網路組態
Enterprise applications

NetApp
May 09, 2024

目錄

網路組態

Oracle 資料庫的邏輯介面設計

Oracle 資料庫需要存取儲存設備。邏輯介面（ Lifs ）是將儲存虛擬機器（ SVM ）連接到網路、然後再連接到資料庫的網路配送。需要適當的 LIF 設計、才能確保每個資料庫工作負載都有足夠的頻寬、而且容錯移轉不會導致儲存服務遺失。

本節概述 LIF 的主要設計原則。如需更完整的文件、請參閱 "ONTAP 網路管理文件"。與資料庫架構的其他層面一樣、儲存虛擬機器（ SVM 、在 CLI 稱為 vserver ）和邏輯介面（ LIF ）設計的最佳選項、在很大程度上取決於擴充需求和業務需求。

建置 LIF 策略時、請考量下列主要主題：

- * 效能。 * 網路頻寬是否足夠？
- * 恢復能力。 * 設計中是否有任何單點故障？
- * 管理能力。 * 網路是否能不中斷地擴充？

這些主題適用於端點對端點解決方案、從主機到交換器、再到儲存系統。

LIF 類型

有多種 LIF 類型。 "LIF 類型的 ONTAP 文件" 提供更完整的本主題資訊、但從功能觀點來看、生命可分為下列群組：

- * 用於管理儲存叢集的叢集與節點管理生命期 * 。
- * SVM 管理階層。 * 允許透過 REST API 或 ONTAPI （也稱為 ZAPI ）存取 SVM 的介面、以執行快照建立或磁碟區調整大小等功能。SnapManager for Oracle （ SMO ）等產品必須能夠存取 SVM 管理 LIF 。
- * 資料生命。 * FC 、 iSCSI 、 NVMe / FC 、 NVMe / TCP 、 NFS 、 或 SMB/CIFS 資料。

> 用於 NFS 流量的資料 LIF 也可透過變更防火牆原則來進行管理 `data` 至 `mgmt` 或其他允許 HTTP 、 HTTPS 或 SSH 的原則。這項變更可避免每部主機的組態設定、以同時存取 NFS 資料 LIF 和個別的管理 LIF 、進而簡化網路組態。儘管這兩者都使用 IP 傳輸協定、但無法同時為 iSCSI 和管理流量設定介面。iSCSI 環境需要個別的管理 LIF 。

SAN LIF 設計

SAN 環境中的 LIF 設計相對簡單、原因有一：多重路徑。所有現代化的 SAN 實作均可讓用戶端透過多個、不受限制的網路路徑存取資料、並選擇最佳的存取路徑或路徑。因此、 LIF 設計的效能更容易因應、因為 SAN 用戶端會在最佳可用路徑之間自動平衡 I/O 負載。

如果路徑無法使用、用戶端會自動選取不同的路徑。因此設計簡易性讓 SAN 的工作更容易管理。這並不表示 SAN 環境總是更容易管理、因為 SAN 儲存設備還有許多其他層面比 NFS 複雜得多。這只是表示 SAN LIF 設計更簡單。

效能

在 SAN 環境中、 LIF 效能的最重要考量是頻寬。例如、雙節點 ONTAP AFF 叢集每個節點有兩個 16GB FC 連接埠、可在每個節點之間提供高達 32GB 的頻寬。

恢復能力

AFF 儲存系統上的 SAN Lifs 不會容錯移轉。如果 SAN LIF 因控制器容錯移轉而失敗、則用戶端的多重路徑軟體會偵測路徑遺失、並將 I/O 重新導向至不同的 LIF 。使用 ASA 儲存系統時、在短暫延遲後將會容錯移轉、但這不會中斷 IO 、因為其他控制器上已經有作用中的路徑。發生容錯移轉程序是為了在所有定義的連接埠上還原主機存取。

管理能力

LIF 移轉是 NFS 環境中較常見的工作、因為 LIF 移轉通常與在叢集周圍重新放置磁碟區有關。當磁碟區移轉至 HA 配對內時、無需在 SAN 環境中移轉 LIF 。這是因為在磁碟區移動完成之後、 ONTAP 會傳送路徑變更通知給 SAN 、而 SAN 用戶端會自動重新最佳化。與 SAN 的 LIF 移轉主要與重大實體硬體變更有關。例如、如果需要不中斷營運的控制器升級、則 SAN LIF 會移轉至新硬體。如果發現 FC 連接埠故障、 LIF 就可以移轉至未使用的連接埠。

設計建議

NetApp 提出下列建議：

- 請勿建立超過所需的路徑。過多的路徑會使整體管理更為複雜、並可能導致部分主機上路徑容錯移轉的問題。此外、有些主機對 SAN 開機等組態有非預期的路徑限制。
- 極少數組態需要四條以上的路徑才能連接到 LUN 。如果擁有 LUN 的節點及其 HA 合作夥伴故障、則無法存取主控 LUN 的集合、因此限制將超過兩個節點的路徑通告至 LUN 的價值。在非主要 HA 配對的節點上建立路徑、在這種情況下並無幫助。
- 雖然可視 LUN 路徑的數量可以透過選擇 FC 區域中包含哪些連接埠來進行管理、但通常較容易在 FC 區域中包含所有潛在目標點、並控制 ONTAP 層級的 LUN 可見度。
- 在 ONTAP 8.3 及更新版本中、選擇性 LUN 對應（ SLM ）功能為預設功能。透過 SLM 、任何新的 LUN 都會自動從擁有基礎 Aggregate 的節點和節點的 HA 合作夥伴通告。這種安排可避免建立連接埠集或設定分區、以限制連接埠存取。每個 LUN 都可在最佳效能和恢復能力所需的最小節點數上使用。
- 如果必須在兩個控制器之外移轉 LUN 、則可以使用新增額外的節點 `lun mapping add-reporting-nodes` 命令，以便在新節點上通告 LUN 。這樣做會建立通往 LUN 的額外 SAN 路徑、以進行 LUN 移轉。但是、主機必須執行探索作業、才能使用新路徑。
- 不要過度擔心間接流量。最好在 I/O 密集環境中避免間接流量、因為每微秒的延遲都是關鍵、但對於一般工作負載而言、可見的效能影響卻微不足道。

NFS LIF 設計

與 SAN 通訊協定不同、 NFS 定義多個資料路徑的能力有限。NFSv4 的平行 NFS （ pNFS ）擴充解決了這項限制、但由於乙太網路速度已達 100GB 、而且在新增其他路徑時、很少會有價值。

效能與恢復能力

雖然測量 SAN LIF 效能主要是從所有主要路徑計算總頻寬、但判斷 NFS LIF 效能需要仔細瞭解確切的網路組態。例如、兩個 10Gb 連接埠可設定為原始實體連接埠、或是可設定為連結集合體控制傳輸協定（ LACP ）介面群組。如果將它們設定為介面群組、則可根據流量是交換還是路由、使用不同的多個負載平衡原則。最後、

Oracle Direct NFS （ DNFS ）提供目前不存在於任何 OS NFS 用戶端的負載平衡組態。

與 SAN 通訊協定不同的是、 NFS 檔案系統需要在通訊協定層恢復能力。例如、 LUN 一律設定為啟用多重路徑、表示儲存系統可使用多個備援通道、每個通道都使用 FC 傳輸協定。另一方面、 NFS 檔案系統則取決於單一 TCP/IP 通道的可用度、而該通道只能在實體層受到保護。這種配置是為何存在連接埠容錯移轉和 LACP 連接埠集合等選項。

在 NFS 環境中、網路傳輸協定層會同時提供效能和恢復能力。因此、這兩個主題彼此交織在一起、必須一起討論。

將生命體繫結至連接埠群組

若要將 LIF 繫結至連接埠群組、請將 LIF IP 位址與一組實體連接埠建立關聯。將實體連接埠集合在一起的主要方法是 LACP 。LACP 的容錯功能相當簡單； LACP 群組中的每個連接埠都會受到監控、並在發生故障時從連接埠群組中移除。不過、對於 LACP 在效能方面的運作方式、有許多誤解：

- LACP 不需要在交換器上進行組態以符合端點。例如、 ONTAP 可設定 IP 型負載平衡、而交換器則可使用 MAC 型負載平衡。
- 使用 LACP 連線的每個端點可以個別選擇封包傳輸連接埠、但無法選擇用於接收的連接埠。這表示從 ONTAP 到特定目的地的流量會連結到特定連接埠、而傳回流量可能會到達不同的介面。但這不會造成問題。
- LACP 不會一直平均分配流量。在擁有許多 NFS 用戶端的大型環境中、通常甚至會使用 LACP 集合中的所有連接埠。不過、環境中的任何一個 NFS 檔案系統都只能使用一個連接埠的頻寬、而非整個集合。
- 雖然 ONTAP 上有資源配置資源配置資源 LACP 原則、但這些原則並不會解決從交換器到主機的連線問題。例如、主機上有四埠 LACP 主幹的組態、 ONTAP 上有四埠 LACP 主幹的組態、仍只能使用單一連接埠讀取檔案系統。雖然 ONTAP 可以透過所有四個連接埠傳輸資料、但目前沒有任何交換器技術可以透過所有四個連接埠從交換器傳送到主機。僅使用一個。

在包含許多資料庫主機的大型環境中、最常見的方法是使用 IP 負載平衡、建立一個包含適當數量 10Gb （或更快）介面的 LACP 集合體。只要有足夠的用戶端、這種方法就能讓 ONTAP 提供所有連接埠的均勻使用。當組態中的用戶端較少時、負載平衡會中斷、因為 LACP 主幹不會動態重新分配負載。

建立連線後、特定方向的流量只會放置在一個連接埠上。例如、對透過四埠 LACP 主幹連接的 NFS 檔案系統執行完整表格掃描的資料庫、只會透過一個網路介面卡（ NIC ）讀取資料。如果只有三個資料庫伺服器在這種環境中、則可能所有三個都從同一個連接埠讀取、而其他三個連接埠則處於閒置狀態。

將生命與實體連接埠繫結

將 LIF 繫結至實體連接埠、可更精細地控制網路組態、因為 ONTAP 系統上的指定 IP 位址一次只與一個網路連接埠相關聯。然後、可透過設定容錯移轉群組和容錯移轉原則來實現恢復能力。

容錯移轉原則和容錯移轉群組

網路中斷期間的生命行為是由容錯移轉原則和容錯移轉群組所控制。不同版本的 ONTAP 已變更組態選項。請參閱 "適用於容錯移轉群組和原則的 ONTAP 網路管理文件" 以取得所部署 ONTAP 版本的特定詳細資料。

ONTAP 8.3 及更高版本可根據廣播網域來管理 LIF 容錯移轉。因此、系統管理員可以定義所有可存取指定子網路的連接埠、並允許 ONTAP 選取適當的容錯移轉 LIF 。這種方法可由部分客戶使用、但由於缺乏可預測性、因此在高速儲存網路環境中有限制。例如、環境可同時包含 1Gb 連接埠、以供例行檔案系統存取、而 10Gb 連接埠則可用於資料檔案 I/O如果兩種連接埠都存在於同一個廣播網域中、 LIF 容錯移轉可能會導致資料檔案 I/O 從 10Gb 連接埠移至 1Gb 連接埠。

總而言之、請考慮下列實務做法：

1. 將容錯移轉群組設定為使用者定義。
2. 將儲存容錯移轉（ SFO ）合作夥伴控制器上的連接埠填入容錯移轉群組、以便在儲存容錯移轉期間、生命體跟隨集合體。如此可避免產生間接流量。
3. 使用效能特性與原始 LIF 相符的容錯移轉連接埠。例如、單一實體 10Gb 連接埠上的 LIF 應包含單一 10Gb 連接埠的容錯移轉群組。四埠 LACP LIF 應容錯移轉至另一個四埠 LACP LIF 。這些連接埠將是廣播網域中定義的連接埠子集。
4. 將容錯移轉原則設為僅限 SFO 合作夥伴。這樣做可確保 LIF 在容錯移轉期間跟隨集合體。

自動還原

設定 `auto-revert` 視需要設定參數。大多數客戶偏好將此參數設為 `true` 讓 LIF 還原至其主連接埠。不過、在某些情況下、客戶將此設定為「假」、表示在將 LIF 傳回其主連接埠之前、可以調查非預期的容錯移轉。

LIF 與 **Volume** 比率

常見的誤解是、磁碟區和 NFS 生命體之間必須有一對一的關係。雖然在叢集中的任何位置移動磁碟區都需要此組態、但絕不會產生額外的互連流量、但絕對不需要此組態。必須考慮叢集間流量、但僅存在叢集間流量並不會造成問題。為 ONTAP 所發佈的許多基準測試主要包括間接 I/O

例如、資料庫專案中包含相對少數的效能關鍵資料庫、只需要總共 40 個磁碟區、可能需要將 1 ： 1 磁碟區轉換為 LIF 策略、這種安排需要 40 個 IP 位址。然後、任何磁碟區都可以連同相關的 LIF 一起移至叢集中的任何位置、而且流量永遠是直接的、即使在微秒層級、也能將每個延遲來源減至最低。

舉例來說、大型託管環境的管理可能更容易、因為客戶與生命的關係是一對一。隨著時間的推移、可能需要將磁碟區移轉至不同的節點、這會造成一些間接流量。但是、除非互連交換器上的網路連接埠飽和、否則效能影響應該無法偵測。如果有疑慮、可以在其他節點上建立新的 LIF 、並在下一個維護時段更新主機、以移除組態中的間接流量。

用於 **Oracle** 資料庫的 **TCP/IP** 和乙太網路組態

ONTAP 上的許多 Oracle 客戶都使用乙太網路、 NFS 、 iSCSI 、 NVMe / TCP 的網路傳輸協定、尤其是雲端。

主機作業系統設定

大多數應用程式廠商文件都包含特定的 TCP 和乙太網路設定、以確保應用程式能以最佳方式運作。這些相同的設定通常足以提供最佳的 IP 型儲存效能。

乙太網路流量控制

這項技術可讓用戶端要求傳送者暫時停止資料傳輸。這通常是因為接收者無法快速處理傳入的資料。一次、要求傳送者停止傳輸的中斷程度比接收者丟棄封包的中斷程度低、因為緩衝區已滿。現今作業系統中使用的 TCP 堆疊已不再如此。事實上、流量控制所造成的問題比解決的問題還多。

近年來、乙太網路流量控制所造成的效能問題不斷增加。這是因為乙太網路流量控制是在實體層運作。如果網路組態允許任何主機作業系統將乙太網路流量控制要求傳送至儲存系統、則所有連線的用戶端都會暫停 I/O 。由於單一儲存控制器服務的用戶端數量不斷增加、因此其中一或多個用戶端傳送流量控制要求的可能性會增加。在擁有廣泛作業系統虛擬化的客戶據點、經常會發現這個問題。

NetApp 系統上的 NIC 不應接收流量控制要求。實現此結果的方法因網路交換器製造商而異。在大多數情況下、可將乙太網路交換器上的流量控制設定為 `receive desired` 或 `receive on`，這意味着流控制請求不會轉發到存儲控制器。在其他情況下、儲存控制器上的網路連線可能不允許停用流程控制。在這些情況下、用戶端必須設定為永遠不要傳送流量控制要求、方法是變更至主機伺服器本身的 NIC 組態、或是變更主機伺服器所連接的交換器連接埠。

* NetApp 建議 * 確保 NetApp 儲存控制器不會接收乙太網路流量控制封包。這通常可以透過設定控制器所連接的交換器連接埠來完成、但有些交換器硬體有限制、可能需要改用用戶端端變更。

MTU 大小

使用巨型框架的結果顯示、透過降低 CPU 和網路成本、可在速度較低的網路中提供一些效能改善、但效益通常並不顯著。

* NetApp 建議 * 盡可能實作巨型框架、以實現任何可能的效能效益、並確保解決方案符合未來需求。

在 10Gb 網路中使用巨型框架幾乎是強制性的。這是因為大多數的 10Gb 實作都達到每秒封包數的限制、而不需要巨型框架、就能達到 10Gb 標誌。使用巨型框架可改善 TCP/IP 處理效率、因為它可讓作業系統、伺服器、NIC 和儲存系統處理較少但較大的封包。效能的改善因 NIC 而異、但成效相當顯著。

對於巨型框架實作、通常但不正確的看法是、所有連線的裝置都必須支援巨型框架、而且 MTU 大小必須與端點對端點相符而是在建立連線時、兩個網路端點會協商最高的雙方可接受的框架大小。在一般環境中、網路交換器的 MTU 大小設為 9216 、 NetApp 控制器設為 9000 、用戶端則設為 9000 和 1514 的混合。支援 9000 MTU 的用戶端可以使用巨型框架、而只支援 1514 的用戶端可以協商較低的值。

在完全交換的環境中、這種配置的問題很少發生。不過、在沒有中繼路由器被迫分割巨型框架的路由環境中、請務必小心。

- NetApp 建議 * 設定下列項目：
- 使用 1 GB 乙太網路（ GbE ）時、巨型框架是理想的選擇、但不是必要的。
- 使用 10GbE 及更快的速度、需要巨型框架才能達到最佳效能。

TCP 參數

三項設定通常設定錯誤： TCP 時間戳記、選擇性認可（ SACK ）和 TCP 視窗縮放。網際網路上的許多過時文件建議停用一或多個這些參數、以改善效能。這項建議在多年前就有一些優點、因為 CPU 功能較低、因此有助於盡可能降低 TCP 處理的成本。

然而、在現代化的作業系統中、停用任何這些 TCP 功能通常會導致無法偵測的效益、同時也可能造成效能受損。在虛擬化網路環境中、效能受損的可能性特別大、因為這些功能是有效處理封包遺失和網路品質變更所必需的。

* NetApp 建議 * 在主機上啟用 TCP 時間戳記、 SACK 和 TCP 視窗縮放功能、而且在任何目前的作業系統中、這三個參數都應該預設為開啟。

Oracle 資料庫的 FC 組態

為 Oracle 資料庫設定 FC SAN 主要是為了遵循日常的 SAN 最佳實務做法。

這包括典型的規劃措施、例如確保主機和儲存系統之間的 SAN 上有足夠的頻寬、使用 FC 交換器廠商所需的 FC 連接埠設定、檢查所有必要裝置之間是否存在所有 SAN 路徑、避免 ISL 爭用、 並使用適當的 SAN 架構監控。

分區

FC 區域不得包含多個啟動器。這種安排一開始可能會運作、但啟動器之間的串擾最終會影響效能和穩定性。

雖然在極少數情況下、來自不同廠商的 FC 目標連接埠行為造成問題、但多目標區域通常被視為安全區域。例如、避免將 NetApp 和非 NetApp 儲存陣列的目標連接埠同時納入同一區域。此外、將 NetApp 儲存系統和磁帶裝置置於同一個區域、更有可能造成問題。

Oracle 資料庫與直接連線 ONTAP 連線

儲存管理員有時偏好從組態中移除網路交換器、以簡化其基礎架構。在某些情況下可能會支援這項功能。

iSCSI 和 NVMe / TCP

使用 iSCSI 或 NVMe / TCP 的主機可以直接連線至儲存系統、並正常運作。原因是路徑。直接連線至兩個不同的儲存控制器、會產生兩個不同的資料流路徑。遺失路徑、連接埠或控制器並不會妨礙其他路徑的使用。

NFS

可以使用直接連線的 NFS 儲存設備、但有很大的限制：如果沒有大量的指令碼工作、容錯移轉將無法運作、這是客戶的責任。

直接連線的 NFS 儲存設備會造成不中斷的容錯移轉複雜化、這是因為本機作業系統上會發生路由。例如、假設主機的 IP 位址為 192.168.1.1/24 、並直接連線至 IP 位址為 192.168.1.50/24 的 ONTAP 控制器。在容錯移轉期間、該位址 192.168.1.50 可以容錯移轉至其他控制器、而且主機可以使用該位址、但主機如何偵測其存在？原來的 192.168.1.1 位址仍然存在於不再連線至作業系統的主機 NIC 上。目的地為 192.168.1.5 的流量將繼續傳送至無法運作的網路連接埠。

第二個 OS NIC 可設定為 19 可以與故障的 over 192.168.1.50 位址進行通訊、但本機路由表預設會使用一個 * 且只有一個 * 位址來與 192.168.1.0/24 子網路通訊。系統管理員可以建立指令碼架構、以偵測失敗的網路連線、並變更本機路由表或使介面正常運作。具體程序取決於所使用的作業系統。

實際上、 NetApp 客戶確實有直接連線的 NFS 、但通常僅適用於容錯移轉期間 IO 暫停的工作負載。使用硬掛載時、在這類暫停期間不應有任何 IO 錯誤。IO 應該會暫停運作、直到服務還原為止、無論是透過容錯回復或手動介入、在主機上的 NIC 之間移動 IP 位址。

FC 直接連線

無法使用 FC 傳輸協定將主機直接連接至 ONTAP 儲存系統。原因是使用 NPIV 。用於識別 FC 網路的 ONTAP FC 連接埠的 WWN 使用稱為 NPIV 的虛擬化類型。任何連接至 ONTAP 系統的裝置都必須能夠辨識 NPIV WWN 。目前沒有任何 HBA 廠商提供可安裝在能夠支援 NPIV 目標的主機上的 HBA 。