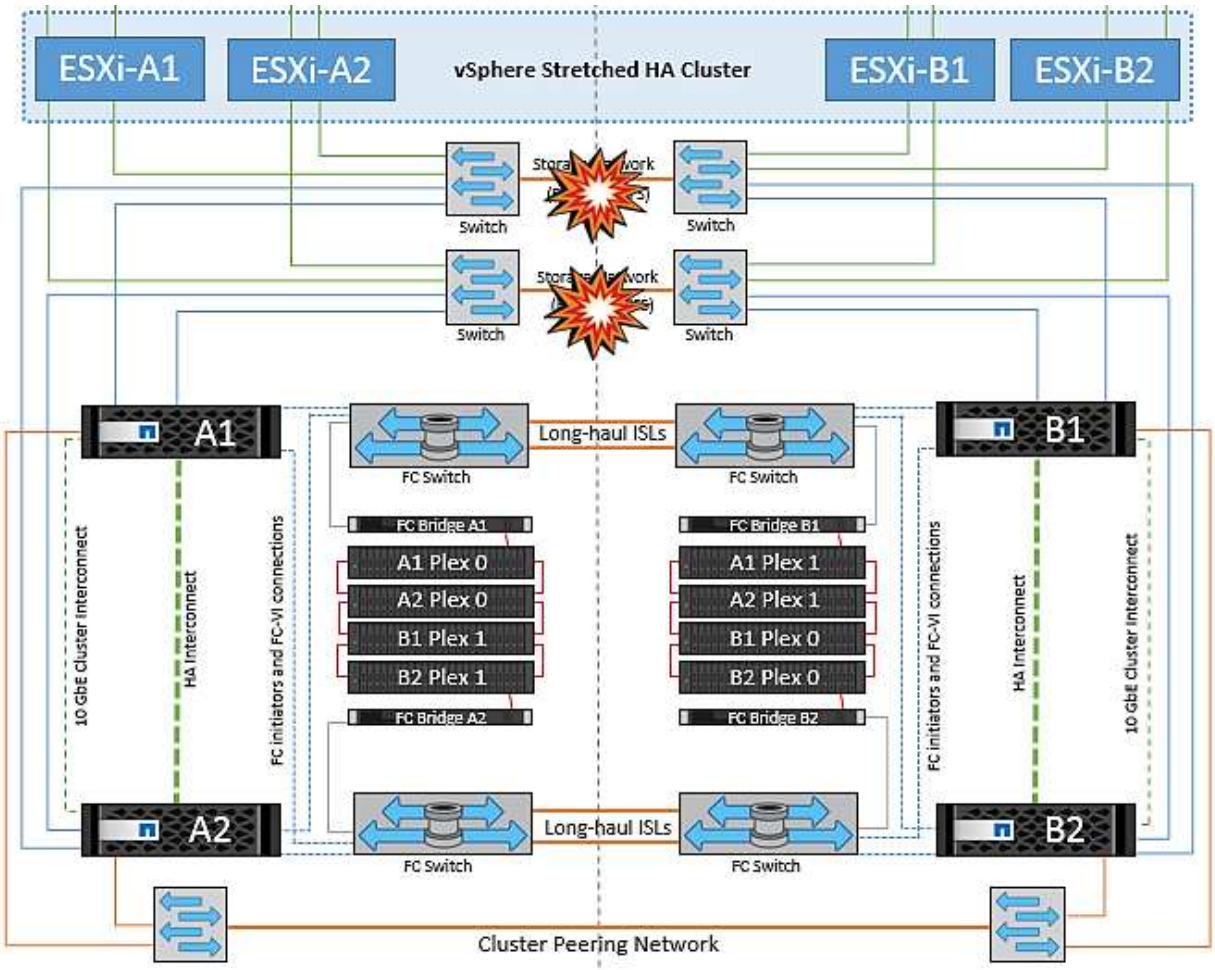
管理网络发生交换机间链路故障



在这种情况下、如果前端主机管理网络的ISL链路发生故障、站点A的ESXi主机将无法与站点B的ESXi主机进行通信这将导致网络分区、因为特定站点上的ESXi主机将无法向HA集群中的主节点发送网络检测点。因此、由于分区、会有两个网段、每个网段中都有一个主节点、用于保护VM免受特定站点中主机故障的影响。

*注意：*在此期间、虚拟机将保持运行状态、MetroCluster行为在这种情况下没有变化。所有数据存储库在其各自的站点中仍保持完好。

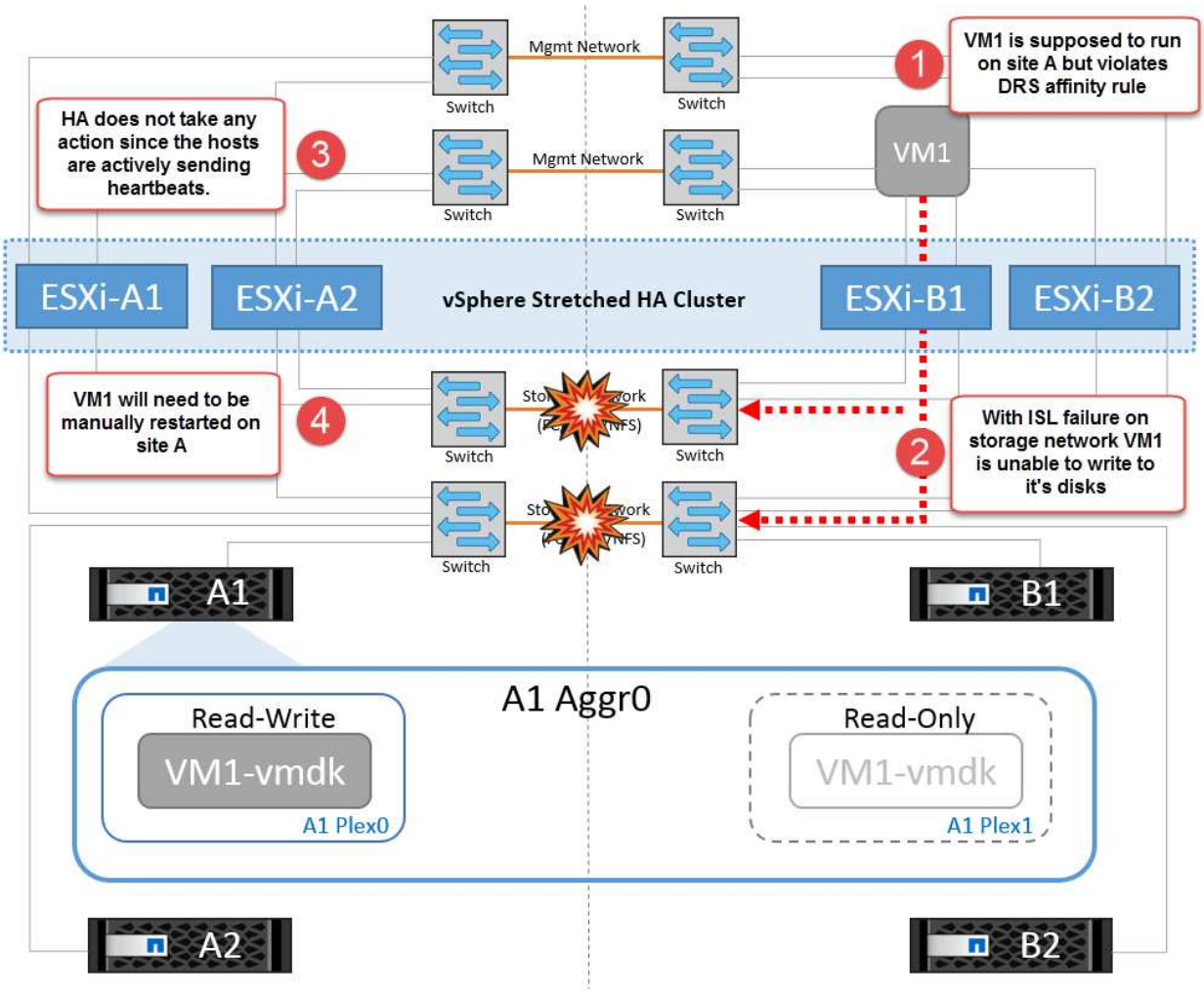
存储网络发生交换机间链路故障



在这种情况下、如果后端存储网络上的ISL链路发生故障、站点A的主机将无法访问站点B上集群B的存储卷或LUN、反之亦然。定义了VMware DRS规则、以便主机-存储站点关联性有利于虚拟机在站点内运行而不会受到影响。

在此期间、虚拟机会在其各自的站点上保持运行状态、并且在此情形下、MetroCluster的行为没有变化。所有数据存储库在其各自的站点中仍保持完好。

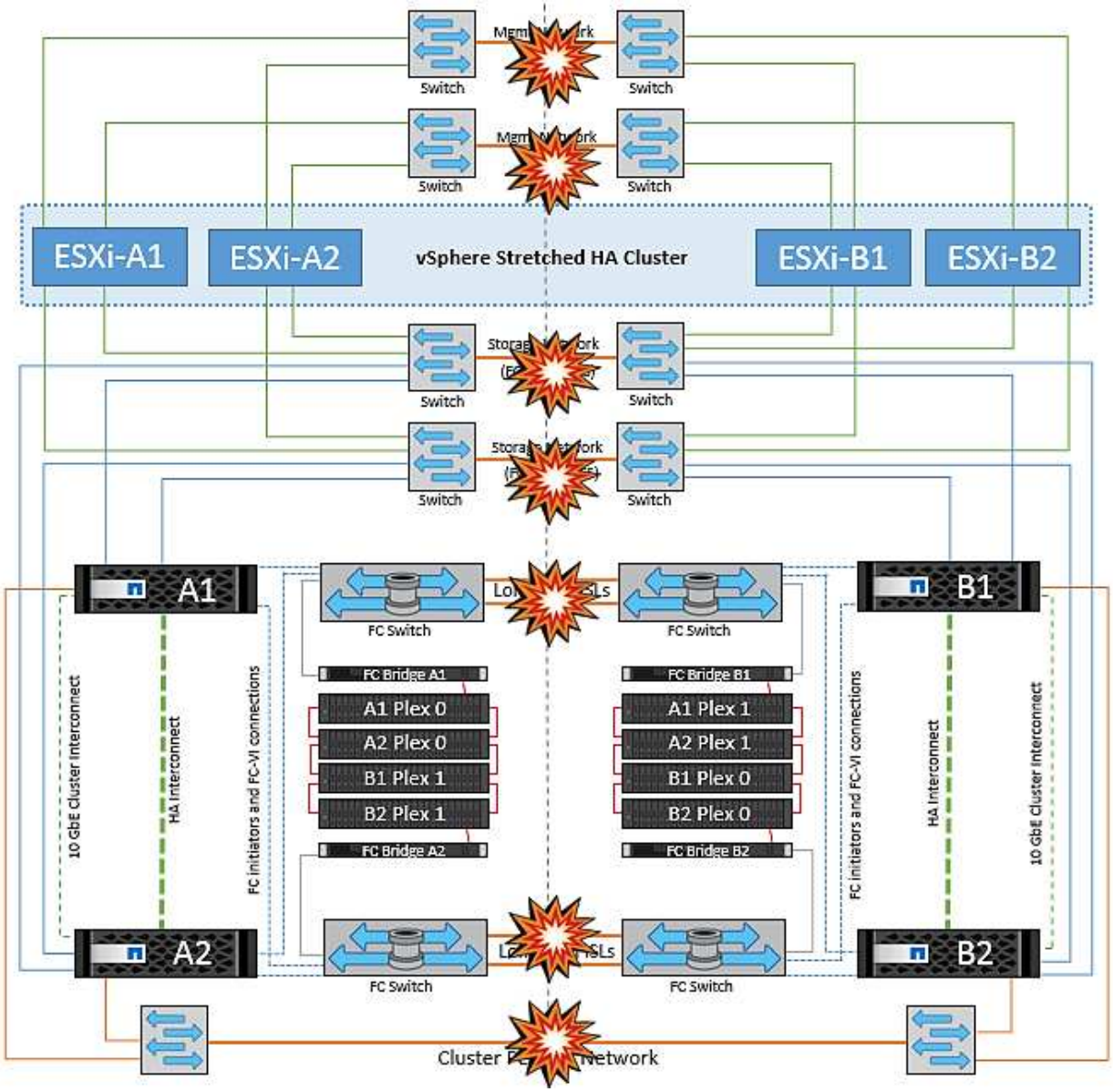
如果出于某种原因违反了相关性规则(例如、VM1本应从站点A运行、而其磁盘位于本地集群A节点上、但却在站点B的主机上运行)、则可以通过ISL链路远程访问虚拟机的磁盘。由于ISL链路故障、在站点B上运行的VM1将无法向其磁盘写入数据、因为存储卷的路径已关闭、并且该特定虚拟机已关闭。在这些情况下、VMware HA不会执行任何操作、因为主机正在主动发送检测信号。这些虚拟机需要在其各自的站点中手动关闭和启动。下图显示了违反DRS关联性规则的虚拟机。



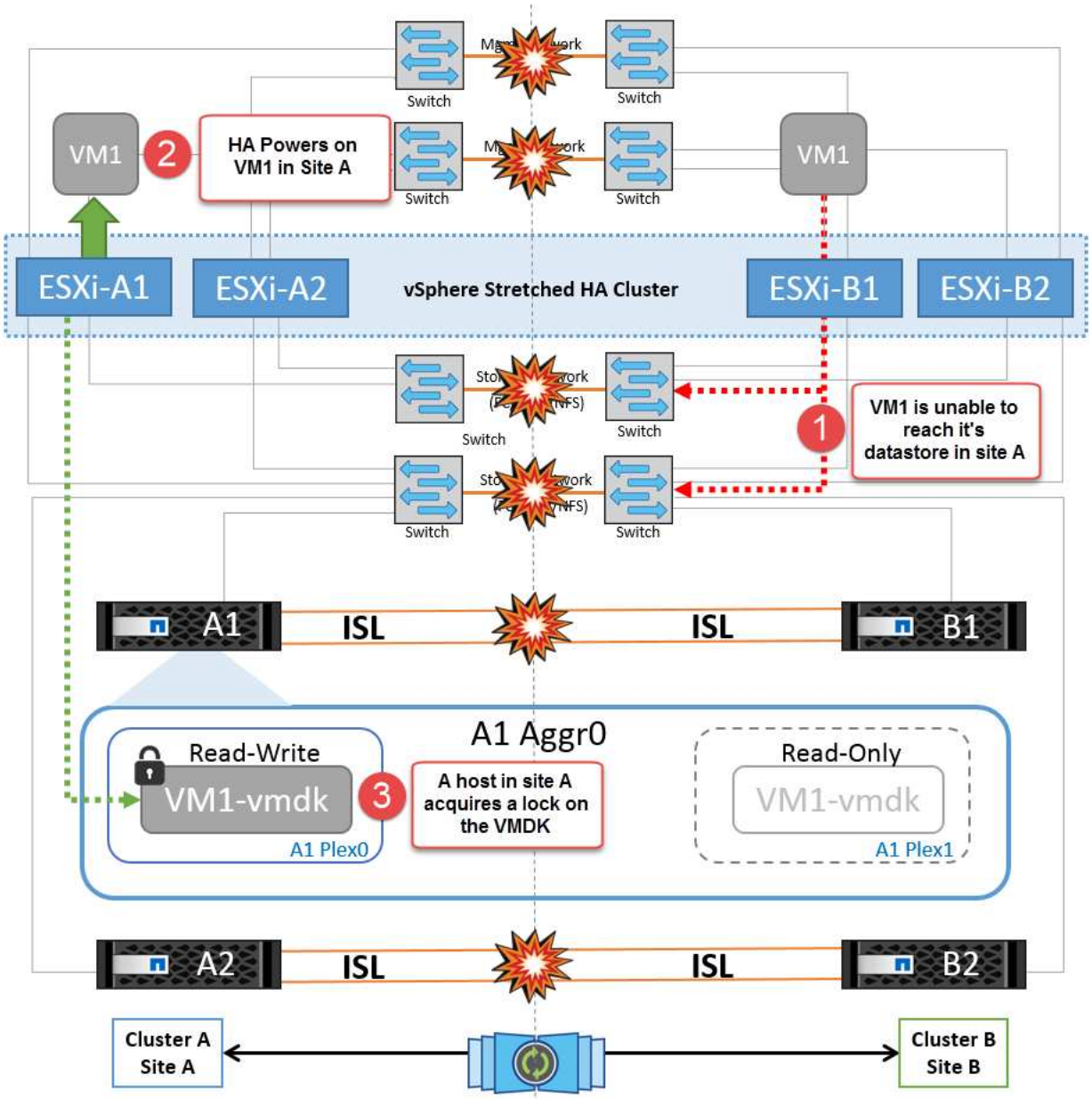
所有交换机间故障或完整数据中心分区

在此场景中、两个站点之间的所有ISL链路均已关闭、并且两个站点彼此隔离。如前文所述(例如、管理网络和存储网络出现ISL故障)、虚拟机不会在完全ISL故障时受到影响。

在站点之间对ESXi主机进行分区后、vSphere HA代理将检查数据存储库检测点、并且在每个站点中、本地ESXi主机将能够将此数据存储库检测点更新到其各自的读写卷/LUN。站点A中的主机将假定站点B中的其他ESXi主机发生故障、因为不存在网络/数据存储库检测点。站点A的vSphere HA将尝试重新启动站点B的虚拟机、但此操作最终将失败、因为存储ISL故障将无法访问站点B的数据存储库。站点B也会出现类似情况



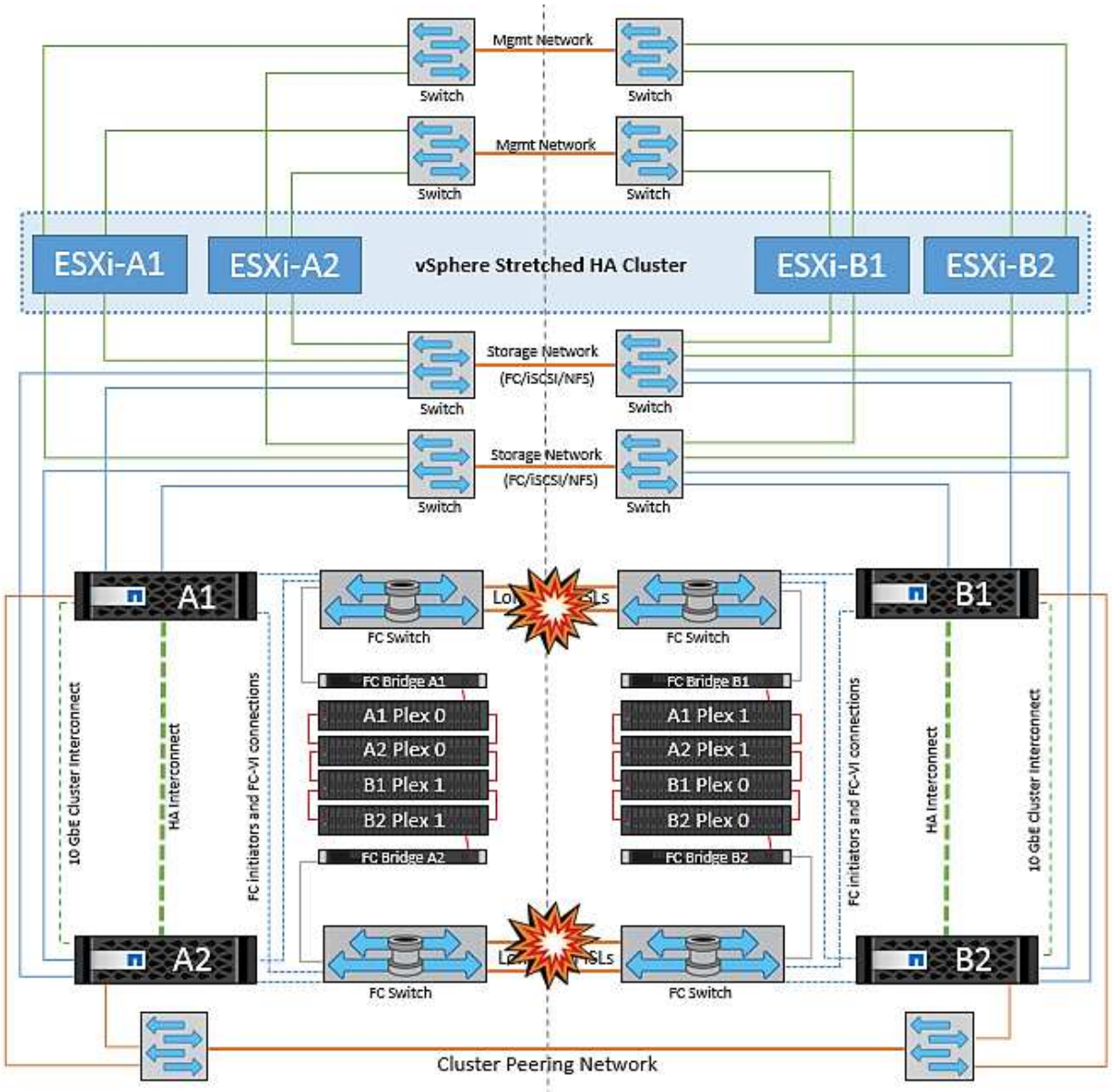
NetApp建议确定是否有任何虚拟机违反了DRS规则。从远程站点运行的任何虚拟机都将关闭、因为它们将无法访问数据存储库、vSphere HA将在本地站点上重新启动该虚拟机。ISL链路恢复联机后、远程站点上运行的虚拟机将被终止、因为不能存在两个使用相同MAC地址运行的虚拟机实例。



NetApp MetroCluster中的两个网络结构上的交换机间链路均出现故障

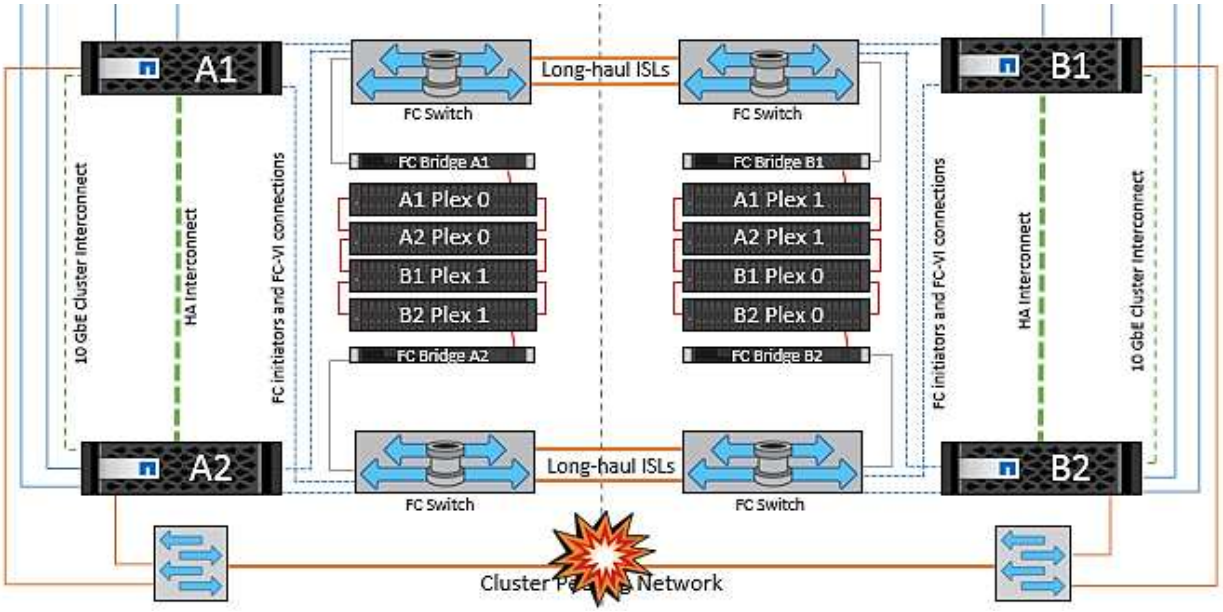
如果一个或多个ISL发生故障、流量将继续通过其余链路。如果两个网络结构上的所有ISO都发生故障、以致于站点之间没有用于存储和NVRAM复制的链路、则每个控制器将继续提供其本地数据。在还原至少一个ISL时、将自动重新同步所有plexes。

在所有ISL关闭后发生的任何写入操作都不会镜像到另一站点。因此、如果在配置处于此状态时发生灾难切换、则会丢失未同步的数据。在这种情况下、需要手动干预才能在切换后进行恢复。如果很可能在很长时间内不会有任何可用的CRL、则管理员可以选择关闭所有数据服务、以避免在发生灾难时需要切换时数据丢失的风险。在至少有一个ISL可用之前、应权衡执行此操作与发生灾难时需要切换的可能性。或者、如果在级联情形下、CRL发生故障、管理员可以在所有链路发生故障之前触发到某个站点的计划内切换。



对等集群链路故障

在对等集群链路故障情形下、由于网络结构的CRL仍处于活动状态、因此两个站点上的数据服务(读取和写入)将继续提供给两个plexs。任何集群配置更改(例如、添加新SVM、在现有SVM中配置卷或LUN)都无法传播到其他站点。这些卷保存在本地CRS元数据卷中、并在对等集群链路还原后自动传播到另一集群。如果需要强制切换才能还原对等集群链路、则在切换过程中、系统将从正常运行的站点上元数据卷的远程复制副本自动重做未完成的集群配置更改。



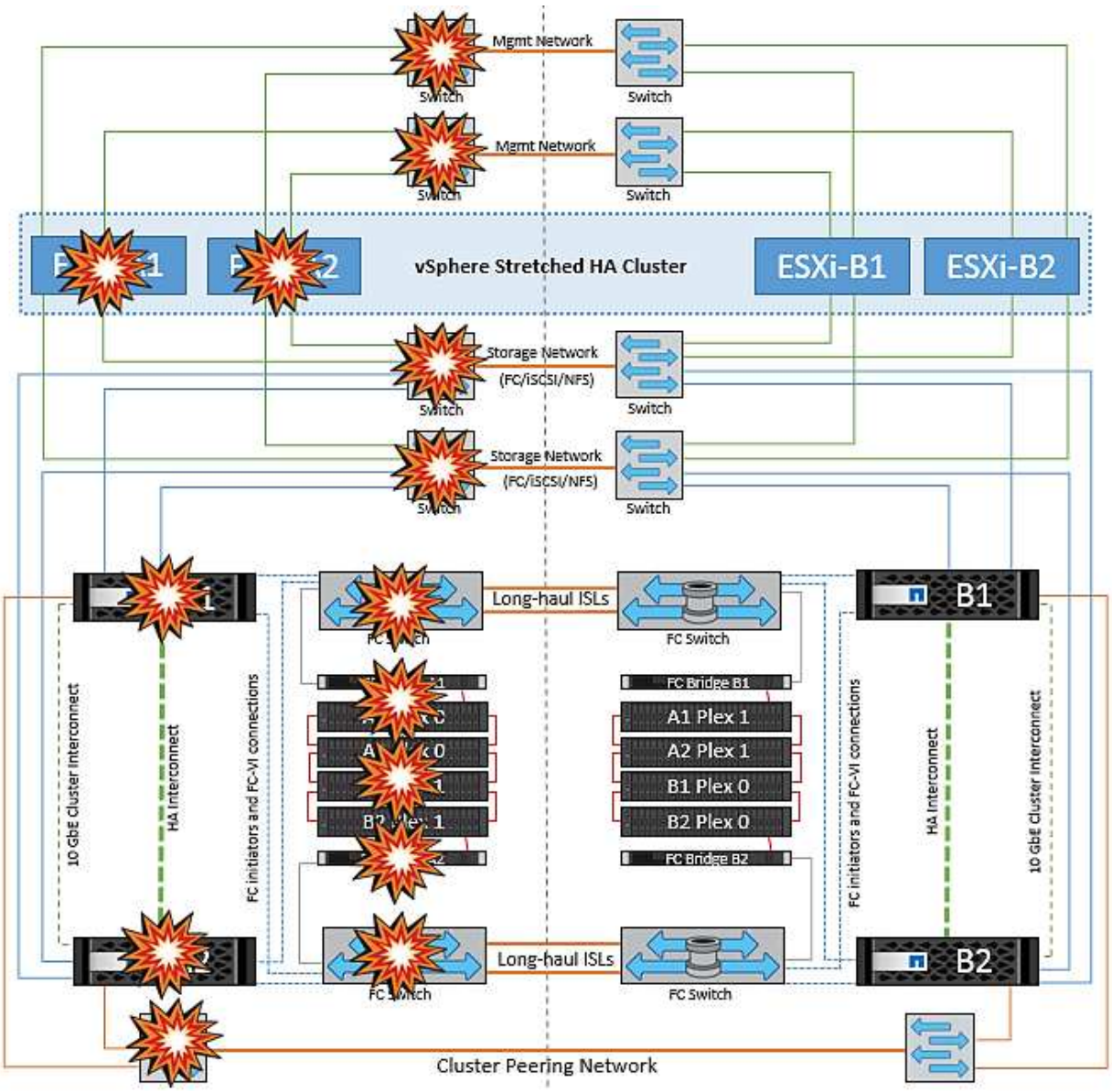
站点完全瘫痪

在完整站点A发生故障的情况下、站点B的ESXi主机无法从站点A的ESXi主机获取网络检测信号、因为它们已关闭。站点B的HA主节点将验证数据存储库检测点是否不存在、并声明站点A的主机出现故障、然后尝试在站点B中重新启动站点A虚拟机在此期间、存储管理员将执行切换以恢复运行正常的站点上故障节点的服务、从而恢复站点B上站点A的所有存储服务当站点A的卷或LUN在站点B上可用后、HA主代理将尝试在站点B中重新启动站点A的虚拟机

如果vSphere HA主代理尝试重新启动虚拟机(包括注册虚拟机并打开虚拟机电源)失败、则会在出现延迟后重试重新启动。重新启动之间的延迟最长可配置为30分钟。vSphere HA尝试这些重新启动的次数最多(默认为六次)。

*注意：*除非布局管理器找到合适的存储、否则HA主节点不会开始尝试重新启动、因此、如果站点完全瘫痪、则应在执行切换后进行。

如果站点A已切换、则可以通过故障转移到运行正常的节点来无缝处理其中一个运行正常的站点B节点的后续故障。在这种情况下、四个节点的工作现在仅由一个节点执行。在这种情况下、恢复将包括向本地节点执行一次恢复。然后、在还原站点A后、将执行切回操作以还原配置的稳定状态操作。


Mgmt Network
Switch
Mgmt Network
Switch
vSphere Stretched HA Cluster
ESXi-B1
ESXi-B2
Storage Network
(FC/iSCSI/NFS)
Switch
Storage Network
(FC/iSCSI/NFS)
Switch
Long-haul ISLs
FC Switch
B1
FC Bridge B1
A1 Plex 1
A2 Plex 1
B1 Plex 0
B2 Plex 0
FC Bridge B2
Long-haul ISLs
FC Switch
B2
10 GbE Cluster Interconnect
HA Interconnect
FC initiators and FC-VI connections
FC initiators and FC-VI connections
HA Interconnect
10 GbE Cluster Interconnect
Cluster Peering Network