Oracle灾难恢复

Enterprise applications

NetApp
February 11, 2026

目录

Oracle灾难恢复

概述

灾难恢复是指在发生灾难性事件(例如火灾、导致存储系统甚至整个站点遭到破坏)后恢复数据服务。

> 本文档可替代先前发布的技术报告_TR-4591：《Oracle数据保护》_和_TR-4592：《基于MetroCluster的Oracle》

灾难恢复可以通过使用SnapMirror轻松复制数据来实现、当然、许多客户每小时更新一次镜像副本。

对于大多数客户而言、灾难恢复不仅需要拥有远程数据副本、还需要能够快速利用这些数据。NetApp提供了两种技术来满足这一需求—MetroCluster和SnapMirror主动同步

MetroCluster是指硬件配置中的ONTAP、其中包括低级同步镜像存储和许多附加功能。MetroCluster等集成解决方案简化了当今复杂的横向扩展数据库、应用程序和虚拟化基础架构。它将多个外部数据保护产品和策略替换为一个简单的中央存储阵列。此外、它还可以在一个集群模式存储系统中提供集成的备份、恢复、灾难恢复和高可用性(HA)功能。

SnapMirror活动同步(SM-AS)基于SnapMirror同步。通过MetroCluster、每个ONTAP控制器都负责将其驱动器数据复制到远程位置。使用SnapMirror主动同步时、您实际上拥有两个不同的ONTAP系统、它们会维护LUN数据的独立副本、但会相互协作、为该LUN提供一个实例。从主机角度来看、它是一个LUN实体。

SM-AS和MCC比较

SM-AS和MetroCluster在整体功能上相似、但在实施RPO = 0复制的方式及其管理方式上存在重要差异。SnapMirror异步和同步也可用作灾难恢复计划的一部分、但它们不是作为HA回配技术而设计的。

- MetroCluster配置更像是一个集成集群、其中的节点分布在各个站点之间。SM-AS的行为类似于两个其他方面独立的集群、它们合作提供选定的RPO = 0同步复制的LUN。
- 在任何给定时间、只能从一个特定站点访问MetroCluster配置中的数据。另一个数据副本位于另一个站点上、但数据是被动的。如果没有存储系统故障转移、则无法访问它。
- MetroCluster和SM-AS执行镜像在不同级别进行。MetroCluster镜像在RAID层执行。使用SyncMirror以镜像格式存储低级别的数据。在LUN、卷和协议层、镜像的使用实际上是不可见的。
- 相反、SM-AS镜像发生在协议层。这两个集群总体上是独立的集群。两个数据副本同步后、这两个集群只需镜像写入即可。在一个集群上进行写入时、该写入会复制到另一个集群。只有在两个站点上的写入均已完成时、才会向主机确认写入。除了此协议拆分行为之外、这两个集群在其他方面都是正常的ONTAP集群。
- MetroCluster的主要角色是大规模复制。您可以复制RPO为0且RTO接近零的整个阵列。这样可以简化故障转移过程、因为故障转移只需执行一项"操作"、而且在容量和IOPS方面扩展得非常好。
- SM-AS的一个关键用例是粒度复制。有时、您不希望将所有数据作为一个单元进行复制、或者您需要能够有选择地对某些工作负载进行故障转移。
- SM-AS的另一个主要用例是主动-主动操作、您希望在位于两个不同位置的两个不同集群上提供完全可用的数据副本、这些集群具有相同的性能特征、如果需要、也不需要在站点间延伸SAN。您的应用程序可以同时在两个站点上运行、这样可以减少故障转移操作期间的整体恢复时间。

MetroCluster

借助**MetroCluster**实现灾难恢复

MetroCluster是ONTAP的一项功能、可通过站点间的RPO = 0同步镜像保护Oracle数据库、并可进行扩展以在一个MetroCluster系统上支持数百个数据库。

它也易于使用。使用MetroCluster并不一定会增加或更改运行企业级应用程序和数据库的任何最佳网络竞赛。

通常的最佳实践仍然适用、如果您的需求仅需要RPO = 0数据保护、则MetroCluster可以满足该需求。但是、大多数客户使用MetroCluster不仅可以实现RROT=0的数据保护、还可以在灾难情形下提高RTO、并在站点维护活动中提供透明的故障转移。

物理架构

要了解Oracle数据库在MetroCluster环境中的运行方式、需要对MetroCluster系统的物理设计进行一些说明。

本文档可替代先前发布的技术报告_TR-4592：《基于MetroCluster的Oracle》

MetroCluster可用于**3**种不同的配置

- 具有IP连接的HA对
- 具有FC连接的HA对
- 具有FC连接的单个控制器

术语"连接"是指用于跨站点复制的集群连接。它不是指主机协议。无论用于集群间通信的连接类型如何、MetroCluster配置均支持所有主机端协议。

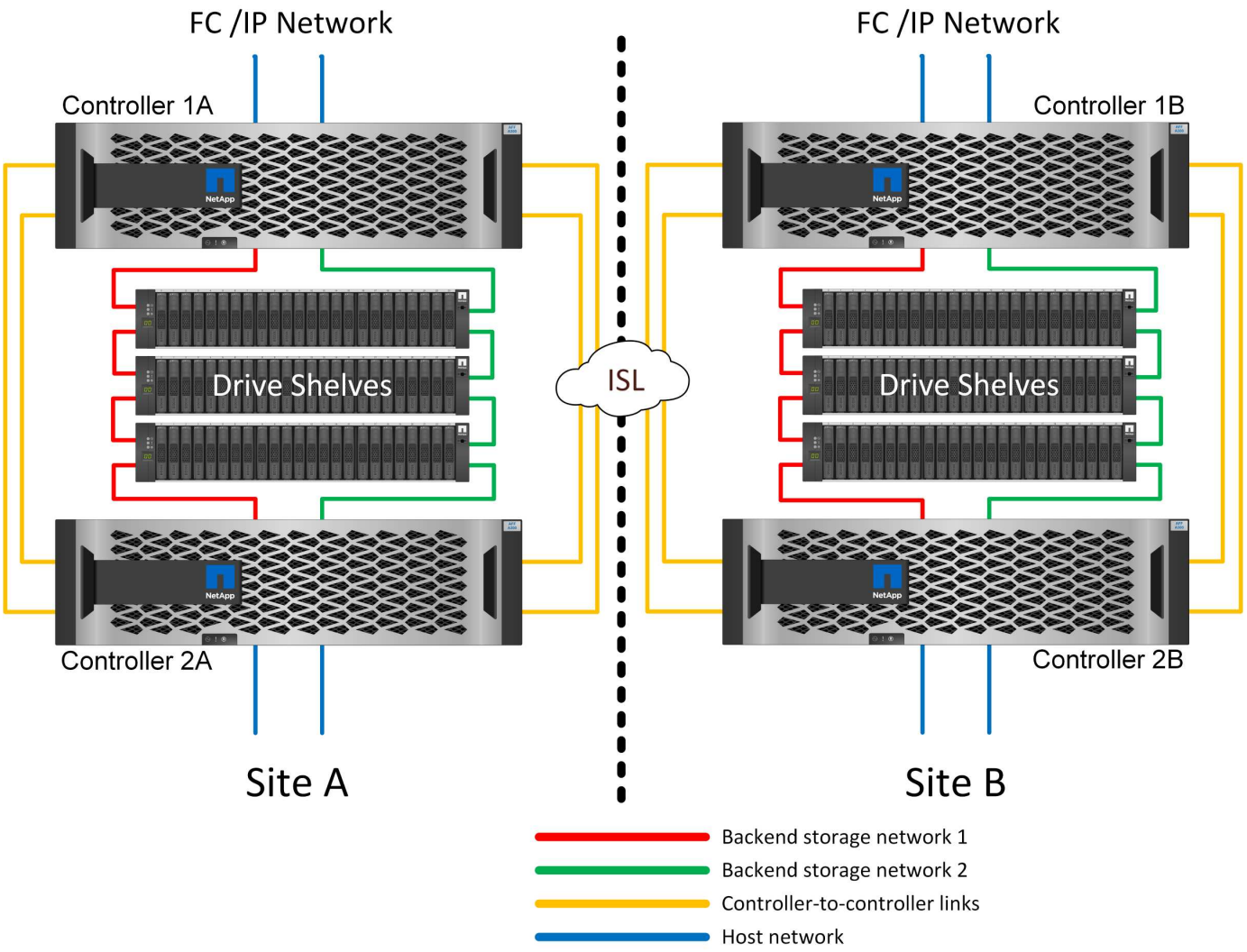
MetroCluster IP

HA对MetroCluster IP配置会在每个站点上使用两个或四个节点。与双节点选项相比、此配置选项会增加复杂性和成本、但它具有一个重要优势：站点内冗余。简单的控制器故障不需要通过WAN访问数据。数据访问仍通过备用本地控制器保持在本地。

大多数客户选择IP连接是因为基础架构要求更简单。过去、使用暗光纤和FC交换机配置高速跨站点连接通常比较容易、但如今、高速、低延迟IP电路更容易获得。

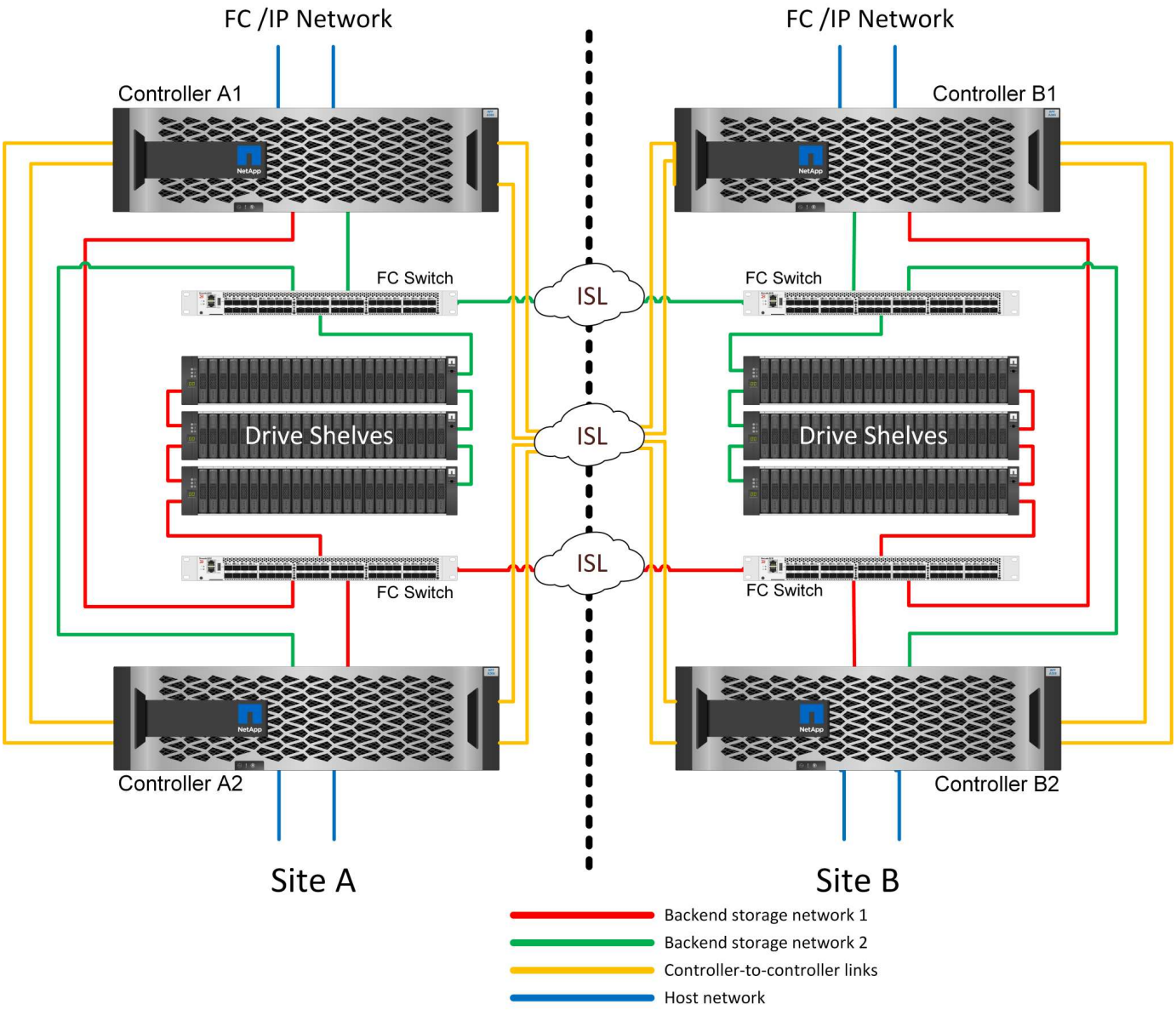
此外、该架构也更加简单、因为只有跨站点连接用于控制器。在FC SAN连接的MetroCluster中、控制器会直接写入另一站点上的驱动器、因此需要更多的SAN连接、交换机和网桥。相反、IP配置中的控制器会通过控制器写入相对的驱动器。

对于追加信息、请参阅ONTAP官方文档和 "MetroCluster IP 解决方案架构和设计"。



HA对FC SAN连接的MetroCluster

HA对MetroCluster FC配置会在每个站点上使用两个或四个节点。与双节点选项相比、此配置选项会增加复杂性和成本、但它具有一个重要优势：站点内冗余。简单的控制器故障不需要通过WAN访问数据。数据访问仍通过备用本地控制器保持在本地。

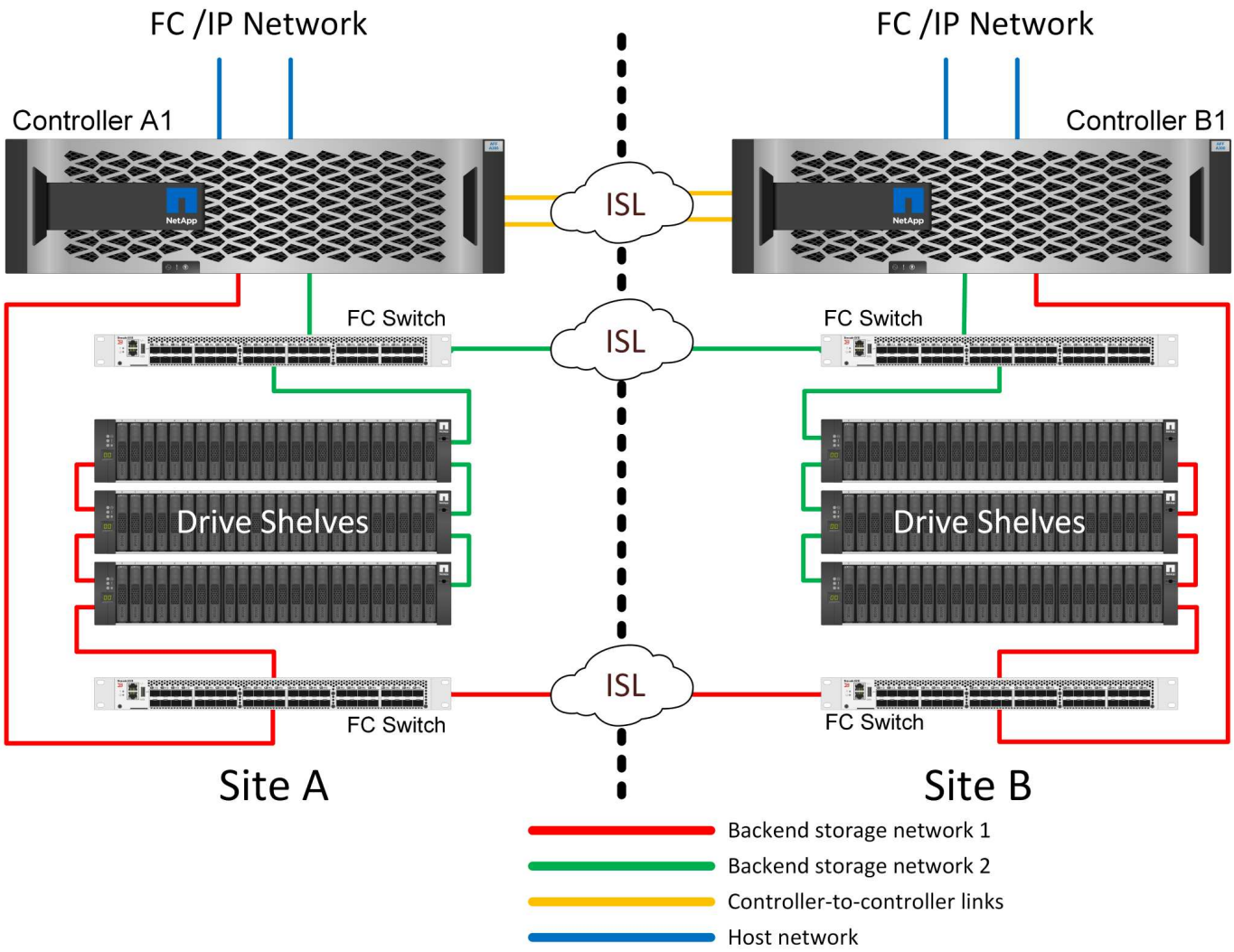


某些多站点基础架构不是为主动-主动操作而设计的、而是更多地用作主站点和灾难恢复站点。在这种情况下、通常最好使用HA对MetroCluster选项、原因如下：

- 尽管双节点MetroCluster集群是一个HA系统、但控制器意外故障或计划内维护要求数据服务必须在相反站点联机。如果站点之间的网络连接无法支持所需的带宽、则性能会受到影响。唯一的选择是同时将各种主机操作系统和相关服务故障转移到备用站点。HA对MetroCluster集群可消除此问题、因为丢失控制器会导致在同一站点内进行简单的故障转移。
- 某些网络拓扑不是为跨站点访问而设计的、而是使用不同的子网或隔离的FC SAN。在这些情况下、双节点MetroCluster集群将不再充当HA系统、因为备用控制器无法向对面站点上的服务器提供数据。要提供完全冗余、需要使用高可用性对MetroCluster选项。
- 如果将双站点基础架构视为一个高可用性基础架构、则适合使用双节点MetroCluster配置。但是、如果系统在站点发生故障后必须长时间运行、则首选HA对、因为它会继续在单个站点中提供HA。

双节点**FC SAN**连接**MetroCluster**

双节点MetroCluster配置仅为每个站点使用一个节点。这种设计比HA对选项更简单、因为需要配置和维护的组件更少。此外、它还降低了布线和FC交换方面的基础架构需求。最后、它还可以降低成本。



这种设计的明显影响是、单个站点上的控制器故障意味着数据可以从另一个站点访问。这种限制不一定是问题。许多企业都拥有多站点数据中心运营、并采用延伸型高速低延迟网络、这些网络本质上充当一个基础架构。在这些情况下、首选配置是双节点版本的MetroCluster。目前、多家服务提供商以PB级的规模使用双节点系统。

MetroCluster故障恢复能力功能

MetroCluster 解决方案 中没有单点故障：

- 每个控制器都有两条通往本地站点上的驱动器架的独立路径。
- 每个控制器都有两条通往远程站点上驱动器架的独立路径。
- 每个控制器都有两条独立的路径连接到另一站点上的控制器。
- 在HA对配置中、每个控制器都有两个指向其本地配对节点的路径。

总之、可以删除配置中的任何一个组件、而不会影响MetroCluster提供数据的能力。这两个选项在故障恢复能力方面的唯一区别是、发生站点故障后、HA对版本仍然是整体HA存储系统。

逻辑架构

要了解Oracle数据库如何在MetroCluster环境alsop中运行、需要对MetroCluster系统的逻

辑功能进行一些说明。

站点故障保护：**NVRAM**和**MetroCluster**

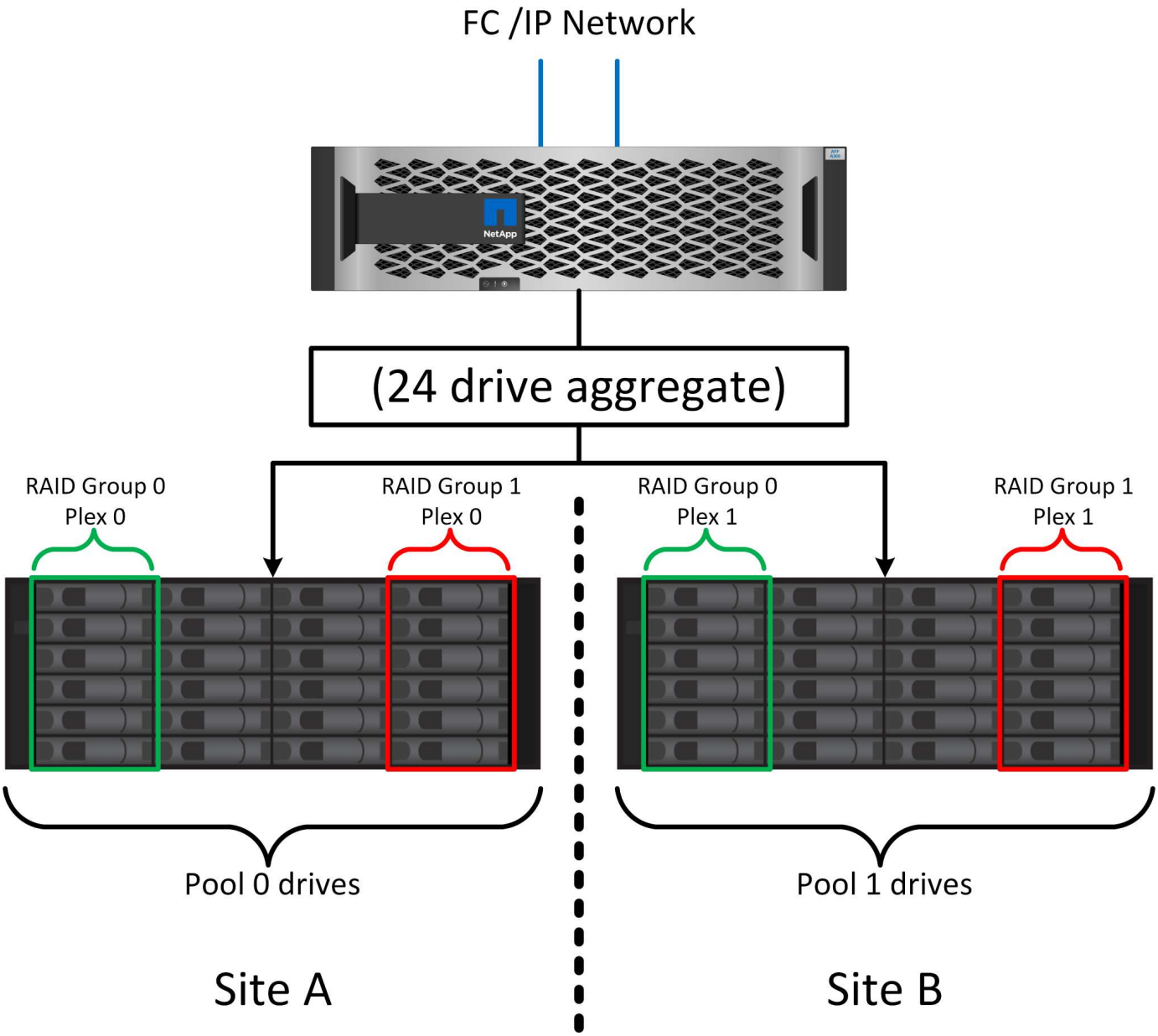
MetroCluster通过以下方式扩展NVRAM数据保护：

- 在双节点配置中、NVRAM数据通过交换机间链路(ISL)复制到远程配对节点。
- 在HA对配置中、NVRAM数据会同时复制到本地配对节点和远程配对节点。
- 写入只有在复制到所有配对项后才会得到确认。此架构通过将NVRAM数据复制到远程配对节点来保护传输中的I/O免受站点故障的影响。驱动器级数据复制不涉及此过程。拥有聚合的控制器负责通过向聚合中的两个plexes写入数据来进行数据复制、但在站点丢失时、仍必须防止传输中I/O丢失。只有当配对控制器必须接管发生故障的控制器时、才会使用复制的NVRAM数据。

站点和磁盘架故障保护：**SyncMirror**和**plexes**

SyncMirror是一种镜像技术、可增强但不会取代RAID DP或RAID-TEC。它会镜像两个独立RAID组的内容。逻辑配置如下：

1. 驱动器会根据位置配置到两个池中。一个池由站点A上的所有驱动器组成、另一个池由站点B上的所有驱动器组成
2. 然后、基于RAID组的镜像集创建一个通用存储池(称为聚合)。从每个站点提取的驱动器数量相等。例如、一个包含20个驱动器的SyncMirror聚合将由站点A的10个驱动器和站点B的10个驱动器组成
3. 给定站点上的每组驱动器都会自动配置为一个或多个完全冗余的RAID DP或RAID-TEC组、而不依赖于镜像的使用。在镜像下使用RAID可提供数据保护、即使在站点丢失后也是如此。



上图显示了一个示例SyncMirror配置。在控制器上创建了一个包含24个驱动器的聚合、其中12个驱动器来自站点A上分配的磁盘架、12个驱动器来自站点B上分配的磁盘架这些驱动器被分组为两个镜像RAID组。RAID组0在站点A上包含一个6驱动器丛、该丛镜像到站点B上的一个6驱动器丛同样、RAID组1在站点A上包含一个6驱动器丛、该丛镜像到站点B上的6驱动器丛

SyncMirror通常用于为MetroCluster系统提供远程镜像、每个站点有一个数据副本。有时、它会用于在单个系统中提供额外的冗余级别。尤其是、它可以提供磁盘架级冗余。驱动器架已包含双电源和控制器、总体比金属板稍多、但在某些情况下、可能需要额外保护。例如、一家NetApp客户为汽车测试期间使用的移动实时分析平台部署了SyncMirror。该系统分为两个物理机架、配有独立的电源和独立的UPS系统。

冗余故障：**NVFAIL**

如前文所述、写入操作只有在至少另一个控制器上记录到本地NVRAM和NVRAM后才会得到确认。此方法可确保硬件故障或断电不会导致传输中I/O丢失如果本地NVRAM发生故障或与其他节点的连接发生故障、则无法再镜像数据。

如果本地NVRAM报告错误、则此节点将关闭。此关闭会导致在使用HA对时故障转移到配对控制器。使用MetroCluster时、行为取决于所选的整体配置、但可能会自动故障转移到远程便签。在任何情况下、数据都不会丢失、因为发生故障的控制器尚未确认写入操作。

站点间连接故障会阻止NVRAM复制到远程节点、这种情况更为复杂。写入操作不再复制到远程节点、因此、如果控制器发生灾难性错误、可能会导致数据丢失。更重要的是、在这些情况下尝试故障转移到其他节点会导致数据丢失。

控制因素是NVRAM是否同步。如果NVRAM已同步、则可以安全地进行节点间故障转移、而不会丢失数据。在MetroCluster配置中、如果NVRAM与底层聚合plexes处于同步状态、则可以安全地继续执行切换、而不会丢失数据。

除非强制执行故障转移或切换、否则ONTAP不允许在数据不同步时执行故障转移或切换。以这种方式强制更改条件即表示数据可能会留在原始控制器中、并且数据丢失是可以接受的。

如果强制执行故障转移或切换、则数据库和其他应用程序尤其容易受到损坏的影响、因为它们在磁盘上维护着更大的内部数据缓存。如果发生强制故障转移或切换、先前确认的更改将被有效丢弃。存储阵列的内容会及时有效地向后跳转、缓存的状态不再反映磁盘上数据的状态。

为了防止出现这种情况、ONTAP允许对卷进行配置、以便针对NVRAM故障提供特殊保护。触发此保护机制后、卷将进入名为NVFAIL的状态。此状态会导致发生原因应用程序崩溃的I/O错误。此崩溃会导致应用程序关闭、以便它们不会使用过时数据。数据不应丢失、因为日志中应存在任何已提交的事务数据。通常的后续步骤是、管理员先完全关闭主机、然后再手动将LUN和卷重新联机。虽然这些步骤可能涉及一些工作、但这种方法是确保数据完整性的最安全方法。并非所有数据都需要这种保护、这就是可以逐个卷配置NVFAIL行为的原因。

HA对和**MetroCluster**

MetroCluster有两种配置：双节点和HA对。就NVRAM而言、双节点配置与HA对的行为相同。如果发生突然故障、配对节点可以重放NVRAM数据、以确保驱动器一致、并确保未丢失任何已确认的写入。

HA对配置也会将NVRAM复制到本地配对节点。简单的控制器故障会导致配对节点上的NVRAM重放、就像不使用MetroCluster的独立HA对一样。如果站点突然完全丢失、远程站点还具有必要的NVRAM、以使驱动器保持一致并开始提供数据。

MetroCluster的一个重要方面是、在正常运行条件下、远程节点无法访问配对节点数据。每个站点本质上都是一个独立的系统、可以承担相反站点的特性。此过程称为切换、其中包括计划内切换、在此过程中、站点操作会无系统地迁移到相反站点。此外、还包括站点丢失以及在灾难恢复过程中需要手动或自动切换的计划外情况。

切换和切回

术语切换和切回是指在MetroCluster配置中的远程控制器之间过渡卷的过程。此过程仅会对远程节点执行适用场景。如果在四卷配置中使用MetroCluster、则本地节点故障转移与前面所述的接管和恢复过程相同。

计划内切换和切回

计划内切换或切回类似于节点之间的接管或交还。此过程包含多个步骤、看起来可能需要几分钟时间、但实际发生的是存储和网络资源的多阶段平稳过渡。控制传输的速度比执行完整命令所需的时间快得多。

接管/交还与切换/切回之间的主要区别在于对FC SAN连接的影响。使用本地接管/备份时、主机会丢失指向本地节点的所有FC路径、并依靠其本机MPIO切换到可用的备用路径。端口不会重新定位。通过切换和切回、控制器上的虚拟FC目标端口将过渡到另一站点。它们实际上暂时不再存在于SAN上、然后重新出现在备用控制器上。

SyncMirror超时

SyncMirror是一种ONTAP镜像技术、可针对磁盘架故障提供保护。如果磁盘架相隔一段距离、则可以实现远程数据保护。

SyncMirror不提供通用同步镜像。结果是可用性更好。某些存储系统使用持续的全镜像或无镜像、有时称为Domino模式。这种形式的镜像在应用程序中受到限制、因为如果与远程站点的连接断开、所有写入活动都必须停止。否则、写入将在一个站点上存在、而在另一个站点上不存在。通常、此类环境会配置为在站点间连接丢失的时间较短(例如30秒)时使LUN脱机。

这种行为适合一小部分环境。但是、大多数应用程序都需要一个解决方案、该系统可以在正常运行条件下提供有保障的同步复制、但可以暂停复制。站点间连接完全断开通常被视为近乎灾难的情况。通常、此类环境会保持联机并提供数据、直到修复连接或正式决定关闭环境以保护数据为止。仅由于远程复制失败而要求自动关闭应用程序的要求并不常见。

SyncMirror支持同步镜像要求、并具有超时的灵活性。如果与远程控制器和/或丛的连接断开、30秒计时器将开始倒计时。当计数器达到0时、写入I/O处理将继续使用本地数据。数据的远程副本可用、但会及时冻结、直到连接恢复为止。重新同步利用聚合级快照使系统尽快恢复到同步模式。

值得注意的是、在许多情况下、在应用程序层实施这种通用的全Domino模式或全无Domino模式复制效果更佳。例如、Oracle DataGuard包括最大保护模式、可保证在任何情况下进行长实例复制。如果复制链路出现故障的时间超过可配置的超时时间、数据库将关闭。

使用光纤连接MetroCluster自动执行无人看管切换

自动无人值守切换(Automatic无人值守切换、AUSO)是一项光纤连接的MetroCluster功能、可提供一种跨站点HA形式。如前文所述、MetroCluster有两种类型：每个站点上一个控制器或每个站点上一个HA对。HA选项的主要优势是、计划内或计划外控制器关闭仍可使所有I/O都位于本地。单节点选项的优势在于降低成本、复杂性和基础架构。

AUSO的主要价值是提高光纤连接MetroCluster系统的HA功能。每个站点都会监控相反站点的运行状况、如果没有节点可提供数据、则AUSO会导致快速切换。在每个站点只有一个节点的MetroCluster配置中、此方法尤其有用、因为它使配置在可用性方面更接近HA对。

AUSO无法在HA对级别提供全面监控。HA对可以提供极高的可用性、因为它包含两根冗余物理缆线、用于节点到节点的直接通信。此外、HA对中的两个节点均可访问冗余环路上的同一组磁盘、从而为一个节点提供另一条路由来监控另一个节点的运行状况。

MetroCluster集群存在于节点间通信和磁盘访问均依赖于站点间网络连接的站点之间。监控集群其余部分的检测信号的能力有限。在另一个站点因网络问题而实际关闭而不是不可用的情况下、AUSO必须区分这种情况。

因此、如果HA对中的控制器检测到因特定原因(例如系统崩溃)而发生的控制器故障、则该控制器可能会提示接管。如果完全断开连接(有时称为丢失检测信号)、它还会提示接管。

只有在原始站点上检测到特定故障时、MetroCluster系统才能安全地执行自动切换。此外、拥有存储系统的控制器必须能够保证磁盘和NVRAM数据保持同步。控制器无法仅因为与源站点断开连接而保证切换的安全性、而源站点仍可正常运行。有关自动执行切换的其他选项、请参见下一节中有关MetroCluster Tieb破碎 机(MCTB)解决方案的信息。

具有光纤连接MetroCluster的MetroCluster Tieb破碎 机

该"NetApp MetroCluster Tieb破碎 机"软件可以在第三个站点上运行、以监控MetroCluster环境的运行状况、发送通知、并在发生灾难时强制执行切换(可选)。有关Tieb破碎 机的完整说明"NetApp 支持站点"，请参见，但MetroCluster Tieb破碎 机的主要用途是检测站点丢失。它还必须区分站点丢失和连接丢失。例如、切换不应

因TiebREAKER无法访问主站点而发生、这就是TiebREAKER同时监控远程站点联系主站点的能力的原因。

使用AUSO自动切换也与MCTB兼容。AUSO反应非常迅速、因为它可以检测特定的故障事件、然后仅在NVRAM和SyncMirror plexes处于同步状态时调用切换。

相反、Tieb破碎 机位于远程位置、因此必须等待计时器经过、然后才能宣布站点停机。Tieb破碎 机最终会检测到由AUSO涵盖的那种控制器故障、但通常、在Tieb破碎 机开始工作之前、AUSO已启动切换、并且可能已完成切换。Tieb破碎 机生成的第二个切换命令将被拒绝。

强制切换时、MCTB软件不会验证NVRAM是否同步和/或plexes是否同步。如果已配置自动切换、则应在维护活动期间禁用、从而导致NVRAM或SyncMirror plexes失去同步。

此外、MCTB可能无法解决导致以下一系列事件的滚动灾难：

1. 站点之间的连接中断30秒以上。
2. SyncMirror复制超时、并且会继续在主站点上执行操作、从而使远程副本过时。
3. 主站点丢失。结果是主站点上存在未复制的更改。因此、切换可能不受欢迎、原因有很多、其中包括：
    - 主站点上可能存在关键数据、这些数据最终可能是可恢复的。允许应用程序继续运行的切换将有效地丢弃这些关键数据。
    - 运行正常的站点上的某个应用程序在站点丢失时使用了主站点上的存储资源、此应用程序可能已缓存数据。切换会导致数据版本过时、与缓存不匹配。
    - 运行正常的站点上的某个操作系统在站点丢失时使用了主站点上的存储资源、此操作系统可能已缓存数据。切换会导致数据版本过时、与缓存不匹配。最安全的方法是、将Tieber4配置为在检测到站点故障时发送警报、然后由某人决定是否强制执行切换。可能需要先关闭应用程序和/或操作系统、才能清除缓存的任何数据。此外、还可以使用NVFAIL设置来添加进一步的保护、并帮助简化故障转移过程。

使用**MetroCluster IP**的**ONTAP**调解器

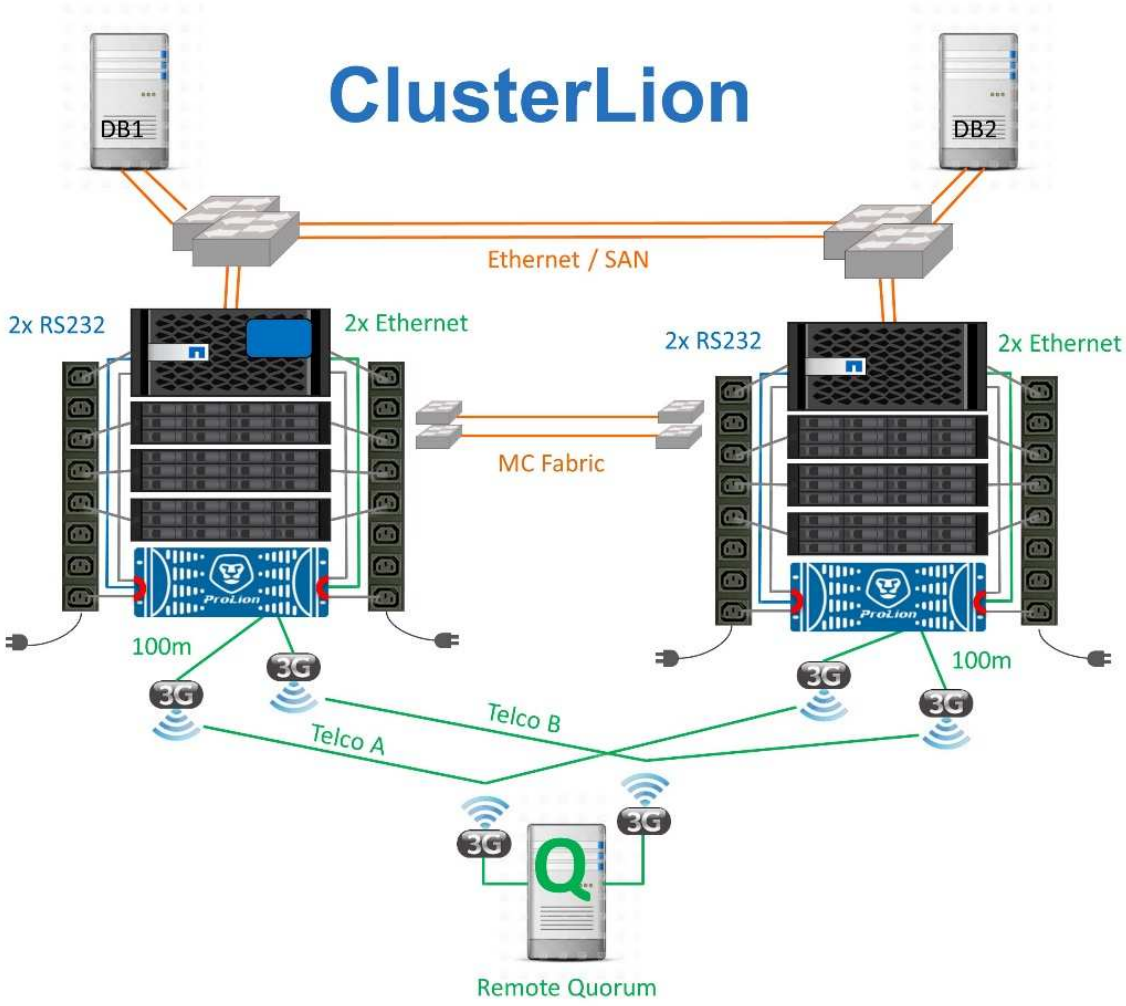
ONTAP调解器可与MetroCluster IP和某些其他ONTAP解决方案结合使用。它的功能与上述MetroCluster Tieb破碎 机软件非常相似、但也包括一项关键功能—执行自动无人值守切换。

光纤连接的MetroCluster可以直接访问相反站点上的存储设备。这样、一个MetroCluster控制器就可以通过从驱动器中读取检测信号数据来监控其他控制器的运行状况。这样、一个控制器就可以识别另一个控制器的故障并执行切换。

相比之下、MetroCluster IP架构会通过控制器-控制器连接独占路由所有I/O；无法直接访问远程站点上的存储设备。这会限制控制器检测故障和执行切换的能力。因此、需要将ONTAP调解器作为Tieb破碎 机设备来检测站点丢失并自动执行切换。

使用**ClusterLion**的虚拟第三站点

ClusterLion是一种高级MetroCluster监控设备、可充当虚拟第三站点。通过这种方法、可以在双站点配置中安全地部署MetroCluster、并提供完全自动化的切换功能。此外、ClusterLion还可以执行额外的网络级监控并执行切换后操作。完整文档可从ProLion获得。



- ClusterLion设备可通过直接连接的以太网和串行缆线监控控制器的运行状况。
- 这两个设备通过冗余3G无线连接相互连接。
- ONTAP控制器的电源通过内部继电器供电。如果站点发生故障、包含内部UPS系统的ClusterLion会在调用切换之前断开电源连接。此过程可确保不会出现脑裂情况。
- ClusterLion会在30秒SyncMirror超时时间内执行切换、或者根本不执行切换。
- 除非NVRAM和SyncMirror plexes的状态保持同步、否则ClusterLion不会执行切换。
- 由于ClusterLion仅在MetroCluster完全同步时执行切换、因此不需要NVFAIL。此配置允许站点范围的环境(例如扩展Oracle RAC)保持联机、即使在计划外切换期间也是如此。
- 支持包括光纤连接MetroCluster和MetroCluster IP

SyncMirror

使用MetroCluster系统进行Oracle数据保护的基础是SyncMirror、这是一种性能最高的横向扩展同步镜像技术。

利用**SyncMirror**实现数据保护

最简单的一个层面是、同步复制意味着、在确认镜像存储之前、必须对镜像存储的两端进行任何更改。例如、如果数据库正在写入日志、或者VMware子系统正在修补、则写入操作绝不能丢失。作为协议级别、在将写入提交

到两个站点上的非易失性介质之前、存储系统不得确认写入。只有这样、才能安全地继续操作、而不会丢失数据。

使用同步复制技术是设计和管理同步复制解决方案的第一步。最重要的注意事项是了解在各种计划内和计划外故障情形下可能发生的情况。并非所有同步复制解决方案都能提供相同的功能。如果您需要的解决方案能够实现零恢复点目标(RPO)、即零数据丢失、则必须考虑所有故障情形。特别是、如果由于站点间连接断开而无法进行复制、则会产生什么预期结果?

SyncMirror数据可用性

MetroCluster复制基于NetApp SyncMirror技术、该技术旨在高效地切换至同步模式和切换至同步模式之外。此功能可满足需要同步复制、但也需要数据服务高可用性的客户的要求。例如、如果与远程站点的连接断开、则通常最好让存储系统继续在未复制的状态下运行。

许多同步复制解决方案只能在同步模式下运行。这种类型的全或全不复制有时称为Domino模式。此类存储系统将停止提供数据、而不是允许本地和远程数据副本处于不同步状态。如果强制中断复制、重新同步可能会非常耗时、并且可能会使客户在重新建立镜像期间完全丢失数据。

SyncMirror不仅可以在无法访问远程站点时无缝切换出同步模式、还可以在恢复连接后快速重新同步到RPO = 0状态。远程站点上的陈旧数据副本也可以在重新同步期间保留在可用状态、从而确保本地和远程数据副本始终存在。

如果需要Domino模式、则NetApp提供SnapMirror同步(SM-S)。此外、还存在应用程序级选项、例如Oracle DataGuard或SQL Server Always On可用性组。可以选择操作系统级磁盘镜像。有关追加信息和选项、请咨询您的NetApp或合作伙伴客户团队。

MetroCluster和NVFAIL

NVFAIL是ONTAP中的一项通用数据完整性功能、旨在最大限度地提高数据库的数据完整性保护。

本节将详细介绍基本ONTAP NVFAIL、以涵盖MetroCluster特定的主题。

使用MetroCluster时、写入操作在至少另一个控制器上登录到本地NVRAM和NVRAM后才会得到确认。此方法可确保硬件故障或断电不会导致传输中I/O丢失如果本地NVRAM发生故障或与其他节点的连接发生故障、则无法再镜像数据。

如果本地NVRAM报告错误、则此节点将关闭。此关闭会导致在使用HA对时故障转移到配对控制器。使用MetroCluster时、行为取决于所选的整体配置、但可能会自动故障转移到远程便签。在任何情况下、数据都不会丢失、因为发生故障的控制器尚未确认写入操作。

站点间连接故障会阻止NVRAM复制到远程节点、这种情况更为复杂。写入操作不再复制到远程节点、因此、如果控制器发生灾难性错误、可能会导致数据丢失。更重要的是、在这些情况下尝试故障转移到其他节点会导致数据丢失。

控制因素是NVRAM是否同步。如果NVRAM已同步、则可以安全地进行节点间故障转移、而不会丢失数据。在MetroCluster配置中、如果NVRAM与底层聚合plexes处于同步状态、则可以安全地继续执行切换、而不会丢失数据。

除非强制执行故障转移或切换、否则ONTAP不允许在数据不同步时执行故障转移或切换。以这种方式强制更改条件即表示数据可能会留在原始控制器中、并且数据丢失是可以接受的。

如果强制执行故障转移或切换、则数据库尤其容易受到损坏的影响、因为数据库在磁盘上维护着更大的内部数据缓存。如果发生强制故障转移或切换、先前确认的更改将被有效丢弃。存储阵列的内容会及时有效地向后跳转、数据库缓存的状态不再反映磁盘上数据的状态。

为了保护应用程序免受这种情况的影响、ONTAP允许对卷进行配置、以便针对NVRAM故障提供特殊保护。触发此保护机制后、卷将进入名为NVFAIL的状态。此状态会导致I/O错误、发生原因应用程序会关闭以使其不使用陈旧数据。不应丢失数据、因为存储系统上仍存在任何已确认的写入、对于数据库、任何已提交的事务数据都应出现在日志中。

通常的后续步骤是、管理员先完全关闭主机、然后再手动将LUN和卷重新联机。虽然这些步骤可能涉及一些工作、但这种方法是确保数据完整性的最安全方法。并非所有数据都需要这种保护、这就是可以逐个卷配置NVFAIL行为的原因。

手动强制**NVFAIL**

要强制与分布在各个站点上的应用程序集群(包括VMware、Oracle RAC等)进行切换、最安全的方法是指定 `-force-nvfail-all` 在命令行中。此选项可作为紧急措施使用、以确保所有缓存数据均已转储。如果主机正在使用最初位于发生灾难的站点上的存储资源、则会收到I/O错误或陈旧的文件句柄 (`ESTALE`)错误。Oracle数据库崩溃、文件系统要么完全脱机、要么切换到只读模式。

切换完成后、 `in-nvfailed-state` 标记、并且LUN需要置于联机状态。完成此活动后、可以重新启动数据库。这些任务可以自动执行、以减少RTO。

dr-force－nvfail

作为一般安全措施、请设置 `dr-force-nvfail` 在正常操作期间可能从远程站点访问的所有卷上的标志、表示它们是故障转移之前使用的活动。此设置的结果是、所选远程卷在进入后将不可用 `in-nvfailed-state` 切换期间。切换完成后、 `in-nvfailed-state` 标记、并且LUN必须置于联机状态。完成这些活动后、可以重新启动应用程序。这些任务可以自动执行、以减少RTO。

结果类似于使用 `-force-nvfail-all` 用于手动切换的标志。但是、受影响的卷数量可以仅限于那些必须防止应用程序或具有陈旧缓存的操作系统访问的卷。

对于不使用的环境、有两个关键要求 `dr-force-nvfail` 在应用程序卷上：

- 在主站点丢失后、强制切换的发生时间不得超过30秒。
- 在执行维护任务期间、或者在SyncMirror plexes或NVRAM复制不同步的任何其他情况下、不得发生切换。第一个要求可通过Tieber4软件来满足、该软件配置为在站点发生故障后30秒内执行切换。此要求并不意味着必须在检测到站点故障后30秒内执行切换。这确实意味着、如果自某个站点确认正常运行后30秒内已过、则不再安全地强制执行切换。

如果已知MetroCluster配置不同步、则可以通过禁用所有自动切换功能来部分满足第二项要求。更好的选择是、使用Tiebre机会 解决方案监控NVRAM复制和SyncMirror plexes的运行状况。如果集群未完全同步、则Tieb破碎机不应触发切换。

NetApp MCTB软件无法监控同步状态、因此、如果MetroCluster因任何原因而不同步、则应将其禁用。ClusterLion具有NVRAM监控和丛监控功能、可以将其配置为不触发切换、除非确认MetroCluster系统已完全同步。

Oracle单实例

如前所述、MetroCluster系统的存在并不一定会增加或更改数据库的任何最佳操作实践。客户MetroCluster系统上当前运行的大多数数据库都是单个实例、并遵循Oracle on ONTAP文档中的建议。

使用预配置的操作系统进行故障转移

SyncMirror在灾难恢复站点提供数据的同步副本、但要使数据可用、需要使用操作系统和相关应用程序。基本自动化可以显著缩短整个环境的故障转移时间。通常会使用Veritas Cluster Server (VCS)等集群软件产品在各个站点之间创建集群、在许多情况下、可以使用简单的脚本来驱动故障转移过程。

如果主节点丢失、则会将集群软件(或脚本)配置为在备用站点使数据库联机。一种方法是、创建为构成数据库的NFS或SAN资源预先配置的备用服务器。如果主站点发生故障、则集群软件或脚本化备用站点将执行一系列类似以下内容的操作：

1. 强制执行MetroCluster切换
2. 发现FC LUN (仅限SAN)
3. 挂载文件系统和/或挂载ASM磁盘组
4. 正在启动数据库

此方法的主要要求是在远程站点上运行操作系统。它必须预配置Oracle二进制文件、这也意味着必须在主站点和备用站点上执行Oracle修补等任务。或者、也可以将Oracle二进制文件镜像到远程站点、并在声明发生灾难时进行挂载。

实际激活操作步骤非常简单。LUN发现等命令只需对每个FC端口执行几个命令即可。文件系统挂载只不过是一个 `mount` 命令、数据库和ASM均可通过CLI使用一个命令来启动和停止。如果在切换之前灾难恢复站点未使用卷和文件系统、则无需设置 `dr-force- nvfail` 卷上。

使用虚拟化操作系统进行故障转移

数据库环境的故障转移可以扩展到包括操作系统本身。理论上、这种故障转移可以使用启动LUN来完成、但大多数情况下、这种故障转移是通过虚拟化操作系统来完成的。操作步骤类似于以下步骤：

1. 强制执行MetroCluster切换
2. 挂载托管数据库服务器虚拟机的数据存储库
3. 启动虚拟机
4. 手动启动数据库或将虚拟机配置为自动启动数据库、例如、ESX集群可以跨越多个站点。发生灾难时、可以在切换后将灾难恢复站点上的虚拟机置于联机状态。只要在发生灾难时托管虚拟化数据库服务器的数据存储库未在使用中、就不需要进行设置 `dr-force- nvfail` 在关联卷上。

Oracle Extended RAC

许多客户通过跨站点扩展Oracle RAC集群来优化其RTO、从而形成完全主动-主动配置。整体设计变得更加复杂、因为它必须包括Oracle RAC的仲裁管理。此外、还可以从两个站点访问数据、这意味着强制切换可能会导致使用过时的数据副本。

尽管两个站点上都存在数据副本、但只有当前拥有聚合的控制器才能提供数据。因此、对于扩展RAC集群、远

程节点必须通过站点到站点连接执行I/O。结果会增加I/O延迟、但这种延迟通常不是问题。RAC互连网络还必须跨站点延伸、这意味着无论如何都需要一个高速、低延迟的网络。如果增加的延迟使发生原因出现问题、则可以主动-被动方式运行集群。然后、需要将I/O密集型操作定向到拥有聚合的控制器本地的RAC节点。然后、远程节点会执行较轻的I/O操作、或者纯粹用作热备用服务器。

如果需要双主动扩展RAC、则应考虑使用SnapMirror主动同步代替MetroCluster。SM-AS复制允许首选使用数据的特定副本。因此、可以构建一个扩展RAC集群、在该集群中、所有读取操作都在本地进行。读取I/O不会跨越站点、从而尽可能地降低延迟。所有写入活动仍必须传输站点间连接、但使用任何同步镜像解决方案时、此类流量都是不可避免的。

如果在Oracle RAC中使用启动LUN (包括虚拟化启动磁盘)、则可能需要更改此 `misscount`参数。有关RAC超时参数的详细信息，请参阅"采用ONTAP的Oracle RAC"。

双站点配置

双站点扩展RAC配置可以提供主动-主动数据库服务、这些服务可以在许多(并非所有)灾难情形下无系统地经受住。

RAC投票文件

在MetroCluster上部署扩展RAC时、首要考虑事项应该是仲裁管理。Oracle RAC有两种管理仲裁的机制：磁盘检测信号和网络检测信号。磁盘检测信号可使用表决文件监控存储访问。对于单站点RAC配置、只要底层存储系统提供HA功能、单个表决资源就足够了。

在早期版本的Oracle中、投票文件放置在物理存储设备上、但在当前版本的Oracle中、投票文件存储在ASM磁盘组中。

NFS支持Oracle RAC。在网格安装过程中、会创建一组ASM进程、以将网格文件使用的NFS位置显示为ASM磁盘组。此过程对最终用户几乎是透明的、安装完成后无需持续进行ASM管理。

双站点配置的第一个要求是、确保每个站点始终可以访问一半以上的表决文件、并确保灾难恢复过程不会中断。在表决文件存储在ASM磁盘组中之前、此任务非常简单、但如今管理员需要了解ASM冗余的基本原则。

ASM磁盘组有三个冗余选项 `external`， `normal`，和 `high`。换言之、未镜像、镜像和三向镜像。名为的新选项 `Flex` 也可用、但很少使用。冗余设备的冗余级别和放置位置控制了在故障情形下发生的情况。例如：

- 将表决文件放置在上 `diskgroup` 使用 `external` 冗余资源可确保在站点间连接断开时逐出一个站点。
- 将表决文件放置在上 `diskgroup` 使用 `normal` 每个站点只有一个ASM磁盘的冗余可确保在站点间连接断开时在两个站点上逐出节点、因为两个站点都不会有多数仲裁。
- 将表决文件放置在上 `diskgroup` 使用 `high` 如果一个站点上有两个磁盘、而另一个站点上有一个磁盘、则可以在两个站点均正常运行且可相互访问时执行主动-主动操作。但是、如果单磁盘站点与网络隔离、则该站点将被逐出。

RAC网络检测信号

Oracle RAC网络检测信号可监控集群互连中的节点可访问情况。要保留在集群中、一个节点必须能够与一半以上的其他节点联系。在双站点架构中、此要求会为RAC节点数创建以下选项：

- 如果在每个站点上放置相同数量的节点、则会在网络连接断开时在一个站点上执行逐出。
- 将N个节点放置在一个站点上、而将N+1个节点放置在另一个站点上、可以确保站点间连接断开会导致站点

中剩余的网络仲裁节点数量增加、而将节点数量减少。

在Oracle 12cR2之前的版本中、无法控制站点丢失期间哪一端会发生逐出。如果每个站点的节点数相等、则逐出操作由主节点控制、主节点通常是要启动的第一个RAC节点。

Oracle 12cR2引入了节点加权功能。通过此功能、管理员可以更好地控制Oracle如何解决脑裂问题。例如、以下命令可为RAC中的特定节点设置首选项：

```
[root@host-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.
```

重新启动Oracle高可用性服务后、配置如下所示：

```
[root@host-a lib]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
```

Node `host-a` 现在指定为关键服务器。如果两个RAC节点彼此隔离、 `host-a` 不会影响、和 `host-b` 被逐出。

有关完整的详细信息、请参见Oracle白皮书《Oracle Clusterware 12c Release 2 Technical Overview》。"

对于12cR2之前的Oracle RAC版本、可通过按如下所示检查CRS日志来识别主节点：

```
[root@host-a ~]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
 [root@host-a ~]# grep -i 'master node' /grid/diag/crs/host-
a/crs/trace/crsd.trc
2017-05-04 04:46:12.261525 :   CRSSE:2130671360: {1:16377:2} Master Change
Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:01:24.979716 :   CRSSE:2031576832: {1:13237:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
2017-05-04 05:11:22.995707 :   CRSSE:2031576832: {1:13237:221} Master
Change Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:28:25.797860 :   CRSSE:3336529664: {1:8557:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
```

此日志指示主节点为 2 和节点 host-a ID为 1。这一事实意味着 host-a 不是主节点。可以使用命令确认主节点的标识 olsnodes -n。

```
[root@host-a ~]# /grid/bin/olsnodes -n
host-a  1
host-b  2
```

ID为的节点 2 为 host-b，即主节点。在每个站点上具有相同节点数的配置中、站点使用 host-b 是指在两组因任何原因丢失网络连接时仍可正常运行的站点。

标识主节点的日志条目可能会在系统中过期。在这种情况下、可以使用Oracle集群注册表(OCR)备份的时间戳。

```
[root@host-a ~]#  /grid/bin/ocrconfig -showbackup
host-b     2017/05/05 05:39:53     /grid/cdata/host-cluster/backup00.ocr
0
host-b     2017/05/05 01:39:53     /grid/cdata/host-cluster/backup01.ocr
0
host-b     2017/05/04 21:39:52     /grid/cdata/host-cluster/backup02.ocr
0
host-a     2017/05/04 02:05:36     /grid/cdata/host-cluster/day.ocr     0
host-a     2017/04/22 02:05:17     /grid/cdata/host-cluster/week.ocr     0
```

此示例显示主节点为 host-b。此外、它还表示主节点与发生了变化 host-a to host-b 5月4日2：05到21：39之间的某个时间。只有在检查了CRS日志后、才能安全地使用这种标识主节点的方法、因为主节点可能在上次OCR备份后发生更改。如果发生了此更改、则OCR日志中应该会显示此更改。

大多数客户都选择一个投票磁盘组来为整个环境提供服务、并在每个站点上选择相同数量的RAC节点。磁盘组应放置在数据库所在的站点上。其结果是、连接断开会导致在远程站点上发生逐出。远程站点将不再具有仲裁、也无法访问数据库文件、但本地站点仍会照常运行。恢复连接后、远程实例可以重新联机。

发生灾难时、需要执行切换、以使运行正常的站点上的数据库文件和表决磁盘组联机。如果灾难允许AUSO触发切换、则不会触发NVFAIL、因为集群已知处于同步状态、并且存储资源正常联机。此操作速度非常快、应在之前完成 disktimeout 期限到期。

由于只有两个站点、因此无法使用任何类型的自动外部中断软件、这意味着强制切换必须手动操作。

三站点配置

使用三个站点构建扩展RAC集群更容易。托管MetroCluster系统一半的两个站点也支持数据库工作负载、而第三个站点则充当数据库和MetroCluster系统的断路器。Oracle TiebREAKER配置可能非常简单、只需将ASM磁盘组的一个成员放置在第三个站点上即可进行表决、也可能包括在第三个站点上运行的实例、以确保RAC集群中的节点数为奇数。

有关在扩展RAC配置中使用NFS的重要信息、请参阅Oracle文档中的"Quorum Failure group"(仲裁故障组)。总之、可能需要修改NFS挂载选项以包括软选项、以确保与托管仲裁资源的第三站点断开连接不会挂起主Oracle服务器或Oracle RAC进程。

SnapMirror活动同步

概述

通过SnapMirror主动同步、您可以构建超高可用性Oracle数据库环境、其中LUN可从两个不同的存储集群访问。

使用SnapMirror主动同步时、不存在数据的"主"和"二级"副本。每个集群都可以从其本地数据副本提供读取IO、并且每个集群都会向其配对集群复制写入。结果是对称IO行为。

除其他选项外、此选项还允许您将Oracle RAC作为扩展集群运行、并在两个站点上运行操作实例。或者、您也可以构建RPO = 0主动-被动数据库集群、在站点中断期间、可以在站点间移动单实例数据库、并且可以通过Pacemker或VMware HA等产品自动执行此过程。所有这些选项的基础都是由SnapMirror主动同步管理的同步复制。

同步复制

在正常操作下、SnapMirror主动同步始终提供RPO = 0的同步副本、但有一个例外。如果无法复制数据、则ONTAP将不再需要复制数据并恢复在一个站点上提供IO、而另一个站点上的LUN将脱机。

存储硬件

与其他存储灾难恢复解决方案不同、SnapMirror主动同步可提供非对称平台灵活性。每个站点的硬件不必相同。通过此功能、您可以调整用于支持SnapMirror活动同步的硬件的大小。如果需要支持完整的生产工作负载、远程存储系统可以与主站点完全相同；但是、如果灾难导致I/O减少、则与远程站点上较小的系统相比、可能会更经济高效。

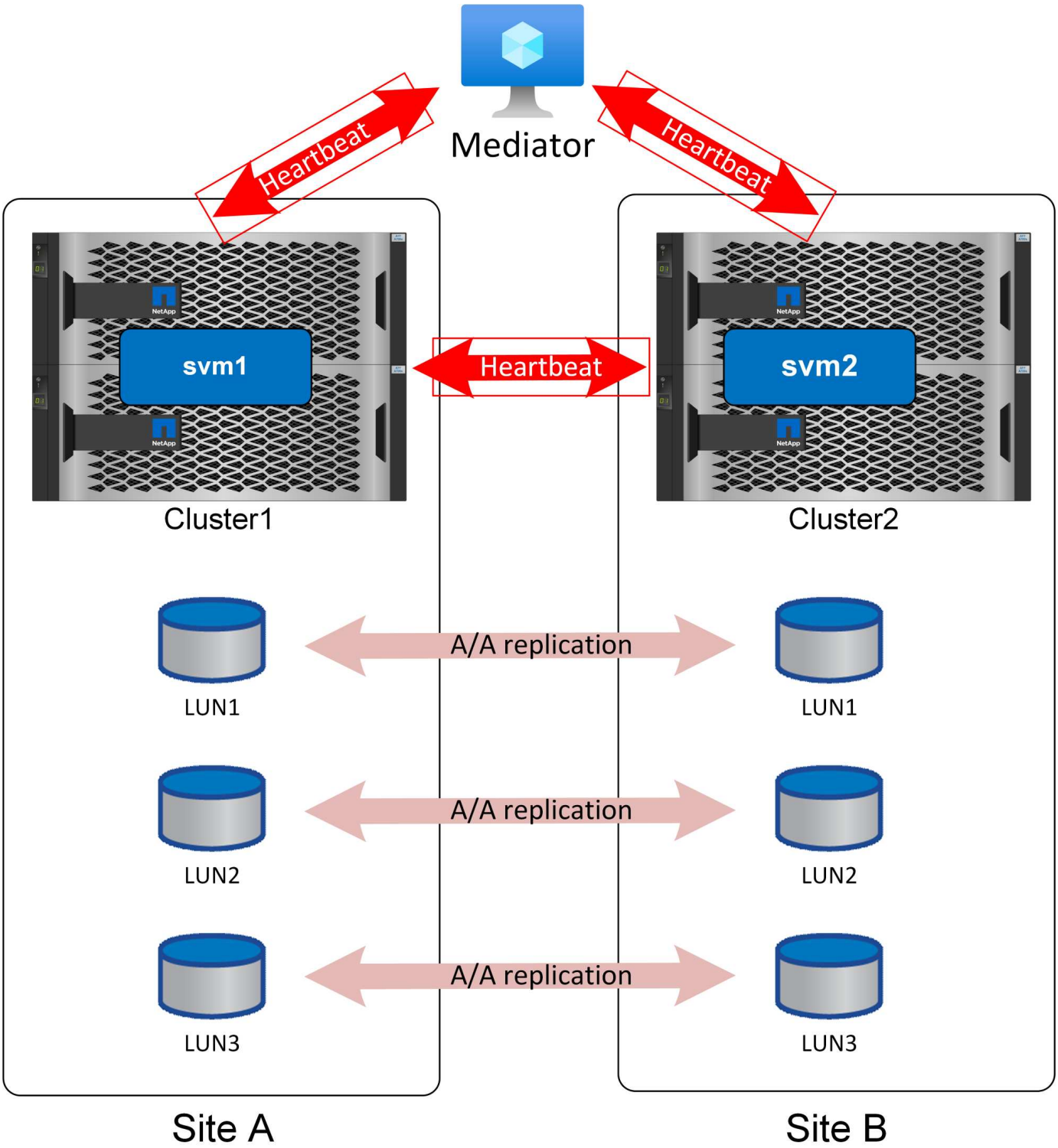
ONTAP调解器

ONTAP调解器是从NetApp支持下载的软件应用程序、通常部署在小型虚拟机上。与SnapMirror活动同步结合使用时、ONTAP调解器不是Tieb破碎 器。它是参与SnapMirror活动同步复制的两个集群的备用通信通道。ONTAP根据通过直接连接和调解器从合作伙伴处收到的响应来推动自动化操作。

ONTAP 调解器

要安全地自动执行故障转移、需要使用调解器。理想情况下、它会放置在独立的第三个站点上、但如果与参与复制的集群之一主机代管、它仍可满足大多数需求。

调解员实际上并不是打破僵局的人，尽管它实际上发挥了这样的作用。中介器有助于确定集群节点的状态，并在站点发生故障时协助自动切换过程。在任何情况下，中介都不会传输数据。



自动化故障转移的第一大挑战是脑裂问题、如果两个站点彼此断开连接、就会出现该问题。应该发生什么？您不希望让两个不同的站点将自己指定为数据的无故障副本、但单个站点如何区分实际丢失相对站点与无法与相反站点通信之间的区别？

这是调解者进入画面的地方。如果放置在第三个站点上、并且每个站点都与该站点建立了单独的网络连接、则每个站点都有一条额外的路径来验证另一个站点的运行状况。再次查看上图、并考虑以下情形。

- 如果调解器发生故障或无法从一个或两个站点访问、会发生什么情况？

- 两个集群仍可通过复制服务所使用的同一链路彼此通信。
- 数据仍会提供RPO = 0保护

- 如果站点A发生故障、会发生什么情况?
  - 站点B将看到两个通信通道关闭。
  - 站点B将接管数据服务、但没有RPO = 0镜像
- 如果站点B发生故障、会发生什么情况?
  - 站点A将看到两个通信通道关闭。
  - 站点A将接管数据服务、但没有RPO = 0镜像

还需要考虑另一种情形：丢失数据复制链路。如果站点之间的复制链路丢失、显然无法执行RPO = 0镜像。那么应该发生什么呢?

这由首选站点状态控制。在SM-AS关系中、其中一个站点是另一个站点的二级站点。这对正常操作没有影响、并且所有数据访问都是对称的、但是如果复制中断、则必须断开连接才能恢复操作。结果是、首选站点将继续操作而不进行镜像、而二级站点将暂停IO处理、直到复制通信恢复为止。

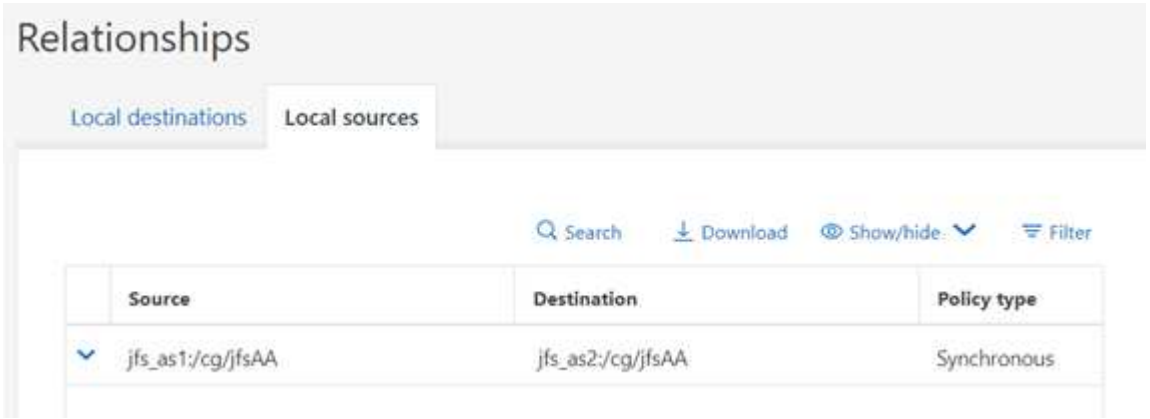
SnapMirror主动同步首选站点

SnapMirror主动同步行为是对称的、但有一个重要例外-首选站点配置。

SnapMirror主动同步会将一个站点视为"源"、而将另一个站点视为"目标"。这意味着单向复制关系、但这不适用于IO行为。复制是双向的、对称的、镜像两端的IO响应时间相同。

该 `source`名称用于控制首选站点。如果复制链路丢失、则源副本上的LUN路径将继续提供数据、而目标副本上的LUN路径将变得不可用、直到重新建立复制并使SnapMirror重新进入同步状态为止。然后、这些路径将恢复提供数据。

可通过SystemManager查看源/目标配置：



或在命令行界面上：

```
Cluster2::> snapmirror show -destination-path jfs_as2:/cg/jfsAA

                           Source Path: jfs_as1:/cg/jfsAA
                     Destination Path: jfs_as2:/cg/jfsAA
                    Relationship Type: XDP
              Relationship Group Type: consistencygroup
                  SnapMirror Schedule: -
               SnapMirror Policy Type: automated-failover-duplex
                    SnapMirror Policy: AutomatedFailOverDuplex
                          Tries Limit: -
                    Throttle (KB/sec): -
                         Mirror State: Snapmirrored
                  Relationship Status: InSync
```

关键在于源是位于第一个Storage Virtual Machine上的SVM。如上所述、术语"源"和"目标"并不表示复制的数据流。两个站点都可以处理写入并将其复制到相反站点。实际上、两个集群都是源和目标。将一个集群指定为源集群的效果只是控制在复制链路丢失时哪个集群作为读写存储系统继续存在。

网络拓扑

统一访问

统一访问网络意味着主机能够访问两个站点(或同一站点中的故障域)上的路径。

SM-AS的一项重要功能是、可以对存储系统进行配置、使其知道主机所在的位置。将LUN映射到给定主机时、您可以指示它们是否接近给定存储系统。

邻近设置

接近是指每个集群的配置、表示特定主机WWN或iSCSI启动程序ID属于本地主机。这是配置LUN访问的第二个可选步骤。

第一步是常规的igrop配置。每个LUN都必须映射到一个igrop、该igrop包含需要访问该LUN的主机的wwn/iSCSI ID。此选项用于控制哪个主机对LUN具有_access_访问权限。

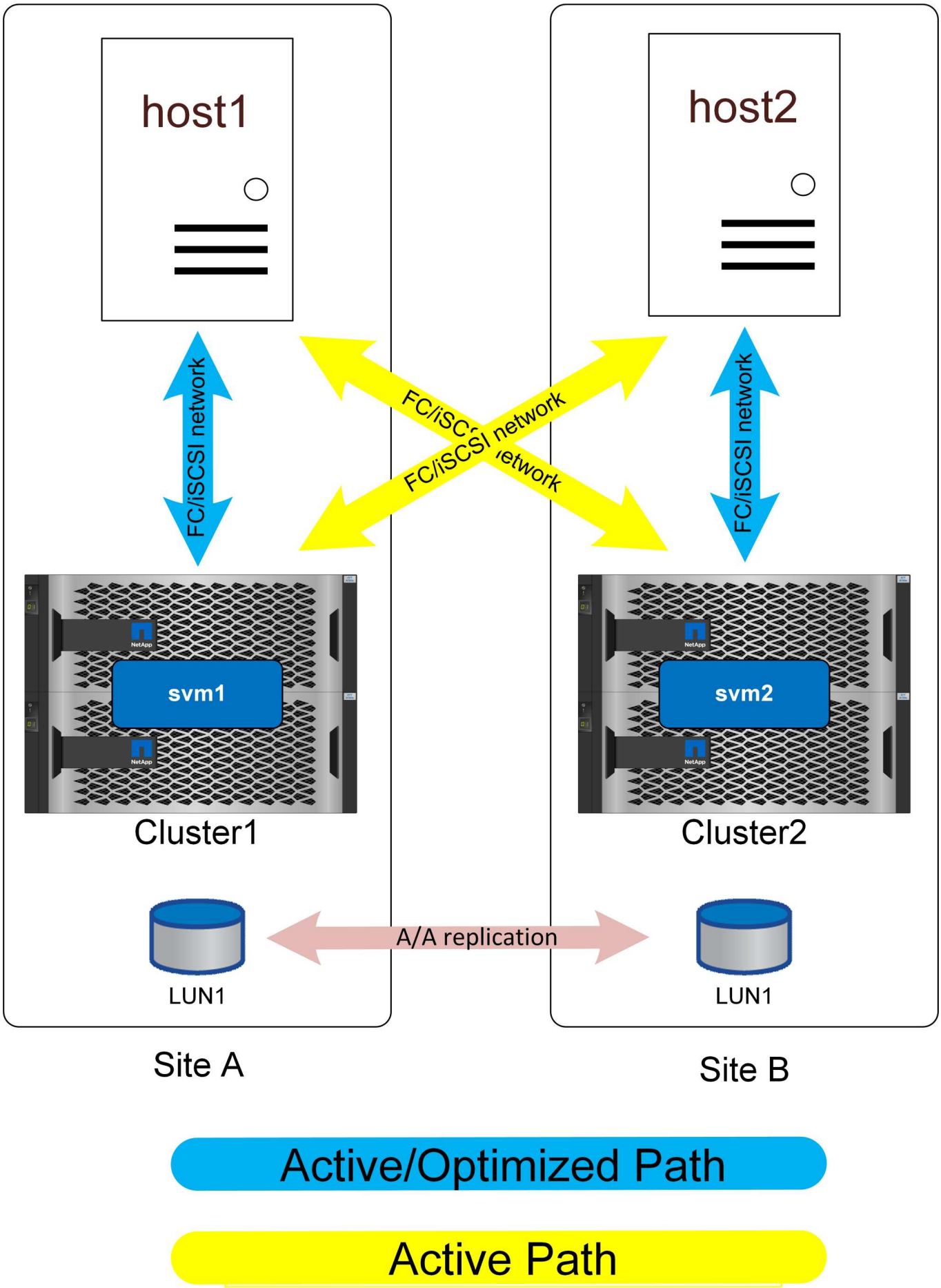
第二个可选步骤是配置主机邻近性。这不控制访问、而是控制_priority_。

例如、可以将站点A的主机配置为访问受SnapMirror活动同步保护的LUN、并且由于SAN跨站点扩展、因此可以使用站点A上的存储或站点B上的存储为该LUN提供路径

如果没有邻近设置、则该主机将平等使用这两个存储系统、因为这两个存储系统都会公布主动/优化路径。如果站点之间的SAN延迟和/或带宽有限、则可能无法实现这一点、您可能希望确保在正常操作期间、每个主机优先使用指向本地存储系统的路径。这可通过将主机的wwn/iSCSI ID作为近端主机添加到本地集群来配置。可通过命令行界面或SystemManager完成此操作。

AFF

对于AFF系统、配置主机邻近性后、路径将如下所示。


host1
host2
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
svm1
svm2
Cluster1
Cluster2
A/A replication
LUN1
LUN1
Site A
Site B
Active/Optimized Path
Active Path

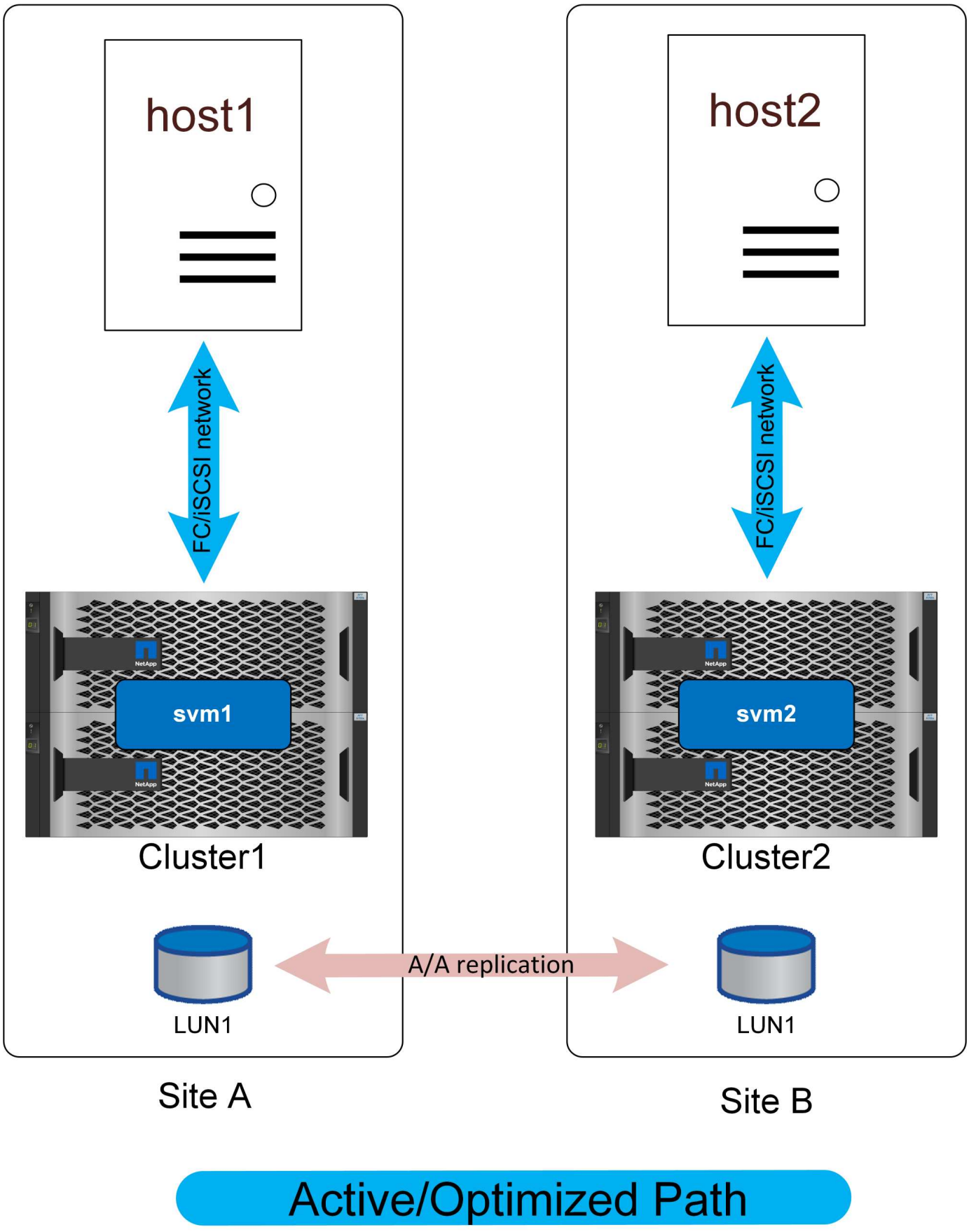
在正常操作下、所有IO均为本地IO。读取和写入操作由本地存储阵列提供。当然、在确认写入IO之前、本地控制器也需要将其复制到远程系统、但所有读取IO都将在本地进行处理、并且不会通过遍历站点间的SAN链路而产生额外延迟。

只有在所有主动/优化路径丢失时、才会使用非优化路径。例如、如果站点A上的整个阵列断电、则站点A上的主机仍可访问站点B上阵列的路径、因此、尽管延迟较长、但仍可保持正常运行。

为了简单起见、这些图中未显示通过本地集群的冗余路径。ONTAP存储系统本身就是HA、因此控制器故障不应导致站点故障。它只会导致受影响站点上使用的本地路径发生更改。

ASA

NetApp ASA系统可在集群上的所有路径之间提供主动-主动多路径功能。这也适用于SM-AS配置。



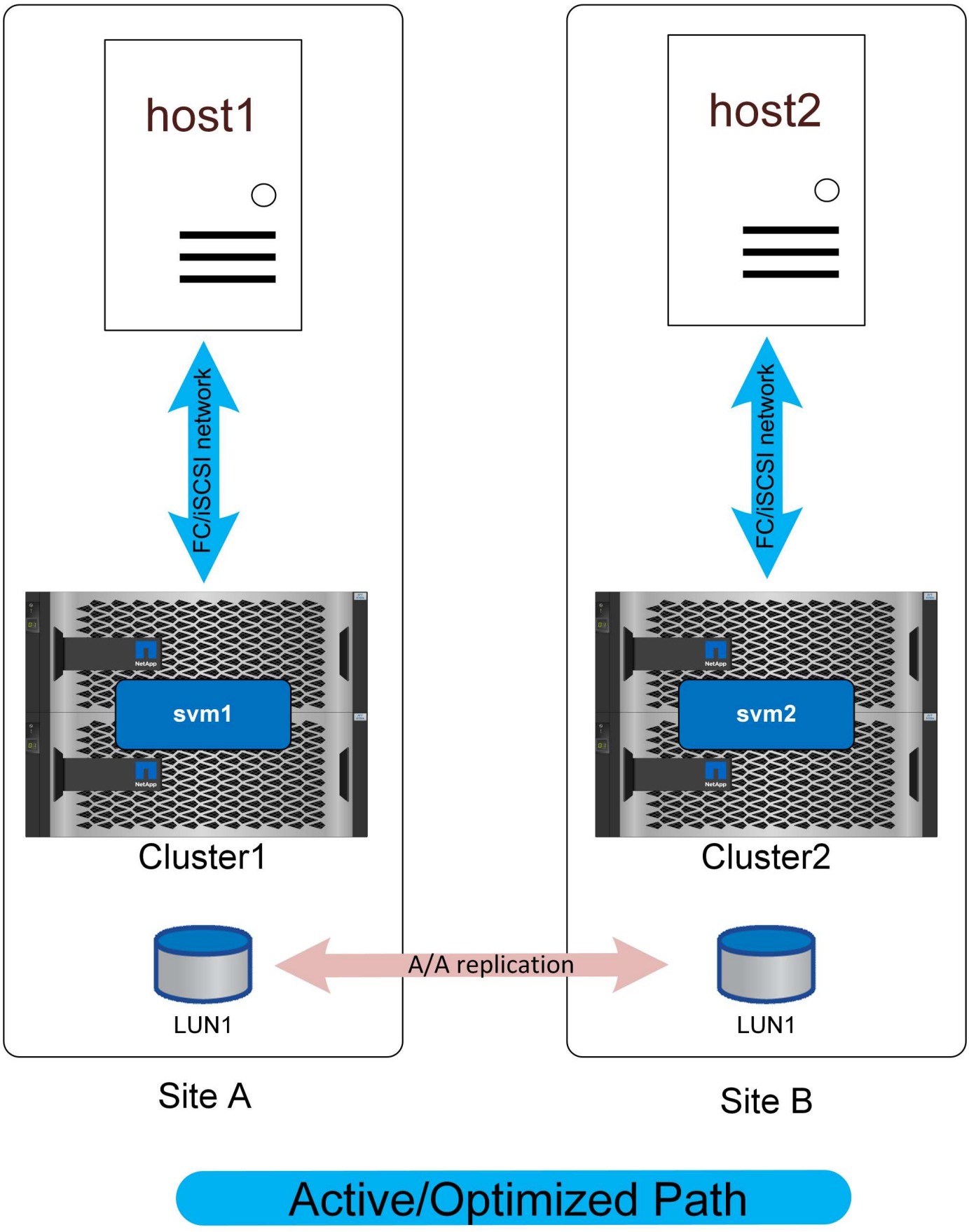
使用非一致访问的ASA配置的工作原理与使用AFF时大致相同。使用统一访问时、IO将跨越WAN。这可能是可取的、也可能不可取。

如果两个站点之间相距100米且具有光纤连接、则WAN上不会出现可检测到的额外延迟、但如果两个站点相距较远、则两个站点上的读取性能都会受到影响。相比之下、使用AFF时、只有在没有可用的本地路径时才会使用这些WAN交叉路径、而且由于所有IO都是本地IO、因此日常性能会更好。使用非一致访问网络的ASA可以获得ASA的成本和功能优势、而不会造成跨站点延迟访问损失。

在低延迟配置中使用SM-A的ASA具有两个有趣的优势。首先、从本质上说、它可以将任何一台主机的性能提高一倍、因为使用两倍路径的控制器可以为IO提供服务。其次、在单站点环境中、它可以提供极高的可用性、因为整个存储系统可能会丢失、而不会中断主机访问。

非一致访问

非一致访问网络意味着每个主机只能访问本地存储系统上的端口。SAN不会跨站点(或同一站点内的故障域)进行扩展。



这种方法的主要优势是SAN的精简性、您无需在网络上延伸SAN。某些客户的站点间连接延迟不足、或者缺少通过站点间网络传输FC SAN流量的基础架构。

非一致访问的缺点是、某些故障情形(包括丢失复制链路)将导致某些主机无法访问存储。如果本地存储连接丢失、则作为单个实例运行的应用程序(例如、本质上仅在任何给定挂载的单个主机上运行的非集群数据库)将失败。数据仍会受到保护、但数据库服务器将无法再访问。它需要在远程站点上重新启动、最好是通过自动化过程重新启动。例如、VMware HA可以在一台服务器上检测到全路径关闭的情况、并在具有可用路径的另一台服务器上重新启动VM。

相比之下、Oracle RAC等集群应用程序可以提供在两个不同站点上同时提供的服务。丢失站点并不意味着整个应用程序服务都会丢失。实例仍可用、并且在正常运行的站点上运行。

在许多情况下、通过站点间链路访问存储的应用程序所产生的额外延迟开销是不可接受的。这意味着统一网络可用性的提高微乎其微、因为如果站点上丢失存储、则无论如何都需要关闭故障站点上的服务。

> 为了简单起见、这些图中未显示通过本地集群的冗余路径。ONTAP存储系统本身就是HA、因此控制器故障不应导致站点故障。它只会导致受影响站点上使用的本地路径发生更改。

Oracle配置

概述

使用SnapMirror主动同步不一定会增加或更改数据库操作的任何最佳实践。

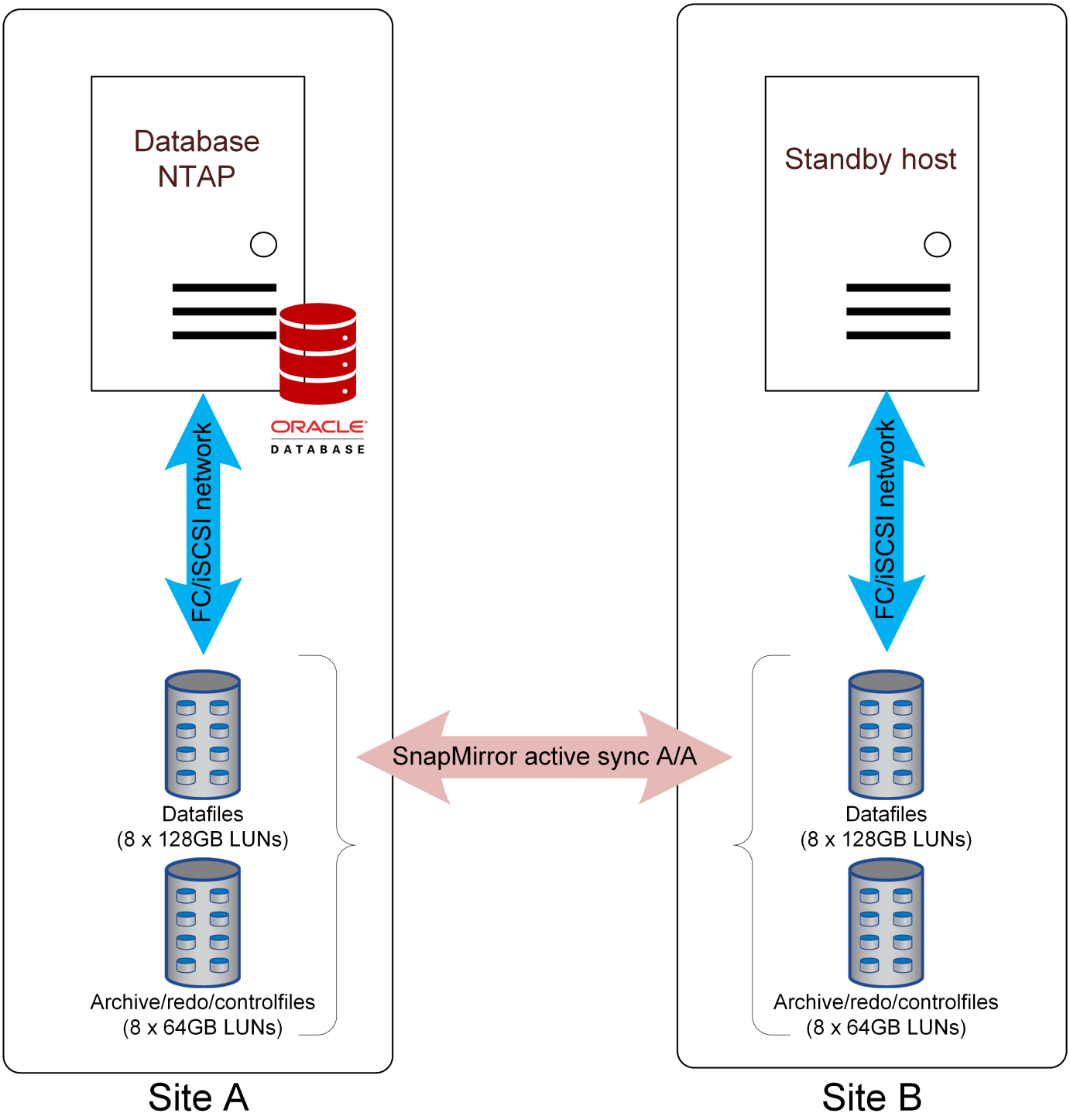
最佳架构取决于业务需求。例如、如果目标是使RPO=0防止数据丢失、但RTO较宽松、则使用Oracle单实例数据库并使用SM-AS复制LUN可能足以满足Oracle许可标准的要求、而且成本更低。远程站点故障不会中断操作、丢失主站点将导致运行正常的站点上的LUN处于联机状态并可供使用。

如果要对RTO进行更严格的配置、则可以通过脚本或PacMaker或Ansv可 等工具实现基本的主动-被动自动化、从而缩短故障转移时间。例如、可以对VMware HA进行配置、使其检测主站点上的VM故障、并激活远程站点上的VM。

最后、为了实现极快的故障转移、可以跨站点部署Oracle RAC。RTO基本上为零、因为数据库将始终处于联机状态、并可在两个站点上使用。

Oracle单实例

下面的示例介绍了使用SnapMirror活动同步复制部署Oracle单实例数据库的许多选项中的一些选项。



使用预配置的操作系统进行故障转移

SnapMirror主动同步可为灾难恢复站点上的数据提供同步副本、但要使数据可用、需要使用操作系统和相关应用程序。基本自动化可以显著缩短整个环境的故障转移时间。通常会使用PacMaker等集群软件产品在站点间创建集群、在许多情况下、可以使用简单的脚本来执行故障转移过程。

如果主节点丢失、则集群软件(或脚本)将使备用站点上的数据库联机。一种选择是、创建为构成数据库的SAN资源预先配置的备用服务器。如果主站点发生故障、则集群软件或脚本化备用站点将执行一系列类似以下内容的操作：

1. 检测主站点故障
2. 发现FC或iSCSI LUN
3. 挂载文件系统和/或挂载ASM磁盘组
4. 正在启动数据库

此方法的主要要求是在远程站点上运行操作系统。它必须预配置Oracle二进制文件、这也意味着必须在主站点和备用站点上执行Oracle修补等任务。或者、也可以将Oracle二进制文件镜像到远程站点、并在声明发生灾难时进行挂载。

实际激活操作步骤非常简单。LUN发现等命令只需对每个FC端口执行几个命令即可。文件系统挂载只不过是一个 `mount`命令、数据库和ASM都可以通过命令行界面使用一个命令来启动和停止。

使用虚拟化操作系统进行故障转移

数据库环境的故障转移可以扩展到包括操作系统本身。理论上、这种故障转移可以使用启动LUN来完成、但大多数情况下、这种故障转移是通过虚拟化操作系统来完成的。操作步骤类似于以下步骤：

1. 检测主站点故障
2. 挂载托管数据库服务器虚拟机的数据存储库
3. 启动虚拟机
4. 手动启动数据库或将虚拟机配置为自动启动数据库。

例如、ESX集群可以跨越多个站点。发生灾难时、可以在切换后将灾难恢复站点上的虚拟机置于联机状态。

存储故障保护

上图显示了的使用"非一致访问"，其中SAN不会跨站点延伸。这可能更易于配置、在某些情况下、这可能是当前SAN功能的唯一选项、但也意味着主存储系统故障将导致数据库中断、直到应用程序进行故障转移为止。

为了提高故障恢复能力，可以使用部署该解决方案"统一访问"。这将允许应用程序使用从另一站点广告的路径继续运行。

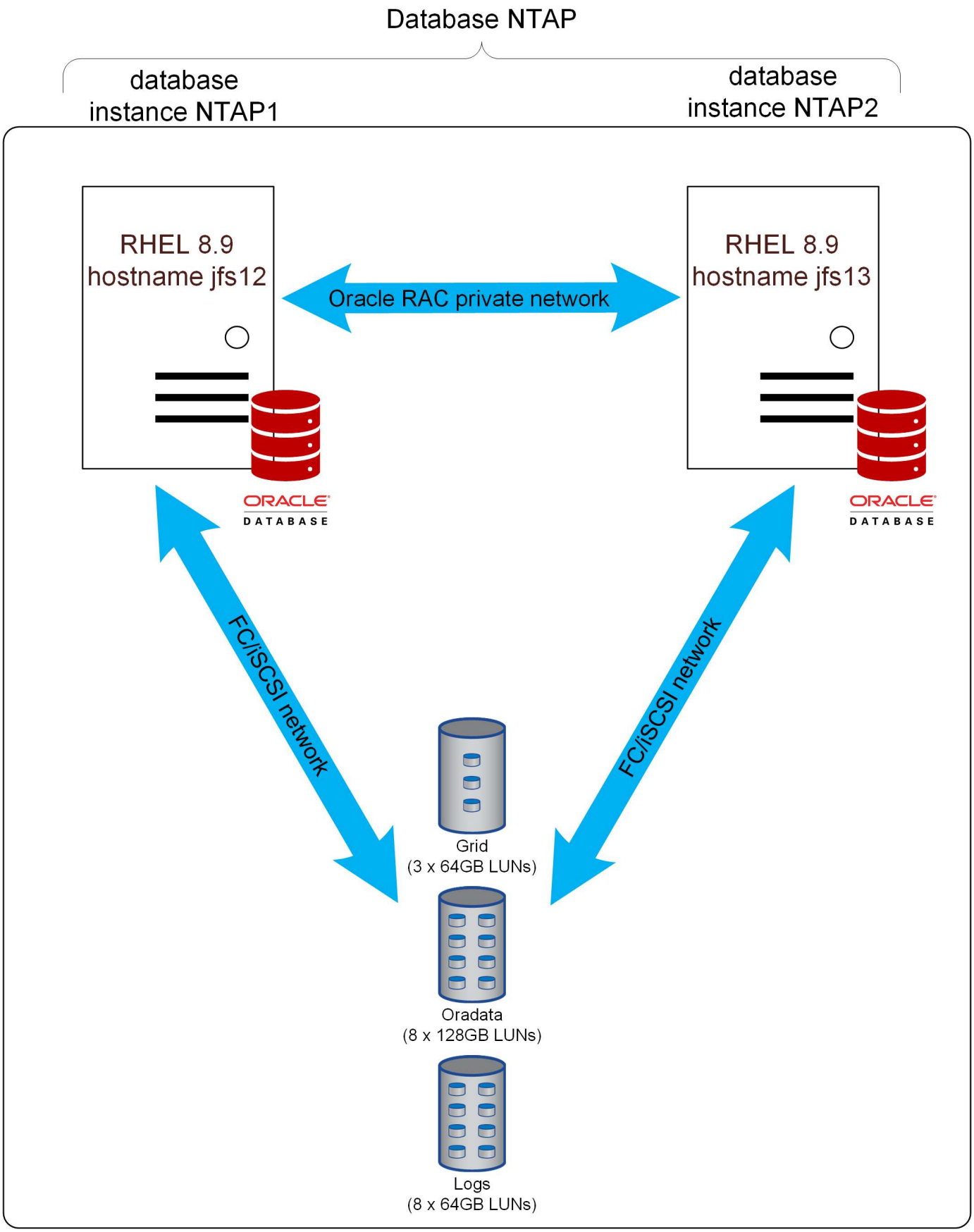
Oracle Extended RAC

许多客户通过跨站点扩展Oracle RAC集群来优化其RTO、从而形成完全主动-主动配置。整体设计变得更加复杂、因为它必须包括Oracle RAC的仲裁管理。

传统的扩展RAC集群模式依靠ASM镜像来提供数据保护。这种方法有效、但也需要大量手动配置步骤、并会给网络基础架构带来开销。相比之下、让SnapMirror主动同步负责数据复制可以显著简化解决方案。同步、中断后重新同步、故障转移和仲裁管理等操作更加简单、而且SAN不需要分布在多个站点上、从而简化了SAN的设计和管理。

Replication

要了解SnapMirror主动同步上的RAC功能、关键在于将存储视为镜像存储上托管的一组LUN。例如：



没有主副本或镜像副本。从逻辑上讲、每个LUN只有一个副本、并且该LUN可在两个不同存储系统上的SAN路径上使用。从主机角度来看、不会发生存储故障转移、而是会发生路径更改。各种故障事件可能会导致LUN的某些路径丢失、而其他路径仍保持联机状态。SnapMirror主动同步可确保在所有操作路径中提供相同的数据。

存储配置

在此示例配置中、ASM磁盘的配置与企业存储上任何单站点RAC配置中的配置相同。由于存储系统提供数据保护、因此会使用ASM外部冗余。

统一访问与非通知访问

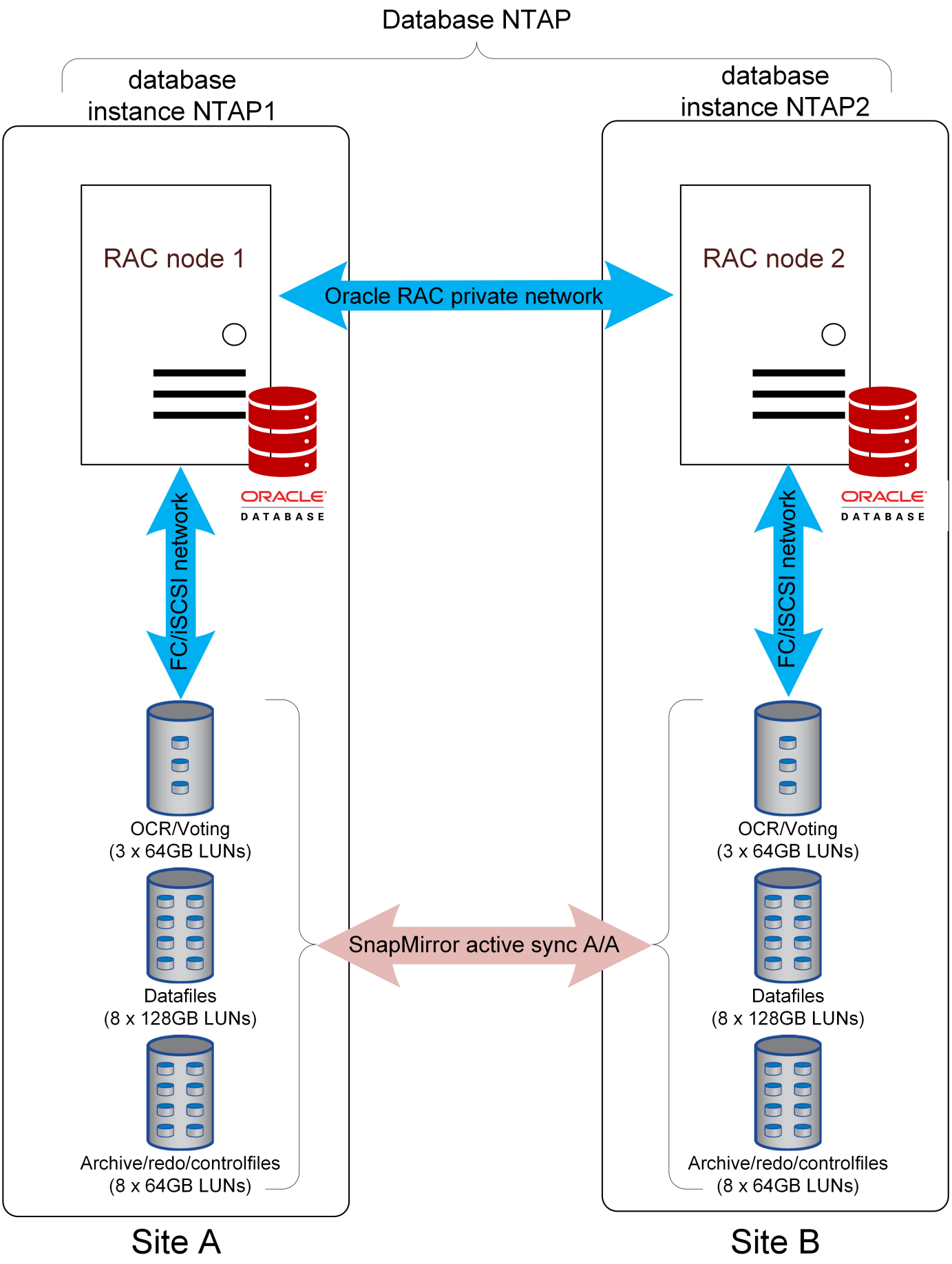
在SnapMirror主动同步模式下使用Oracle RAC最重要的注意事项是使用统一访问还是非统一访问。

统一访问意味着每个主机都可以看到两个集群上的路径。非一致访问表示主机只能查看本地集群的路径。

这两个选项都不是特别建议的、也不建议采用。有些客户可以随时使用暗光纤连接站点、而有些客户则没有这种连接、或者他们的SAN基础架构不支持远程ISL。

非一致访问

从SAN的角度来看、非一致性访问更易于配置。


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
ORACLE DATABASE
ORACLE DATABASE
FC/iSCSI network
FC/iSCSI network
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

此方法的主要缺点"非一致访问"是、站点间ONTAP连接断开或存储系统丢失将导致一个站点上的数据库实例丢失。这显然不是理想的做法、但作为交换更简单的SAN配置、这种风险可能是可以接受的。

统一访问

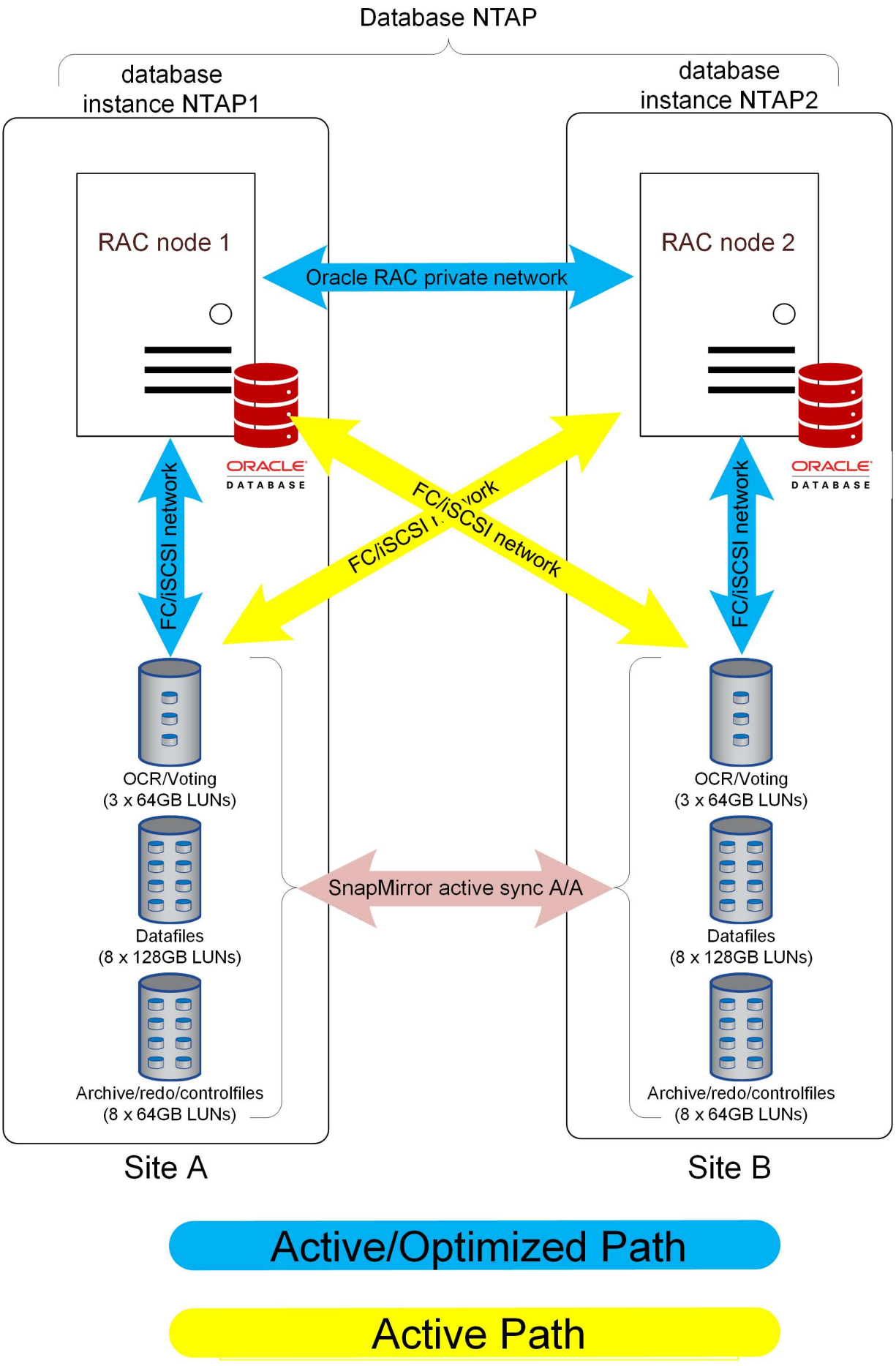
统一访问要求跨站点扩展SAN。主要优势是、丢失存储系统不会导致数据库实例丢失。相反、它会导致当前正在使用的路径发生多路径更改。

可以通过多种方式配置非一致性访问。

> 在下图中、还存在一些在简单控制器故障期间会使用的活动但非优化路径、但为了简化图示、这些路径不会显示出来。

具有邻近设置的**AFF**

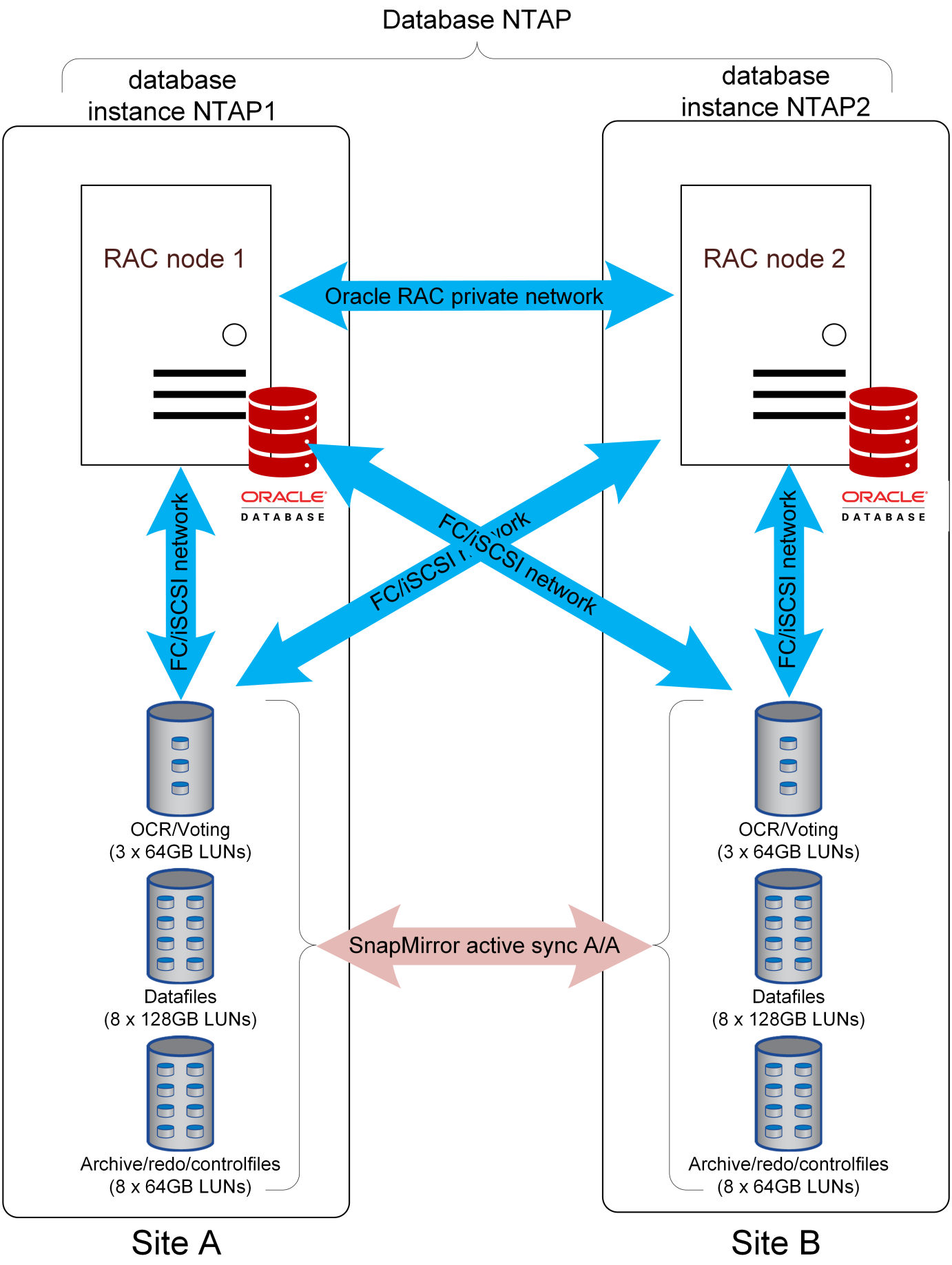
如果站点间延迟较长、则可以为AFF系统配置主机邻近设置。这样、每个存储系统就可以了解哪些主机是本地主机、哪些主机是远程主机、并相应地分配路径优先级。


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B
Active/Optimized Path
Active Path

在正常操作下、每个Oracle实例都会优先使用本地主动/优化路径。这样、所有读取操作都将由块的本地副本处理。这样可以尽可能地降低延迟。写入IO也会通过路径向下发送到本地控制器。在确认之前、仍然必须复制IO、因此、通过站点到站点网络仍会产生额外的延迟、但在同步复制解决方案中无法避免这种情况。

不带邻近设置的**ASA / AFF**

如果站点之间没有明显延迟、则可以在不配置主机邻近设置的情况下配置AFF系统、也可以使用ASA。


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

每个主机都可以使用两个存储系统上的所有操作路径。这样、每个主机就可以利用两个集群(而不仅仅是一个集群)的性能潜能、从而显著提高性能。

使用ASA时、不仅会将两个集群的所有路径视为活动路径并进行了优化、而且配对控制器上的路径也会处于活动状态。结果将始终是整个集群上的全活动SAN路径。

ASA系统也可用于非统一访问配置。由于不存在跨站点路径、因此IO跨越ISL不会对性能产生任何影响。

RAC Tieb破碎 机

虽然使用SnapMirror主动同步的扩展RAC在IO方面是对称架构、但有一个例外情况是连接到脑裂管理。

如果复制链路丢失且两个站点都没有仲裁、会发生什么情况？应该发生什么？此问题既适用于Oracle RAC、也适用于ONTAP行为。如果无法在各个站点之间复制更改、而您希望恢复操作、则其中一个站点必须继续运行、而另一个站点必须不可用。

"ONTAP 调解器"可在ONTAP层满足此要求。RAC分Tieb破碎 功能有多个选项。

Oracle Tieburkers

管理脑裂Oracle RAC风险的最佳方法是使用奇数个RAC节点、最好使用第三个站点的Tieb破碎 机。如果第三个站点不可用、则可以将Tieb破碎 机实例放置在两个站点中的一个站点上、从而有效地将其指定为首选的幸存站点。

Oracle和CSS_critical

如果节点数为偶数、则默认Oracle RAC行为是、集群中的一个节点将被视为比其他节点更重要。具有较高优先级节点的站点将不受站点隔离的影响、而另一站点上的节点将被逐出。优先级基于多个因素、但您也可以使用设置来控制此行为 `css_critical`。

在该架构中"示例"、RAC节点的主机名是jfs12和jfs13。的当前设置 `css_critical`如下：

```
[root@jfs12 ~]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.

[root@jfs13 trace]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.
```

如果要将带有jfs12的站点作为首选站点、请在站点A节点上将此值更改为yes、然后重新启动服务。

```
[root@jfs12 ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.

[root@jfs12 ~]# /grid/bin/crsctl stop crs
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'jfs12'
CRS-2673: Attempting to stop 'ora.crsd' on 'jfs12'
CRS-2790: Starting shutdown of Cluster Ready Services-managed resources on
server 'jfs12'
CRS-2673: Attempting to stop 'ora.ntap.ntappdb1.pdb' on 'jfs12'
…
CRS-2673: Attempting to stop 'ora.gipcd' on 'jfs12'
CRS-2677: Stop of 'ora.gipcd' on 'jfs12' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'jfs12' has completed
CRS-4133: Oracle High Availability Services has been stopped.

[root@jfs12 ~]# /grid/bin/crsctl start crs
CRS-4123: Oracle High Availability Services has been started.
```

故障情形

概述

要规划完整的SnapMirror主动同步应用程序架构、需要了解SM-AS如何在各种计划内和计划外故障转移场景中做出响应。

在以下示例中、假设站点A已配置为首选站点。

复制连接丢失

如果SM-AS复制中断、则无法完成写入IO、因为集群无法将更改复制到相反站点。

站点**A (**首选站点**)**

首选站点上的复制链路故障会导致写入IO处理暂停大约15秒、因为ONTAP会在确定复制链路确实无法访问之前重试复制的写入操作。15秒后、站点A系统将恢复读取和写入IO处理。SAN路径不会更改、LUN将保持联机状态。

站点 **B**

由于站点B不是SnapMirror主动同步首选站点、因此其LUN路径将在大约15秒后变得不可用。

存储系统故障

存储系统故障的结果与丢失复制链路的结果几乎相同。正常运行的站点应出现大约15秒的IO暂停。15秒过后、IO将照常在该站点上恢复。

调解器丢失

调解器服务不直接控制存储操作。它可用作集群之间的备用控制路径。它主要用于自动执行故障转移、而不存在脑裂情况的风险。在正常操作下、每个集群都会将更改复制到其配对集群、因此、每个集群都可以验证配对集群是否联机并提供数据。如果复制链路失败、复制将停止。

安全自动故障转移需要调解器的原因是、否则存储集群将无法确定双向通信丢失是网络中断还是实际存储故障所致。

调解器为每个集群提供一个备用路径、以验证其配对集群的运行状况。具体情形如下：

- 如果集群可以直接与其配对集群联系、则复制服务将正常运行。无需执行任何操作。
- 如果首选站点无法直接或通过调解器与其配对站点联系、则会假定配对站点实际不可用或已隔离、并且其LUN路径已脱机。然后、首选站点将继续释放RPO = 0状态、并继续处理读写IO。
- 如果非首选站点无法直接与其配对站点联系、但可以通过调解器与其联系、则它会使其路径脱机、并等待复制连接返回。
- 如果非首选站点无法直接联系其配对站点或无法通过操作调解器联系其配对站点、则会假定配对站点实际不可用或已隔离、并且其LUN路径已脱机。然后、非首选站点将继续释放RPO = 0状态、并继续处理读写IO。它将承担复制源的角色、并成为新的首选站点。

如果调解器完全不可用：

- 复制服务因任何原因发生故障(包括非首选站点或存储系统发生故障)、都会导致首选站点释放RPO = 0状态并恢复读写IO处理。非首选站点将使其路径脱机。
- 首选站点发生故障将导致中断、因为非首选站点无法验证对等站点是否真正脱机、因此非首选站点无法安全地恢复服务。

正在还原服务

解决故障(例如、还原站点间连接或启动故障系统)后、SnapMirror活动同步端点将自动检测是否存在故障复制关系、并将其恢复为RPO = 0状态。重新建立同步复制后、故障路径将再次联机。

在许多情况下、集群模式应用程序会自动检测故障路径的返回情况、这些应用程序也会恢复联机。在其他情况下、可能需要进行主机级SAN扫描、或者可能需要手动将应用程序恢复联机。它取决于应用程序及其配置方式、通常、此类任务可以轻松实现自动化。ONTAP本身具有自我修复能力、不需要任何用户干预即可恢复RPO = 0存储操作。

手动故障转移

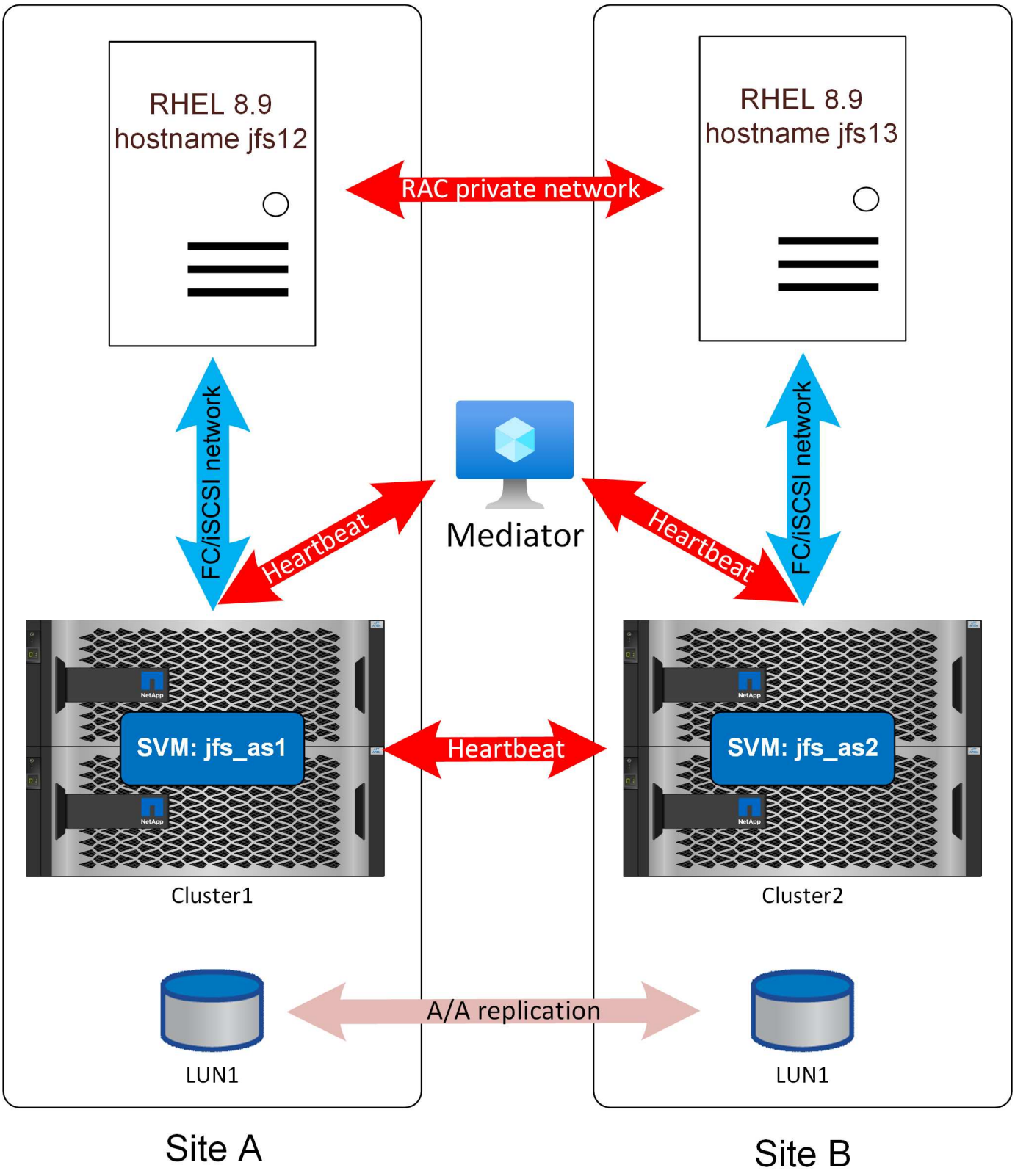
更改首选站点只需简单的操作即可。在集群之间切换复制行为的权限时、IO将暂停一两秒钟、但IO不会受到影响。

架构示例

本节中显示的详细故障示例基于下面所示的架构。

> 这只是SnapMirror主动同步上的Oracle数据库的众多选项之一。之所以选择此设计、是因为它展示了一些更复杂的情形。

在此设计中，假设站点A设置在"首选站点"。



RAC互连故障

丢失Oracle RAC复制链路会产生与丢失SnapMirror连接类似的结果、只是默认情况下超时时间较短。在默认设置下、Oracle RAC节点在丢失存储连接后将等待200秒后才会被逐出、但在丢失RAC网络检测信号后只会等待30秒。

CRS消息与以下所示类似。您可以看到30秒的超时时间。由于在位于站点A的jfs12上设置了css_critical,因此该站点将继续运行,而站点B上的jfs13将被逐出。

```
2024-09-12 10:56:44.047 [ONMD(3528)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 6.980 seconds
2024-09-12 10:56:48.048 [ONMD(3528)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.980 seconds
2024-09-12 10:56:51.031 [ONMD(3528)]CRS-1607: Node jfs13 is being evicted
in cluster incarnation 621599354; details at (:CSSNM00007:) in
/gridbase/diag/crs/jfs12/crs/trace/onmd.trc.
2024-09-12 10:56:52.390 [CRSD(6668)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:33194;', interface list of remote node 'jfs13' is
'192.168.30.2:33621;'.
2024-09-12 10:56:55.683 [ONMD(3528)]CRS-1601: CSSD Reconfiguration
complete. Active nodes are jfs12 .
2024-09-12 10:56:55.722 [CRSD(6668)]CRS-5504: Node down event reported for
node 'jfs13'.
2024-09-12 10:56:57.222 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'Generic'.
2024-09-12 10:56:57.224 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'ora.NTAP'.
```

SnapMirror通信失败

如果SnapMirror活动同步复制链路无法完成写入IO、因为集群无法将更改复制到相反站点。

站点 **A**

在站点A上、复制链路发生故障会导致写入IO处理暂停大约15秒、因为ONTAP会在确定复制链路确实无法运行之前尝试复制写入。15秒后、站点A上的ONTAP集群将恢复读写IO处理。SAN路径不会更改、LUN将保持联机状态。

站点 **B**

由于站点B不是SnapMirror主动同步首选站点、因此其LUN路径将在大约15秒后变得不可用。

复制链路在时间戳15：19：44处断开。当200秒超时(由Oracle RAC参数disktimeout控制)接近时、Oracle RAC发出的第一条警告会在100秒后到达。

```
2024-09-10 15:21:24.702 [ONMD(2792)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99340 milliseconds.
2024-09-10 15:22:14.706 [ONMD(2792)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49330 milliseconds.
2024-09-10 15:22:44.708 [ONMD(2792)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19330 milliseconds.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.716 [ONMD(2792)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.731 [OCSSD(2794)]CRS-1652: Starting clean up of CRSD
resources.
```

达到200秒表决磁盘超时时间后、此Oracle RAC节点将从集群中退出并重新启动。

网络互连总故障

如果站点之间的复制链路完全丢失、则SnapMirror活动同步和Oracle RAC连接都将中断。

Oracle RAC脑裂检测依赖于Oracle RAC存储检测信号。如果丢失站点间连接导致RAC网络检测信号和存储复制服务同时丢失、则RAC站点将无法通过RAC互连或RAC投票磁盘进行跨站点通信。如果节点数为偶数、则可能会在默认设置下逐出这两个站点。具体行为取决于事件顺序以及RAC网络和磁盘检测信号轮询的时间。

双站点中断的风险可以通过两种方式来解决。首先、"Tieb破碎 机"可以使用配置。

如果第三个站点不可用、则可以通过调整RAC集群上的mscount参数来解决此风险。在默认设置下、RAC网络检测信号超时为30秒。RAC通常会使用此方法来确定发生故障的RAC节点并将其从集群中删除。它还可以连接到投票磁盘检测信号。

例如、如果反铲切断了承载Oracle RAC和存储复制服务的站点间流量的管道、则会开始30秒的错误计数倒计时。如果RAC首选站点节点无法在30秒内与另一站点重新建立联系、并且也无法使用投票磁盘在同一30秒窗口内确认另一站点已关闭、则首选站点节点也将被清除。结果是数据库完全中断。

根据发生错误计数轮询的时间、30秒可能不足以使SnapMirror活动同步超时并允许首选站点上的存储在30秒窗口到期之前恢复服务。这30秒的窗口时间可以增加。

```
[root@jfs12 ~]# /grid/bin/crsctl set css misscount 100
CRS-4684: Successful set of parameter misscount to 100 for Cluster
Synchronization Services.
```

此值允许首选站点上的存储系统在错误计数超时过期之前恢复操作。这样、只会逐出已删除LUN路径的站点上的节点。以下示例：

```
2024-09-12 09:50:59.352 [ONMD(681360)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 49.570 seconds
2024-09-12 09:51:10.082 [CRSD(682669)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:46039;', interface list of remote node 'jfs13' is
'192.168.30.2:42037;'.
2024-09-12 09:51:24.356 [ONMD(681360)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 24.560 seconds
2024-09-12 09:51:39.359 [ONMD(681360)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 9.560 seconds
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8011: reboot advisory message
from host: jfs13, component: cssagent, with time stamp: L-2024-09-12-
09:51:47.451
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8013: reboot advisory message
text: oracssdagent is about to reboot this node due to unknown reason as
it did not receive local heartbeats for 10470 ms amount of time
2024-09-12 09:51:48.925 [ONMD(681360)]CRS-1632: Node jfs13 is being
removed from the cluster in cluster incarnation 621596607
```

Oracle支持部门强烈建议您不通过更改msscount或disktimeout参数来解决配置问题。但是、在许多情况下、包括SAN启动、虚拟化和存储复制配置、更改这些参数是有保证的、也是不可避免的。例如、如果您的SAN或IP网络出现稳定性问题、导致RAC逐出、则应修复底层问题、而不对msscount或disktimeout值收费。更改超时以解决配置错误会掩盖问题、而不会解决问题。根据底层基础架构的设计方面更改这些参数以正确配置RAC环境的做法有所不同、并且与Oracle支持声明一致。在SAN启动中、通常会将Msscount调整为最大200、以匹配磁盘超时。有关更多信息、请参见"此链接。"。

站点故障

存储系统或站点故障的结果与丢失复制链路的结果几乎相同。正常运行的站点应在写入时发生大约15秒的IO暂停。15秒过后、IO将照常在该站点上恢复。

如果仅存储系统受到影响、则故障站点上的Oracle RAC节点将丢失存储服务、并在逐出和后续重新启动之前输入相同的200秒磁盘超时时间。

```
2024-09-11 13:44:38.613 [ONMD(3629)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99750 milliseconds.
2024-09-11 13:44:51.202 [ORAAGENT(5437)]CRS-5011: Check of resource "NTAP"
failed: details at "(:CLSN00007:)" in
"/gridbase/diag/crs/jfs13/crs/trace/crsd_oraagent_oracle.trc"
2024-09-11 13:44:51.798 [ORAAGENT(75914)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 75914
2024-09-11 13:45:28.626 [ONMD(3629)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49730 milliseconds.
2024-09-11 13:45:33.339 [ORAAGENT(76328)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 76328
2024-09-11 13:45:58.629 [ONMD(3629)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19730 milliseconds.
2024-09-11 13:46:18.630 [ONMD(3629)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-11 13:46:18.631 [ONMD(3629)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.638 [ONMD(3629)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.651 [OCSSD(3631)]CRS-1652: Starting clean up of CRSD
resources.
```

丢失存储服务的RAC节点上的SAN路径状态如下所示：

```
oradata7 (3600a0980383041334a3f55676c697347) dm-20 NETAPP,LUN C-Mode
size=128G features='3 queue_if_no_path pg_init_retries 50' hwhandler='1
alua' wp=rw
|-+- policy='service-time 0' prio=0 status=enabled
| `- 34:0:0:18 sdam 66:96  failed faulty running
`-+- policy='service-time 0' prio=0 status=enabled
  `- 33:0:0:18 sdaj 66:48  failed faulty running
```

Linux主机检测到路径丢失的速度比200秒快得多、但从数据库角度来看、在默认Oracle RAC设置下、与故障站点上主机的客户端连接仍会冻结200秒。只有在逐出完成后、才会恢复完整数据库操作。

同时、另一站点上的Oracle RAC节点将记录另一个RAC节点的丢失情况。否则，它将继续照常运作。

```
2024-09-11 13:46:34.152 [ONMD(3547)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 14.020 seconds
2024-09-11 13:46:41.154 [ONMD(3547)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 7.010 seconds
2024-09-11 13:46:46.155 [ONMD(3547)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.010 seconds
2024-09-11 13:46:46.470 [OHASD(1705)]CRS-8011: reboot advisory message
from host: jfs13, component: cssmonit, with time stamp: L-2024-09-11-
13:46:46.404
2024-09-11 13:46:46.471 [OHASD(1705)]CRS-8013: reboot advisory message
text: At this point node has lost voting file majority access and
oracssdmonitor is rebooting the node due to unknown reason as it did not
receive local hearbeats for 28180 ms amount of time
2024-09-11 13:46:48.173 [ONMD(3547)]CRS-1632: Node jfs13 is being removed
from the cluster in cluster incarnation 621516934
```

调解器故障

调解器服务不直接控制存储操作。它可用作集群之间的备用控制路径。它主要用于自动执行故障转移、而不存在脑裂情况的风险。

在正常操作下、每个集群都会将更改复制到其配对集群、因此、每个集群都可以验证配对集群是否联机并提供数据。如果复制链路失败、复制将停止。

安全自动化操作需要调解器的原因是、否则存储集群将无法确定双向通信丢失是网络中断还是实际存储故障所致。

调解器为每个集群提供一个备用路径、以验证其配对集群的运行状况。具体情形如下：

- 如果集群可以直接与其配对集群联系、则复制服务将正常运行。无需执行任何操作。
- 如果首选站点无法直接或通过调解器与其配对站点联系、则会假定配对站点实际不可用或已隔离、并且其LUN路径已脱机。然后、首选站点将继续释放RPO = 0状态、并继续处理读写IO。
- 如果非首选站点无法直接与其配对站点联系、但可以通过调解器与其联系、则它会使其路径脱机、并等待复制连接返回。
- 如果非首选站点无法直接联系其配对站点或无法通过操作调解器联系其配对站点、则会假定配对站点实际不可用或已隔离、并且其LUN路径已脱机。然后、非首选站点将继续释放RPO = 0状态、并继续处理读写IO。它将承担复制源的角色、并成为新的首选站点。

如果调解器完全不可用：

- 复制服务因任何原因出现故障都会导致首选站点释放RPO = 0状态、并恢复读写IO处理。非首选站点将使其路径脱机。
- 首选站点发生故障将导致中断、因为非首选站点无法验证对等站点是否真正脱机、因此非首选站点无法安全

地恢复服务。

服务还原

SnapMirror可以自行恢复。SnapMirror主动同步将自动检测复制关系是否存在故障、并将其恢复到RPO = 0状态。重新建立同步复制后、路径将再次联机。

在许多情况下、集群模式应用程序会自动检测故障路径的返回情况、这些应用程序也会恢复联机。在其他情况下、可能需要进行主机级SAN扫描、或者可能需要手动将应用程序恢复联机。

这取决于应用程序及其配置方式、通常、此类任务可以轻松实现自动化。SnapMirror主动同步本身可以自行修复、在电源和连接恢复后、不需要任何用户干预即可恢复RPO = 0存储操作。

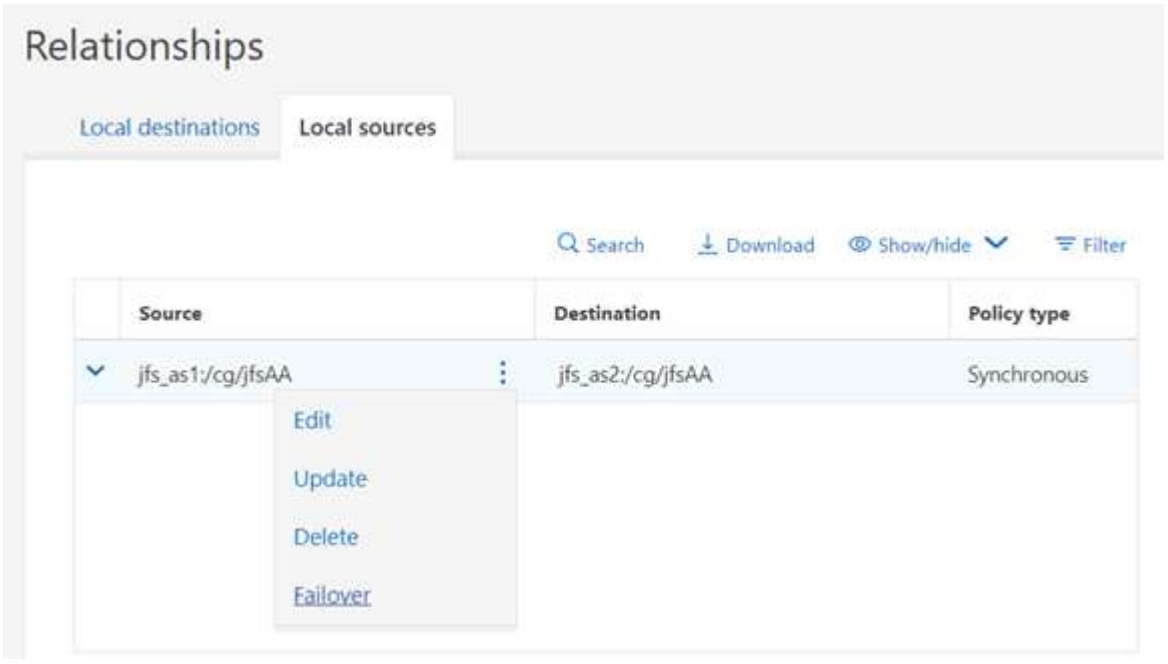
手动故障转移

术语"故障转移"并不是指使用SnapMirror活动同步进行复制的方向、因为它是一种双向复制技术。相反、"故障转移"是指发生故障时、哪个存储系统将成为首选站点。

例如、您可能希望在关闭站点进行维护之前或执行灾难恢复测试之前执行故障转移以更改首选站点。

更改首选站点只需简单的操作即可。在集群之间切换复制行为的权限时、IO将暂停一两秒钟、但IO不会受到影响。

GUI示例：



通过命令行界面将其更改回的示例：

```
Cluster2::> snapmirror failover start -destination-path jfs_as2:/cg/jfsAA
[Job 9575] Job is queued: SnapMirror failover for destination
"jfs_as2:/cg/jfsAA                      ".

Cluster2::> snapmirror failover show

Source    Destination                                           Error
Path      Path        Type     Status    start-time end-time   Reason
--------- ----------  -------- --------- ---------- ---------- ----------
jfs_as1:/cg/jfsAA
          jfs_as2:/cg/jfsAA
                      planned  completed 9/11/2024  9/11/2024
                                         09:29:22   09:29:32

The new destination path can be verified as follows:


Cluster1::> snapmirror show -destination-path jfs_as1:/cg/jfsAA

                           Source Path: jfs_as2:/cg/jfsAA
                        Destination Path: jfs_as1:/cg/jfsAA
                       Relationship Type: XDP
                 Relationship Group Type: consistencygroup
                  SnapMirror Policy Type: automated-failover-duplex
                       SnapMirror Policy: AutomatedFailOverDuplex
                             Tries Limit: -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```