Oracle 災難恢復

Enterprise applications

NetApp
February 11, 2026

目錄

Oracle 災難恢復

總覽

災難恢復是指在發生災難性事件（例如破壞儲存系統甚至整個站台的火災）之後還原資料服務。

本文件取代先前發佈的技術報告 _TR-4591 ： Oracle Data Protection _ 和 _TR-4592 ： Oracle on MetroCluster 。 _

當然、災難恢復可以透過使用 SnapMirror 簡單複寫資料來完成、許多客戶會在每小時更新鏡射複本。

對於大多數客戶而言、 DR 不只需要擁有遠端資料複本、還需要能夠快速使用該資料。NetApp 提供兩種技術來滿足這種需求： MetroCluster 和 SnapMirror 主動同步

MetroCluster 指的是硬體組態中的 ONTAP 、其中包括低階同步鏡射儲存設備和許多其他功能。MetroCluster 等整合式解決方案可簡化現今複雜的橫向擴充資料庫、應用程式及虛擬化基礎架構。它以一個簡單的中央儲存陣列取代多種外部資料保護產品和策略。它也能在單一叢集式儲存系統中提供整合式備份、還原、災難恢復和高可用度（ HA ）。

SnapMirror 主動式同步（ SM-AS ）是以 SnapMirror 同步為基礎。使用 MetroCluster 、每個 ONTAP 控制器都負責將其磁碟機資料複寫到遠端位置。有了 SnapMirror 主動式同步、您基本上擁有兩個不同的 ONTAP 系統、可維護 LUN 資料的獨立複本、但可以合作呈現該 LUN 的單一執行個體。從主機的角度來看、這是單一 LUN 實體。

SM 與 MCC 比較

SM-AS 和 MetroCluster 在整體功能上相似，但在實作 RPO =0 複寫的方式及其管理方式上有重要差異。SnapMirror 非同步和同步也可作為 DR 計畫的一部分使用，但它們並非設計為 HA 重新安裝技術。

- MetroCluster 組態更像是一個整合式叢集，節點分散在各個站台上。SM-AS 的運作方式類似於兩個原本是相互不同的叢集，這些叢集正在合作為選取的 RPO = 0 同步複寫 LUN 提供服務。
- MetroCluster 組態中的資料只能在任何指定時間從特定站台存取。另一個資料複本位於另一個站台，但資料是被動的。如果沒有儲存系統容錯移轉，就無法存取。
- MetroCluster 和 SM-AS 執行鏡像會在不同層級執行。MetroCluster 鏡射是在 RAID 層執行。低階資料會使用 SyncMirror 以鏡射格式儲存。在 LUN ，磁碟區和傳輸協定層，使用鏡像幾乎是不可見的。
- 相反地， SM-AS 鏡射則發生在傳輸協定層。這兩個叢集都是整體上的不相關叢集。資料的兩個複本同步後，兩個叢集只需鏡射寫入。當某個叢集發生寫入時，它會複寫到另一個叢集。只有當兩個站台的寫入作業完成時，才會將寫入內容確認給主機。除了這種傳輸協定分割行為之外，這兩個叢集都是正常的 ONTAP 叢集。
- MetroCluster 的主要角色是大規模複寫。您可以使用 RPO=0 和接近零的 RTO 來複寫整個陣列。這簡化了容錯移轉程序，因為只有一件事需要容錯移轉，而且在容量和 IOPS 方面的擴充能力極佳。
- SMAS 的一個關鍵使用案例是精細複寫。有時候您不想將所有資料複寫為單一單元，或者您需要選擇性地容錯移轉特定工作負載。
- 另一個用於 SM-AS 的重要使用案例是用於雙主動式作業，您想要在兩個不同位置的兩個不同叢集上提供完全可用的資料複本，而且效能特性相同，如果需要，也不需要在各個站台之間擴充 SAN 。您可以讓應用程式同時在兩個站台上執行，以降低容錯移轉作業期間的整體 RTO 。

MetroCluster

使用 **MetroCluster** 進行災難恢復

MetroCluster 是一項 ONTAP 功能、可在各個站台之間使用 RPO=0 同步鏡射來保護您的 Oracle 資料庫、並可在單一 MetroCluster 系統上擴充至支援數百個資料庫。

使用也很簡單。使用 MetroCluster 並不一定會增加或變更任何用於營運企業應用程式和資料庫的最佳競賽。

通常的最佳實務做法仍適用、如果您的需求只需要 RPO = 0 資料保護、則 MetroCluster 會滿足您的需求。然而、大多數客戶不僅使用 MetroCluster 來保護 RPO =0 資料、還能在災難期間改善 RTO 、並在站台維護活動中提供透明的容錯移轉。

實體架構

瞭解 Oracle 資料庫在 MetroCluster 環境中的運作方式、需要對 MetroCluster 系統的實體設計進行一些說明。

本文件取代先前發佈的技術報告 _TR-4592 ： Oracle on MetroCluster 。 _

MetroCluster 可在 **3** 種不同組態中使用

- HA 可與 IP 連線配對
- HA 可與 FC 連線配對
- 單一控制器、具備 FC 連線能力

術語「連線」是指用於跨站台複寫的叢集連線。它並不指主機協定。無論叢集間通訊所使用的連線類型為何、 MetroCluster 組態中的所有主機端通訊協定都會如常支援。

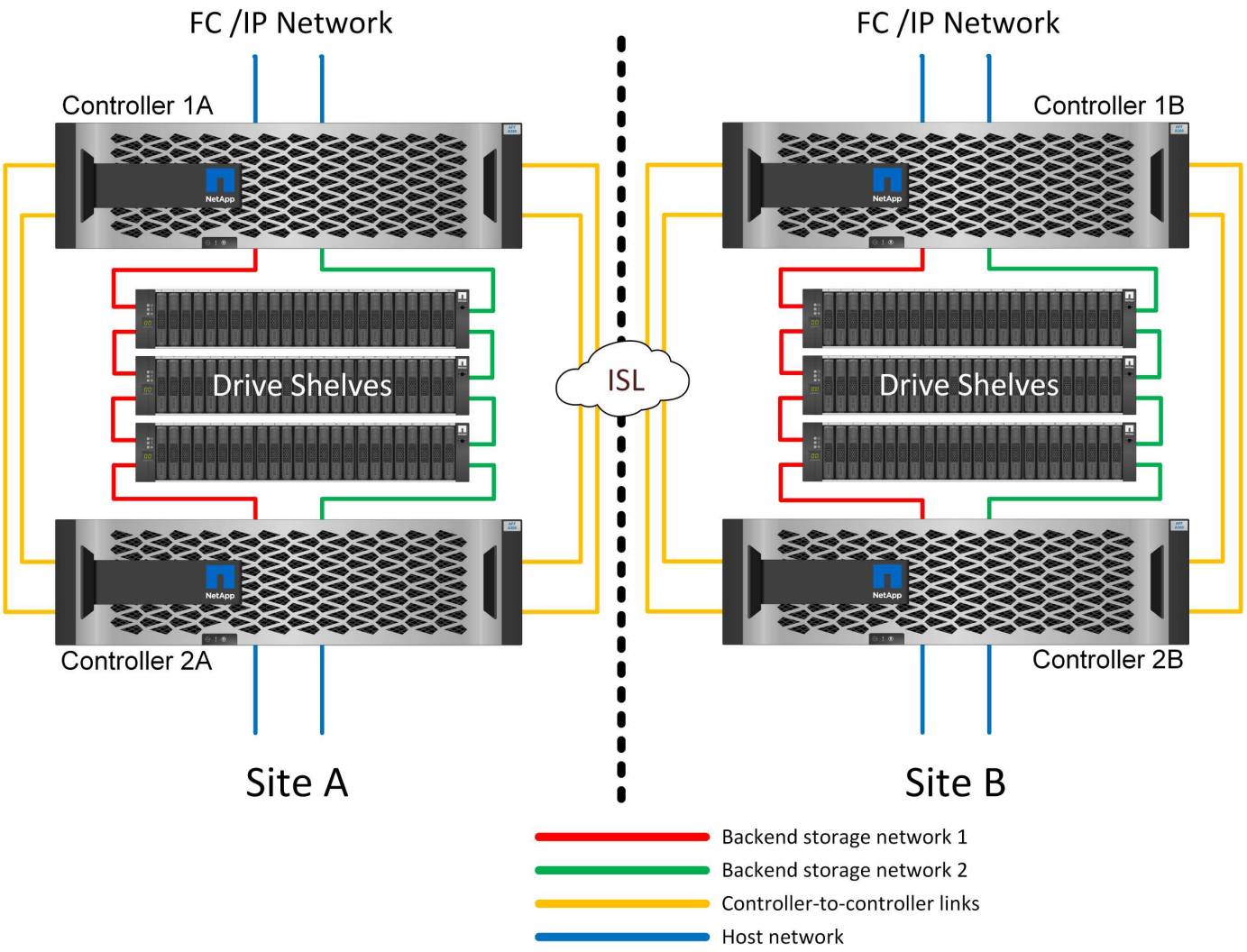
知識產權**MetroCluster**

HA 配對 MetroCluster IP 組態每個站台使用兩或四個節點。此組態選項可增加與雙節點選項相關的複雜度和成本、但它提供重要的優點：站台內備援。簡單的控制器故障不需要透過 WAN 存取資料。透過替代本機控制器、資料存取仍保持在本機狀態。

大多數客戶都選擇 IP 連線、因為基礎架構需求較為簡單。過去、高速跨站台連線通常較容易使用深色光纖和 FC 交換器進行配置、但如今、高速、低延遲的 IP 電路更容易使用。

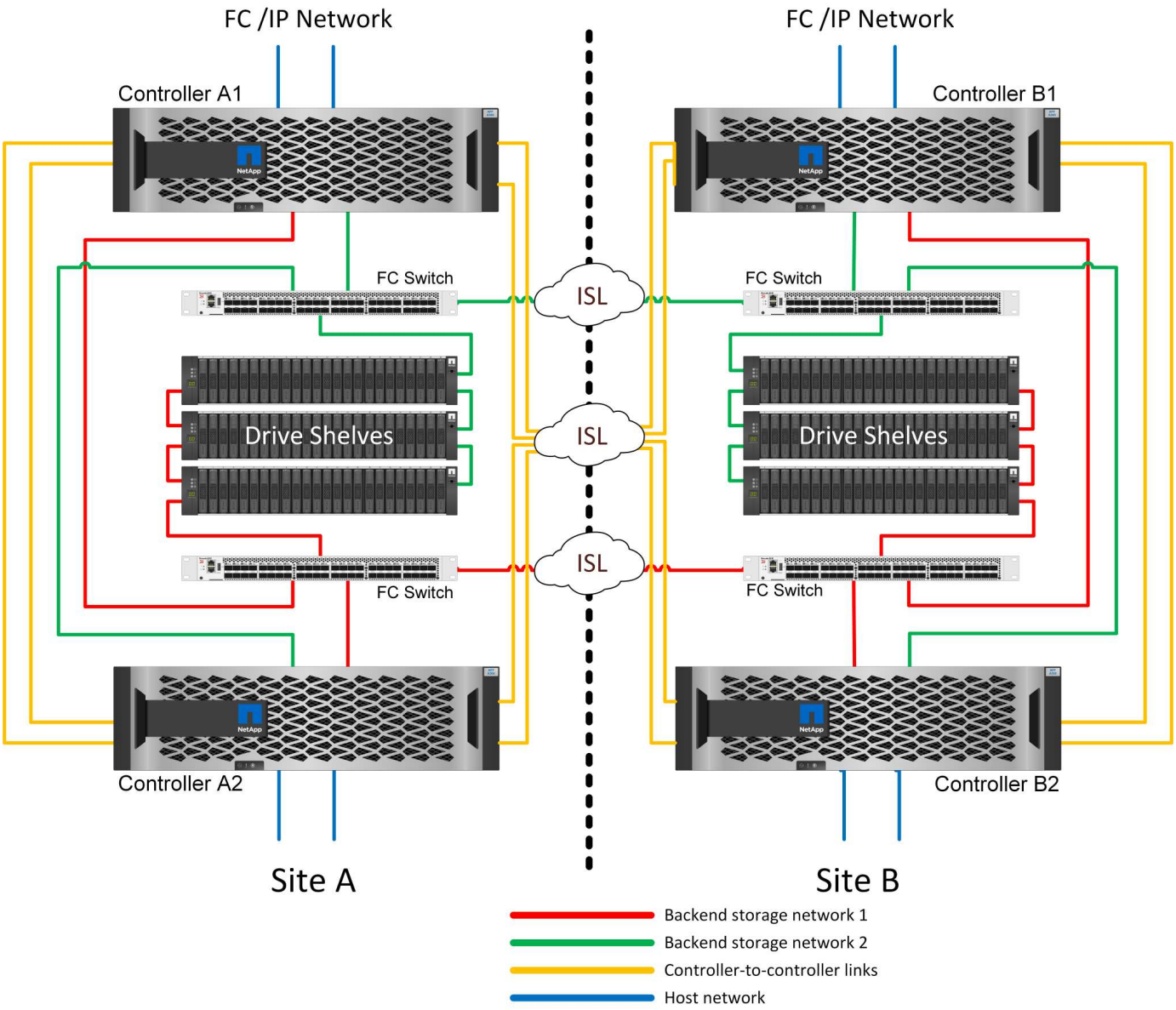
由於唯一的跨站台連線適用於控制器、因此架構也更簡單。在 FC SAN 附加 MetroCluster 中、控制器會直接寫入另一個站台上的磁碟機、因此需要額外的 SAN 連線、交換器和橋接器。相反地、 IP 組態中的控制器會透過控制器寫入相對的磁碟機。

如需其他資訊、請參閱 ONTAP 正式文件和 "SIP解決方案架構與設計MetroCluster"。



HA 配對 **FC SAN** 附加 **MetroCluster**

HA 配對 MetroCluster FC 組態每個站台使用兩個或四個節點。此組態選項可增加與雙節點選項相關的複雜度和成本、但它提供重要的優點：站台內備援。簡單的控制器故障不需要透過 WAN 存取資料。透過替代本機控制器、資料存取仍保持在本機狀態。

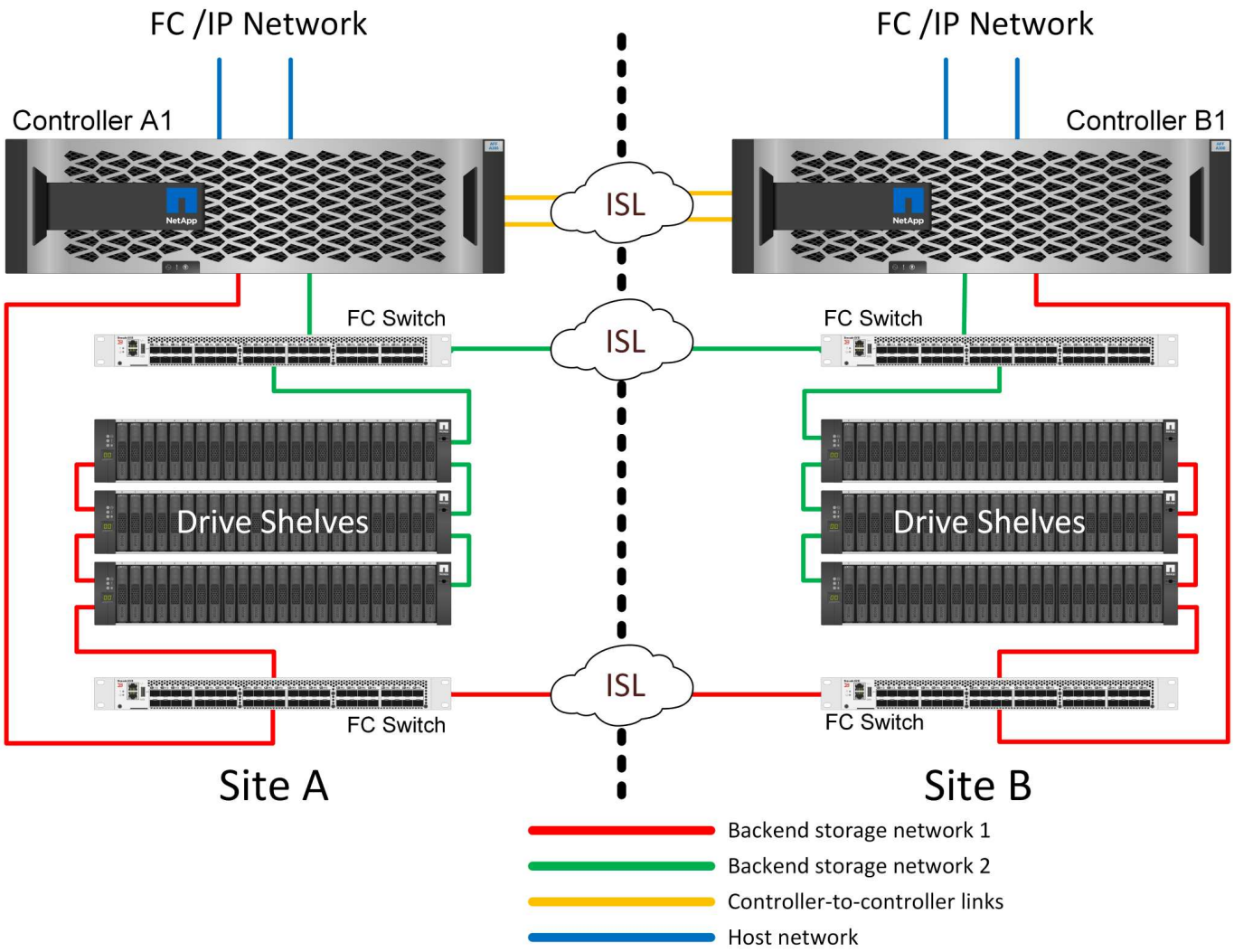


有些多站台基礎架構並非設計用於主動式作業、而是更多用於主要站台和災難恢復站台。在這種情況下、 HA 配對 MetroCluster 選項通常較為理想、原因如下：

- 雖然雙節點 MetroCluster 叢集是 HA 系統、但控制器意外故障或規劃的維護作業需要資料服務必須在相反的站台上線。如果站台之間的網路連線能力不支援所需的頻寬、效能就會受到影響。唯一的選項是將各種主機作業系統和相關服務容錯移轉至替代站台。HA 配對 MetroCluster 叢集可消除此問題、因為遺失控制器會導致同一個站台內的簡單容錯移轉。
- 有些網路拓撲並非設計用於跨站台存取、而是使用不同的子網路或隔離的 FC SAN 。在這種情況下、雙節點 MetroCluster 叢集不再作為 HA 系統運作、因為替代控制器無法將資料提供給位於相反站台的伺服器。HA 配對 MetroCluster 選項是提供完整備援的必要條件。
- 如果將雙站台基礎架構視為單一的高可用度基礎架構、則雙節點 MetroCluster 組態很適合。不過、如果系統在站台故障後必須長時間運作、則最好使用 HA 配對、因為它會繼續在單一站台內提供 HA 。

雙節點 **FC SAN** 附加 **MetroCluster**

雙節點 MetroCluster 組態每個站台僅使用一個節點。此設計比 HA 配對選項簡單、因為要設定和維護的元件較少。此外、它也降低了佈線和 FC 交換方面的基礎架構需求。最後、它能降低成本。



這項設計的明顯影響是、控制器在單一站台上故障、表示資料可從另一個站台取得。這種限制不一定是個問題。許多企業都有多站台資料中心作業、並有延伸、高速、低延遲的網路、基本上是一個基礎架構。在這些情況下、MetroCluster 的雙節點版本是慣用的組態。多家服務供應商目前以 PB 規模使用雙節點系統。

MetroCluster 恢復功能

MetroCluster 解決方案沒有單點故障：

- 每個控制器都有兩條通往本機站台磁碟櫃的路徑。
- 每個控制器都有兩條通往遠端站台磁碟機櫃的路徑。
- 每個控制器都有兩條通往另一個站台上控制器的路徑。
- 在 HA 配對組態中、每個控制器都有兩條路徑通往本機合作夥伴。

總而言之、您可以移除組態中的任何一個元件、而不會影響 MetroCluster 提供資料的能力。這兩個選項之間恢復能力的唯一差異是 HA 配對版本在站台故障後仍是整個 HA 儲存系統。

邏輯架構

瞭解 Oracle 資料庫在 MetroCluster 環境中的運作方式需要對 MetroCluster 系統的邏輯功

能進行一些說明。

站台故障保護： **NVRAM** 和 **MetroCluster**

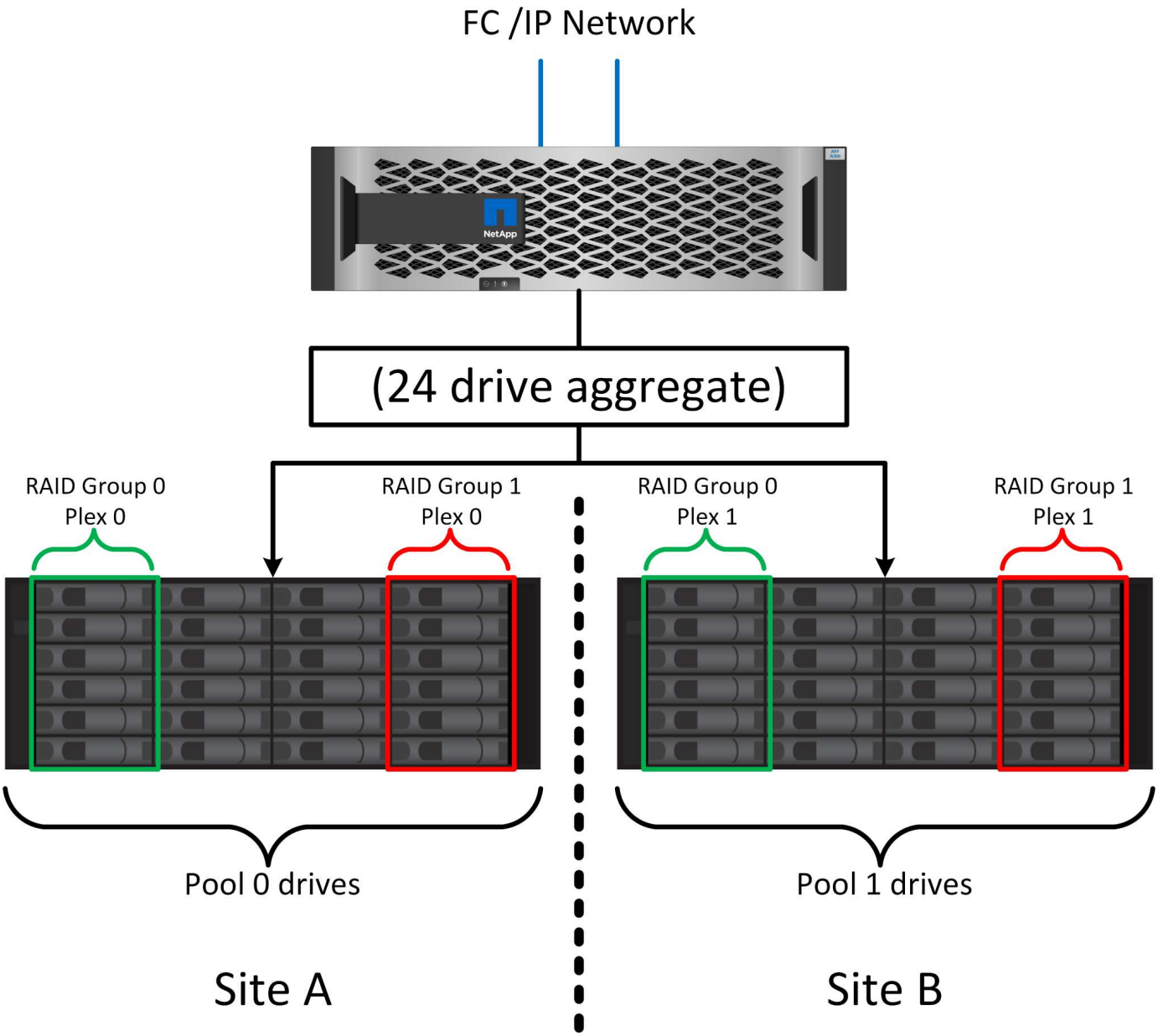
MetroCluster 以下列方式擴充 NVRAM 資料保護：

- 在雙節點組態中、 NVRAM 資料會使用交換器間連結（ ISL ）複寫到遠端合作夥伴。
- 在 HA 配對組態中、 NVRAM 資料會同時複寫到本機合作夥伴和遠端合作夥伴。
- 寫入內容必須複寫到所有合作夥伴、才能予以確認。此架構可將 NVRAM 資料複寫至遠端合作夥伴、保護機上 I/O 不受站台故障影響。此程序不涉及磁碟機層級的資料複寫。擁有該集合體的控制器負責將資料複寫至集合體中的兩個叢集、但在站台遺失時仍必須保護資料、避免在執行中遺失 I/O 。只有當合作夥伴控制器必須接管故障控制器時、才會使用複寫的 NVRAM 資料。

站台和機櫃故障保護： **SyncMirror** 和叢

SyncMirror 是一項鏡射技術、可增強但不取代 RAID DP 或 RAID-TEC 。它會鏡射兩個不同 RAID 群組的內容。邏輯組態如下：

1. 磁碟機會根據位置設定成兩個集區。一個集區由站台 A 上的所有磁碟機組成、第二個集區由站台 B 上的所有磁碟機組成
2. 接著會根據鏡射的 RAID 群組集建立通用儲存池（稱為 Aggregate ）。從每個站台擷取的磁碟機數量相等。例如、 20 個磁碟機的 SyncMirror Aggregate 將由站台 A 的 10 個磁碟機和站台 B 的 10 個磁碟機組成
3. 指定站台上的每組磁碟機都會自動設定為一個或多個完全備援的 RAID DP 或 RAID-TEC 群組、而不受鏡像的使用影響。在鏡射下使用 RAID 、即使在站台遺失之後、也能提供資料保護。



上圖說明 SyncMirror 組態範例。在控制器上建立了 24 個磁碟機的集合體、其中 12 個磁碟機來自於站台 A 上配置的機櫃、 12 個磁碟機來自站台 B 上配置的機櫃磁碟機分為兩個鏡射 RAID 群組。RAID 群組 0 包含站台 A 的 6 磁碟機叢、鏡射到站台 B 的 6 磁碟機叢同樣地、 RAID 群組 1 也包含站台 A 的 6 磁碟叢叢、鏡射到站台 B 的 6 磁碟叢叢

SyncMirror 通常用於提供 MetroCluster 系統的遠端鏡射、每個站台都有一份資料複本。有時候、它是用來在單一系統中提供額外的備援層級。特別是提供機架層級的備援。磁碟機櫃已包含雙電源供應器和控制器、整體上比金屬板稍多、但在某些情況下、可能需要額外的保護。例如、有一位 NetApp 客戶部署 SyncMirror 、用於汽車測試期間使用的行動即時分析平台。系統分為兩個實體機架、分別隨附獨立的電源饋送和獨立的 UPS 系統。

備援故障： **NVFAIL**

如前所述、寫入必須先登入本機 NVRAM 及至少一個其他控制器上的 NVRAM 、才會被確認。此方法可確保硬體故障或停電不會導致機內 I/O 遺失如果本機 NVRAM 故障或連線至其他節點失敗、則資料將不再鏡射。

如果本機 NVRAM 回報錯誤、節點會關機。當使用 HA 配對時、此關機會導致容錯移轉至合作夥伴控制器。使

用 MetroCluster 時、行為取決於所選的整體組態、但可能會導致自動容錯移轉至遠端記事。無論如何、由於發生故障的控制器尚未確認寫入作業、因此不會遺失任何資料。

站台對站台連線故障會封鎖 NVRAM 複寫至遠端節點、這種情況更為複雜。寫入不再複寫到遠端節點、因此如果控制器發生災難性錯誤、可能會導致資料遺失。更重要的是、在這些情況下、嘗試容錯移轉至其他節點會導致資料遺失。

控制因素是 NVRAM 是否同步。如果 NVRAM 已同步、則節點對節點容錯移轉可安全地繼續進行、不會有資料遺失的風險。在 MetroCluster 組態中、如果 NVRAM 和基礎 Aggregate plex 同步、則可以安全地繼續進行轉換、而不會有資料遺失的風險。

除非強制進行容錯移轉或切換、否則 ONTAP 不允許在資料不同步時進行容錯移轉或切換。以這種方式強制變更條件、即表示資料可能會留在原始控制器中、而且資料遺失是可以接受的。

如果強制進行容錯移轉或切換、資料庫和其他應用程式尤其容易毀損、因為它們會在磁碟上保留較大的內部資料快取。如果發生強制容錯移轉或切換、先前確認的變更將會有效捨棄。儲存陣列的內容會有效地及時向後跳轉、而且快取狀態不再反映磁碟上資料的狀態。

為了避免這種情況發生、 ONTAP 允許設定磁碟區、以針對 NVRAM 故障提供特殊保護。觸發時、此保護機制會導致磁碟區進入稱為 NVFAIL 的狀態。此狀態會導致 I/O 錯誤、導致應用程式當機。這項當機會導致應用程式關機、使其不使用過時的資料。資料不應遺失、因為記錄中應存在任何已認可的交易資料。通常的後續步驟是讓系統管理員在手動將 LUN 和磁碟區重新上線之前、先完全關閉主機。雖然這些步驟可能涉及一些工作、但這種方法是確保資料完整性的最安全方法。並非所有資料都需要這項保護、因此 NVFAIL 行為可依每個磁碟區設定。

HA 配對與 MetroCluster

MetroCluster 提供兩種組態：雙節點和 HA 配對。雙節點組態在 NVRAM 上的運作方式與 HA 配對相同。如果發生突然故障、合作夥伴節點可以重新執行 NVRAM 資料、以確保磁碟機一致、並確保沒有遺失任何已確認的寫入資料。

HA 配對組態也會將 NVRAM 複寫到本機合作夥伴節點。簡單的控制器故障會在合作夥伴節點上重新執行 NVRAM 、而獨立 HA 配對則不使用 MetroCluster 。萬一突然完全遺失站台、遠端站台也需要 NVRAM 、才能讓磁碟機保持一致、開始提供資料。

MetroCluster 的一個重要層面是、在正常作業條件下、遠端節點無法存取合作夥伴資料。每個站台基本上都是一個可假設對方站台特性的個別系統。此程序稱為「轉換」、包含計畫性的轉換、可在不中斷營運的情況下、將站台作業移轉至另一個站台。它也包括站台遺失的非計畫性情況、以及災難恢復需要手動或自動切換。

切換與切換

術語切換和切換是指在 MetroCluster 組態中、在遠端控制器之間轉換磁碟區的程序。此程序僅適用於遠端節點。在四個磁碟區組態中使用 MetroCluster 時、本機節點容錯移轉是先前所述的相同接管和恢復程序。

計畫性切換與切換

規劃的切換或切換類似於節點之間的接管或恢復。此程序有多個步驟、可能需要幾分鐘的時間、但實際發生的是儲存設備和網路資源的多階段順暢轉換。控制傳輸的速度比執行完整命令所需的時間快得多。

接管 / 恢復與切換 / 切換回復之間的主要差異在於對 FC SAN 連線能力的影響。使用本機接管 / 恢復功能、主機會遺失通往本機節點的所有 FC 路徑、並仰賴其原生 MPIO 來切換至可用的替代路徑。連接埠不會重新定位。透過切換和切換、控制器上的虛擬 FC 目標連接埠會轉換到另一個站台。它們在 SAN 上實際上已經停用一段時間、然後重新出現在替代控制器上。

SyncMirror 逾時

SyncMirror 是一項 ONTAP 鏡射技術、可針對機櫃故障提供保護。當機櫃之間相隔一段距離時、就能獲得遠端資料保護。

SyncMirror 無法提供通用同步鏡像。因此、可用度更高。有些儲存系統使用固定的全或全自動鏡射、有時稱為 Domino 模式。這種形式的鏡像在應用程式中受到限制、因為如果與遠端站台的連線中斷、所有寫入活動都必須停止。否則、寫字會存在於某個站台、但不會存在於另一個站台。一般而言、如果站台對站台連線中斷超過一段短時間（例如 30 秒）、這類環境就會設定為使 LUN 離線。

這種行為是小型環境子集的理想選擇。不過、大多數應用程式都需要一套解決方案、能夠在正常作業條件下提供保證同步複寫、但能夠暫停複寫。站台對站台連線能力完全中斷通常被視為近乎災難的情況。一般而言、這類環境會保持在線上狀態並提供資料、直到連線能力修復或正式決定關閉環境以保護資料為止。純粹因為遠端複寫失敗而需要自動關閉應用程式、這是不尋常的。

SyncMirror 支援同步鏡射需求、並可靈活調整逾時時間。如果與遠端控制器和 / 或叢的連線中斷、 30 秒定時器就會開始倒數。當計數器達到 0 時、會使用本機資料繼續寫入 I/O 處理。資料的遠端複本可以使用、但會在連線恢復之前、及時凍結。重新同步利用 Aggregate 層級快照、將系統儘快恢復至同步模式。

值得注意的是、在許多情況下、這種通用的「全或全無」 Domino 模式複寫功能更適合在應用程式層上實作。例如、 Oracle DataGuard 包括最大保護模式、可在任何情況下保證執行個體的長時間複寫。如果複寫連結失敗超過可設定的逾時時間、資料庫就會關閉。

使用 Fabric 附加 MetroCluster 自動進行無人值守切換

自動無人值守切換（ AUSO ）是一項 Fabric 附加 MetroCluster 功能、可提供一種跨站台 HA 的形式。如前所述、 MetroCluster 有兩種類型：每個站台上只有一個控制器、或每個站台上有一個 HA 配對。HA 選項的主要優點是、計畫性或非計畫性控制器關機仍可讓所有 I/O 成為本機。單一節點選項的優勢在於降低成本、複雜度和基礎架構。

AUSO 的主要價值在於改善 Fabric 附加 MetroCluster 系統的 HA 功能。每個站台都會監控相對站台的健全狀況、如果沒有節點仍可提供資料、 AUSO 就會導致快速的轉換。這種方法在每個站台只有一個節點的 MetroCluster 組態中特別有用、因為在可用度方面、它使組態更接近 HA 配對。

AUSO 無法在 HA 配對層級提供全方位監控。HA 配對可提供極高的可用度、因為它包含兩條備援實體纜線、可用於直接節點對節點通訊。此外、 HA 配對中的兩個節點都能存取備援迴圈上的同一組磁碟、為一個節點提供另一條路由來監控另一個節點的健全狀況。

MetroCluster 叢集存在於站台之間、節點對節點通訊和磁碟存取都仰賴站台對站台網路連線。監控叢集其餘部分的活動訊號的能力有限。AUSO 必須區分其他站台實際停機、而非因為網路問題而無法使用的情況。

因此、如果 HA 配對中的控制器偵測到因特定原因（例如系統異常）而發生的控制器故障、就會提示接管。如果連線完全中斷、也可能會提示接管、有時也稱為「失去心跳」。

只有在原始站台偵測到特定故障時、 MetroCluster 系統才能安全地執行自動切換。此外、擁有儲存系統所有權的控制器必須能夠保證磁碟和 NVRAM 資料同步。控制器無法保證進行變更的安全性、因為它與來源站台失去接觸、而該站台仍可運作。如需將交換作業自動化的其他選項、請參閱下一節中的 MetroCluster tiebreaker （ MCTB ）解決方案資訊。

MetroCluster tiebreaker 搭配網路附加 MetroCluster

此"NetApp MetroCluster tiebreaker"軟體可在第三個站台上執行，以監控 MetroCluster 環境的健全狀況，傳送通知，並在災難情況下強制切換。您可以在上找到"NetApp 支援網站"有關 tiebreaker 的完整說明，但 MetroCluster tiebreaker 的主要用途是偵測站台遺失。它還必須區分站台遺失和連線中斷。例如、不應因為斷路

器無法到達主要站台而進行切入、這就是為什麼斷路器也會監控遠端站台與主要站台聯絡的能力。

與 AUSO 的自動切換功能也相容於 MCTB 。AUSO 反應非常迅速、因為它的設計是偵測特定故障事件、然後只有在 NVRAM 和 SyncMirror 叢同步時才叫用切入。

相反地、斷路器位於遠端位置、因此必須等到定時器結束後才會宣告站台停機。tiebreaker 最終會偵測 AUSO 涵蓋的控制器故障類型、但一般而言、 AUSO 已經開始進行開關作業、而且可能會在 tiebreaker 運作之前完成開關作業。產生的第二個來自 tiebreaker 的切換命令將會遭到拒絕。

當強制切入時， MCTB 軟體無法驗證 NVRAM 是否與 / 或叢同步。如果已設定自動切換、則應在維護活動期間停用、導致 NVRAM 或 SyncMirror 叢同步中斷。

此外、 MCTB 可能無法因應導致下列事件順序的滾動災難：

1. 站台之間的連線中斷超過 30 秒。
2. SyncMirror 複寫逾時、且作業會繼續在主要站台上執行、使遠端複本過時。
3. 主站台會遺失。結果是主站台上存在未複寫的變更。因此、由於下列幾個原因、可能不希望進行任何一次的重新操作：
   - 關鍵資料可能會出現在主要站台上、而且該資料最終可能會恢復。允許應用程式繼續作業的轉換作業、將會有效捨棄該關鍵資料。
   - 當站台遺失時、使用主要站台上儲存資源的仍在運作中站台上的應用程式可能已快取資料。切入會導致資料的過時版本與快取不相符。
   - 當發生站台遺失時、使用主要站台上儲存資源的仍在運作中站台上的作業系統、可能已快取資料。切入會導致資料的過時版本與快取不相符。最安全的選項是將斷路器設定為在偵測到站台故障時傳送警示、然後讓人員決定是否強制進行轉換。應用程式和（或）作業系統可能需要先關機、才能清除任何快取資料。此外、 NVFAIL 設定也可用於新增進一步的保護、並協助簡化容錯移轉程序。

ONTAP Mediator 搭配 MetroCluster IP

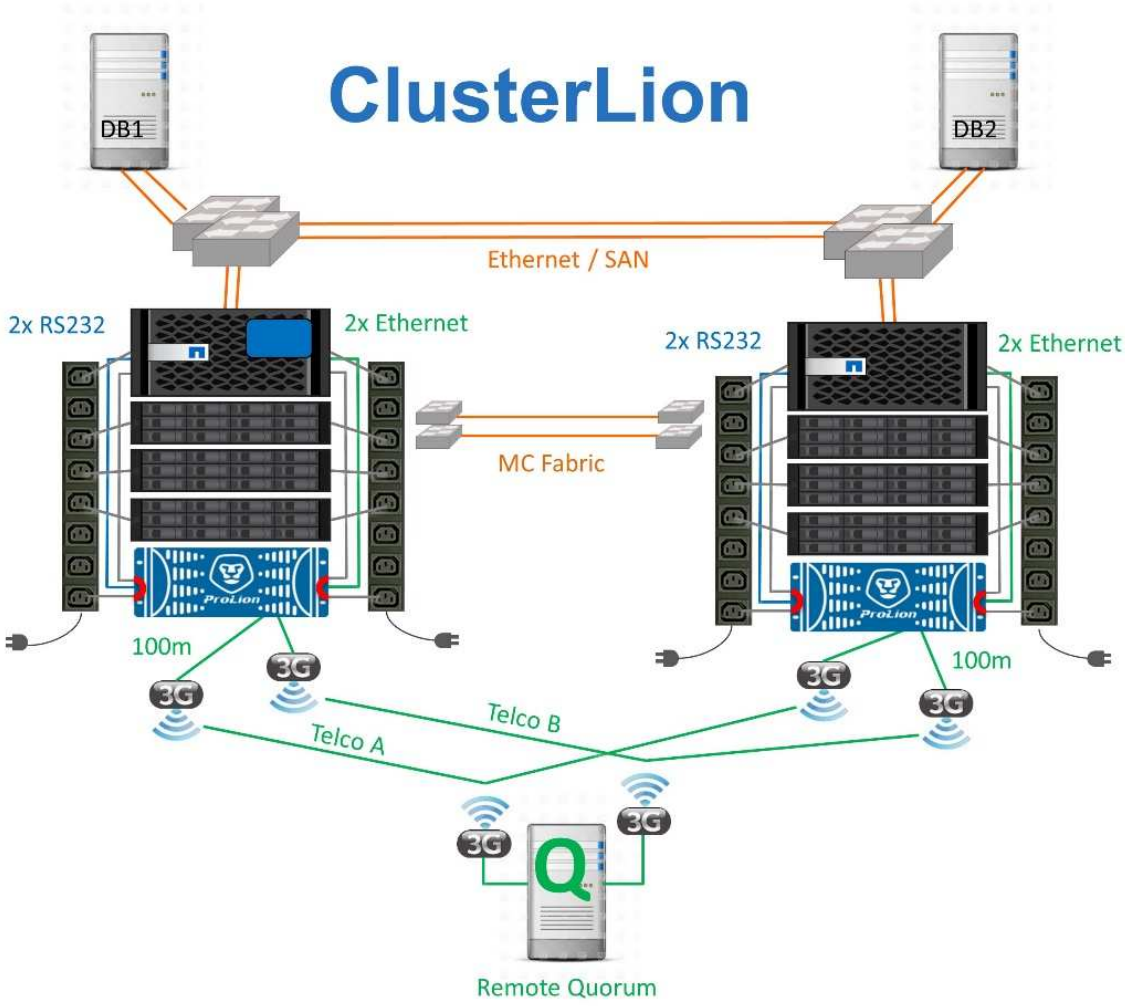
ONTAP Mediator 可搭配 MetroCluster IP 和某些其他 ONTAP 解決方案使用。它是一項傳統的斷路器服務，就像上述的 MetroCluster tiebreaker 軟體一樣，但也包含一項重要功能，即執行自動無人值守的移除。

光纖連接的 MetroCluster 可直接存取位於相對站台的儲存裝置。這可讓一個 MetroCluster 控制器從磁碟機讀取心跳資料、以監控其他控制器的健全狀況。這可讓一個控制器辨識另一個控制器的故障、並執行切換。

相反地、 MetroCluster IP 架構只會透過控制器控制器連線路由所有 I/O 、而無法直接存取遠端站台上的儲存裝置。這會限制控制器偵測故障和執行轉換的能力。因此、 ONTAP Mediator 必須作為斷路器裝置、才能偵測站台遺失並自動執行轉換。

使用 ClusterLion 的虛擬第三站點

ClusterLion 是一款先進的 MetroCluster 監控設備、可作為虛擬第三站點使用。此方法可讓 MetroCluster 安全部署在雙站台組態中、並具備全自動的轉換功能。此外、 ClusterLion 還能執行額外的網路層級監控、並執行後置作業。完整文件可從 ProLion 取得。



- ClusterLion 設備會使用直接連接的乙太網路和序列纜線來監控控制器的健全狀況。
- 這兩台設備透過備援的 3G 無線連線彼此連線。
- ONTAP 控制器的電源會透過內部中繼路由傳送。發生站台故障時、包含內部 UPS 系統的 ClusterLion 會先切斷電源連線、然後再啟動切入。此程序可確保不會發生任何大腦分割狀況。
- ClusterLion 會在 30 秒 SyncMirror 逾時內執行切換、或完全不執行。
- 除非 NVRAM 和 SyncMirror 叢集的狀態同步、否則 ClusterLion 不會執行切入。
- 由於 ClusterLion 只會在 MetroCluster 完全同步時執行切入、因此不需要 NVFAIL 。此組態可讓擴充 Oracle RAC 等站台跨距環境保持連線、即使在非計畫性的轉換期間亦然。
- 支援包括光纖連接的 MetroCluster 和 MetroCluster IP

SyncMirror

SyncMirror 是 MetroCluster 系統的 Oracle 資料保護基礎、是最大效能的橫向擴充同步鏡射技術。

使用 **SyncMirror** 保護資料

在最簡單的層級上、同步複寫表示必須先對鏡射儲存設備的兩側進行任何變更、然後才會被確認。例如、如果資料庫正在寫入記錄檔、或是正在修補 VMware 來賓作業系統、則寫入作業絕不能遺失。作爲一種協議級別，在

兩個站點上的非易失性介質被認可之前，存儲系統不得確認寫入內容。只有這樣、在不遺失資料的風險下繼續作業是安全的。

使用同步複寫技術是設計和管理同步複寫解決方案的第一步。最重要的考量是瞭解在各種計畫性和非計畫性失敗案例中可能發生的情況。並非所有同步複寫解決方案都提供相同的功能。如果您需要提供零恢復點目標（ RPO ）的解決方案、亦即零資料遺失、則必須考慮所有故障情況。特別是、當站台之間的連線中斷而無法進行複寫時、預期會產生什麼結果？

SyncMirror 資料可用度

MetroCluster 複寫是以 NetApp SyncMirror 技術為基礎、其設計旨在有效率地切換至同步模式及從同步模式切換到同步模式。這項功能符合要求同步複寫、但也需要高可用度資料服務的客戶需求。例如、如果中斷與遠端站台的連線、通常最好讓儲存系統繼續以非複寫狀態運作。

許多同步複寫解決方案只能以同步模式運作。這種類型的全或全無複寫有時稱為 Domino 模式。這類儲存系統會停止提供資料、而不允許資料的本機和遠端複本進行非同步處理。如果複寫被強制中斷、重新同步可能會非常耗時、而且可能會讓客戶在重新建立鏡像期間暴露在完全資料遺失的風險中。

SyncMirror 不僅可以在無法連線到遠端站台時、無縫切換至同步模式、也可以在連線恢復時、快速重新同步至 RPO = 0 狀態。遠端站台的資料過時複本也可在重新同步期間保留為可用狀態、以確保資料的本機和遠端複本隨時都存在。

在需要 Domino 模式的情況下、 NetApp 提供 SnapMirror 同步（ SM-S ）。應用程式層級選項也存在、例如 Oracle DataGuard 或 SQL Server Always On Availability Groups 。作業系統層級的磁碟鏡射可以是一個選項。如需其他資訊和選項、請洽詢您的 NetApp 或合作夥伴客戶團隊。

MetroCluster 和 NVFAIL

NVFAIL 是 ONTAP 中的一般資料完整性功能、其設計可讓資料庫發揮最大的資料完整性保護。

本節將進一步說明基本的 ONTAP NVFAIL 、以涵蓋 MetroCluster 專屬主題。

使用 MetroCluster 時、寫入必須登入至少一個其他控制器的本機 NVRAM 和 NVRAM 、才能被確認。此方法可確保硬體故障或停電不會導致機內 I/O 遺失如果本機 NVRAM 故障或連線至其他節點失敗、則資料將不再鏡射。

如果本機 NVRAM 回報錯誤、節點會關機。當使用 HA 配對時、此關機會導致容錯移轉至合作夥伴控制器。使用 MetroCluster 時、行為取決於所選的整體組態、但可能會導致自動容錯移轉至遠端記事。無論如何、由於發生故障的控制器尚未確認寫入作業、因此不會遺失任何資料。

站台對站台連線故障會封鎖 NVRAM 複寫至遠端節點、這種情況更為複雜。寫入不再複寫到遠端節點、因此如果控制器發生災難性錯誤、可能會導致資料遺失。更重要的是、在這些情況下、嘗試容錯移轉至其他節點會導致資料遺失。

控制因素是 NVRAM 是否同步。如果 NVRAM 已同步、則節點對節點容錯移轉可安全地繼續進行、而不會有資料遺失的風險。在 MetroCluster 組態中、如果 NVRAM 和基礎 Aggregate plex 同步、則在不遺失資料的情況下繼續進行轉換是安全的。

除非強制進行容錯移轉或切換、否則 ONTAP 不允許在資料不同步時進行容錯移轉或切換。以這種方式強制變更條件、即表示資料可能會留在原始控制器中、而且資料遺失是可以接受的。

如果強制進行容錯移轉或切換、則資料庫特別容易遭到毀損、因為資料庫會在磁碟上保留較大的內部資料快取。

如果發生強制容錯移轉或切換、先前確認的變更將會有效捨棄。儲存陣列的內容會有效地及時向後跳轉、而且資料庫快取的狀態不再反映磁碟上資料的狀態。

為了保護應用程式不受這種情況影響、 ONTAP 允許設定磁碟區、以針對 NVRAM 故障提供特殊保護。觸發時、此保護機制會導致磁碟區進入稱為 NVFAIL 的狀態。此狀態會導致 I/O 錯誤、導致應用程式關機、使其不使用過時的資料。資料不應遺失、因為儲存系統上仍有任何已確認的寫入資料、而資料庫則應在記錄中顯示任何已認可的交易資料。

通常的後續步驟是讓系統管理員在手動將 LUN 和磁碟區重新上線之前、先完全關閉主機。雖然這些步驟可能涉及一些工作、但這種方法是確保資料完整性的最安全方法。並非所有資料都需要這項保護、因此 NVFAIL 行為可依每個磁碟區設定。

手動強制 **NVFAIL**

最安全的選項是透過指定來強制轉換跨站台散佈的應用程式叢集（包括 VMware 、 Oracle RAC 及其他）`-force-nvfail-all` 在命令列。此選項可作為緊急措施使用、以確保所有快取資料均已清除。如果主機使用的儲存資源原本位於災難性站台上、則會收到 I/O 錯誤或過時的檔案處理 (`ESTALE`）錯誤。Oracle 資料庫當機、檔案系統可能完全離線、或切換至唯讀模式。

在完成重新操作之後、 `in-nvfailed-state` 需要清除旗標、且 LUN 必須置於線上。完成此活動後、即可重新啟動資料庫。這些工作可以自動化、以降低 RTO 。

dr-force-nvfail

作為一般安全措施、請設定 `dr-force-nvfail` 在所有可能在正常作業期間從遠端站台存取的磁碟區上加上旗標、表示這些磁碟區是在容錯移轉之前使用的活動。此設定的結果是、選取的遠端磁碟區在進入時無法使用 `in-nvfailed-state` 在進行重新操作時。在完成重新操作之後、 `in-nvfailed-state` 旗標必須清除、且 LUN 必須置於線上。這些活動完成後、即可重新啟動應用程式。這些工作可以自動化、以降低 RTO 。

結果就像使用 `-force-nvfail-all` 手動切換的旗標。然而、受影響的磁碟區數量可能僅限於必須受到保護的磁碟區、不受具有過時快取的應用程式或作業系統的影響。

對於不使用的環境、有兩項關鍵需求 `dr-force-nvfail` 在應用程式磁碟區上：

- 在主站台遺失後、強制進行的重新操作不得超過 30 秒。
- 在維護工作期間、或是在 SyncMirror 叢或 NVRAM 複寫不同步的任何其他情況下、切勿進行切入。第一項需求可以透過使用已設定為在站台故障 30 秒內執行轉換的斷路器軟體來達成。這並不表示切入作業必須在偵測站台故障的 30 秒內執行。這表示、如果站台確認運作已過 30 秒、就不再安全地強制進行轉換。

第二項需求可在已知 MetroCluster 組態不同步時停用所有自動切換功能、以部分滿足。更好的選擇是擁有可監控 NVRAM 複寫和 SyncMirror 叢的健全狀況的斷路器解決方案。如果叢集未完全同步、則斷路器不應觸發切入。

NetApp MCTB 軟體無法監控同步處理狀態、因此當 MetroCluster 因任何原因而未同步時、應該停用同步處理狀態。ClusterLion 確實包含 NVRAM 監控和叢監視功能、除非 MetroCluster 系統確認完全同步、否則可將其設定為不觸發切入。

Oracle 單一執行個體

如前所述、 MetroCluster 系統的存在並不一定會新增或變更任何操作資料庫的最佳實務做法。目前在客戶 MetroCluster 系統上執行的大多數資料庫都是單一執行個體、並遵循

Oracle on ONTAP 文件中的建議。

使用預先設定的作業系統進行容錯移轉

SyncMirror 在災難恢復站點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。例如 Veritas Cluster Server （ VCS ）等叢集件產品通常用於在站台之間建立叢集、而且在許多情況下、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）會設定為在替代站台上線資料庫。其中一個選項是建立預先設定為 NFS 或 SAN 資源的備用伺服器、以供組成資料庫。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. 強制 MetroCluster 進行重新操作
2. 執行 FC LUN 探索（僅限 SAN ）
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。檔案系統掛載只不過是 `mount` 只需一個命令、即可在 CLI 上啟動和停止資料庫和 ASM 。如果在切換之前、磁碟區和檔案系統並未在災難恢復站台上使用、則無需設定 `dr-force- nvfail` 在磁碟區上。

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. 強制 MetroCluster 進行重新操作
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 叢集可能跨越站台。在發生災難時、虛擬機器可在移至災難恢復站台後上線。只要主控虛擬化資料庫伺服器的資料存放區在災難發生時並未使用、就不需要設定 `dr-force- nvfail` 在相關的磁碟區上。

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。此外、從兩個站台存取資料、這表示強制轉換可能會導致使用過時的資料複本。

雖然兩個站台上都有資料複本、但只有目前擁有 Aggregate 的控制器才能提供資料。因此、使用擴充的 RAC 叢集時、遠端節點必須透過站台對站台連線來執行 I/O 。結果會增加 I/O 延遲、但這種延遲通常不是問題。RAC 互連網路也必須延伸至站台、這表示無論如何都需要高速、低延遲的網路。如果增加的延遲確實造成問題、則叢集可以主動被動方式運作。接著、需要將 I/O 密集作業導向至擁有該集合體的控制器本機的 RAC 節點。然後、遠端節點會執行較輕的 I/O 作業、或純粹作為暖待機伺服器使用。

如果需要雙主動式擴充 RAC 、則應考慮使用 SnapMirror 主動式同步來取代 MetroCluster 。SM-AS 複寫可讓您偏好資料的特定複本。因此、可以內建擴充 RAC 叢集、讓所有讀取作業都在本機進行。讀取 I/O 永遠不會跨越網站、因此可提供最低的延遲。所有寫入活動仍必須傳輸站台間連線、但任何同步鏡射解決方案都無法避免此類流量。

如果開機 LUN （包括虛擬化開機磁碟）與 Oracle RAC 搭配使用、則 `misscount`可能需要變更參數。如需 RAC 逾時參數的詳細資訊"Oracle RAC 搭配 ONTAP"、請參閱。

雙站台組態

雙站台擴充 RAC 組態可提供雙主動式資料庫服務、可在不中斷營運的情況下、在許多（但並非全部）災難案例中順利運作。

RAC 投票檔案

在 MetroCluster 上部署擴充 RAC 時、首先應考慮仲裁管理。Oracle RAC 有兩種機制可管理仲裁：磁碟心跳和網路心跳。磁碟心跳會使用投票檔案來監控儲存設備存取。只要基礎儲存系統提供 HA 功能、單一投票資源就足以搭配單一站台 RAC 組態。

在早期版本的 Oracle 中、投票檔案會放置在實體儲存裝置上、但在目前版本的 Oracle 中、投票檔案會儲存在 ASM 磁碟群組中。

NFS 支援 Oracle RAC 。在網格安裝程序期間、會建立一組 ASM 程序、將用於網格檔案的 NFS 位置顯示為 ASM 磁碟群組。此程序對終端使用者來說幾乎透明、安裝完成後不需要持續進行 ASM 管理。

雙站台組態的第一項需求是確保每個站台都能以保證不中斷災難恢復程序的方式存取超過半數的投票檔案。這項工作在投票檔案儲存在 ASM 磁碟群組之前很簡單、但現在管理員必須瞭解 ASM 備援的基本原則。

ASM 磁碟群組有三種備援選項 `external`、 `normal`和 `high`。換句話說、非鏡射、鏡射和 3 向鏡射。名為的較新選項 `Flex` 也可以使用、但很少使用。備援裝置的備援層級和放置位置可控制故障情況發生的情況。例如：

- 將投票檔案放在上 `diskgroup` 與 `external` 如果站台間連線中斷、備援資源保證可收回一個站台。
- 將投票檔案放在上 `diskgroup` 與 `normal` 如果站台間連線中斷、每個站台只有一個 ASM 磁碟的備援功能可確保兩個站台的節點遷離、因為兩個站台都不會有大部分的仲裁。
- 將投票檔案放在上 `diskgroup` 與 `high` 當兩個站台都可以運作且彼此可連線時、一個站台上有兩個磁碟和另一個站台上的單一磁碟的備援功能可讓雙主動式作業運作。但是、如果單一磁碟站台與網路隔離、則該站台會被逐出。

RAC 網路心跳

Oracle RAC 網路活動訊號可監控叢集互連中的節點可連性。若要保留在叢集中、節點必須能夠連絡其他節點的一半以上。在雙站台架構中、此需求會為 RAC 節點數建立下列選項：

- 如果每個站台放置相同數量的節點、則會在網路連線中斷時、在某個站台上造成遷離。
- 在另一個站台上放置 N 個節點、在另一個站台上放置 N+1 個節點、可確保站台之間的連線中斷、導致站台的網路仲裁中剩餘節點數量較多、而節點移出數量較少的站台。

在 Oracle 12cR2 之前、無法控制哪一方在站台遺失時會發生遷離。當每個站台的節點數量相等時、會由主要節點控制遷離、這通常是第一個要開機的 RAC 節點。

Oracle 12cR2 引進節點加權功能。這項功能可讓管理員更有效地控制 Oracle 如何解決大腦分裂狀況。例如、下列命令可設定 RAC 中特定節點的偏好設定：

```
[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.
```

重新啟動 Oracle 高可用度服務後、組態如下所示：

```
[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
```

節點 `host-a` 現已指定為關鍵伺服器。如果兩個 RAC 節點是隔離的、 `host-a` 生存、和 `host-b` 被逐出。

如需完整詳細資料、請參閱 Oracle 白皮書《 Oracle Clusterware 12c Release 2 Technical Overview 》。」

對於 12cR2 之前的 Oracle RAC 版本、可透過檢查 CRS 記錄來識別主節點、如下所示：

```
[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
```

此記錄表示主節點為 `2` 和節點 `host-a` ID 為 `1`。這意味著 `host-a` 不是主節點。您可以使用命令確認主節點的身分識別 `olsnodes -n`。

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

識別碼為的節點 2 是 host-b，這是主節點。在每個站台上節點數量相等的組態中、站台為 host-b 如果這兩組因為任何原因而失去網路連線、則該站台仍可生存。

識別主節點的記錄項目可能會超出系統的使用期限。在這種情況下、可以使用 Oracle 叢集登錄（ OCR ）備份的時間戳記。

```
[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
```

此範例顯示主節點是 host-b。它也表示主節點的變更來源 host-a 至 host-b 5 月 4 日下午 2 ： 05 至 21 ： 39 之間。這種識別主節點的方法只有在也檢查了 CRS 記錄檔時才安全使用、因為主節點可能自上一次的 OCR 備份後變更。如果發生此變更、則應可在 OCR 記錄中看到。

大多數客戶選擇單一投票磁碟群組來服務整個環境、以及每個站台上相同數量的 RAC 節點。磁碟群組應放置在包含資料庫的網站上。結果是連線中斷會導致遠端站台被逐出。遠端站台將不再擁有仲裁、也無法存取資料庫檔案、但本機站台會繼續如常運作。連線恢復後、遠端執行個體即可重新上線。

發生災難時、需要進行轉換、才能讓資料庫檔案和投票磁碟群組在正常運作的網站上線。如果災難允許 AUSO 觸發切換、則不會觸發 NVFAIL 、因為已知叢集處於同步狀態、且儲存資源正常上線。AUSO 是一項非常快速的作業、應在完成之前完成 disktimeout 期間過期。

由於只有兩個站台、因此無法使用任何類型的自動外部中斷軟體、這表示強制切換必須是手動操作。

三站台組態

擴充的 RAC 叢集可更輕鬆地建構三個站台。裝載 MetroCluster 系統每一半的兩個站台也支援資料庫工作負載、而第三個站台則是資料庫和 MetroCluster 系統的斷路器。Oracle tiebreaker 組態可能只需在第三站台上放置用於投票的 ASM 磁碟群組成員、也可能在第三站台上加入作業執行個體、以確保 RAC 叢集中有奇數個節點。

有關在擴展 RAC 配置中使用 NFS 的重要信息，請參閱 Oracle 文檔中的“ quorum failure group （仲裁故障組）”。總而言之、 NFS 掛載選項可能需要修改以包含軟選項、以確保主仲裁資源所在的第三站台連線中斷、不會使主 Oracle 伺服器或 Oracle RAC 程序掛起。

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。

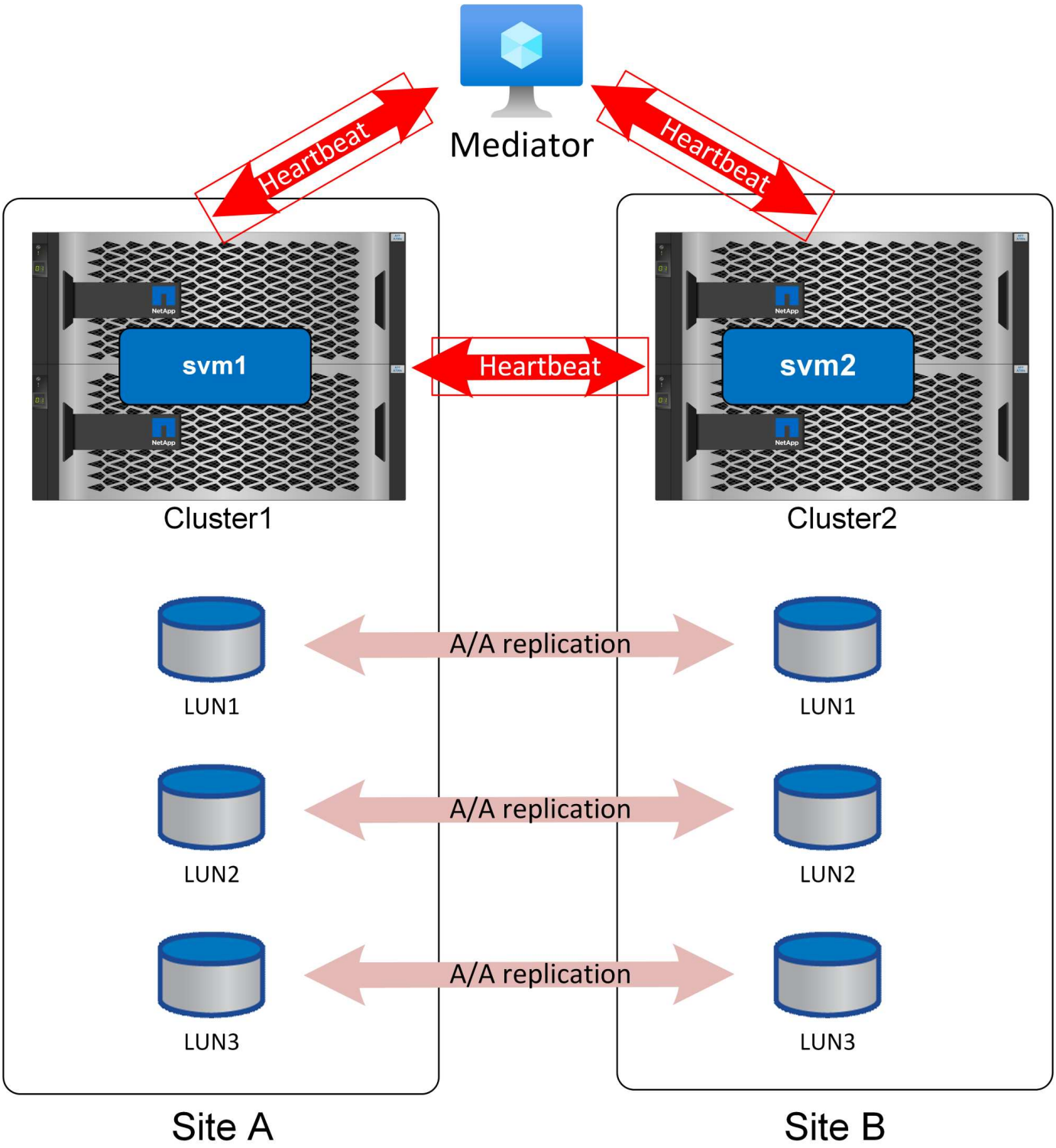
中間器**ONTAP**

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。

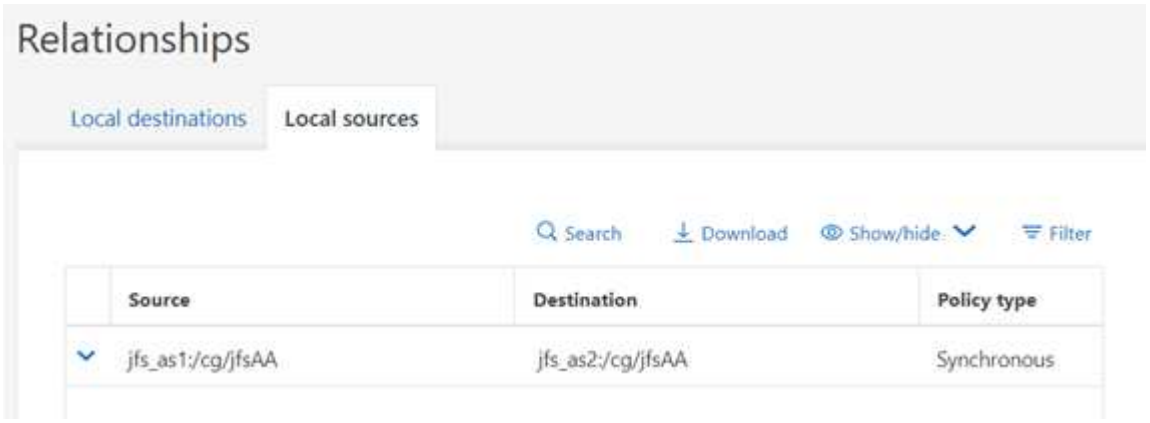
SnapMirror 主動式同步偏好的站台

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* 。

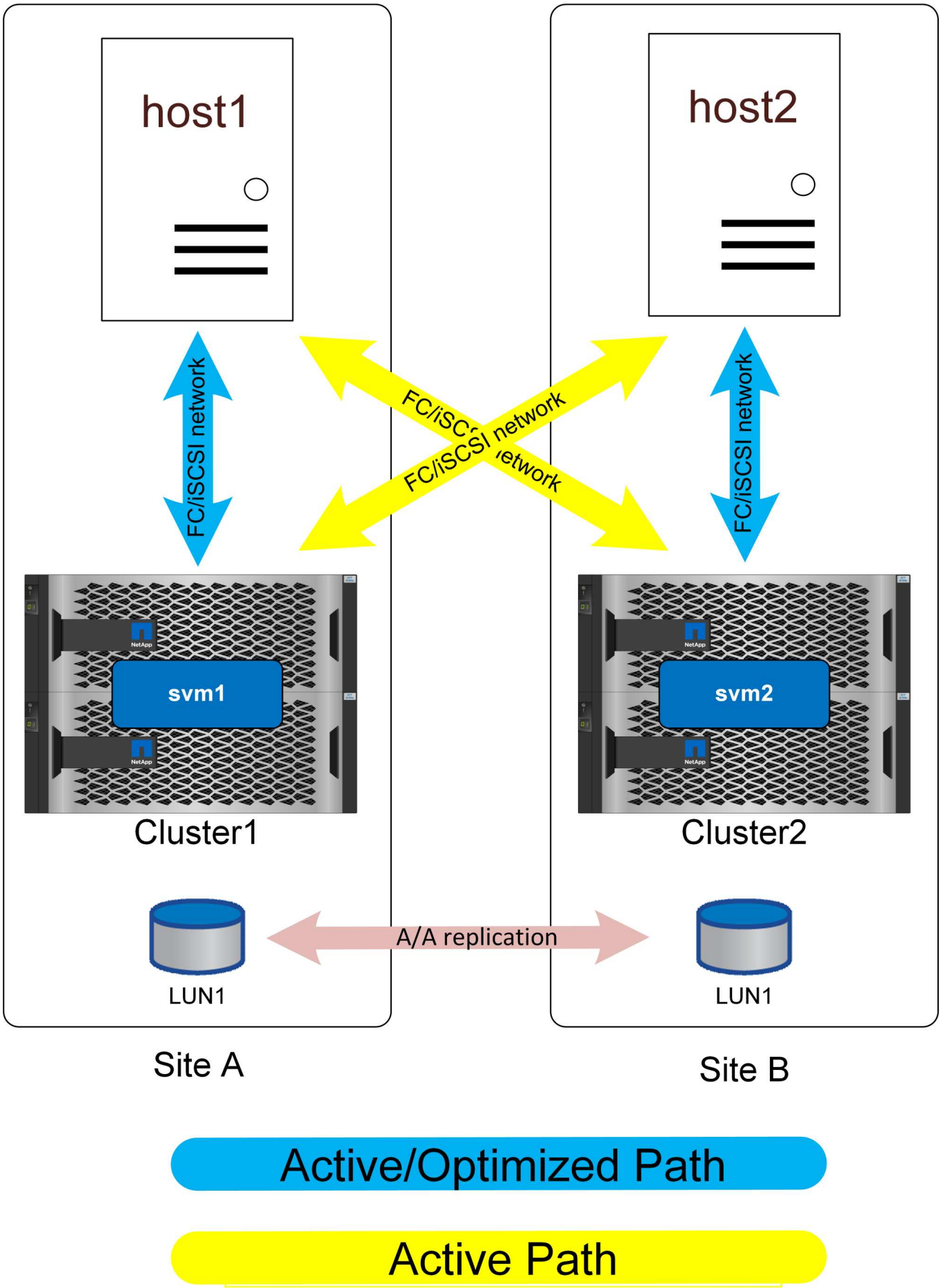
第二個選用步驟是設定主機鄰近度。這無法控制存取、而是控制 *priority* 。

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

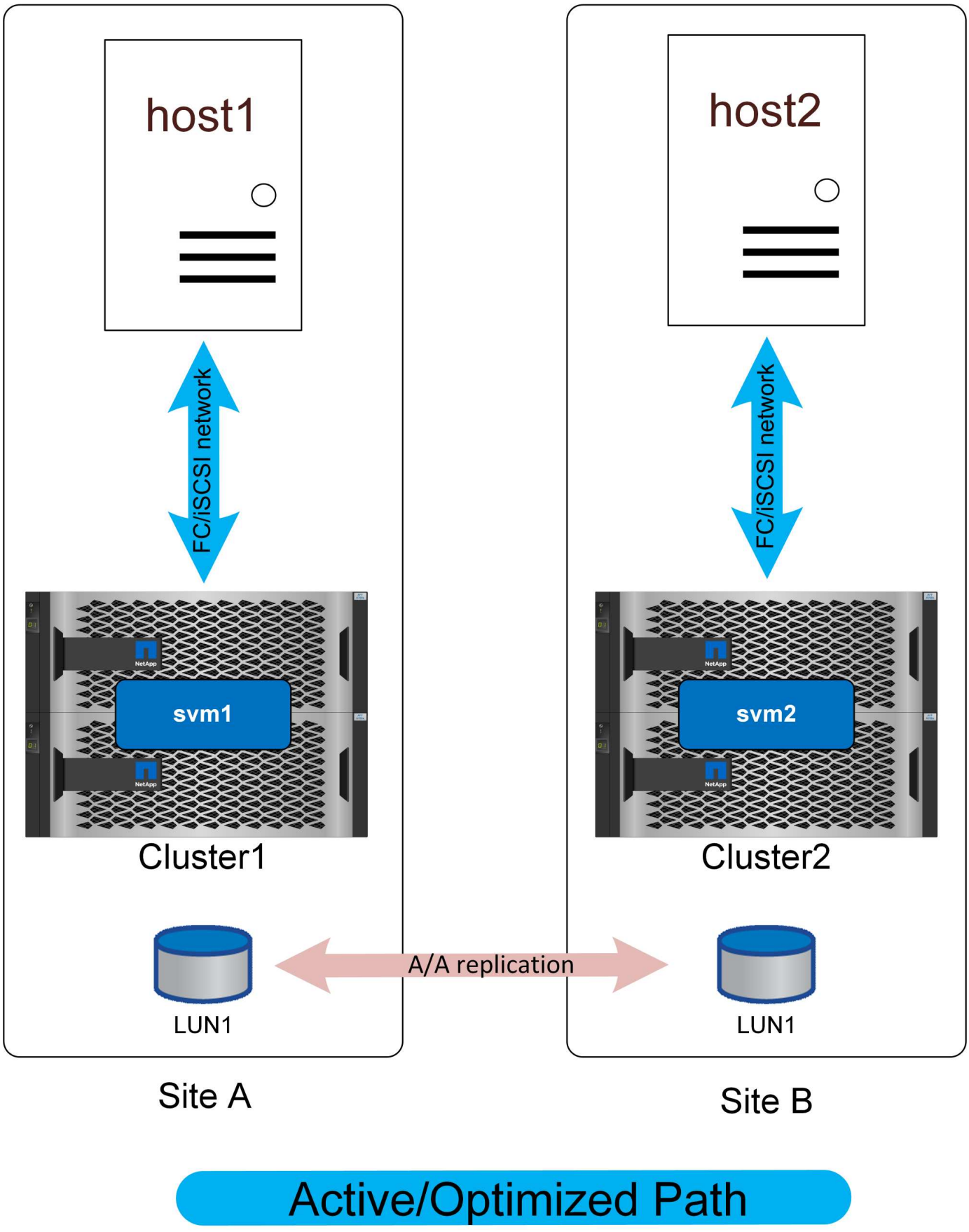
在正常作業中、所有 IO 都是本機 IO 。從本機儲存陣列提供讀取和寫入服務。寫入 IO 當然也需要由本機控制器複寫到遠端系統、然後才會被確認、但所有讀取 IO 都會在本機上提供服務、而且不會因在站台之間瀏覽 SAN 連結而產生額外延遲。

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。



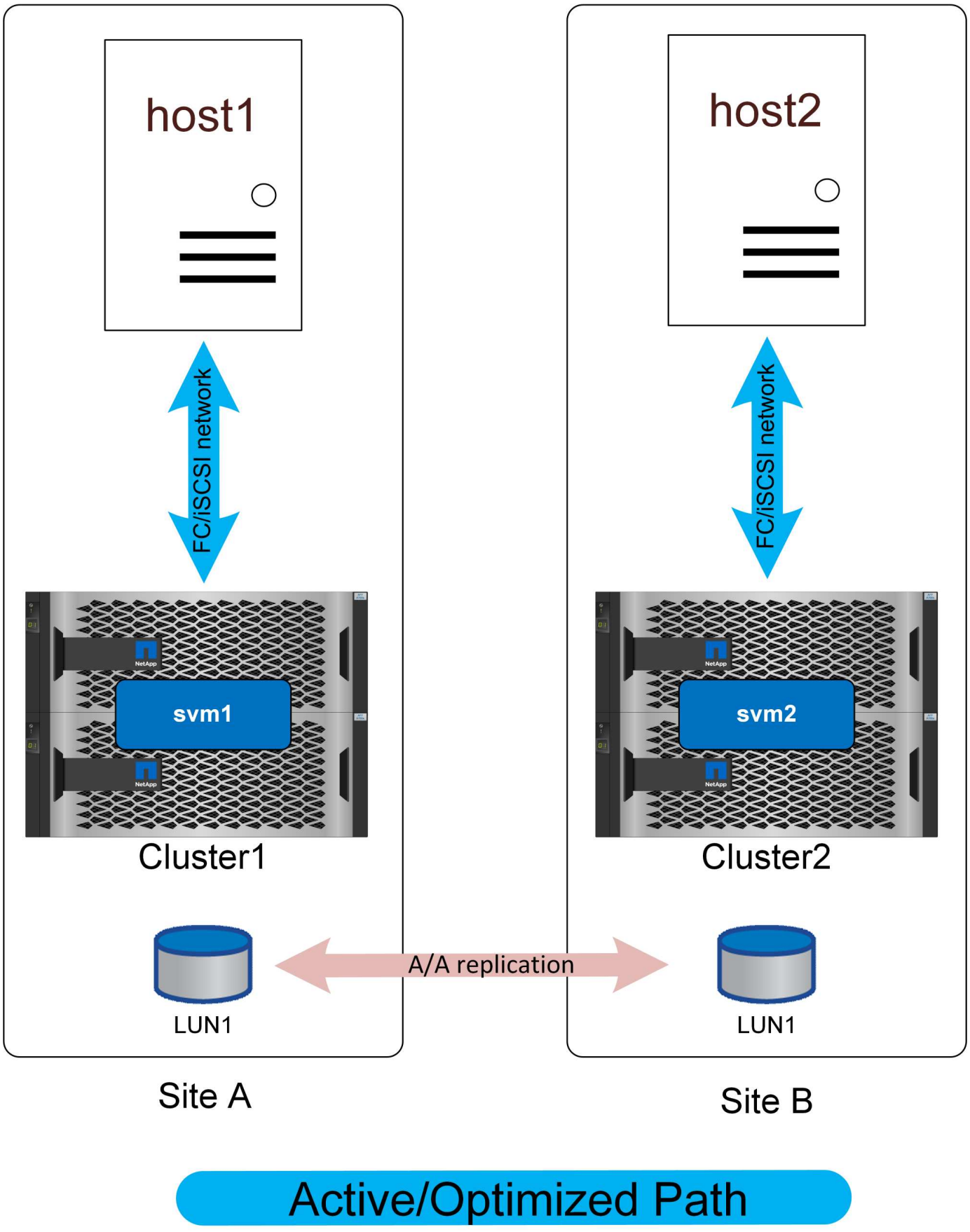
具有非統一存取權的 ASA 組態、其運作方式與 AFF 大致相同。透過統一存取、 IO 就能跨越 WAN 。這可能是或不理想的做法。

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。

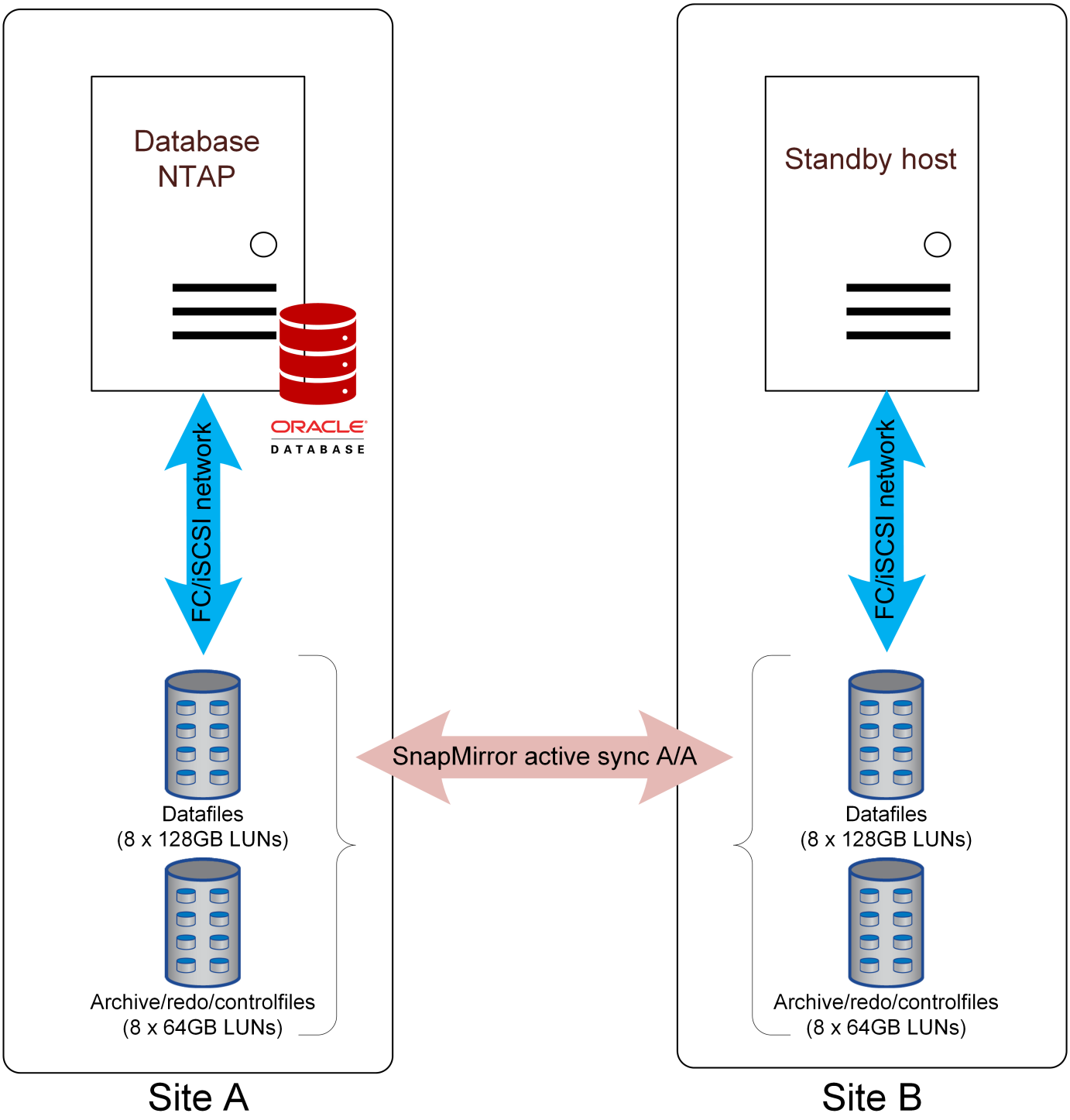
最佳架構取決於業務需求。例如、如果目標是在資料遺失的情況下提供 RPO =0 保護、但 RTO 是放寬的、則使用 Oracle 單一執行個體資料庫並以 SM 方式複寫 LUN 、可能就足以滿足 Oracle 授權標準的要求、而且成本也較低。遠端站台的故障不會中斷作業、而主站台的遺失會導致仍在運作中的站台產生 LUN 、而這些 LUN 已在線上且可供使用。

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。

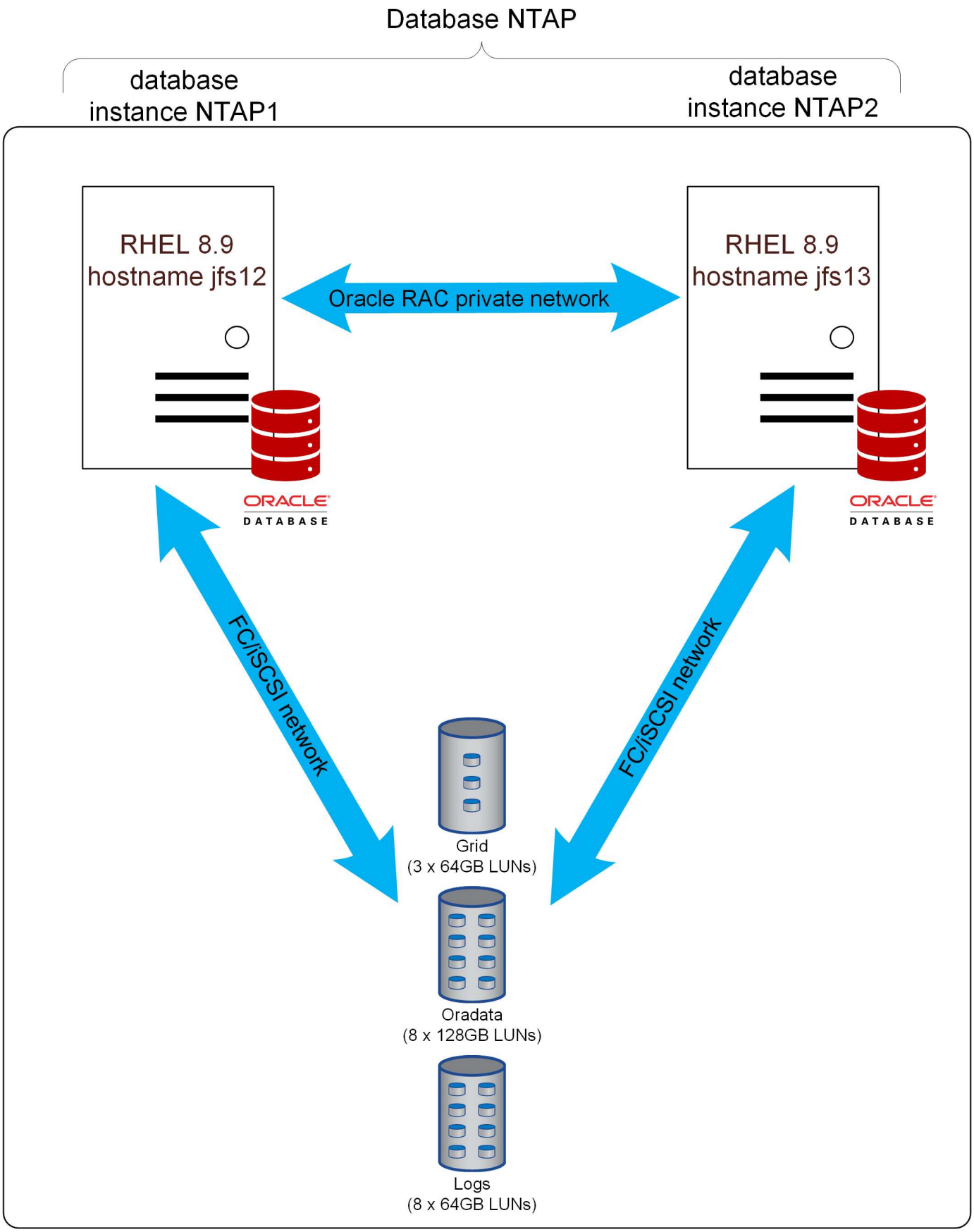
Oracle Extended RAC

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

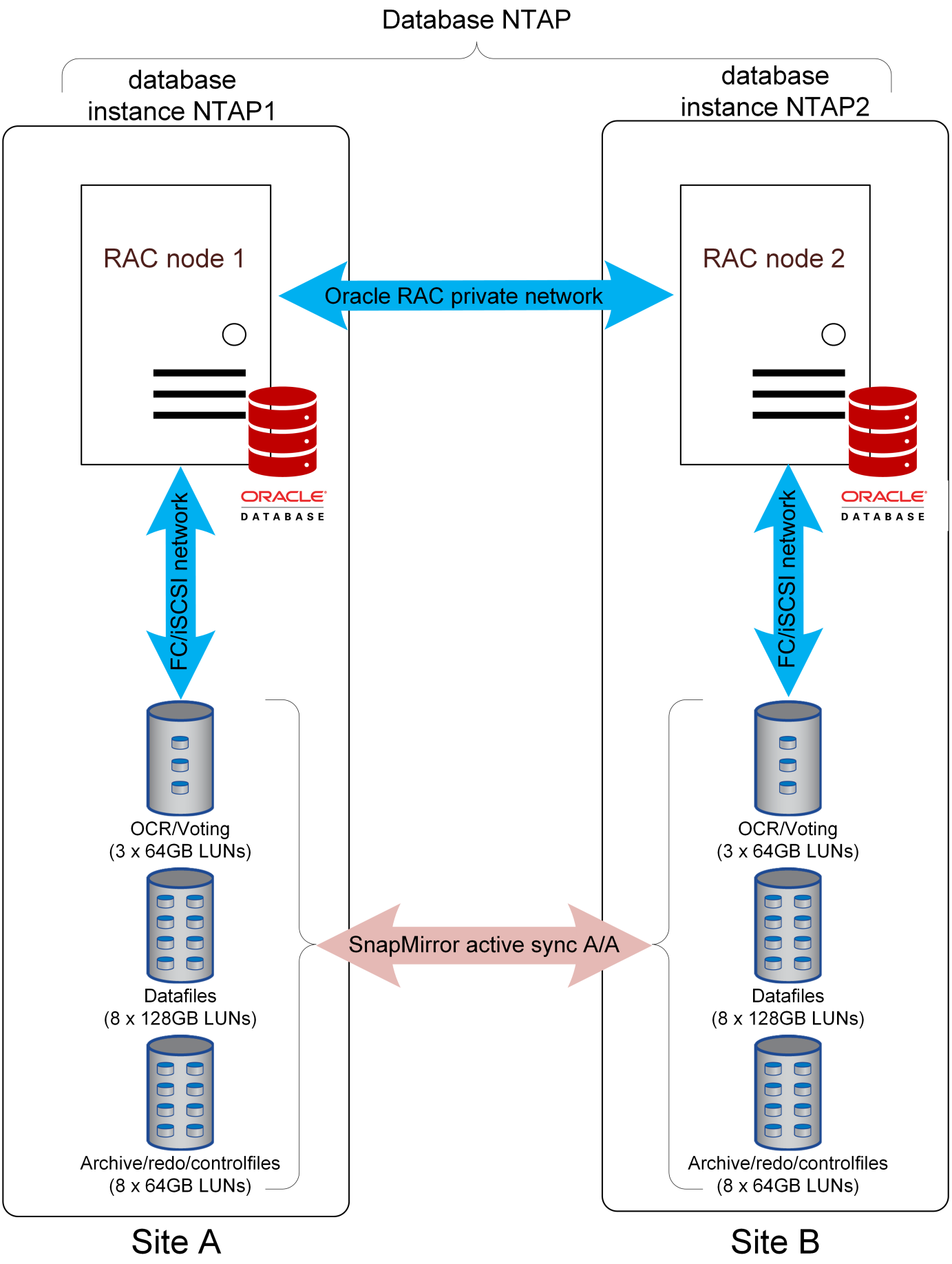
在 SnapMirror 主動式同步上使用 Oracle RAC 最重要的考量、是使用統一或非統一存取。

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

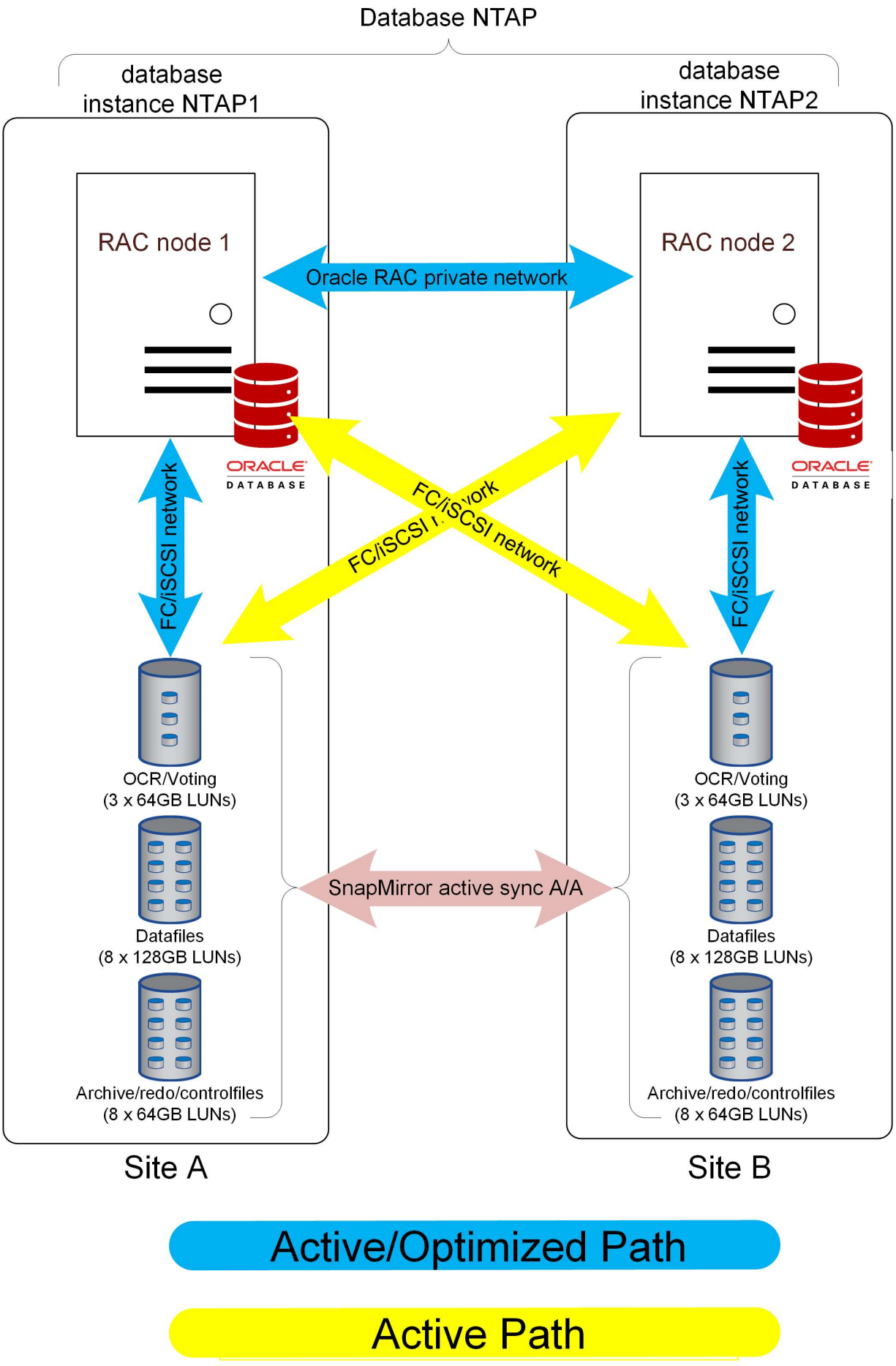
統一存取需要將 SAN 延伸至各個站台。主要優點是儲存系統的遺失不會導致資料庫執行個體遺失。相反地、它會導致路徑目前正在使用的多重路徑變更。

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**

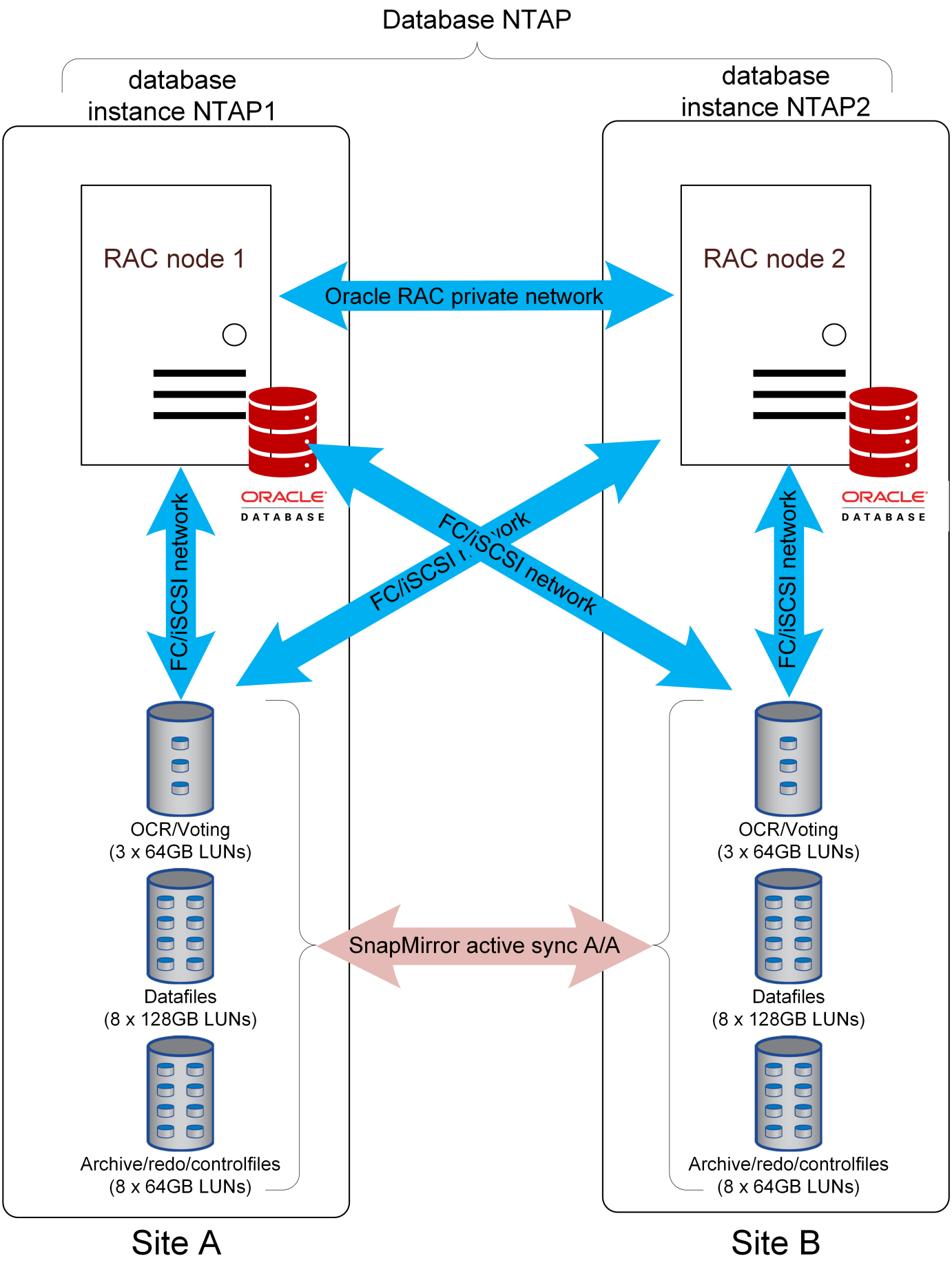
如果站台之間存在嚴重延遲、則可以使用主機鄰近設定來設定 AFF 系統。如此一來、每個儲存系統就能知道哪些主機是本機主機、哪些是遠端主機、並適當地指派路徑優先順序。


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

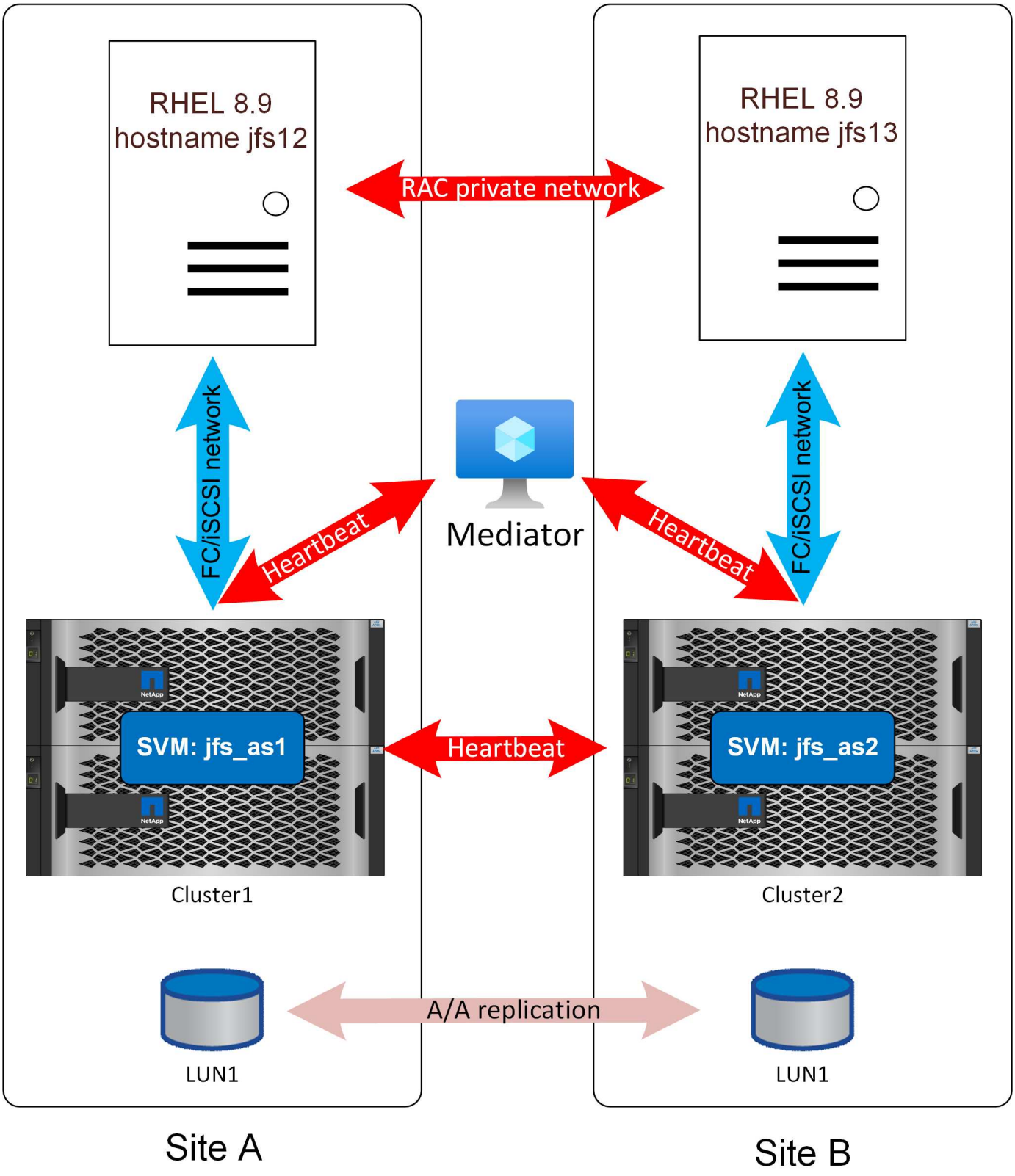
變更偏好的站台需要簡單的操作。IO 會暫停一秒或兩秒、作為叢集之間複寫行為切換的權限、但 IO 不會受到影響。

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。

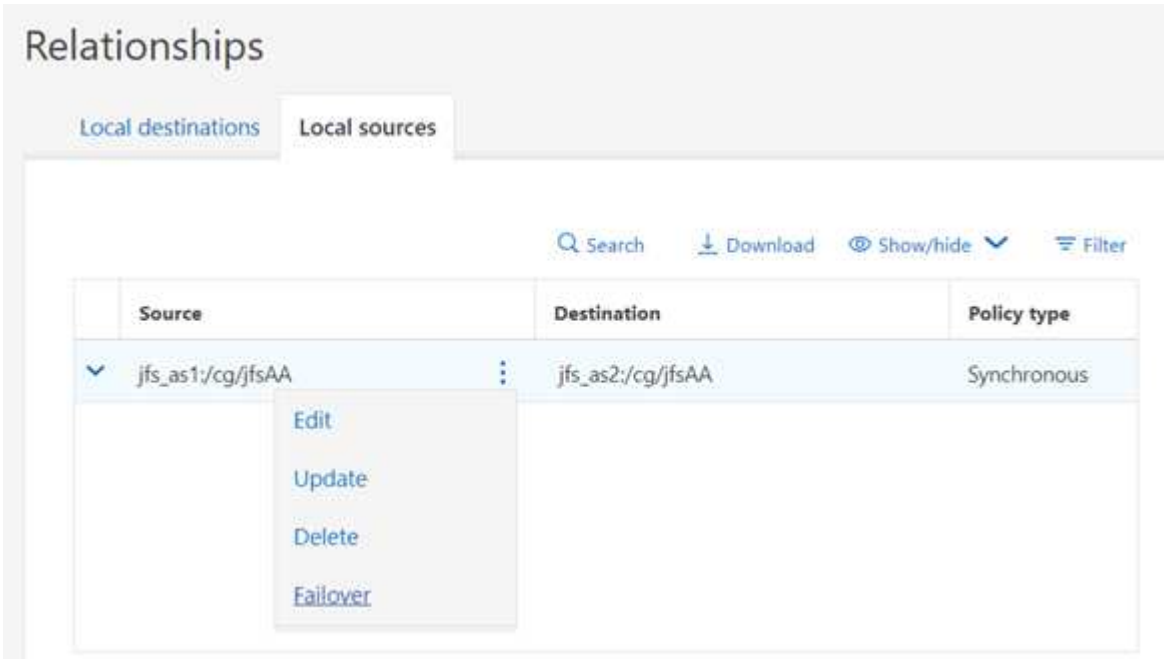
手動容錯移轉

「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
```