MetroCluster

Enterprise applications

NetApp
May 09, 2024

目录

MetroCluster

MetroCluster物理架构和Oracle数据库

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

[注意]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。

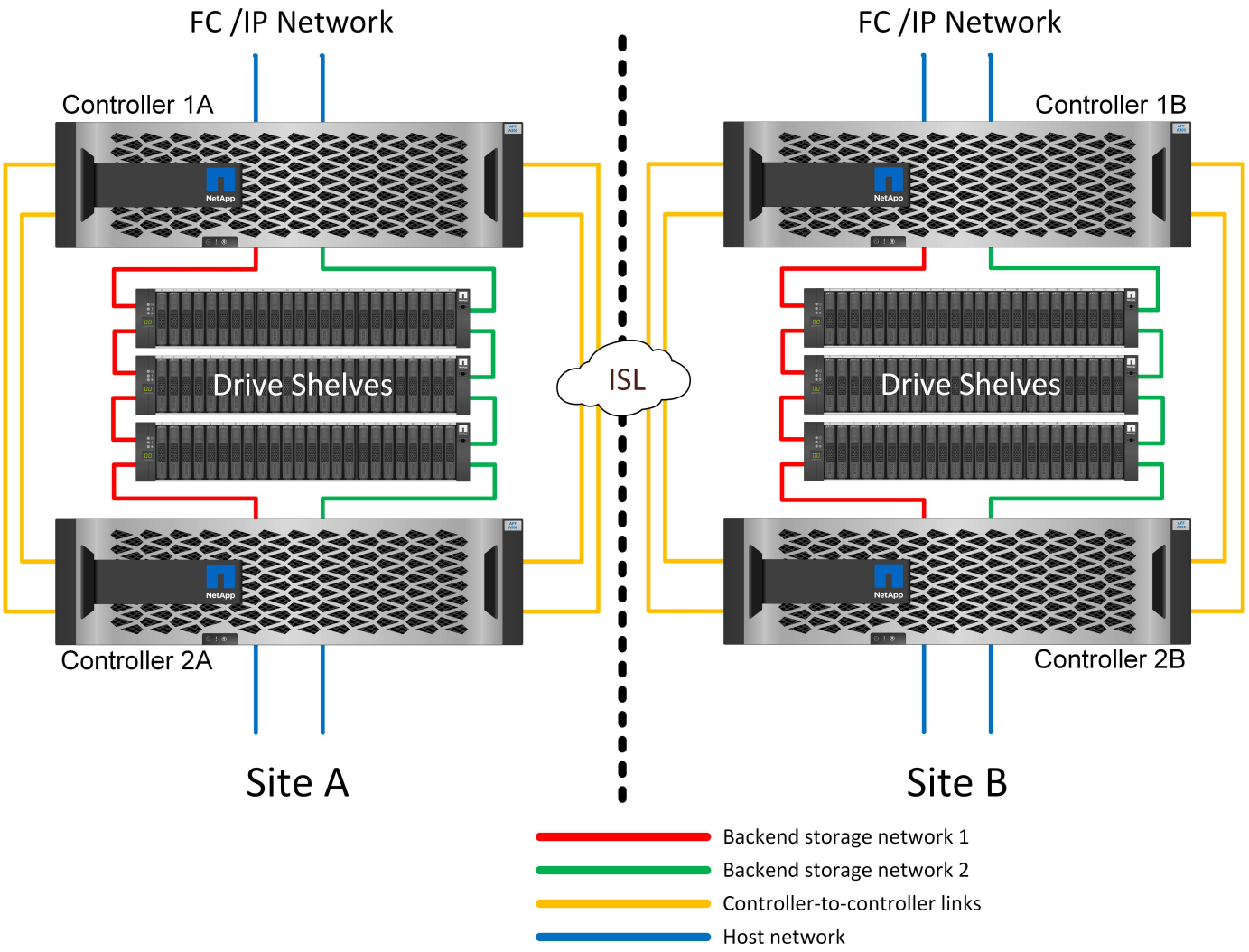
MetroCluster IP

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。

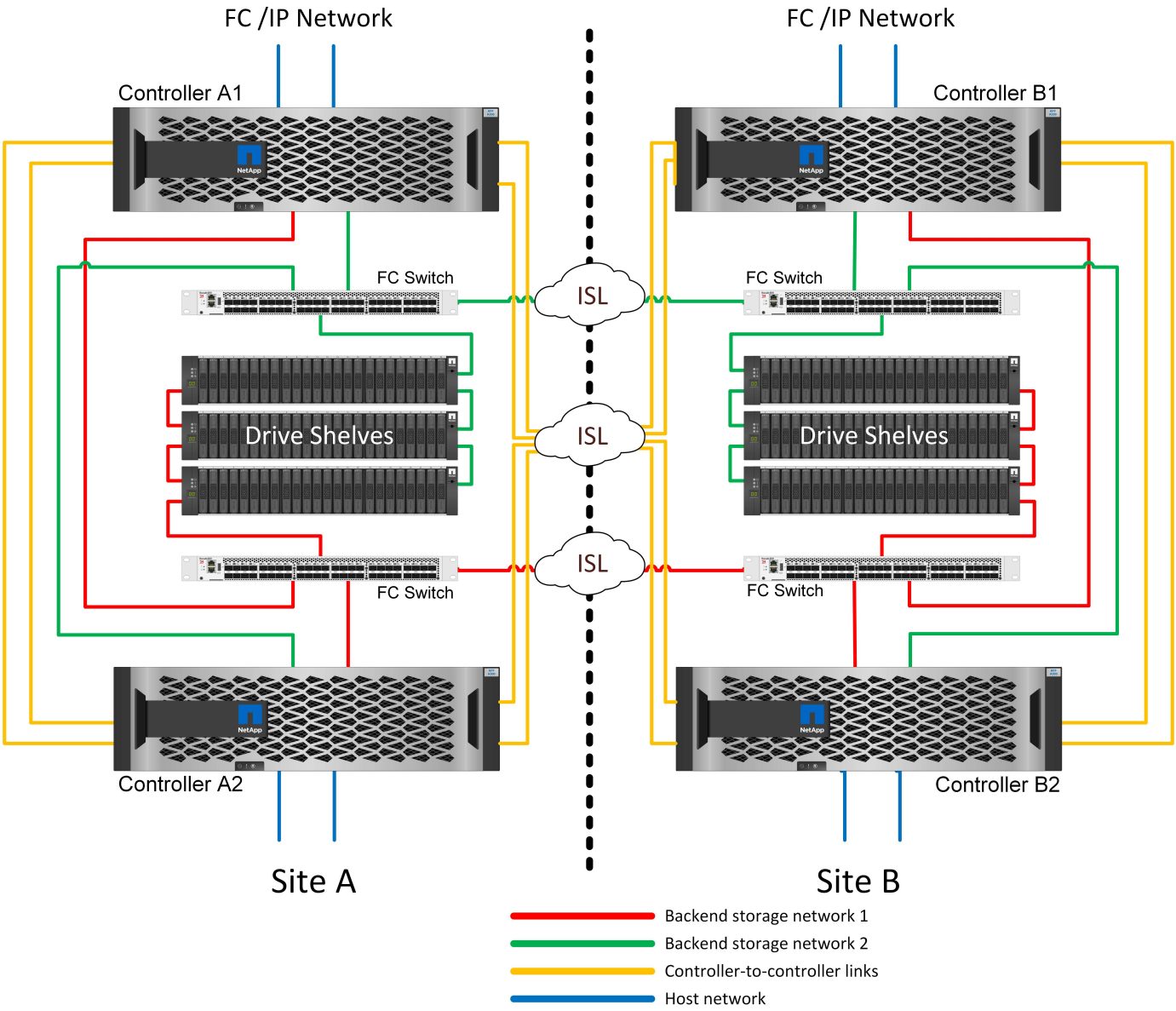
此外、该架构也更加简单、因为只有跨站点连接用于控制器。在FC SAN连接的MetroCluster中、控制器会直接写入另一站点上的驱动器、因此需要更多的SAN连接、交换机和网桥。相反、IP配置中的控制器会通过控制器写入相对的驱动器。

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。

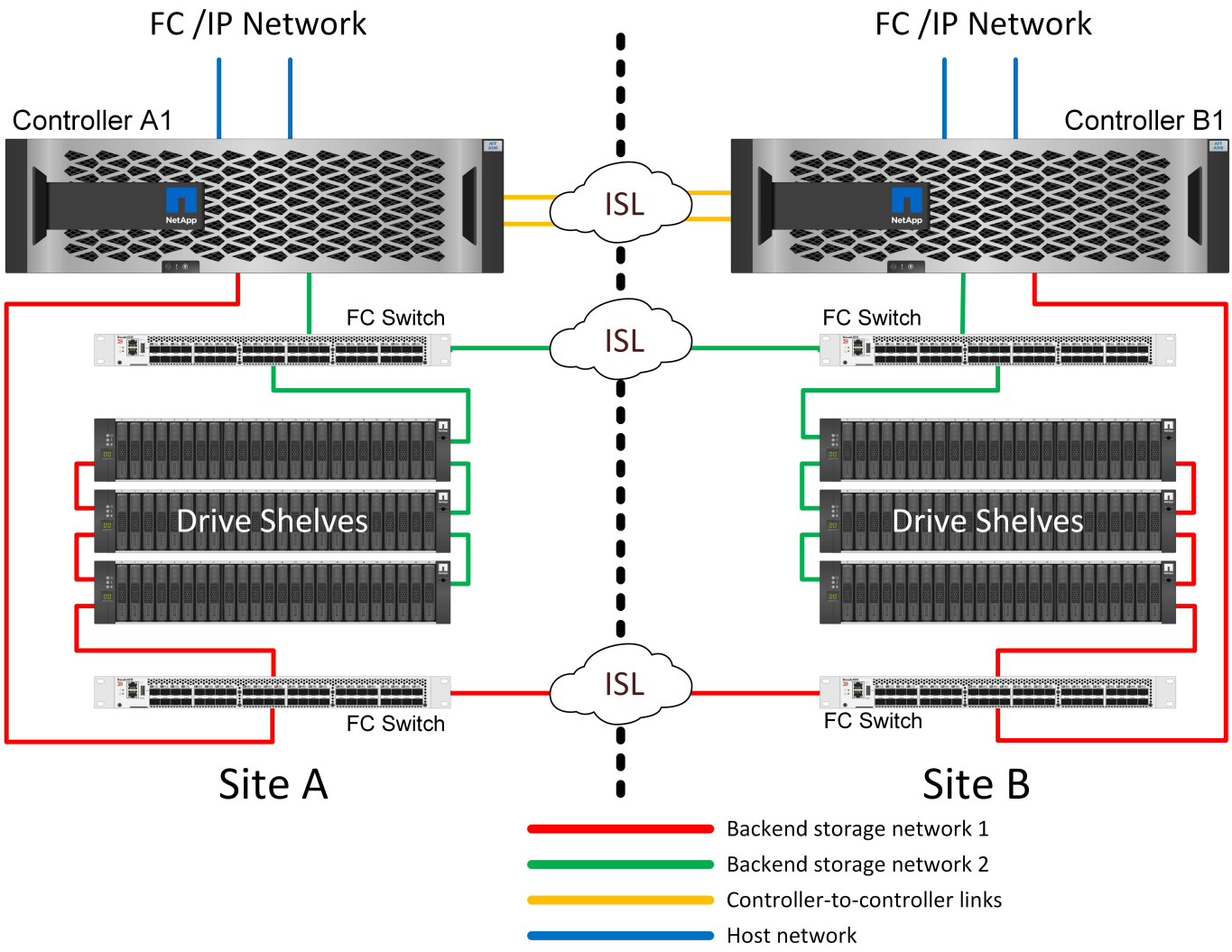


某些多站点基础架构不是为主动-主动操作而设计的、而是更多地用作主站点和灾难恢复站点。在这种情况下、通常最好使用HA对MetroCluster选项、原因如下：

- 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。

MetroCluster逻辑架构和Oracle数据库

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

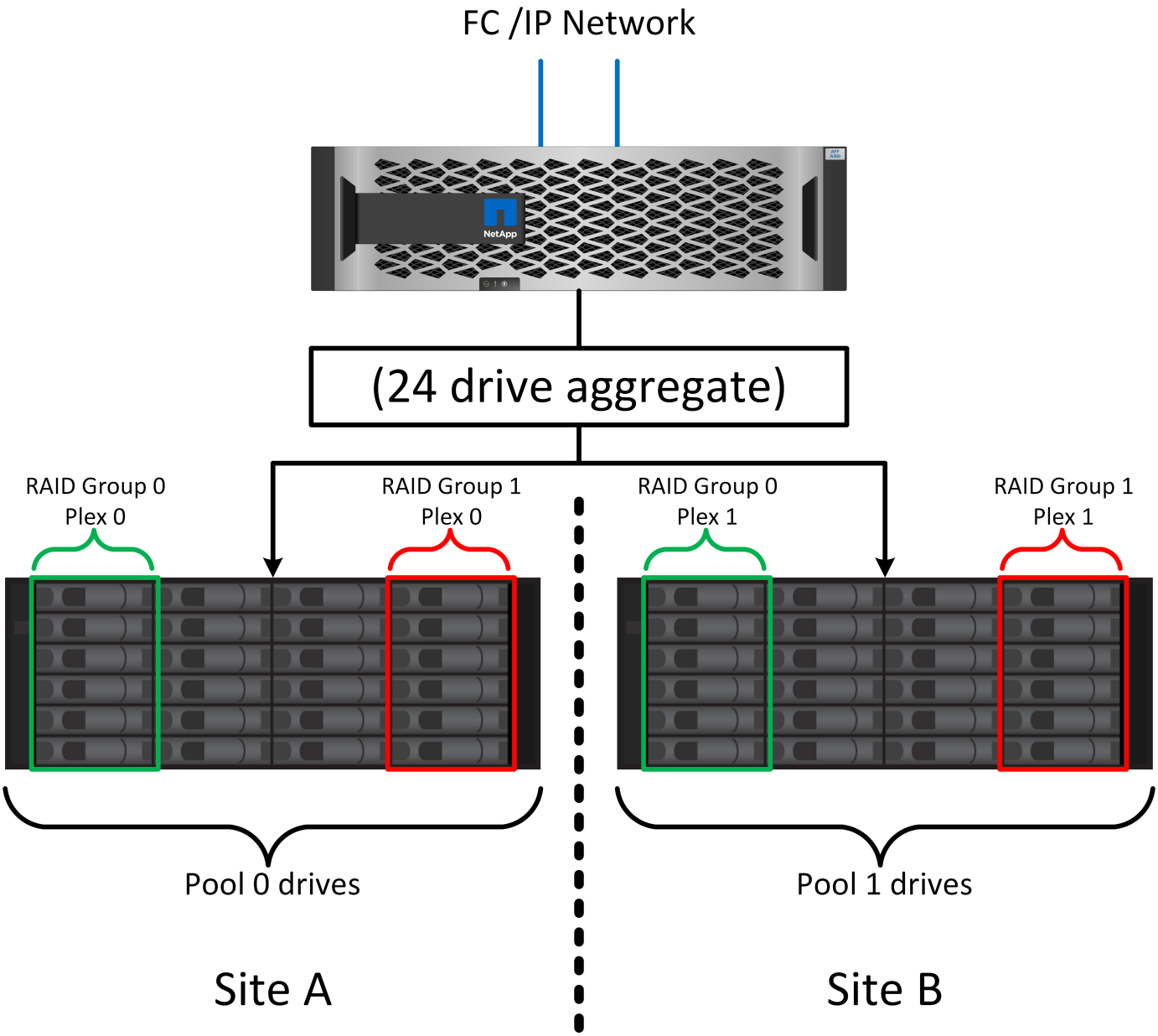
MetroCluster通过以下方式扩展NVRAM数据保护：

- 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

。 "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 机的完整问题描述、请参见 "NetApp 支持站点"

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。

*注意：*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

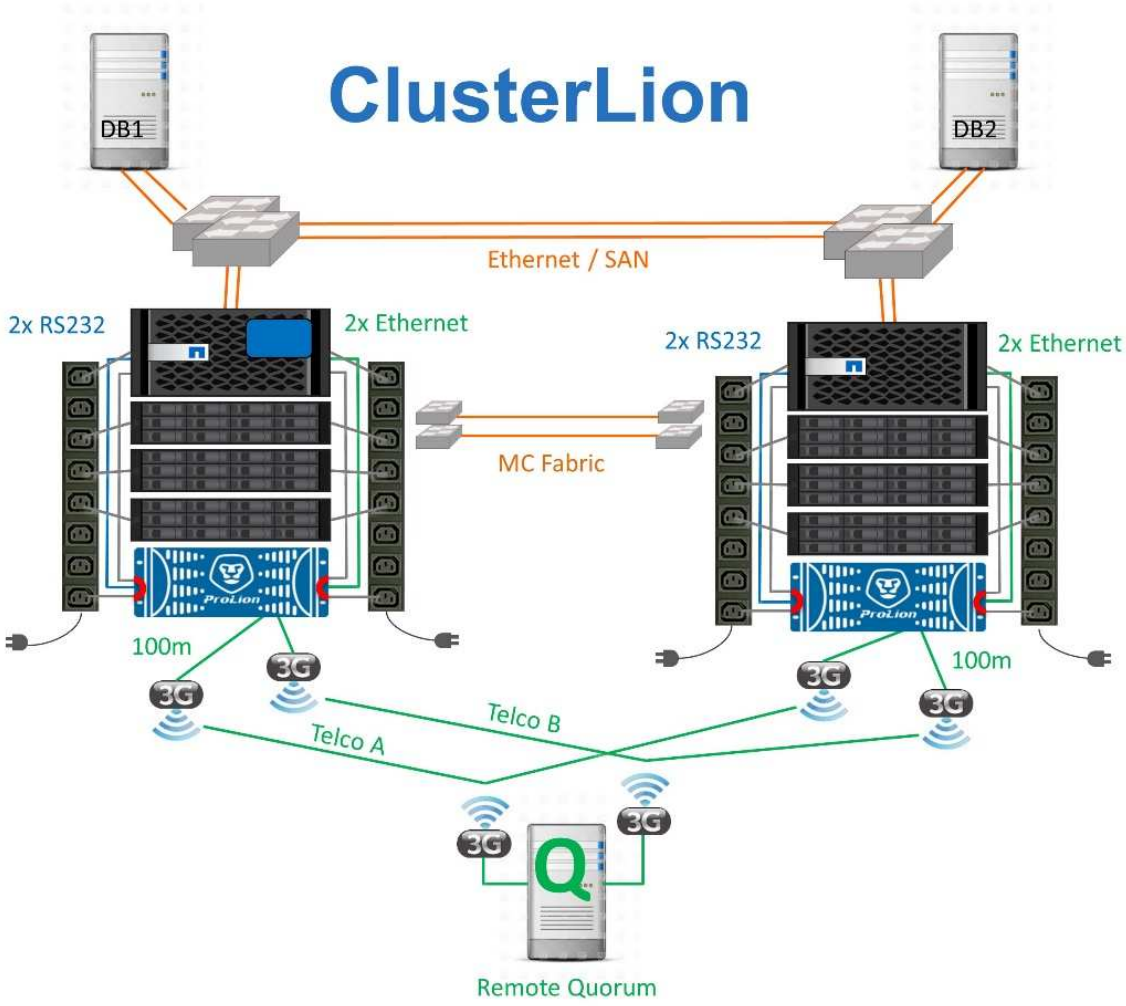
ONTAP调解器可与MetroCluster IP和某些其他ONTAP解决方案结合使用。它的功能与上述MetroCluster Tieb破碎 机软件非常相似、但也包括一项关键功能、即执行自动无人值守切换。

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

采用SyncMirror的Oracle数据库

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、还提供了应用程序级选项、例如Oracle DataGuard或主机端磁盘镜像的扩展超时。有关追加信息和选项、请咨询您的NetApp或合作伙伴客户团队。

使用MetroCluster进行Oracle数据库故障转移

```
 Metrocluster is an ONTAP feature that can protect your Oracle databases
with RPO=0 synchronous mirroring across sites, and it scales up to support
hundreds of databases on a single MetroCluster system. It's also simple to
use. The use of MetroCluster does not necessarily add to or change any
best practices for operating a enterprise applications and databases.
```

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。

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。Oracle RAC、Veritas Cluster Server (VCS)或VMware HA等集群软件产品通常用于在站点间创建集群、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应用程序联机。一种选择是、创建为构成应用程序的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
4. 正在启动应用程序

此方法的主要要求是在远程站点上运行操作系统。它必须预配置应用程序二进制文件、这也意味着必须在主站点和备用站点上执行修补等任务。或者、可以将应用程序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数据库、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。

基于MetroCluster的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。

基于MetroCluster的扩展Oracle 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。

如果需要主动-主动扩展RAC、则应考虑使用ASM镜像代替MetroCluster。ASM镜像允许首选使用特定的数据副本。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 `misscount` 可能需要更改参数。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。