主機組態

Enterprise applications

NetApp
May 09, 2024

目錄

主機組態

使用 IBM AIX 的 Oracle 資料庫

IBM AIX 與 ONTAP 上 Oracle 資料庫的組態主題。

並行 I/O

若要在 IBM AIX 上達到最佳效能、必須同時使用 I/O如果沒有並行 I/O 、效能限制很可能是因為 AIX 執行序列化的原子 I/O 、這會產生重大的負荷。

NetApp 最初建議使用 `cio` 掛載選項可強制在檔案系統上使用並行 I/O 、但此程序有缺點、不再需要。自從推出 AIX 5.2 和 Oracle 10gR1 之後、 AIX 上的 Oracle 就可以開啟個別檔案來同時執行 IO 、而不是強制在整個檔案系統上同時執行 I/O 。

啟用並行 I/O 的最佳方法是設定 `init.ora` 參數 `filesystemio_options` 至 `setall`。這樣做可讓 Oracle 開啟特定檔案、以與並行 I/O 搭配使用

使用 `cio` 作爲掛載選項，強制使用並行 I/O ，這可能會產生負面影響。例如、強制並行 I/O 會停用檔案系統上的預先讀取、這可能會損害 Oracle 資料庫軟體以外的 I/O 效能、例如複製檔案和執行磁帶備份。此外、 Oracle GoldenGate 和 SAP BR* Tools 等產品與使用不相容 `cio` 裝載選項搭配特定版本的 Oracle 。

- NetApp 建議 * 下列事項：
- 請勿使用 `cio` 檔案系統層級的掛載選項。而是透過使用來啟用並行 I/O `filesystemio_options=setall`。
- 請僅使用 `cio` 如果無法設定掛載選項、則應選擇掛載選項 `filesystemio_options=setall`。

AIX NFS 裝載選項

下表列出 Oracle 單一執行個體資料庫的 AIX NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
控制檔 資料檔案 重作記錄	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

下表列出 RAC 的 AIX NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144
控制檔 資料檔案 重作記錄	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac
CRS/Voting	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac
專屬 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144
共享 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr

單一執行個體與 RAC 掛載選項之間的主要差異在於新增 noac 至掛載選項。此新增功能可停用主機作業系統快取、使 RAC 叢集中的所有執行個體都能一致地檢視資料狀態。

雖然使用 cio 裝載選項和 init.ora 參數 filesystemio_options=setall 停用主機快取的效果相同、仍需使用 noac。 noac 為共享的必要項目 ORACLE_HOME 部署以促進 Oracle 密碼檔案和等檔案的一致性 spfile 參數檔。如果 RAC 叢集中的每個執行個體都有專用的 ORACLE_HOME，則不需要此參數。

AIX jfs/JFS2 掛載選項

下表列出 AIX jfs/JFS2 掛載選項。

檔案類型	掛載選項
ADR 首頁	預設值
控制檔 資料檔案 重作記錄	預設值
Oracle_Home	預設值

使用 AIX 之前 hdisk 任何環境中的裝置（包括資料庫）都請檢查參數 queue_depth。此參數不是 HBA 佇列深度、而是與個別主機的 SCSI 佇列深度相關 hdisk device. Depending on how the LUNs are configured, the value for `queue_depth 效能可能太低。測試顯示最佳值為 64 。

使用 HP-UX 的 Oracle 資料庫

適用於 HP-UX with ONTAP 上 Oracle 資料庫的組態主題。

HP-UX NFS 掛載選項

下表列出單一執行個體的 HP-UX NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
控制檔 資料檔案 重作記錄	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,forcedirectio, nointr,suid
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid

下表列出適用於 RAC 的 HP-UX NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,noac,suid
控制檔 資料檔案 重作記錄	rw, bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio,suid
CRS/ 投票	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac, forcedirectio,suid
專屬 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
共享 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid

單一執行個體與 RAC 掛載選項之間的主要差異在於新增 `noac` 和 `forcedirectio` 至掛載選項。此新增功能可停用主機作業系統快取、使 RAC 叢集中的所有執行個體都能一致地檢視資料狀態。雖然使用 `init.ora` 參數 `filesystemio_options=setall` 停用主機快取的效果相同、仍需使用 `noac` 和 `forcedirectio`。

原因 `noac` 為共享的必要項目 `ORACLE_HOME` 部署是為了促進檔案的一致性、例如 Oracle 密碼檔案和 spfiles 。如果 RAC 叢集中的每個執行個體都有專用的 `ORACLE_HOME`，不需要此參數。

HP-UX VxFS 掛載選項

對於託管 Oracle 二進位檔的檔案系統、請使用下列掛載選項：

```
delaylog,nodatainlog
```

對於包含資料檔案、重做記錄檔、歸檔記錄檔和控制檔的檔案系統、若 HP-UX 版本不支援並行 I/O 、請使用下

列掛載選項：

```
nodatainlog,mincache=direct,convosync=direct
```

支援並行 I/O （ VxFS 5.0.1 及更新版本、或 ServiceGuard Storage Management Suite ）時、請針對包含資料檔案、重作記錄檔、歸檔記錄檔和控制檔的檔案系統、使用這些掛載選項：

```
delaylog,cio
```

參數 `db_file_multiblock_read_count` 在 VxFS 環境中尤其重要。Oracle 建議在 Oracle 10g R1 及更新版本中保留此參數、除非另有特別指示。Oracle 8KB 區塊大小的預設值為 128 。如果此參數的值強制為 16 或更少、請移除 `convosync=direct` 裝載選項、因為它可能會損害連續 I/O 效能。此步驟會損害其他效能層面、只有在的價值下才應採取 `db_file_multiblock_read_count` 必須從預設值變更。

使用 Linux 的 Oracle 資料庫

Linux 作業系統專屬的組態主題。

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。

Linux NFS 裝載選項

下表列出單一執行個體的 Linux NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144`
控制檔 資料檔案 重作記錄	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr`
Oracle_Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr`

下表列出 RAC 的 Linux NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,actimeo=0`
控制檔 資料檔案 重作記錄	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,actimeo=0`
CRS/ 投票	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,actimeo=0`
專屬 `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144`
共享 `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,actimeo=0`

單一執行個體與 RAC 掛載選項之間的主要差異在於新增 `actimeo=0` 至掛載選項。此新增功能可停用主機作業系統快取、使 RAC 叢集中的所有執行個體都能一致地檢視資料狀態。雖然使用 `init.ora` 參數 `filesystemio_options=setall` 停用主機快取的效果相同、仍需使用 `actimeo=0`。

原因 `actimeo=0` 為共享的必要項目 `ORACLE_HOME` 部署是為了促進檔案的一致性、例如 Oracle 密碼檔案和 spfiles 。如果 RAC 叢集中的每個執行個體都有專用的 `ORACLE_HOME`，則不需要此參數。

一般而言、非資料庫檔案的掛載應該與單一執行個體資料檔案所使用的選項相同、不過特定應用程式可能有不同的需求。避免使用掛載選項 `noac` 和 `actimeo=0` 如果可能、因為這些選項會停用檔案系統層級的預先讀取和緩衝處理。這可能會對擷取、轉譯和載入等程序造成嚴重的效能問題。

存取與 **GetAttr**

有些客戶指出、存取和 GetAttr 等極高層級的其他 IOPS 、可能會主導他們的工作負載。在極端情況下、讀取和寫入等作業可低於總作業量的 10% 。這是任何包含使用的資料庫的正常行為 `actimeo=0` 和/或 `noac` 在 Linux 上、因為這些選項會導致 Linux 作業系統持續從儲存系統重新載入檔案中繼資料。存取和 GetAttr 等作業是從資料庫環境中的 ONTAP 快取提供服務的低影響作業。它們不應被視為真正的 IOPS 、例如讀寫、而會對儲存系統產生真正的需求。不過、這些其他 IOPS 確實會產生一些負載、尤其是在 RAC 環境中。若要解決這種情況、請啟用 DNFS 、以略過作業系統緩衝區快取、避免這些不必要的中繼資料作業。

Linux Direct NFS

另一個掛載選項、稱為 `nosharecache`（ a ） DNFS 啟用時、需要使用、（ b ）來源磁碟區多次裝載於具有巢狀 NFS 裝載的單一伺服器（ c ）上。此組態主要出現在支援 SAP 應用程式的環境中。例如、 NetApp 系統上的單一磁碟區可能會有位於的目錄 `/vol/oracle/base` 再來一次 `/vol/oracle/home`。如果 `/vol/oracle/base` 安裝於 `/oracle` 和 `/vol/oracle/home` 安裝於 `/oracle/home`，結果是來自相同來源的巢狀 NFS 掛載。

作業系統可以偵測到這個事實 `/oracle` 和 `/oracle/home` 位於同一個磁碟區、即相同的來源檔案系統。作業系統接著會使用相同的裝置控制代碼來存取資料。這樣做可以改善 OS 快取和某些其他作業的使用、但會干擾 DNFS 。如果 DNFS 必須存取檔案、例如 `spfile`、開啟 `/oracle/home`，它可能會錯誤地嘗試使用錯誤的資料路徑。結果是 I/O 作業失敗。在這些組態中、新增 `nosharecache` 裝載選項至任何與該主機上其他 NFS 檔案系統共用來源 FlexVol 磁碟區的 NFS 檔案系統。這樣做會強制 Linux 作業系統為該檔案系統分配獨立的裝置控制代碼。

Linux Direct NFS 和 Oracle RAC

使用 DNFS 對 Linux 作業系統上的 Oracle RAC 有特殊的效能優勢、因為 Linux 沒有強制直接 I/O 的方法、而 RAC 需要此方法才能在節點之間保持一致。因應措施是 Linux 需要使用 `actimeo=0` 掛載選項、會使檔案資料從作業系統快取立即過期。此選項會強制 Linux NFS 用戶端持續重新讀取屬性資料、進而損害延遲並增加儲存控制器的負載。

啟用 DNFS 會略過主機 NFS 用戶端、避免此損害。多家客戶在啟用 DNFS 時、報告 RAC 叢集的效能大幅提升、 ONTAP 負載大幅降低（特別是其他 IOPS ）。

Linux Direct NFS 和 oranfstab 檔案

在 Linux 上搭配多重路徑選項使用 DNFS 時、必須使用多個子網路。在其他作業系統上、可使用建立多個 DNFS 通道 `LOCAL` 和 `DONTROUTE` 可在單一子網路上設定多個 DNFS 通道的選項。不過、這在 Linux 上無法正常運作、因此可能會產生非預期的效能問題。在 Linux 中、用於 DNFS 流量的每個 NIC 都必須位於不同的子網路上。

I/O 排程器

Linux 核心可讓您以低層級控制 I/O 排程封鎖裝置的方式。Linux 各版本的預設值差異極大。測試顯示、截止日期通常會提供最佳結果、但有時 NOOP 會稍微好一點。效能差異最小、但如果需要從資料庫組態擷取最大可能效能、請測試這兩個選項。在許多組態中、 CFQ 是預設值、而且已證明資料庫工作負載的效能有重大問題。

如需設定 I/O 排程器的指示、請參閱相關的 Linux 廠商文件。

多重路徑

部分客戶在網路中斷期間遭遇當機、因為多重路徑常駐程式未在其系統上執行。在最新版本的 Linux 上、作業系統的安裝程序和多重路徑常駐程式可能會讓這些作業系統容易受到此問題的影響。套件已正確安裝、但未設定為在重新開機後自動啟動。

例如、 RHEL5.5 上多重路徑常駐程式的預設值可能如下所示：

```
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:off   3:off   4:off   5:off   6:off
```

您可以使用下列命令來修正此問題：

```
[root@host1 iscsi]# chkconfig multipathd on
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

ASM 鏡像

ASM鏡射可能需要變更Linux多重路徑設定、以允許ASM辨識問題並切換至其他故障群組。大部分關於「不完整」的ASM組態ONTAP 都使用外部備援、這表示資料保護是由外部陣列提供、而ASM不會鏡射資料。某些站台使用具有一般備援的ASM來提供雙向鏡像、通常是跨不同站台。

中顯示的 Linux 設定 "NetApp 主機公用程式文件" 包含會導致 I/O 無限期佇列的多重路徑參數這表示 LUN 裝置上沒有作用中路徑的 I/O 會在 I/O 完成所需的時間內等待。這通常是很理想的做法、因為 Linux 主機會在 SAN 路徑變更完成、 FC 交換器重新開機或儲存系統完成容錯移轉所需的時間內等待。

這種不受限制的佇列行為會導致 ASM 鏡像發生問題、因為 ASM 必須收到 I/O 故障、才能在替代 LUN 上重試 I/O 。

在 Linux 中設定下列參數 `multipath.conf` 用於 ASM 鏡像的 ASM LUN 檔案：

```
polling_interval 5
no_path_retry 24
```

這些設定會為 ASM 裝置建立 120 秒的逾時。逾時會計算為 `polling_interval` * `no_path_retry` 秒。在某些情況下可能需要調整確切的值、但 120 秒的逾時時間應足以滿足大部分的使用需求。具體而言、 120 秒的時間應該能讓控制器接管或恢復、而不會產生 I/O 錯誤、導致故障群組離線。

較低 `no_path_retry` 此值可縮短 ASM 切換至替代故障群組所需的時間、但這也會增加在維護活動（例如控制器接管）期間不必要的容錯移轉風險。仔細監控 ASM 鏡像狀態、即可降低風險。如果發生不必要的容錯移轉、只要執行重新同步的速度相對較快、鏡像就能快速重新同步。如需更多資訊、請參閱 ASM Fast Mirror Resync 上的 Oracle 說明文件、以瞭解所使用的 Oracle 軟體版本。

Linux xfs 、 ext3 和 ext4 掛載選項

* NetApp 建議 * 使用預設掛載選項。

使用 ASMLib/AFD 的 Oracle 資料庫（ ASM 篩選器驅動程式）

使用 AFD 和 ASMLib 的 Linux 作業系統專屬組態主題

ASMLib 區塊大小

ASMLib 是選用的 ASM 管理程式庫和相關公用程式。其主要值是將 LUN 或 NFS 型檔案標記為具有人類可讀標籤的 ASM 資源。

ASMLib 的最新版本會偵測稱為每個實體區塊指數（ LBPPBE ）的邏輯區塊的 LUN 參數。ONTAP SCSI 目標直

到最近才回報此值。現在會傳回一個值、表示偏好 4KB 區塊大小。這不是區塊大小的定義、但它是使用 LBPPBE 的任何應用程式的提示、可能會更有效率地處理特定大小的 I/O 。不過、 ASMLib 會將 LBPPBE 解譯為區塊大小、並在建立 ASM 裝置時持續標記 ASM 標頭。

此程序可能會以多種方式造成升級和移轉問題、全部是因為無法在同一個 ASM 磁碟群組中混合使用不同區塊大小的 ASMLib 裝置。

例如、較舊的陣列通常回報 LBPPBE 值為 0 、或根本沒有回報此值。ASMLib 會將此解譯為 512 位元組的區塊大小。較新的陣列會被解譯為具有 4KB 區塊大小。無法在同一個 ASM 磁碟群組中混合使用 512 位元組和 4KB 的裝置。這樣做會阻止用戶使用兩個陣列中的 LUN 或使用 ASM 作爲遷移工具來增加 ASM 磁盤組的大小。在其他情況下、 RMAN 可能不允許在具有 512 位元組區塊大小的 ASM 磁碟群組和具有 4KB 區塊大小的 ASM 磁碟群組之間複製檔案。

首選的解決方案是修補 ASMLib 。Oracle 錯誤 ID 為 13999609 、而 Oracle 錯誤 ID 則存在於 oracleas-support-2.1.8-1 及更高版本中。此修補程式可讓使用者設定參數 `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` 至 `true` 在中 `/etc/sysconfig/oracleasm` 組態檔。這樣做會阻止 ASMLib 使用 LBPPBE 參數、這表示新陣列上的 LUN 現在會被辨識為 512 位元組區塊裝置。

> ⓘ 此選項不會變更先前由 ASMLib 戳記的 LUN 區塊大小。例如、如果具有 512 位元組區塊的 ASM 磁碟群組必須移轉至回報 4KB 區塊的新儲存系統、則選項 `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` 必須先設定、才能使用 ASMLib 標記新的 LUN 。 如果裝置已被 oracleasm 戳記、則必須先重新格式化、然後再重新設定新的區塊大小。首先、請使用取消設定裝置 `oracleasm deletedisk`、然後使用清除裝置的前 1GB `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`。最後、如果裝置先前已分割、請使用 `kpartx` 命令移除過時的分割區、或只是重新開機作業系統。

如果無法修補 ASMLib 、可以從組態中移除 ASMLib 。這項變更會造成中斷、需要在 ASM 磁碟上加蓋戳記、並確定 `asm_diskstring` 參數設定正確。不過、這項變更並不需要移轉資料。

ASM Filter Drive （ AFD ）區塊大小

AFD 是選用的 ASM 管理程式庫、正在取代 ASMLib 。從儲存角度來看、它與 ASMLib 非常類似、但它還包含其他功能、例如能夠封鎖非 Oracle I/O 、以降低使用者或應用程式錯誤可能毀損資料的機會。

裝置區塊大小

如同 ASMLib 、 AFD 也會讀取 LUN 參數每個實體區塊指數（ LBPPBE ）的邏輯區塊、並依預設使用實體區塊大小、而非邏輯區塊大小。

如果將 AFD 新增至現有組態、而 ASM 裝置已格式化為 512 位元組區塊裝置、則可能會造成問題。AFD 驅動程式會將 LUN 辨識為 4K 裝置、而 ASM 標籤與實體裝置之間的不符將會妨礙存取。同樣地、移轉也會受到影響、因為無法在同一個 ASM 磁碟群組中混合使用 512 位元組和 4KB 的裝置。這樣做會阻止用戶使用兩個陣列中的 LUN 或使用 ASM 作爲遷移工具來增加 ASM 磁盤組的大小。在其他情況下、 RMAN 可能不允許在具有 512 位元組區塊大小的 ASM 磁碟群組和具有 4KB 區塊大小的 ASM 磁碟群組之間複製檔案。

解決方案很簡單 - AFD 包含一個參數、可控制它是否使用邏輯區塊或實體區塊大小。這是影響系統上所有裝置的全域參數。若要強制 AFD 使用邏輯區塊大小、請設定 `options oracleafd oracleafd_use_logical_block_size=1` 在中 `/etc/modprobe.d/oracleafd.conf` 檔案：

多重路徑傳輸大小

最近的 Linux 核心變更會強制執行傳送至多重路徑裝置的 I/O 大小限制、而 AFD 則不遵守這些限制。然後會拒

絕 I/O 、導致 LUN 路徑離線。結果是無法安裝 Oracle Grid 、設定 ASM 或建立資料庫。

解決方案是在 ONTAP LUN 的 multipath.conf 檔案中手動指定傳輸長度上限：

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

即使目前沒有問題、如果使用 AFD 來確保未來的 Linux 升級不會意外造成問題、也應設定此參數。

使用 Microsoft Windows 的 Oracle 資料庫

Microsoft Windows with ONTAP 上 Oracle 資料庫的組態主題。

NFS

Oracle 支援搭配直接 NFS 用戶端使用 Microsoft Windows 。這項功能提供 NFS 管理效益的途徑、包括跨環境檢視檔案、動態調整磁碟區大小、以及使用較便宜的 IP 傳輸協定。如需在 Microsoft Windows 上使用 DNFS 安裝及設定資料庫的詳細資訊、請參閱正式的 Oracle 文件。不存在任何特殊的最佳實務做法。

SAN

為達到最佳壓縮效率、請確保 NTFS 檔案系統使用 8K 或更大的分配單元。使用 4K 分配單元（通常是預設）會對壓縮效率造成負面影響。

Oracle 資料庫與 Solaris

特定於 Solaris OS 的組態主題。

Solaris NFS 掛載選項

下表列出單一執行個體的 Solaris NFS 掛載選項。

檔案類型	掛載選項
ADR 首頁	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
控制檔 資料檔案 重作記錄	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`

檔案類型	掛載選項
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid

的用途 llock 已獲證實、可移除在儲存系統上取得和釋放鎖定的相關延遲、大幅提升客戶環境的效能。在將多個伺服器設定為掛載相同檔案系統的環境中、請謹慎使用此選項、並將 Oracle 設定為掛載這些資料庫。雖然這是非常不尋常的組態、但只有少數客戶使用。如果第二次意外啟動某個執行個體、可能會因為 Oracle 無法偵測到外部伺服器上的鎖定檔案而導致資料毀損。NFS 鎖定不會提供保護；如同 NFS 第 3 版一樣、它們只是建議事項。

因為 llock 和 forcedirectio 參數是互斥的、這一點很重要 filesystemio_options=setall 存在於中 init.ora 檔案就是這樣 directio 已使用。如果沒有此參數、就會使用主機作業系統緩衝區快取、而且效能可能會受到負面影響。

下表列出了 Solaris NFS RAC 掛載選項。

檔案類型	掛載選項
ADR 首頁	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,noac
控制檔 資料檔案 重作記錄	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
CRS/ 投票	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
專屬 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
共享 ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid

單一執行個體與 RAC 掛載選項之間的主要差異在於新增 noac 和 forcedirectio 至掛載選項。此新增功能可停用主機作業系統快取、使 RAC 叢集中的所有執行個體都能一致地檢視資料狀態。雖然使用 init.ora 參數 filesystemio_options=setall 停用主機快取的效果相同、仍需使用 noac 和 forcedirectio。

原因 actimeo=0 為共享的必要項目 ORACLE_HOME 部署是為了促進檔案的一致性、例如 Oracle 密碼檔案和 spfiles 。如果 RAC 叢集中的每個執行個體都有專用的 ORACLE_HOME，不需要此參數。

Solaris UFS 掛載選項

NetApp 強烈建議您使用記錄掛載選項、以便在 Solaris 主機當機或 FC 連線中斷時保留資料完整性。記錄掛載選項也可保留 Snapshot 備份的使用性。

Solaris ZFS

必須仔細安裝和設定 Solaris ZFS 、才能提供最佳效能。

mvector

Solaris 11 變更了 IT 處理大型 I/O 作業的方式、可能會在 SAN 儲存陣列上造成嚴重的效能問題。NetApp 錯誤報告 630173 「 Solaris 11 ZFS 效能回歸」中詳細說明了此問題。" 解決方案是變更名為的 OS 參數 `zfs_mvector_max_size`。

以 root 執行下列命令：

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

如果這項變更發生任何非預期的問題、您可以以 root 執行下列命令、輕鬆地將其還原：

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

核心

可靠的 ZFS 效能需要修補 Solaris 核心、以因應 LUN 對齊問題。此修正程式是隨 Solaris 10 中的修補程式 147440-19 和適用於 Solaris 11 的 SRU 10.5 一起推出的。只能將 Solaris 10 及更新版本與 ZFS 搭配使用。

LUN 組態

若要設定 LUN 、請完成下列步驟：

1. 建立類型的 LUN `solaris`。
2. 安裝所指定的適當主機公用程式套件（ Huk ） "NetApp互通性對照表工具IMT （不含）"。
3. 請依照 Huk 中的說明進行操作、完全符合上述說明。以下概述基本步驟、但請參閱 "最新文件" 以瞭解正確的程序。
    a. 執行 `host_config` 更新的公用程式 `sd.conf/sdd.conf` 檔案：這樣做可讓 SCSI 磁碟機正確探索 ONTAP LUN 。
    b. 請遵循所提供的指示 `host_config` 啟用多重路徑輸入 / 輸出（ MPIO ）的公用程式。
    c. 重新開機。此步驟是必要步驟、以便在整個系統中辨識任何變更。
4. 分割 LUN 並確認它們已正確對齊。請參閱「附錄 B ： WAFL 校準驗證」、瞭解如何直接測試及確認校準。

zPools

只有在中的步驟之後才應建立 zpool "LUN 組態" 執行。如果程序未正確執行、可能會因為 I/O 對齊而導致嚴重的效能降低。ONTAP 的最佳效能要求 I/O 必須與磁碟機上的 4K 邊界對齊。在 zpool 上建立的檔案系統使用有效的區塊大小、並透過稱為的參數加以控制 `ashift`，您可以執行命令來檢視 `zdb -C`。

的價值 `ashift` 預設為 9 、表示 2^9 或 512 位元組。為了獲得最佳效能 `ashift` 值必須為 12 （ 2^12=4K ）。此值是在創建 zpool 時設置的，不能更改，這意味着 zpool 中的數據 `ashift` 除 12 個以外、應將資料複製到新

建立的 zPool 、以進行移轉。

建立 zPool 之後、請驗證的值 `ashift` 繼續之前。如果值不是 12 、則表示未正確探索到 LUN 。銷毀 zpool 、確認相關主機公用程式文件中顯示的所有步驟均已正確執行、然後重新建立 zPool 。

zPools 和 Solaris LDoms

Solaris LDoms 還需要確保 I/O 對齊正確無誤。雖然 LUN 可能會被正確發現為 4K 裝置、但 LDOM 上的虛擬 vdsk 裝置不會繼承 I/O 網域的組態。以該 LUN 為基礎的 vdsk 預設為 512 位元組區塊。

需要額外的組態檔案。首先、必須針對 Oracle 錯誤 15824910 修補個別的 LDOM 、才能啟用其他組態選項。此修補程式已移轉至所有目前使用的 Solaris 版本。一旦 LDOM 獲得修補、就可以依照下列方式設定新的正確對齊 LUN ：

1. 識別要在新的 zPool 中使用的 LUN 或 LUN 。在此範例中、它是 c2d1 裝置。

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. 擷取要用於 ZFS Pool 的裝置之 VDC 執行個體：

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. 編輯 `/platform/sun4v/kernel/drv/vdc.conf`：

```
block-size-list="1:4096";
```

這表示裝置執行個體 1 的區塊大小為 4096 。

另一個範例是假設需要將 vdsk 執行個體 1 至 6 設定為 4K 區塊大小和 `/etc/path_to_inst` 內容如下：

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. 最終結果 `vdc.conf` 檔案應包含下列項目：

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

注意
設定 VC.conf 並建立 vdsk 之後、必須重新啟動 LDOM 。無法避免此步驟。區塊大小變更只會在重新開機後生效。繼續使用 zpool 組態、並確保如前所述、移位已正確設定為 12 。

ZFS Intent Log （ ZIL ）

一般而言、沒有理由在不同的裝置上找到 ZFS Intent Log （ ZIL ）。記錄檔可以與主集區共用空間。獨立 ZIL 的主要用途是使用缺乏現代儲存陣列寫入快取功能的實體磁碟機。

logbias

設定 `logbias` 託管 Oracle 資料的 ZFS 檔案系統參數。

```
zfs set logbias=throughput <filesystem>
```

使用此參數可降低整體寫入層級。根據預設值、寫入的資料會先提交至 ZIL 、然後再提交至主儲存池。此方法適用於使用純磁碟機組態的組態、包括 SSD 型 ZIL 裝置和主儲存池的旋轉媒體。這是因為它允許在可用的最低延遲媒體上、在單一 I/O 交易中進行認可。

使用包含其快取功能的現代化儲存陣列時、通常不需要使用此方法。在極少數情況下、可能需要在單一交易中寫入記錄檔、例如由高度集中、對延遲敏感的隨機寫入所組成的工作負載。寫入放大的形式會產生影響、因為記錄的資料最終會寫入主儲存池、導致寫入活動加倍。

直接 I/O

許多應用程式（包括 Oracle 產品）都可以啟用直接 I/O 、藉此略過主機緩衝區快取此策略無法在 ZFS 檔案系統中正常運作。雖然會略過主機緩衝區快取、但 ZFS 本身仍會繼續快取資料。使用 Fio 或 Sio 等工具執行效能測試時、這項動作可能會產生誤導性的結果、因為很難預測 I/O 是否到達儲存系統、或是是否在作業系統中本機快取。此動作也會讓使用此類模擬測試來比較 ZFS 效能與其他檔案系統的情況變得非常困難。實際上、在真實使用者工作負載下、檔案系統效能幾乎沒有任何差異。

多個 zPools

必須在 zpool 層級執行快照型備份、還原、複製及歸檔 ZFS 型資料、而且通常需要多個 zPools 。zpool 類似於 LVM 磁碟群組、應使用相同的規則進行設定。例如、資料庫的配置最好是存放在資料檔案上 `zpool1` 以及駐留在上的歸檔記錄、控制檔和重做記錄 `zpool2`。此方法允許標準熱備份、將資料庫置於熱備份模式、然後是的快照 `zpool1`。接著會從熱備份模式移除資料庫、強制進行記錄歸檔、並建立快照 `zpool2` 已建立。還原作業需要卸載 zfs 檔案系統、並在執行 SnapRestore 還原作業之後、將 zPool 完全離線。然後可以重新上線並恢復資料庫。

filesystemio_options

Oracle 參數 `filesystemio_options` 使用 ZFS 的方式不同。如果 `setall` 或 `directio` 使用時、寫入作業會同步並略過 OS 緩衝區快取、但讀取會由 ZFS 進行緩衝。此動作會導致效能分析方面的困難、因為有時會被 ZFS 快取攔截和服務 I/O 、使儲存延遲和總 I/O 比預期的要少。

版權資訊

商標資訊