Sun Cluster 概念指南 (適用於 Solaris 作業系統)

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054
U.S.A.

文件號碼：819–2064–10
2005 年 8 月，修訂版 A

Adobe PostScript

050816@12762

目錄

前言

「Sun™ Cluster 概念指南 (適用於 Solaris 作業系統)」同時包含適用於基於 SPARC® 和 x86 的系統的 SunPlex™ 系統的概念和參考資訊。

備註 – 在本文件中，「x86」一詞指 Intel 32 位元系列的微處理器晶片和 AMD 製造的相容微處理器晶片。

SunPlex 系統包含構成 Sun 叢集解決方案的所有硬體和軟體元件。

本文件主要針對接受過關於 Sun Cluster 軟體訓練的有經驗的系統管理員。請不要將本文件當做規劃作業或售前指引。您應該已經決定您的系統需求並購買了適當的設備與軟體之後，再閱讀本文件。

若要理解本書說明的概念，您應該具備 Solaris™ 作業系統的知識和 Sun Cluster 系統使用的容體管理程式軟體的專業技術。

備註 – Sun Cluster 軟體在兩個平台 (SPARC 與 x86上) 上執行。本文件中的資訊適用於這兩個平台，除非在特定章節、小節、備註、項目符號、圖形、表格或範例中另行指定。

印刷排版慣例

下表描述本書在印刷排版上所作的變更。

表 P–1 印刷排版慣例

字體*	意義	範例
`AaBbCc123`	指令、檔案及目錄的名稱；螢幕畫面輸出。	請編輯您的`.login`檔案。 請使用 `ls -a` 列出所有檔案。 `machine_name% you have mail.`
`AaBbCc123`	您所鍵入的內容(與螢幕畫面輸出相區別)。	`machine_name%` **`su`** `Password:`
AaBbCc123	保留未譯的新的字彙或術語、要強調的詞。	要刪除檔案，請鍵入 `rm` *filename*。 （注：在聯機狀態下，有些需要強調的詞以黑體顯示。）
術語強調變數	新的字彙或術語、要強調的詞。 將用實際的名稱或數值取代的指令行變數。	請執行**修補程序分析**。 請**不要**儲存此檔案。
「AaBbCc123」	用於書名及章節名稱。	請閱讀「使用者指南」中的第 6 章。

* 瀏覽器中的設定可能會與這些設定不同。

指令範例中的 Shell 提示符號

下表顯示用於
C shell、Bourne shell 和 Korn shell 的預設系統提示符號以及超級使用者提示符號。

表 P–2 Shell 提示

Shell	提示
C shell 提示符號	`machine_name%`
C shell 超級使用者提示符號	`machine_name#`
Bourne shell 和 Korn shell 提示符號	`$`
Bourne shell 和 Korn shell 超級使用者提示符號	`#`

相關說明文件

有關 Sun Cluster 相關主題的資訊可從下表中列出的文件中獲得。所有 Sun Cluster 文件均可從 http://docs.sun.com 取得。

主題	文件
概述	「Sun Cluster 簡介 (適用於 Solaris 作業系統)」
概念	「Sun Cluster 概念指南 (適用於 Solaris 作業系統)」
硬體安裝與管理	「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」
	個別硬體管理指南
軟體安裝	「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」
資料服務安裝與管理	「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」
	個別資料服務指南
資料服務開發	「Sun Cluster 資料服務開發者指南 (適用於 Solaris 作業系統)」
系統管理	「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」
錯誤訊息	「Sun Cluster Error Messages Guide for Solaris OS」
指令和功能參考	「Sun Cluster Reference Manual for Solaris OS」

如需 Sun Cluster 文件的完整清單，請參閱 http://docs.sun.com 上關於您的 Sun Cluster 軟體發行版本之版本說明。

線上存取 Sun 文件

docs.sun.com℠ 網站可讓您存取 Sun 線上技術文件。您可以瀏覽 docs.sun.com 的歸檔檔案或搜尋特定書名或主題。其 URL 爲 `http://docs.sun.com`。

訂購 Sun 說明文件

Sun Microsystems 提供列印的所選產品文件。如需文件清單與訂購方式，請參閱 `http://docs.sun.com` 上的「購買書面說明文件」。

取得說明

如果在安裝或使用 Sun Cluster 系統時遇到問題，請聯絡您的服務供應商並提供以下資訊：

- 您的姓名和電子郵件地址 (如果有的話)
- 您的公司名稱、地址和電話號碼
- 您系統的機型和序號
- 作業環境的版次編號 (例如，Solaris 9)
- Sun Cluster 軟體的版次編號 (例如，3.1 8/05)

使用下列指令收集您系統上每一個節點的相關資訊，提供給您的服務供應商：

指令	功能
`prtconf -v`	顯示系統記憶體的大小及報告周邊裝置的相關資訊
`psrinfo -v`	顯示處理器的相關資訊
`showrev -p`	報告安裝了哪些修補程式
SPARC：`prtdiag -v`	顯示系統診斷資訊
`scinstall -pv`	顯示 Sun Cluster 軟體發行版本和套裝軟體版本資訊
`scstat`	提供叢集狀態的快照
`scconf -p`	列示叢集配置資訊
`scrgadm -p`	顯示有關已安裝資源、資源群組與資源類型的資訊

同時還請提供 `/var/adm/messages` 檔案的內容。

產品訓練

Sun Microsystems 透過由講師主持的課程和自學課程來提供關於多種 Sun 技術的訓練。如需有關 Sun 提供的訓練課程的資訊並希望報名加入班級，請造訪 Sun Microsystems 訓練 `http://training.sun.com/`。

第 1 章

簡介與概觀

Sun Cluster 系統是一種整合的硬體和用於建立高度可用的和可延伸服務的 Sun Cluster 軟體解決方案。

「Sun Cluster 概念指南 (適用於 Solaris 作業系統)」爲 Sun Cluster 文件的初級讀者提供所需的概念資訊。這些讀者包括

- 安裝與維修叢集硬體的服務供應商
- 安裝、配置和管理 Sun Cluster 軟體的系統管理員
- 開發適用於目前 Sun Cluster 產品所不包含的應用程式容錯移轉和可延伸服務的應用程式開發人員

使用本書與整套 Sun Cluster 文件以全面瞭解 Sun Cluster 系統。

本章

- 提供 Sun Cluster 系統的介紹和高階簡介
- 說明 Sun Cluster 讀者的數種視角
- 指出在使用 Sun Cluster 系統之前需要瞭解的重要概念
- 將重要概念與包含程序和相關資訊的 Sun Cluster 文件相對應
- 對應叢集相關作業至包含用來完成那些作業程序的說明文件

Sun Cluster 系統簡介

Sun Cluster 系統將 Solaris 作業系統延伸成爲叢集作業系統。叢集或診測裝置是一組鬆散式結合的運算節點，提供網路服務或應用程式的單一用戶端檢視，包括資料庫、網路服務和檔案服務。

每一個叢集節點均爲一個獨立的伺服器，可執行其本身的處理程序。這些處理程序可以互相通訊，有如形成 (對網路用戶端而言) 一個單一系統，協力將應用程式、系統資源和資料提供給使用者。

叢集可提供比傳統單一伺服器系統更多項的優勢。這些優勢包括支援故障轉移和可延伸服務的支援、模組成長的能力，以及比傳統硬體容錯系統低的導入成本。

Sun Cluster 系統的目標包括：

- 減少或免除因爲軟體或硬體故障所造成的當機時間
- 確保對一般使用者的資料和應用程式的可用性，不論是否爲一般使單一伺服器系統當機的那種故障
- 增加節點至叢集，讓服務延伸至額外的處理器，以增加應用程式的效率
- 強化系統的可用性，讓您可以執行維護作業而不需要關閉整個叢集

如需有關容錯性和高度可用性的更多資訊，請參閱「Sun Cluster 簡介 (適用於 Solaris 作業系統)」中的「透過 Sun Cluster 使應用程式具有高度可用性」。

請參閱第 79 頁的 「高度可用性常見問題」，以瞭解關於高度可用性的問題與解答。

Sun Cluster 系統的三種視角

本節說明 Sun Cluster 系統的三種不同的視角和重要概念以及每種視角的相關文件。以下是下專業人員的典型視角：

- 硬體安裝與維修人員
- 系統管理員
- 應用程式開發人員

硬體安裝和維修視角

對於硬體維修專業人員而言，Sun Cluster 系統就像是一組常用的硬體，包括伺服器、網路和儲存體。這些元件全部以電纜連接在一起，因此使得每一個元件均具有備份而不會有單點故障存在。

重要概念 – 硬體

硬體維修專業人員需要理解以下叢集概念。

- 叢集硬體配置和電纜佈線
- 安裝與維修 (新增、移除、更換)：
  - 網路介面元件 (配接卡、連接、電纜)
  - 磁碟介面卡
  - 磁碟陣列

    - 磁碟機
    - 管理主控台和主控台存取裝置
- 設定管理主控台和主控台存取裝置

更多的硬體概念資訊

下列各節包含前述重要概念的相關資料：

- 第 20 頁的 「叢集節點」
- 第 22 頁的 「多重主機裝置」
- 第 23 頁的 「Local Disks（本機磁碟）」
- 第 23 頁的 「叢集交互連接」
- 第 24 頁的 「公用網路介面」
- 第 24 頁的 「用戶端系統」
- 第 25 頁的 「管理主控台」
- 第 24 頁的 「主控台存取裝置」
- 第 26 頁的 「SPARC: 適用於 SPARC 的叢集化配對拓樸」
- 第 28 頁的 「SPARC: 適用於 SPARC 的 N+1 (星狀) 拓樸」

供硬體專業人員使用的 Sun Cluster 文件

以下 Sun Cluster 文件包含與硬體維修概念相關的程序和資訊：

「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」

系統管理員視角

對於系統管理員而言，Sun Cluster 系統是一組以電纜連接在一起的伺服器 (節點) 和共用儲存裝置。系統管理員會注意執行特定作業的軟體：

- 與 Solaris 軟體整合的專用叢集軟體，用來監視叢集節點之間的連接
- 用來監視在叢集節點上執行的使用者應用程式運作狀況的專用軟體
- 設定和管理磁碟的容體管理軟體
- 可以讓所有節點存取所有儲存裝置 (即使未直接連接到磁碟) 的專用叢集軟體
- 可以讓檔案像是本機連接至該節點的方式出現於每個節點上的專用叢集軟體

重要概念 – 系統管理

系統管理員需要瞭解下列概念與程序：

- 硬體和軟體元件之間的相互作用
- 安裝和配置叢集的一般流程，包括：
    - 安裝 Solaris 作業系統

    - 安裝和配置 Sun Cluster 軟體
    - 安裝和配置容體管理程式
    - 安裝和配置應用軟體使其成爲具備叢集功能
    - 安裝和配置 Sun Cluster 資料服務軟體
- 新增、移除、更換和維修叢集硬體與軟體元件的叢集管理程序
- 修改配置以增進效能

更多的管理員概念資訊

下列各節包含前述重要概念的相關資料：

- 第 33 頁的 「管理介面」
- 第 34 頁的 「叢集時間」
- 第 34 頁的 「高可用性框架」
- 第 37 頁的 「全域裝置」
- 第 38 頁的 「磁碟裝置群組」
- 第 40 頁的 「全域名稱空間」
- 第 42 頁的 「Cluster File Systems（叢集檔案系統）」
- 第 44 頁的 「磁碟路徑監視」
- 第 48 頁的 「關於故障隔離」
- 第 56 頁的 「資料服務」

供系統管理員使用的 Sun Cluster 文件

以下 Sun Cluster 文件包含與系統管理概念相關的程序和資訊：

- 「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」
- 「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」
- 「Sun Cluster Error Messages Guide for Solaris OS」
- 「Sun Cluster 3.1 8/05 版本說明 (適用於 Solaris 作業系統)」
- 「Sun Cluster 3.0-3.1 Release Notes Supplement」

應用程式開發人員視角

Sun Cluster 系統爲以下應用程式提供**資料服務**：Oracle、NFS、DNS、Sun™ Java System Web Server、Apache Web Server (在基與 SPARC 系統上) 和 Sun Java System Directory Server 等。資料服務是透過配置 Sun Cluster 軟體控制下的常用應用程式建立的。Sun Cluster 軟體提供啓動、停止和監視應用程式的配置檔案和管理方法。如果您需要建立新的容錯移轉或可延伸服務，您可以使用 Sun Cluster 應用程式編程介面 (API) 和資料服務啓用技術 API (DSET API)，來開發可以使其應用程式在叢集上作爲資料服務執行的必要的配置檔案和管理方法。

重要概念 – 應用程式開發

應用程式開發人員需要理解以下內容：

- 應用程式的特性，以決定其是否可以被當作故障轉移或可延伸的資料服務來執行。

- Sun Cluster API、DSET API 和「通用」資料服務。開發人員需要決定最適合於編寫用於爲叢集環境配置其應用程式的程式或程式檔的工具。

更多的應程式開發人員概念資訊

下列各節包含前述重要概念的相關資料：

- 第 56 頁的 「資料服務」
- 第 64 頁的 「資源、資源群組與資源類型」
- 第 4 章

供應用程式開發人員使用的 Sun Cluster 文件

以下 Sun Cluster 文件包含與應用程式開發人員概念相關的程序和資訊：

- 「Sun Cluster 資料服務開發者指南 (適用於 Solaris 作業系統)」
- 「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」

Sun Cluster 系統作業

所有 Sun Cluster 系統作業都需要具備某些概念背景。下列表格提供作業與說明作業步驟之說明文件的進階概觀。本書中有關的概念章節說明概念如何對應至這些作業。

表 1–1 對應作業：將使用者作業對應至文件

作業	操作說明
安裝叢集硬體	「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」
將 Solaris 軟體安裝於叢集上	「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」
SPARC：安裝 Sun™ Management Center 軟體	「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」
安裝和配置 Sun Cluster 軟體	「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」
安裝和配置容體管理軟體	「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」 您的容體管理說明文件

表 1–1 對應作業：將使用者作業對應至文件　*(續)*

作業	操作說明
安裝和配置 Sun Cluster 資料服務	「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」
維修叢集硬體	「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」
管理 Sun Cluster 軟體	「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」
管理容體管理軟體	「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」和容體管理文件
管理應用程式軟體	您的應用程式說明文件
問題辨別與建議的使用者動作	「Sun Cluster Error Messages Guide for Solaris OS」
建立新的資料服務	「Sun Cluster 資料服務開發者指南 (適用於 Solaris 作業系統)」

第 2 章

硬體服務提供者的重要概念

本章說明有關 Sun Cluster 系統配置的硬體元件的重要概念。涵蓋的主題包括以下內容：

Sun Cluster 系統硬體和軟體元件

本資訊主要是針對硬體服務供應商。這些概念可以協助服務供應商在安裝、配置或維修叢集硬體之前，瞭解各硬體元件之間的關係。叢集系統管理員可能也會發現，這項資訊對於安裝、配置和管理叢集軟體是很有用的。

叢集是由數個硬體元件所組成，包括：

- 具有本機磁碟 (未共用) 的叢集節點
- 多重主機儲存體 (節點之間共用磁碟)
- 可移除式媒體 (磁帶和 CD-ROM)
- 叢集互連
- 公用網路介面
- 用戶端系統
- 管理主控台

- 主控台存取裝置

Sun Cluster 系統可以讓您將這些元件合併爲多種配置。下列各節將說明這些配置。

- 第 26 頁的 「SPARC: 適用於 SPARC 的 Sun Cluster 拓樸」
- 第 30 頁的 「x86: 適用於 x86 的 Sun Cluster 」

如需範例雙節點叢集配置的圖例，請參閱「Sun Cluster 簡介 (適用於 Solaris 作業系統)」中的「Sun Cluster 硬體環境」。

叢集節點

叢集節點是一種同時執行 Solaris Operating System 和 Sun Cluster 軟體的機器。叢集節點還是叢集的目前成員 (**叢集成員**)，或潛在成員。

- SPARC： Sun Cluster 軟體在叢集中支援一到十六個節點。請參閱第 26 頁的 「SPARC: 適用於 SPARC 的 Sun Cluster 拓樸」，以瞭解所支援的節點配置。
- x86： Sun Cluster 軟體在叢集中支援兩個節點。請參閱第 30 頁的 「x86: 適用於 x86 的 Sun Cluster 」，以瞭解所支援的節點配置。

叢集節點一般連接到一個或多個多重主機裝置。未連接到多重主機裝置的節點使用叢集檔案系統來存取多重主機裝置。例如，一個可延伸的服務配置可以讓節點不需要直接連接到多重主機裝置便可處理請求。

另外，平行資料庫配置中的節點共用對所有磁碟的並行。

- 請參閱第 22 頁的 「多重主機裝置」，以取得有關並行存取磁碟的資訊。
- 請參閱第 26 頁的 「SPARC: 適用於 SPARC 的叢集化配對拓樸」和第 30 頁的 「x86: 適用於 x86 的叢集化配對拓撲」，以取得有關平行資料庫配置的更多資訊。

叢集中的所有節點會依照一般名稱，即叢集名稱 (用來存取和管理叢集)，來加以分群。

公用網路配接卡會將節點連接到公用網路，以供用戶端存取叢集。

叢集成員與叢集中的其他節點透過一個或多個實體上獨立的網路進行通訊。此組實體上獨立的網路是被視爲**叢集交互連接**。

當另一個節點加入或離開叢集時，叢集中的每個節點都會知道。此外，叢集中的每個節點也都知道本機正在執行的資源，以及在其他叢集節點上執行的資源。

相同叢集中的節點必須有類似的處理程序、記憶體和 I/O 能力，以便啓動故障轉移，而不至於大幅降低效能。因爲可能發生容錯移轉，每個節點必須有足夠的額外容量，以承擔所有的備份或次要節點的工作負荷量。

每一個節點會啓動其個別的 root (/) 檔案系統。

叢集硬體成員的軟體元件

若要作爲叢集成員運作，節點必須安裝有以下軟體：

- Solaris 作業系統
- Sun Cluster 軟體
- 資料服務應用程式
- 容體管理 (Solaris Volume Manager™ 或 VERITAS Volume Manager)

  使用獨立磁碟 (RAID) 硬體冗餘陣列的配置是一個例外。此配置可能不需要軟體容體管理程式，如 Solaris Volume Manager 或 VERITAS Volume Manager。
- 請參閱「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」，以取得有關如何安裝 Solaris 作業系統、Sun Cluster 和容體管理軟體的資訊。
- 請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」，以取得有關如何安裝和配置資料服務的資訊。
- 請參閱第 3 章，以取得有關上述軟體元件的概念資訊。

下圖提供共同運作以建立 Sun Cluster 軟體環境之軟體元件的高階觀點。

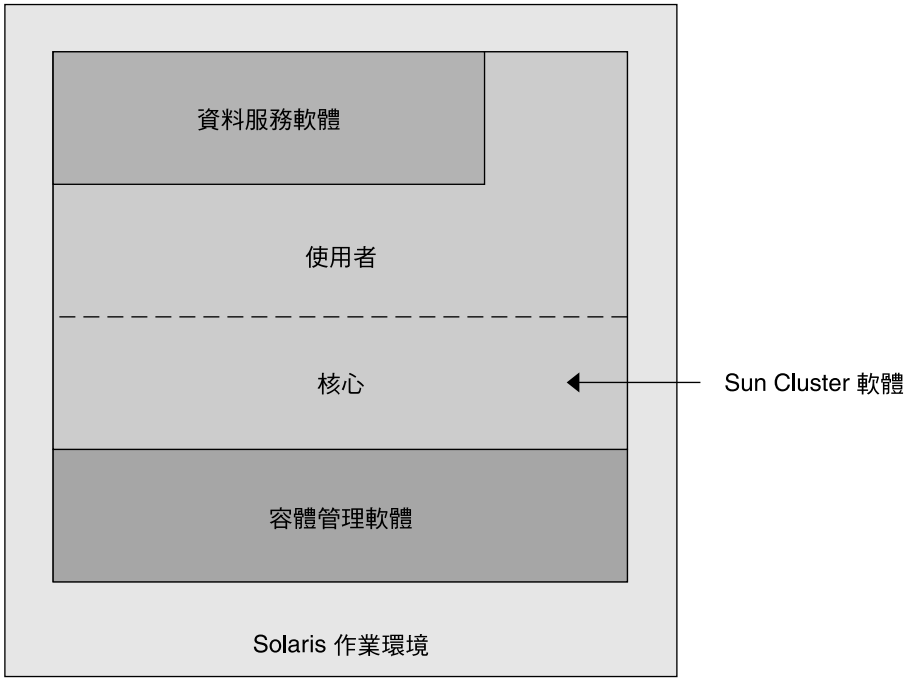


圖 **2–1** Sun Cluster 軟體元件的高階關係

請參閱第 4 章，以瞭解有關叢集成員的問題與解答。

多重主機裝置

一次可以連接至多個節點的磁碟是多重主機裝置。在 Sun Cluster 環境中，多重主機儲存體可使磁碟具有高度可用性。Sun Cluster 軟體需要雙節點叢集多重主機儲存體以建立法定數目。超過兩個節點以上的叢集不需要法定裝置。如需有關法定數目的更多資訊，請參閱第 47 頁的 「法定數目和法定裝置」。

多重主機裝置有下列特性。

- 單一節點故障的容錯性。
- 能夠儲存應用程式資料、應用程式二進位代碼及配置檔案。
- 防止節點故障。如果用戶端透過某個節點請求資料且節點發生故障，則會將切換移轉這些請求，以使用與同一磁碟有直接連線的其他節點。
- 透過「控制」磁碟的主要節點全域存取，或透過本機路徑直接並行存取。目前，Oracle Real Application Clusters Guard 是唯一使用直接並行存取的應用程式。

容體管理程式爲多重主機裝置的資料冗餘的鏡像配置或 RAID-5 配置做了準備。目前，Sun Cluster 支援 Solaris Volume Manager 和 VERITAS Volume Manager，作爲容體管理程式，僅在基於 SPARC 的叢集中和數種硬體 RAID 平台上的 RDAC RAID-5 硬體控制器中可用。

透過磁碟鏡像與磁碟平行儲存將重主機裝置結合起來，可以防止節點故障和單個磁碟故障。

請參閱第 4 章，以瞭解有關多重主機儲存體的問題與解答。

多重初始端 SCSI

本節僅適用於 SCSI 儲存體，不適用於多重主機裝置的「光纖通道」儲存體。

在獨立式伺服器中，伺服器節點是以連接此伺服器至特定 SCSI 匯流排的 SCSI 主機配接卡電路，來控制 SCSI 匯流排活動。此 SCSI 主機配接卡電路即爲 *SCSI* **初始端** *(SCSI initiator)*。這個電路起始此 SCSI 匯流排的所有匯流排活動。SCSI 主機配接卡的預設 SCSI 位址在 Sun 系統中是 7。

叢集配置利用多重主機裝置在多重伺服器節點之間共用儲存體。叢集儲存體由單端或差動 SCSI 裝置組成時，配置稱爲多重初始端 SCSI。這個詞彙所隱含的意義，即 SCSI 匯流排上存在一個以上的 SCSI 初始端。

SCSI 規格要求 SCSI 匯流排上的每個裝置均具有唯一的 SCSI 位址。(主機配接卡也是 SCSI 匯流排上的裝置。)多重初始端環境中的預設硬體配置導致衝突，原因是所有 SCSI 主機配接卡均預設爲 7。

若要解決衝突，在每個 SCSI 匯流排上，留下其中一個 SCSI 主機配接卡的 SCSI 位址爲 7，並將其他的主機配接卡設定爲未用的 SCSI 位址。請適當地規劃指定這些"未用的" SCSI 位址，包括目前和最後未使用的位址。將來不使用的位址範例，是安裝新磁碟到空磁碟插槽以便增加儲存體。

在大部分配置中，第二主機配接卡的可用 SCSI 位址爲 6。

您可以使用下列工具中的一種來設定 `scsi-initiator-id` 特性，以變更這些爲主機配接卡選取的 SCSI 位址：

- `eeprom`(1M)
- 基於 SPARC 的系統上的 OpenBoot PROM
- BIOS 在基於 x86 的系統上啓動之後，您選擇執行的 SCSI 公用程式

您可以全域式或以個別主機配接卡的方式，來設定節點的這個特性。關於爲每個 SCSI 主機配接卡設定唯一 `scsi-initiator-id` 的說明包含在「Sun Cluster 3.0-3.1 With SCSI JBOD Storage Device Manual for Solaris OS」中。

Local Disks（本機磁碟）

本機磁碟是僅連接至單一節點的磁碟。因此，不能防止本機磁碟發生節點故障 (非高度可用)。然而，所有磁碟 (包括本機磁碟) 均包含在全域名稱空間中並且均配置爲**全域裝置**。因此，從所有的叢集節點可以看到磁碟本身。

您可以透過將本機磁碟上的檔案系統置於全域掛載點下，以使其對其他節點可用。如果目前裝載這些整體檔案系統之其中一個檔案系統的節點故障，所有節點均會遺失該檔案系統的存取。使用容體管理程式可讓您鏡像這些磁碟，如此磁碟故障就不會導致這些檔案系統成爲無法存取，但是容體管理程式無法防止節點故障。

請參閱第 37 頁的 「全域裝置」一節，以取得有關全域裝置的更多資訊。

可移除的媒體

叢集中支援如磁帶機和 CD-ROM 光碟機的抽換式媒體。通常，安裝、配置和服務這些裝置的方法與在非叢集環境中相同。在 Sun Cluster 中，這些裝置均配置爲全域裝置，所以每個裝置均可從叢集的任何節點進行存取。請參閱「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」，以取得有關安裝和配置可移除式媒體的資訊。

請參閱第 37 頁的 「全域裝置」一節，以取得有關全域裝置的更多資訊。

叢集交互連接

叢集交互連接是用於在叢集節點之間傳輸叢集私有通訊與資料服務通訊的裝置實體配置。由於交互連接廣泛使用於叢集私有通訊，所以會限制效能。

只有叢集節點可以連接至叢集交互連接。Sun Cluster 安全性模型假定只有叢集節點對叢集互連具有實體存取權。

必須使用叢集互連透過至少兩個實體上獨立的備援網路或路徑連線所有節點，以避免單點故障。任何兩個節點之間可以有多個實體上獨立的網路 (二到六個)。

叢集交互連接由三個硬體元件組成：配接卡、接點與電纜。下表說明各個硬體元件。

- 配接卡 – 駐留在每個叢集節點內的網路介面卡。其名稱由裝置名稱構成，其後緊跟實體單元編號，例如 qfe2。某些配接卡僅有一個實體網路連接，但其他配接卡 (像 qfe 卡) 具有多個實體連接。某些配接卡同時會包含網路介面和儲存介面。

  具有多重介面的網路配接卡在整個配接卡出現故障時會成爲單一故障點。爲擁有最大的可用性，請規劃您的叢集，使兩個節點之間的唯一路徑不會依賴單一的網路配接卡。
- 接點 – 駐留在叢集節點外的切換點。接點執行傳輸和交換功能，讓您連線兩個以上的節點。在雙節點的叢集中，您不需要接點，因爲透過多餘的實體電纜連接至每個節點上的冗餘配接卡，節點可以直接彼此連接。大於兩個節點的配置一般會需要接點。
- 電纜 – 安裝在兩個網路配接卡或配接卡與接點之間的實體連接。

請參閱第 4 章，以瞭解有關叢集互連的問題與解答。

公用網路介面

用戶端透過公用網路介面連接至叢集。每一個網路配接卡可以連接至一或多個公用網路，這要根據配接卡是否有多重硬體介面而定。您可以設定節點來包含已配置的多重公用網路介面卡，使多重卡都處於使用中狀態，並且彼此作爲故障轉移的備份。如果配接卡中的一個發生故障，則將呼叫 Internet Protocol (IP) 網路多重路徑 軟體以將該發生故障的介面容錯移轉至群組中的其他配接卡。

公用網路介面的叢集不需要特別的硬體注意事項。

請參閱第 4 章，以瞭解有關公用網路的問題與解答。

用戶端系統

用戶端系統包括工作站或透過公用網路存取叢集的其他伺服器。用戶端程式使用在叢集中執行的伺服器端應用程式所提供的資料或其他服務。

用戶端系統不具高可用性。叢集上的資料和應用程式則具高可用性。

請參閱第 4 章，以瞭解有關用戶端系統的問題與解答。

主控台存取裝置

對於所有的叢集節點，您必須擁有主控台存取權。若要獲得主控台存取權，請使用以下裝置中的一種：

- 與叢集硬體一同購買的終端機集訊機
- Sun Enterprise E10000 伺服器上的系統服務處理器 (SSP) (適用於基於 SPARC 的叢集)

- Sun Fire™ 伺服器上的系統控制器 (同樣適用於基於 SPARC 的叢集)
- 可以存取每個節點上的 `ttya` 的另一種裝置

來自 Sun 之受支援的終端機集線器只有一個，而是否使用此支援的 Sun 終端機集線器是可選擇的。終端機集線器允許使用 TCP/IP 網路來存取每一個節點上的 `/dev/console`。結果是從網路上任意位置的遠端工作站，以主控台層次來存取每一個節點。

系統服務處理器 (SSP) 爲 Sun Enterprise E1000 伺服器提供主控台存取權。SSP 是一種位於配置爲支援 Sun Enterprise E1000 伺服器的乙太網路上的機器。SSP 是 Sun Enterprise E1000 伺服器管理主控台。使用「Sun Enterprise E10000 網路主控台」功能，網路上的任何工作站皆可開啓主機主控台階段作業。

其他主控台存取方法包括其他終端機集訊機，從其他節點和無智型終端機的 `tip(1)` 串列埠存取。您可以使用 Sun™ 鍵盤和監視器，或其他串列埠裝置 (如果您的硬體服務供應商支援這些裝置)。

管理主控台

您可以使用專屬的 UltraSPARC® 工作站或 Sun Fire V65x 伺服器，稱爲**管理主控台**，以管理使用中的叢集。通常，您在管理主控台上安裝和執行管理工具軟體，如叢集控制面板 (CCP) 和適用於 Sun Management Center 產品 (僅可與基於 SPARC 的叢集一同使用) 的 Sun Cluster 模組。使用 CCP 下的 `cconsole` 可讓您一次連接一個以上的節點主控台。如需有關使用 CCP 的更多資訊，請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 1 章「管理 Sun Cluster 的簡介」。

管理主控台並非叢集節點。管理主控台用於對叢集節點的遠端存取，可透過公用網路，也可選擇性地透過基於網路的終端機集線器進行。如果您的叢集是由 Sun Enterprise E10000 平台組成的，則必須從管理主控台登入 SSP 並使用 `netcon(1M)` 指令進行連接。

一般您會配置沒有監視器的節點。然後，透過 `telnet` 階段作業從管理主控台存取節點的主控台。管理主控台連接至終端機集訊機，並從終端機集訊機連接至節點的串列埠。如果是 Sun Enterprise E1000 伺服器，則從系統服務處理器連接。請參閱第 24 頁的「主控台存取裝置」，以取得更多資訊。

Sun Cluster 不需要專屬的管理主控台，但是使用專屬的主控台有以下優點：

- 在同一機器上將主控台和管理工具分組，達到中央化叢集管理
- 讓您的硬體服務供應商可較快速地解決問題

請參閱第 4 章，以瞭解有關管理主控台的問題與解答。

SPARC: 適用於 SPARC 的 Sun Cluster 拓樸

拓撲是連接叢集節點和叢集中所使用儲存體平台的連接機制。Sun Cluster 軟體支援所有符合以下規範的拓樸。

- 由基於 SPARC 的系統組成的 Sun Cluster 環境在叢集中最多支援十六個節點，無論您實作的儲存配置如何。
- 共用的儲存裝置可以連接至儲存裝置所支援數目的節點。
- 共用的儲存體不需要連接至叢集的所有節點。不過，它們必須至少連接至兩個節點。

Sun Cluster 軟體不要求您使用特定的拓樸配置叢集。透過說明下列拓撲來提供論述叢集連接機制的語彙。這些拓撲是典型的連接機制。

- 叢集化對
- Pair+N
- N+1 (星狀)
- N*N (可延伸的)

以下各節包含說明每一種拓撲架構的圖表。

SPARC: 適用於 SPARC 的叢集化配對拓樸

叢集化配對拓樸是兩對或多對在單一叢集管理框架下作業的節點。在此配置中，故障轉移僅發生於配對之間。然而，所有的節點均透過叢集互連進行連接，並在 Sun Cluster 軟體控制下運作。您可能會使用這種拓撲架構，在某個配對上執行平行資料庫應用程式，而在另一個配對上執行故障轉移或可延伸的應用程式。

使用叢集檔案系統，您還可以使用雙組配置。兩個以上的節點可以執行可縮放式服務或並列資料庫，即使所有節點並未直接連接至儲存應用程式的磁碟。

下圖說明叢集化配對配置。

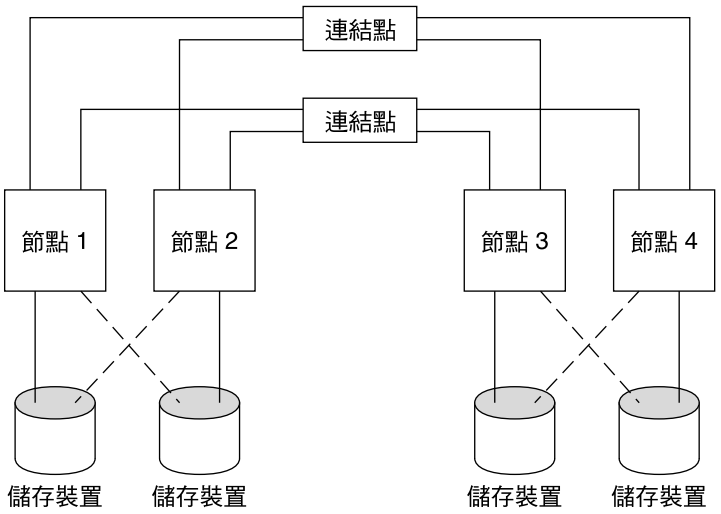


圖 **2–2** SPARC: 叢集化配對拓撲

SPARC: 適用於 SPARC 的 Pair+N 拓樸

此 pair+N 拓撲中包含一對直接連接至共用儲存體的節點與附加節點集，它們使用叢集交互連接來存取共用儲存體，其本身並不具備直接連接。

下圖展示 pair+N 拓撲，其中四個節點的兩個 (節點 3 和節點 4) 使用叢集交互連接來存取儲存體。此項配置可加以擴展，以便納入其他並未具有可直接存取共用儲存體的節點。

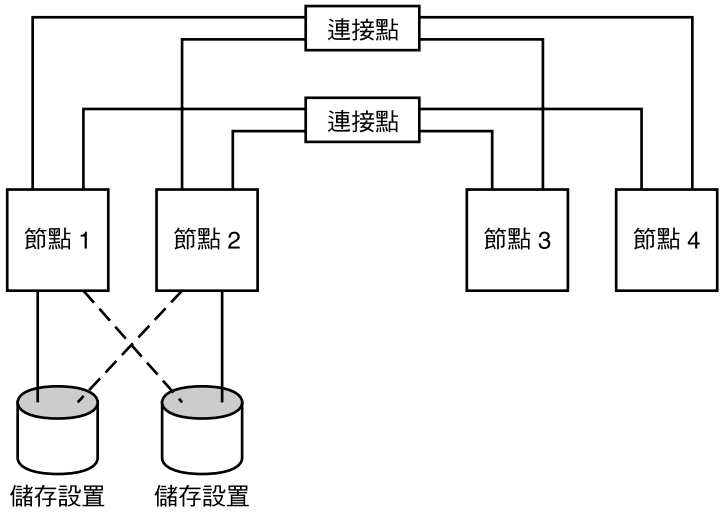


圖 2–3 Pair+N 拓撲

SPARC: 適用於 SPARC 的 N+1 (星狀) 拓樸

N+1 拓撲架構包括一些主要節點和一個次要節點。您不需要配置相同的主要節點和次要節點。主要節點主動提供應用程式服務。在等待主要節點故障時，次要節點不需要閒置。

次要節點在配置中是唯一實際連接至所有多重主機儲存體的節點。

如果主要節點上發生故障，Sun Cluster 會將資源容錯移轉至次要節點，直至切換 (自動或手動) 回到主要節點。

次要節點必須時常保有足夠的額外 CPU 容量，以便在主要節點之一故障時處理負載。

下圖說明 N+1 配置。

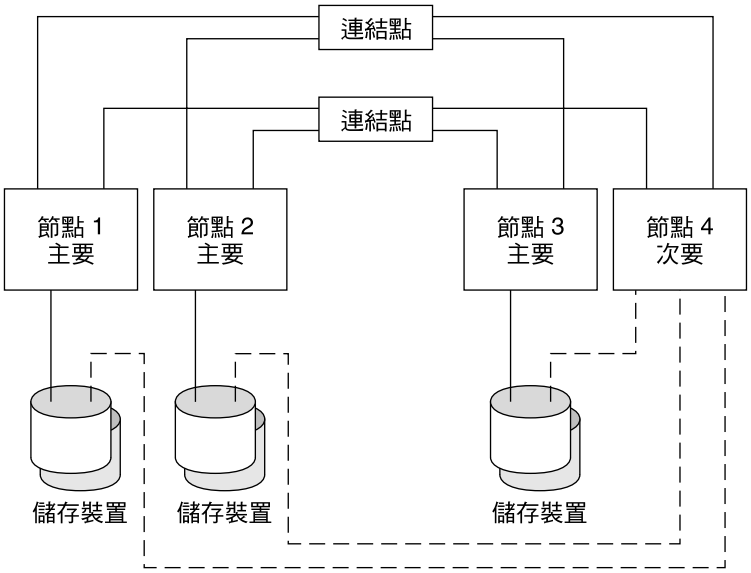


圖 2–4 SPARC: N+1 拓撲

SPARC: 適用於 SPARC 的 N*N (可延伸的) 拓樸

N*N 拓樸使叢集中的每個共用儲存裝置都可以連接至叢集中的每個節點。此拓樸使高度可用的應用程式可以從一個節點容錯移轉至另一個節點而不會發生服務降級。如果發生容錯移轉，則新的節點可以使用本機路徑而非私有互連存取儲存裝置。

下圖說明 N*N 配置。

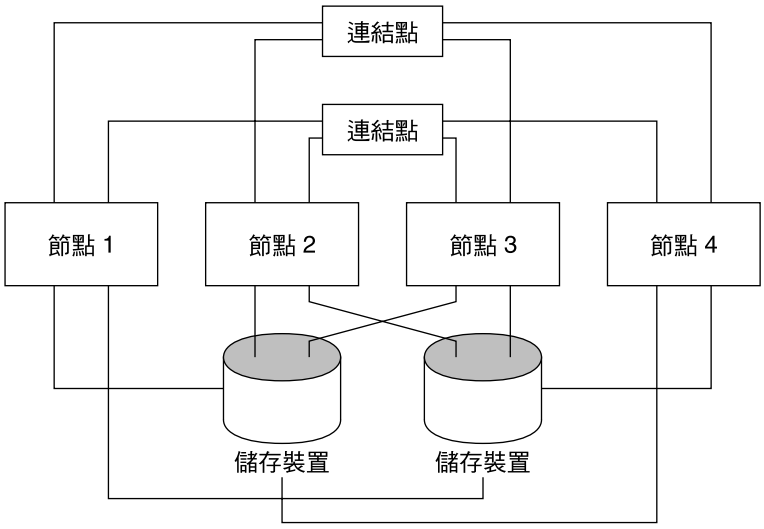


圖 **2–5** SPARC: N*N 拓撲

x86: 適用於 x86 的 Sun Cluster

拓撲是連接叢集節點和叢集中所使用儲存體平台的連接機制。Sun Cluster 支援符合下列準則的所有拓撲。

- 由基於 x86 的系統所組成的 Sun Cluster 支援叢集中的兩個節點。
- 共用儲存裝置必須連接至這兩個節點。

Sun Cluster 不需要您透過特定拓撲配置一個叢集。透過說明下列叢集化配對拓撲 (是由基於 x86 節點所組成的叢集之唯一拓撲)，來提供論述叢集連接機制的詞彙。此拓撲是典型的連接機制。

下面一節包含拓撲圖表範例。

x86: 適用於 x86 的叢集化配對拓撲

叢集化配對拓撲是在單一叢集管理框架下作業的兩個節點。在此配置中，故障轉移僅發生於配對之間。然而，所有的節點均透過叢集互連進行連接，並在 Sun Cluster 軟體控制下運作。您可以使用這種拓撲在配對上執行平行資料庫、故障轉移或可延伸的應用程式。

下圖說明叢集化配對配置。

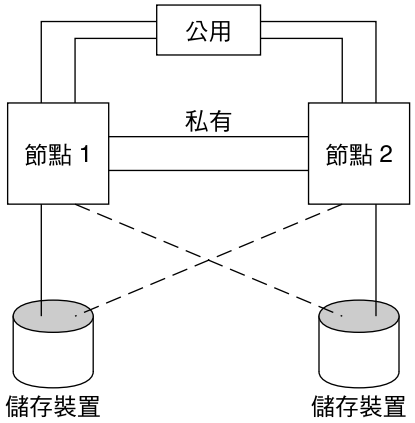


圖 **2–6** x86: 叢集化配對拓撲

第 3 章

針對系統管理員和應用程式開發者的重要概念

本章說明有關 Sun Cluster 系統的軟體元件的重要概念。涵蓋的主題包含：

本資訊主要是供使用 Sun Cluster API 和 SDK 的系統管理員和應用程式開發人員參考。叢集系統管理員可以使用本資訊來準備安裝、配置和管理叢集軟體。應用程式開發人員可以使用這些資訊來瞭解將要利用的叢集環境。

管理介面

您可以從數種使用者介面中選擇一種，以安裝、配置和管理 Sun Cluster 系統。您可以透過 SunPlex Manager 圖形使用者介面 (GUI) 或透過歸檔指令行介面來完成系統管理作業。在指令行介面的頂端有一些公用程式 (如 `scinstall` 和 `scsetup`)，可以簡化選取的安裝和配置作業。Sun Cluster 系統亦有一個模組作爲 Sun Management Center 的一部分執行，爲特定的叢集作業提供 GUI。此模組僅可在基於 SPARC 的叢集中使用。請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理工具」，以取得管理介面的完整說明。

叢集時間

叢集中所有節點的時間均必須同步。不論您是否將叢集節點與任何外在的時間來源同步化，對於叢集操作而言並不重要。Sun Cluster 系統使用網路時間協定 (NTP) 同步化各節點間的時鐘。

一般而言，系統時鐘在傾刻之間變更並不會造成問題。然而，如果您在使用中的叢集上執行 `date(1)`、`rdate(1M)`，或 `xntpdate(1M)`(交談式，或在 `cron` 指令集之內)，您可以強制進行比傾刻更久的時間變更來同步化系統時鐘與時間來源。這種強制變更可能會導致檔案修改時間戳記有問題或混淆 NTP 服務。

在每個叢集節點上安裝 Solaris 作業系統時，您都有機會變更該節點預設的時間和日期設定。一般而言，您可以接受出廠預設值。

使用 `scinstall(1M)` 安裝 Sun Cluster 軟體時，安裝程序中的一個步驟是爲叢集配置 NTP。Sun Cluster 軟體提供一個範本檔案，`ntp.cluster` (請參閱已安裝的叢集節點上的 `/etc/inet/ntp.cluster`)，該檔案會在所有叢集節點之間建立對等關係。一個節點定義爲「喜好的」節點。由專用的主電腦名稱和跨叢集交互連接時發生的時間同步化來識別節點。如需有關如何配置叢集 NTP 的說明，請參閱「Sun Cluster 軟體安裝指南（適用於 Solaris 作業系統）」中的第 2 章「安裝和配置 Sun Cluster 軟體」。

另外一種方式是，您可以在叢集之外設定一或多部 NTP 伺服器，並變更 `ntp.conf` 檔案以反映該配置。

在正常作業中，您應該不會需要調整叢集的時間。然而，您安裝 Solaris 作業系統時時間設定不正確，並且您希望變更時間，「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 7 章「管理叢集」中包含有關該作業的程序。

高可用性框架

Sun Cluster 系統使使用者和資料之間的「路徑」上所有元件 (包括網路介面、應用程式本身、檔案系統和多重主機裝置) 都具有高度可用性。一般而言，如果叢集元件在系統內有任何單一(軟體或硬體) 故障之後仍然存在，就具有高度可用性。

下表顯示了各種 Sun Cluster 元件故障 (硬體和軟體均包括) 和內建於高度可用性框架中的各種回復。

表 3–1 Sun Cluster 故障偵測和回復的層級

故障的叢集元件	軟體復原	硬體恢復
資料服務	HA API、HA 框架	不適用
公用網路配接卡	Internet Protocol (IP) 網路多重路徑	多重公用網路配接卡
叢集檔案系統	主要與次要複製	多重主機裝置
鏡像的多重主機裝置	容體管理 (Solaris Volume Manager 和 VERITAS Volume Manager，它們僅在基於 SPARC 的叢集中可用)	硬體 RAID-5 (例如，Sun StorEdge™ A3x00)
整體裝置	主要與次要複製	至裝置的多重路徑，叢集傳輸接點
私有網路	HA 傳輸軟體	多重私有硬體獨立網路
節點	CMM，failfast 驅動程式	多重節點

Sun Cluster 軟體的高度可用性框架可以快速偵測到節點故障，並會在叢集剩餘節點上爲框架資源建立一個新的等效伺服器。框架資源隨時皆可使用。不受當機節點影響的框架資源在回復期間完全可用。此外，已故障節點的框架資源一經恢復之後，便會成爲可使用。已回復的框架資源不必等待所有其他的框架資源完成回復。

大多數高度可用的框架資源都可回復爲使用此資源的應用程式 (資料服務) 而不需設定。框架資源存取的語意會在各項節點故障時被完整地保留。應用程式完全不會偵測到框架資源伺服器已經移至其他節點。透過使用連結至單一節點的檔案、裝置和磁碟容體，該節點的故障對剩餘節點上的程式是完全不需設定的。如果存在到另一節點上的磁碟的替代硬體路徑，則不需進行設定。其中的一個範例便是使用具有連到多重節點的通訊埠的多重主機裝置。

叢集成員關係監視器

爲了讓資料免於毀損，所有的節點必須對叢集成員達成一致的協議。必要時，CMM 會爲了回應故障而協調叢集服務 (應用程式) 的叢集重新配置。

CMM 從叢集傳輸層接收有關連接到其他節點的資訊。在重新配置期間，CMM 使用叢集交互連接來交換狀態資訊。

偵測到叢集成員關係變更後，CMM 會執行叢集的同步化配置。在已同步的配置中，叢集資源可能會根據新的叢集成員關係重新分配。

與先前的 Sun Cluster 軟體發行版本不同，CMM 完全在核心中執行。

請參閱第 48 頁的 「關於故障隔離」，以取得有關叢集如何防止自身分割爲多重單獨叢集的更多資訊。

Failfast 機制

如果 CMM 偵測到某節點具有嚴重問題，它便會通知叢集框架強制關閉 (當機) 該節點並將其從叢集成員關係中移除。發生此情況的機制稱爲 *failfast*。Failfast 會以兩種方式關閉節點。

- 如果一個節點離開叢集，然後在沒有法定數目的情況下嘗試啓動一個新叢集，則它將被「隔離」，無法存取共用磁碟。請參閱第 48 頁的 「關於故障隔離」，以取得有關 failfast 的此種用途之詳細資訊。
- 如果一個或多個叢集特定的常駐程式終止 (`clexecd`、`rpc.pmfd`、`rgmd` 或 `rpc.ed`)，則 CMM 會偵測到故障並且節點將會當機。

叢集常駐程式終止導致節點當機時，該節點的主控台上將會顯示與以下訊息類似的訊息。

```
panic[cpu0]/thread=40e60: Failfast: Aborting because "pmfd" died 35 seconds ago.
409b8 cl_runtime:__0FZsc_syslog_msg_log_no_argsPviTCPCcTB+48 (70f900, 30, 70df54, 407acc, 0)
%l0-7: 1006c80 000000a 000000a 10093bc 406d3c80 7110340 0000000 4001 fbf0
```

發生當機之後，節點可能會重新啓動並嘗試重新加入叢集。或者，如果叢集由基於 SPARC 的系統組成，則節點可能會保持在 OpenBoot™ PROM (OBP) 提示符號狀態。節點的下一個動作由 `auto-boot?` 參數的設定決定。您可以在 OpenBoot PROM `ok` 提示符號中使用 `eeprom`(1M) 設定 `auto-boot?`。

Cluster Configuration Repository (CCR，叢集配置儲存庫)

CCR 使用兩階段確定演算法作爲更新之用：更新必須在所有叢集成員上都成功完成，否則該更新將會回復。CCR 使用叢集交互連接來套用分散式更新。

注意 – 雖然 CCR 是由文字檔所組成，請絕對不要手動編輯 CCR 檔案。每一個檔案均含有總和檢查記錄，以確保節點之間的一致性。手動更新 CCR 檔案會導致節點或整個叢集停止運作。

CCR 依賴 CMM 來保證叢集只有在到達法定數目時才能執行。CCR 負責驗證整個叢集的資料一致性，必要時執行復原，以及促使資料的更新。

全域裝置

Sun Cluster 系統使用**全域裝置**來提供從任何節點對叢集中任何裝置之叢集範圍內的、高度可用的存取，無論裝置實體上連接到何處。通常，如果節點在提供對全域裝置的存取期間發生故障，Sun Cluster 軟體會自動探索到此裝置的其他路徑並將存取重新導向至該路徑。Sun Cluster 全域裝置包含磁碟、CD-ROM 與磁帶。然而，Sun Cluster 軟體支援的多埠式全域裝置僅限於磁碟。從而，CD-ROM 和磁帶裝置目前不是高度可用的裝置。每部伺服器上的本機磁碟亦不是多埠式，因此不是高可用性裝置。

叢集可以自動爲叢集中的每個磁碟、CD-ROM 和磁帶裝置指定唯一的 ID。這種指定可讓叢集中的任何節點對各個裝置進行一致存取。整體裝置名稱空間是保存於 `/dev/global` 目錄。請參閱第 40 頁的 「全域名稱空間」，以取得更多資訊。

多埠式整體裝置提供一條以上的裝置路徑。因爲多重主機磁碟是由多個節點宿主的磁碟裝置群組的一部分，所以多重主機磁碟具有高度可用性。

裝置 ID 和 DID 虛擬驅動程式

Sun Cluster 軟體透過稱爲 DID 虛擬驅動程式的建構來管理全域裝置。此驅動程式用於將唯一的 ID 自動指定給叢集中的每個裝置，包括多重主機磁碟、磁帶機和 CD-ROM。

DID 虛擬驅動程式是叢集的全域裝置存取功能的主要部分。DID 驅動程式會探測叢集的所有節點並建立唯一磁碟裝置的清單，爲每個裝置指定唯一的主要和次要編號，這些編號在叢集的所有節點上是一致的。對全域裝置的存取是利用唯一的裝置 ID，而不是利用傳統的 Solaris 裝置 ID 執行的，例如用於磁碟的 `c0t0d0`。

此方法可以確保存取磁碟的所有應用程式 (例如使用原始裝置的容體管理程式或應用程式) 在叢集中使用一致的路徑。這種一致性對多重主機磁碟而言特別重要，因爲每個裝置的本機主要編號和次要編號會隨著節點不同而改變，因此也會變更 Solaris 裝置命名慣例。例如，`Node1` 可能將某個多重主機磁碟識別爲 `c1t2d0`，而 `Node2` 可能會將同一磁碟完全不同地識別爲 `c3t2d0`。DID 驅動程式會指定全域名稱 (例如 `d10`)，節點會改用該名稱，爲每個節點指定一致的多重主機磁碟對映。

您可以透過 `scdidadm(1M)` 和 `scgdevs(1M)` 來更新和管理裝置 ID。如需更多資訊，請參閱下列線上說明手冊：

- `scdidadm`(1M)
- `scgdevs`(1M)

磁碟裝置群組

在 Sun Cluster 系統中，所有多重主機裝置均必須受 Sun Cluster 軟體控制。您首先要在多重主機磁碟上建立容體管理程式磁碟群組：Solaris Volume Manager 磁碟組或 VERITAS Volume Manager 磁碟群組 (僅可在基於 SPARC 的叢集中使用)。然後，將容體管理程式磁碟群組註冊爲**磁碟裝置群組**。磁碟裝置群組是一種整體裝置類型。此外，Sun Cluster 軟體自動爲叢集中的每個磁碟和磁帶裝置建立原始的磁碟裝置群組。不過這些叢集裝置群組仍會維持離線狀態，除非您以整體裝置來存取它們。

註冊爲 Sun Cluster 系統提供有關哪些節點具有特定容體管理程式磁碟群組的路徑的資訊。在此，容體管理程式磁碟群組會變成可由叢集內做全域存取。如果一個以上的節點可以寫至 (主控) 磁碟裝置群組，儲存在此磁碟裝置群組上的資料就變得高度可用了。該高度可用的磁碟裝置群組可以用於容納叢集檔案系統。

備註 – 磁碟裝置群組與資源群組無關。一個節點可以控制資源群組 (代表資料服務程序群組)，而另一個節點可以控制正在被資料服務存取的磁碟群組。然而，最佳方法是將儲存特定應用程式的資料之磁碟裝置群組和存放該應用程式資源的資源群組 (應用程式常駐程式) 的資源群組保持在同一節點上。請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的「Relationship Between Resource Groups and Disk Device Groups」，以取得有關磁碟裝置群組和資源群組之間的關聯之更多資訊。

節點使用磁碟裝置群組時，容體管理程式磁碟群組將成爲「全域」群組，因爲其爲基礎磁碟提供多重路徑支援。實體連接到多重主機磁碟的每一個叢集節點均會提供磁碟裝置群組的路徑。

Disk Device Group Failover（磁碟裝置群組防故障備用模式）

因爲磁碟機殼連接至一個以上的節點，當目前主控裝置群組的節點故障時，仍可透過替代路徑來存取該外殼中的所有磁碟裝置群組。主控裝置群組的節點故障不會影響裝置群組的存取，但是在執行恢復與一致性檢查的期間除外。在這段期間內，所有的要求均會暫停執行 (對於應用程式爲透明的)，直到系統恢復使用裝置群組爲止。

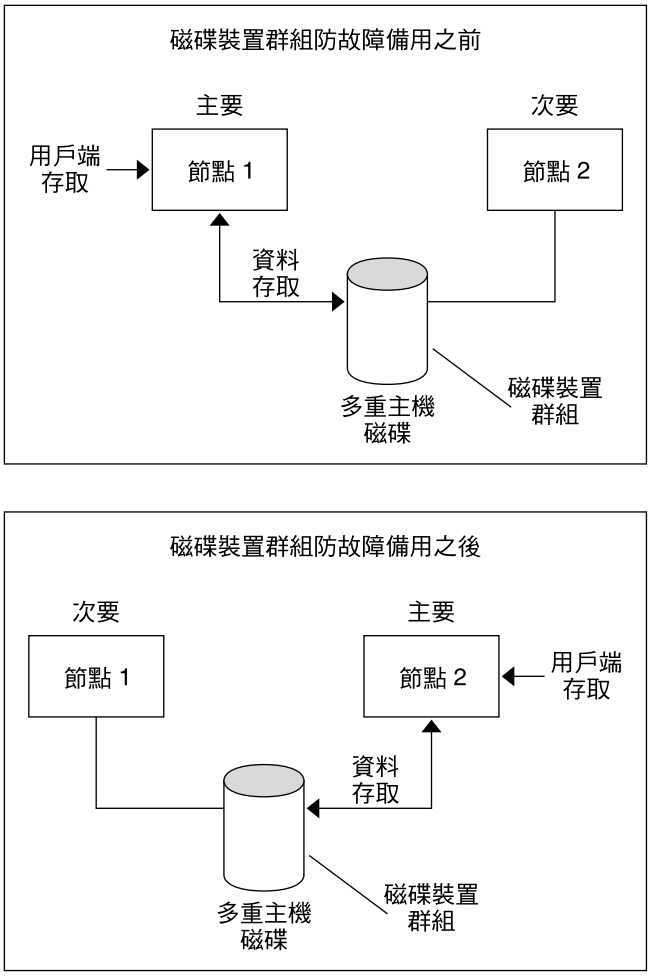


圖 3–1 容錯移轉之前和之後的磁碟裝置群組

多埠式磁碟裝置群組

本節說明可以讓您在多埠式磁碟配置中平衡效能和可用性的磁碟裝置群組特性。Sun Cluster 軟體提供了兩個用於配置多埠式磁碟配置特性：`preferenced` 和 `numsecondaries`。您可以使用 `preferenced` 特性控制發生容錯移轉時節點嘗試採取控制的順序。使用 `numsecondaries` 特性，可爲裝置群組設定所需數目的次要節點。

當主要節點發生故障且沒有符合的次要節點可以成爲主要節點時，會認爲高度可用的服務當機。如果發生服務容錯移轉並且 `preferenced` 特性爲 `true`，則節點會按照節點清單中的順序選取次要節點。由該特性設定的節點清單可以定義節點嘗試採取主要控制或是從備用轉爲次要的順序。您可以使用 `scsetup`(1M) 公用程式動態變更裝置服務的喜好設定。與附屬服務提供者 (例如全域檔案系統) 相關的喜好設定將與裝置服務的喜好設定相同。

在正常作業期間，次要節點是主要節點的核對點。在多埠式磁碟配置中，對每個次要節點進行核對點作業將導致叢集效能降低和記憶體的耗用。備用節點支援用於將檢查點導致的效能降低和記憶體經常性耗用時間降到最低。依預設，磁碟裝置群組有一個主要節點和一個次要的節點。剩餘的可用提供者節點成爲備用節點。如果發生容錯移轉，則次要節點將成爲主要節點，而節點清單中優先權最高的節點將成爲次要節點。

所需次要節點的數目可以設定爲 1 與裝置群組中可作業非主要提供者節點數目之間的任何整數。

備註 – 如果您使用的是 Solaris Volume Manager，您必須建立磁碟裝置群組才能將 `numsecondaries` 特性設定爲預設值之外的數值。

預設的所需裝置服務次要節點數目爲 1。由副本框架維護的次要提供者的實際數目爲所需數目，除非可作業非主要提供者的數目少於所需數目。如果您要在配置中增加或移除節點，則必須更改 `numsecondaries` 特性並仔細檢查節點清單。維護節點清單和次要節點的所需數目可以防止配置的次要節點數目與框架允許的實際數目之間發生衝突。

- (Solaris Volume Manager) 使用 Solaris Volume Manager裝置群組的 `metaset`(1M) 指令以及 `preferenced` 和 `numsecondaries` 特性的設定，管理配置中節點的增加和移除。
- (Veritas Volume Manager) 使用 VxVM 磁碟裝置群組的 `scconf`(1M) 指令以及 `preferenced` 和 `numsecondaries` 特性的設定，管理配置中節點的增加和移除。
- 請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理叢集檔案系統簡介」，以取得有關變更磁碟裝置群組特性的程序資訊。

全域名稱空間

啓用全域裝置的 Sun Cluster 軟體機制是**全域名稱空間**。全域名稱空間包含 `/dev/global/` 階層結構和容體管理程式名稱空間。全域名稱空間反映多重主機磁碟和本機磁碟 (以及任何其他的叢集裝置，如 CD-ROM 和磁帶)，並提供多重主機磁碟的多重故障轉移路徑。每個實體連接至多重主機磁碟的節點會爲叢集中任何的節點提供儲存體的路徑。

通常，對於 Solaris Volume Manager，容體管理程式名稱空間位於 `/dev/md/`*diskset*`/dsk` (以及 `rdsk`) 目錄中。對於 Veritas VxVM，容體管理程式名稱空間位於 `/dev/vx/dsk/`*disk-group* 和 `/dev/vx/rdsk/`*disk-group* 目錄中。這些名稱空間分別由在整個叢集匯入的每個 Solaris Volume Manager 磁碟組和每個 VxVM 磁碟群組之目錄組成。其中的每個目錄均包含該磁碟組或磁碟群組中的每個中介裝置或容體的裝置節點。

在 Sun Cluster 系統中，本機容體管理程式名稱空間中的每個裝置節點均替代爲 `/global/.devices/node@`*nodeID* 檔案系統中裝置節點的符號連結，其中 *nodeID* 是代表叢集中的節點的整數。Sun Cluster 軟體繼續在其標準位置以符號連結表示容體管理程式裝置。全域名稱空間和標準容體管理程式名稱空間均可由任何叢集節點使用。

全域名稱空間的優勢包含以下各項：

- 每個節點保持完全獨立，而在裝置管理模型中可有一點變更。
- 裝置可以選擇性地成爲整體。
- 協力廠商連結產生器繼續運作。
- 給定本機裝置名稱，提供簡易的對應，以獲得其整體名稱。

區域和全域名稱空間範例

下表顯示多重主機磁碟 (`c0t0d0s0`) 的本機和全域名稱空間之間的對應。

表 3–2 本機和全域名稱空間對應

元件或路徑	本機節點名稱空間	全域名稱空間
Solaris logical name (Solaris 邏輯名稱)	`/dev/dsk/c0t0d0s0`	`/global/.devices/node@`*nodeID*`/dev/dsk/c0t0d0s0`
DID name (DID 名稱)	`/dev/did/dsk/d0s0`	`/global/.devices/node@`*nodeID*`/dev/did/dsk/d0s0`
Solaris Volume Manager	`/dev/md/`*diskset*`/dsk/d0`	`/global/.devices/node@`*nodeID*`/dev/md/`*diskset*`/dsk/d0`
SPARC：VERITAS Volume Manager	`/dev/vx/dsk/`*disk-group*`/v0`	`/global/.devices/node@`*nodeID*`/dev/vx/dsk/`*disk-group*`/v0`

全域名稱空間是在安裝和更新的每次重新配置重新開機時自動產生。您也可以執行 `scgdevs(1M)` 指令來產生全域名稱空間。

Cluster File Systems（叢集檔案系統）

叢集檔案系統具備下述功能：

- 檔案存取位置是透明的。程序可以開啓位於系統中任何位置的檔案。所有節點上的程序均可以使用相同的路徑名稱找到檔案。

  備註 – 當叢集檔案系統讀取檔案時，並不會更新這些檔案上的存取時間。

- 使用一致的通訊協定來保持 UNIX 檔案存取語意，即使檔案是從多個節點並行地被存取。
- 廣泛的快取是與 zero-copy bulk I/O 移動一起使用，使檔案資料的移動更有效率。
- 叢集檔案系統透過使用 `fcntl(2)` 介面提供高度可用的建議檔案鎖定功能。透過使用叢集檔案系統上的建議檔案鎖定功能，在多個叢集節點上執行的應用程式可以同步化資料的存取。檔案鎖可立即由離開叢集的節點，以及維持鎖定時故障的應用程式加以回復。
- 即使發生故障時，仍可確保資料的持續存取。只要磁碟的路徑仍然是作業中，應用程式不會受到故障的影響。這項保證適用於原始磁碟存取和所有的檔案系統作業。
- 叢集檔案系統獨立於基礎檔案系統及容體管理軟體。叢集檔案系統可讓任何受支援的磁碟檔案系統都是全域的。

您可以使用 `mount -g` 將檔案系統全域掛載在全域裝置上或使用 `mount` 將其本機掛載在全域裝置上。

程式可以從叢集中的任何節點，透過相同的檔名 (例如，`/global/foo`) 來存取叢集檔案系統中的檔案。

叢集檔案系統會裝載於所有叢集成員上。您不能將叢集檔案系統裝載於叢集成員的子集上。

叢集檔案系統並非不同的檔案系統類型。用戶端驗證基礎檔案系統 (如 UFS)。

使用叢集檔案系統

在 Sun Cluster 系統中，所有多重主機磁碟均置入磁碟裝置群組中，這些群組可以是 Solaris Volume Manager 磁碟組、VxVM 磁碟群組或是不受軟體式容體管理程式控制的個別磁碟。

要使叢集檔案系統爲高度可用，基礎的磁碟儲存體必須連結一個以上的節點。因此，成爲叢集檔案系統的本機檔案系統 (即儲存於節點本機磁碟上的檔案系統) 並不具有高度可用性。

您可以掛載叢集檔案系統，方法與掛載檔案系統相同：

- **手動** – 使用 `mount` 指令和 `-g` 或 `-o global` 掛載選項從指令行掛載叢集檔案系統，例如：

  ```
  SPARC：# mount -g /dev/global/dsk/d0s0 /global/oracle/data
  ```

- **自動** – 使用 `global` 掛載選項在 `/etc/vfstab`檔案中建立一個項目，以在啓動時掛載叢集檔案系統。然後在所有節點的 `/global` 目錄下建立裝載點。目錄 `/global` 是建議使用的位置，而並非要求的位置。以下是來自 `/etc/vfstab` 檔案之叢集檔案系統的範例行：

```
SPARC：/dev/md/oracle/dsk/d1 /dev/md/oracle/rdsk/d1 /global/oracle/data ufs 2 yes global,logging
```

備註 – 因爲 Sun Cluster 軟體沒有強制叢集檔案系統的命名策略，您可以將所有叢集檔案系統的掛載點建立在同一目錄下，例如 `/global/`*disk-device-group*，以簡化管理作業。請參閱「Sun Cluster 3.1 9/04 Software Collection for Solaris OS (SPARC Platform Edition)」和「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」，以取得更多資訊。

`HAStoragePlus` 資源類型

`HAStoragePlus` 資源類型是設計用於使非全域檔案系統配置 (如 UFS 和 VxFS) 具有高度可用性的。使用 `HAStoragePlus` 將您的本機檔案系統整合到 Sun Cluster 環境中，並且使檔案系統具有高度可用性。`HAStoragePlus` 提供附加的檔案系統功能，例如檢查、掛載和強制卸載，這些功能使 Sun Cluster 可以容錯移轉本機檔案系統。本機檔案系統必須位於已啓動切換保護移轉的全域磁碟群組中，才能進行故障轉移。

請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的「Enabling Highly Available Local File Systems」，以取得有關如何使用 `HAStoragePlus` 資源類型的資訊。

`HAStoragePlus` 還可用於同步化資源和資源所依賴的磁碟裝置群組的啓動。如需更多資訊，請參閱第 64 頁的 「資源、資源群組與資源類型」。

`Syncdir` 掛載選項

您可以將 `syncdir` 掛載選項用於使用 UFS 作爲基礎檔案系統的叢集檔案系統。然而，如果您不指定 `syncdir`，效能會明顯改善。如果您指定 `syncdir`，則會保證寫入與 POSIX 相容。如果您不指定 `syncdir`，您將遇到與 NFS 檔案系統中相同的運作方式。例如，不使用 `syncdir`，您可能關閉檔案時才能發覺空間不足的狀況。使用 `syncdir` (和 POSIX 行爲)，便可在寫入作業期間發覺空間不足的狀況。如果不指定 `syncdir` 而遇到問題的情況會很少。

如果您使用的是基於 SPARC 的叢集，則 VxFS 沒有與 UFS `syncdir` 掛載選項等效的掛載選項。未指定 `syncdir` 掛載選項時，VxFS 運作方式與 UFS 的相同。

請參閱第 80 頁的 「檔案系統常見問題」，以瞭解有關全域裝置和叢集檔案系統的常見問題。

磁碟路徑監視

目前發行版本的 Sun Cluster 軟體支援磁碟路徑監視 (DPM)。本節提供了有關 DPM、DPM 常駐程式的概念資訊，以及用於監視磁碟路徑的管理工具。請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」，以取得有關如何監視、取消監視和檢查磁碟路徑狀態的程序資訊。

備註 – 如果節點執行的是 Sun Cluster 3.1 10/03 發行版本之前的軟體版本，則該節點上不支援 DPM。當進行滾動升級時，請勿使用 DPM 指令。在升級了所有節點後，節點必須在線上才能使用 DPM 指令。

DPM 簡介

DPM 可以透過監視次要磁碟路徑的可用性，來提昇故障轉移和切換保護移轉的整體可信賴性。使用 `scdpm` 指令來驗證某個資源在切換之前所使用的磁碟路徑之可用性。`scdpm` 指令隨附的選項可以讓您監視叢集中單一節點或所有節點的路徑。請參閱 `scdpm`(1M) 線上手冊，以取得有關指令行選項的更多資訊。

DPM 元件是從 `SUNWscu` 套件安裝的。`SUNWscu` 套裝軟體是透過標準的 Sun Cluster 安裝程序安裝的。請參閱 `scinstall`(1M) 線上手冊，以取得關於安裝介面的詳細資訊。下表說明了 DPM 元件的預設安裝位置。

位置	元件
常駐程式	`/usr/cluster/lib/sc/scdpmd`
指令行介面	`/usr/cluster/bin/scdpm`
共用檔案庫	`/user/cluster/lib/libscdpm.so`
常駐程式狀態檔 (在執行期間建立)	`/var/run/cluster/scdpm.status`

多重執行緒 DPM 常駐程式在每個節點上執行。當節點啓動時，DPM 常駐程式 (`scdpmd`) 由 `rc.d` 程序檔啓動。如果出現問題，則此常駐程式將由 `pmfd` 管理並自動重新啓動。下列清單說明初始啓動時 `scdpmd` 的工作方式。

備註 – 在啓動時，每個磁碟路徑的狀態都將初始化爲 UNKNOWN。

1. DPM 常駐程式從之前的狀態檔案或從 CCR 資料庫收集磁碟路徑和節點名稱資訊。請參閱第 36 頁的 「Cluster Configuration Repository (CCR，叢集配置儲存庫)」，以取得有關 CCR 的更多資訊。啓動 DPM 常駐程式之後，可以強制此常駐程式從指定的檔案名稱讀取受監視磁碟的清單。
2. DPM 常駐程式可初始化通訊介面 (如指令行介面)，以回應來自此常駐程式外部元件的要求。
3. DPM 常駐程式可使用 `scsi_inquiry` 指令，每隔 10 分鐘在受監視的清單中偵測每個磁碟路徑。將鎖定每個項目，以防止通訊介面存取被修改項目的內容。
4. DPM 常駐程式會通知 Sun Cluster Event Framework 並透過 UNIX `syslogd`(1M) 機制記錄路徑的新狀態。

備註 – `pmfd` (1M) 會報告與常駐程式相關的所有錯誤。API 的所有功能均傳回 `0` 表示成功，傳回 `-1` 表示發生任何故障。

DPM 常駐程式會監視透過多重路徑驅動程式可見的邏輯路徑 (例如 Sun StorEdge Traffic Manager、HDLM 和 PowerPath) 的可用性。將不監視這些驅動程式管理的個別實體路徑，因爲多重路徑驅動程式可遮罩 DPM 常駐程式的個別故障。

監視磁碟路徑

本節說明了監視叢集內磁碟路徑的兩種方法。第一種方法由 `scdpm` 指令提供。使用該指令，可監視、取消監視或顯示叢集內磁碟路徑的狀態。此指令對列印故障磁碟清單和從檔案監視磁碟路徑也非常有用。

SunPlex Manager 圖形使用者介面 (GUI) 提供了監視叢集內磁碟路徑的第二種方法。SunPlex Manager 提供了叢集內受監視磁碟路徑的拓撲檢視。此檢視每 10 分鐘更新一次，以提供關於失敗偵測的數目。請將 SunPlex Manager GUI 提供的資訊與 `scdpm`(1M) 指令配合使用，來管理磁碟路徑。請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 10 章「利用圖形化使用者介面管理 Sun Cluster」，以取得有關 SunPlex Manager 的資訊。

使用 `scdpm` 指令監視磁碟路徑

`scdpm`(1M) 指令提供了可讓您執行下列作業的 DPM 管理指令：

- 監視新的磁碟路徑
- 取消監視磁碟路徑
- 從 CCR 資料庫中重新讀取配置資料
- 從指定的檔案中讀取要監視或取消監視的磁碟

- 報告叢集內某個磁碟路徑或所有磁碟路徑的狀態
- 列印可從節點存取的所有磁碟路徑

從任何使用中節點發出具有磁碟路徑引數的 `scdpm`(1M) 指令，以便對叢集執行 DPM 管理作業。磁碟路徑引數總是由節點名稱與磁碟名稱構成。節點名稱不是必須的，並且如果未指定節點名稱則預設爲 `all`。下列表格說明了磁碟路徑的命名慣例。

備註 – 極力建議您使用全域磁碟路徑名稱，因爲全域磁碟路徑名稱在整個叢集中是一致的。UNIX 磁碟路徑名稱在整個叢集中是不一致的。一個磁碟的 UNIX 磁碟路徑在叢集節點之間可以不同。磁碟路徑可以在一個節點上爲 `c1t0d0`，而在另一個節點上爲 `c2t0d0`。如果您使用 UNIX 磁碟路徑名稱，請在發出 DPM 指令之前，使用 `scdidadm -L` 指令將 UNIX 磁碟路徑名稱對應至全域磁碟路徑名稱。請參閱 `scdidadm`(1M) 線上手冊。

表 3–3 範例磁碟路徑名稱

名稱類型	範例磁碟路徑名稱	描述
整體磁碟路徑	`schost-1:/dev/did/dsk/d1`	`schost-1` 節點上的磁碟路徑 `d1`
`all:d1`	叢集內所有節點上的磁碟路徑 `d1`	
UNIX 磁碟路徑	`schost-1:/dev/rdsk/c0t0d0s0`	`schost-1` 節點上的磁碟路徑 `c0t0d0s0`
`schost-1:all`	`schost-1` 節點上的所有磁碟路徑	
所有磁碟路徑	`all:all`	叢集中所有節點上的全部磁碟路徑

使用 SunPlex Manager 監視磁碟路徑

SunPlex Manager 可讓您執行下列基本的 DPM 管理作業：

- 監視磁碟路徑
- 取消監視磁碟路徑
- 檢視叢集中所有磁碟路徑的狀態

請參考 SunPlex Manager 線上說明，以取得關於如何使用 SunPlex Manager 來執行磁碟路徑管理的程序資訊。

法定數目和法定裝置

本節包含下列主題：

備註 – 如需 Sun Cluster 軟體支援作爲法定裝置的特定裝置清單，請聯絡您的 Sun 服務供應商。

由於叢集節點共用資料與資源，因此叢集永遠不能分割爲同時處於使用中的單個分割區，因爲多個使用中的分割區可能導致資料毀損。叢集成員關係監視器 (CMM) 與法定數目演算法保證同一叢集在任何時候均最多有一個實例處於作業中，即使分割了叢集互連亦是如此。

如需有關法定數目和 CMM 的說明，請參閱「Sun Cluster 簡介 (適用於 Solaris 作業系統)」中的「叢集成員關係」。

叢集分割區中產生的兩種問題：

- Slipt brain
- Amnesia

當節點間的叢集互連遺失且該叢集被分割成子叢集時會出現 Split Brain。每個分割區都「認爲」自己是唯一的分割區，因爲一個分割區中的節點無法與其他分割區中的節點進行通訊。

關機後叢集重新啓動時 (其中叢集配置資料比關機時還舊) 會發生 Amnesia。當您不是在上一次起作用的節點上啓動叢集時可能會發生此問題。

Sun Cluster 軟體透過以下方法避免 Split Brain 與 Amnesia：

- 爲每個節點指定一票
- 託管作業中叢集的多數投票

具有多數投票的分割區獲得**法定數目**，可以進行運作。在一個叢集中配置兩個以上的節點時，該多數投票機制會防止 Split Brain 與 Amnesia。但是，在一個叢集中配置兩個以上的節點時，僅僅計數節點票數是不夠的。在兩個節點的叢集中，票數爲兩票。如果此類包含兩個節點的叢集被分割，則其中一個分割區需要外部投票才能獲得法定數目。該外部投票由**法定裝置**提供。

關於法定票數

使用 `scstat -q` 指令來確定以下資訊：

- 配置的投票總數
- 目前票數
- 法定要求票數

如需有關此指令的更多資訊，請參閱 `scstat`(1M)。

節點與法定裝置均會向叢集投票以形成法定數目。

節點根據節點狀態進行投票：

- 當節點啓動並成爲叢集成員時，其票數爲 *1*。
- 安裝節點時，其票數爲 *0*。
- 當系統管理?將節點置於維護狀態時，節點的票數爲 *0*。

法定裝置根據與該裝置連接的票數進行投票。當您配置法定裝置時，Sun Cluster 軟體爲法定裝置指定票數 *N*-1，其中 *N* 是連接至此法定裝置的票數。例如，與兩個有非零票數節點連線的法定裝置，擁有一票法定票數 (二減一)。

如果滿足以下兩個條件之一，則法定裝置就會投票：

- 至少有一個目前與法定裝置連接的節點是叢集成員。
- 至少有一個目前與法定裝置連接的節點正在啓動，且該節點爲最後一個叢集分割區的成員才能擁有該法定裝置。

您在安裝叢集的過程中配置法定裝置，或者稍後使用「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 5 章「管理法定數目」中說明的程序進行配置。

關於故障隔離

叢集的主要問題是導致叢集被分割的故障 (稱爲 *Split Brain*)。發 split brain 時，不是所有節點均可通訊，所以個別節點或節點子集可能會嘗試形成個別的叢集或子集叢集。每個子集或分割區都可能會「認爲」自己對多重主機裝置擁有唯一的存取權和所有權。當多個節點嘗試寫入磁碟時會發生資料毀損。

故障隔離藉由實際防止磁碟存取來限制節點存取多重主機裝置。當節點離開叢集時 (故障或被分割)，故障隔離可確保節點不會再存取碟。只有目前的成員可以存取磁碟，因此維持了資料的完整性。

磁碟裝置服務爲使用多重主機裝置的服務提供容錯移轉功能。當目前作爲磁碟裝置群組的主要節點 (所有者) 的叢集成員發生故障或無法連線時，將會選擇新的主要節點。新的主要節點可以讓對磁碟裝置的存取繼續，而只發生短暫中斷。在此過程中，舊的主要節點必須喪失對裝置的存取權才能啓動新的主要節點。然而，當成員退出叢集且接觸不到時，叢集就無法通知該主要節點釋放裝置。因此，您需要一個方法讓存活的成員可以從故障的成員接手控制和存取整體裝置。

Sun Cluster 系統使用 SCSI 磁碟保留來實現故障隔離。使用 SCSI 保留，便可以將發生故障的節點與多重主機裝置「隔離」，防止它們存取這些磁碟。

SCSI-2 磁碟保留支援一種形式的保留，它會授予對所有連接到該磁碟的節點的存取權 (當不存在保留時)。或者，存取權被限制爲某一單一節點 (存放該保留的節點)。

當叢集成員偵測到另一個節點在叢集交互連接上已經不再進行通訊，即會起始隔離程序來防止其他節點存取共用磁碟。發生此故障隔離時，隔離的節點將會當機，並在其主控台上顯示「保留衝突」訊息。

如果發現節點不再是叢集成員，則在此節點和其他節點間共用的所有磁碟上觸發 SCSI 保留。隔離的節點可能不「發覺」其已被隔離，並且如果其嘗試存取共用磁碟，則會偵測到保留和當機。

用於故障隔離之 Failfast 機制

叢集框架用於確保故障的節點無法重新啓動和開始寫入共用儲存體的機制稱爲 *failfast*。

叢集成員的節點對於它們有存取權的磁碟，包括法定數目的磁碟，會連續啓用特定的 ioctl，也就是 MHIOCENFAILFAST。Ioctl 是用於磁碟機的指令。Ioctl 使節點可以在因磁碟被其他節點保留而無法存取時將自己當機。

MHIOCENFAILFAST ioctl 會讓磁碟機檢查從每次讀取和寫入節點發送給磁碟的 `Reservation_Conflict` 錯誤代碼時傳回的錯誤。Ioctl 會在背景中定期地對磁碟發出測試作業，以檢查 `Reservation_Conflict`。如果傳回 `Reservation_Conflict`，則前景與背景的控制流程路徑都會發生錯誤。

對於 SCSI-2 磁碟而言，保留並不是永久性的 — 它們並不能在節點重新啓動時存活。對於具有 Persistent Group Reservation (PGR) 的 SCSI-3 磁碟而言，保留資訊是儲存在磁碟上，並且在節點重新啓動後仍會保留。Failfast 機制工作方式相同，無論您使用的是 SCSI-2 磁碟還是 SCSI-3 磁碟。

如果節點在叢集中失去與其他節點的連接，並且也不是可達法定容量的分割區，它會被其他節點強制從叢集中移除。另一可達法定容量之分割區部分的節點，將共用磁碟保留。如果節點嘗試存取共用磁碟次數不是法定數目，則 failfast 機制會導致其收到保留衝突並當機。

在當機之後，該節點可能重新啓動並嘗試重新連結叢集，或者停留在 OpenBoot™ PROM (OBP) 提示符號處 (如果叢集由基於 SPARC 的系統組成)。採用的動作由 `auto-boot?` 參數的設定所決定。在基於 SPARC 的叢集中，您可以在 OpenBoot PROM `ok` 提示符號中使用 `eeprom`(1M) 設定 `auto-boot?`。或者，您可以在基於 x86 的叢集中使用在 BIOS 啓動後選擇性執行的 SCSI 公用程式設定此參數。

關於法定數目配置

以下清單包含關於法定數目配置的事實：

- 法定裝置可包含使用者資料。
- 在 *N* 法定裝置中的每一個都連接到 `1` 到 *N* 個節點和 `N+1` 節點的 N+1 配置中，`1` 到 *N* 的全部節點或任何 *N*/**2** 節點當機的情況下叢集仍可運作。此可用性假定法定裝置運作正常。

- 在單一法定裝置連接到所有節點的 *N* 節點配置中，*N*-**1** 節點中任何節點當機的情況下叢集任可正常運作。此可用性假定法定裝置運作正常。
- 在 *N* 個節點的配置 (其中，單一法定裝置連線至所有節點) 中，如果所有叢集節點均可用，則該法定裝置發生故障時，叢集可免於當機。

如果要瞭解需要避免的法定配置之範例，請參閱第 54 頁的 「不正確的法定數目配置」。如果要瞭解建議使用的法定配置之範例，請參閱第 51 頁的 「建議使用的法定數目配置」。

遵守法定裝置需求

您必須遵守以下需求。如果忽略了這些需求，則您可能會降低叢集的可用性。

- 請確保 Sun Cluster 軟體支援您的特定裝置作爲法定裝置。

備註 – 如需 Sun Cluster 軟體支援作爲法定裝置的特定裝置清單，請聯絡您的 Sun 服務供應商。

Sun Cluster 軟體支援兩種類型的法定裝置：

- 支援 SCSI-3 PGR 保留的多重主機共用磁碟
- 支援 SCSI-2 保留的雙重主機共用磁碟

- 在雙節點配置中，您必須配置至少一個法定裝置以確保其他節點發生故障後單一節點可以繼續運作。請參閱圖 3–2。

如果要瞭解有關需要避免的法定數目配置之範例，請參閱第 54 頁的 「不正確的法定數目配置」。如果要瞭解建議使用的法定數目配置之範例，請參閱第 51 頁的 「建議使用的法定數目配置」。

遵照法定裝置最佳方法

請使用以下資訊來爲您的拓樸評估最佳法定配置：

- 您是否具有可連線至叢集所有節點的裝置？
  - 如果有，請將該裝置配置爲唯一的法定裝置。您**無**需配置其他法定裝置，因爲您的配置已是最佳配置。

注意 – 如果您忽略此需求又新增另一個法定裝置，額外的法定裝置會降低叢集的可用性。

  - 如果沒有，請配置您的雙埠裝置。

- 請確保由法定裝置投票的總票數嚴格少於由節點投票的總票數。否則，如果所有磁碟均不可用，則節點無法形成叢集 (即使所有節點都在正常運作)。

  備註 – 在特定環境下，您可能需要降低叢集總體可用性以滿足您的需要。在這些情形下，可以忽略此最佳方式。然而，不遵守此最佳方式會降低整體可用性。例如，在第 54 頁的 「非典型的法定數目配置」中簡述的配置中，叢集的可用性較低：法定票數超出了節點票數。叢集具有以下特性，即如果失去對 `Nodes A` 和 `Node B` 之間共用儲存體的存取權，則整個叢集發生故障。

  請參閱第 54 頁的 「非典型的法定數目配置」，以瞭解有關此最佳方法的例外情況。
- 請指定共用儲存裝置存取權之每個節點對之間的法定裝置。此法定數目配置會加速故障隔離程序。請參閱第 52 頁的 「多於兩個節點的配置中的法定數目」。
- 一般而言，如果新增法定裝置使得叢集總票數爲偶數，則會降低叢集整體可用性。
- 加入節點或節點當機後，法定裝置會稍微減慢重新配置的速度。因此，除非必要，否則請不要增加更多的法定裝置。

如果要瞭解需要避免的法定數目配置之範例，請參閱第 54 頁的 「不正確的法定數目配置」。如果要瞭解建議使用的法定數目配置之範例，請參閱第 51 頁的 「建議使用的法定數目配置」。

建議使用的法定數目配置

本節提供了建議使用的法定數目配置之範例。如果要瞭解應避免的法定數目配置之範例，請參閱第 54 頁的 「不正確的法定數目配置」。

雙節點配置中的法定數目

需要有兩票法定票數才能形成包含兩個節點的叢集。這兩票可以從兩個叢集節點獲得，或者從一個節點和一個法定裝置獲得。

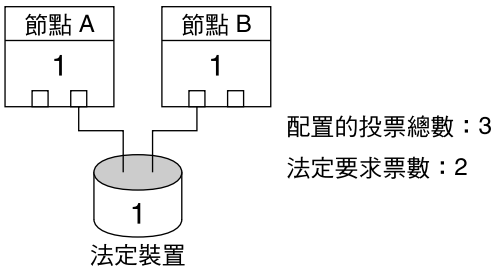
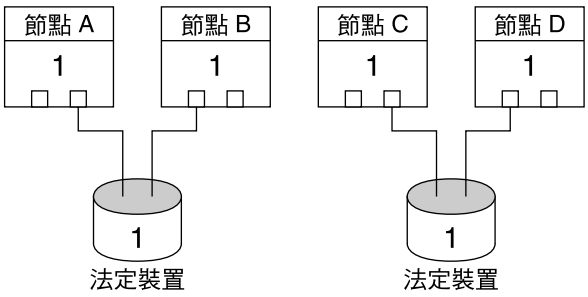


圖 3–2 雙節點配置

多於兩個節點的配置中的法定數目

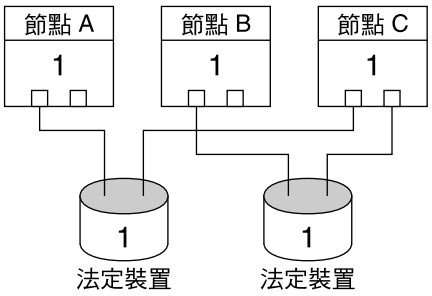
您可以配置多於兩個節點而不包含法定裝置的叢集。然而，如果這樣，則必須使用叢集中的大多數節點才能啓動叢集。



配置的投票總數：6

法定要求票數：4

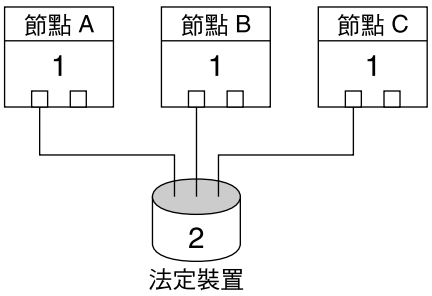
在此配置中，每一對均必須可用，其中一對才可存活。



配置的投票總數：5

法定要求票數：3

在此配置中，通常將應用程式配置為在節點 A 和節點 B 上運行，並使用節點 C 作為緊急備援節點。



配置的投票總數：5

法定要求票數：3

在此配置中，任何一個或多個節點與該法定裝置的組合均可形成一個叢集。

非典型的法定數目配置

圖 3–3 假定在您正在 Node A 和 Node B 上執行關鍵作業應用程式 (如 Oracle database)。如果節點 A 與節點 B 不可用，且無法存取共用資料，則您可能想要使整個叢集當機。否則，該配置爲次佳配置，因爲它沒有提供高度可用性。

如需有關與異常相關的最佳方法的資訊，請參閱第 50 頁的 「遵照法定裝置最佳方法」。

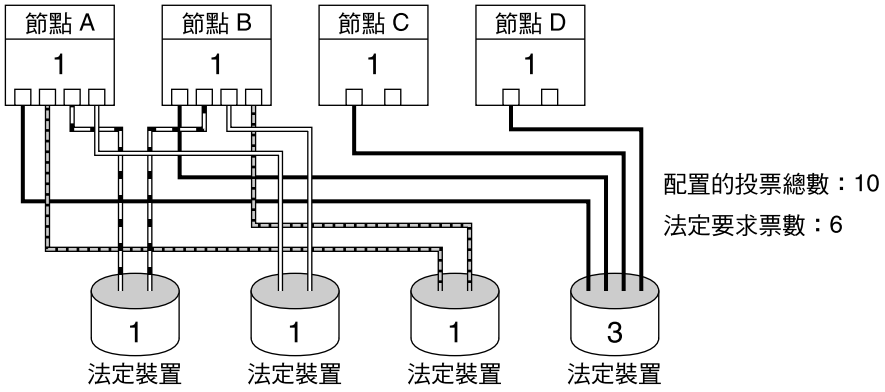
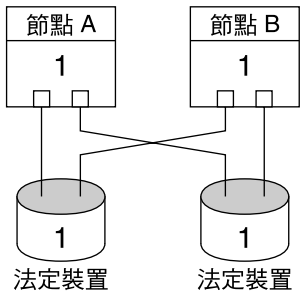


圖 3–3 非典型的配置

不正確的法定數目配置

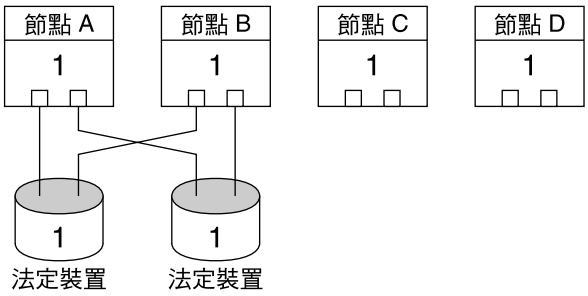
本節提供了應避免的法定數目配置之範例。如果要瞭解建議使用的法定數目配置之範例，請參閱第 51 頁的 「建議使用的法定數目配置」。



配置的投票總數：4

法定要求票數：3

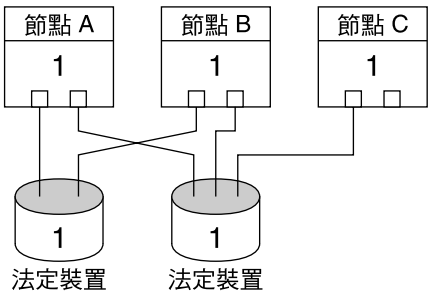
此配置違背了法定裝置票數應該嚴格少於節點票數的最佳方式。



配置的投票總數：6

法定要求票數：4

此配置違背了您不應增加法定裝置以使總票數為偶數的最佳方式。 此配置未增加可用性。



配置的投票總數：5

法定要求票數：3

此配置違背了法定裝置票數應該嚴格少於節點票數的最佳方式。

資料服務

專有名詞**資料服務**說明的是已配置爲在叢集上而不是在單一伺服器上執行的應用程式，例如 Sun Java System Web Server 或 Oracle。資料服務由一個應用程式、專用的 Sun Cluster 配置檔案以及控制應用程式以下動作的 Sun Cluster 管理方法組成。

- 啓動
- 停止
- 監視並採用校正措施

如需有關資料服務類型的資訊，請參閱「Sun Cluster 簡介 (適用於 Solaris 作業系統)」中的「資料服務」。

圖 3–4 將在單一應用程式伺服器 (單一伺服器模型) 上執行的應用程式與在叢集 (叢集伺服器模型) 上執行的同一應用程式進行比較。這兩種配置之間的唯一差異就是叢集應用程式會執行得更快並且可用性更高。

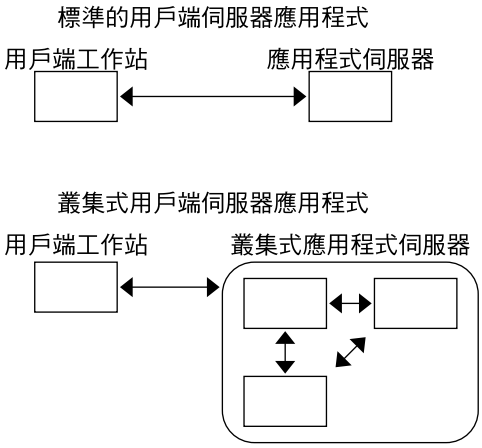


圖 3–4 標準主從式配置與叢集主從式配置

在單一伺服器模型中，您將應用程式配置爲透過特定的公用網路介面 (主機名稱) 存取伺服器。主機名稱與實體伺服器有關。

在叢集伺服器模型中，公用網路介面爲**邏輯主機名稱**或**共用位址**。**網路資源**一詞用於指代邏輯主機名稱和共用位址。

某些資料服務要求您將邏輯主機名稱或共用位址指定爲網路介面。邏輯主機名稱和共用位址不能互換。其他資料服務則容許您指定邏輯主機名稱或共用位址。請參考每個資料服務的安裝和配置，以取得有關必須指定的介面類型的詳細資訊。

網路資源不與特定的實體伺服器相關。網路資源可以在實體伺服器之間遷移。

網路資源與一個節點 (**主要節點**) 初始相關。如果主要節點發生故障，則網路資源和應用程式資源將容錯移轉至其他叢集節點 (次要節點)。當網路資源發生故障轉移時，只要稍有延誤，應用程式資源就繼續在次要節點上執行。

圖 3–5 將單一伺服器模型與叢集伺服器模型進行比較。請注意，在叢集伺服器模型中，網路資源 (在此例中爲邏輯主機名稱) 可於兩或多個叢集節點間移動。應用程式被配置爲使用此邏輯主機名稱，而非與特定伺服器相關的主機名稱。

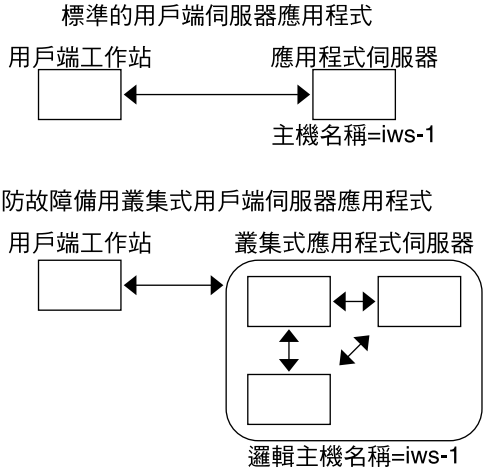


圖 3–5 固定主機名稱與邏輯主機名稱

共用位址也與一個節點初始相關。此節點稱爲全域介面節點。共用位址 (稱爲**全域介面**) 作爲叢集的單一網路介面。

邏輯主機名稱模型和可延伸服務模型之間的差異是：在後者中，每個節點也都在其迴路介面上主動配置了共用位址。此配置使資料服務的多個實例可以同時在多個節點上處於使用中的狀態。「可延伸的服務」一詞表示，您可藉由新增附加的叢集節點來爲應用程式提供更多 CPU 能力，其效能也隨之延伸。

如果全域介面節點發生故障，則可以在也在執行該應用程式實例的其他節點上啓動共用位址 (從而使此節點成爲新的全域介面節點)。但共用位址也可能發生故障轉移而移轉至另一個先前未執行應用程式的節點。

圖 3–6 將單一伺服器配置與叢集可延伸服務配置進行比較。請注意，在可延伸服務配置中，共用位址存在於所有節點上。與邏輯主機名稱用於移轉資料服務方式類似的是，應用程式被配置爲使用此共用位址而不是與特定伺服器相關的主機名稱。

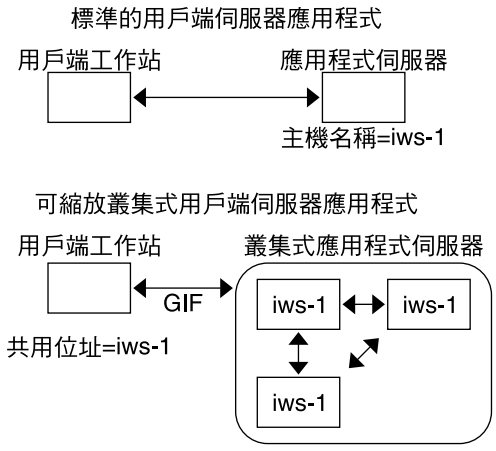


圖 3–6 固定主機名稱與共用位址

資料服務方法

Sun Cluster 軟體提供了一組服務管理方法。這些方法在 資源群組管理員 (RGM) 的控制下執行，它會使用這些方法啓動、停止和監視叢集節點上的應用程式。這些方法配合叢集框架軟體和多重主機裝置，可讓應用程式成爲防故障備用或可延伸的資料服務。

RGM 也會管理叢集內的資源，包括應用程式的實例和網路資源 (邏輯主機名稱和共用位址)。

除 Sun Cluster 軟體提供的方法之外，Sun Cluster 系統也提供了 API 和數種資料服務開發工具。這些工具使應用程式開發人員可以開發所需的資料服務方法，以使用 Sun Cluster 軟體使其他應用程式作爲高度可用的資料服務執行。

故障轉移資料服務

如果正在執行資料服務的節點 (主要節點) 故障，該服務會移轉至其他運作中的節點而不需要使用者介入。容錯移轉服務使用**容錯移轉資源群組**，它是應用程式實例和網路資源 (**邏輯主機名稱**) 的容器。邏輯主機名稱是 IP 位址，其可以在某個節點上配置，稍後在原始節點上自動配置下線並在其他節點上配置。

對於故障轉移資料服務，應用程式實例僅在單一節點上執行。如果故障監視器偵測到錯誤，則會嘗試在同一節點上重新啓動該實例，或在其他節點上啓動該實例 (容錯移轉)。結果取決於資料服務是如何配置的。

可延伸的資料服務

可延伸的資料服務具有在多重節點上的使用中實例之潛力。可延伸服務使用以下兩個資源群組：

- 含有應用程式資源的**可延伸資源群組**。
- 包含可延伸服務所依賴的網路資源 (**共用位址**) 的容錯移轉資源群組。

可延伸資源群組可以在多重節點上成爲線上，所以即可一次執行多個服務實例。放置共用位址的故障轉移資源群組一次只在一個節點上啓動成爲線上。宿主可延伸服務的所有節點均使用相同的共用位址宿主服務。

服務請求透過單一網路介面 (全域介面) 進入叢集。會根據**負載平衡策略**設定的數個預先定義的演算法之一將這些請求將分配到各節點。叢集可以使用平衡資料流量策略，來均衡各個節點之間的服務負載。存放其他共用位址的不同節點上可以存在多個全域介面。

對於可延伸的服務，應用程式實例可同時在數個節點上執行。如果放置整體介面的節點故障，該整體介面會轉移至另一個節點。如果正在執行的應用程式實例發生故障，則該實例將嘗試在同一節點上重新啓動。

如果無法在同一節點上重新啓動應用程式實例，就會配置另一個未使用的節點來執行此服務，該服務便轉移至未使用的節點。否則，該服務將繼續在剩餘的節點上執行，可能導致服務流量降低。

備註 – 每個應用程式實例的 TCP 狀態是保存在具有該實例的節點上，而不是在整體介面節點上。因此，整體介面節點的故障並不會影響連接。

圖 3–7 顯示了容錯移轉和可延伸資源群組的範例，以及兩者之間存在的可延伸服務的相依性。此範例顯示三個資源群組。容錯移轉資源群組包含高度可用的 DNS 之應用程式資源，以及高度可用的 DNS 和高度可用的 Apache Web Server (僅可在基於 SPARC 的叢集中使用) 所使用的網路資源。可延伸資源群組僅包含 Apache Web Server 的應用程式實例。請注意，可延伸和容錯移轉資源群組 (實線) 之間存在資源群組相依性。此外，所有 Apache 應用程式資源都依賴網路資源 `schost-2`，該資源爲共用位址 (虛線)。

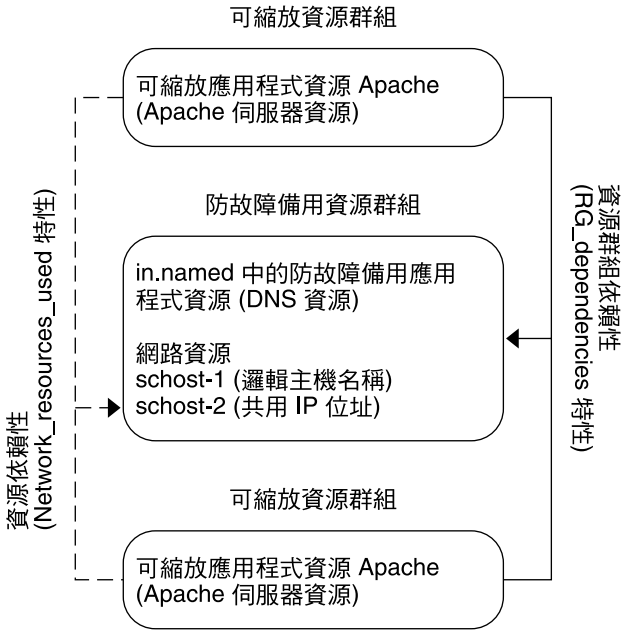


圖 3–7 SPARC: 故障轉移與可延伸的資源群組範例

平衡資料流量策略

平衡資料流量可以在回應時間及產量上增進可延伸服務的效能。可延伸資料服務有兩類。

- Pure
- Sticky

Pure 服務可以使其任何實例回應用戶端請求。*Sticky* 服務可以使一個用戶端向同一實例傳送請求。那些要求不會重新導向至其他實例。

Pure 服務使用加權平衡資料流量策略。在此平衡資料流量策略下，依預設用戶端要求會平均地分配給叢集中的伺服器實例。例如，在一個三節點叢集中，假定每個節點的權重爲 1。則每個節點將代表服務處理所有來自用戶端的請求的 1/3。管理員可以透過 `scrgadm(1M)` 指令介面或 SunPlex Manager GUI 隨時變更權重。

Sticky 服務具有兩種類型：*ordinary sticky* 和 *wildcard sticky*。Sticky 服務使在多個 TCP 連線上的並行應用程式級階段作業可以用 in-state 記憶體 (應用程式階段作業狀態)。

Ordinary sticky 服務使用戶端可以共用多個並行 TCP 連線之間的狀態。稱該用戶端對偵聽單一連接埠的伺服器實例「sticky」。只要實例維持啓動與可存取的狀態，且此服務處於線上狀態時負載平衡策略爲變更，便可保證用戶端的所有請求均傳送至同一伺服器實例。

例如，用戶端上的 Web 瀏覽器使用三個不同的 TCP 連線透過連接埠 80 連線至的共用 IP 位址。然而，連線會在服務中交換它們之間快取的階段作業資訊。

Sticky 策略的一般化會延伸至在背景中並且在同一實例上交換階段作業資訊的多個可延伸服務。當這些服務在背景中並且在同一實例上交換階段作業資訊時，則稱該用戶端對同一節點上偵聽不同連接埠的多個伺服器實例「sticky」。

例如，電子商務網站上的某位顧客透過在連接埠 80 上使用 HTTP 將購物車裝滿商品。然後該顧客切換至連接埠 443 上的 SSL 來傳送安全資料以使用信用卡支付購物車內的商品。

Wildcard sticky 服務使用動態指定的連接埠號碼，但仍然希望用戶端請求傳送到同一節點。用戶端在具有同一 IP 位址的連接埠上「sticky wildcard」。

這種策略的典型範例是被動模式 FTP。例如，某用戶端連線至連接埠 21 上的 FTP 伺服器。伺服器則會指示該用戶端連線回動態連接埠範圍內的偵聽程式埠伺服器。此 IP 位址的所有請求均轉發至同一節點伺服器透過控制資訊通知用戶端。

依預設，對於其中的每個 sticky 策略，加權式負載平衡策略均有效。因而，會將用戶端的初始請求導向至負載平衡器指定的實例。用戶端爲實例正在其上執行的節點建立關聯之後，會有條件地將以後的請求導向至該實例。節點必須可以存取且負載平衡策略必須未變更。

有關特定的負載平衡策略的其他詳細資訊如下。

- 加權式。這項載入會按照指定的加權值來分配到各種節點。此策略可以透過使用 `Load_balancing_weights` 特性的 `LB_WEIGHTED` 值設定。如果節點的權重未明確設定時，則此節點的權重將預設爲「一」。

  加權策略可將一定百分比的用戶端通訊重新導向某個特定節點。假定 X=權重且 A=所有使用中節點的總權重，則使用中的節點的所有新連線中大約 X/A 的連線將會導向至該使用中的節點。然而，連線的總數必須足夠大。此策略不針對個別的要求。

  請注意，此策略並非全體循環式。round-robin 策略總是會使來自某個用戶端的每個請求都傳送至其他節點。例如，第一個請求會傳送至節點 1，第二個請求會傳送至節點 2，以此類推。
- Sticky。在此策略中，會於配置應用程式資源時知道通訊埠集合。此策略可以透過使用 `Load_balancing_policy` 資源特性的 `LB_STICKY` 值設定。
- Sticky-wildcard。此策略是一般「Sticky」策略的超集合。對於透過 IP 位址識別的可延伸服務，由伺服器爲其指定連接埠 (並且事先並不知道)。通訊埠可能會變更。此策略可以透過使用 `Load_balancing_policy` 資源特性的 `LB_STICKY_WILD` 值設定。

故障回復設定

資源群組因故障轉移，從某個節點移轉至另一個節點。發生容錯移轉時，原來的次要節點將成爲新的主要節點。故障回復設定可以指定當原來的主要節點回到線上後將發生的動作。此選項是要使原來的主要節點再次成爲主要節點 (故障回復) 或維持目前的主要節點。您可以使用 `failback` 資源群組特性設定指定要使用的選項。

如果原來存放資源群組的節點發生故障並反復重新啓動，則設定故障回復會導致資源群組的可用性降低。

資料服務錯誤監視器

每個 Sun Cluster 資料服務均提供一個故障監視器，可定期探測資料服務以確定其運作狀態。故障監視器驗證應用程式常駐程式是否在執行，以及用戶端是否正在接受服務。根據探測傳回的資訊，可以啓動預先定義的動作，如重新啓動常駐程式或進行容錯移轉。

項o新的資料服務

Sun 提供配置檔案與管理方法範本，讓您得以使各種應用式在叢集中以故障轉移或可延伸的服務來運作。如果 Sun 未提供您要作爲容錯移轉或可延伸服務執行的應用程式，您還可使用替代方案。使用 Sun Cluster API 或 DSET API 將應用程式配置爲作爲容錯移轉或可延伸服務執行。然而，並非所有應用程式都可以成爲可延伸服務。

可延伸服務的特徵

一組可以確定應用程式是否可以成爲可延伸服務的條件。若要確定應用程式是否可以成爲可延伸服務，請參閱「Sun Cluster 資料服務開發者指南 (適用於 Solaris 作業系統)」中的「分析應用程式的適當性」。以下對這組條件進行了總結。

- 首先，這樣的服務由一個或多個伺服器**實例**組成。每一個實例執行於不同的叢集節點上。同一節點無法執行相同服務的兩個或多個實例。
- 其次，如果服務提供外部邏輯資料儲存區，則作業時應謹慎。多個伺服器實例對此儲存區同的並行存取必須同步化，以避免更新或讀取資料在變更時遺失。請注意，「外部」是用於與處於記憶體中狀態的儲存區相區分。而專有名詞「邏輯」表示該儲存區作爲單一實體出現 (儘管其本身可能是複製的)。此外，邏輯資料服務儲存區具有此該特性，即只要伺服器實例更新儲存區其他實例便立即可以「看到」該更新。

  Sun Cluster 系統透過叢集檔案系統和全域原始分割區提供此外部儲存體。 例如，假設服務會寫入新的資料到外部登錄檔，或就地修改現存的資料。此伺服器的多個實例執行時，每個實例擁有對此外部記錄檔的存取權，並且每個實例可以同時存取此記錄檔。每一個實例必須將此登錄的存取同步化，否則實例會互相干擾。可以使用一般 Solaris 檔案 `fcntl(2)` 和 `lockf(3C)` 鎖定該服務，以達到所需的同步化。

  此儲存區類型的另一範例是後端資料庫，例如基於 SPARC 的或 Oracle 的高度可用的 Oracle Real Application Clusters Guard。此類型的後端資料庫伺服器透過使用資料庫查詢或更新作業事件來提供內建同步化。因此，多個伺服器實例無需實作其自己的同步化。

Sun IMAP 伺服器就是非可延伸服務的服務之範例。服務會更新儲存體，但是該儲存體是私有的，而且當多個 IMAP 實例寫入此儲存體時，會因爲未同步化而彼此覆寫。IMAP 伺服器必須要改寫，以同步化並行存取。

- 最後，請注意實例可包含與其他實例分離的私有資料。在這種情況下，服務無需將並行存取同步化，因爲資料是私有的，並且只有該實例可以處理它。在這種情況下，您必須注意不要將此私有資料儲存在叢集檔案系統下，因爲此資料會變爲可全域存取的資料。

資料服務 API 與資料服務檔案庫 API

Sun Cluster 系統提供以下項目，以使應用程式具有高度可用性：

- 作爲 Sun Cluster 系統的一部分提供的資料服務
- 資料服務 API
- 資料服務的開發程式庫 API
- 「一般」資料服務

「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」說明如何安裝和配置 Sun Cluster 系統隨附的資料服務。「Sun Cluster 3.1 9/04 Software Collection for Solaris OS (SPARC Platform Edition)」說明如何在 Sun Cluster 框架下將使其他應用程式具有高度可用性。

Sun Cluster API 使應用程式開發人員能夠開發可以啓動和停止資料服務實例的故障監視器和程式檔。使用這些工具，應用程式可以實作實爲容錯移轉或可延伸服務。Sun Cluster 系統提供「通用」資料服務。使用此通用資料服務可以快速產生應用程式所需的啓動和停止方法，並可以該資料服務實作爲容錯移轉或可延伸服務。

使用資料服務通訊的叢集交互連接

叢集在節點之間必須具備多網路連接，以形成叢集交互連接。Sun Cluster 軟體使用多重互連以達到以下目標：

- 確保高度可用性
- 改善效能

對於內部流量 (例如檔案系統資料或延伸服務資料)，訊息以 round-robin 方式透過所有的可用互連平行儲存。叢集交互連接也可以用於應用程式，以便在節點之間建立高可用性通訊。例如，分散式應用程式可能會有元件在多個需要通訊的節點上執行。如果使用叢集交互連接而不是公用傳輸，可以防制個別連結的故障。

要在節點之間使用叢集互連進行通訊，應用程式必須使用在 Sun Cluster 安裝過程中配置的私有主機名稱。例如，如果 `node 1` 的私有主機名稱是 `clusternode1-priv`，則使用該名稱透過叢集互連與 `node 1` 通訊。使用此名稱開啓的 TCP 通訊端透過叢集互連佈置路由，並且如果網路發生故障還可以重新佈置路由而不需設定。

因爲您可以在 Sun Cluster 安裝過程中配置私有主機名稱，所以叢集互連會使用您在安裝時選擇的任何名稱。若要確定實際名稱，請使用 `scha_cluster_get`(3HA) 指令和 `scha_privatelink_hostname_node` 引數。

應用程式通訊和內部叢集通訊均透過所有互連平行儲存。因爲應用程式與內部叢集流量共用叢集互連，所以應用程式可用的頻寬取決於其他叢集流量使用的頻寬。如果發生故障，內部流量和應用程式流量將透過所有可用的互連平行儲存。

還會爲每個節點指定一個固定的 pernode 位址。此 pernode 位址綁定在 `clprivnet` 驅動程式上。該 IP 位址對映至以下節點的私有主機名稱：`clusternode1-priv`。如需有關 Sun Cluster 私有網路驅動程式的資訊，請參閱 `clprivnet`(7) 線上手冊。

如果應用程式要求在各方面均一致的 IP 位址，則請在用戶端和伺服器上均進行配置，以將應用程式連結至 pernode 位址。則所有顯示的連線便均來自並傳回 pernode 位址。

資源、資源群組與資源類型

資料服務利用了多種類型的**資源**：應用程式 (例如 Sun Java System Web Server 或 Apache Web Server) 使用應用程式所依賴的網路位址 (邏輯主機名稱和共用位址)。應用程式和網路資源形成受 RGM 管理的基本單位。

資料服務式資源類型。例如，Sun Cluster HA for Oracle 屬於資源類型 `SUNW.oracle-server`，而 Sun Cluster HA for Apache 屬於資源類型 `SUNW.apache`。

資源是在整個叢集中定義的**資源類型**的個體化。定義了數種資源類型。

網路資源屬於 `SUNW.LogicalHostname` 或 `SUNW.SharedAddress` 資源類型。這兩種資源類型由 Sun Cluster 軟體預先註冊。

`HAStorage` 和 `HAStoragePlus` 資源類型用於將資源和其所依賴的磁碟裝置群組的啓動同步化。這些資源類型可以確保在資料服務啓動前，叢集檔案系統的掛載點、全域裝置和裝置群組名稱的路徑均爲可用。如需更多資訊，請參閱「*Data Services Installation and Configuration Guide*」中的「Synchronizing the Startups Between Resource Groups and Disk Device Groups」。在 Sun Cluster 3.0 5/02 中，`HAStoragePlus` 資源類型已經可用，並增加了其他功能，從而使本機檔案系統具有高度可用性。如需有關此功能的更多資訊，請參閱第 43 頁的 「`HAStoragePlus` 資源類型」。

RGM 管理的資源被置入到稱爲**資源群組**的群組中，以便可以將其作爲一個整體來管理。如果在資源群組上啓動了故障轉移或切換保護移轉，則資源群組會被當作一個單位來遷移。

備註 – 當您使含有應用程式資源的資源群組線上運作時，則此應用程式便會啓動。資料服務啓動方法等待應用程式進入執行狀態後才會成功結束。判斷應用程式何時啓動與執行的方式，與資料服務故障監視器判斷資料服務是否仍在服務用戶端的方式相同。請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」，以取得有關此程序的更多資訊。

資源群組管理員 (RGM)

RGM 可控制資料服務 (應用程式) 作爲資源 (由**資源類型**實作來管理)。這些實施由 Sun 提供，或由開發人員以一般資料服務範本、資料服務開發檔案庫 API (DSDL API) 或資源管理 API (RMAPI) 所建立。叢集管理員建立並管理稱爲**資源群組**的容器中的資源。RGM 停止和啓動所選取節點上的資源群組，以回應叢集成員變更。

RGM 作用於**資源**及**資源群組**。 RGM 動作會使致資源和資源群組在上線和離線狀態之間切換。第 65 頁的 「資源及資源群組狀態與設定值」一節中提供了有關可以適用於資源和資源群組的狀態和設定的完整說明。

請參閱第 66 頁的 「資料服務專案配置」，以取得有關如何在 RGM 控制下啓動 Solaris 專案的資訊。

資源及資源群組狀態與設定值

管理者將靜態設定值套用到資源與資源群組中。這些設定值只可經由管理動作來變更。RGM 將資源群組在動態「狀態」之間切換。這些設定值與狀態的說明列於下述清單中。

- 管理或不管理 – 這些都是僅套用在資源群組上的全叢集設定值。資源群組由 RGM 管理。`scrgadm(1M)` 指令可用於使 RGM 管理或不管理資源群組。這些設定值不會隨著叢集再配置而變更。

  在建立第一個資源群組時，它是不被管理的。必須先管理資源群組，該群組中的資源才能進入使用中的狀態。

  在某些資料服務 (如可延伸 Web 伺服器) 中，必須在網路資源啓動之前以及停止後進行作業。此工作是由 initialization (INIT) 及 finish (FINI) 資料服務方法來達成。INIT 方法只有在資源所在的資源群組在被管理狀態時才會執行。

  當資源群組由不管理移向管理的狀態時，任何用於群組已註冊的 INIT 方法都會在群組的資源上執行。

  當資源群組由管理移向不管理的狀態時，任何已註冊的 INIT 方法都會被呼叫以執行清除。

  INIT 和 FINI 方法最常用於可延伸服務的網路資源。然而，您可以將這些方法用於任何不由此應用程式執行的初始化和清除工作。
- 啓用或停用 – 這些都是套用至資源的全叢集設定值。可以使用 `scrgadm(1M)` 指令來啓用或停用資源。這些設定值不會隨著叢集再配置而變更。

資源的正常設定值爲，它在系統中是啓用且主動執行的。

如果您希望使資源在所有叢集上均不可用，請停用此資源。停用的資源不作爲一般用途。

- 線上或離線 – 這些都是套用於資源與資源群組的動態狀態。

  上線和離線狀態隨透過切換移轉或容錯移轉過程中的叢集重新配置步驟進行的叢集作業事件而改變。您還可以透過管理動作來變更這些狀態。使用 `scswitch(1M)` 指令變更資源或資源群組的上線或離線狀態。

  在任何時間，故障轉移資源或資源群組只能在一個節點上爲線上。可延伸的資源或資源群組可以在某些節點上處於線上狀態，而在其他節點上處於離線狀態。在切換保護移轉或故障轉移期間，資源群組及其群組內的資源會在一個節點上離線，然後在另一個節點上連線。

  如果資源群組處於離線狀態，則其所有資源均處於離線狀態。如果資源群組處於上線狀態，則其啓用的所有資源均處於上線狀態。

  資源群組含有數種資源，在各資源間具有相依性。這些相依性要求資源要以特定次序連到線上及離開線上。連到線上及離開線上的方法，可能對於各個資源會花費不同的時間。因有資源相依性及隆1與結束的時間差異，在叢集重新配置時單一資源群組內的資源會有不同的線上及離線狀態。

資源和資源群組特性

您可以爲 Sun Cluster 資料服務配置資源和資源群組的特性值。標準特性常見於所有資料服務中。延伸特性則特定於個別的資料服務。部分標準和延伸特性是以預設值配置的，所以您不需要修改它們。其他特性則需要在建立和配置資源時加以設定。各資料服務的說明文件會指定可設定哪些資源特性，及設定的方式。

標準特性是用來配置通常與任何特定資料服務無關的資源和資源群組特性。如果要瞭解標準特性集，請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的附錄 A「Standard Properties」。

RGM 延伸特性提供了諸如應用程式二進位檔案及配置檔案之位置的資訊。您要依照資料服務的配置方式來修改延伸特性。在資料服務的個別指南中描述了延伸特性集。

資料服務專案配置

當使用 RGM 使資料服務上線運作後，可以將其配置爲以 Solaris 專案名稱啓動。此配置可將 RGM 管理的資源或資源群組與 Solaris 專案 ID 關聯起來。從資源或資源群組至專案 ID 的對映使您可以使用 Solaris 作業系統中提供複雜的控制項，以管理叢集內的工作負荷量和消耗量。

備註 – 僅當您所執行的 Sun Cluster 軟體當前的發行版本不低於 Solaris 9 時才可以執行此配置。

在 Sun Cluster 環境中使用 Solaris 管理功能，可以保證在與其他應用程式共用節點時，最重要的應用程式獲得優先權。如果您已經合併了服務或應用程式已經進行了故障轉移，則應用程式可能會共用一個節點。使用此處說明的管理功能可以防止低優先權的應用程式過度消耗系統資源 (如 CPU 時間)，以提高重要應用程式的可用性。

備註 – 此功能的 Solaris 文件將 CPU 時間、程序、作業和類似元件稱爲「資源」。同時，Sun Cluster 文件將受 RGM 控制的實體稱爲「資源」。在下一節中，專有名詞「資源」是指受 RGM 控制的 Sun Cluster 實體。在該節中，專有名詞「供應」是指 CPU 時間、程序和作業。

本節提供配置資料服務以在指定的 Solaris 9 `project`(4) 中啓動程序的概念說明。本節還說明數種容錯移轉分析藍本和有關打算使用 Solaris 提供的管理功能的建議。

如需有關該管理功能詳細的概念和程序文件，請參閱「System Administration Guide: Network Services」中的第 1 章「Network Service (Overview)」。

當配置資源和資源群組以在叢集中使用 Solaris 管理功能，請使用以下高階程序：

1. 將應用程式配置爲資源的一部分。
2. 將資源配置爲資源群組的一部分。
3. 啓用資源群組中的資源。
4. 使資源群組受管理。
5. 爲資源群組建立 Solaris 專案。
6. 配置標準特性，以將資源群組名稱與在步驟 5 中建立的專案名稱相關聯。
7. 讓資源群組上線運作。

若要配置標準 `Resource_project_name` 或 `RG_project_name` 特性以將 Solaris 專案 ID 與資源或資源群組相關聯，請使用 `-y` 選項和 `scrgadm`(1M) 指令。將特性值設定爲資源或資源群組。請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的附錄 A「Standard Properties」，以瞭解特性定義。請參閱 `r_properties`(5) 和 `rg_properties`(5)，以取得特性說明。

指定的名稱必須在專案資料庫 (`/etc/project`) 中，並且必須將超級使用者配置爲已命名的專案之成員。請參閱「System Administration Guide: Solaris Containers-Resource Management and Solaris Zones」中的第 2 章「Projects and Tasks (Overview)」，以取得有關專案名稱資料庫的概念資訊。請參閱 `project`(4)，以取得專案檔案語法的說明。

若 RGM 使資源或資源群組線上運作，它便啓動了專案名稱下的相關程序。

備註 – 使用者可以隨時將資源或資源群組與專案關聯起來。然而，直到資源或資源群組離線並使用 RGM 重新使其線上運作之後，新的專案名稱才會生效。

啓動專案名稱下的資源與資源群組可讓您配置下列功能，以便在整個叢集內管理系統供給品。

- 延伸記帳 – 以作業或程序爲基礎，提供記錄耗用量的靈活方式。延伸記帳可讓您檢查歷史使用情況，並評估用於未來工作量的容量需求。
- 控制 – 提供限制系統供給品的機制。可以防止程序、作業與專案耗用大量指定的系統供給品。
- 公平共用排程 (FSS) – 可根據工作量的重要性，控制在它們之間分配可用的 CPU 時間。工作量重要性採用您指定給每個工作量的 CPU 時間份額數來表示。請參閱以下線上手冊，以取得更多資訊。
  - `dispadmin`(1M)
  - `priocntl`(1)
  - `ps`(1)
  - `FSS`(7)
- 儲存區 – 可依據應用程式需求，使用互動式應用程式的分割區。可以使用儲存區來分割可支援多個不同軟體應用程式的伺服器。使用儲存區使得可以預測針對每個應用程式回應的可能性更大。

確定專案配置的需求

在您配置資料服務以在 Sun Cluster 環境中使用 Solaris 提供的控制項之前，您必須決定如何在切換移轉或容錯移轉間控制和追蹤資源。在配置新的專案之前識別叢集中的相依性。例如，資源與資源群組依賴磁碟裝置群組。

使用由 `scrgadm`(1M) 配置的 `nodelist`、`failback`、`maximum_primaries` 和 `desired_primaries` 資源群組特性識別資源群組之節點清單特性。

- 如需資源群組和磁碟裝置群組之間節點清單相依性之概要討論，請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的「Relationship Between Resource Groups and Disk Device Groups」。
- 如需詳細的特性說明，請參閱 `rg_properties`(5)。

使用由 `scrgadm`(1M) 和 `scsetup`(1M) 配置的 `preferenced` 特性和 `failback` 特性確定磁碟裝置群組節點清單優先權。

- 如需有關 `preferenced` 特性的概念資訊，請參閱第 39 頁的 「多埠式磁碟裝置群組」。
- 如需程序資訊，請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理磁碟裝置群組」 中的「如何變更磁碟裝置特性」。
- 如需有關節點配置和容錯移轉與可延伸資料服務之運作方式的概念資訊，請參閱第 19 頁的 「Sun Cluster 系統硬體和軟體元件」。

如果您以相同方式配置所有的叢集節點，將在主要節點與次要節點上以相同方式執行使用限制。對於所有節點上配置檔案中所有的應用程式，專案的配置參數無需相同。至少該應用程式所有潛在主要節點上的專案資料庫必須可以存取所有與應用程式相關聯的專案。假定應用程式 1 由 *phys-schost-1* 控制，但會可能切換移轉至或容錯移轉至 *phys-schost-2* 或 *phys-schost-3*。與應用程式 1 相關聯的專案必須可以在所有三個節點 (*phys-schost-1*、*phys-schost-2* 和 *phys-schost-3*) 上進行存取。

備註 – 專案資料庫資訊可以是本機 `/etc/project` 資料庫檔案或者可以儲存在 NIS 對映或 LDAP 目錄服務中。

Solaris 作業系統可以靈活配置使用參數，並且 Sun Cluster 強制的限制很少。配置選項取決於網站的需要。在配置系統之前，請考慮下列章節中的一般準則。

設定每個程序的虛擬記憶體限制

請將 `process.max-address-space` 控制設定爲以每個程序爲基礎來限制虛擬記憶體。請參閱 `rctladm`(1M)，以取得有關設定 `process.max-address-space` 值的詳細資訊。

當您使用 Sun Cluster 軟體的管理控制項時，適當地配置記憶體限制以防止不必要的應用程式容錯移轉和應用程式的「交替」效果。一般情況，請遵守以下規範。

- 不要將記憶體限制設定得太低。

  當應用程式達到它的記憶體限制時，它可能會發生故障轉移。若達到虛擬記憶體限制可以產生非預期的結果，則此準則對於資料庫應用程式而言尤其重要。

- 不要在主要節點及次要節點上以相同方式設定記憶體限制。

  當應用程式達到記憶體限制並將故障轉移至具有相同記憶體限制的次要節點時，相同的限制可導致交替效果。在次要節點上，將記憶體限制設定得稍微高些。記憶體限制的差異可幫助防止交替情形的發生，並爲系統管理員提供依需要調整參數的時間。

- 請使用資源管理記憶體限制來平衡資料流量。

  例如，您可以使用記憶體限制來防止發生錯誤的應用程式耗用過多的交換空間。

故障轉移方案

您可以配置管理參數，以便專案配置 (`/etc/project`) 中的分配可在一般的叢集作業中以及在切換保護移轉或故障轉移情形下運作。

下列章節爲方案範例。

- 前兩節「具有兩個應用程式的兩個節點叢集」與「具有三個應用程式的兩個節點叢集」顯示了全體節點的故障轉移方案。
- 「僅資源群組的故障轉移」一節闡明了僅一個應用程式的故障轉移作業。

在 Sun Cluster 環境中，將應用程式配置爲資源的一部分然後將資源配置爲資源群組 (RG) 的一部分。如果發生故障，則資源群組連同與其相關聯的應用程式都將容錯移轉至其他節點。在下列範例中不明確顯示資源。假定每個資源僅有一個應用程式。

備註 – 故障轉移以 RGM 中設定的個人喜好節點清單順序發生。

下列範例具有這些限制：

- 應用程式 1 (App-1) 在資源群組 RG-1 中配置。
- 應用程式 2 (App-2) 在資源群組 RG-2 中配置。
- 應用程式 3 (App-3) 在資源群組 RG-3 中配置。

雖然指定的份額數相同，但分配給每個應用程式的 CPU 時間百分比將在故障轉移後發生變更。此百分比取決於節點上執行的應用程式數目，以及指定給每個使用中應用程式的份額數。

在這些情形下，假定下列配置。

- 在共用專案下配置所有應用程式。
- 每個資源僅有一個應用程式。
- 應用程式是節點上唯一處於使用中的程序。
- 在叢集的每個節點上配置專案資料庫的方式相同。

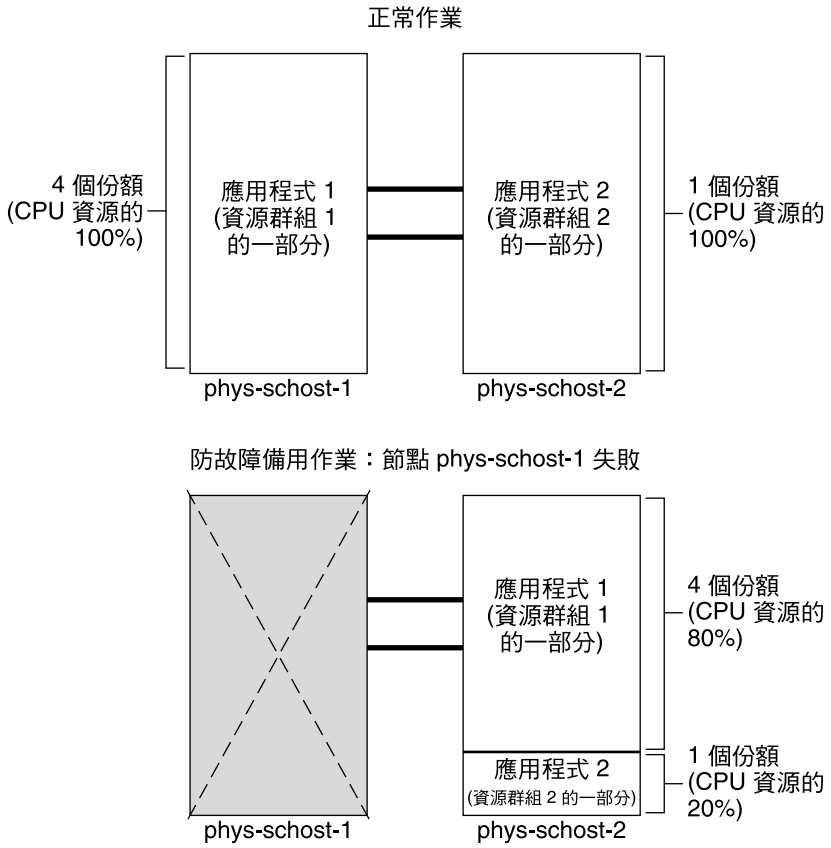
具有兩個應用程式的兩個節點叢集

您可以在一個雙節點叢集上配置兩個應用程式，以保證每個實體主機 (*phys-schost-1* 和 *phys-schost-2*) 作爲一個應用程式的預設主要節點。每個實體主機可作爲另一個實體主機的次要節點。與應用程式 1 和應用程式 2 相關聯的所有專案必須出現在兩個節點上的專案資料庫檔案中。當叢集正常執行時，每個應用程式將在其預設主控者上執行，在此位置上將藉由管理設備爲每個應用程式分配所有 CPU 時間。

發生故障轉移或切換保護移轉之後，兩個應用程式將在單一節點上執行，在該節點上將按照配置檔案中的指定爲它們分配份額。例如，`/etc/project` 檔案中的此項目指定爲應用程式 1 配置 4 個份額而爲應用程式 2 配置 1 個份額。

```
Prj_1:100:project for App-1:root::project.cpu-shares=(privileged,4,none)
Prj_2:101:project for App-2:root::project.cpu-shares=(privileged,1,none)
```

下圖展示了此配置的一般作業與故障轉移作業。指定的份額數沒有變更。然而，每個應用程式的可用 CPU 時間比例會變更。該比例取決於爲每個需要 CPU 時間的程序指定的份額數。



具有三個應用程式的兩個節點叢集

在一個包含三個應用程式的雙節點叢集上，您可以將一個實體主機配置爲 (*phys-schost-1*) 一個應用程式的預設主要節點。您可以將第二個實體主機配置爲 (*phys-schost-2*) 剩餘兩個應用程式的預設主要節點。假定每個節點上都有以下範例專案資料庫檔案。當發生故障轉移或切換保護移轉時，專案資料庫檔案不會變更。

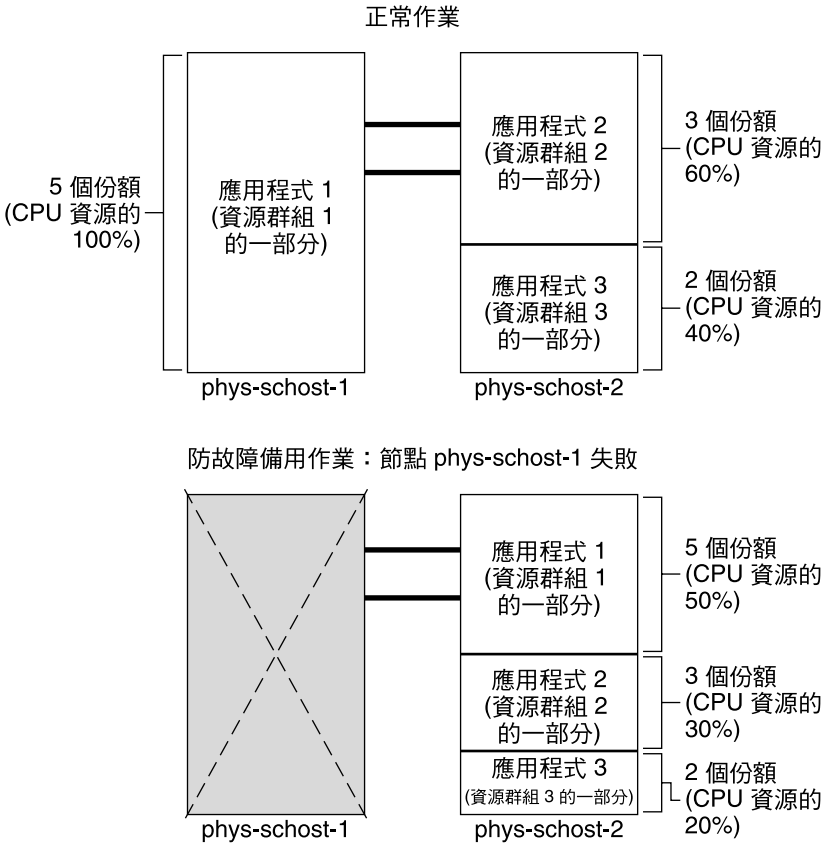
```
Prj_1:103:project for App-1:root::project.cpu-shares=(privileged,5,none)
Prj_2:104:project for App_2:root::project.cpu-shares=(privileged,3,none)
Prj_3:105:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

當叢集正常執行時，將在應用程式 1 的預設主控者 *phys-schost-1* 上爲其分配 5 個份額。此數相當於 CPU 時間的 100%，因爲它是該節點上需要 CPU 時間的唯一應用程式。在應用程式 2 和 3 的預設主節點 *phys-schost-2* 上，分別爲應用程式 2 和 3 配置 3 個份額和 2 個份額。在一般作業期間，應用程式 2 將收到 60% 的 CPU 時間，應用程式 3 將收到 40% 的 CPU 時間。

如果發生容錯移轉或切換移轉並且應用程式 1 切換移轉至 *phys-schost-2*，則三個應用程式的份額數保持不變。不過，將依據專案資料庫檔案重新分配 CPU 資源的百分比。

- 應用程式 1 具有 5 個份額，將收到 CPU 的 50%。
- 應用程式 2 具有 3 個份額，將收到 CPU 的 30%。
- 應用程式 3 具有 2 個份額，將收到 CPU 的 20%。

下圖展示了此配置的一般作業與故障轉移作業。



僅資源群組的故障轉移

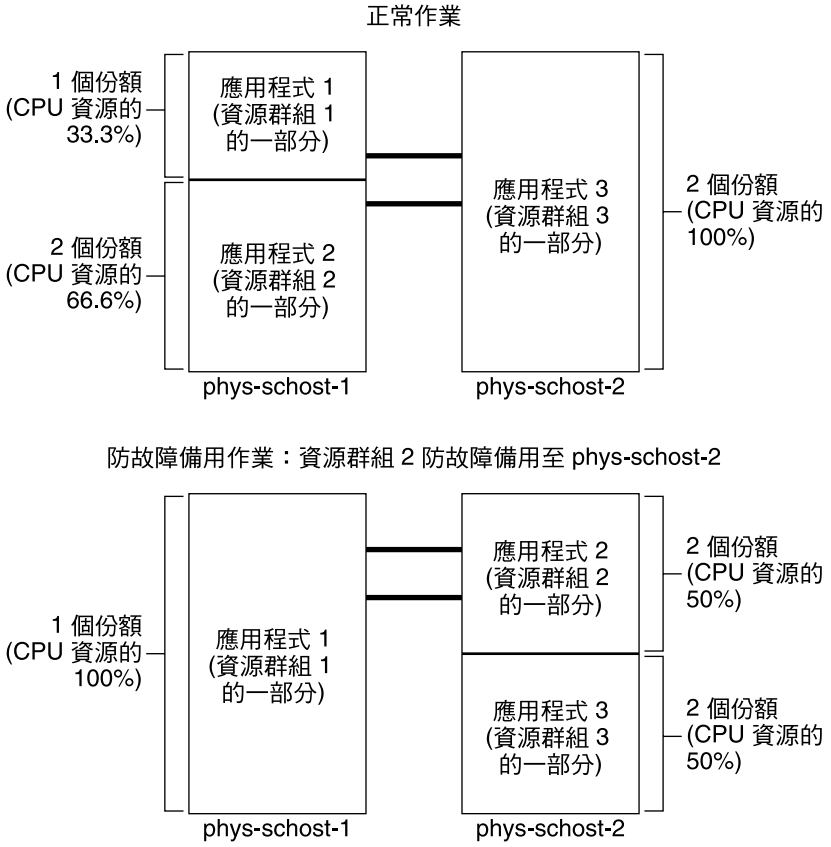
在多個資源群組使用同一個預設主控者的配置中，資源群組 (及其關聯應用程式) 可以發生故障轉移或切換至次要節點。同時在叢集內執行預設主控者。

備註 – 在故障轉移期間，將按照次要節點上配置檔案中的指定，爲發生故障轉移的應用程式分配資源。在此範例中，主要節點和次要節點上的專案資料庫檔案具有相同的配置。

例如，此範例配置檔案指定爲應用程式 1 分配 1 個份額，爲應用程式 2 分配 2 個份額以及爲應用程式 3 分配 2 個份額。

```
Prj_1:106:project for App_1:root::project.cpu-shares=(privileged,1,none)
Prj_2:107:project for App_2:root::project.cpu-shares=(privileged,2,none)
Prj_3:108:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

下圖說明了此配置的正常作業和容錯移轉作業，其中包含應用程式 2 的 RG-2 容錯移轉至 *phys-schost-2*。請注意指定的份額數不會變更。然而，每個應用程式的可用 CPU 時間會變更，這取決於爲每個需要 CPU 時間的應用程式指定的份額數。



公用網路配接卡和 Internet Protocol (IP) 網路多重路徑

用戶端透過公用網路來將要求送至叢集。每個叢集節點均透過一對公用網路配接卡連接至至少一個公用網路。

Sun Cluster 上的 Solaris 網際網路通訊協定 (IP) 網路多重路徑軟體提供一種用於監視公用網路配接卡，並在偵測到故障時將 IP 位址從一個配接卡容錯移轉至其他配接卡的基本機制。每個叢集節點均有自己的 Internet Protocol (IP) 網路多重路徑 配置，該配置可能與其他叢集節點上的配置不同。

公用網路配接卡已經置入 *IP* **多重路徑群組** (多重路徑群組) 中。每個多重路徑群組均有一個或多個公用網路配接卡。多重路徑群組中的每個配接卡均可以處於使用中狀態。或者，您可以配置待機介面，該介面處於非使用中狀態，除非發生容錯移轉。

`in.mpathd` 多重路徑常駐程式使用測試 IP 位址來偵測故障並進行修復。如果透過多重路徑常駐程式在其中一個配接卡上偵測到錯誤，將發生故障轉移。所有網路存取將從發生錯誤的配接卡容錯移轉至多重路徑群組中的其他可以正常運作的配接卡。因此，常駐程式可以維護節點的公用網路可連結性。如果您配置了待機介面，則常駐程式會選擇該待機介面。否則，常駐程式會選擇 IP 位址數最小的介面。因爲發生的是配接卡層級的容錯移轉，更高層級的連線 (如 TCP) 未受影響，只是在容錯移轉過程中出現短暫的暫態延遲。IP 位址的容錯移轉成功完成後，將會發送 ARP 廣播。因此，常駐程式可以維護遠端用戶端的連結。

備註 – 由於 TCP 的擁塞回復特性，TCP 端點會在成功容錯移轉後進一步延遲。某些區段可能會在容錯移轉過程中遺失，從而啓動 TCP 中的擁塞控制機制。

多重路徑群組可提供邏輯主機名稱與共用位址資源的建置區塊。您也可以另外建立邏輯主機名稱與共用位址資源的多重路徑群組，來監視叢集節點的公用網路連接性。節點上的相同多重路徑群組可以擁有任意數目的邏輯主機名稱或共用位址資源。如需有關邏輯主機名稱和共用位址資源的更多資訊，請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」。

備註 – Internet Protocol (IP) 網路多重路徑 機制的設計是爲了偵測和遮罩配接卡故障。該設計並不是供管理員用於使用 `ifconfig(1M)` 以移除一個邏輯 (或共用) IP 位址，以進行故障回復。Sun Cluster 軟體將邏輯和共用 IP 位址視爲由 RGM 管理的資源。管理員增加或移除 IP 位址的正確方法是使用 `scrgadm`(1M) 修改包含資源的資源群組。

如需有關 Solaris 實作 IP 網路多重路徑的更多資訊，請參閱叢集上所安裝的 Solaris 作業系統之相應文件。

作業系統發行版本	操作說明
Solaris 8 作業系統	「IP Network Multipathing Administration Guide」
Solaris 9 作業系統	「IP Network Multipathing Administration Guide」中的第 1 章「IP Network Multipathing (Overview)」
Solaris 10 作業系統	「System Administration Guide: IP Services」中的第 VI部分「IPMP」

SPARC: 動態重新配置支援

Sun Cluster 3.1 8/05 對動態重新配置 (DR) 軟體功能的支援正處於不斷取得進展的開發階段。本節說明了關於 DR 功能之 Sun Cluster 3.1 8/05 支援的概念和注意事項。

文件中說明的 Solaris?DR 功能的所有需求、程序和限制也適用於 Sun Cluster DR 支援 (作業環境的靜態作業除外)。因此，在使用 Sun Cluster 軟體的 DR 功能**之前**，請先閱讀 Solaris DR 功能的文件。您應該仔細閱讀在 DR 拆離作業過程中影響非網路 IO 裝置的問題。

「Sun Enterprise 10000 Dynamic Reconfiguration User Guide」和「Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual」(在Solaris 8 on Sun Hardware 或 Solaris 9 on Sun Hardware集合中) 均可從 http://docs.sun.com 進行下載。

SPARC: 動態重新配置一般說明

DR 功能可以在正在執行系統中進行作業 (例如移除系統硬體)。DR 程序用於確保連續的系統作業，無需停止系統或中斷叢集可用性。

DR 在板層次上作業。因此，DR 作業會影響板上所有的元件。每個板可以包含多個元件，包括 CPU、記憶體以及磁碟裝置、磁帶機與網路連接的周邊介面。

移除包含使用中的元件的板會導致系統錯誤。在移除板之前，DR 子系統可查詢其他子系統 (如 Sun Cluster)，以確定是否正在使用板上的元件。如果 DR 子系統發現一個板正在使用中，將不執行 DR 移除板的作業。因此，執行 DR 移除板作業一定是安全的，因爲 DR 子系統會拒絕在包含使用中的元件的板上執行的作業。

DR 增加板的作業也一定是安全的。新加入板上的 CPU 與記憶體會由系統自動納入服務中。然而，系統管理員必須手動將叢集配置爲主動使用新增板上的元件。

備註 – DR 子系統具有數個層次。如果較低層次報告一個錯誤，則較高層次也將報告一個錯誤。然而，較低層級報告特定錯誤時，較高層級會報告 `Unknown error`。您可以安全地忽略此錯誤。

下列章節說明了用於不同裝置類型的 DR 注意事項。

SPARC: CPU 裝置的 DR 叢集注意事項

Sun Cluster 軟體不會由於 CPU 裝置而拒絕 DR 移除板作業。

若接著執行 DR 加入板作業，加入板上的 CPU 裝置將自動納入系統作業中。

SPARC: 記憶體的 DR 叢集注意事項

根據 DR 的用途，請注意記憶體的兩種類型。

- 核心記憶體機架
- 非核心記憶體機架

這兩種類型僅在用法上不同，其實際硬體相同。核心記憶體機架是 Solaris 作業體統所使用的記憶體。Sun Cluster 軟體不支援在包含核心記憶體機架的板上執行移除板作業並會拒絕任何此類作業。對其他非核心記憶體機架執行 DR 移除板作業時，Sun Cluster 軟體不會拒絕此作業。若接著執行關係到記憶體的 DR 加入板作業，加入板上的記憶體將自動納入系統作業中。

SPARC: 磁碟和磁帶裝置的 DR 叢集注意事項

Sun Cluster 會拒絕主要節點之使用中磁碟機上的 DR 移除板作業。DR 移除板作業可以在主要節點中的非使用中的磁碟機上和次要節點中的任何磁碟機上執行。DR 作業完成後，叢集資料存取會像之前一樣繼續。

備註 – Sun Cluster 會拒絕影響法定裝置可用性的 DR 作業，如需有關法定裝置及在其上執行 DR 作業的程序之注意事項，請參閱第 77 頁的 「SPARC: 法定裝置的 DR 叢集注意事項」。

請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「動態重新配置法定裝置」，以取得有關如何執行這些動作的詳細說明。

SPARC: 法定裝置的 DR 叢集注意事項

如果對含有配置爲法定裝置的介面之板執行 DR 移除板作業，則 Sun Cluster 軟體拒絕此作業。Sun Cluster 軟體還會識別可能受到此作業影響的法定裝置。您必須先將此裝置作爲法定裝置停用，然後才可以執行 DR 移除板作業。

請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 5 章「管理法定數目」，以取得有關如何管理法定數目的詳細說明。

SPARC: 叢集互連介面的 DR 叢集注意事項

如果對含有使用中的叢集互連介面執行 DR 移除板作業，則 Sun Cluster 軟體會拒絕此作業。Sun Cluster 軟體還會識別可能受到此作業影響的介面。您必須使用 Sun Cluster 管理工具停用使用中的介面 DR 作業才能成功。

注意 – Sun Cluster 軟體要求每個叢集節點至少具有一條可以正常運作的到其他各叢集節點的路徑。請勿停用私有交互連接介面支援任何叢集節點的最後路徑。

請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理叢集交互連接」，以取得有關如何執行這些動作的詳細說明。

SPARC: 公用網路介面的 DR 叢集注意事項

如果對含有使用中的公用網路介面的板執行 DR 移除板作業，則 Sun Cluster 軟體會拒絕此作業。Sun Cluster 軟體還會識別可能受到此作業影響的介面。在您移除含有使用中的網路介面的主機板之前，請使用 `if_mpadm`(1M) 指令以將此介面上所有的流量切換移轉到多重路徑群組中的其他可以正常運作的介面上。

注意 – 如果您在停用的網路配接卡上執行 DR 移除作業時其餘網路配接卡發生故障，則可用性會受到影響。其餘的配接卡沒有空間可以爲 DR 作業的持續時間進行故障轉移。

請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理公用網路」，以取得有關如何在公用網路介面上執行 DR 移除作業的詳細資訊。

第 4 章

常見問題

本章包含有關 Sun Cluster 系統最常見問題的解答。問題是依照主題來排列。

高度可用性常見問題

問題: 到底什麼是高可用性系統？

答案: Sun Cluster 系統將高度可用性 (HA) 定義爲叢集保持應用程式執行的能力。甚至發生一般會導致伺服器系統不可用的故障時，應用程式仍可執行。

問題: 叢集是利用何種處理程序來提供高可用性？

答案: 藉由故障轉移的處理程序，叢集框架提供高可用性的環境。故障轉移是叢集所執行的一系列步驟，可將應用程式從故障節點移轉至叢集中的另一個可作業節點上。

問題: 故障轉移與可延伸的資料服務之間的差異爲何？

答案: 高度可用的資料服務共有兩種類型：

- 防故障備用
- 可延伸的

故障轉移資料服務表示應用程式一次僅在叢集中的一個主要節點上執行。其他的節點可能執行其他的應用程式，但是每個應用程式僅執行於單一節點上。如果主要節點發生故障，則在此發生故障的節點上執行的應用程式會容錯移轉至其他節點並繼續執行。

可延伸服務將應用程式分散在多個節點，以建立單一、邏輯的服務。可延伸服務會利用其執行所在的整個叢集中的節點與處理器數目。

對於各個應用程式，一個節點擁有叢集的實體介面。此節點稱爲「整體介面 (GIF) 節點」。叢集中可以存在多個 GIF 節點。每個 GIF 節點都擁有一個或多個可延伸服務可以使用的邏輯介面。這些邏輯介面稱爲**整體介面**。一個 GIF 節點擁有用於處理針對特定應用程式之所有要求的整體介面，並可將這些要求派送至應用程式伺服器正在執行的多重節點上。如果 GIF 節點發生故障，則整體介面將故障轉移至存活節點。

如果應用程式在其上執行的任何節點發生故障，則應用程式會繼續在其他節點上執行，同時效能會有所降低。此過程會繼續，直至發生故障的節點返回到叢集中。

檔案系統常見問題

問題: 是否將執行一個或多個叢集節點作爲高度可用的 NFS 伺服器可以並其將他叢集節點作爲用戶端？

答案: 不，不要做回送裝載。

問題: 是否可以將叢集檔案系統用於不在 資源群組管理員 控制下的應用程式？

答案: 可以。然而，沒有 RGM 的控制，應用程式其上執行的節點發生故障後需要手動將其重新啓動。

問題: 是否所有節點檔案系統均必須在 `/global` 目錄下具有掛載點？

答案: 不是。然而，將叢集檔案系統放在相同的裝載點之下 (如 `/global/`)，會使這些檔案系統的組織和管理有所改善。

問題: 使用叢集檔案系統和匯出 NFS 檔案系統之間的差異是什麼？

答案: 存在數處差異：

1. 叢集檔案系統支援整體裝置。NFS 不支援遠端存取裝置。
2. 叢集檔案系統擁有全域名稱空間。只需要一個裝載指令。至於 NFS，您必須在每一個節點載設檔案系統。
3. 叢集檔案系統快取檔案的機會多於 NFS。例如，當從多個節點對檔案進行讀取、寫入、檔案鎖定或非同步 I/O 存取時，叢集檔案系統會快取檔案。
4. 建置叢集檔案系統，是爲了利用提供遠程 DMA 和零複製功能的未來快速叢集交互連接。
5. 如果您變更叢集檔案系統中某個檔案的特性 (例如，使用 `chmod(1M)`)，此變更會立即反映到所有節點。使用匯出的 NFS 檔案系統，則此變更會花費更長的時間。

問題: 檔案系統 `/global/.devices/node@`*nodeID* 會顯示在我的叢集節點上。我可使用此系統檔，以儲存我想要讓其爲高可用及整體的資料嗎？

答案: 這些系統檔會儲存整體裝置的名稱空間。這些檔案系統不是用於一般用途的。其爲全域時，絕不會以全域方式存取 — 每個節點僅存取其自己的全域裝置名稱空間。假如節點當機了，其他節點就無法存取當機節點的名稱空間。這些檔案系統不具高可用性。它們不應用來儲存需爲整體或高可用的資料

容體管理常見問題

問題: 是否需要鏡像所有的磁碟裝置？

答案: 對於要作爲高可用性的磁碟裝置，必須要進行鏡像，或使用 RAID-5 硬體。所有的資料服務應該使用高可用性磁碟裝置，或裝載於高可用性磁碟裝置上的叢集檔案系統。這樣的配置可以容忍單一磁碟故障。

問題: 我可對本機磁碟 (開機磁碟) 使用一個容體管理程式，而對多重主機磁碟使用不同的容體管理程式嗎？

答案: SPARC：此配置受 Solaris Volume Manager 軟體 (管理本機磁碟) 和 VERITAS Volume Manager (管理多重主機磁碟) 支援。但並不支援其他組合。

x86： 不，不支援此配置，因爲在基於 x86 的叢集中僅支援 Solaris Volume Manager。

資料服務常見問題

問題: 哪些 Sun Cluster 資料服務是可用的？

答案: 「Sun Cluster 3.1 8/05 版本說明 (適用於 Solaris 作業系統)」中的「支援的產品」中提供了所支援的資料服務之清單。

問題: Sun Cluster 資料服務支援哪些版本的應用程式？

答案: 「Sun Cluster 3.1 8/05 版本說明 (適用於 Solaris 作業系統)」中的「支援的產品」中提供了所支援的應用程式版本之清單。

問題: 我是否可寫入自己的資料服務？

答案: 可以。請參閱「Sun Cluster 資料服務開發者指南 (適用於 Solaris 作業系統)」中的第 11 章「DSDL API 函數」，以取得更多資訊。

問題: 在建立網路資源時，我是否該指定數字型的 IP 位址或主機名稱？

答案: 指定網路資源，最好是使用 UNIX 主機名稱，而非數字型 IP 位址。

問題: 在建立網路資源時，使用邏輯主機名稱 (LogicalHostname 資源) 或共用的位址 (SharedAddress 資源) 之間的差異是什麼？

答案: 除 Sun Cluster HA for NFS 的情況之外，文件建議在 `Failover` 模式資源群組中使用 `LogicalHostname` 資源時，便會交替使用 `SharedAddress` 資源或 `LogicalHostname` 資源。使用 `SharedAddress` 資源會導致一些額外的經常性耗用時間因爲叢集網路軟體是針對 `SharedAddress` 而非 `LogicalHostname` 進行配置的。

當您配置了可延伸和容錯移轉兩種資料服務，並希望用戶端能夠透過使用同一主機名稱存取這兩種服務時，使用 SharedAddress 資源的優勢將得到證明。在這種情況下，SharedAddress 資源連同容錯移轉應用程式資源包含在一個資源群組中。可延伸服務資源包含在單獨的資源群組中，並配置爲使用 SharedAddress 資源。然後可延伸和容錯移轉服務便均可使用配置在 SharedAddress 資源中的同一組主機名稱/位址。

公用網路常見問題

問題: Sun Cluster 系統支援哪些公用網路配接卡？

答案: 目前，Sun Cluster 系統支援乙太網路 (10/100BASE-T 和 1000BASE-SX Gb) 公用網路配接卡。因爲未來可能會支援新的介面，請洽詢您的 Sun 業務代表，以取得最新的資訊。

問題: 在故障轉移中 MAC 位址扮演的角色是什麼？

答案: 發生故障轉移時，會產生新的「位址解析度通訊協定 (Address Resolution Protocol，ARP)」封包並廣播到網路上。這些 ARP 封包包含新的 MAC 位址 (節點移轉後的新實體配接卡的位址) 和舊的 IP 位址。當網路上的其他機器接收到上述封包中的一個封包之後，該封包會從其 ARP 快取中清除舊的 MAC-IP 對映，而使用新對映。

問題: Sun Cluster 系統是否支援設定 `local-mac-address?=true` ？

答案: 可以。實際上，IP 網路多重路徑要求 `local-mac-address?` **必須**設定爲 `true`。

您可以在以 SPARC 爲基礎的叢集中的 OpenBoot PROM `ok` 提示符號中，使用 `eeprom`(1M) 來設定`local-mac-address?`。您還可以在基於 x86 的叢集中，使用 BIOS 啓動後選擇性執行的 SCSI 公用程式來設定 MAC 位址。

問題: Internet Protocol (IP) 網路多重路徑 在配接卡之間執行切換時要延遲多久？

答案: 延遲可以達數分鐘。因爲當 Internet Protocol (IP) 網路多重路徑 執行切換備用時，此作業會傳送免費的 ARP。然而，您無法確保用戶端和叢集間的路由器將使用免費的 ARP。因此，直到在此路由器上 IP 位址的 ARP 快取項目逾時，項目才會使用無效的 MAC 位址。

問題: 偵測到網路配接卡故障的速度有多快？

答案: 預設的故障偵測時間爲 10 秒。演算法嘗試符合此故障偵測時間，但實際時間取決於網路負載。

叢集成員常見問題

問題: 所有的叢集成員是否需要相同的 root 密碼？

答案: 每個叢集成員不需要有相同的 root 密碼。然而，所有的節點使用相同的 root 密碼可以簡化您的節點管理工作。

問題: 節點啓動的順序是否相當重要？

答案: 在大多數情況下，啓動順序並不重要。但對於防止 amnesia 是很重要的。例如，如果節點 2 是法定裝置的所有者，而且節點 1 關機，接著您又將節點 2 關機，則您必須先啓動節點 2 再啓動節點 1。此順序可以防止您意外啓動含有過期的叢集配置資訊的節點。請參閱第 48 頁的 「關於故障隔離」，以取得有關 amnesia 的詳細資訊。

問題: 我是否需要在叢集節點中鏡像本機磁碟？

答案: 可以。雖然並不要求鏡像，但是鏡像叢集節點的磁碟可以防止非鏡像的磁碟故障使節點當機。鏡像叢集節點的區域磁碟的缺點，是需要較多的系統管理負擔。

問題: 叢集成員備份的問題有哪些？

答案: 您可以對叢集使用多種備份方法。一種方法是將連結有磁帶機或程式庫的節點作爲備份節點。然後使用叢集檔案系統來備份資料。請勿連接此節點至共用磁碟。

請參閱「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的第 9 章「備份與復原叢集」，以取得關於如何備份和修復資料的更多資訊。

問題: 節點何時正常到足以作爲次要節點？

答案: Solaris 8 和 Solaris 9：

在重新啓動後，當節點顯示登入提示時，此節點正常，足以成爲次要節點。

Solaris 10：

如果 `multi-user-server` 里程碑正在執行，則節點正常，完全可以成爲次要節點。

```
# svcs -a | grep multi-user-server:default
```

叢集儲存體常見問題

問題: 什麼原因讓多重主機儲存體具備高可用性？

答案: 多重主機儲存體具有高度可用性，因爲由於鏡像 (或由於硬體式 RAID-5 控制器) 而可以在單一磁碟遺失的情況下繼續運作。因爲多重主機儲存裝置具有一個以上的主機連接，也可以承受失去它所連接的單一節點。另外，從每個節點到貼附儲存體的冗餘路徑可提供主機匯流排配接卡、電纜或磁碟控制器故障的公差。

叢集互連常見問題

問題: Sun Cluster 系統支援哪些叢集互連？

答案: 目前，Sun Cluster 系統支援以下叢集互連：

- 基於 SPARC 和 x86 的兩種叢集中的乙太網路 (100BASE-T Fast Ethernet 和 1000BASE-SX Gb)
- 基於 SPARC 和 x86 的兩種叢集中的 Infiniband
- 僅基於 SPARC 的叢集中的 SCI

問題: 「電纜」和傳輸「路徑」之間有何差異？

答案: 叢集傳輸電纜是透過使用傳輸配接器和交換器進行配置的。電纜是以元件對元件方式連接配接卡和切換器。叢集拓樸管理者使用可用的電纜來建立節點之間的點對點傳輸路徑。電纜並不會直接對應至傳輸路徑。

電纜是由管理者做靜態的"啓用"和"停用"。電纜有「狀況」(啓用或停用)，但非「狀態」。如果電纜是啓用的，其就如同尚未配置。停用的電纜無法用作傳輸路徑。這些電纜是未經測試的，因此其狀況不明。您可以使用 `scconf -p` 指令獲取電纜狀況。

傳輸路徑並非由叢集拓樸管理者動態建立的。傳輸路徑的"狀態"是由拓樸管理者決定。路徑可以的狀態可以爲「上線」或「離線」。您可以使用 `scstat(1M)` 指令獲取傳輸路徑的狀態。

請考慮下述具四條電纜的兩個節點叢集範例。

```
node1:adapter0      to switch1, port0
node1:adapter1      to switch2, port0
node2:adapter0      to switch1, port1
node2:adapter1      to switch2, port1
```

透過這四條電纜可以形成兩條可能的路徑。

```
node1:adapter0    to node2:adapter0
node2:adapter1    to node2:adapter1
```

用戶端系統常見問題

問題: 與叢集配合使用是否需要考慮任何特殊的用戶端需求或限制？

答案: 用戶端系統連接至叢集的方式與連線至其他伺服器的方式相同。在某些情況下，視資料服務應用程式而定，您可能需要安裝用戶端軟體或執行其它配置變更，使得用戶端可以連接至資料服務應用程式。請參閱「Sun Cluster Data Services Planning and Administration Guide for Solaris OS」中的第 1 章「Planning for Sun Cluster Data Services」，以取得有關用戶端配置需求的更多資訊。

管理主控台常見問題

問題: Sun Cluster 系統是否需要管理主控台？

答案: 可以。

問題: 管理主控台必須專屬於叢集，或者可以用於其他作業嗎？

答案: Sun Cluster 系統不需要專屬的管理主控台，但使用專屬的主控台會有以下優點：

- 在同一機器上將主控台和管理工具分組，達到中央化叢集管理
- 讓您的硬體服務供應商可較快速地解決問題

問題: 管理主控台的位置需要「靠近」叢集嗎，例如在同一房間中？

答案: 請洽詢您的硬體服務供應商。供應商可能會要求主控台的位置要儘量靠近叢集。將主控台置於同一房間中，並無技術上的原因。

問題: 管理主控台是否可以服務多個叢集，是否要首先滿足距離要求？

答案: 可以。您可以從單一管理主控台來控制多個叢集。您也可以在叢集之間共用單一的終端機集線器。

終端機集訊機和系統服務處理器常見問題

問題: Sun Cluster 系統是否需要終端機集訊機？

答案: 所有以 Sun Cluster 3.0 開始的軟體發行版本均不需要終端機集訊機便可執行。與需要終端機集訊機來作爲故障隔離之用的 Sun Cluster 2.2 產品不同，之後的產品不依靠終端機集訊機。

問題: 我發現大部分 Sun Cluster 伺服器均使用終端機集訊機，而 Sun Enterprise E1000 伺服器卻不用。爲什麼？

答案: 終端機集線器對大部分的伺服器而言，實際上是一個串列對 Ethernet 轉換器。終端機集訊機的主控台連接埠是串列埠。Sun Enterprise E1000 伺服器沒有串列主控台。「系統服務處理器」(SSP) 是主控台，是透過 Ethernet 或 `jtag` 通訊埠。對於 Sun Enterprise E1000 伺服器，始終將 SSP 用於主控台。

問題: 使用終端機集線器有些什麼樣的好處？

答案: 使用終端機集訊機可以從網路上任何地方的遠端工作站對每個節點進行主控台層級的存取。即使節點位於基與 SPARC 的節點上的 OpenBoot PROM (OBP) 中或基於 x86 的節點上的啓動子系統中，仍然會提供此存取權。

問題: 如果使用 Sun 不支援的終端機集訊機，則需要瞭解些什麼才能證明要使用終端機集訊機符合要求？

答案: Sun 支援的終端機集訊機與其他主控台裝置之間的主要差異是 Sun 終端機集訊機具有特殊的韌體。此韌體會防止終端機集訊機在主控台啓動時向其傳送中斷。如果您有可以傳送中斷或可解譯爲主控台中斷的訊號至主控台的主控台裝置，則中斷將關閉節點。

問題: 是否可以在不重新啓動的情況下釋放 Sun 支援的終端機集訊機上鎖定的連接埠？

答案: 可以。請注意需要重設的通訊埠編號並鍵入下列指令：

```
telnet tc
輸入 Annex 通訊埠名稱或編號：cli
annex: su -
annex# admin
admin : reset port_number
admin : quit
annex# hangup
#
```

請參閱以下手冊，以取得有關如何配置和管理 Sun 支援的終端機集訊機的更多資訊。

- 「Sun Cluster 系統管理指南（適用於 Solaris 作業系統）」中的「管理 Sun Cluster 簡介」
- 「Sun Cluster 3.0-3.1 Hardware Administration Manual for Solaris OS」中的第 2 章「Installing and Configuring the Terminal Concentrator」

問題: 萬一終端機集線器本身故障，要怎麼辦？我必須要有另一個備用的嗎？

答案: 不需要。如果終端機集線器故障，您並不會失去任何叢集可用性。但是您會失去連接節點主控台的能力，直到集線器回復服務爲止。

問題: 如果我眞的使用終端機集線器，其安全性如何？

答案: 通常，終端機集訊機連結到系統管理員使用的小型網路上，而非用於供其他客戶端存取的網路上。您可以藉由限制該特定網路的存取權來控制安全性。

問題: SPARC：如何藉由磁帶機或磁碟機使用動態重新配置？

答案: 執行下列步驟：

- 判斷磁碟機或磁帶機是否爲使用中裝置群組的一部分。如果磁碟機不是使用中裝置群組的一部分，您可以在其上執行移除 DR 作業。
- 如果 DR 移除板作業可能會影響到使用中的磁碟機或磁帶機，系統會拒絕該作業，並指出可能會被該作業影響的磁碟機。如果磁碟機是使用中的裝置群組的一部分，請移至第 76 頁的 「SPARC: 磁碟和磁帶裝置的 DR 叢集注意事項」。
- 判斷磁碟機是主要節點的元件還是次要節點的元件。如果磁碟機是次要節點的元件，便可以在其上執行 DR 移除作業。
- 如果磁碟機是主要節點的元件，則必須先切換主要節點與次要節點，然後才能在該裝置上執行 DR 移除作業。

注意 – 如果您在次要節點上執行 DR 作業時，現行的主要節點發生故障，叢集可用性將會受到影響。除非提供新的次要節點，否則主要節點沒有地方可以進行故障轉移。

索引