SnapMirror 主動同步

Enterprise applications

NetApp
February 11, 2026

目錄

SnapMirror 主動同步

總覽

SnapMirror 主動式同步可讓您建置超高可用度的 Oracle 資料庫環境、其中 LUN 可從兩個不同的儲存叢集取得。

使用 SnapMirror 主動式同步時、資料不會有「主要」和「次要」複本。每個叢集都可以從其本機資料複本提供讀取 IO 、而且每個叢集都會將寫入複寫到其合作夥伴。結果是對稱 IO 行為。

除了其他選項之外、這可讓您將 Oracle RAC 當作延伸叢集來執行、並在兩個站台上執行作業執行個體。或者、您也可以建置 RPO = 0 主動被動式資料庫叢集、在站台中斷期間、可在站台之間移動單一執行個體資料庫、而此程序可透過 Pacemaker 或 VMware HA 等產品來自動化。所有這些選項的基礎都是由 SnapMirror 主動式同步管理的同步複寫。

同步複寫

在正常作業中、 SnapMirror 主動式同步可隨時提供 RPO = 0 同步複本、但有一個例外。如果資料無法複寫、ONTAP 將釋出複寫資料的需求、並在另一個站台上的 LUN 離線時、繼續在一個站台上提供 IO 服務。

儲存硬體

與其他儲存災難恢復解決方案不同、 SnapMirror 主動式同步提供非對稱式平台靈活度。每個站台的硬體不一定相同。此功能可讓您調整支援 SnapMirror 主動同步所用硬體的大小。如果遠端儲存系統需要支援完整的正式作業工作負載、則它可以與主要站台相同、但如果災難導致 I/O 減少、遠端站台上較小的系統可能會更具成本效益。

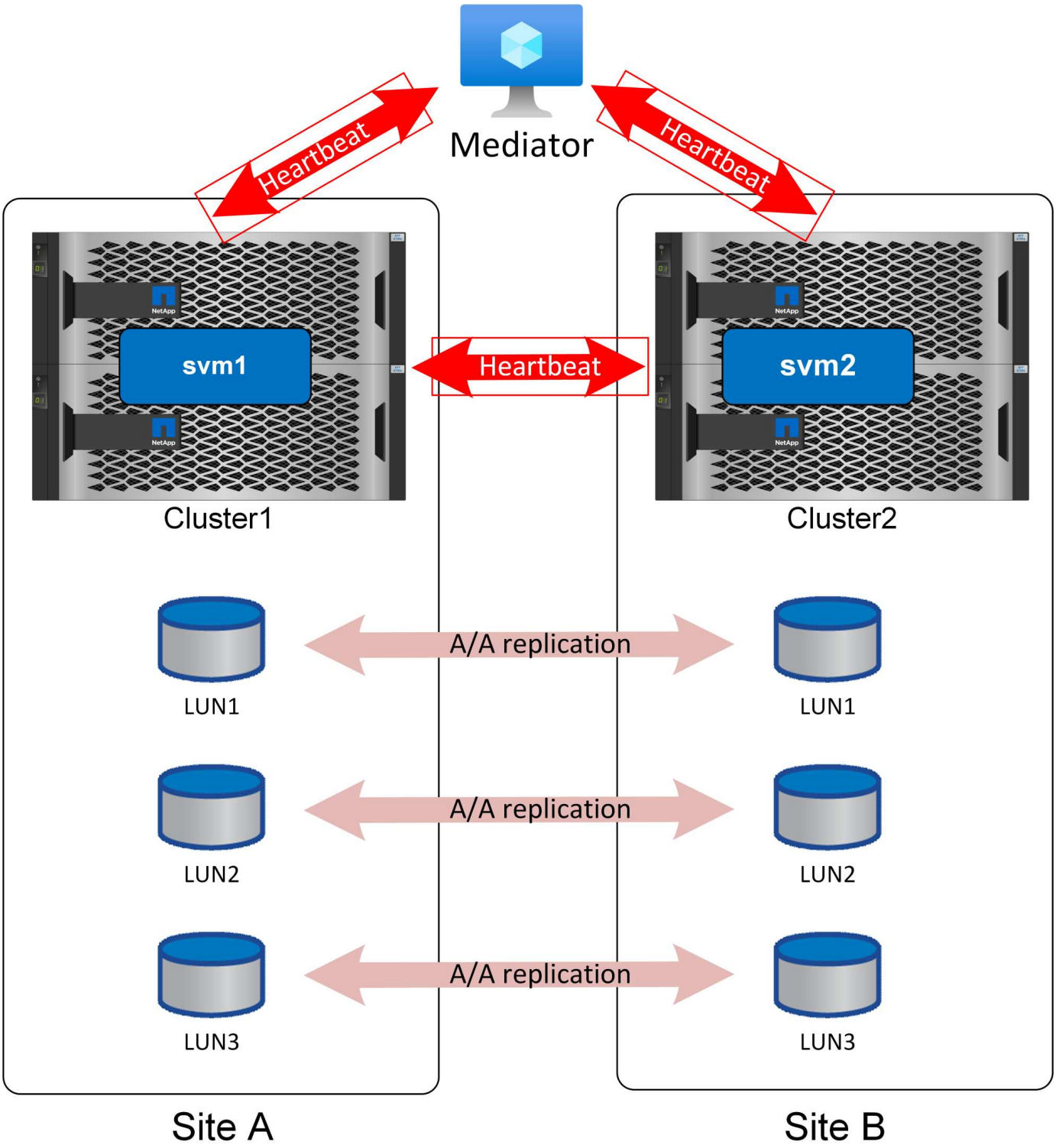
中間器**ONTAP**

ONTAP Mediator 是從 NetApp 支援下載的軟體應用程式、通常部署在小型虛擬機器上。ONTAP Mediator 與 SnapMirror 主動式同步搭配使用時、並不是一種斷路器。它是參與 SnapMirror 主動同步複寫之兩個叢集的替代通訊通道。自動化作業由 ONTAP 根據合作夥伴透過直接連線和協調員所收到的回應來驅動。

資訊媒體**ONTAP**

安全自動化容錯移轉需要中介程序。理想情況下、它會放置在獨立的第三站台、但如果與參與複寫的叢集之一共存、則仍能滿足大多數需求。

調解員實際上並不是決勝者，儘管它實際上發揮了這樣的作用。中介器有助於確定叢集節點的狀態，並在站點發生故障時協助自動切換流程。中介在任何情況下都不會傳輸資料。



自動化容錯移轉的第一項挑戰是大腦分離問題、如果兩個站台彼此之間的連線中斷、就會發生這個問題。應該發生什麼事？您不想讓兩個不同的網站自行指定為資料的保存複本、但如何讓單一網站分辨相對網站的實際損失與無法與相對網站通訊的差異？

這是調解者輸入圖片的地方。如果放置在第三個站台上、而且每個站台都有與該站台的個別網路連線、則每個站台都有額外的路徑來驗證對方的健全狀況。請再次查看上圖、並思考下列案例。

- 如果調解器故障或無法從一個或兩個站台連線、會發生什麼情況？

- 這兩個叢集仍可透過複寫服務所使用的相同連結彼此通訊。
- 資料仍以 RPO = 0 保護提供

- 如果站台 A 故障會發生什麼情況？
  - 站台 B 會看到兩個通訊通道都中斷。
  - 站台 B 將接管資料服務、但不使用 RPO=0 鏡射
- 如果站台 B 故障會發生什麼情況？
  - 站台 A 會看到兩個通訊通道都中斷。
  - 站台 A 會接管資料服務、但不會使用 RPO=0 鏡射

還有一個案例需要考量：資料複寫連結遺失。如果站台之間的複寫連結遺失、 RPO=0 鏡射顯然是不可能的。那麼應該發生什麼事？

這是由偏好的站台狀態所控制。在 SM 合夥關係中、其中一個站台是次要站台。這對正常作業沒有影響、所有資料存取都是對稱的、但如果複寫中斷、則必須中斷連結才能恢復作業。結果是首選站台將在不進行鏡射的情況下繼續作業、而次要站台將停止 IO 處理、直到複寫通訊恢復為止。

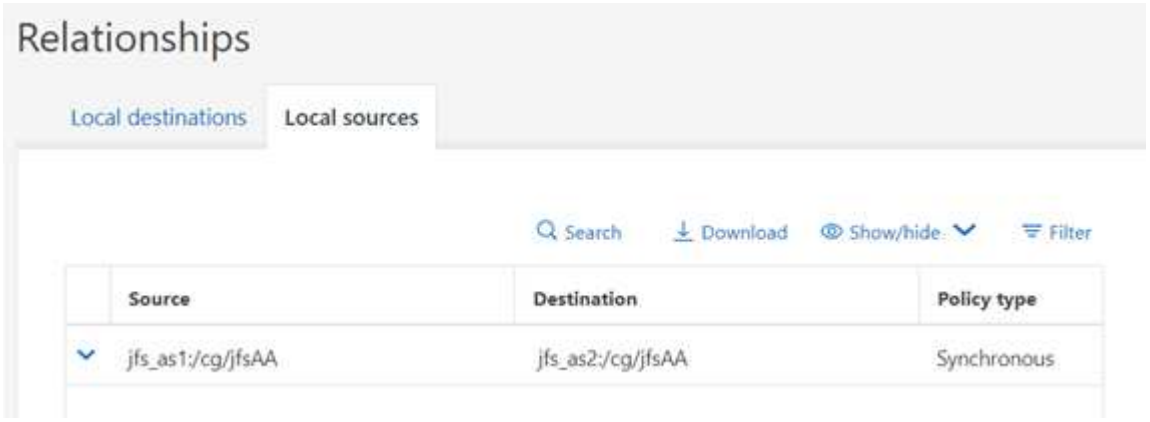
SnapMirror 主動式同步偏好的站台

SnapMirror 主動式同步處理行為是對稱的、但有一個重要的例外是偏好的站台組態。

SnapMirror 作用中同步將一個站台視為「來源」、另一個則視為「目的地」。這表示單向複寫關係、但這不適用於 IO 行為。複寫是雙向的、對稱的、而且在鏡像的兩側、 IO 回應時間相同。

> `source`指定是控制偏好的站台。如果複寫連結遺失、來源複本上的 `LUN` 路徑將繼續提供資料、而目的地複本上的 `LUN` 路徑將無法使用、直到 `SnapMirror` 重新建立複寫並重新進入同步狀態為止。然後路徑將恢復服務資料。

來源 / 目的地組態可透過 SystemManager 檢視：



或在 CLI ：

```
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
```

關鍵在於來源為叢集 1 上的 SVM 。如上所述、「來源」和「目的地」兩詞並未說明複寫資料的流程。這兩個站台都可以處理寫入作業、並將其複寫到另一個站台。實際上、兩個叢集都是來源和目的地。將一個叢集指定為來源的效果、只是控制在複寫連結遺失時、哪個叢集仍保留為讀寫儲存系統。

網路拓撲

統一存取

統一存取網路意味著主機可以存取兩個站台（或同一個站台內的故障網域）上的路徑。

SM — as 的一項重要功能是能夠設定儲存系統、以瞭解主機的位置。將 LUN 對應至指定主機時、您可以指出 LUN 是否接近指定的儲存系統。

特殊警示點設定

特殊警示是指每個叢集的組態、表示特定主機 WWN 或 iSCSI 啟動器 ID 屬於本機主機。這是設定 LUN 存取的第二個選用步驟。

第一步是一般的 igroup 組態。每個 LUN 都必須對應至包含需要存取該 LUN 之主機的 WWN/iSCSI ID 的 igroup 。這會控制哪些主機擁有對 LUN 的 *access* 。

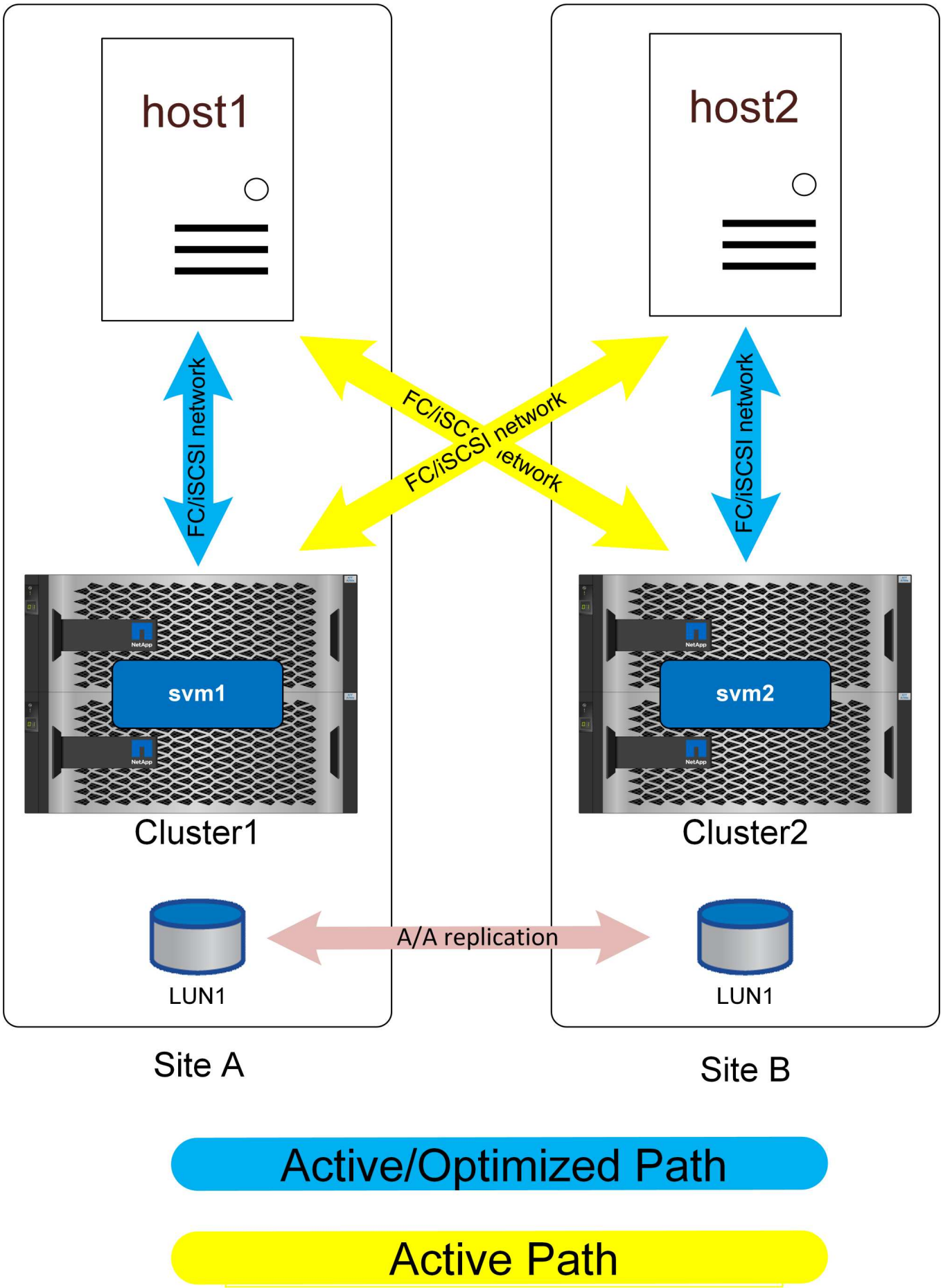
第二個選用步驟是設定主機鄰近度。這無法控制存取、而是控制 *priority* 。

例如、站台 A 的主機可能設定為存取受 SnapMirror 主動式同步保護的 LUN 、而且由於 SAN 延伸至站台、因此該 LUN 可使用站台 A 上的儲存設備或站台 B 上的儲存設備來存取路徑

如果沒有特殊警示點設定、則該主機會同時使用兩個儲存系統、因為這兩個儲存系統都會通告主動 / 最佳化的路徑。如果站台之間的 SAN 延遲和 / 或頻寬受到限制、這可能無法進行設計、您可能希望確保在正常作業期間、每個主機都優先使用本機儲存系統的路徑。這是透過將主機 WWN/iSCSI ID 新增至本機叢集做為近端主機來設定的。這可以在 CLI 或 SystemManager 上完成。

AFF

在 AFF 系統中、設定主機鄰近時、路徑會如下所示。


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

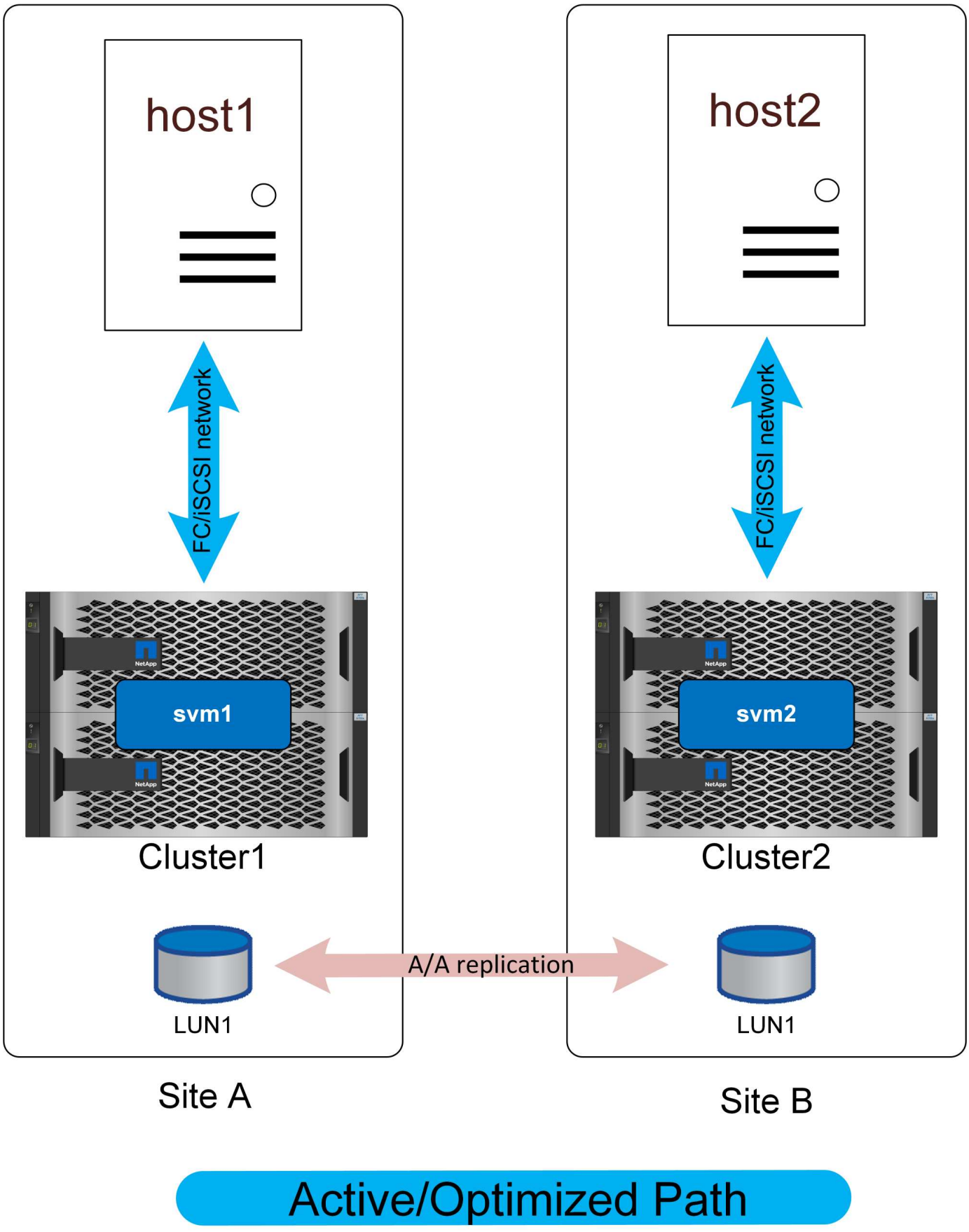
在正常作業中、所有 IO 都是本機 IO 。從本機儲存陣列提供讀取和寫入服務。寫入 IO 當然也需要由本機控制器複寫到遠端系統、然後才會被確認、但所有讀取 IO 都會在本機上提供服務、而且不會因在站台之間瀏覽 SAN 連結而產生額外延遲。

只有當所有主動 / 最佳化路徑都遺失時、才會使用非最佳化路徑。例如、如果站台 A 上的整個陣列失去電力、站台 A 的主機仍能存取站台 B 上陣列的路徑、因此仍可繼續運作、雖然延遲會較高。

由於簡單起見、本機叢集有多個備援路徑未顯示在這些圖表中。ONTAP 儲存系統本身就是 HA 、因此控制器故障不應導致站台故障。只會導致受影響網站上使用本機路徑的變更。

ASA

NetApp ASA 系統可跨叢集上的所有路徑提供雙主動式多重路徑。這也適用於 SM-AS 組態。



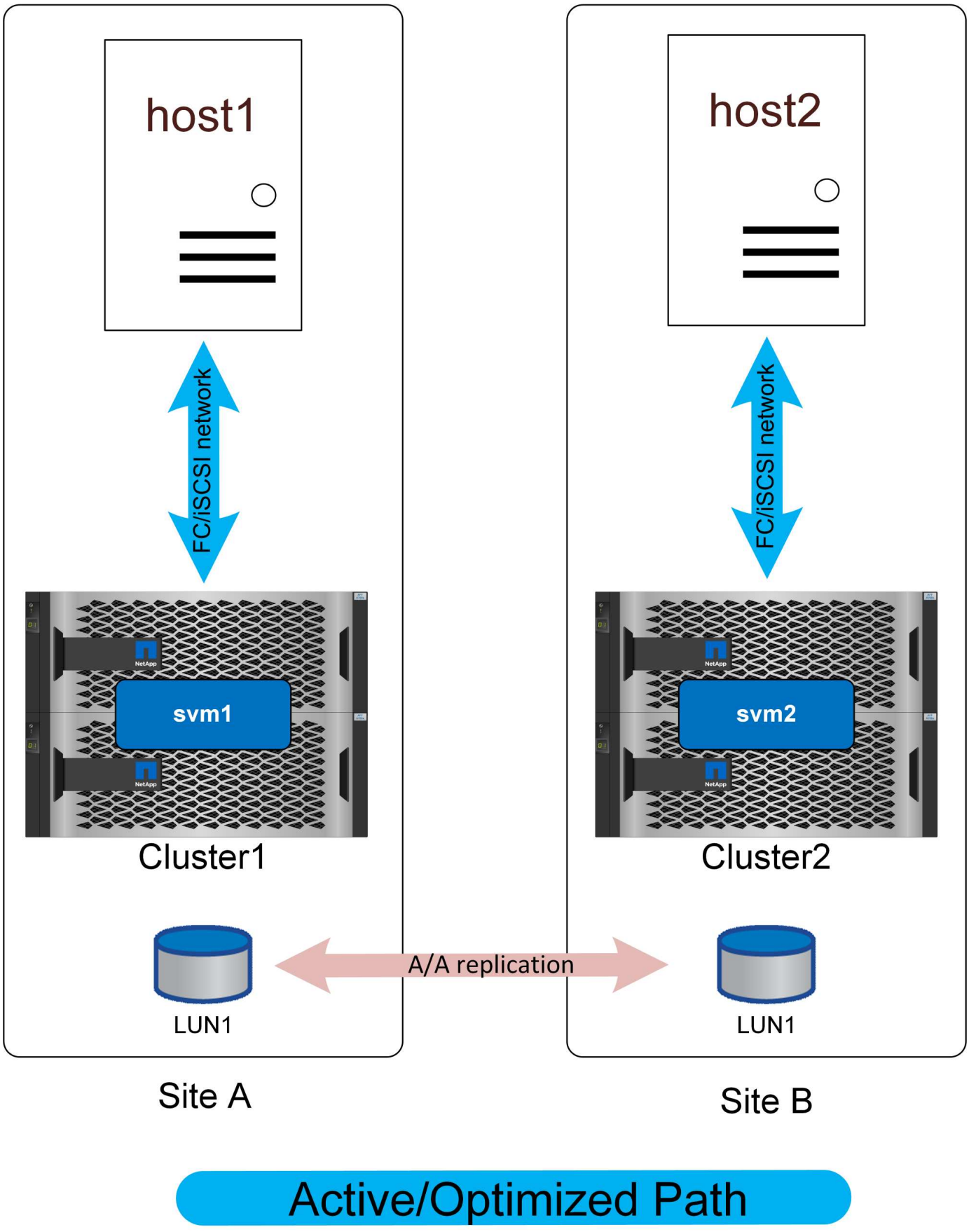
具有非統一存取權的 ASA 組態、其運作方式與 AFF 大致相同。透過統一存取、 IO 就能跨越 WAN 。這可能是或不理想的做法。

如果兩個站台相距 100 公尺、且具備光纖連線能力、則不應偵測到透過 WAN 的額外延遲、但如果站台相距很遠、則兩個站台的讀取效能都會受到影響。與此相反、 AFF 只有在沒有可用的本機路徑時、才會使用這些 WAN 路徑、而且因為所有 IO 都是本機 IO 、所以日常效能會更好。使用非統一存取網路的 ASA 可讓您選擇取得 ASA 的成本和功能效益、而不會造成跨站台延遲存取的損失。

採用低延遲組態的 ASA 具有兩項有趣的優點。首先、它基本上是 * 任何單一主機的效能加倍 * 、因為 IO 可以由兩倍多的控制器使用兩倍的路徑來提供服務。其次、在單一站台環境中、它提供極高的可用度、因為整個儲存系統可能會遺失、而不會中斷主機存取。

不一致的存取

非統一存取網路表示每部主機只能存取本機儲存系統上的連接埠。SAN 不會延伸至站台（或同一站台內的故障網域）。



這種方法的主要優點是 SAN 簡易性、您無需透過網路擴充 SAN 。有些客戶在站台之間沒有足夠的低延遲連線、或缺乏基礎架構、無法透過站台間網路來通道 FC SAN 流量。

不一致存取的缺點是、某些失敗情況（包括遺失複寫連結）會導致部分主機失去儲存設備的存取權。以單一執行個體執行的應用程式、例如原本只在任何指定掛載的單一主機上執行的非叢集資料庫、如果本機儲存連線中斷、就會失敗。資料仍會受到保護、但資料庫伺服器將無法再存取。需要在遠端站台上重新啟動、最好是透過自動化程序來重新啟動。例如、 VMware HA 可偵測一部伺服器上的所有路徑停機情況、並在另一部可用路徑的伺服器上重新啟動 VM 。

相反地、叢集式應用程式（例如 Oracle RAC ）可提供在兩個不同站台同時可用的服務。失去站台並不代表整個應用程式服務都會遺失。執行個體仍可在仍正常運作的站台上執行。

在許多情況下、透過站台對站台連結存取儲存設備的應用程式額外延遲成本是不可接受的。這表示統一網路的可用度提升到最低、因為站台上的儲存設備遺失、可能導致仍需要關閉該故障站台上的服務。

由於簡單起見、本機叢集有多個備援路徑未顯示在這些圖表中。ONTAP 儲存系統本身就是 HA 、因此控制器故障不應導致站台故障。只會導致受影響網站上使用本機路徑的變更。

Oracle 組態

總覽

使用 SnapMirror 主動式同步不一定會新增或變更任何操作資料庫的最佳實務做法。

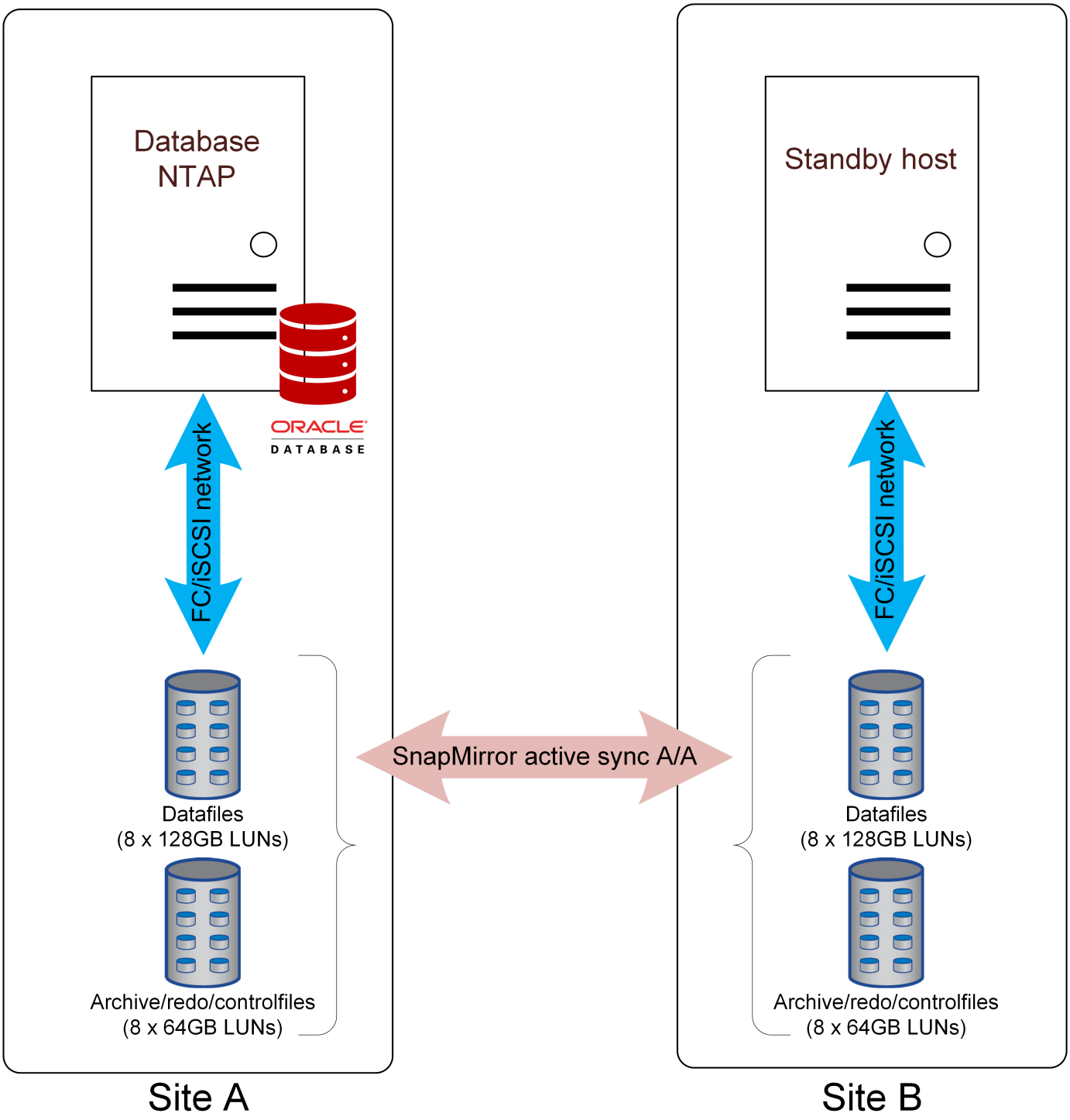
最佳架構取決於業務需求。例如、如果目標是在資料遺失的情況下提供 RPO =0 保護、但 RTO 是放寬的、則使用 Oracle 單一執行個體資料庫並以 SM 方式複寫 LUN 、可能就足以滿足 Oracle 授權標準的要求、而且成本也較低。遠端站台的故障不會中斷作業、而主站台的遺失會導致仍在運作中的站台產生 LUN 、而這些 LUN 已在線上且可供使用。

如果 RTO 更嚴格、則透過指令碼或叢集軟體（例如 Pacemaker 或 Ansible ）進行基本主動被動式自動化、可縮短容錯移轉時間。例如、可將 VMware HA 設定為偵測主要站台上的 VM 故障、並在遠端站台上啟用 VM 。

最後、為了實現極快速的容錯移轉、 Oracle RAC 可部署在各個站台上。RTO 基本上為零、因為資料庫會隨時在線上、且可在兩個站台上使用。

Oracle 單一執行個體

以下說明的範例顯示部署具有 SnapMirror 作用中同步複寫之 Oracle 單一執行個體資料庫的許多選項。



使用預先設定的作業系統進行容錯移轉

SnapMirror 主動式同步功能可在災難恢復站台上提供資料的同步複本、但要讓資料可用、則需要作業系統和相關應用程式。基本自動化可大幅改善整體環境的容錯移轉時間。叢集件產品（例如 Pacemaker ）通常用於在站台之間建立叢集、在許多情況下、容錯移轉程序可以使用簡單的指令碼來驅動。

如果主節點遺失、叢集軟體（或指令碼）將會在替代站台上線。其中一個選項是建立預先針對組成資料庫的 SAN 資源所預先設定的待命伺服器。如果主站台發生故障、叢集軟體或指令碼替代方案會執行類似下列的一系列動作：

1. 偵測主要站台故障
2. 執行 FC 或 iSCSI LUN 的探索
3. 掛載檔案系統和 / 或掛載 ASM 磁碟群組
4. 啟動資料庫

此方法的主要需求是在遠端站台上執行作業系統。它必須預先設定 Oracle 二進位檔、這也表示 Oracle 修補等工作必須在主要站台和待命站台上執行。或者、 Oracle 二進位檔可鏡射至遠端站台、並在宣告災難時掛載。

實際的啟動程序很簡單。LUN 探索等命令每個 FC 連接埠只需要幾個命令。檔案系統掛載只是一個 `mount`命令、只要一個命令、即可在 CLI 上啟動和停止資料庫和 ASM 。

使用虛擬化作業系統進行容錯移轉

資料庫環境的容錯移轉可延伸至包含作業系統本身。理論上、此容錯移轉可以使用開機 LUN 來完成、但通常是使用虛擬化的作業系統來完成。此程序類似於下列步驟：

1. 偵測主要站台故障
2. 裝載託管資料庫伺服器虛擬機器的資料存放區
3. 啟動虛擬機器
4. 手動啟動資料庫、或將虛擬機器設定為自動啟動資料庫。

例如、 ESX 叢集可以跨越站台。在發生災難時、虛擬機器可在移至災難恢復站台後上線。

儲存設備故障保護

上圖顯示的用途"不一致的存取"、其中 SAN 並未延伸至各個站台。這可能比較容易設定、在某些情況下、可能是目前 SAN 功能唯一的選項、但也表示主要儲存系統故障會導致資料庫中斷、直到應用程式容錯移轉為止。

為了獲得更高的恢復能力、您可以使用部署解決方案"統一存取"。如此可讓應用程式繼續使用從另一站點廣告的路徑運作。

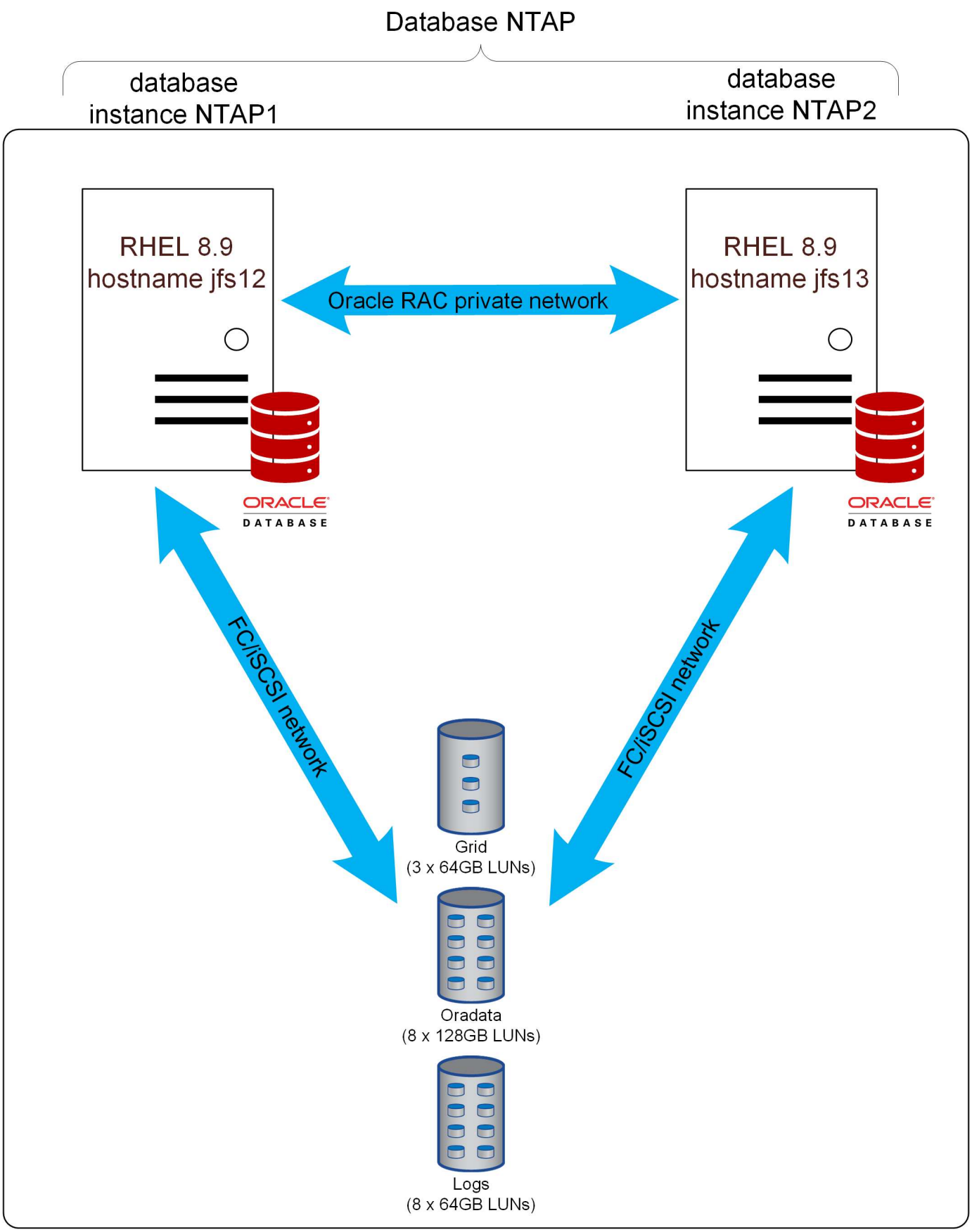
Oracle Extended RAC

許多客戶透過在各個站台之間延伸 Oracle RAC 叢集來最佳化 RTO 、進而實現完全主動式的組態。整體設計變得更複雜、因為它必須包含 Oracle RAC 的仲裁管理。

傳統的延伸 RAC 叢集式仰賴 ASM 鏡射來提供資料保護。這種方法可行、但也需要大量手動設定步驟、並對網路基礎架構造成負擔。相反地、讓 SnapMirror 主動式同步處理負責資料複寫、可大幅簡化解決方案。同步、中斷後重新同步、容錯移轉和仲裁管理等作業都變得更簡單、而且 SAN 不需要分散在各個站台上、如此就能簡化 SAN 的設計與管理。

複寫

瞭解 SnapMirror 主動式同步上的 RAC 功能的關鍵在於將儲存裝置視為單一 LUN 集、並以鏡射儲存設備為主控。例如：



沒有主要複本或鏡射複本。從邏輯上來說、每個 LUN 只有一個複本、而位於兩個不同儲存系統上的 SAN 路徑上則有該 LUN 可用。從主機的角度來看、沒有儲存容錯移轉、而是有路徑變更。當其他路徑保持連線時、各種故障事件可能會導致通往 LUN 的特定路徑遺失。SnapMirror 主動式同步可確保所有作業路徑都能使用相同的資

料。

儲存組態

在此範例組態中、 ASM 磁碟的組態與企業儲存設備上任何單站台 RAC 組態的組態相同。由於儲存系統提供資料保護、因此會使用 ASM 外部備援。

統一存取與不通知存取

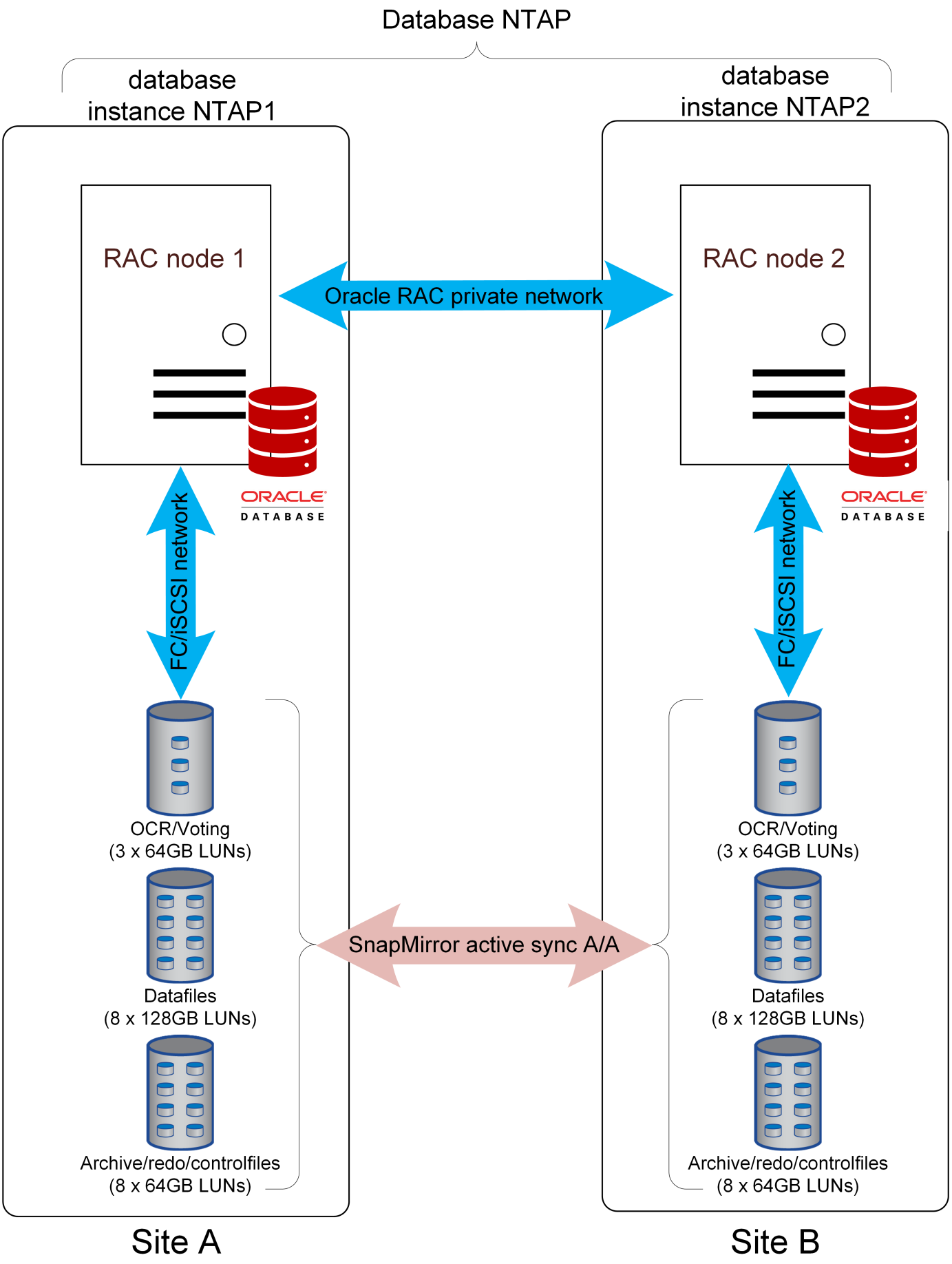
在 SnapMirror 主動式同步上使用 Oracle RAC 最重要的考量、是使用統一或非統一存取。

統一存取意味著每個主機都可以看到兩個叢集上的路徑。非統一存取表示主機只能看到本機叢集的路徑。

這兩個選項都不是特別推薦或不鼓勵的。有些客戶可以隨時連線到網站、有些客戶可能沒有這種連線能力、或是他們的 SAN 基礎架構不支援長距離 ISL 。

不一致的存取

從 SAN 的角度來看、不一致的存取更容易設定。


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

此方法的主要缺點"不一致的存取"是、站台對站台 ONTAP 連線中斷或儲存系統遺失、將導致一個站台的資料庫執行個體遺失。這顯然不是理想的做法、但在交換較簡單的 SAN 組態時、這可能是可接受的風險。

統一存取

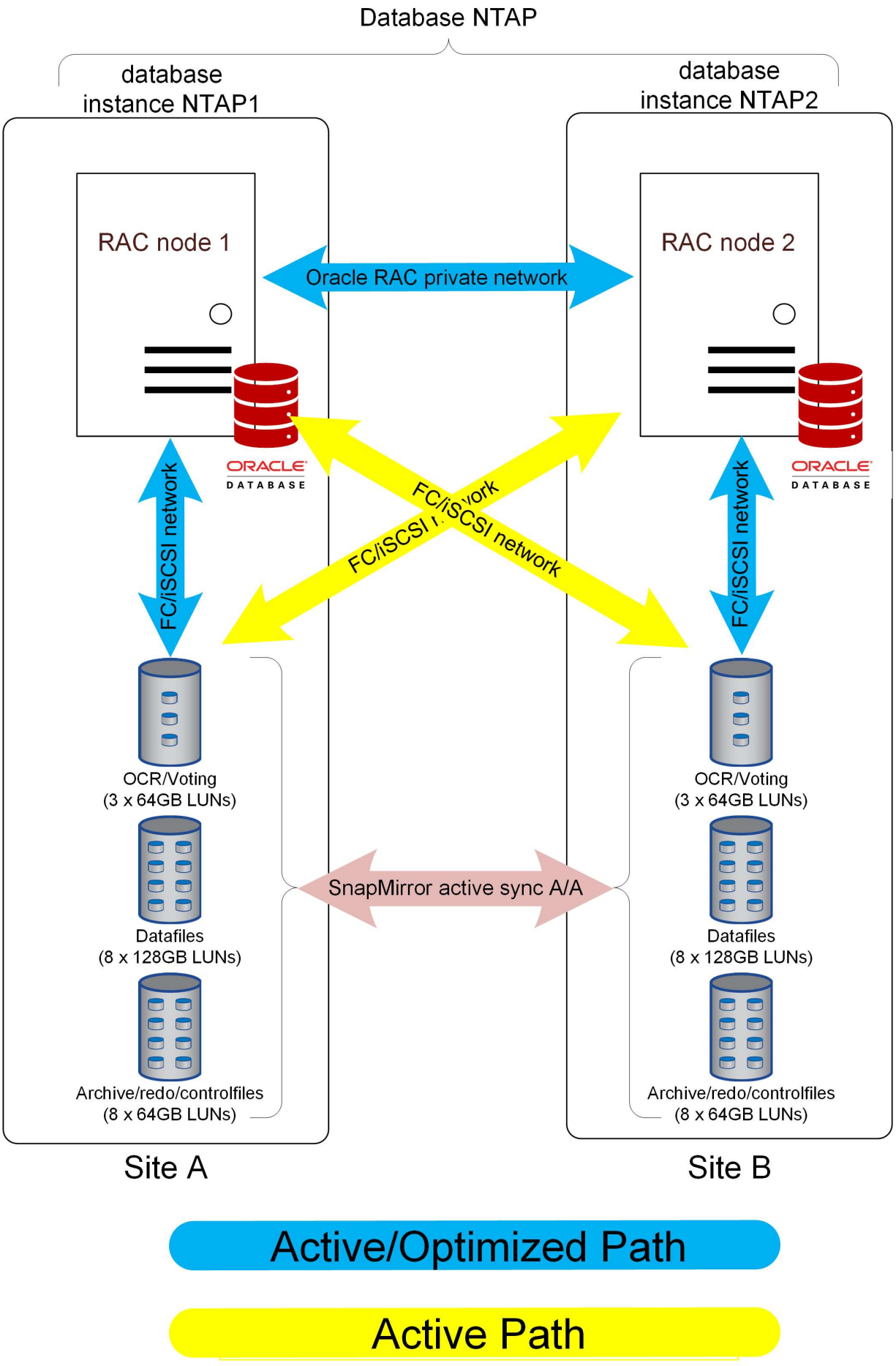
統一存取需要將 SAN 延伸至各個站台。主要優點是儲存系統的遺失不會導致資料庫執行個體遺失。相反地、它會導致路徑目前正在使用的多重路徑變更。

有幾種方法可以設定不一致的存取。

在下圖中、也會出現一些作用中但未最佳化的路徑、這些路徑會在簡單的控制器故障期間使用、但這些路徑並不代表簡化圖表的目的。

具有鄰近設定的 **AFF**

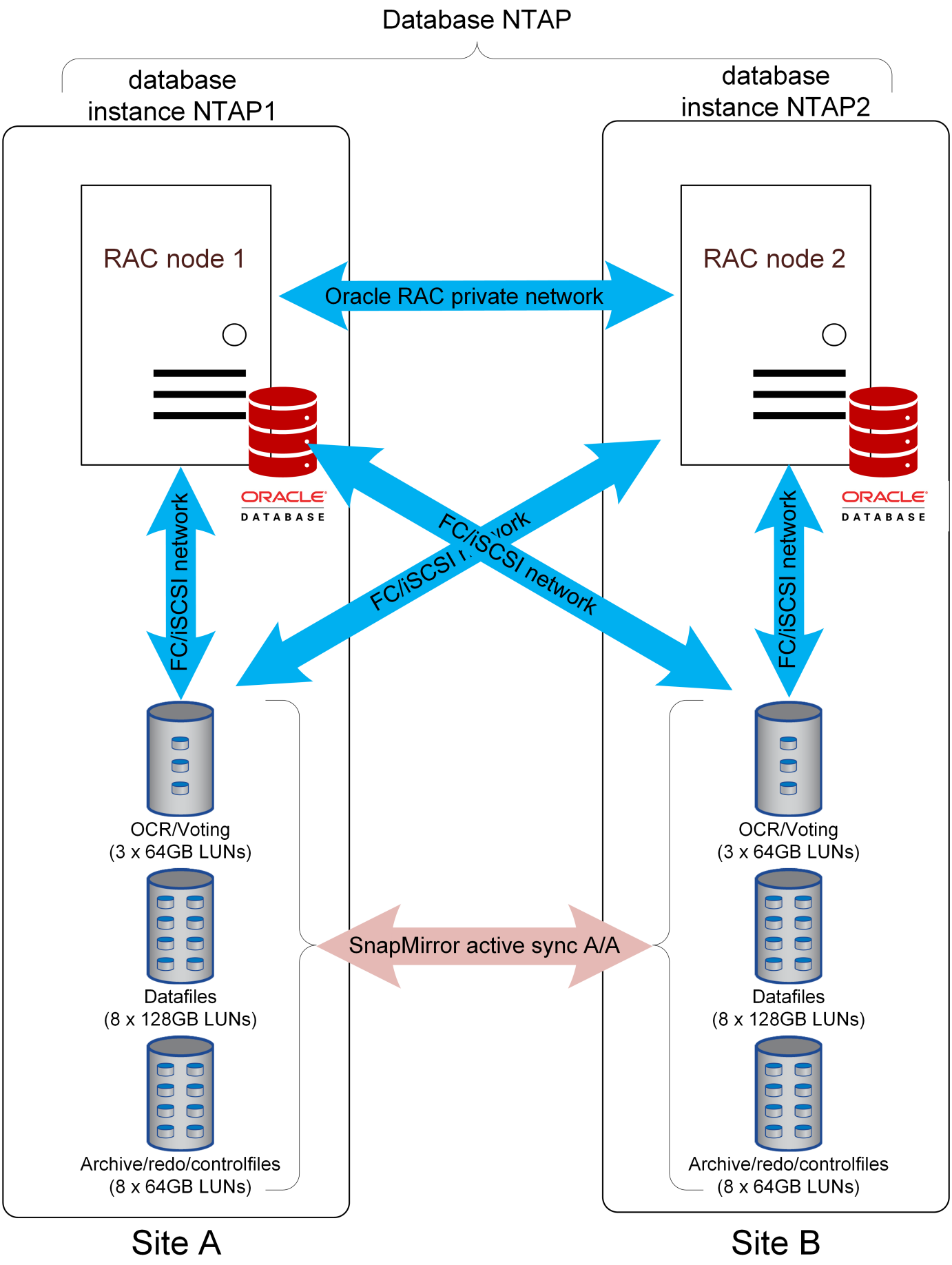
如果站台之間存在嚴重延遲、則可以使用主機鄰近設定來設定 AFF 系統。如此一來、每個儲存系統就能知道哪些主機是本機主機、哪些是遠端主機、並適當地指派路徑優先順序。


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

在正常作業中、每個 Oracle 執行個體都會優先使用本機主動 / 最佳化路徑。結果是所有讀取都會由區塊的本機複本提供服務。這會產生最低的可能延遲。寫入 IO 也同樣會向下傳送至本機控制器的路徑。在確認之前必須複寫 IO 、因此仍會產生跨越站台對站台網路的額外延遲、但在同步複寫解決方案中無法避免這種情況。

ASA / AFF 不含感應設定

如果站台之間沒有明顯的延遲、則可在不設定主機鄰近設定的情況下設定 AFF 系統、或使用 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

每個主機都可以使用兩個儲存系統上的所有作業路徑。如此一來、每部主機就能充分發揮兩個叢集的效能潛力、而不只是一個叢集、進而大幅提升效能。

有了 ASA 、不僅兩個叢集的所有路徑都會視為作用中且最佳化、而且合作夥伴控制器上的路徑也會是作用中的。結果是整個叢集上的全作用中 SAN 路徑、

ASA 系統也可用於非統一存取組態。由於不存在跨站台路徑、因此透過 ISL 的 IO 不會影響效能。

RAC tiebreaker

雖然使用 SnapMirror 主動式同步的延伸 RAC 是 IO 的對稱架構、但有一個例外是連線至大腦分割管理。

如果複寫連結遺失且兩個站台都沒有仲裁、會發生什麼情況？應該發生什麼事？此問題同時適用於 Oracle RAC 和 ONTAP 行為。如果無法跨站台複寫變更、而您想要恢復作業、則其中一個站台必須生存、另一個站台必須無法使用。

可"資訊媒體ONTAP"在 ONTAP 層解決此需求。RAC 中斷有多個選項。

Oracle tiebreaker

管理分離式 Oracle RAC 風險的最佳方法、是使用奇怪數量的 RAC 節點、最好是使用第三站台的斷路器。如果第三個站台無法使用、則可將斷路器執行個體放置在兩個站台的其中一個站台、有效地將其指定為慣用的生存者站台。

Oracle 和 CSS_critical

對於偶數個節點、預設的 Oracle RAC 行為是叢集中的其中一個節點將被視為比其他節點更重要。具有較高優先順序節點的站台會在站台隔離後繼續運作、而另一個站台上的節點則會被移除。優先順序是以多個因素為基礎、但您也可以使用設定來控制此行為 `css_critical`。

在"範例"架構中、 RAC 節點的主機名稱為 jfs12 和 jfs13 。的目前設定 `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.
```

如果您想要將具有 jfs12 的站台設為慣用站台、請在站台 A 節點上將此值變更為「是」、然後重新啟動服務。

```
[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.
```

故障案例

總覽

規劃完整的 SnapMirror 主動式同步應用程式架構時、需要瞭解 SM-AS 在各種計畫性和非計畫性容錯移轉案例中的回應方式。

針對下列範例、假設站台 A 已設定為慣用站台。

喪失複寫連線能力

如果 SM-AS 複寫中斷、寫入 IO 就無法完成、因為叢集無法將變更複寫到相反的站台。

站台 **A** （慣用站台）

偏好的站台上的複寫連結失敗、在寫入 IO 處理中會有大約 15 秒的暫停、因為 ONTAP 會在判斷複寫連結確實無法連線之前、重試複寫的寫入作業。15 秒後、站台 A 系統會恢復讀寫 IO 處理。SAN 路徑不會變更、 LUN 也會保持連線。

站台**B**

由於站台 B 不是 SnapMirror 作用中同步偏好的站台、因此其 LUN 路徑將在大約 15 秒後變成無法使用。

儲存系統故障

儲存系統故障的結果與遺失複寫連結的結果幾乎完全相同。當仍在運作的站台發生 IO 暫停約 15 秒。一旦超過 15 秒、 IO 就會像往常一樣繼續在該站台上進行。

調解員遺失

中介服務無法直接控制儲存作業。它可作為叢集之間的替代控制路徑。它主要用於自動化容錯移轉、而不會有發生分裂的風險。在正常作業中、每個叢集都會將變更複寫到其合作夥伴、因此每個叢集都可以驗證合作夥伴叢集是否在線上並提供資料。如果複寫連結失敗、複寫就會停止。

安全自動容錯移轉需要協調員、因為否則儲存叢集就無法判斷雙向通訊是否因為網路中斷或實際儲存設備故障而中斷。

中介程序為每個叢集提供替代路徑、以驗證其合作夥伴的健全狀況。案例如下：

- 如果叢集可以直接聯絡其合作夥伴、複寫服務就可以運作。無需採取任何行動。
- 如果偏好的站台無法直接聯絡其合作夥伴或透過中介人聯絡、則會假設該合作夥伴實際上無法使用、或是被隔離、並已將其 LUN 路徑離線。接著、偏好的站台會繼續釋放 RPO=0 狀態、並繼續處理讀取和寫入 IO 。
- 如果非偏好的站台無法直接聯絡其合作夥伴、但可以透過協調器聯絡、則會使其路徑離線、並等待複寫連線的恢復。
- 如果非偏好的站台無法直接或透過營運協調員聯絡其合作夥伴、則會假設該合作夥伴實際上無法使用、或是被隔離、並已將其 LUN 路徑離線。然後、非偏好的站台會繼續釋放 RPO = 0 狀態、並繼續處理讀取和寫入 IO 。它將扮演複寫來源的角色、並將成為新的慣用站台。

如果調解器完全無法使用：

- 複寫服務因任何原因而失敗、包括非慣用站台或儲存系統故障、將導致偏好的站台釋放 RPO = 0 狀態、並恢復讀寫 IO 處理。非慣用站台將使其路徑離線。
- 偏好的站台故障將導致中斷、因為非偏好的站台將無法驗證相對站台是否確實離線、因此非偏好的站台無法安全恢復服務。

還原服務

解決故障（例如還原站台對站台連線或啟動故障系統）後、 SnapMirror 作用中同步端點會自動偵測是否存在錯誤的複寫關係、並將其恢復至 RPO=0 狀態。重新建立同步複寫後、故障路徑將再次上線。

在許多情況下、叢集式應用程式會自動偵測失敗路徑的傳回、這些應用程式也會重新上線。在其他情況下、可能需要主機層級的 SAN 掃描、或是需要手動將應用程式恢復上線。這取決於應用程式及其設定方式、一般而言、這類工作可以輕鬆自動化。ONTAP 本身具有自我修復功能、不應需要任何使用者介入、即可恢復 RPO = 0 儲存作業。

手動容錯移轉

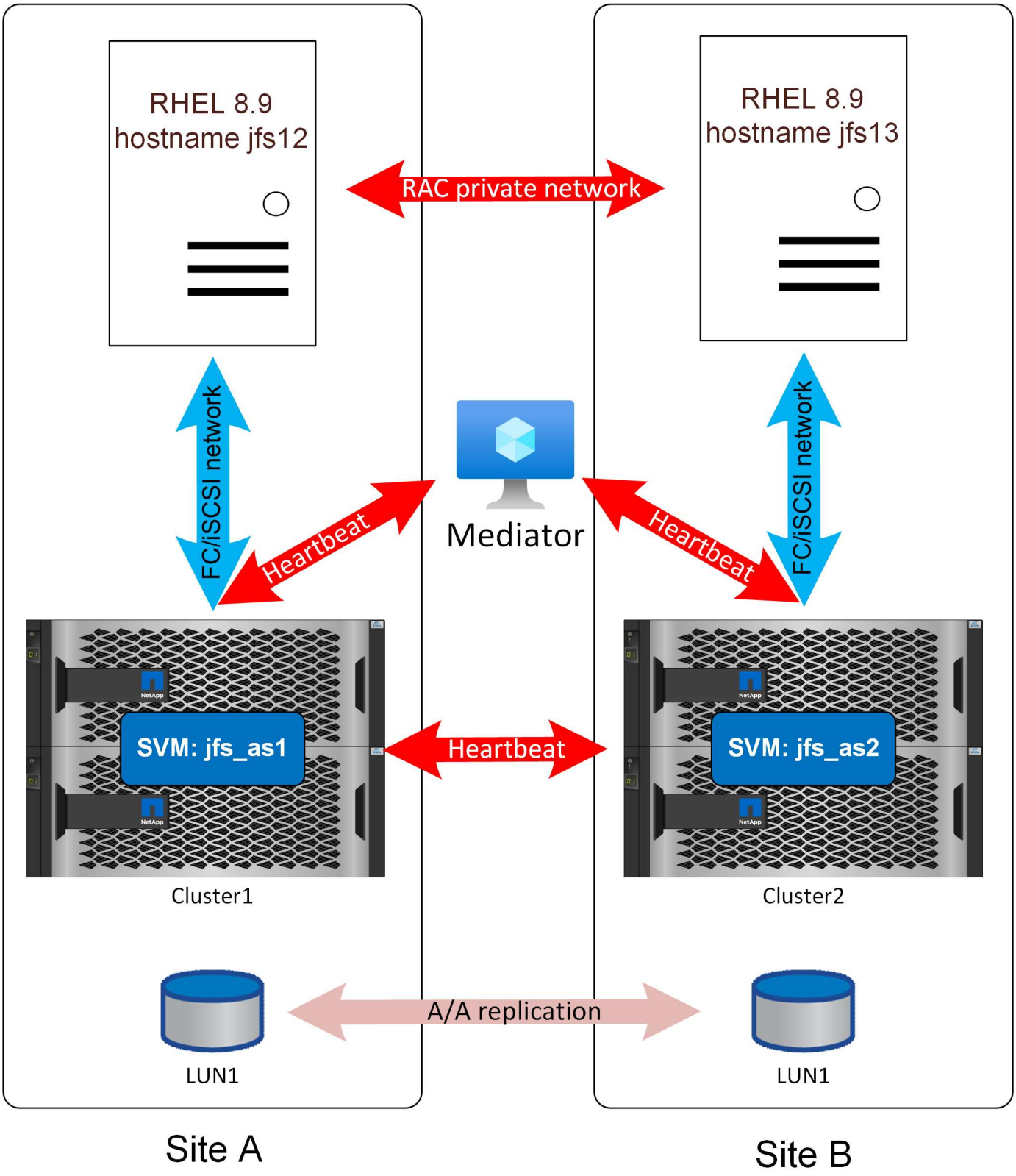
變更偏好的站台需要簡單的操作。IO 會暫停一秒或兩秒、作為叢集之間複寫行為切換的權限、但 IO 不會受到影響。

範例架構

本節所示的詳細故障範例是根據下列架構而定。

這只是 SnapMirror Active Sync 上 Oracle 資料庫的眾多選項之一。選擇此設計是因為它說明了一些較複雜的案例。

在此設計中，假設站台 A 是在設定的"偏好的網站"。



RAC 互連故障

喪失 Oracle RAC 複寫連結會產生類似於 SnapMirror 連線中斷的結果、但預設會縮短逾時時間。在預設設定下、 Oracle RAC 節點會在遺失儲存連線後等待 200 秒後才會消失、但在 RAC 網路心跳中斷後、只會等待 30 秒。

CRS 訊息類似於下列訊息。您可以看到 30 秒的逾時時間。由於 CSS_critical 設定在站台 A 上的 jfs12 上、這將是要生存的站台、而站台 B 上的 jfs13 將被逐出。

```
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 通訊失敗

如果 SnapMirror 主動式同步複寫連結、則無法完成寫入 IO 、因為叢集無法將變更複寫到另一個站台。

站台A

複寫連結失敗的站台 A 在寫入 IO 處理中會暫停約 15 秒、因為 ONTAP 會在判斷複寫連結確實無法運作之前、嘗試複寫寫入內容。經過 15 秒後、站台 A 上的 ONTAP 叢集會恢復讀寫 IO 處理。SAN 路徑不會變更、 LUN 也會保持連線。

站台B

由於站台 B 不是 SnapMirror 作用中同步偏好的站台、因此其 LUN 路徑將在大約 15 秒後變成無法使用。

複寫連結的時間是 15 ： 19 ： 44 。Oracle RAC 的第一個警告會在 200 秒逾時（由 Oracle RAC 參數 disktimeout 控制）接近 100 秒後到達。

```
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.
```

達到 200 秒投票磁碟逾時後、此 Oracle RAC 節點將自行從叢集移除並重新開機。

網路互連性總故障

如果站台之間的複寫連結完全遺失、則 SnapMirror 主動式同步和 Oracle RAC 連線都會中斷。

Oracle RAC SPLIT 偵測功能與 Oracle RAC 儲存設備活動訊號有關。如果站台對站台連線中斷導致 RAC 網路心跳和儲存複寫服務同時中斷、結果是 RAC 站台無法透過 RAC 互連或 RAC 投票磁碟進行跨站台通訊。在預設設定下、這兩個站台可能會被移除、因此產生一組偶數的節點。具體行為將取決於事件順序、 RAC 網路和磁碟心跳輪詢的時間。

雙站台中斷的風險可透過兩種方式解決。首先、"斷路器"可以使用組態。

如果第三站台無法使用、則可調整 RAC 叢集上的「錯誤數」參數來解決此風險。根據預設值、 RAC 網路心跳逾時為 30 秒。這通常是 RAC 用來識別故障的 RAC 節點、並將其從叢集中移除。它也與投票磁碟活動訊號有連線。

例如、如果反鏟切斷了傳輸 Oracle RAC 和儲存複寫服務站台間流量的處理通道、則 30 秒錯過計數倒數將會開始。如果 RAC 偏好的站台節點無法在 30 秒內重新與另一個站台建立連絡、而且也無法使用投票磁碟來確認對方站台在相同的 30 秒內停機、則偏好的站台節點也會被移除。結果是資料庫完全中斷。

視發生錯誤數輪詢的時間而定、 30 秒可能不足以讓 SnapMirror 作用中同步逾時、並允許首選站台上的儲存設備在 30 秒的時間過期之前恢復服務。這 30 秒的時間範圍可以增加。

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

此值可讓偏好的站台上的儲存系統在錯誤計數逾時過期之前恢復作業。然後、只會將 LUN 路徑移除站台上的節點移除。範例如下：

```
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 Support 強烈建議您不要變更錯誤數或磁碟逾時參數、以解決組態問題。不過、在許多情況下、變更這些參數可能是必要且不可避免的、包括 SAN 開機、虛擬化及儲存複寫組態。例如、如果 SAN 或 IP 網路發生穩定性問題、導致 RAC 遷離、您應該修正基礎問題、而不要收取錯誤數或磁碟逾時的值。為了解決組態錯誤而變更逾時是掩蓋問題、而非解決問題。根據基礎架構的設計層面、變更這些參數以正確設定 RAC 環境、是不同的、且與 Oracle 支援聲明一致。使用 SAN 開機時、通常會調整到最大 200 的 miskcount 、以符合磁碟逾時。如需其他資訊、請參閱"此連結"。

站台故障

儲存系統或站台故障的結果與遺失複寫連結的結果幾乎相同。當仍在運作的站台寫入時、IO 應該會暫停約 15 秒。一旦超過 15 秒、 IO 就會像往常一樣繼續在該站台上進行。

如果只有儲存系統受到影響、故障站台上的 Oracle RAC 節點將會遺失儲存服務、並在遷離和後續重新開機之

前、輸入相同的 200 秒磁碟逾時倒數。

```
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.
```

RAC 節點上遺失儲存服務的 SAN 路徑狀態如下：

```
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 主機偵測到路徑遺失速度快於 200 秒、但從資料庫的角度來看、故障站台上的用戶端連線仍會在預設 Oracle RAC 設定下凍結 200 秒。完整資料庫作業只會在遷離完成後恢復。

同時、另一個站台上的 Oracle RAC 節點會記錄其他 RAC 節點的遺失。否則、它會繼續如常運作。

```
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
```

中介故障

中介服務無法直接控制儲存作業。它可作為叢集之間的替代控制路徑。它主要用於自動化容錯移轉、而不會有發生分裂的風險。

在正常作業中、每個叢集都會將變更複寫到其合作夥伴、因此每個叢集都可以驗證合作夥伴叢集是否在線上並提供資料。如果複寫連結失敗、複寫就會停止。

安全自動化作業需要協調員的原因、是因為儲存叢集無法判斷雙向通訊是否因為網路中斷或實際儲存設備故障而中斷。

中介程序為每個叢集提供替代路徑、以驗證其合作夥伴的健全狀況。案例如下：

- 如果叢集可以直接聯絡其合作夥伴、複寫服務就可以運作。無需採取任何行動。
- 如果偏好的站台無法直接聯絡其合作夥伴或透過中介人聯絡、則會假設該合作夥伴實際上無法使用、或是被隔離、並已將其 LUN 路徑離線。接著、偏好的站台會繼續釋放 RPO=0 狀態、並繼續處理讀取和寫入 IO 。
- 如果非偏好的站台無法直接聯絡其合作夥伴、但可以透過協調器聯絡、則會使其路徑離線、並等待複寫連線的恢復。
- 如果非偏好的站台無法直接或透過營運協調員聯絡其合作夥伴、則會假設該合作夥伴實際上無法使用、或是被隔離、並已將其 LUN 路徑離線。然後、非偏好的站台會繼續釋放 RPO = 0 狀態、並繼續處理讀取和寫入 IO 。它將扮演複寫來源的角色、並將成為新的慣用站台。

如果調解器完全無法使用：

- 由於任何原因而導致複寫服務失敗、將導致首選站台釋放 RPO = 0 狀態、並恢復讀寫 IO 處理。非慣用站台將使其路徑離線。

- 偏好的站台故障將導致中斷、因為非偏好的站台將無法驗證相對站台是否確實離線、因此非偏好的站台無法安全恢復服務。

服務還原

SnapMirror 可以自我修復。SnapMirror 主動式同步會自動偵測是否存在錯誤的複寫關係、並將其恢復至 RPO = 0 狀態。重新建立同步複寫後、路徑將再次上線。

在許多情況下、叢集式應用程式會自動偵測失敗路徑的傳回、這些應用程式也會重新上線。在其他情況下、可能需要主機層級的 SAN 掃描、或是需要手動將應用程式恢復上線。

這取決於應用程式及其設定方式、一般而言、這類工作可以輕鬆自動化。SnapMirror 主動式同步本身是自行修正的、在電源和連線恢復後、不應需要任何使用者介入、即可恢復 RPO = 0 儲存作業。

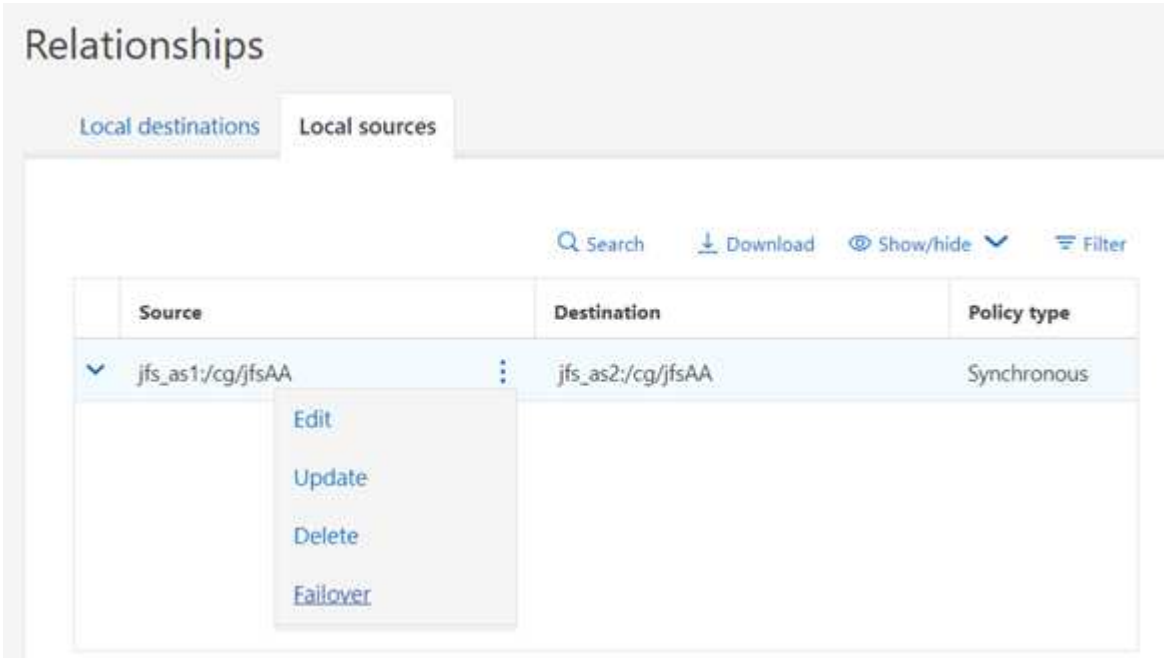
手動容錯移轉

「容錯移轉」一詞並不表示使用 SnapMirror 主動式同步進行複寫的方向、因為它是雙向複寫技術。相反地、「容錯移轉」是指在發生故障時、哪個儲存系統是偏好的站台。

例如、您可能想要在關閉站台進行維護之前、或是在執行 DR 測試之前、執行容錯移轉以變更偏好的站台。

變更偏好的站台需要簡單的操作。IO 會暫停一秒或兩秒、作為叢集之間複寫行為切換的權限、但 IO 不會受到影響。

GUI 範例：



透過 CLI 將其改回的範例：

```
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
```