儲存組態
Enterprise applications

NetApp
May 09, 2024

目錄

儲存組態

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。

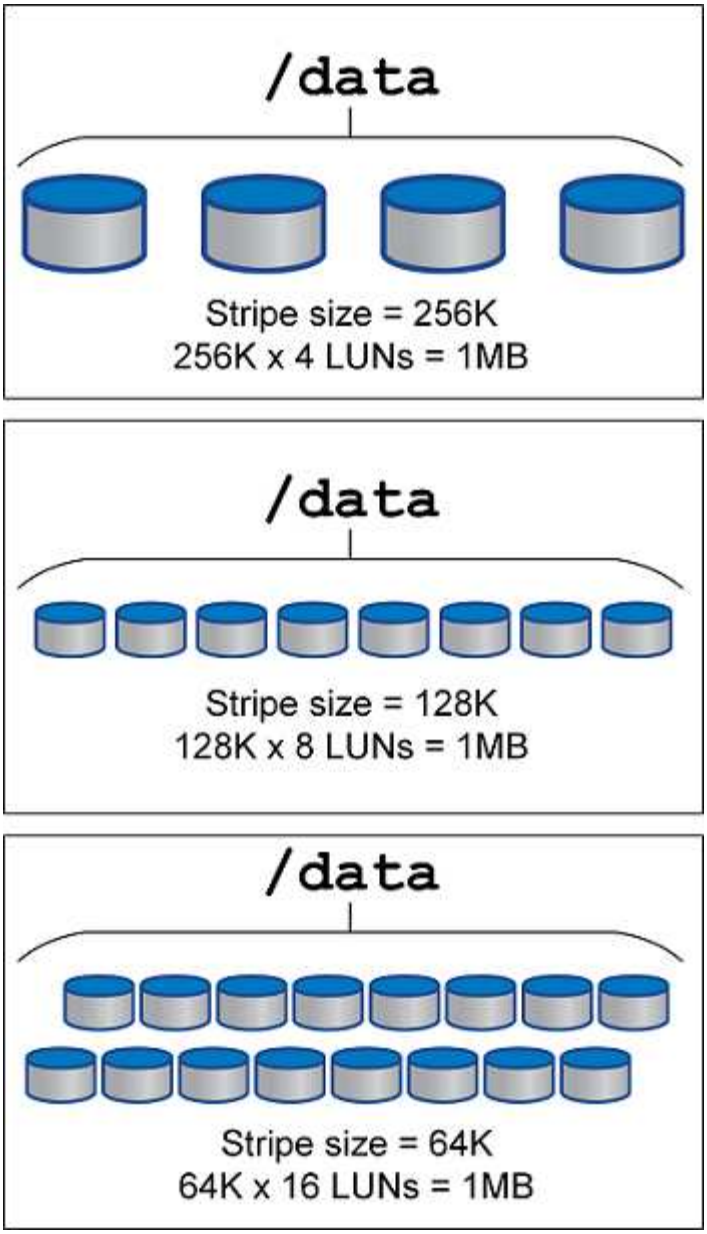
由於必須事先知道 I/O 模式、因此使用旋轉磁碟機進行分拆會更困難。如果串列區塊延展未針對真正的 I/O 模式正確調整、則等量區塊配置可能會損害效能。使用 Oracle 資料庫、特別是搭配 All Flash 組態、分拆作業更容易設定、並經證實可大幅提升效能。

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 。



NFS

Oracle 資料庫的 NFS 組態

NetApp 已提供企業級 NFS 儲存設備超過 30 年、由於其簡易性、隨著雲端型基礎架構的推向、其使用量也不斷增加。

NFS 傳輸協定包含多個不同需求的版本。如需 ONTAP 的 NFS 組態完整說明、請參閱 "TR-4067 ONTAP 最佳實務做法上的 NFS"。下列各節涵蓋一些較重要的需求和一般使用者錯誤。

NFS 版本

NetApp 必須支援作業系統 NFS 用戶端。

- NFSv3 支援符合 NFSv3 標準的作業系統。

- Oracle DNFS 用戶端支援 NFSv3 。
- 所有遵循 NFSv4 標準的作業系統都支援 NFSv4 。
- NFSv4.1 和 NFSv4.2 需要特定的作業系統支援。請參閱 "NetApp IMT" 適用於支援的作業系統。
- Oracle DNFS 支援 NFSv4.1 需要 Oracle 12.2.0.2 或更高版本。

。 "NetApp 支援對照表" 對於 NFSv3 和 NFSv4 、不包含特定的作業系統。一般支援所有遵守 RFC 的作業系統。搜尋線上 IMT 以取得 NFSv3 或 NFSv4 支援時、請勿選取特定的作業系統、因為不會顯示任何相符項目。一般原則隱含支援所有作業系統。

Linux NFSv3 TCP 插槽表

TCP 插槽表是與主機匯流排介面卡（ HBA ）佇列深度相當的 NFSv3 。這些表格可控制任何時間都可以處理的NFS作業數量。預設值通常為16、這對於最佳效能而言太低。相反的問題發生在較新的Linux核心上、這會自動將TCP插槽表格限制增加到要求使NFS伺服器飽和的層級。

為了達到最佳效能並避免效能問題、請調整控制 TCP 插槽表的核心參數。

執行 `sysctl -a | grep tcp.*.slot_table` 並觀察下列參數：

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

所有 Linux 系統都應該包括在內 `sunrpc.tcp_slot_table_entries`但只有部分包含在內 `sunrpc.tcp_max_slot_table_entries`。兩者都應設為 128 。

注意
若未設定這些參數、可能會對效能造成重大影響。在某些情況下、效能會受到限制、因為 Linux 作業系統沒有發出足夠的 I/O在其他情況下、隨著 Linux 作業系統嘗試發出的 I/O 數量超過可服務的數量、 I/O 延遲也會增加。

ADR 和 NFS

部分客戶回報的效能問題是由於中的資料 I/O 過多所造成 `ADR` 位置。問題通常不會在累積許多效能資料之前發生。I/O 過多的原因不明、但此問題似乎是 Oracle 處理程序重複掃描目標目錄以進行變更所致。

移除 `noac` 和/或 `actimeo=0` 掛載選項可執行主機作業系統快取、並降低儲存 I/O 層級。

* NetApp 建議 * 不要放置 `ADR` 檔案系統上的資料 `noac` 或 `actimeo=0` 因為效能問題很可能會發生。獨立 `ADR` 如有必要、請將資料移至不同的掛載點。

NFS-rootonly 和 mount-rootonly

ONTAP 包含一個稱為的 NFS 選項 `nfs-rootonly` 控制伺服器是否接受來自高連接埠的 NFS 流量連線。為了安全起見、只有 root 使用者可以使用低於 1024 的來源連接埠來開啟 TCP/IP 連線、因為這類連接埠通常是保留給作業系統使用、而非使用者處理程序。此限制有助於確保 NFS 流量來自實際的作業系統 NFS 用戶端、而非模擬 NFS 用戶端的惡意程序。Oracle DNFS 用戶端是 userspace 驅動程式、但程序是以 root 執行、因此通常

不需要變更的值 `nfs-rootonly`。連線是從低連接埠建立。

。 `mount-rootonly` 選項僅適用於 NFSv3 。它控制是否從大於 1024 的連接埠接受 RPC 掛載呼叫。使用 DNFS 時、用戶端會再次以 root 執行、因此能夠開啟低於 1024 的連接埠。此參數無效。

透過 NFS 4.0 及更高版本開啟與 DNFS 連線的程序不會以 root 執行、因此需要 1024 以上的連接埠。。 `nfs-rootonly` 參數必須設為停用、 DNFS 才能完成連線。

如果 `nfs-rootonly` 啟用時、結果會在掛載階段開啟 DNFS 連線時暫停。sqlplus 輸出類似於以下內容：

```
SQL>startup
ORACLE instance started.
Total System Global Area 4294963272 bytes
Fixed Size                  8904776 bytes
Variable Size             822083584 bytes
Database Buffers         3456106496 bytes
Redo Buffers                7868416 bytes
```

參數可變更如下：

```
Cluster01::> nfs server modify -nfs-rootonly disabled
```

在極少數情況下、您可能需要將 NFS-rootonly 和 mount-rootonly 都變更為停用。如果伺服器正在管理大量的 TCP 連線、則可能沒有低於 1024 的連接埠可用、而且作業系統必須使用較高的連接埠。需要變更這兩個 ONTAP 參數、才能完成連線。

NFS 匯出原則：超級使用者和 **setuid**

如果 Oracle 二進位檔位於 NFS 共用區、則匯出原則必須包含超級使用者和 setuid 權限。

用於一般檔案服務（例如使用者主目錄）的共享 NFS 匯出通常會佔用 root 使用者。這表示已掛載檔案系統的主機上 root 使用者的要求、會重新對應為具有較低權限的不同使用者。這有助於防止特定伺服器上的根使用者存取共用伺服器上的資料、進而保護資料安全。在共享環境中， setuid 位元也可能是安全性風險。setuid 位元可讓處理程序以不同於使用者的身分執行、而非以使用者的身分執行指令。例如、 root 擁有的 Shell 指令碼搭配 setuid 位元、會以 root 執行。如果其他使用者可以變更該 Shell 指令碼、任何非 root 使用者都可以透過更新指令碼、以 root 身分發出命令。

Oracle 二進位檔包含 root 擁有的檔案、並使用 setuid 位元。如果在 NFS 共用上安裝 Oracle 二進位檔、匯出原則必須包含適當的超級使用者和 setuid 權限。在以下範例中、這兩項規則都包含在內 `allow-suid` 及許可 `superuser` （ root ）使用系統驗證來存取 NFS 用戶端。

```
Cluster01::> export-policy rule show -vserver vserver1 -policyname orabin
-fields allow-suid,superuser
vserver   policyname ruleindex superuser allow-suid
--------- ---------- --------- --------- ----------
vserver1  orabin     1         sys       true
```

NFSv4/4.1 組態

對於大多數應用程式、 NFSv3 和 NFSv4 之間的差異很小。應用程式 I/O 通常是非常簡單的 I/O 、而且不會從 NFSv4 中提供的某些進階功能中獲得顯著效益。較高版本的 NFS 不應從資料庫儲存的角度視為「升級」、而應視為包含其他功能的 NFS 版本。例如、如果需要 Kerberos 隱私模式（ krb5p ）的端點對端安全性、則需要 NFSv4 。

* 如果需要 NFSv4 功能、 NetApp 建議 * 使用 NFSv4.1 。NFSv4.1 中的 NFSv4 傳輸協定有一些功能性增強功能、可改善某些邊緣情況的恢復能力。

切換至 NFSv4 比單純將掛載選項從 ves=3 變更為 ves=4.1 更複雜。如需更完整的 NFSv4 組態與 ONTAP 說明、包括作業系統設定指南、請參閱 "TR-4067 ONTAP 最佳實務做法上的 NFS"。本 TR 的下列各節說明使用 NFSv4 的一些基本要求。

NFSv4 網域

NFSv4/4.1 組態的完整說明已超出本文件的範圍、但常見的問題之一是網域對應不相符。從系統管理員的角度來看、 NFS 檔案系統的行為似乎正常、但應用程式會報告某些檔案的權限和 / 或 setuid 錯誤。在某些情況下、系統管理員不正確地判斷應用程式二進位檔的權限已受損、並在實際問題是網域名稱時執行 chown 或 chmod 命令。

NFSv4 網域名稱是在 ONTAP SVM 上設定：

```
Cluster01::> nfs server show -fields v4-id-domain
vserver   v4-id-domain
--------- ------------
vserver1  my.lab
```

主機上的 NFSv4 網域名稱是在中設定 `/etc/idmap.cfg`

```
[root@host1 etc]# head /etc/idmapd.conf
[General]
#Verbosity = 0
# The following should be set to the local NFSv4 domain name
# The default is the host's DNS domain name.
Domain = my.lab
```

網域名稱必須相符。如果沒有、則會在中顯示類似下列的對應錯誤 `/var/log/messages`：

```
Apr 12 11:43:08 host1 nfsidmap[16298]: nss_getpwnam: name 'root@my.lab'
does not map into domain 'default.com'
```

應用程式二進位檔（例如 Oracle 資料庫二進位檔）包含 root 擁有的具有 setuid 位元的檔案、這表示 NFSv4 網域名稱不相符會導致 Oracle 啟動失敗、並會發出呼叫檔案擁有權或權限的警告 `oradism`、位於 `$ORACLE_HOME/bin` 目錄。其內容應如下所示：

```
[root@host1 etc]# ls -l /orabin/product/19.3.0.0/dbhome_1/bin/oradism
-rwsr-x--- 1 root oinstall 147848 Apr 17  2019
/orabin/product/19.3.0.0/dbhome_1/bin/oradism
```

如果此檔案的擁有權為 nobody 、則可能是 NFSv4 網域對應問題。

```
[root@host1 bin]# ls -l oradism
-rwsr-x--- 1 nobody oinstall 147848 Apr 17  2019 oradism
```

若要修正此問題、請參閱 `/etc/idmap.cfg` 根據 ONTAP 上的 vv4 識別碼網域設定來建立檔案、並確保檔案一致。如果沒有、請進行必要的變更、然後執行 `nfsidmap -c`，然後等待一段時間讓變更傳播。接著、檔案擁有權應正確辨識為 root 。如果使用者嘗試執行 `chown root` 在 NFS 網域設定修正之前、可能需要在這個檔案上執行 `chown root` 再一次。

Oracle directNFS

Oracle 資料庫可使用 NFS 的方式有兩種。

首先、它可以使用以作業系統一部分的原生 NFS 用戶端所掛載的檔案系統。這有時稱為核心 NFS 或 kNFS 。NFS 檔案系統是由 Oracle 資料庫安裝及使用、與任何其他應用程式使用 NFS 檔案系統的方式完全相同。

第二種方法是 Oracle Direct NFS （ DNFS ）。這是在 Oracle 資料庫軟體中實作 NFS 標準。它不會改變 DBA 設定或管理 Oracle 資料庫的方式。只要儲存系統本身有正確的設定、就應該對 DBA 團隊和終端使用者透明使用 DNFS 。

啟用 DNFS 功能的資料庫仍會掛載一般的 NFS 檔案系統。資料庫開啟後、 Oracle 資料庫會開啟一組 TCP/IP 工作階段、並直接執行 NFS 作業。

Direct NFS

Oracle Direct NFS 的主要值是略過主機 NFS 用戶端、並直接在 NFS 伺服器上執行 NFS 檔案作業。啟用它只需要變更 Oracle 磁碟管理程式（ ODM ）程式庫。Oracle 說明文件中提供此程序的說明。

使用 DNFS 可大幅提升 I/O 效能、並降低主機和儲存系統的負載、因為 I/O 是以最有效率的方式執行。

此外、 Oracle DNFS 還包含 * 選項 * 、可用於網路介面多重路徑和容錯功能。例如、兩個 10Gb 介面可以結合在一起、以提供 20Gb 的頻寬。如果某個介面發生故障、則會在另一個介面上重試 I/O 。整體作業與 FC 多重路徑非常類似。多重路徑在數年前是最常見的標準、那就是 1 個乙太網路。10Gb NIC 足以應付大多數 Oracle 工作負載、但如果需要更多 10Gb NIC 、則可加以連結。

使用 DNFS 時、必須安裝 Oracle Doc 1495104.1 中所述的所有修補程式。如果無法安裝修補程式、則必須評估環境、確保該文件中所述的錯誤不會造成問題。在某些情況下、無法安裝所需的修補程式會導致無法使用 DNFS 。

請勿將 DNFS 用於任何類型的循環名稱解析、包括 DNS 、 DDNS 、 NIS 或任何其他方法。這包括 ONTAP 中可用的 DNS 負載平衡功能。當使用 DNFS 的 Oracle 資料庫將主機名稱解析為 IP 位址時、後續查詢時不得變更。這可能會導致 Oracle 資料庫當機、並可能導致資料毀損。

直接 **NFS** 和主機檔案系統存取

使用 DNFS 有時會導致依賴掛載在主機上的可見檔案系統的應用程式或使用者活動發生問題、因為 DNFS 用戶端會從主機作業系統不定期存取檔案系統。DNFS 用戶端可以在不瞭解作業系統的情況下建立、刪除及修改檔案。

使用單一執行個體資料庫的掛載選項時、會啟用檔案和目錄屬性的快取、這也表示目錄內容會快取。因此、DNFS 可以建立檔案、而且在作業系統重新讀取目錄內容和讓使用者看到檔案之前、會有短暫的延遲。這通常不是問題、但在極少數情況下、 SAP BR*Tools 等公用程式可能會發生問題。如果發生這種情況、請變更掛載選項、以使用 Oracle RAC 的建議來解決此問題。這項變更會導致停用所有主機快取。

只有在使用（ a ） DNFS 時才變更掛載選項、且（ b ）檔案可見度延遲所造成的問題。如果未使用 DNFS 、在單一執行個體資料庫上使用 Oracle RAC 掛載選項會導致效能降低。

請參閱附註 `nosharecache` 在中 "Linux NFS 裝載選項" 針對可能產生異常結果的 Linux 特定 DNFS 問題。

Oracle 資料庫和 NFS 會租用和鎖定

NFSv3 無狀態。這實際上意味著 NFS 伺服器（ ONTAP ）無法追蹤哪些檔案系統是掛載的、由誰掛載、或哪些鎖定是真的就位。

ONTAP 確實有一些功能會記錄掛載嘗試、因此您可以知道哪些用戶端可能正在存取資料、而且可能會出現諮詢鎖定、但這項資訊並不保證 100% 完整。無法完成、因為追蹤 NFS 用戶端狀態並非 NFSv3 標準的一部分。

NFSv4 狀態

相反地、 NFSv4 是有狀態的。NFSv4 伺服器會追蹤哪些用戶端正在使用哪些檔案系統、哪些檔案存在、哪些檔案和 / 或檔案區域被鎖定等 這表示 NFSv4 伺服器之間需要定期通訊、才能保持狀態資料最新。

NFS 伺服器所管理的最重要狀態是 NFSv4 鎖定和 NFSv4 租賃、它們彼此之間有很大的關聯。您必須瞭解每個項目本身的運作方式、以及它們彼此之間的關係。

NFSv4 鎖定

有了 NFSv3 、鎖定是建議事項。NFS 用戶端仍可修改或刪除「鎖定」檔案。NFSv3 鎖本身不會過期、必須將其移除。這會造成問題。例如、如果叢集式應用程式會建立 NFSv3 鎖定、而其中一個節點發生故障、您該怎麼做？您可以在仍在運作的節點上對應用程式進行編碼、以移除鎖定、但您如何知道這是安全的？可能是「故障」節點可以運作、但無法與叢集的其他部分通訊？

有了 NFSv4 、鎖定的持續時間有限。只要持有鎖定的用戶端繼續與 NFSv4 伺服器簽入、就不允許其他用戶端取得這些鎖定。如果用戶端無法使用 NFSv4 進行存回、伺服器最終會撤銷鎖定、而其他用戶端則能要求並取得鎖定。

NFSv4 租賃

NFSv4 鎖定與 NFSv4 租用相關聯。當 NFSv4 用戶端與 NFSv4 伺服器建立連線時、它會取得租用。如果用戶端取得鎖定（鎖定類型眾多）、則鎖定會與租用相關聯。

此租用具有定義的逾時時間。根據預設、 ONTAP 會將逾時值設為 30 秒：

```
Cluster01::*> nfs server show -vserver vserver1 -fields v4-lease-seconds

vserver   v4-lease-seconds
--------- ----------------
vserver1  30
```

這表示 NFSv4 用戶端需要每 30 秒與 NFSv4 伺服器簽入一次、才能續約。

租賃會自動由任何活動續約、因此如果用戶端正在工作、就不需要執行額外作業。如果某個應用程式變得很安靜、而且沒有真正的工作、則需要改為執行某種保持活動狀態的作業（稱為順序）。基本上只是說：「我還在這裏、請重新整理我的租約。」

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
NFSv3 無狀態。這並不需要用戶端的通訊。NFSv4
可設定狀態、一旦租用期間結束、租用即會過期、鎖定會被撤銷、而鎖定的檔案會提供給其他用戶端使用。
```

有了 NFSv3 、您可以四處移動網路纜線、重新啟動網路交換器、進行組態變更、並確保不會發生任何問題。應用程式通常只會耐心等待網路連線再次運作。

有了 NFSv4 、您有 30 秒的時間（除非您已在 ONTAP 中增加該參數的值）來完成工作。如果您超過此上限、您的租約將會逾時。這通常會導致應用程式當機。

舉例來說、如果您有 Oracle 資料庫、而且網路連線中斷（有時稱為「網路分割區」）超過租用逾時、您就會使資料庫當機。

以下是 Oracle 警示記錄中發生這種情況的範例：

```
2022-10-11T15:52:55.206231-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00202: control file: '/redo0/NTAP/ctrl/control01.ctl'
ORA-27072: File I/O error
Linux-x86_64 Error: 5: Input/output error
Additional information: 4
Additional information: 1
Additional information: 4294967295
2022-10-11T15:52:59.842508-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00206: error in writing (block 3, # blocks 1) of control file
ORA-00202: control file: '/redo1/NTAP/ctrl/control02.ctl'
ORA-27061: waiting for async I/Os failed
```

如果您查看系統記錄檔、您應該會看到以下幾個錯誤：

```
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
```

記錄訊息通常是問題的第一個徵象、而非應用程式凍結。通常、在網路中斷期間、您完全看不到任何內容、因為程序和作業系統本身都遭到封鎖、無法嘗試存取 NFS 檔案系統。

網路重新運作後、就會出現錯誤。在上述範例中、一旦重新建立連線、作業系統就會嘗試重新取得鎖定、但時間太晚了。租約已到期、鎖定已移除。這會導致一個錯誤、該錯誤會傳播到 Oracle 層、並導致警示記錄中出現訊息。根據資料庫的版本和組態、您可能會看到這些模式的變化。

總之、 NFSv3 可容忍網路中斷、但 NFSv4 更敏感、並規定了一段定義的租用期。

如果無法接受 30 秒的逾時、該怎麼辦？如果您管理一個動態變化的網路、在其中重新啟動交換器或重新放置纜線、導致網路偶爾中斷、該怎麼辦？您可以選擇延長租用期、但是否需要說明 NFSv4 寬限期。

NFSv4 寬限期

如果 NFSv3 伺服器重新開機、幾乎可以立即為 IO 服務。它並未維持任何形式的用戶端狀態。結果是、 ONTAP 接管作業通常似乎接近即時。當控制器準備好開始提供資料時、就會傳送 ARP 給網路、以表示拓撲的變化。客戶端通常幾乎立即檢測到這種情況，數據恢復流動。

不過 NFSv4 會短暫暫停。這只是 NFSv4 運作方式的一部分。

NFSv4 伺服器需要追蹤租用、鎖定、以及使用何種資料的人員。如果 NFS 伺服器出現問題並重新開機、或停電一段時間、或在維護活動期間重新啟動、則會導致租約 / 鎖定、而其他用戶端資訊也會遺失。伺服器需要先找出哪個用戶端正在使用哪些資料、才能恢復作業。這就是寬限期的開始。

如果您突然關閉 NFSv4 伺服器的電源、當恢復 IO 時、嘗試恢復 IO 的用戶端會收到回應、基本上說：「我遺失了租用 / 鎖定資訊。您是否要重新登錄鎖定？」 這就是寬限期的開始。ONTAP 預設為 45 秒：

```
Cluster01::> nfs server show -vserver vserver1 -fields v4-grace-seconds

vserver   v4-grace-seconds
--------- ----------------
vserver1  45
```

結果是、在重新啟動之後、控制器會暫停 IO 、而所有用戶端都會回收租用和鎖定。寬限期結束後、伺服器將恢復 IO 作業。

租用逾時與寬限期比較

寬限期與租用期間已連線。如上所述、預設的租用逾時為 30 秒、這表示 NFSv4 用戶端必須至少每 30 秒與伺服器簽入一次、否則就會遺失租約、進而導致鎖定。存在寬限期、可讓 NFS 伺服器重建租用 / 鎖定資料、預設為 45 秒。ONTAP 要求寬限期比租用期長 15 秒。如此可確保設計為至少每 30 秒續約的 NFS 用戶端環境、在重新啟動後能夠與伺服器簽入。45 秒的寬限期可確保所有預期至少每 30 秒續約一次的客戶都有機會續約。

如果無法接受 30 秒的逾時、您可以選擇延長租用期。如果您想要將租用逾時延長至 60 秒、以便承受 60 秒的網路中斷、您必須將寬限期延長至至少 75 秒。ONTAP 要求比租用期高 15 秒。這表示您將會在控制器容錯移轉期間經歷更長的 IO 暫停時間。

這通常不是問題。一般使用者每年只會更新 ONTAP 控制器一或兩次、而且由於硬體故障而造成的非計畫性容錯移轉極少。此外、如果您的網路發生 60 秒網路中斷的可能性、而您需要將租用逾時時間延長至 60 秒、那麼您可能不會反對罕見的儲存系統容錯移轉、導致暫停時間也達 75 秒。您已確認網路暫停超過 60 秒、而且速度較快。

使用 **Oracle** 資料庫進行 **NFS** 快取

如果存在下列任一掛載選項、則會停用主機快取：

```
cio, actimeo=0, noac, forcedirectio
```

這些設定可能會嚴重影響軟體安裝、修補及備份 / 還原作業的速度。在某些情況下、尤其是叢集式應用程式、這些選項是必要的、因為必須在叢集中的所有節點之間提供快取一致性。在其他情況下、客戶誤用這些參數、結果是不必要的效能損害。

許多客戶在安裝或修補應用程式二進位檔時、會暫時移除這些掛載選項。如果使用者在安裝或修補程序過程中確認沒有其他處理程序正在使用目標目錄、則可安全地執行此移除。

Oracle 資料庫的 **NFS** 傳輸大小

根據預設、 ONTAP 將 NFS I/O 大小限制為 64K 。

大多數應用程式和資料庫的隨機 I/O 使用的區塊大小要小得多、遠低於 64K 的最大值。大型區塊 I/O 通常是平行處理的、因此 64K 的最大值也不是取得最大頻寬的限制。

有些工作負載的上限為 64K 、因此會造成限制。特別是、如果資料庫執行的 I/O 數量較少、但容量較大、則備份或還原作業或資料庫完整表格掃描等單執行緒作業、會更快、更有效率地執行。ONTAP 的最佳 I/O 處理大小為 256k 。

指定 ONTAP SVM 的最大傳輸大小可變更如下：

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

注意
切勿將 ONTAP 上允許的傳輸大小上限降至低於目前掛載之 NFS 檔案系統的 rsize/wsize 值。這可能會在某些作業系統中造成當機或甚至資料毀損。例如、如果 NFS 用戶端目前設定為 rsize/wsize 65536 、則 ONTAP 最大傳輸大小可在 65536 到 1048576 之間調整、因為用戶端本身受到限制、因此沒有任何影響。將傳輸大小上限降至 65536 以下可能會損害可用度或資料。

Oracle 資料庫與 NVFAIL

NVFAIL 是 ONTAP 中的一項功能、可確保災難性容錯移轉案例期間的完整性。

資料庫在儲存設備容錯移轉事件期間容易受損、因為它們會維持大型內部快取。如果災難性事件需要強制 ONTAP 容錯移轉或強制 MetroCluster 切換、無論整體組態的健全狀況為何、都可能會有效捨棄先前確認的變更。儲存陣列的內容會及時向後跳轉、而且資料庫快取的狀態不再反映磁碟上資料的狀態。這種不一致會導致資料毀損。

快取可能發生在應用程式或伺服器層。例如、 Oracle Real Application Cluster （ RAC ）組態、主站台和遠端站台上的伺服器都處於作用中狀態、可在 Oracle SGA 中快取資料。強制切入作業會導致資料遺失、因此資料庫可能會發生毀損、因為儲存在 SGA 中的區塊可能與磁碟上的區塊不符。

較不明顯的快取用途是在作業系統檔案系統層。來自掛載 NFS 檔案系統的區塊可能會快取到作業系統中。或者、以位於主要站台上的 LUN 為基礎的叢集式檔案系統、可以掛載到遠端站台的伺服器上、然後再次快取資料。在這些情況下、 NVRAM 故障或強制接管或強制性的作業系統、可能會導致檔案系統毀損。

ONTAP 使用 NVFAIL 及其相關設定、保護資料庫和作業系統不受此案例影響。

ASM 回收公用程式和 ONTAP 零區塊偵測

啟用即時壓縮時、 ONTAP 可有效移除寫入檔案或 LUN 的歸零區塊。Oracle ASM 回收公用程式（ ASRU ）等公用程式的運作方式是將零寫入未使用的 ASM 範圍。

這可讓 DBA 在資料刪除後回收儲存陣列上的空間。ONTAP 會攔截零並取消分配 LUN 的空間。回收程序非常快速、因為儲存系統中沒有寫入資料。

從資料庫的角度來看、 ASM 磁碟群組包含零、讀取 LUN 的這些區域會產生零串流、但 ONTAP 不會將零儲存在磁碟機上。而是進行簡單的中繼資料變更、在內部將 LUN 的歸零區域標記為任何資料的空白。

由於類似的原因、涉及零位資料的效能測試無效、因為零區塊實際上並未在儲存陣列內以寫入方式處理。

使用 ASRU 時、請確定已安裝所有 Oracle 建議的修補程式。