FC SAN

Enterprise applications

NetApp
May 03, 2024

目錄

FC SAN

Oracle 資料庫 I/O 的 LUN 對齊

LUN 對齊是指針對基礎檔案系統配置最佳化 I/O 。

在 ONTAP 系統上、儲存設備是以 4KB 為單位進行組織。資料庫或檔案系統 8KB 區塊應對應至兩個 4KB 區塊。如果 LUN 組態發生錯誤、在任一方向將對齊移至 1KB 、則每個 8KB 區塊會存在於三個不同的 4KB 儲存區塊、而非兩個。這種安排會導致延遲增加、並導致在儲存系統中執行額外的 I/O 。

對齊也會影響 LVM 架構。如果在整個磁碟機裝置上定義邏輯磁碟區群組內的實體磁碟區（不建立分割區）、 LUN 上的前 4KB 區塊會與儲存系統上的前 4KB 區塊對齊。這是正確的對齊方式。磁碟分割發生問題、因為它們會移轉作業系統使用 LUN 的起始位置。只要偏移量以 4KB 的整體單位移動、 LUN 就會對齊。

在 Linux 環境中、在整個磁碟機裝置上建立邏輯磁碟區群組。當需要磁碟分割時、請執行檢查對齊 `fdisk -u` 並驗證每個分割區的開始時間為八個之倍數。這表示分割區從八個 512 位元組磁區的倍數開始、即 4KB 。

另請參閱一節中有關壓縮區塊對齊的討論 "效率"。任何與 8KB 壓縮區塊邊界對齊的配置、也會與 4KB 邊界對齊。

錯誤對齊警告

資料庫重做 / 交易記錄通常會產生未對齊的 I/O 、導致 ONTAP 上未對齊 LUN 的錯誤警告。

記錄會以不同大小的寫入方式、連續寫入記錄檔。不符合 4KB 界限的記錄寫入作業通常不會造成效能問題、因為下一個記錄寫入作業會完成區塊。結果是 ONTAP 幾乎能將所有寫入作業視為完整的 4KB 區塊來處理、即使某些 4KB 區塊中的資料是以兩個不同的作業來寫入。

使用公用程式（例如）來驗證對齊 `sio` 或 `dd` 可在定義的區塊大小下產生 I/O 。您可以使用檢視儲存系統上的 I/O 對齊統計資料 `stats` 命令。請參閱 "WAFL 對齊驗證" 以取得更多資訊。

在 Solaris 環境中進行對齊更為複雜。請參閱 "SAN主機組態ONTAP" 以取得更多資訊。

注意
在 Solaris x86 環境中、由於大多數組態都有多層分割區、因此請格外注意正確的對齊方式。Solaris x86 分割區磁碟片通常位於標準主開機記錄分割區表格的上方。

Oracle 資料庫 LUN 規模調整和 LUN 數量

選擇最佳 LUN 大小和要使用的 LUN 數量、對於 Oracle 資料庫的最佳效能和管理性至關重要。

LUN 是 ONTAP 上的虛擬化物件、存在於託管集合體中的所有磁碟機中。因此、 LUN 的效能不受其大小影響、因為無論選擇何種大小、 LUN 都會充分發揮彙總的效能潛力。

為了方便起見、客戶可能想要使用特定大小的 LUN 。例如、如果資料庫建置在由兩個 LUN 組成的 LVM 或 Oracle ASM 磁碟群組上、每個 LUN 均為 1TB 、則該磁碟群組必須以 1TB 為增量來擴充。最好是從八個 LUN （每個 LUN 爲 500GB ）構建磁盤組，以便可以以更小的增量來增加磁盤組。

我們不鼓勵建立通用標準 LUN 大小的做法、因為這樣做可能會使管理變得複雜。例如、當資料庫或資料存放區的範圍介於 1TB 到 2TB 時、 100GB 的標準 LUN 大小可能運作良好、但大小為 20TB 的資料庫或資料存放區需要 200 個 LUN 。這表示伺服器重新開機時間較長、不同 UI 中需要管理的物件較多、而 SnapCenter 等產品必須在許多物件上執行探索。使用較少、較大的 LUN 可避免此類問題。

- LUN 數量比 LUN 大小更重要。
- LUN 大小大多由 LUN 數需求控制。
- 避免建立超過所需數量的 LUN 。

LUN 計數

與 LUN 大小不同、 LUN 數量確實會影響效能。應用程式效能通常取決於透過 SCSI 層執行平行 I/O 的能力。因此、兩個 LUN 的效能優於單一 LUN 。使用 LVM （例如 Veritas VxVM 、 Linux LVM2 或 Oracle ASM ）是提高平行度的最簡單方法。

NetApp 客戶通常從 LUN 數量增加到 16 個以上獲得最小的效益、不過測試 100% SSD 環境時、隨機 I/O 非常繁重、這已證實可進一步改善至 64 個 LUN 。

- NetApp 建議 * 下列事項：

一般而言、四到十六個 LUN 足以支援任何特定資料庫工作負載的 I/O 需求。由於主機 SCSI 實作的限制、少於四個 LUN 可能會造成效能限制。

Oracle 資料庫 LUN 放置

資料庫 LUN 在 ONTAP 磁碟區內的最佳放置方式、主要取決於如何使用各種 ONTAP 功能。

磁碟區

與剛接觸 ONTAP 的客戶混淆的一個常見點是使用 FlexVols 、通常稱為「 Volume 」。

磁碟區不是 LUN 。這些詞彙與許多其他廠商產品（包括雲端供應商）同義。ONTAP Volume 是簡單的管理容器。它們本身不會提供資料、也不會佔用空間。它們是檔案或 LUN 的容器、可改善及簡化管理、尤其是大規模管理。

磁碟區和 LUN

相關 LUN 通常位於單一磁碟區中。例如、需要 10 個 LUN 的資料庫通常會將所有 10 個 LUN 放在同一個磁碟區上。

- 使用 LUN 對磁碟區的比例 1 ： 1 表示每個磁碟區有一個 LUN 、這是 * 非 * 正式最佳實務做法。
- 而是應將磁碟區視為工作負載或資料集的容器。每個磁碟區可能只有一個 LUN 、或者可能有許多 LUN 。正確的答案取決於管理需求。
- 在不必要數量的磁碟區之間分散 LUN 、可能會導致額外的額外負荷和排程問題、例如快照作業、 UI 中顯示的物件過多、並導致在達到 LUN 限制之前達到平台磁碟區限制。

磁碟區、 LUN 和快照

Snapshot 原則和排程會放置在磁碟區上、而非 LUN 上。如果資料集由 10 個 LUN 組成、則當這些 LUN 位於同一個磁碟區中時、只需要單一快照原則。

此外、在單一磁碟區中共同定位給定資料集的所有相關 LUN 、可提供原子快照作業。例如、如果基礎 LUN 全部放在單一磁碟區上、則位於 10 個 LUN 上的資料庫、或是由 10 個不同作業系統組成的 VMware 應用程式環境、都可以作為單一且一致的物件加以保護。如果將快照放在不同的磁碟區上、則即使同時排程、快照仍可能保持 100% 同步。

在某些情況下、由於恢復需求、相關的 LUN 集可能需要分割成兩個不同的磁碟區。例如、資料庫可能有四個 LUN 用於資料檔案、兩個 LUN 用於記錄。在這種情況下、具有 4 個 LUN 的資料檔案磁碟區和具有 2 個 LUN 的記錄磁碟區可能是最佳選擇。原因在於可進行的可恢復性是不相關的。例如、資料檔案磁碟區可以選擇性地還原為較早的狀態、這表示所有四個 LUN 都會還原為快照狀態、而記錄磁碟區與其重要資料則不會受到影響。

Volume 、 LUN 和 SnapMirror

SnapMirror 原則和作業就像快照作業一樣、是在磁碟區上執行、而不是在 LUN 上執行。

在單一磁碟區中共同定位相關 LUN 、可讓您建立單一 SnapMirror 關係、並透過單一更新來更新所有包含的資料。與快照一樣、更新也將是一項原子作業。SnapMirror 目的地將保證擁有來源 LUN 的單一時間點複本。如果 LUN 分散在多個磁碟區、則複本可能彼此一致、也可能不一致。

磁碟區、 LUN 和 QoS

雖然 QoS 可以選擇性地套用至個別 LUN 、但通常在磁碟區層級設定 QoS 會比較容易。例如、指定 ESX 伺服器中的來賓所使用的所有 LUN 都可以放置在單一磁碟區上、然後就可以套用 ONTAP 調適性 QoS 原則。結果是將每 TB IOPS 的自我擴充限制套用至所有 LUN 。

同樣地、如果資料庫需要 10 萬次 IOPS 、而且佔用 10 個 LUN 、則在單一磁碟區上設定單一的 10 萬次 IOPS 限制、比在每個 LUN 上設定 10 個個別的 10K IOPS 限制更容易。

多重 Volume 配置

在某些情況下、跨多個磁碟區散佈 LUN 可能會有幫助。主要原因是控制器分段。例如、 HA 儲存系統可能會裝載單一資料庫、其中需要每個控制器的完整處理與快取潛力。在這種情況下、典型的設計是將一半的 LUN 放在控制器 1 的單一磁碟區、而另一半的 LUN 則放在控制器 2 的單一磁碟區中。

同樣地、控制器分段也可用於負載平衡。HA 系統託管 100 個資料庫、每個資料庫各有 10 個 LUN 、每個資料庫可在兩個控制器上接收 5 個 LUN 磁碟區。如此一來、每個控制器就能以對稱的方式進行對稱載入、同時還能配置額外的資料庫。

不過、這些範例都不涉及 1 ： 1 的磁碟區對 LUN 比率。目標仍然是透過在磁碟區中共同定位相關 LUN 來最佳化管理性。

其中一個例子是、 1 ： 1 LUN 對磁碟區比率非常合理、其中每個 LUN 可能真正代表單一工作負載、而且每個工作負載都需要個別管理。在這種情況下、 1 ： 1 的比率可能是最佳的。

Oracle 資料庫 LUN 調整大小和以 LVM 為基礎的調整大小

當 SAN 型檔案系統達到容量上限時、有兩個選項可以增加可用空間：

- 增加 LUN 的大小
- 將 LUN 新增至現有的磁碟區群組、並擴充內含的邏輯磁碟區

雖然 LUN 調整大小是增加容量的選項、但通常最好使用 LVM 、包括 Oracle ASM 。LVM 存在的主要原因之一、是為了避免需要調整 LUN 大小。使用 LVM 時、多個 LUN 會結合在一個虛擬儲存池中。從該池中切出的邏輯卷由 LVM 管理，可以輕鬆調整大小。另一項優點是在所有可用 LUN 之間分配給定的邏輯磁碟區、以避免在特定磁碟機上出現熱點。通常可以使用 Volume Manager 將邏輯磁碟區的基礎範圍重新放置到新的 LUN 、以執行透明移轉。

使用 **Oracle** 資料庫的 **LVM** 分拆

LVM 分拆是指在多個 LUN 之間分配資料。如此一來、許多資料庫的效能大幅提升。

在快閃磁碟機時代之前、使用區塊延展來協助克服旋轉磁碟機的效能限制。例如、如果作業系統需要執行 1MB 讀取作業、則從單一磁碟機讀取 1MB 的資料時、需要大量的磁碟機磁頭搜尋和讀取、因為 1MB 會緩慢傳輸。如果將 1MB 的資料分散在 8 個 LUN 上、則作業系統可能會同時執行 8 個 128K 讀取作業、並縮短完成 1MB 傳輸所需的時間。

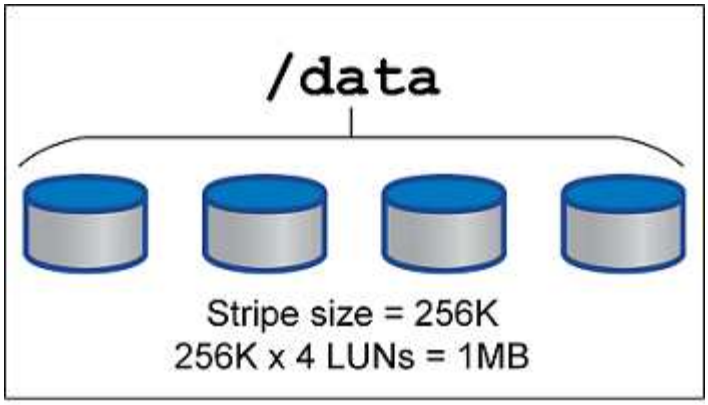
由於必須事先知道 I/O 模式、因此使用旋轉磁碟機進行分拆會更困難。如果串列區塊延展未針對真正的 I/O 模式正確調整、則等量區塊配置可能會損害效能。使用 Oracle 資料庫、特別是搭配 All Flash 組態、分拆作業更容易設定、並經證實可大幅提升效能。

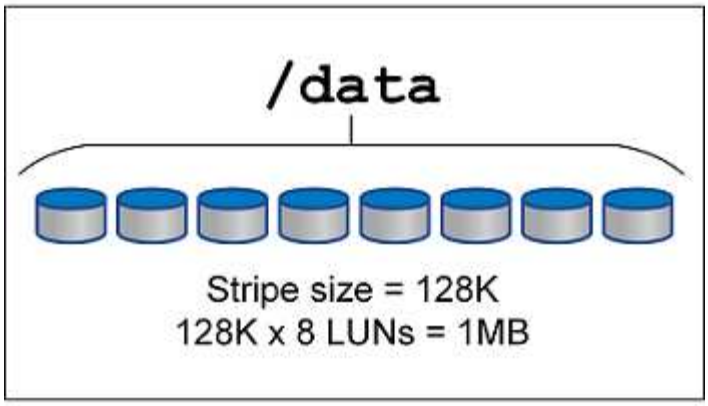
依預設、邏輯磁碟區管理程式（例如 Oracle ASM 等量磁碟區）不屬於原生 OS LVM 。其中有些 LUN 會將多個 LUN 連結在一起、成為串連的裝置、導致資料檔案存在於一台 LUN 裝置上、而只存在於一台 LUN 裝置上。這會造成熱點。其他 LVM 實作預設為分散式擴充。這與分拆類似、但卻是比較粗糙的。磁碟區群組中的 LUN 會切成大型片段、稱為區段、通常以百萬位元組為單位測量、然後邏輯磁碟區會分佈在這些區段中。結果是對檔案進行隨機 I/O 、應該能在 LUN 之間妥善分配、但連續 I/O 作業的效率卻不如以前那麼高。

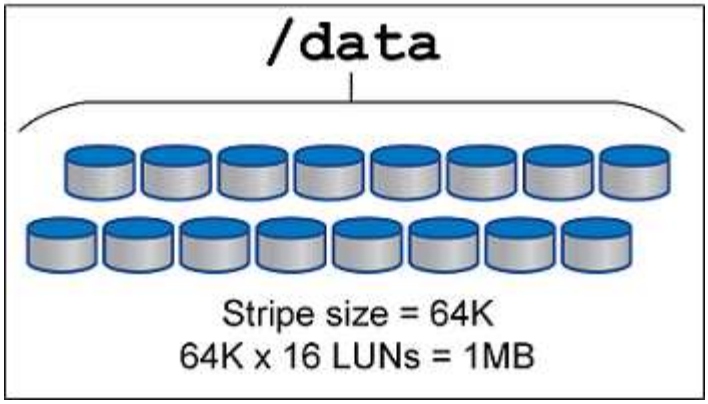
效能密集的應用程式 I/O 幾乎總是（ a ）以基本區塊大小為單位、或（ b ） 1 MB 。

等量分配組態的主要目標是確保單一檔案 I/O 可作為單一單元執行、而多區塊 I/O 的大小應為 1MB 、可在等量磁碟區中的所有 LUN 之間平均平行處理。這表示等量磁碟區大小不得小於資料庫區塊大小、且等量磁碟區大小乘以 LUN 數量應為 1MB 。

下圖顯示等量磁碟區大小和寬度調校的三個可能選項。選擇 LUN 數量以滿足上述效能需求、但在所有情況下、單一等量磁碟區內的總資料為 1MB 。


/data
Stripe size = 256K
256K x 4 LUNs = 1MB


/data
Stripe size = 128K
128K x 8 LUNs = 1MB


/data
Stripe size = 64K
64K x 16 LUNs = 1MB