SnapMirror活动同步

Enterprise applications

NetApp
February 11, 2026

This PDF was generated from https://docs.netapp.com/zh-cn/ontap-apps-dbs/oracle/oracle-dr-smas-overview.html on February 11, 2026. Always check docs.netapp.com for the latest.

目录

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。

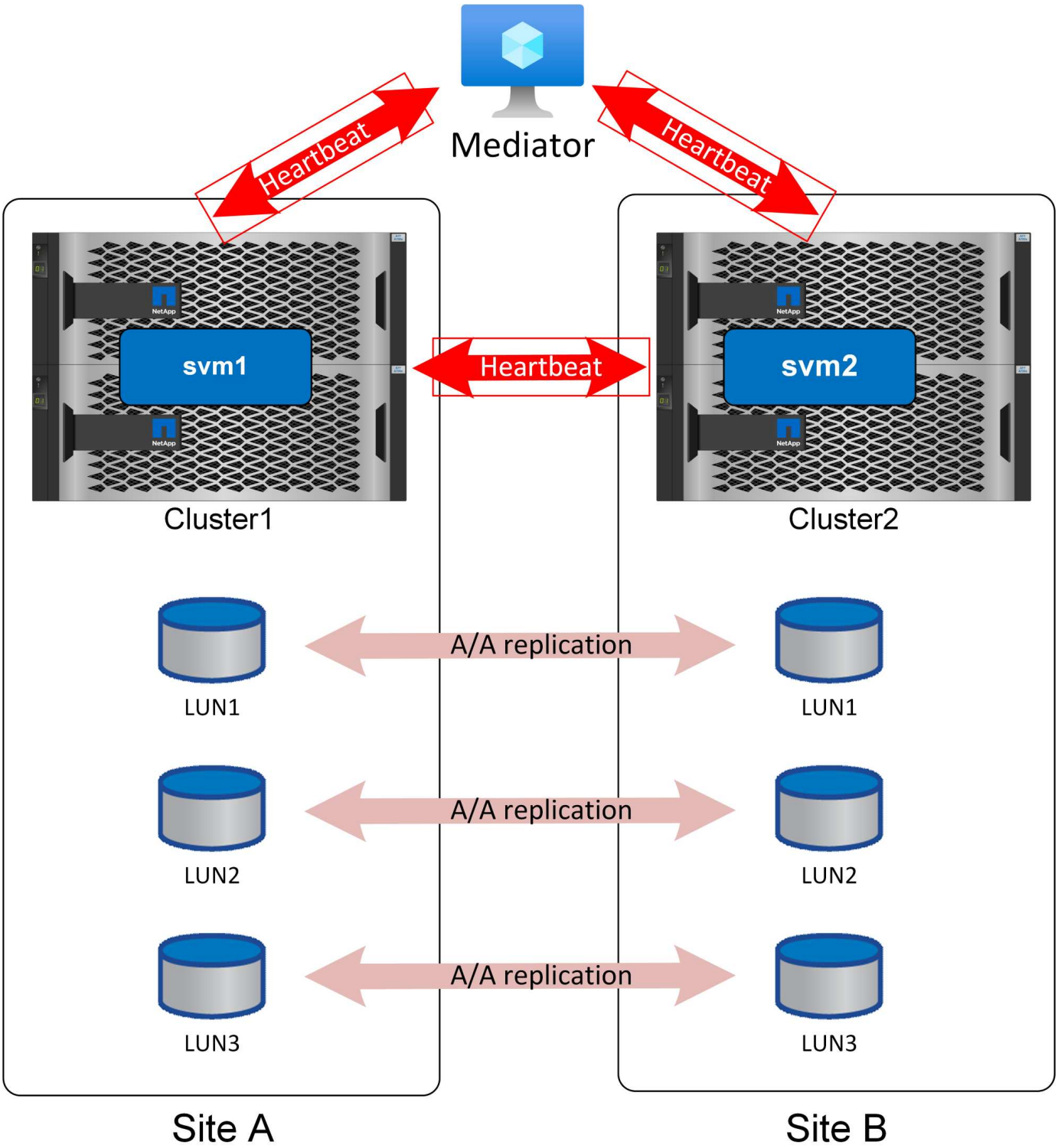
ONTAP调解器

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。

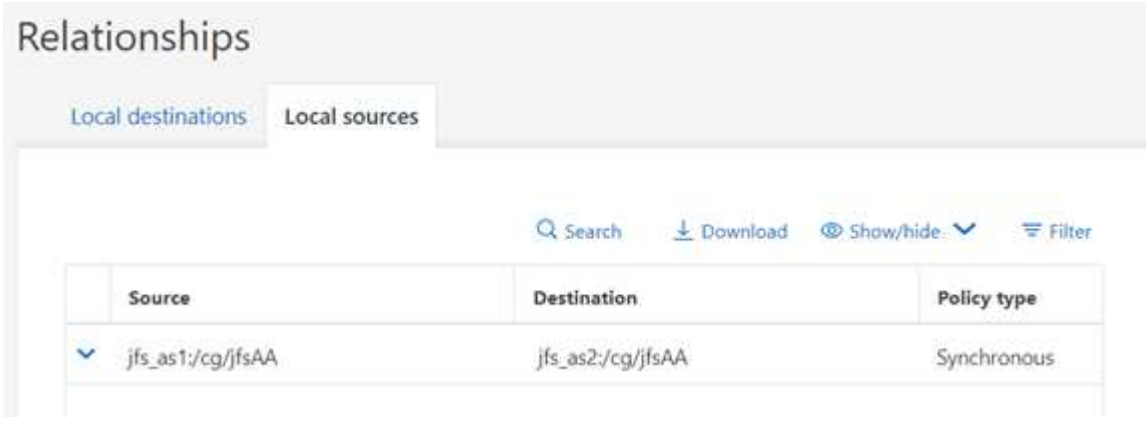
SnapMirror主动同步首选站点

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。

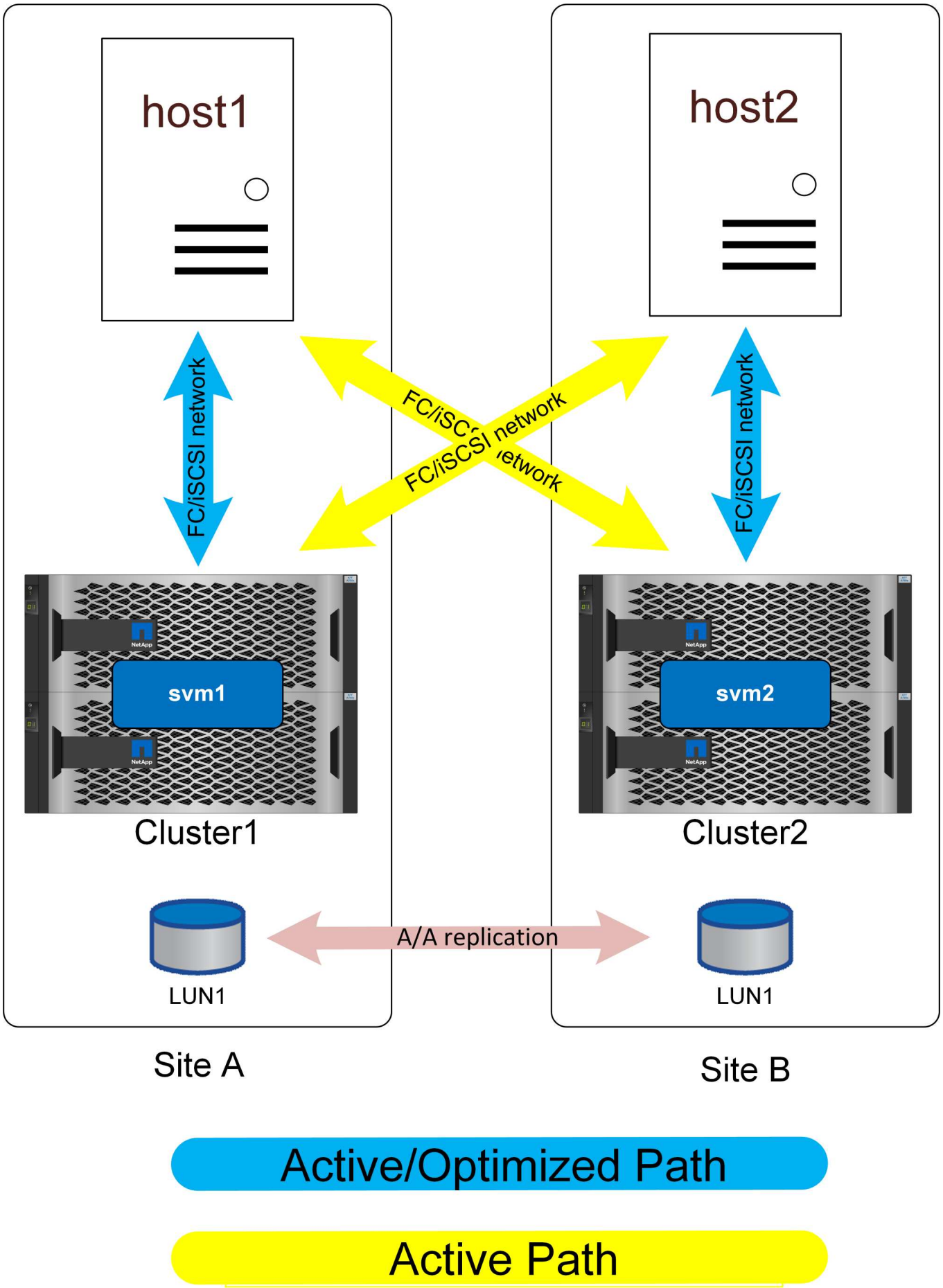
第二个可选步骤是配置主机邻近性。这不控制访问、而是控制_priority_。

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

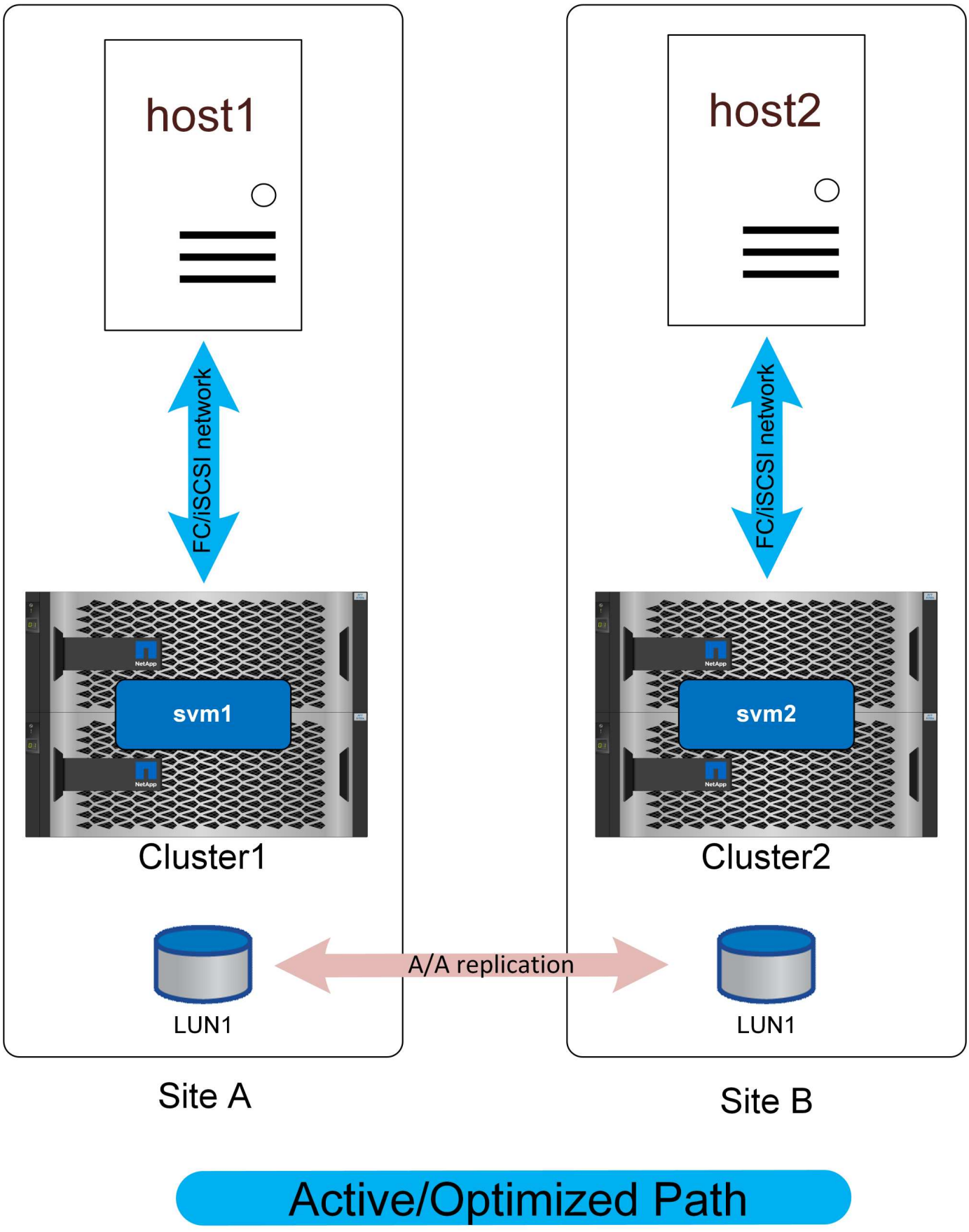
在正常操作下、所有IO均为本地IO。读取和写入操作由本地存储阵列提供。当然、在确认写入IO之前、本地控制器也需要将其复制到远程系统、但所有读取IO都将在本地进行处理、并且不会通过遍历站点间的SAN链路而产生额外延迟。

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。



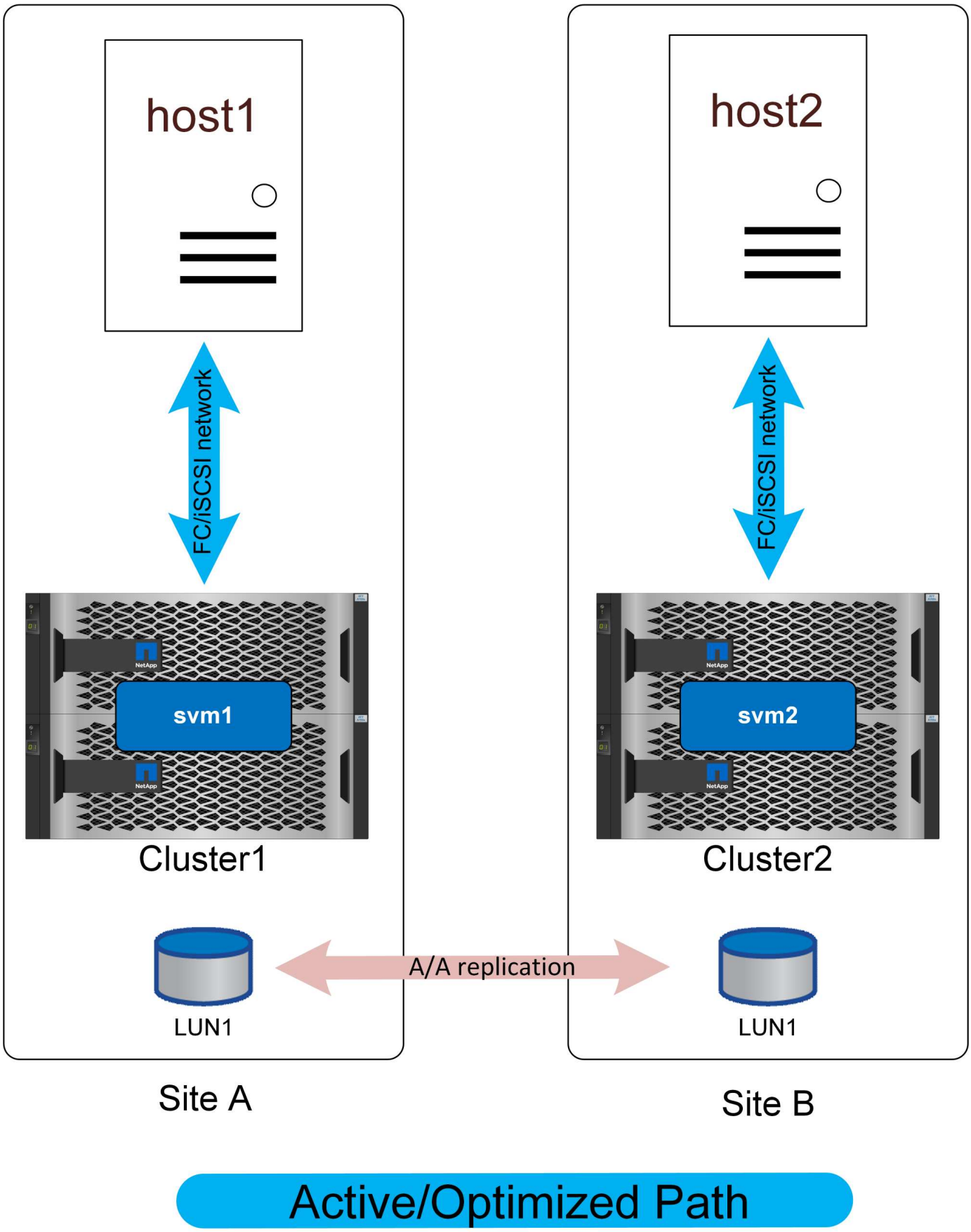
使用非一致访问的ASA配置的工作原理与使用AFF时大致相同。使用统一访问时、IO将跨越WAN。这可能是可取的、也可能不可取。

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。

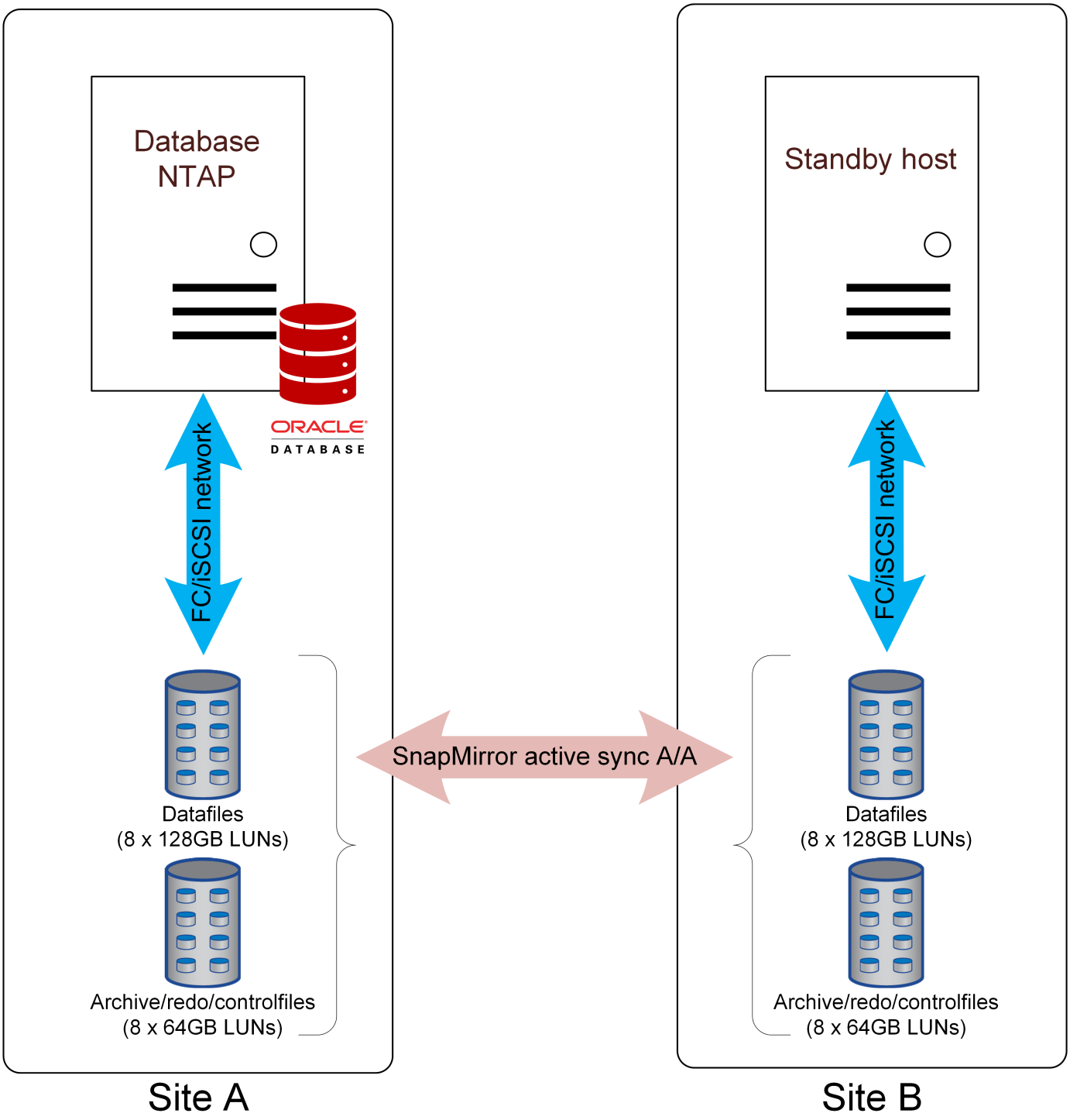
最佳架构取决于业务需求。例如、如果目标是使RPO=0防止数据丢失、但RTO较宽松、则使用Oracle单实例数据库并使用SM-AS复制LUN可能足以满足Oracle许可标准的要求、而且成本更低。远程站点故障不会中断操作、丢失主站点将导致运行正常的站点上的LUN处于联机状态并可供使用。

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。

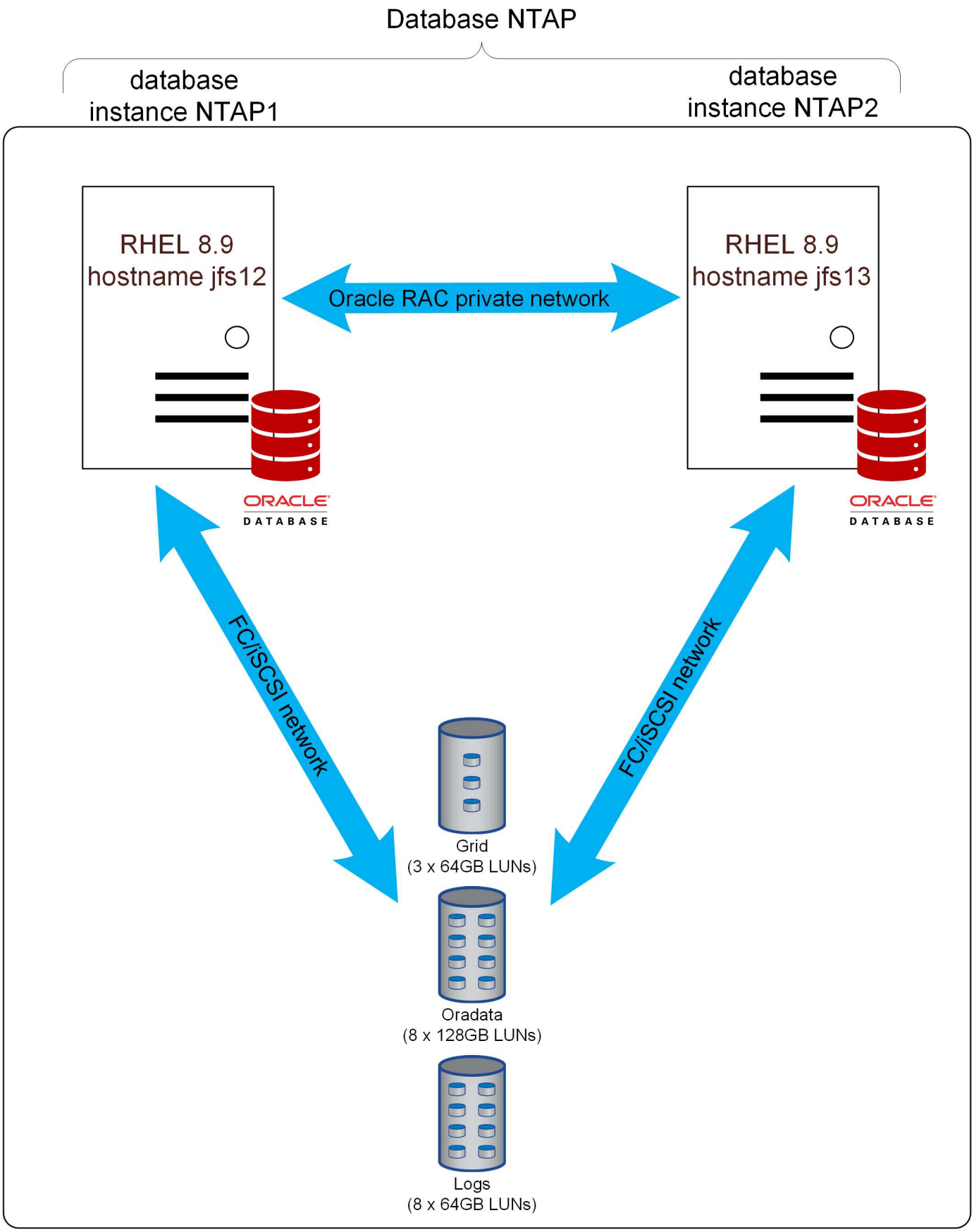
Oracle Extended RAC

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

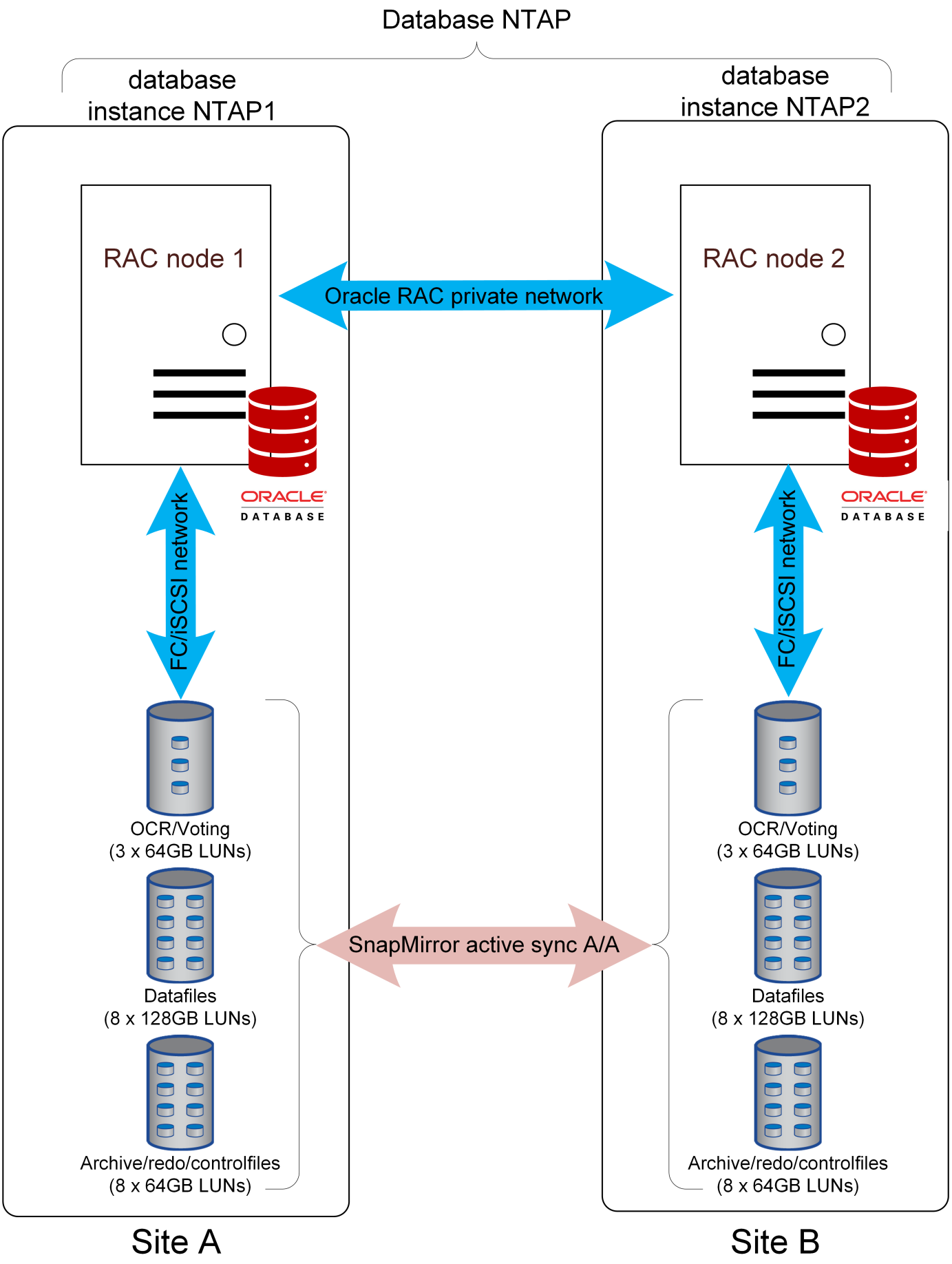
在SnapMirror主动同步模式下使用Oracle RAC最重要的注意事项是使用统一访问还是非统一访问。

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

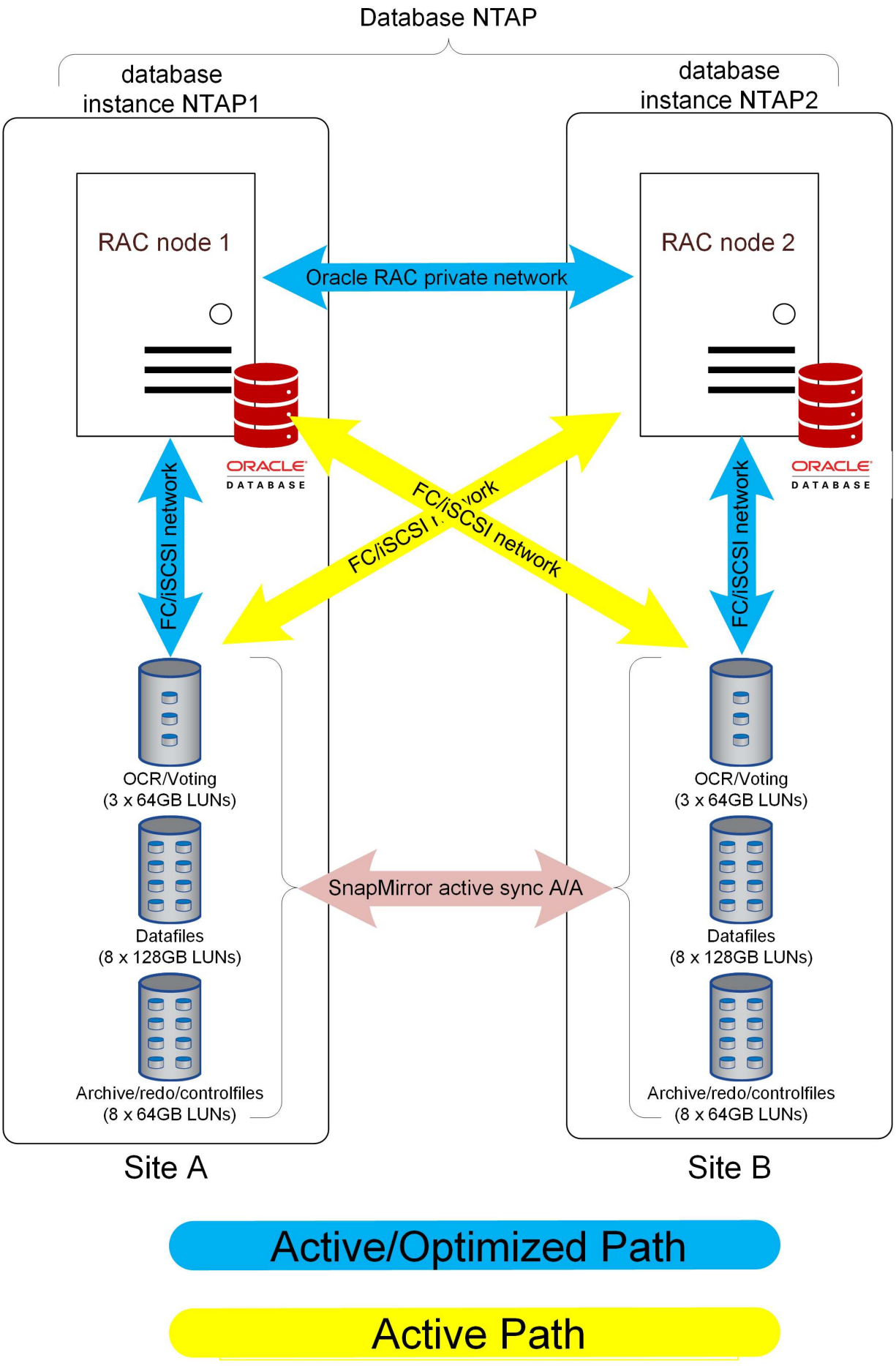
统一访问要求跨站点扩展SAN。主要优势是、丢失存储系统不会导致数据库实例丢失。相反、它会导致当前正在使用的路径发生多路径更改。

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**

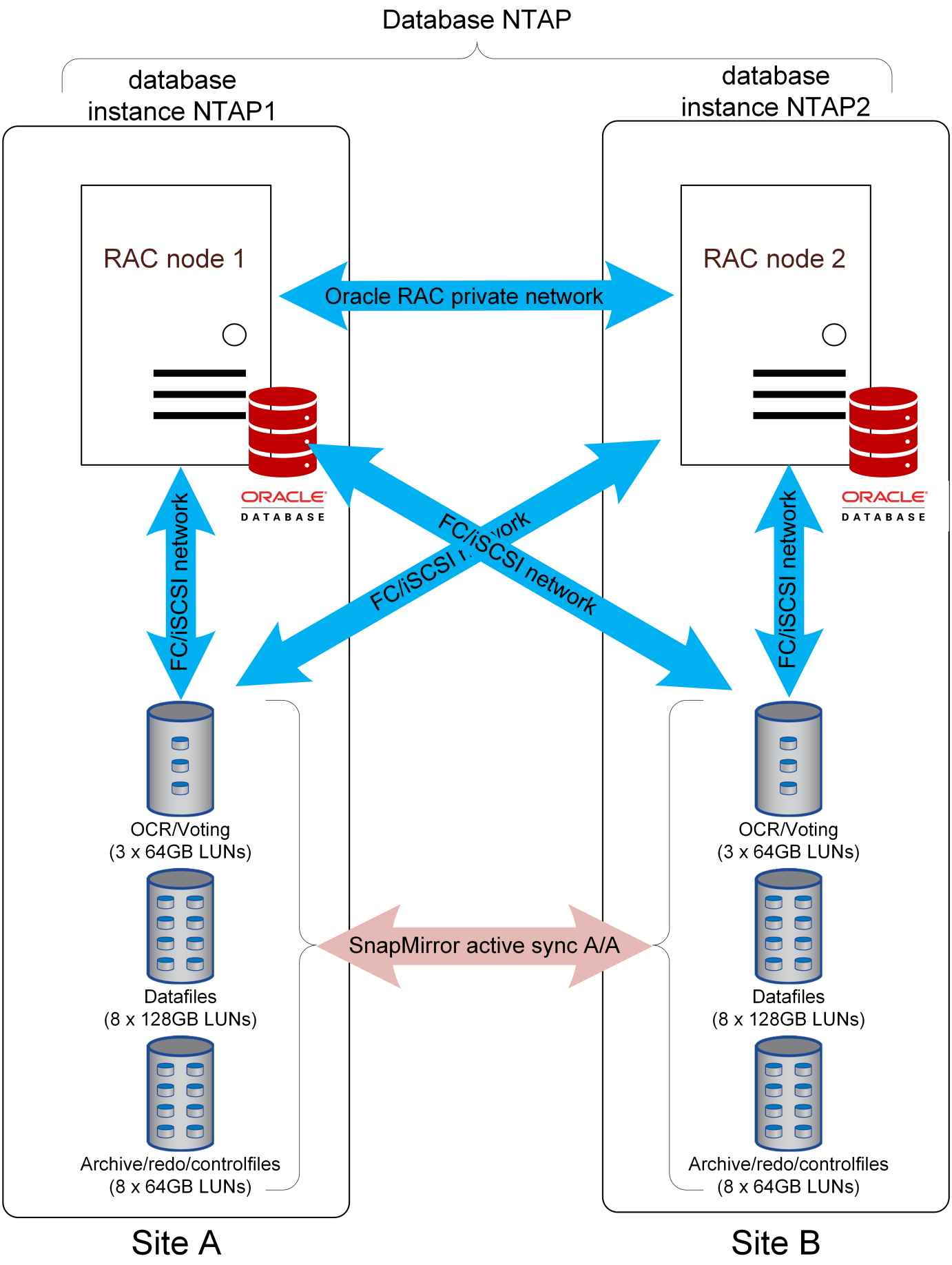
如果站点间延迟较长、则可以为AFF系统配置主机邻近设置。这样、每个存储系统就可以了解哪些主机是本地主机、哪些主机是远程主机、并相应地分配路径优先级。


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

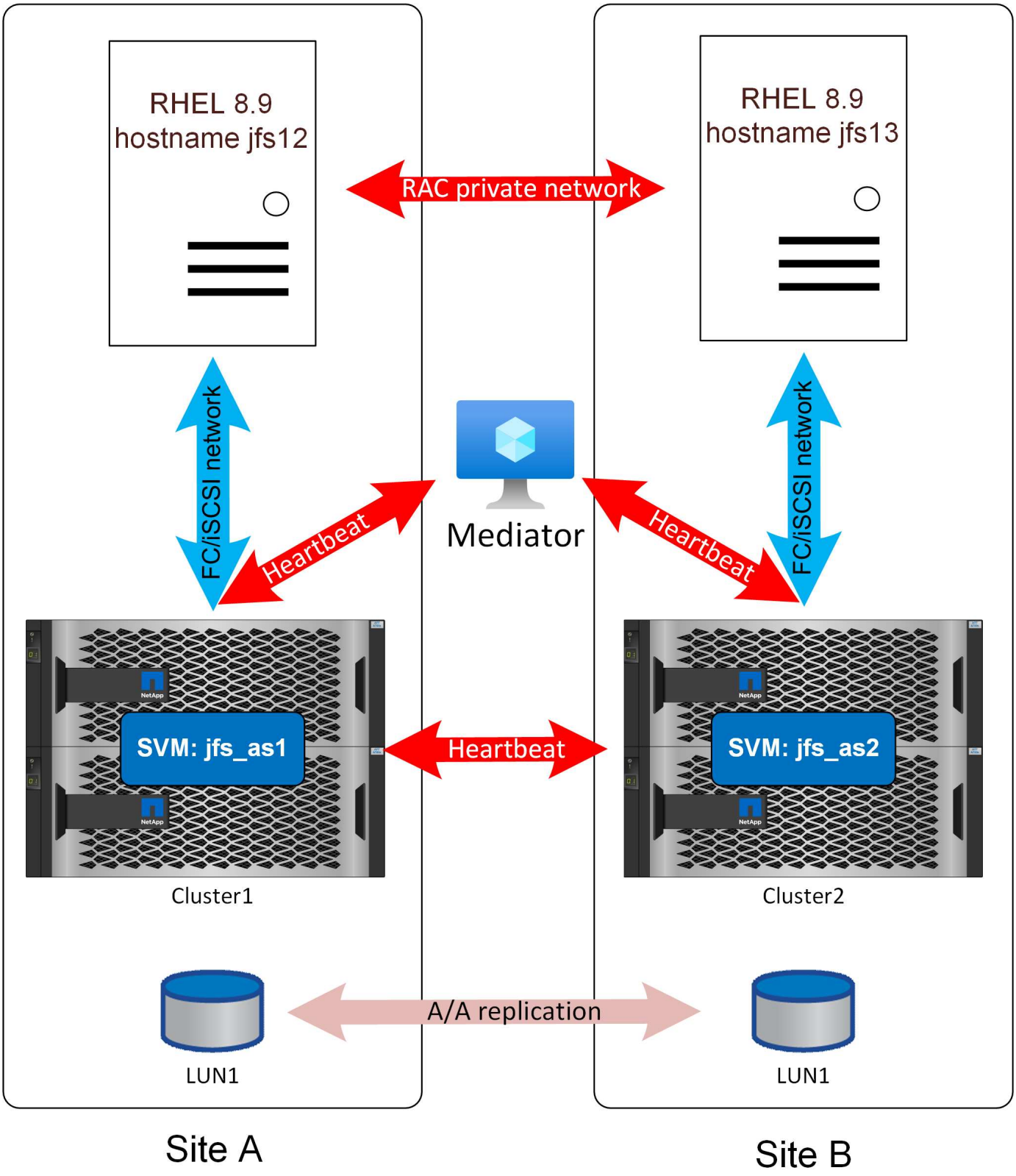
更改首选站点只需简单的操作即可。在集群之间切换复制行为的权限时、IO将暂停一两秒钟、但IO不会受到影响。

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。

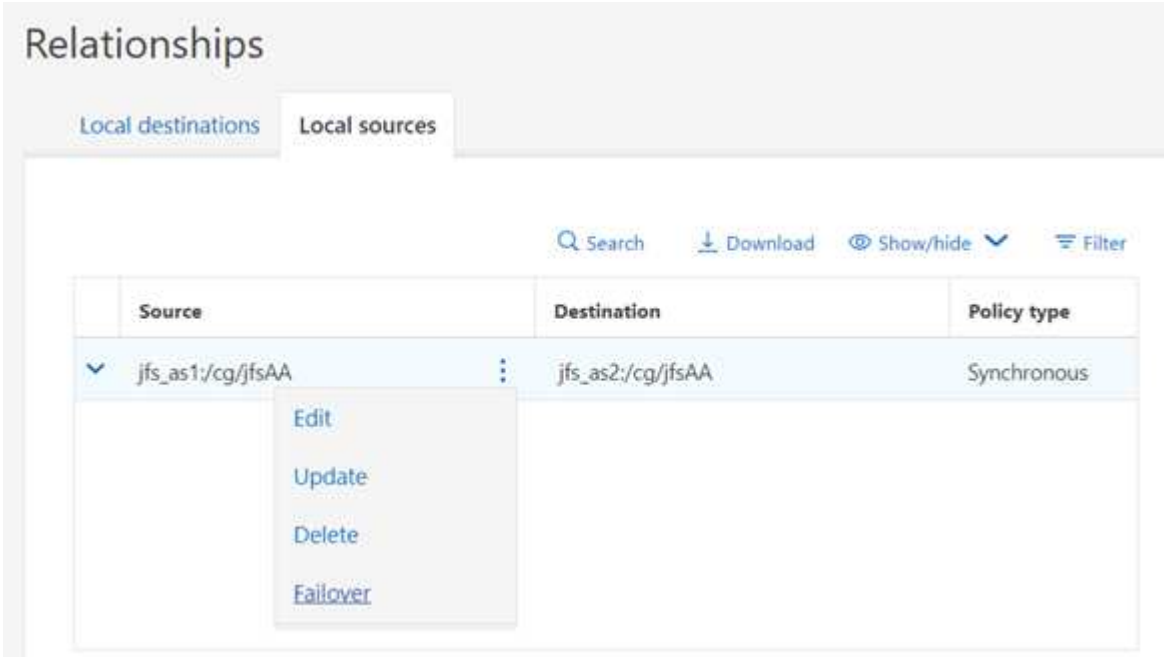
手动故障转移

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
```

版权信息

版权所有 © 2026 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。 美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

商标信息

NetApp、NetApp 标识和 http://www.netapp.com/TM 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。