MetroCluster

Enterprise applications

NetApp
May 09, 2024

目錄

MetroCluster

MetroCluster 實體架構和 Oracle 資料庫

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

[注意] 「連線」一詞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。

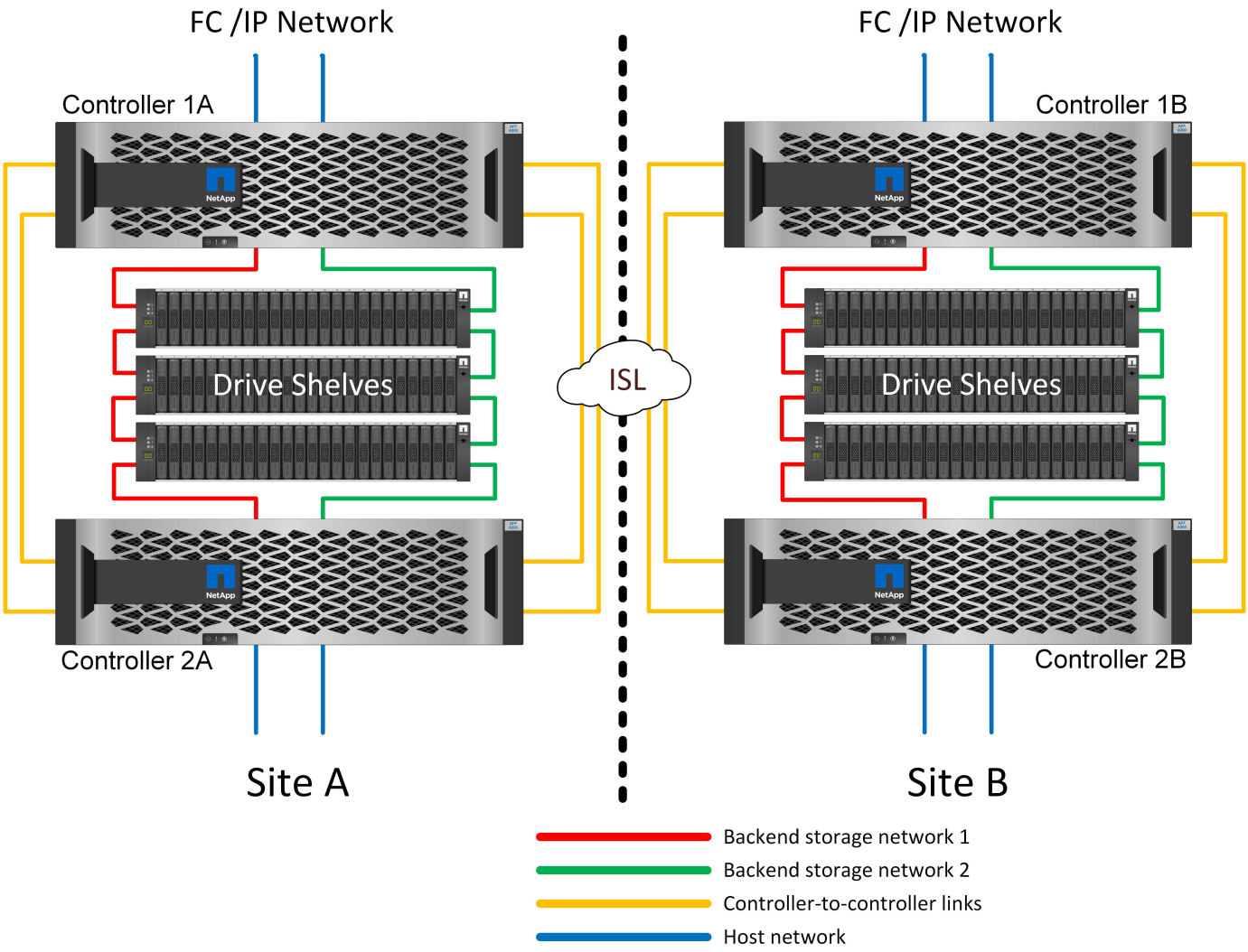
知識產權MetroCluster

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。

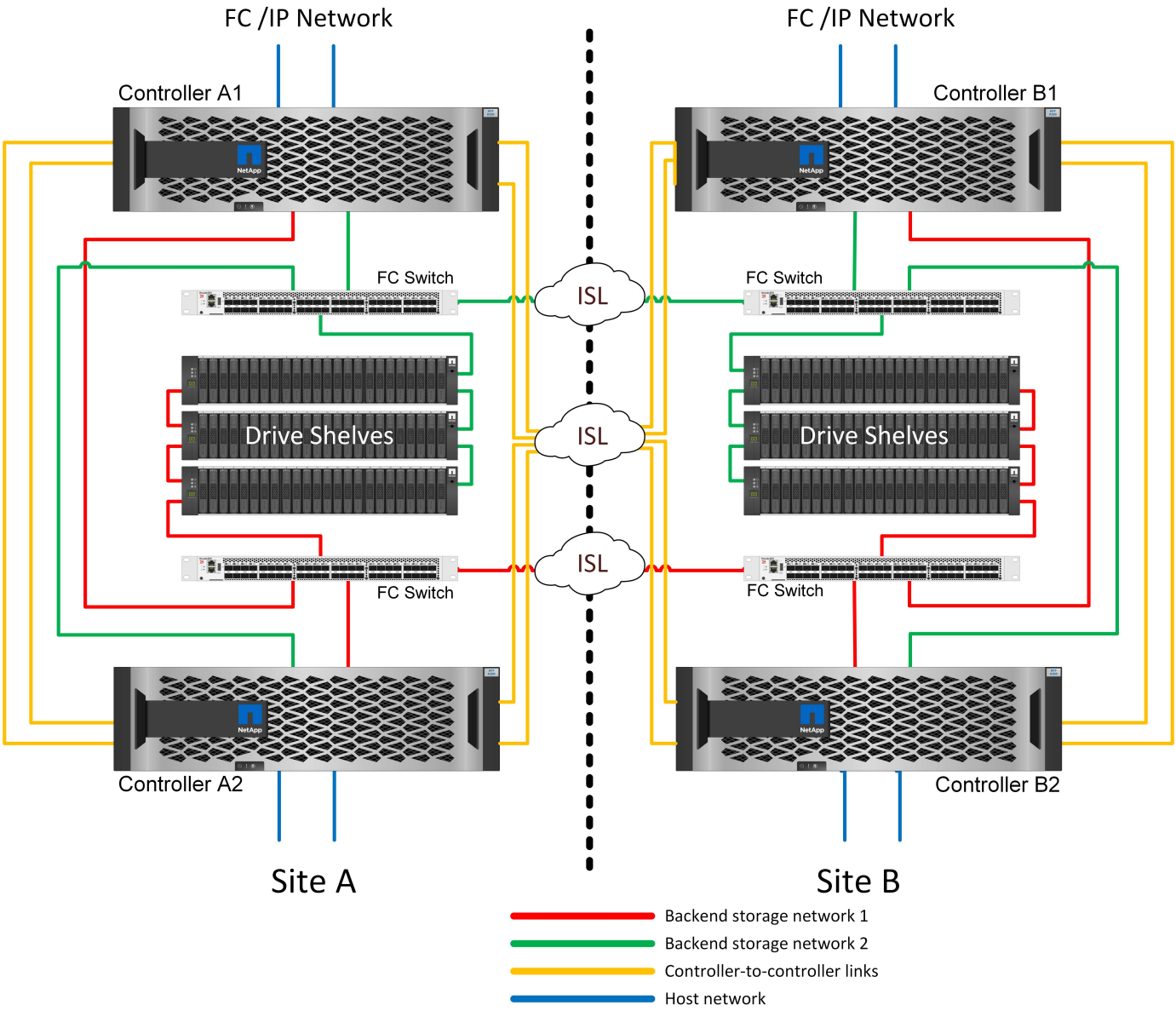
由於唯一的跨站台連線適用於控制器、因此架構也更簡單。在 FC SAN 附加 MetroCluster 中、控制器會直接寫入另一個站台上的磁碟機、因此需要額外的 SAN 連線、交換器和橋接器。相反地、 IP 組態中的控制器會透過控制器寫入相對的磁碟機。

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。

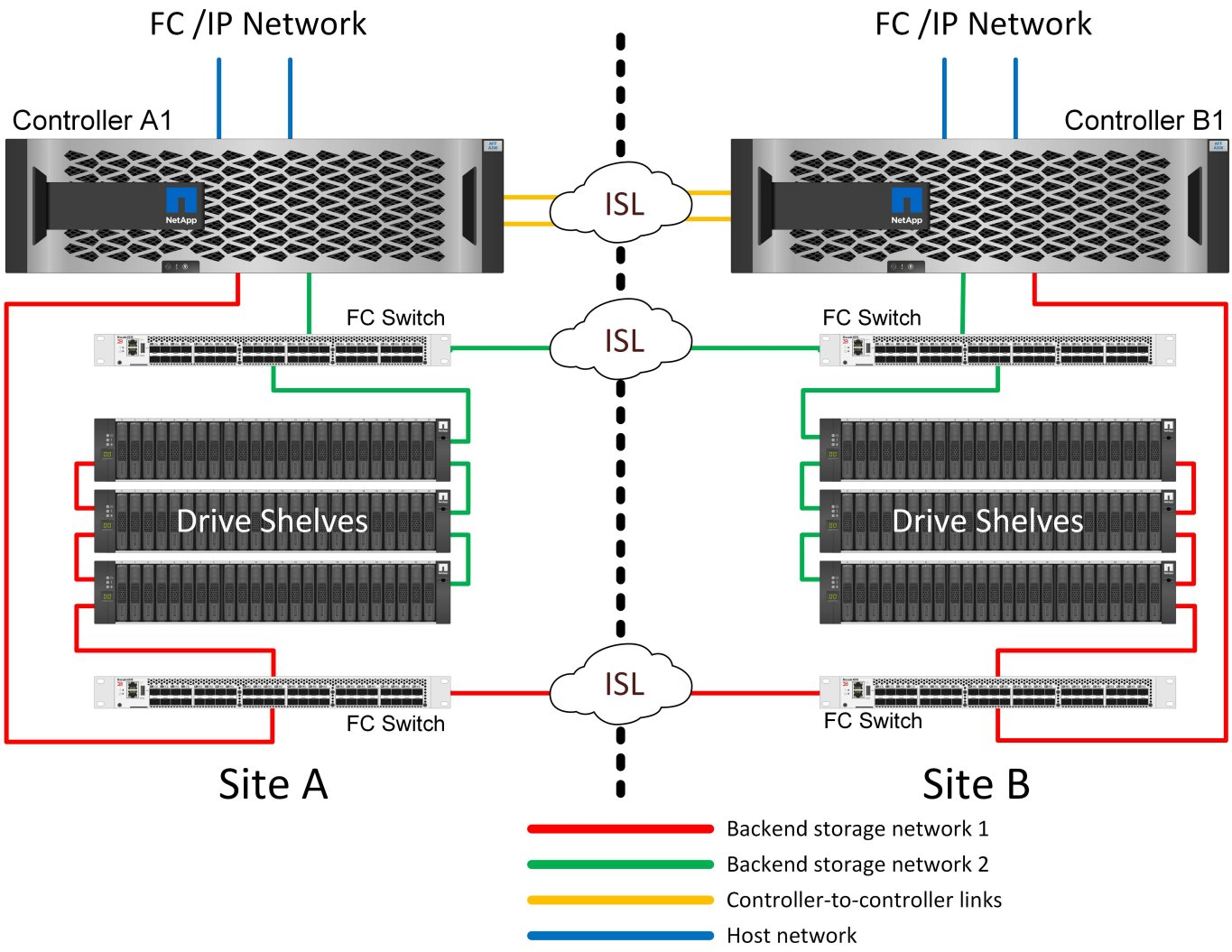


有些多站台基礎架構並非設計用於主動式作業、而是更多用於主要站台和災難恢復站台。在這種情況下、 HA 配對 MetroCluster 選項通常較為理想、原因如下：

- 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。

MetroCluster 邏輯架構和 Oracle 資料庫

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

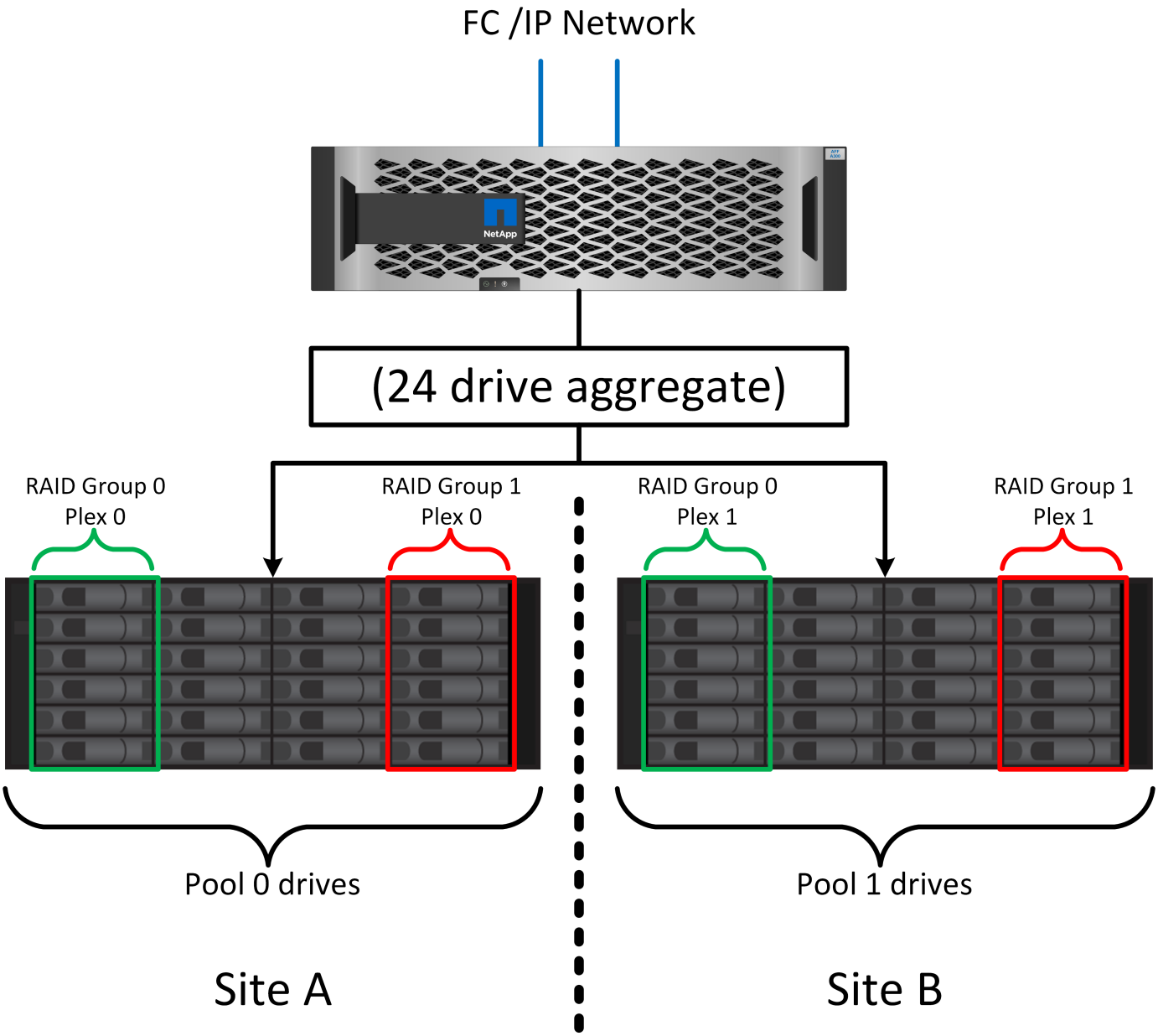
MetroCluster 以下列方式擴充 NVRAM 資料保護：

- 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

。 "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。如需有關斷路器的完整說明、請參閱 "NetApp 支援網站"但 MetroCluster 斷路

器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。

- 注意： * 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**

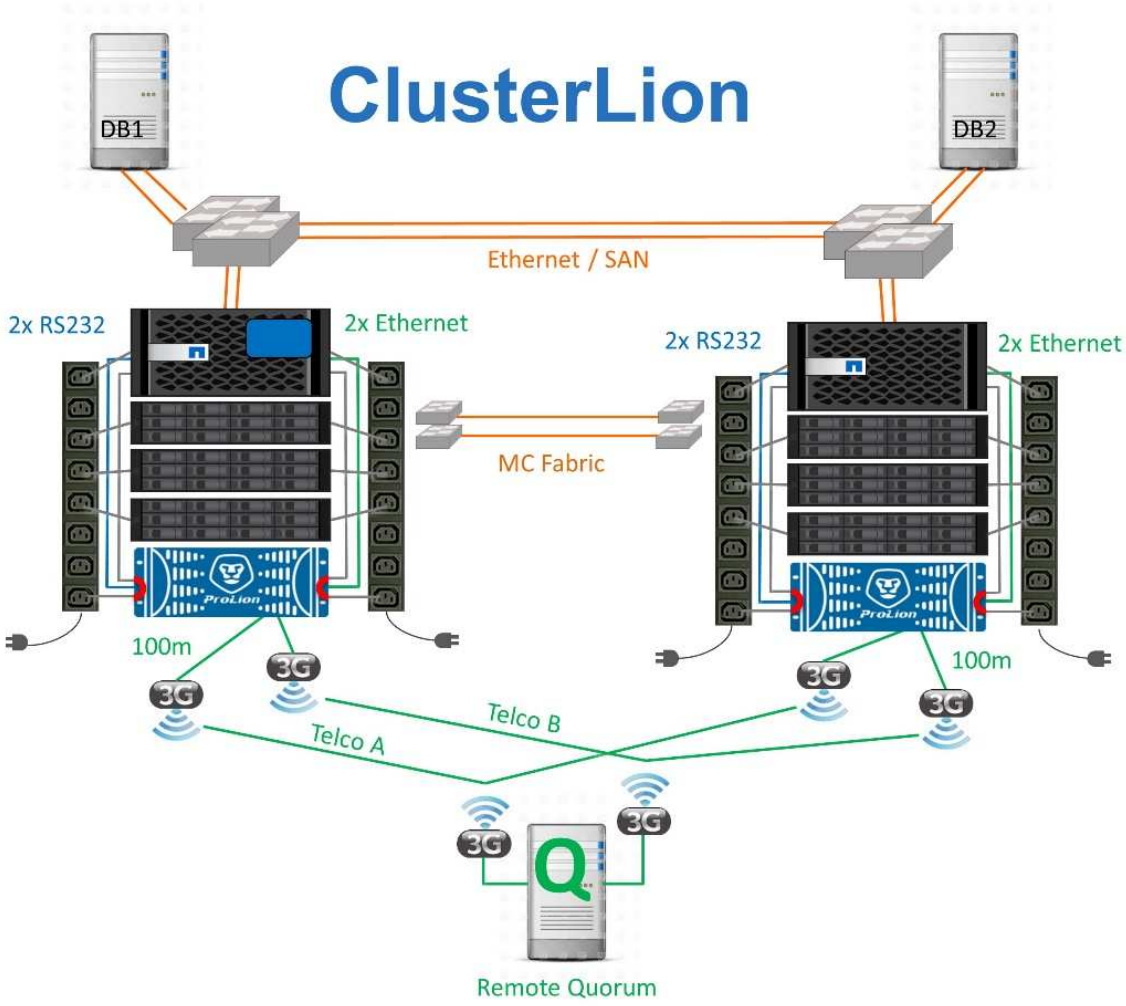
ONTAP Mediator 可搭配 MetroCluster IP 和某些其他 ONTAP 解決方案使用。它是一項傳統的斷路器服務、就像上述的 MetroCluster tiebreaker 軟體一樣、但也包含一項重要功能、即執行自動無人值守的移除。

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 的 Oracle 資料庫

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 或主機端磁碟鏡射的延長逾時。如需其他資訊和選項、請洽詢您的 NetApp 或合作夥伴客戶團隊。

使用 **MetroCluster** 進行 **Oracle** 資料庫容錯移轉

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

通常的最佳實務做法仍適用、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。

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。Oracle RAC 、 Veritas 叢集伺服器（ VCS ）或 VMware HA 等叢集式產品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）會設定為在替代站台上線應用程式。其中一個選項是建立預先針對構成應用程式的 NFS 或 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. 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
4. 啟動應用程式

此方法的主要需求是在遠端站台上執行作業系統。它必須預先設定應用程式二進位檔、也就是說、修補等工作必須在主要站台和待命站台上執行。或者、應用程式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。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 資料庫、 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。

MetroCluster 上的 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。

MetroCluster 上的延伸 Oracle 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 ASM 鏡像來取代 MetroCluster 。ASM 鏡像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、請使用 `misscount` 可能需要變更參數。如需 RAC 逾時參數的詳細資訊、請參閱 "Oracle RAC 搭配 ONTAP"。

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。

版權資訊

商標資訊