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: 819–0167–10
2004 年 9 月，修訂版 A

050104@10536

目錄

前言

「*Sun*™ *Cluster* **概念指南** *(***適用於** *Solaris* **作業系統***)*」包含基於 SPARC™ 與 x86 的系統上的 SunPlex™ 系統之概念性與參考資訊。

注意 – 在本文件中，「x86」指 Intel 32 位元系列的微處理器晶片和由 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。

此文件主要是針對在 Sun Cluster 軟體上接受過訓練且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。

若要瞭解本書中說明的概念，您應該具備 Solaris™ 作業環境的知識，以及用於 SunPlex 系統的容體管理程式軟體技術。

注意 – 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

下表描述了本書中所用到的印刷排版變更。

表 P–1 印刷排版慣例

字體或符號	涵義	範例
`AaBbCc123`	指令、檔案和目錄的名稱，或是電腦螢幕的輸出	編輯您的 `.login` 檔案。 使用 `ls -a` 列出所有檔案。 `machine_name% you have mail.`
`AaBbCc123`	您輸入的內容，對照電腦螢幕上的輸出	`machine_name%` **`su`** `Password:`
AaBbCc123	指令行預留位置：用實際名稱或值取代	移除檔案的指令是 `rm` *filename*。
AaBbCc123	書名、新專有名詞，以及要強調的專有名詞	請閱讀「**使用者指南**」中的第 6 章。 這些選項稱爲**類別**選項。 請**不要**儲存檔案。 (強調有時在線上以粗體顯示。)

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 提示符號

Shell	提示符號
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 說明文件可在 `http://docs.sun.com` 上找到。

主題	文件
簡介	「*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.x 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 軟體版本的版本說明。

線上存取 Sun 說明文件

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 是 `http://docs.sun.com`。

訂購 Sun 說明文件

Sun Microsystems 提供列印的選取產品說明文件。若要瞭解文件清單及其訂購方法，請參閱 `http://docs.sun.com/?l=zh_TW` 上的「購買書面文件」。

取得說明

如果在安裝或使用 SunPlex 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 4/04)

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 軟體版次與套件版本資訊
`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

同時提供 `/var/adm/messages` 檔案的內容。

第 1 章

簡介與概觀

SunPlex 系統爲一整合的硬體與 Sun Cluster 軟體解決方案，用於建立高可用性及可延伸的服務。

「Sun Cluster 概念指南 (適用於 Solaris 作業系統)」提供 SunPlex 文件主要讀者所需的概念資訊。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
- 開發目前 Sun Cluster 產品所未包含的應用程式故障轉移和可延伸服務的應用程式開發人員

本書配合其餘的 SunPlex 說明文件集使用，提供 SunPlex 系統的完整概觀。

本章

- 提供 SunPlex 系統的簡介和進階概觀
- 說明 SunPlex 讀者的各種觀點
- 指出在使用 SunPlex 系統之前需要瞭解的重要概念
- 對應重要概念至包含程序和相關資訊的 SunPlex 說明文件
- 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

SunPlex 系統簡介

SunPlex 系統將 Solaris 作業環境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。

SunPlex 系統的目標是：

- 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

如需有關故障偏差與高可用性的更多資訊，請參閱「*Sun Cluster* **簡介** *(***適用於** *Solaris* **作業系統***)*」中的「使應用程式在 Sun Cluster 中具有高可用性」。

請參閱 第 75 頁的 「高可用性 FAQ」，以取得關於高可用性的問題與解答。

SunPlex 系統的三個觀點

本節說明 SunPlex 系統的三種不同觀點，以及每個觀點的相關重要概念和說明文件。這些觀點來自：

- 硬體安裝與維修人員
- 系統管理員
- 應用程式設計師

硬體安裝與服務觀點

對於硬體維修人員而言，SunPlex 系統就像是常用硬體的集合，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

硬體維修人員需要瞭解下列叢集概念。

- 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

建議的硬體概念參考文件

下列各節包含前述重要概念的相關資料：

- 第 18 頁的 「叢集節點」
- 第 19 頁的 「多重主機裝置」
- 第 21 頁的 「本機磁碟」
- 第 21 頁的 「叢集交互連接」
- 第 22 頁的 「公用網路介面」
- 第 22 頁的 「用戶端系統」
- 第 23 頁的 「管理主控台」
- 第 22 頁的 「主控台存取裝置」
- 第 24 頁的 「SPARC: 叢集化配對拓撲」
- 第 25 頁的 「SPARC: N+1 (星狀) 拓撲」

相關的 SunPlex 說明文件

下列 SunPlex 說明文件包括與硬體維修概念相關的程序和資訊：

「*Sun Cluster 3.x Hardware Administration Manual for Solaris OS*」

系統管理員觀點

對於系統管理員而言，SunPlex 系統就像是一組以電纜連接在一起的伺服器 (節點)，共用儲存裝置。系統管理員會看見：

- 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 作業環境

  - 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

建議的系統管理員概念參考文件

下列各節包含前述重要概念的相關資料：

- 第 29 頁的 「管理介面」
- 第 30 頁的 「叢集時間」
- 第 30 頁的 「高可用性框架」
- 第 32 頁的 「整體裝置」
- 第 33 頁的 「磁碟裝置群組」
- 第 36 頁的 「全域名稱空間」
- 第 37 頁的 「叢集檔案系統」
- 第 40 頁的 「磁碟路徑監視」
- 第 44 頁的 「關於故障隔離」
- 第 52 頁的 「資料服務」

相關的 SunPlex 說明文件 – 系統管理員

下列 SunPlex 文件包括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 9/04* **版本說明** *(***適用於** *Solaris* **作業系統***)*」
- 「*Sun Cluster 3.x Release Notes Supplement*」

應用程式設計師觀點

SunPlex 系統爲 Oracle (在基於 SPARC 的系統上)、NFS、DNS、Sun™ Java System Web Server (以前稱爲 Sun Java System Web Server)、Apache Web Server (在基於 SPARC 的系統上)、Sun Java System Directory Server (以前稱爲 Sun Java System Directory Server) 之類的應用程式提供**資料服務**。資料服務是藉由配置 Sun Cluster 軟體控制下之常用應用程式而建立的。Sun Cluster 軟體提供了啓動、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故障轉移或可延伸的服務，可以使用 SunPlex 應用程式設計介面 (API) 與資料服務啓用技術 API (DSET API) 來開發必要的配置檔案及管理方法，它們可讓其應用程式作爲叢集上的資料服務來執行。

重要概念 – 應用程式設計師

應用程式設計師需要瞭解下列各項：

- 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 及「一般」資料服務。程式設計師需要決定哪一種工具最適合用來撰寫程式或程序檔，以配置其用於叢集環境的應用程式。

建議的應用程式設計師概念參考文件

下列各節包含前述重要概念的相關資料：

- 第 52 頁的 「資料服務」
- 第 60 頁的 「資源、資源群組與資源類型」
- 第 4 章

相關的 SunPlex 說明文件 – 應用程式設計師

下列 SunPlex 文件包括與應用程式設計師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*」

SunPlex 系統作業

所有 SunPlex 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

執行此作業...	使用此說明文件...
安裝叢集硬體	「*Sun Cluster 3.x 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)**」

表 1–1 對應作業：將使用者作業對應至文件　　*(續)*

執行此作業...	**使用此說明文件...**
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
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.x 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***)*」與容體管理文件
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

重要概念 – 硬體服務供應商

本章說明有關 SunPlex 系統配置的硬體元件的重要概念。涵蓋的主題包含：

SunPlex 系統的硬體與軟體元件

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)
- 抽換式媒體 (磁帶和 CD-ROM)
- 叢集交互連接
- 公用網路介面
- 用戶端系統
- 管理主控台
- 主控台存取裝置

SunPlex 系統可以讓您將這些元件結合成各種配置，請參閱第 23 頁的 「SPARC: Sun Cluster 拓撲範例」之說明。

如需兩個節點叢集配置範例的圖例，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

叢集節點是執行 Solaris 作業環境及 Sun Cluster 軟體的機器，也是叢集 (**叢集成員**) 的目前成員或潛在成員。

SPARC: Sun Cluster 軟體可讓您在一個叢集中有二到八個節點。請參閱 第 23 頁的 「SPARC: Sun Cluster 拓撲範例」，以取得支援的節點配置。

x86: Sun Cluster 軟體讓您能夠在一個叢集中包含兩個節點。請參閱第 27 頁的 「x86: Sun Cluster 拓撲範例」，以取得受支援的節點配置。

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要求。

另外，平行資料庫配置中的節點可共用對所有磁碟的並行存取。請參閱第 19 頁的 「多重主機裝置」與第 3 章，以取得平行資料庫配置的詳細資訊。

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。

叢集成員透過一個或多個實際上獨立的網路與叢集中的其他節點進行通訊。此組實體上獨立的網路被視爲**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。由於可能發生故障轉移，所以每個節點必須有足夠的額外容量，可以作爲備份或次要節點來接管所有節點的工作負荷。

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，必須安裝下列軟體：

- Solaris 作業環境
- 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**)」，以取得關於如何安裝 Solaris 作業環境、Sun Cluster 與容體管理軟體的資訊。
- 請參閱「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」，以取得關於如何安裝與配置資料服務的資訊。
- 請參閱第 3 章，以取得前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檢視。

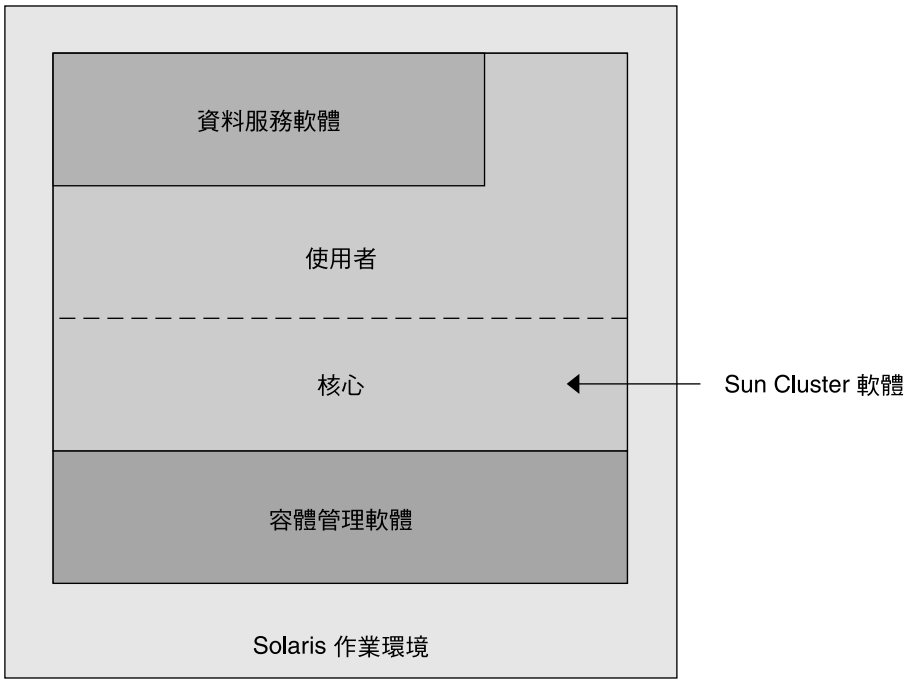


圖 2–1 Sun Cluster 軟體元件的高階關係

請參閱第 4 章，以取得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，多重主機儲存體可讓磁碟高度可用。Sun Cluster 要求包含兩個節點的叢集使用多重主機儲存體，以建立法定數目。多於三個節點的叢集不需要多重主機儲存體。

多重主機裝置有下列特性。

- 它們可容許單一節點故障。
- 它們儲存應用程式資料，也可儲存應用程式的二進位檔案與配置檔案。
- 它們對於節點故障做出保護。如果用戶端要求是透過某個節點來存取資料而該節點故障，這些要求會切換爲使用另一個可直接連接同一磁碟的節點。
- 多重主機磁碟是透過"主控"磁碟的主要節點來進行全域存取，或透過本機路徑直接進行並行存取。目前 Oracle Real Application Clusters 是唯一一個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資料冗餘提供鏡像配置或 RAID-5 配置。目前的 Sun Cluster 支援 Solaris Volume Manager™ 與 VERITAS Volume Manager，後者僅適用於作爲容體管理程式的基於 SPARC 的叢集以及數個硬體 RAID 平台上的 RDAC RAID-5 硬體控制器。

結合多重主機裝置和磁碟鏡像與資料分置，可以防止節點故障和個別磁碟故障。

請參閱第 4 章，以取得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。當叢集儲存體是由單端或差動式 SCSI 裝置所組成時，該配置即爲多重初始端 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 規格需要 SCSI 匯流排上的每一個裝置均具有一個唯一的 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。

您可以使用下列任一工具設定 `scsi-initiator-id` 特性，來變更爲這些主機配接卡選取的 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。如需有關爲每一個 SCSI 主機配接卡設定唯一 `scsi-initiator-id` 的說明，請參閱「*Sun Cluster Hardware Collection*」中有關各磁碟附件的章節。

本機磁碟

本機磁碟是僅連接至單一節點的磁碟。因此，沒有節點故障的保護 (不具高可用性)。然而，所有的磁碟 (包括本機磁碟) 均含括於全域名稱空間中，並且配置爲**整體裝置**。因此，從所有的叢集節點可以看到磁碟本身。

您可以將本機磁碟上的檔案系統放在整體裝載點下，讓其他節點使用。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 32 頁的 「整體裝置」一節，以取得關於全域裝置的更多資訊。

抽換式媒體

叢集中支援如磁帶機和 CD-ROM 光碟機的抽換式媒體。一般而言，您安裝、配置和維修這些裝置的方式與在非叢集環境的方式相同。這些裝置被配置爲 Sun Cluster 中的整體裝置，所以每一個裝置均可從叢集的任何節點來存取。請參考「*Sun Cluster 3.x Hardware Administration Manual for Solaris OS*」，以取得關於安裝與配置可移除媒體的資訊。

請參閱第 32 頁的 「整體裝置」一節，以取得關於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 安全性模型假設僅叢集節點具有叢集交互連接的實體存取權限。

必須使用叢集交互連接 (透過至少兩個實體獨立的冗餘網路或路徑) 來連接所有節點，以避免單一故障點。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 – 位於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。部分網路卡還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。執行通行和轉換功能，讓您將兩個以上的節點連接在一起。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 – 兩個網路配接卡之間或配接卡與接點之間的實體連接。

請參閱第 4 章，以取得有關叢集交互連接的問題與解答。

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，使多重卡都處於作用中狀態，並且彼此作爲故障轉移的備份。如果其中一個配接卡發生故障，將呼叫 IP Network Multipathing 軟體，對有缺陷的介面執行故障轉移至群組中的另一個配接卡。

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，以取得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。用戶端程式使用由執行於叢集上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，以取得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。若要取得主控台存取，請使用與叢集硬體一起購買的終端機集線器、Sun Enterprise E10000™ 伺服器 (用於基於 SPARC 的叢集) 上的「系統服務處理器 (SSP)」、Sun Fire™ 伺服器 (用於基於 SPARC 的叢集) 上的系統控制器，或是可以存取每個節點上 `ttya` 的其他裝置。

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)」提供 Sun Enterprise E10000 server 的主控台存取。SSP 是 Ethernet 網路上的機器，配置爲支援 Sun Enterprise E10000 server。SSP 是 Sun Enterprise E10000 server 的管理主控台。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線器，從另一個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

您可以使用專用的 UltraSPARC® 工作站或者 Sun Fire™ V65x 伺服器 (也稱**管理主控台**) 來管理作用中的叢集。通常，您在管理主控台上安裝和執行管理工具軟體，例如 Sun Management Center™ 產品 (僅與基於 SPARC 的叢集配合使用) 的「叢集控制面板 (CCP)」和 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 的詳細資訊，請參閱「*Sun Cluster System Administration Guide*」。

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 平台所組成，您必須能夠從管理主控台登入「系統服務處理器」(SSP)，並使用 `netcon(1M)` 指令連接。

一般您會配置沒有監視器的節點。然後，透過管理主控台的 `telnet` 階段作業來存取節點的主控台，管理主控台連接至終端機集線器，並從終端機集線器連接至節點的串列埠。(如果是 Sun Enterprise E10000 server，您要從「系統服務處理器」連接。)請參閱第 22 頁的 「主控台存取裝置」，以取得詳細資訊。

Sun Cluster 不需要專用的管理主控台，但是使用專用主控台可以有以下優點：

- 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，以取得有關管理主控台的問題與解答。

SPARC: Sun Cluster 拓撲範例

拓撲是指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。

- 無論您實施的儲存配置爲何，由基於 SPARC 的系統所組成的 Sun Cluster 最多支援叢集中的八個節點。
- 共用的儲存體可以連接至儲存體所支援數目的節點。

- 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 不需要您透過特定拓撲配置一個叢集。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。

- 叢集化配對
- 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: 叢集化配對拓撲

叢集化配對拓樸架構是二個或以上的節點配對，在單一叢集管理框架之下運作。在此配置中，故障轉移僅發生於配對之間。然而，所有的節點以叢集交互連接來連接，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。

利用叢集檔案系統，您也可以讓兩個配對的配置，其中有兩個以上的節點執行可延伸服務或平行資料庫，即使所有的節點均未直接連接儲存應用資料的磁碟。

下圖說明叢集化配對配置。

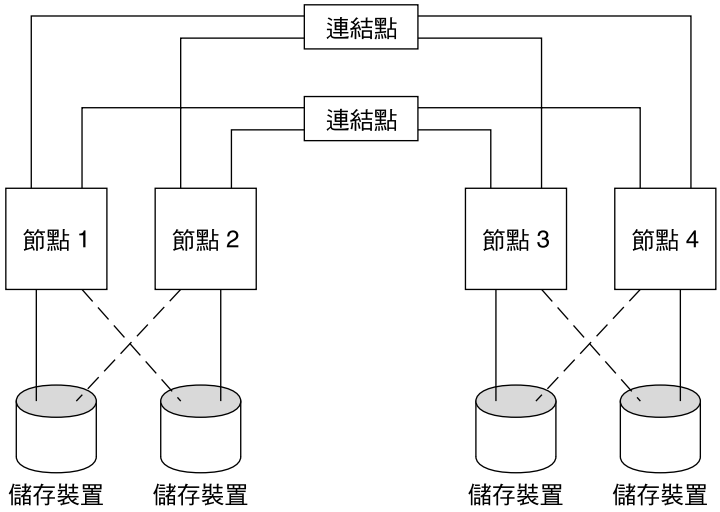


圖 2–2 SPARC: 叢集化配對拓撲

SPARC: Pair+N 拓撲

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。

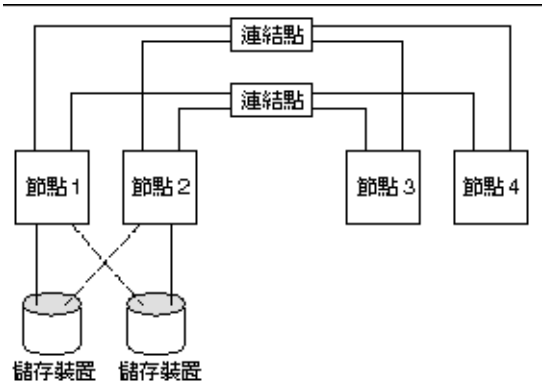


圖 2–3 SPARC: Pair+N 拓撲

SPARC: N+1 (星狀) 拓撲

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 會移轉資源至次要節點以繼續運作，直到轉換 (自動或手動) 回到主要節點爲止。

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。

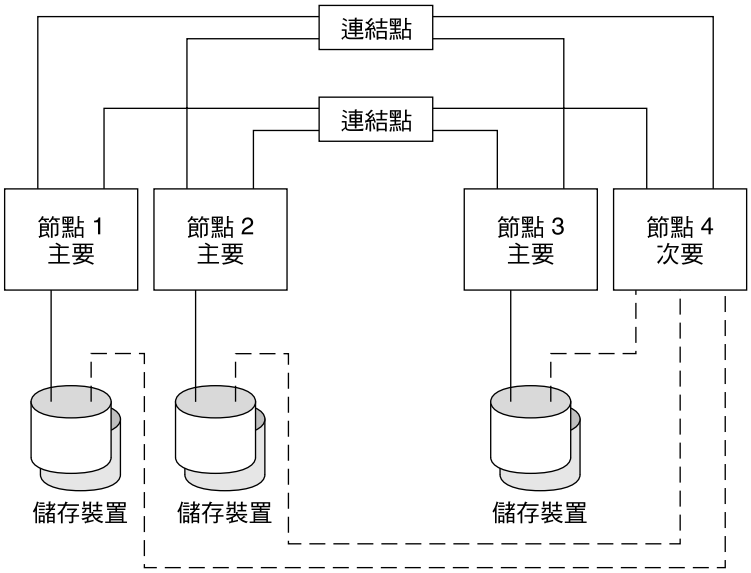


圖 2–4 SPARC: N+1 拓撲

SPARC: N*N (可延伸的) 拓撲

N*N 拓撲可讓叢集中的每個共用儲存體連接至叢集中的每個節點。此拓撲可讓高度可用的應用程式從一個節點故障轉移至另一個節點，而不會降低服務品質。發生故障轉移時，新節點可以使用本機路徑 (而不是專用交互連接) 存取儲存體。

下圖說明 N*N 配置。

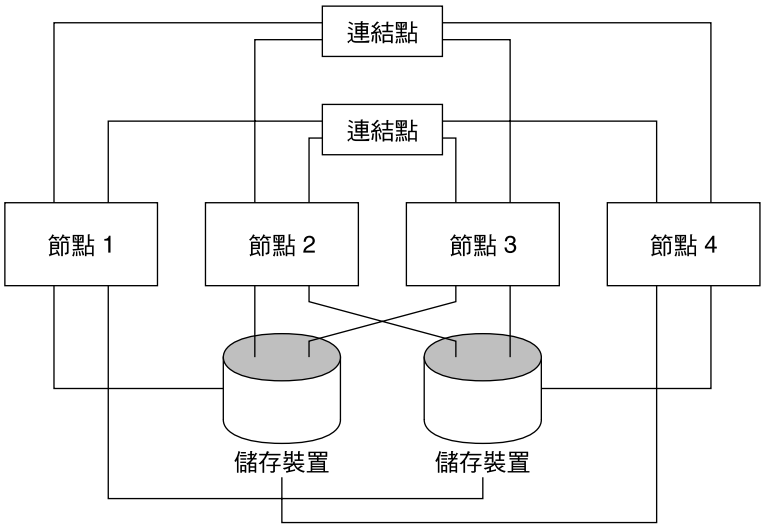


圖 2–5 SPARC: N*N 拓撲

x86: Sun Cluster 拓撲範例

拓撲是指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。
- 共用儲存體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: 叢集化配對拓撲

叢集化配對拓撲是指在單一叢集管理架構下作業的兩個節點。在此配置中，故障轉移僅發生於配對之間。然而，所有的節點以叢集交互連接來連接，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。

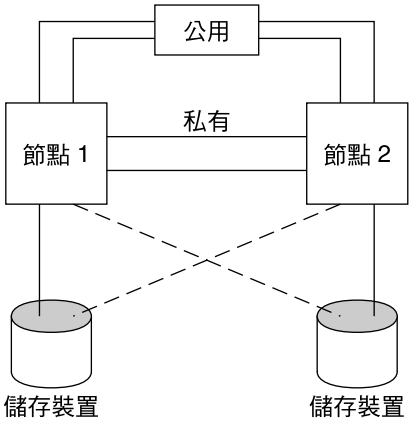


圖 2–6 x86: 叢集化配對拓撲

第 3 章

重要概念 – 管理和應用程式開發

本章說明有關 SunPlex 系統的軟體元件的重要概念。涵蓋的主題包含：

此資訊主要爲使用 SunPlex API 和 SDK 的系統管理員與應用程式開發人員提供參考。叢集系統管理員可以利用此資訊來準備安裝、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

您可以選擇從數個使用者介面安裝、配置和管理 SunPlex 系統的方式。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是一些公用程式 (例如 `scinstall` 與 `scsetup`)，可簡化所選安裝與配置作業。SunPlex 系統也有一個模組，它作爲 Sun Management Center 的一部分來執行，爲特定叢集作業提供 GUI。此模組僅可在基於 SPARC 的叢集中使用。請參考「*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描述。

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。SunPlex 系統使用「網路時間通訊協定 (NTP)」來同步化各節點的時鐘。

一般而言，系統時鐘在傾刻之間變更並不會造成問題。然而，如果您在作用中的叢集上執行 `date(1)`、`rdate(1M)` 或 `xntpdate(1M)` (互動式，或在 `cron` 程序檔之內)，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。

當您在每一個叢集節點上安裝 Solaris 作業環境時，您有機會變更節點的預設時間及日期設定。一般而言，您可以接受出廠預設值。

當您使用 `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 軟體提供範本檔 `ntp.cluster` (請參閱已安裝叢集節點上的 `/etc/inet/ntp.cluster`)，它可在所有叢集節點之間建立同級關係，並使一個節點成爲「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。如需有關如何將叢集配置爲使用 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「安裝與配置 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，如果在安裝 Solaris 作業環境時時間設定不正確，想要進行變更，變更時間設定的程序可在「*Sun Cluster* **系統管理指南** *(***適用於** *Solaris* **作業系統***)*」中的「管理叢集」中找到。

高可用性框架

SunPlex 系統使得使用者和資料間「路徑」上的所有元件均具有高度可用性，包括網路介面、應用程式本身、檔案系統和多重主機裝置。一般而言，如果系統內有任何單一 (軟體或硬體) 故障，叢集元件就具有高度可用性。

下表顯示了 SunPlex 元件故障 (硬體故障與軟體故障) 的種類，以及建立在高可用性框架中的恢復種類。

表 3–1 SunPlex 故障偵測與恢復的層次

故障的叢集元件	軟體復原	硬體恢復
資料服務	HA API、HA 框架	N/A
公用網路配接卡	IP Network Multipathing	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 與 VERITAS Volume Manager，後者僅可在基於 SPARC 的叢集中使用)	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 軟體的高可用性框架可以快速地偵測到某個節點故障，並且爲叢集中剩餘節點上的框架資源建立一個新的相等伺服器。框架資源隨時皆可使用。未受故障節點影響的框架資源，在恢復時完全可加以使用。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透明地恢復爲使用此資源的應用程式 (資料服務)。框架資源存取的語意會在各項節點故障時被完整地保留。應用程式無法辨識出框架資源伺服器已移到另一個節點。只要從另一節點到磁碟存在著另一個替代的硬體路徑，對於在使用檔案、裝置以及連接到此節點的磁碟容體上的程式而言，單一節點的故障便是完全透明。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。

叢集成員身份監視器

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。

在偵測到叢集成員變更之後，CMM 會執行叢集的同步化配置，此時可能會根據新的叢集成員而重新分配叢集資源。

與舊版次 Sun Cluster 軟體不同，CMM 完全在核心程式中執行。

請參閱第 44 頁的 「關於故障隔離」，以取得有關叢集如何保護自己免於被分割成多個單獨叢集的更多資訊。

Failfast 機制

如果 CMM 偵測到某節點發生緊急問題，則它會呼叫叢集框架以強制關閉 (當機) 節點，然後從叢集成員身份中移除該節點。發生此情況的機制稱爲 *failfast*。Failfast 會導致節點以兩種方式關閉。

- 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 44 頁的 「關於故障隔離」，以取得有關 failfast 此種用法的詳細資訊。
- 如果一個或多個叢集特定的常駐程式掛掉 (`clexecd`、`rpc.pmfd`、`rgmd` 或 `rpc.ed`)，CMM 會偵測到此故障，而節點會混亂。 當叢集常駐程式的失效導致節點當機時，在該節點的主控台上會顯示類似下列內容的訊息。

```
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
```

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。您可以在 OpenBoot PROM `ok` 提示符號處，使用 `eeprom`(1M) 來設定 `auto-boot?`。

叢集配置儲存庫 (CCR)

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成功完成，否則更新會轉返。CCR 使用叢集交互連接來套用分散式更新。

Caution – 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。

整體裝置

SunPlex 系統使用**整體裝置**來提供全叢集、高可用性存取叢集中任何裝置 (從任意節點)，不管裝置實體連接的位置。一般而言，如果節點是在提供整體裝置的存取時發生故障，Sun Cluster 軟體會自動探尋該裝置的其他路徑，並將存取重新導向到此路徑。SunPlex 整體裝置包含磁碟、CD-ROM 與磁帶。然而，磁碟是唯一支援多埠的整體裝置。這代表 CD-ROM 和磁帶裝置目前不是高可用性裝置。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各磁碟、CD-ROM 和磁帶裝置指定唯一的 ID。這項指定可以讓人從叢集的任何節點一致存取各個裝置。整體裝置名稱空間是保存於 `/dev/global` 目錄。請參閱第 36 頁的 「全域名稱空間」，以取得詳細資訊。

多埠式整體裝置提供一條以上的裝置路徑。如果是多重主機磁碟，因爲磁碟是由一個節點以上所共有之磁碟裝置群組的一部分，因此多重主機磁碟具備高可用性。

裝置 ID (DID)

Sun Cluster 軟體藉由建構裝置 ID (DID) 虛擬驅動程式來管理整體裝置。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。

裝置 ID (DID) 虛擬驅動程式是叢集的整體裝置存取功能的主要部分。DID 驅動程式會測試叢集的所有節點，並建置唯一磁碟裝置的清單，指定每個裝置唯一的主要編號和次要編號，在叢集的所有節點間是一致的。整體裝置的存取是利用由 DID 驅動程式所指定的唯一裝置 ID 來執行的，而不是透過傳統的 Solaris 裝置 ID (例如磁碟的 `c0t0d0`) 來執行的。

這種方式可以確保存取磁碟的任何應用程式 (如容體管理程式或使用原始裝置的應用程式) 可以使用一致的叢集存取路徑。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，節點 `1` 可能將多主機磁碟視爲 `c1t2d0`，節點 2 可能將同一磁碟視爲完全不同的其他名稱 `c3t2d0`。DID 驅動程式會指定一個整體名稱 (如 d10)，而節點則改用此名稱，提供了每個節點一致的多重主機磁碟對應。

您是透過 `scdidadm(1M)` 和 `scgdevs(1M)` 來更新和管理裝置 ID。請參閱以下線上說明手冊以取得更多資訊：

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

磁碟裝置群組

在 SunPlex 系統中，所有的多重主機裝置必須受 Sun Cluster 軟體的控制。首先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 (僅可用於基於 SPARC 的叢集中)。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。

註冊提供有關何種節點具有哪個容體管理程式磁碟群組路徑的 SunPlex 系統資訊。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。高度可用的磁碟裝置群組可用來包含磁碟檔案系統。

注意 – 磁碟裝置群組與資源群組無關。某個節點可以主控一個資源群組 (代表一群資料服務處理程序)，而另外一個節點則可以主控資料服務所存取的磁碟群組。然而，最佳的方式是將儲存特定應用程式之資料的磁碟裝置群組，以及包含應用程式之資源 (應用程式常駐程式) 的資源群組保存在同一節點上。請參考「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」中的「Relationship Between Resource Groups and Disk Device Groups」，以取得關於磁碟裝置群組與資源群組之間關聯性的更多資訊。

藉由磁碟裝置群組，容體管理程式磁碟群組便成爲「整體」，因爲它提供對基礎磁碟的多重路徑支援。實體連接到多重主機磁碟的每一個叢集節點，均提供了一個磁碟裝置群組的路徑。

磁碟裝置群組故障轉移

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。

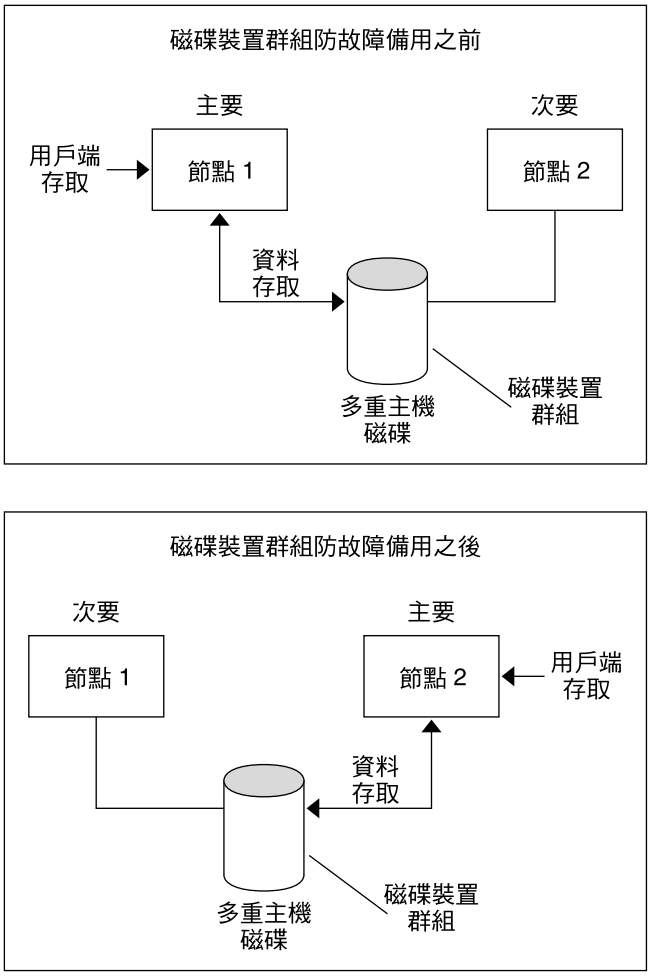


圖 3–1 磁碟裝置群組故障轉移

多埠式磁碟裝置群組

本節說明了可讓您平衡多埠式磁碟配置中的效能與可用性的磁碟裝置群組特性。Sun Cluster 軟體提供了用來配置多埠式磁碟配置的兩個特性：`preferenced` 與 `numsecondaries`。您可以使用 `preferenced` 特性來控制當發生故障轉移時節點嘗試採用控制的順序。使用 `numsecondaries` 特性，可爲裝置群組設定所需數目的次要節點。

若主要節點當機，並且沒有合格的次要節點可以提昇爲主要節點，則高度可用的服務將被視爲失效。如果發生服務故障轉移，並且 `preferenced` 特性爲 `true`，則節點將按照節點清單中的順序選取一個次要節點。設定的節點清單定義了節點將嘗試採用主要控制的順序或採用從備用節點至次要節點之轉換的順序。您可以使用 `scsetup`(1M) 公用程式動態變更裝置服務的個人喜好。與獨立服務提供者關聯的個人喜好 (如全域檔案系統)，將成爲裝置服務的個人喜好。

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。已執行備用節點支援，以最小化由核對點作業導致的效能降低程度和記憶體耗用。依預設，磁碟裝置群組將具備一個主要節點和一個次要節點。其餘可用的提供者節點將以備用狀態在線上提供。如果發生故障轉移，次要節點將成爲主要節點，節點清單中優先權最高的節點將成爲次要節點。

需要的次要節點數目，可以設定爲介於一與裝置群組中作業非主要提供者節點數目之間的任何整數。

注意 – 如果您使用的是 Solaris 容體管理程式，必須先建立磁碟裝置群組，然後才可以將 `numsecondaries` 特性設定爲非預設值的數目。

預設的所需裝置服務次要節點數目爲 1。複製框架保留的次要提供者實際數目爲需要的數目，除非作業非主要提供者的數目少於需要的數目。如果您要在配置中新增或移除節點，需要更改 `numsecondaries` 特性並仔細檢查節點清單。保留節點清單以及需要的次要節點數目將防止已配置次要節點數目與框架允許的實際數目之間發生衝突。將 `metaset`(1M) 指令用於 Solaris Volume Manager 裝置群組，如果您使用 Veritas Volume Manager，則將 `scconf`(1M) 指令用於 VxVM 磁碟裝置群組，並配合使用 `preferenced` 與 `numsecondaries` 特性設定，管理在配置中加入和移除節點。請參考「*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

啓用整體裝置的 Sun Cluster 軟體機制稱爲**全域名稱空間**。全域名稱空間包括 `/dev/global/` 階層以及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。實際連接多重主機磁碟的每一個節點，均提供一條儲存體路徑給叢集中的任何節點。

通常，對於 Solaris Volume Manager 而言，容體管理程式名稱空間位於 `/dev/md/`*diskset*`/dsk` (與 `rdsk`) 目錄中。對於 Veritas VxVM 而言，容體管理程式名稱空間位於 `/dev/vx/dsk/`*disk-group* 與 `/dev/vx/rdsk/`*disk-group* 目錄中。這些名稱空間由各自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。每個目錄對該磁碟組或磁碟群組中的每個 metadevice 或容體均含一個裝置節點。

在 SunPlex 系統中，本機容體管理程式名稱空間中的每個裝置節點均會被置換爲 `/global/.devices/node@` *nodeID* 檔案系統中裝置節點的符號連結，其中 *nodeID* 是代表叢集中節點的整數。Sun Cluster 軟體仍繼續在其標準位置展示容體管理程式裝置，例如符號連結。全域名稱空間和標準容體管理程式均可由任何叢集節點使用。

全域名稱空間的優點包括：

- 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。

本機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/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。

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。處理程序可以開啓位於系統任何位置的檔案，而且所有節點上的處理程序均可使用相同的路徑名稱來尋找檔案。

注意 – 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。
- 叢集檔案系統使用 `fcntl(2)` 介面來提供高度可用的建議檔案鎖定功能。藉由使用叢集檔案系統檔案上的建議檔鎖定功能，在多重叢集節點上執行的應用程式便得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。

您可以藉由全域的 `mount -g` 或藉由本機的 `mount` 在整體裝置上裝載檔案系統。

程式可以透過相同的檔案名稱 (例如，`/global/foo`)，從叢集中的任何節點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。亦即，用戶端可以看見基礎檔案系統 (例如，UFS)。

使用叢集檔案系統

在 SunPlex 系統中，所有多重主機磁碟均置於磁碟裝置群組中，群組可以爲 Solaris Volume Manager 磁碟組、VxVM 磁碟群組或不受基於軟體的容體管理程式所控制的個別磁碟。

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。

至於一般檔案系統，您可以用兩種方式裝載叢集檔案系統：

- **手動**—使用 `mount` 指令和 `-g` 或 `-o global` 裝載選項，從指令行裝載叢集檔案系統，例如：

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

- **自動**— 在 `/etc/vfstab` 檔案中建立具有 `global` 裝載選項的項目，於啓動時裝載叢集檔案系統。然後在所有節點的 `/global` 目錄下建立裝載點。`/global` 目錄是建議位置，並非基本要求。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
```

注意 – 因爲 Sun Cluster 軟體沒有強制叢集檔案系統的命名策略，您可以在同一個目錄 (如 `/global/`*disk-device-group*) 下建立所有叢集檔案系統的裝載點以簡化管理作業。請參閱「*Sun Cluster 3.1 9/04 Software Collection for Solaris OS (SPARC Platform Edition)*」與「*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` 資源類型的設計是要讓非全域的檔案系統配置 (例如 UFS 及 VxFS) 具有高度可用性。使用 `HAStoragePlus` 即可將您的本機檔案系統整合到 Sun Cluster 環境中，並讓檔案系統具有高度可用性。`HAStoragePlus` 提供額外的檔案系統功能 (如檢查、裝載以及強制卸載)，可讓 Sun Cluster 在本機檔案系統上進行故障轉移。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` 也可用於同步化資源與其依賴的磁碟裝置群組的啓動。如需詳細資訊，請參閱第 60 頁的 「資源、資源群組與資源類型」。

Syncdir 裝載選項

`syncdir` 裝載選項可用於將 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`，效能就會明顯改善。如果您指定 `syncdir`，此項寫入便保證相容於 POSIX。如果沒有指定，您所看到的功能，將會與 UFS 檔案系統相同。例如，在某些情況下，沒有 `syncdir`，一直到關閉檔案您才會發覺出空間不足的狀況。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。由於您未指定 `syncdir` 而出現問題的情況極少發生，因此我們建議您不要指定 `syncdir`，這樣效能會得到提昇。

如果您使用的是基於 SPARC 的叢集，則 Veritas VxFS 不具有同 UFS 的 `syncdir` 裝載選項對等的裝載選項。未指定 `syncdir` 裝載選項時，VxFS 的行爲會與 UFS 的行爲相同。

請參閱第 76 頁的 「檔案系統 FAQ」，以取得有關整體裝置和叢集檔案系統的常見問題。

磁碟路徑監視

目前版次的 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。請參考「*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。

注意 – 在執行 Sun Cluster 3.1 4/04 軟體之前發行的舊版本的節點上不支援 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。

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。藉由 `scdpm` 指令提供的選項，可讓您監視叢集中單一節點或所有節點的磁碟路徑。請參閱 `scdpm`(1M) 線上援助頁，以取得有關指令行選項的詳細資訊。

DPM 元件是從 `SUNWscu` 套件安裝的。`SUNWscu` 套件是依照標準 Sun Cluster 安裝程序安裝的。請參閱 `scinstall` (1M) 線上援助頁，以取得安裝介面詳細資訊。下表說明了 DPM 元件的預設安裝位置。

\u4f4d\u7f6e	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。當某個節點啓動時，DPM 常駐程式 (`scdpmd`) 將由 `rc.d` 程序檔啓動。如果發生問題，此常駐程式將由 `pmfd` 管理並自動重新啓動。以下清單說明了 `scdpmd` 在初次啓動時的運作方式。

注意 – 在啓動時，每個磁碟路徑的狀態都將初始化爲 `UNKNOWN`。

---

1. DPM 常駐程式可從上一個狀態檔或 CCR 資料庫中收集磁碟路徑與節點名稱資訊。請參考第 32 頁的 「叢集配置儲存庫 (CCR)」，以取得關於 CCR 的詳細資訊。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 常駐程式可通知 Sun Cluster 事件框架，並透過 UNIX `syslogd`(1M) 機制來記錄路徑的新狀態。

注意 – 關於此常駐程式的所有錯誤均由 `pmfd` (1M) 報告。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 常駐程式可監視透過多重路徑驅動程式 (如 MPxIO、HDLM 與 PowerPath) 可看到的邏輯路徑的可用性。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。此指令對於列印故障磁碟的清單以及從檔案監視磁碟路徑也很有用。

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。請參考「*Sun Cluster* **系統管理指南 (適用於** *Solaris* **作業系統)**」中的「藉由圖形化使用者介面來管理 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

從任何作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。如果未指定任何節點，則不需要節點名稱，而預設爲 `all`。下列表格說明了磁碟路徑的命名慣例。

注意 – 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) 線上援助頁。

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
- 檢視叢集內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。

法定數目與法定裝置

本節包含下列主題：

注意 – 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。

如需關於 CMM 的更多資訊，請參閱「*Sun Cluster* **簡介 (適用於** *Solaris* **作業系統)**」中的「叢集成員關係」。

叢集分割區會導致兩種問題：

- Split 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。因爲一個分割區中的節點無法與其他分割區中的節點通信，所以每個分割區會認爲其自身是唯一的分割區。

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。當您在某節點 (該節點不在最後使用的叢集分割區內) 上啓動該叢集時，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

如需有關該指令的更多資訊，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。

節點依據節點的狀態投票：

- 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*。
- 當系統管理員將節點置於維護狀態時，節點的票數爲 *0*。

法定裝置根據連線至該裝置的投票數目進行投票。當您配置法定裝置時，Sun Cluster 軟體會爲法定裝置指定票數 *N*-1，其中 *N* 是與法定裝置連線的投票數目。例如，連線至兩個非零票數之節點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。

您可以在叢集安裝過程中配置法定裝置，也可在稍後利用「*Sun Cluster* **系統管理指南** *(***適用於** *Solaris* **作業系統***)*」的「管理法定數目」中所述的程序來配置法定裝置。

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*)。發生此情形時，不是所有的節點均可通訊，所以個別節點或節點子集可能會嘗試形成個別或子集叢集。每個子集或分割區可能認爲自己擁有唯一的多重主機裝置存取和所有權。嘗試寫入磁碟的多個節點會導致資料毀損。

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磁碟。只有目前的成員可以存取磁碟，因此維持了資料的完整性。

磁碟裝置服務爲利用多重主機裝置的服務提供防故障備用功能。當目前是磁碟裝置群組的主要 (所有者) 叢集成員故障或無法到達時，會選出新的主要成員，繼續提供磁碟裝置群組的存取，期間只出現輕微的中斷情形。在此處理程序期間，啓動新的主要成員之前，舊的主要成員會放棄存取裝置。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。

SunPlex 系統使用 SCSI 磁碟保留來實施故障隔離。使用 SCSI 保留，便可以將故障節點與多重主機裝置「隔離」，防止它們存取這些磁碟。

SCSI-2 磁碟保留支援一種保留形式，授與存取權給所有連接磁碟的節點 (沒有保留存在) 或限制單一節點的存取權 (握有保留的節點)。

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。當發生此故障隔離時，一般會使隔離節點發生混亂，並在其主控台上出現「保留衝突」訊息。

偵測到有節點不再是叢集成員時，會放置 SCSI 保留在此節點與其他節點之間共用的所有磁碟上，所以就發生保留衝突的狀況。隔離節點可能不知道，自己已被隔離，而且如果它嘗試存取其中一個共用磁碟，就會偵測到保留和當機。

故障隔離的 Failfast 機制

叢集框架用來確保故障節點無法重新啓動與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。此 ioctl 爲磁碟驅動式的指示詞，會讓節點在無法存取已被保留爲其他節點之用的磁碟時，有能力自我混亂。

MHIOCENFAILFAST ioctl 會使驅動程式檢查由節點發佈給磁碟的每次讀取與寫入的錯誤傳回，以取得 `Reservation_Conflict` 錯誤碼。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`，前景與背景的控制流路徑都會混亂。

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。不管您是否有 SCSI-2 磁碟或 SCSI-3 磁碟，failfast 機制的運作都一樣。

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，在共用磁碟上放置了保留，且當沒有法定容量的節點嘗試存取共用磁碟時，它會收到保留衝突並且由於 failfast 機制而混亂。

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。您可以在基於 SPARC 的叢集中，於 OpenBoot PROM `ok` 提示符號處，使用 `eeprom`(1M) 來設定 `auto-boot?`，或者在基於 x86 的叢集中，於 BIOS 啓動後使用您所選擇執行的 SCSI 公用程式來設定。

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+1 配置中 (其中 *N* 個法定裝置中的每一個均連線到其中一個 `1` 至 *N* 個節點以及第 `N+1` 個節點)，當所有 `1` 至 *N* 個節點或 *N*/ **2** 個節點中的任何一個節點失效時，叢集將不會當機。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，當 *N*- **1** 個節點中的任何一個節點當機失效時，叢集可免於當機。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。

如需要避免使用的法定數目配置範例，請參閱第 51 頁的 「不正確的法定數目配置」。
如需建議使用的法定數目配置範例，請參閱第 48 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。若不遵守，可能會危及叢集的可用性。

- 請確保 Sun Cluster 軟體支援您的特定裝置作爲法定裝置。

注意 – 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
- 在包含兩個節點的配置中，您必須至少配置一個法定裝置，以確保在一個節點發生故障時，另一個節點可以繼續工作。請參閱圖 3–2。

如需要避免使用的法定數目配置範例，請參閱第 51 頁的 「不正確的法定數目配置」。
如需建議使用的法定數目配置範例，請參閱第 48 頁的 「建議使用的法定數目配置」。

遵守法定裝置最佳方式

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**不**需要配置其他法定裝置，因爲您的配置即爲最佳配置。

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，即使所有節點均正常運作，如果所有磁碟不可用，節點亦無法形成叢集。

注意 – 有時，在特定環境下，爲了滿足您的需要，也許可以降低叢集整體可用性。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 50 頁的 「非典型的法定數目配置」中概述的配置中，叢集的可用性降低：法定票數超出了節點票數。叢集具有以下特性：遺失節點 A 與節點 B 之間共用儲存體的存取權時，整個叢集將發生故障。

請參閱第 50 頁的 「非典型的法定數目配置」，以取得此最佳方式之例外情況的資訊。

- 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。該法定數目配置會加速故障隔離程序。請參閱第 49 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。

如需要避免使用的法定數目配置範例，請參閱第 51 頁的 「不正確的法定數目配置」。
如需建議使用的法定數目配置範例，請參閱第 48 頁的 「建議使用的法定數目配置」。

建議使用的法定數目配置

如需要避免使用的法定數目配置範例，請參閱第 51 頁的 「不正確的法定數目配置」。

兩個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。這兩票可以來自兩個叢集節點，或一個節點和一個法定裝置。

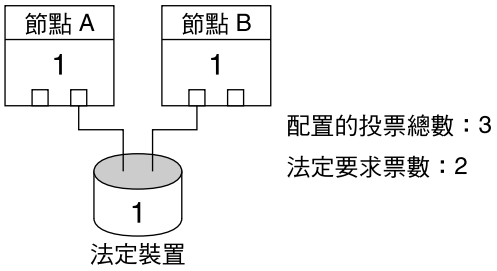
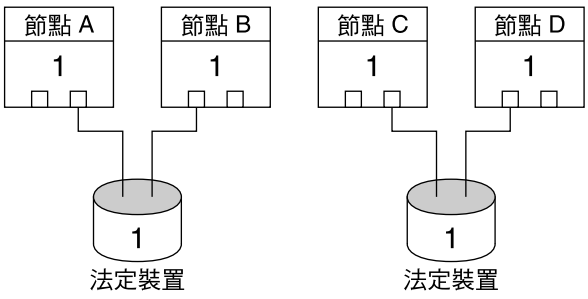


圖 3–2 兩個節點的配置

多於兩個節點的配置中的法定數目

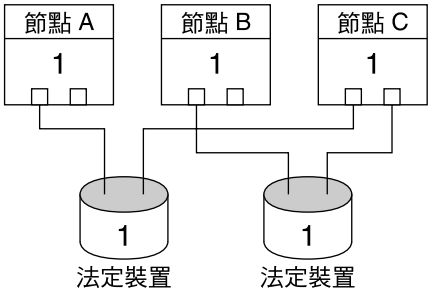
可以在無法定裝置的情況下配置多於兩個節點的叢集。但是，這樣做將無法在叢集中不具備多數節點的情況下啓動叢集。



配置的投票總數：6

法定要求票數：4

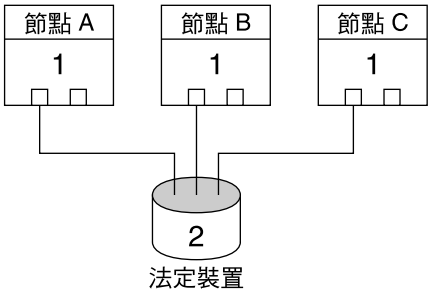
在此配置中，每一對均必須可用，其中一對才可存活。



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 假定您正在節點 A 與節點 B 上運行對任務至關重要的應用程式 (例如，Oracle 資料庫)。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。

如需關於與此例外相關的最佳方式之資訊，請參閱第 46 頁的 「遵守法定裝置最佳方式」。

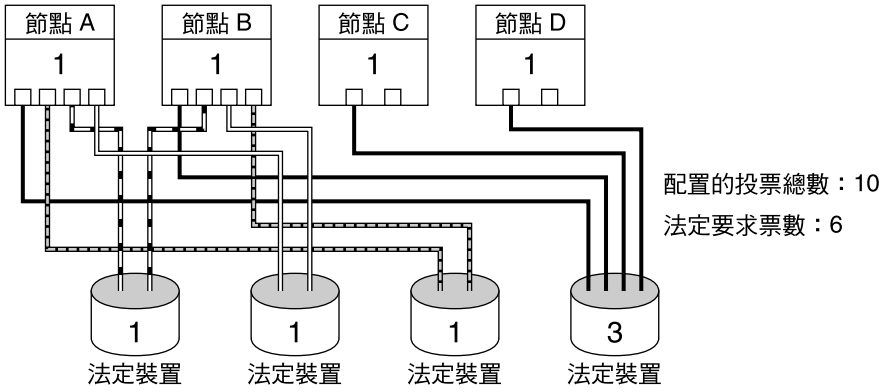
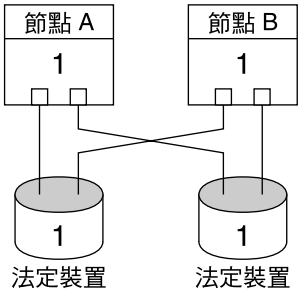


圖 3–3 非典型的配置

不正確的法定數目配置

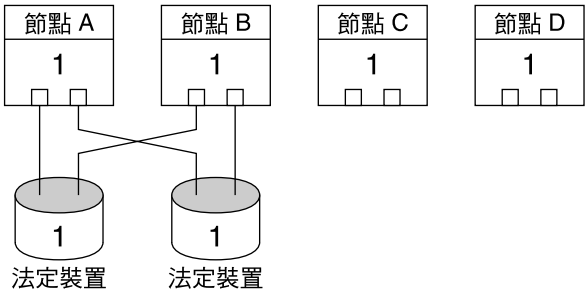
如需建議使用的法定數目配置範例，請參閱第 48 頁的 「建議使用的法定數目配置」。



配置的投票總數：4

法定要求票數：3

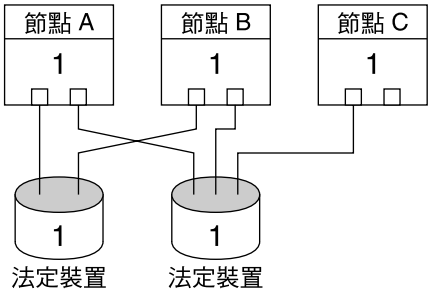
此配置違背了法定裝置票數應該嚴格少於節點票數的最佳方式。



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

資料服務一詞說明協力廠商應用程式，例如 Sun Java System Web Server (以前稱爲 Sun Java System Web Server)；對於基於 SPARC 的叢集而言，說明 Oracle，它已被配置爲在叢集上執行，而不是在單一伺服器上執行。資料服務包含一個應用程式、專用的 Sun Cluster 配置檔案以及 Sun Cluster 管理方法，可控制應用程式的下列動作。

- 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 將執行於單一應用程式伺服器上的應用程式 (單一伺服器模型) 與執行於叢集上的同一個應用程式 (叢集伺服器模型) 進行比較。請注意，就使用者的觀點而言，這兩種配置除了叢集應用程式可能執行較快且較爲高度可用以外，並無任何差別。

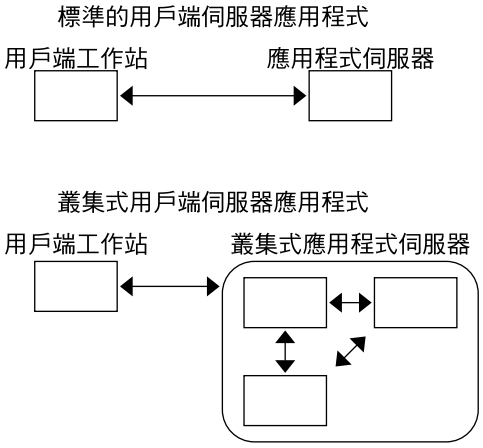


圖 3–4 標準與叢集用戶端/伺服器配置

在單一伺服器模型中，可配置應用程式，以透過特定的公用網路介面 (一個主機名稱) 來存取伺服器。主機名稱與實體伺服器有關。

在叢集伺服器模型中，公用網路介面是一個**邏輯主機名稱**或一個**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 — 它們是不能互相交換的。其他資料服務則容許您指定邏輯主機名稱或共用位址。請參考各資料服務的安裝與配置檔，以取得您必須指定的介面類型的詳細資訊。

網路資源與特定的實體伺服器無關 — 它可在實體伺服器之間遷移。

網路資源最初與一個節點 (**主要節點**) 相關聯。如果主要節點發生故障，則網路資源與應用程式資源會故障轉移至其他叢集節點 (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 比較單一伺服器模型與叢集伺服器模型。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。

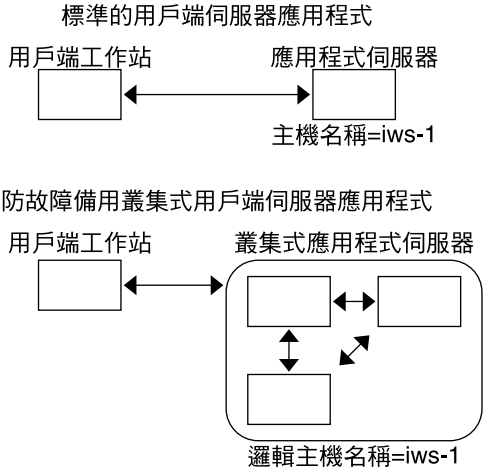


圖 3–5 固定主機名稱與邏輯主機名稱

共用位址最初也與一個節點相關聯。此節點稱爲整體介面節點。共用位址用來作爲叢集的單一網路介面。稱之爲**整體介面**。

邏輯主機名稱模型與可延伸服務模型的差異，在於後者的每個節點在其回送介面中亦主動配置有共用位址。此配置可使同時在數個節點上作用中的資料服務具有多重實例。「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。

如果整體介面節點發生故障，可將共用位址提供到也在執行應用程式實例的另一個節點上 (從而使此另一個節點成爲新的整體介面節點)。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 比較了單一伺服器配置與叢集可延伸服務配置。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。

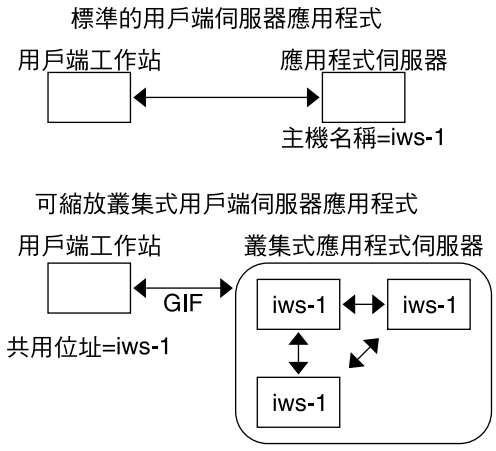


圖 3–6 固定主機名稱與共用位址

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 Resource Group Manager (RGM) 控制下執行，Resource Group Manager (RGM) 使用這些方法來啓動、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)。

除了 Sun Cluster 軟體提供的方法，SunPlex 系統還提供了 API 與數種資料服務開發工具。這些工具可讓應用程式設計師藉由 Sun Cluster 軟體來開發讓其他應用程式作爲高度可用資料服務執行所需的資料服務方法。

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。故障轉移服務使用**故障轉移資源群組**，它是應用程式實例資源與網路資源 (**邏輯主機名稱**) 的儲存區。邏輯主機名稱是 IP 位址，可以在某個節點配置上線，稍後自動在原始節點配置下線，並在其他節點配置上線。

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。如果錯誤監視器偵測到錯誤，則會嘗試於同一節點重新啓動實例，或於其他節點啓動實例 (故障轉移)，視資料服務的配置方式而定。

可延伸資料服務

可延伸的資料服務具有在多重節點上的作用中實例之潛力。可延伸的服務使用兩種資源群組：包含應用程式資源的**可延伸資源群組**，以及包含可延伸服務所依賴的網路資源 (**共用位址**) 之故障轉移資源群組。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。放置可延伸服務的所有節點，均使用相同的共用位址來放置服務。

服務要求經由單一網路介面 (整體介面) 進入叢集，並且根據**平衡資料流量策略**所設定的數種預先定義演算法之一來分配給節點。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。請注意，在不同的節點上可能有多重的整體介面主控其他共用的位址。

對於可延伸服務，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。如果此項應用程式實例失敗時，此實例會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，服務會繼續在剩餘的節點上執行，可能造成服務產量的降低。

注意 – 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 顯示的範例是可延伸服務的故障轉移和可延伸資源群組，以及它們之間的相依性。此範例顯示三個資源群組。故障轉移資源群組包含高可用性 DNS 的應用程式資源，以及高可用性 DNS 和高可用性 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，可延伸的資源群組與故障轉移資源群組之間存在資源群組相依性 (實線)，並且所有 Apache 應用程式資源都依賴作爲共用位址的網路資源 `schost-2` (虛線)。

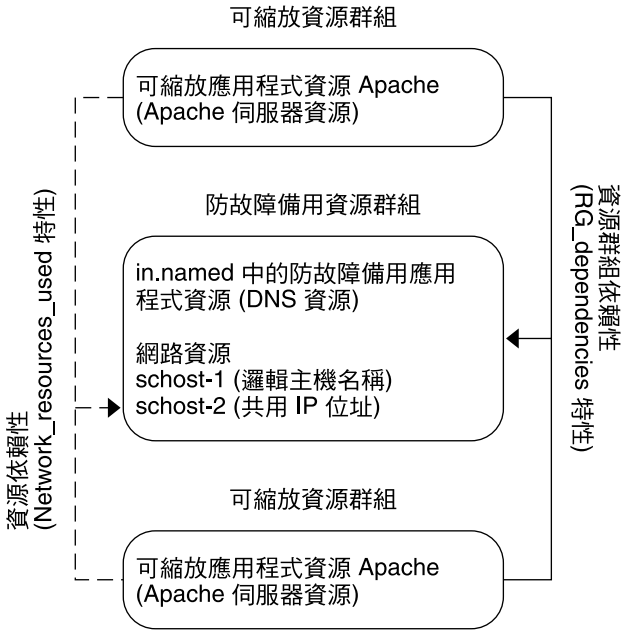


圖 3–7 SPARC: 故障轉移與可延伸的資源群組範例

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。

可延伸資料服務的類別有兩種：*Pure* 與 *Sticky*。Pure 服務是，它的任何實例均可回應用戶端要求。Sticky 服務是用戶端傳送要求給相同實例的服務。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，在三節點的叢集中，我們假設每一個節點的權重是 1。每一個節點代表該服務，分別服務 1/3 的任何用戶端要求。管理員可以透過 `scrgadm(1M)` 指令介面或 SunPlex Manager GUI 來隨時變更權重。

Sticky 服務有兩種方式，即 *Ordinary Sticky* 與 *Wildcard Sticky*。Sticky 服務允許在多個 TCP 連接上並行處理應用程式層次階段作業，以共用 in-state 記憶體 (應用程式階段作業狀態)。

Ordinary Sticky 服務允許用戶端共用多個並行 TCP 連接之間的狀態。就偵聽單一通訊埠的該伺服器實例而言，用戶端被稱爲「Sticky」。只要該實例維持啓動與可存取的狀態，且當此服務在線上時，平衡資料流量策略未曾改變，即可保證用戶端的所有要求均會到達相同的伺服器實例。

例如，用戶端上的網際網路瀏覽器使用三種不同的 TCP 連線連接到共用 IP 位址的 80 通訊埠，但是連線是在服務時交換快取的階段作業資訊。

一般化的 Sticky 策略擴展至多重可延伸服務，在相同實例上進行幕後交換階段作業資訊。當這些服務在相同實例上於幕後交換階段作業資訊時，用戶端被稱爲「Sticky」(就偵聽不同通訊埠的同一個節點上的多個伺服器實例而言)。

例如，電子商務網站上的客戶使用一般的 HTTP (80 通訊埠) 將物品填入其購物車，但是會切換至 SSL (443 通訊埠) 傳送安全性資料，以使用信用卡付購物車中物品的帳款。

Wildcard Sticky 服務使用動態指定的通訊埠編號，但是仍然希望用戶端要求會到達相同的節點。用戶端在相關的同一 IP 位址之通訊埠爲「Sticky wildcard」。

這種策略的典型範例是被動模式 FTP。用戶端連接至 FTP 伺服器的 21 通訊埠，然後被伺服器通知以動態埠範圍連接回至接收埠伺服器。對此 IP 位址的所有要求，均會轉遞至伺服器經由控制資訊通知用戶端的同一節點。

請注意，對此每一種 Sticky 策略，依預設都會使用加權平衡資料流量策略，因此用戶端的起始要求會被導向平衡資料流量程式所指定的實例。在用戶端建立與執行實例之節點的關係之後，只要該節點是可存取的，且平衡資料流量策略未變更，則後續的要求會被導向該實例。

特定平衡資料流量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。可使用 `Load_balancing_weights` 特性的 `LB_WEIGHTED` 值來設定此策略。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。如果指定 X=權重與 A=所有作用中節點的總權重，當連接總數足夠大時，作用中的節點便可預期將新連接總數的大約 X/A 導向作用中節點。此策略不針對個別的要求。

  請注意，此策略並非全體循環式。全體循環式策略一定會將用戶端的每個要求送至不同的節點：將第一個要求送到節點 1，將第二個要求送到節點 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。可使用 `Load_balancing_policy` 資源特性的 `LB_STICKY` 值來設定此策略。
- Sticky-wildcard。此策略是一般「Sticky」策略的超集合。對於以 IP 位址來識別的可延伸服務而言，是由伺服器來指定通訊埠 (而且事先無法知道)。通訊埠可能會變更。可使用 `Load_balancing_policy` 資源特性的 `LB_STICKY_WILD` 值來設定此策略。

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。發生此情況時，原來的次要節點就成爲新的主要節點。故障回復設定指定當原來的主要節點回到線上時會採取的動作。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。您可使用故障回復資源群組特性設定來指定需要的選項。

在某些情況下，假如放置資源群組的原始節點重複故障和重新開機，設定故障回復可能會造成資源群組的可用性降低。

資料服務故障監視器

每個 SunPlex 資料服務均提供了故障監視器，定期地測試資料服務以判斷其運作狀況。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。根據測試所傳回的資訊，可以起始預先定義的動作，如重新啓動常駐程式或進行故障轉移。

開發新的資料服務

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。如果您要當作故障轉移或可延伸服務來執行的應用程式，目前不是由 Sun 所提供，您可以使用 API 或 DSDL API，將您的應用程式配置成爲故障轉移或可延伸的服務。

有一套準則可用來斷定應用程式是否可成爲故障轉移服務。特定的準則在 SunPlex 文件中有所說明，其中說明了可用於您的應用程式的 API。

在此，我們提供一些準則來協助您瞭解，您的服務是否可以利用可延伸資料服務架構。請參閱第 55 頁的 「可延伸資料服務」一節，以取得關於可延伸服務的更多一般資訊。

滿足下列準則的新服務，則可以使用可延伸服務。如果現存的服務不完全符合這些準則，可能需要改寫些某些部分，使服務能夠符合準則。

可延伸資料服務具有下列特性。首先，這類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。

第二，如果服務提供外部邏輯資料儲存體，從多部伺服器對此儲存體做並行存取時，必須同步化，以避免將之變更時遺失更新或讀取資料。請注意，我們用「外部」來區別儲存處與記憶體內部狀態，而使用「邏輯」是因爲儲存處顯示爲單一實體，雖然其自身可進行複製。此外，此邏輯資料儲存體具有當任何伺服器實例更新儲存體時其他實例會立即看到更新的特性。

SunPlex 系統透過其叢集檔案系統與其整體原始分割區來提供此類的外部儲存體。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。在執行此服務的多個實例時，每個實例均存取此外部登錄，而且可能同時存取此登錄。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。服務可以透過 `fcntl(2)` 和 `lockf(3C)` 的一般 Solaris 檔案鎖定，來達到所需的同步化。

此類型儲存體的另一個範例是後端資料庫，例如基於 SPARC 的叢集之高度可用 Oracle 或 Oracle Real Application Clusters。請注意，這種後端資料庫伺服器使用資料庫查詢或更新異動來提供內建的同步化，因此多重伺服器實例不需要實作自己的同步化。

目前不是可延伸服務的服務範例，是 Sun 的 IMAP 伺服器。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，實例可能會具有與其他實例的資料區隔的私有資料。在此情況下，服務不需要關心自己的同步化並行存取，因爲資料是私有的，而且只有該實例可以操作資料。因此，您必須愼防將此私有資料儲存在叢集檔案系統之下，因爲它可能會變成可全域存取。

資料服務 API 與資料服務開發檔案庫 API

SunPlex 系統提供下列項目，可使應用程式具備高可用性：

- 作爲 SunPlex 系統一部分提供的資料服務
- 資料服務 API
- 資料服務發展檔案庫 API
- 「一般」資料服務

「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」描述了如何安裝與配置隨 SunPlex 系統提供的資料服務。「*Sun Cluster 3.1 9/04 Software Collection for Solaris OS (SPARC Platform Edition)*」描述了如何裝備其他應用程式，使其在 Sun Cluster 框架下高度可用。

Sun Cluster API 可讓應用程式設計師開發可啓動與停止資料服務實例的故障監視器及程序檔。利用這些工具，應用程式可以變成具備故障轉移和可延伸資料服務。另外，SunPlex 系統還提供可用來快速產生應用程式必需的啓動與停止方法的「一般」資料服務，以使其作爲故障轉移或可延伸服務來執行。

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。叢集軟體可使用多重交互連接來達到高可用性以及增進效能。對於內部通訊 (例如檔案系統資料或可延伸服務資料) 而言，會以全體循環式將訊息散置在所有可用的交互連接中。

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交互連接進行通訊，應用程式必須使用安裝叢集時配置的專用主機名稱。例如，如果節點 1 的專用主機名稱是 `clusternode1-priv`，請使用該名稱以透過節點 1 的叢集交互連接進行通訊。使用此名稱開啓的 TCP 插槽透過叢集交互連接進行路由，並可在出現網路故障時透明式地重新路由。

請注意，由於專用主機名稱可以在安裝時配置，因此叢集交互連接可使用當時選取的任何名稱。可使用 `scha_privatelink_hostname_node` 引數從 `scha_cluster_get` (3HA) 取得實際名稱。

在應用程式層次使用叢集交互連接時，每一對節點之間使用單一的交互連接，但若可能的話，不同的節點配對之間應使用個別的交互連接。例如，考慮到應用程式在三個基於 SPARC 的節點上執行，並透過叢集交互連接來進行通訊。節點 1 與節點 2 之間的通訊可能在 `hme0` 介面上進行，而節點 1 與節點 3 之間的通訊可能在介面 `qfe1` 上進行。也就是說，任意二個節點之間的應用程式通訊將限制於單一交互連接，內部叢集通訊則散置在所有的交互連接。

請注意，應用程式和內部叢集通訊共用交互連接，因此應用程式可用的頻寬是由其他叢集通訊所使用的頻寬來決定。在發生故障時，內部通訊可以在其餘交互連接中循環，而發生故障的交互連接上的應用程式連接也可以切換到運作中的交互連接。

有兩種類型的位址支援叢集交互連接，專用主機名稱上的 `gethostbyname`(3N) 通常傳回兩個 IP 位址。第一個位址稱爲**邏輯** *pairwise* **位址**，第二個位址稱爲 **邏輯** *pernode* **位址**。

每一對節點會被指派個別的邏輯 pairwise 位址。這個小型邏輯網路支援連接的故障轉移。每一個節點還會被指派一個固定的 pernode 位址。也就是說，在每個節點上 `clusternode1-priv` 的邏輯 pairwise 位址不同，而它的邏輯 pernode 位址是相同的。不過，一個節點本身不會有 pairwise 位址，因此，節點 1 上的 `gethostbyname` (`clusternode1-priv`) 僅傳回邏輯 pernode 位址。

請注意，透過叢集交互連接接受連接、然後出於安全性原因而驗證 IP 位址的應用程式，必須檢查從 `gethostbyname` 傳回的所有 IP位址，而不僅僅檢查第一個 IP 位址。

如果您要求應用程式在各個點都是一致的 IP 位址，請將應用程式配置爲在用戶端以及伺服器都是連結到 pernode 位址，這樣所有的連接看起來都會是透過 pernode 位址往來。

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 (以前稱爲 Sun Java System Web Server) 或 Apache Web Server，均使用這些應用程式所依賴的網路位址 (邏輯主機名稱與共用位址)。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 `SUNW.oracle-server` 資源類型；而 Sun Cluster HA for Apache 屬於 `SUNW.apache` 資源類型。

注意 – 資源類型 `SUNW.oracle-server` 僅在基於 SPARC 的叢集中使用。

資源是在整個叢集中定義的**資源類型**的個體化。有數種已定義的資源類型。

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

SUNW.HAStorage 及 HAStoragePlus 資源類型用來將資源的啓動與該資源所依賴之磁碟裝置群組的啓動同步化。它可確保在資料服務啓動之前，叢集檔案系統裝載點的路徑、整體裝置和裝置群組名稱是可用的。如需詳細資訊，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 中已經可以使用 HAStoragePlus 資源類型，並在此資源中新增了另一項可讓本機檔案系統具備高可用性的功能。如需有關此功能的詳細資訊，請參閱第 39 頁的 「HAStoragePlus 資源類型」。)

RGM 管理的資源會分成群組，稱爲**資源群組**，讓群組可以一個單位的方式來管理。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。

注意 – 當您將包含應用程式資源的資源群組啓動爲線上時，即會啓動應用程式。資料服務啓動方法會等到應用程式啓動並執行之後，才順利結束。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。請參考「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」，以取得有關本程序的更多資訊。

Resource Group Manager (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 Sun Java System Web Server 資源管理 API (RMAPI) 所建立。叢集管理員可建立和管理儲存區中的資源，稱爲**資源群組**。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 動作可使資源及資源群組在線上狀態與離線狀態之間移動。可套用至資源與資源群組的狀態與設定之完整說明列於第 61 頁的 「資源和資源群組的狀態與設定」一節中。請參考第 60 頁的 「資源、資源群組與資源類型」，以取得關於如何在 RGM 控制下啓動資源管理專案的資訊。

資源和資源群組的狀態與設定

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 在動態的「狀態」間移動資源群組。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。可以使用 scrgadm(1M) 指令，讓 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。若要讓群組中的任何資源成爲作用的，它就必須是被管理的。

  在某些資料服務中，諸如可延伸式 Web 伺服器，在設定網路資源前與停止網路資源後都有工作要做。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 及 FINI 方法的最常使用方式是用於可延伸服務的網路資源，但也可用於應用式不做的初始化或清除工作。

- 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。

  如果出於某種原因，您要使資源在所有叢集節點上均不可用，則請停用資源。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。

  在切換保護移轉或故障轉移期間，這些狀態隨著經由叢集重新配置步驟而發生的叢集轉換而變更。亦可經由管理動作來加以變更。`scswitch(1M)` 可用於變更資源或資源群組的線上狀態或離線狀態。

  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。

  假如一個資源群組爲離線的，則它所有的資源均爲離線的。假如一個資源群組爲線上的，則它所有的資源均爲線上的。

  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l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。

資源及資源群組特性

您可以爲 SunPlex 資料服務的資源及資源群組配置特性值。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。如需標準特性集，請參閱「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」中的「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 使資料服務處於線上狀態時，可將此資料服務配置爲在 Solaris 專案名稱下啓動。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。若將資源或資源群組對應至專案 ID，便可讓您使用 Solaris 環境中可用的複雜控制，以管理叢集內的工作量與耗用量。

注意 – 僅當您將目前版次的 Sun Cluster 軟體與 Solaris 9 配合執行時，才可以執行此配置。

在叢集環境中使用 Solaris 管理功能，可讓您確定已經爲最重要的應用程式提供了優先權 (當它與其他應用程式共用節點時)。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。若使用此處所述的管理功能，可能會透過防止其他低優先權應用程式過度耗用系統供給品 (如 CPU 時間)，來提高重要應用程式的可用性。

注意 – 此功能的 Solaris 說明文件說明了 CPU 時間、程序、作業以及與「資源」相似的元件。同時，Sun Cluster 說明文件使用「資源」一詞來說明處於 RGM 控制下的實體。以下一節將使用「資源」一詞來表示處於 RGM 控制下的 Sun Cluster 實體，使用「供給品」一詞來表示 CPU 時間、程序以及作業。

本節提供配置資料服務以啓動所指定 Solaris 9 `project`(4) 中程序的概念性說明。本節也說明了幾種故障轉移方案與建議，以計劃使用 Solaris 環境提供的管理功能。如需有關管理功能的概念與程序詳細說明文件，請參考「*Solaris 9 System Administrator Collection*」中的「*System Administration Guide: Resource Management and Network Services*」。

若配置資源及資源群組以在叢集內使用 Solaris 管理功能，請考慮使用以下高階程序：

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步驟 5 中建立的專案關聯起來。
7. 讓資源群組上線運作。

若要配置標準的 `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，請將 `-y` 選項與 `scrgadm`(1M) 指令配合使用。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「Standard Properties」，以取得特性定義。請參考 `r_properties`(5) 與 `rg_properties`(5)，以取得對特性的描述。

指定的專案名稱必須存在於專案資料庫 (/etc/project) 中，超級使用者必須配置爲已命名專案的成員。請參考「*Solaris 9 System Administrator Collection*」內「*System Administration Guide: Resource Management and Network Services*」中的「Projects and Tasks」，以取得關於專案名稱資料庫的概念資訊。請參考 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。

注意 – 使用者可以隨時將資源或資源群組與專案關聯起來。不過，只有使用 RGM 將資源或資源群組離線，然後重新使其線上運作，新的專案名稱才可會有效。

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。請參考 dispadmin (1M)，以取得關於將 FSS 設定爲預設排程程式的指令行說明。另請參閱 priocntl(1)、ps(1)、FSS(7)，以取得詳細資訊。
- 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

在配置資料服務以在 Sun Cluster 環境中使用 Solaris 提供的控制之前，您必須決定要如何在整個切換保護移轉或故障轉移中控制與追蹤資源。請先考慮識別叢集內的相依性，然後再配置一個新專案。例如，資源與資源群組依賴磁碟裝置群組。請使用 nodelist、failback、maximum_primaries 與 desired_primaries 資源群組特性 (用 scrgadm (1M) 配置) 來識別資源群組的 nodelist 優先權。請參考「*Sun Cluster Data Services Planning and Administration Guide for Solaris OS*」中的「Relationship Between Resource Groups and Disk Device Groups」，以取得資源群組與磁碟裝置群組之間節點清單相依性的簡短論述。如需詳細的特性說明，請參考 rg_properties (5)。

請使用 preferenced 與 failback 特性 (用 scrgadm(1M) 與 scsetup(1M) 配置) 來確定磁碟裝置群組的節點清單優先權。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***)*」的「管理磁碟裝置群組」中的「如何變更磁碟裝置特性」。請參考第 17 頁的 「SunPlex 系統的硬體與軟體元件」，以取得關於節點配置與故障轉移及可延伸資料服務之行爲的概念資訊。

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。在所有節點上的配置檔案中，所有應用程式的專案配置參數無需完全相同。與應用程式關聯的所有專案必須至少可透過該應用程式所有潛在主控者上的專案資料庫來存取。假設應用程式 1 由 *phys-schost-1* 主控，但可以潛在地切換至或故障轉移至 *phys-schost-2* 或 *phys-schost-3*。與應用程式 1 關聯的專案必須可在所有三個節點上 (*phys-schost-1*、*phys-schost-2* 與 *phys-schost-3*) 存取。

注意 – 專案資料庫資訊可以爲本機的 `/etc/project` 資料庫檔，也可以儲存在 NIS 對應或 LDAP 目錄服務中。

Solaris 環境允許靈活配置使用參數，並且 Sun Cluster 施加極少的限制。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。請參考 `rctladm`(1M)，以取得關於設定 `process.max-address-space` 值的詳細資訊。

在 Sun Cluster 中使用管理控制時，請適當配置記憶體限制，以防止應用程式發生不必要的故障轉移以及「交替」效果。一般而言：

- 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。

在叢集環境中，可將應用程式配置爲資源的一部分，並將資源配置爲資源群組 (RG) 的一部分。發生故障時，資源群組及其關聯的應用程式會將故障轉移至另一個節點。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。

注意 – 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，以及指定給每個作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作用中的程序。
- 在叢集的每個節點上配置專案資料庫的方式相同。

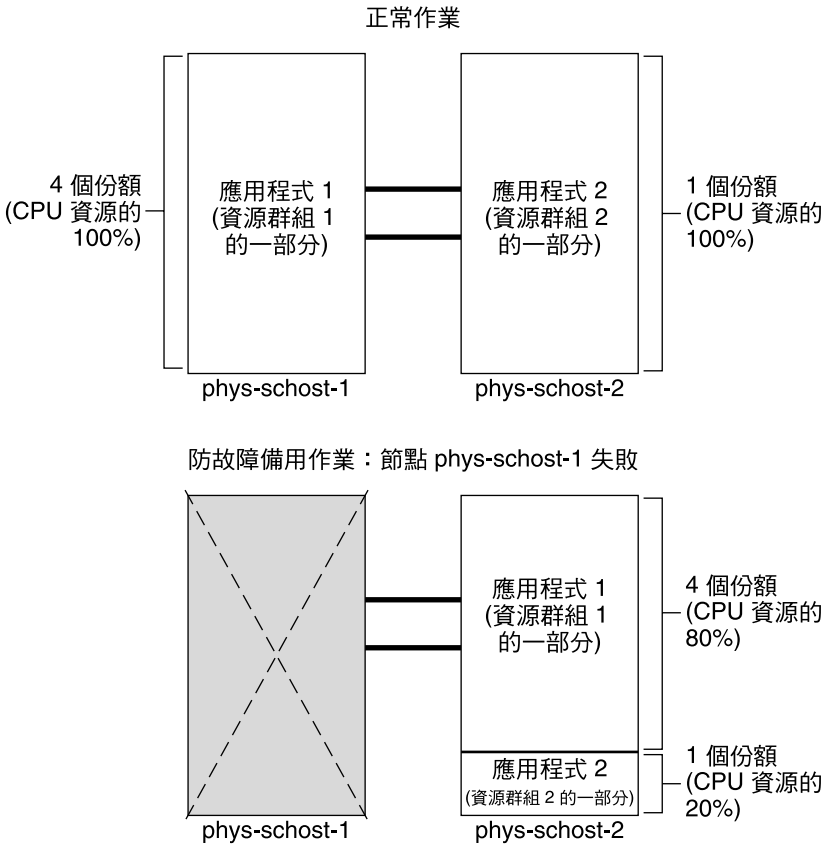
具有兩個應用程式的兩個節點叢集

您可以在兩個節點的叢集上配置兩個應用程式，以確保每個實體主機 (*phys-schost-1*、*phys-schost-2*) 作爲一個應用程式的預設主控者。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 及應用程式 2 關聯的所有專案必須在兩個節點的專案資料庫檔案中提供。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，在 `/etc/project` 檔案中，此項目指定了爲應用程式 1 分配了 4 個份額，爲應用程式 2 分配了 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。不過，每個應用程式可用的 CPU 時間百分比可以變更，這要取決於指定給需要 CPU 時間的每個程序的份額數。



具有三個應用程式的兩個節點叢集

在具有三個應用程式的兩個節點叢集上，您可以將一個實體主機 (*phys-schost-1*) 配置爲一個應用程式的預設主控者，將第二個實體主機 (*phys-schost-2*) 配置爲其餘兩個應用程式的預設主控者。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。

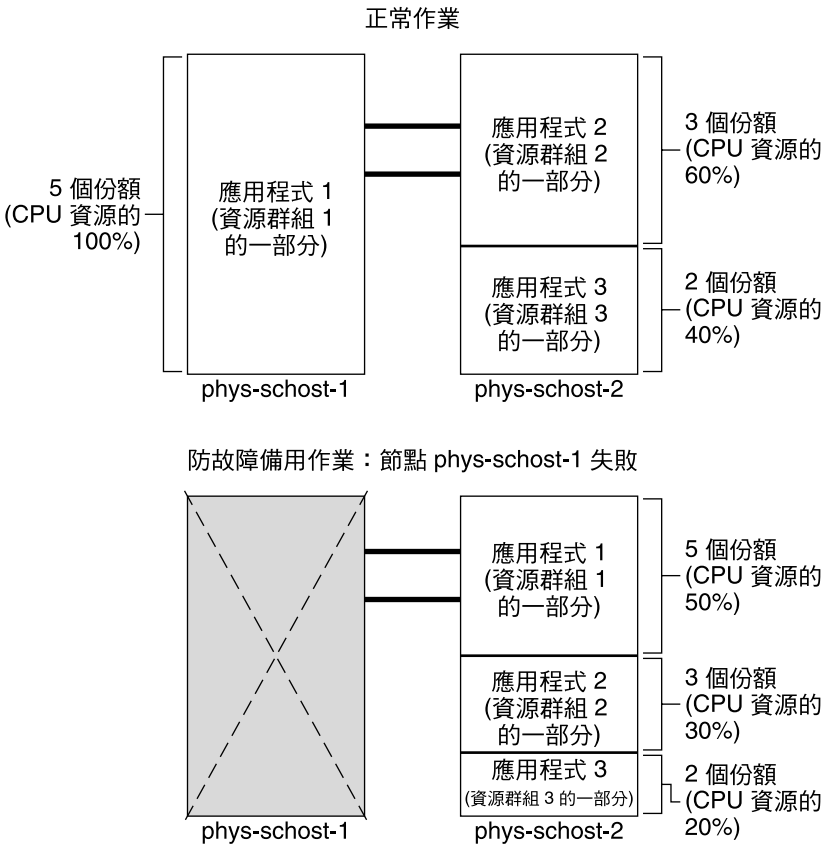
```
Prj_1:103:project for App-1:root::project.cpu-shares=(privileged,5,none)
Prj_2:104:project for App_2:root::project.cpu-shares=(privileged,3,none)
Prj_3:105:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

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。應用程式 2 與 3 在各自的預設主控者 *phys-schost-2* 上分別分配了 3 個與 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。

如果發生故障轉移或切換保護移轉，並將應用程式 1 切換至 *phys-schost-2*，所有三個應用程式的份額將保持相同。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

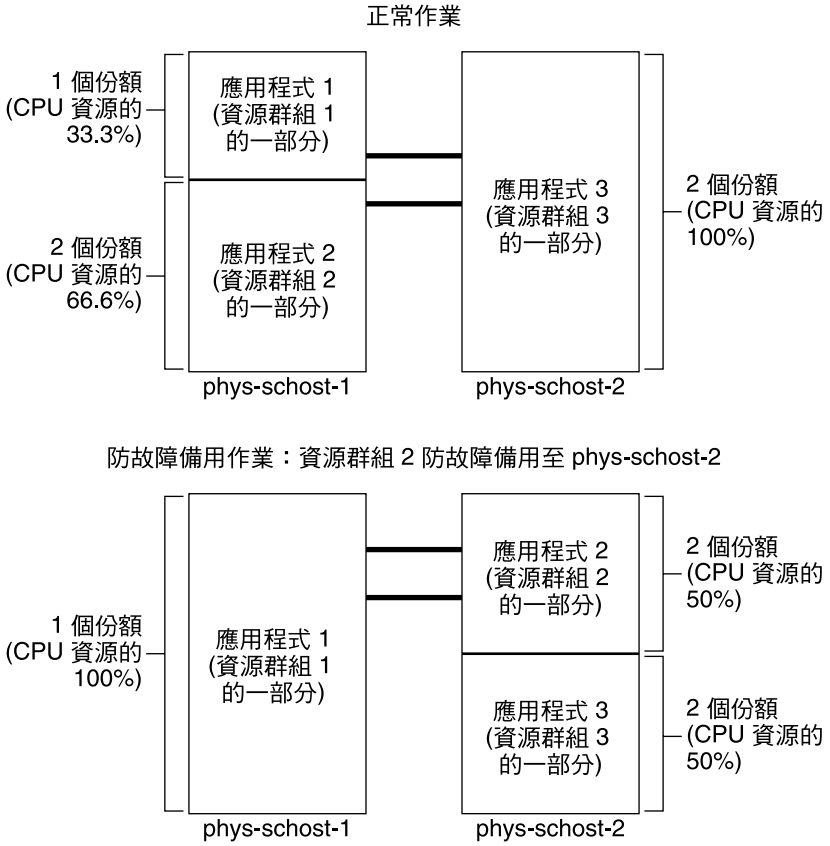
在多個資源群組使用同一個預設主控者的配置中，資源群組 (及其關聯應用程式) 可以發生故障轉移或切換至次要節點。同時在叢集內執行預設主控者。

注意 – 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，主要節點與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)
```

下圖展示了此配置的一般作業與故障轉移作業，其中包含應用程式 2 的 RG-2 可將故障轉移至 *phys-schost-2*。請注意指定的份額數不會變更。然而，每個應用程式可用的 CPU 時間之百分比可以變更，這要取決於指定給需要 CPU 時間的每個應用程式的份額數。



公用網路配接卡與 IP Network Multipathing

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 網際網路協定 (IP) 網路多重路徑軟體提供監視公用網路配接卡並將 IP 位址故障從一個配接卡轉移至另一個配接卡 (當偵測到錯誤時) 的基本機制。每一個叢集節點均擁有自己的 IP Network Multipathing 配置，該配置可能與其他叢集節點上的此種配置不同。

可將公用網路配接卡組織到 *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都可以處於作用中，或者您可以配置備用介面 (除非發生故障轉移，否則處於非作用中)。`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。所有網路存取都可將故障從發生錯誤的配接卡轉移到多重路徑群組中另一個功能正常的配接卡，從而維護節點的公用網路連接性。如果配置了備用介面，常駐程式將選擇此備用介面。否則，`in.mpathd` 將選擇具有最少 IP 位址數目的介面。因爲失效保護是發生在配接卡介面層次，所以較高層次的連接 (如 TCP) 不受影響，但是在失效保護期間的短暫延遲除外。若 IP 位址的故障轉移成功完成，將發送免費的 ARP 廣播。由此將維護與遠端用戶端的連接性。

注意 – 因爲 TCP 的壅塞恢復特性，TCP 端點在故障轉移成功之後可以承受更進一步的延遲，其中部分區段可能會在故障轉移期間遺失，因而啓動 TCP 的壅塞控制機制。

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*」。

注意 – IP Network Multipathing 機制的設計是爲了偵測和遮罩配接卡故障。其設計目的不是爲了使用 `ifconfig(1M)` 從管理員回復以移除其中一個邏輯 (或共用) IP 位址。Sun Cluster 軟體將邏輯與共用 IP 位址檢視爲 RGM 管理的資源。管理者增加或移除 IP 位

如需關於 IP 網路多重路徑 Solaris 實作的詳細資訊，請參閱安裝在叢集上的 Solaris 作業環境之適當說明文件。

作業環境版次	如需相關說明，請參閱...
Solaris 8 作業環境	「*IP Network Multipathing Administration Guide*」
Solaris 9 作業環境	「*System Administration Guide: IP Services*」中的「IP Network Multipathing Topics」

SPARC: 動態重新配置支援

正在以遞增方式分階段地開發 Sun Cluster 3.1 4/04 對動態重新配置 (DR) 軟體功能的支援。本節說明關於 Sun Cluster 3.1 4/04 支援 DR 功能的概念與注意事項。

請注意，爲 Solaris?DR 功能形成文件的所有需求、程序及限制，也適用於 Sun Cluster DR 支援 (作業環境暫停運作除外)。因此，在使用搭配 Sun Cluster 軟體的 DR 功能**之前**，請先參閱 Solaris DR 功能的說明文件。此外，還需特別注意DR 分解作業時會影響非網路 IO 裝置的問題。「*Sun Enterprise 10000 Dynamic Reconfiguration User Guide*」與「*Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual*」(從「*Solaris 8 on Sun Hardware*」或「*Solaris 9 on Sun Hardware*」集合中) 都可從 `http://docs.sun.com` 進行下載。

SPARC: 動態重新配置一般說明

DR 功能允許在執行中的系統內執行諸如移除系統硬體的作業。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 作業將影響板上的所有元件。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。

移除包含作用中元件的板將導致系統錯誤。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。由於 DR 子系統不對包含作用中元件的板執行作業，因此執行 DR 移除板作業永遠是安全的。

DR 加入板作業也永遠是安全的。新加入板上的 CPU 與記憶體會由系統自動納入服務中。不過，系統管理員必須手動配置叢集，以便以現用方式使用新加入板上的元件。

注意 – 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，較低層次報告特定錯誤時，較高層次將報告「未知的錯誤」。系統管理員應該忽略由較高層次報告的「未知的錯誤」。

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 叢集考慮事項

由於存在 CPU 裝置，因此 Sun Cluster 軟體不會拒絕執行 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 叢集考慮事項

出於 DR 目的，需要考慮兩種類型的記憶體。這兩種類型僅在用法上不同，其實際硬體相同。

作業系統使用的記憶體稱爲核心記憶體機架。在包含核心記憶體機架的板上，Sun Cluster 軟體不支援移除板作業，並將拒絕執行任何此種作業。若 DR 移除板作業關係到記憶體而非核心記憶體機架，則 Sun Cluster 將不會拒絕此作業。

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: 磁碟與磁帶機的 DR 叢集注意事項

Sun Cluster 會拒絕主要節點之作用中磁碟機上的 DR 移除板作業。DR 移除板作業可以在主要節點的非作用中磁碟機以及次要節點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。

注意 – Sun Cluster 會拒絕影響法定裝置可用性的 DR 作業，如需關於法定裝置與在其上執行 DR 作業之程序的注意事項，請參閱第 72 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「工作表：對法定裝置的動態重新配置」，以取得有關如何執行這些動作的詳細資訊。

SPARC: 法定裝置的 DR 叢集注意事項

如果 DR 移除板作業掛細到包含裝置 (爲法定數目配置) 介面的板，Sun Cluster 將拒絕此作業並識別將受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「工作表：對法定裝置的動態重新配置」，以取得有關如何執行這些動作的詳細資訊。

SPARC: 叢集交互連接介面的 DR 叢集注意事項

如果 DR 移除板作業關係到包含作用中叢集交互連接介面的板，Sun Cluster 會拒絕該作業，並指出可能會被該作業影響的介面。您必須使用 Sun Cluster 管理工具來停用作用中介面，然後 DR 作業才可以接著執行 (另請參閱下面的警告)。

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

Caution – Sun Cluster 要求每個叢集節點和其他所有叢集節點至少要有一個作業路徑。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。

SPARC: 公用網路介面的 DR 叢集注意事項

如果 DR 移除板作業關係到包含作用中公用網路介面的板，Sun Cluster 會拒絕該作業，並指出可能會被該作業影響的介面。在使用提供的作用中網路介面移除一個板之前，必須透過 `if_mpadm`(1M) 指令，首先將該介面上的所有通訊切換至多重路徑群組中另一個功能正常的介面。

Caution – 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」，以取得關於如何在公用網路介面上執行 DR 移除作業的詳細說明。

第 4 章

常見問題

INDEXTERM-343

本章包含有關 SunPlex 系統最常見問題的解答。問題是依照主題來排列。

高可用性 FAQ

- **到底什麼是高可用性系統？**

  SunPlex 系統將高可用性 (HA) 定義爲，即使發生一般可造成伺服器系統無法使用的故障，叢集仍可保持應用程式啓動並執行的能力。

- **叢集是利用何種處理程序來提供高可用性？**

  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。

- **故障轉移資料服務與可延伸資料服務之間的差異為何？**

  高可用性的資料服務有兩類，亦即故障轉移和可延伸。

  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。如果主要節點故障，在故障節點上執行的應用程式會移轉至另一個節點繼續執行。

  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」。叢集中可以有多個 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。

  如果應用程式所執行的任一節點故障，應用程式會繼續在其他的節點上執行，其中部分效能會降低，直到故障節點返回叢集之後才改善。

檔案系統 FAQ

- **我是否可以作為用戶端，執行一個或多個作為包含其他叢集節點的高度可用 NFS 伺服器的叢集節點？**

  不，不要做回送裝載。

- **是否可以將叢集檔案系統用於不在 Resource Group Manager 控制下的應用程式？**

  可以。然而，沒有 RGM 的控制，應用程式需要在其執行的節點發生故障後，以手動方式重新啓動。

- **是否所有叢集檔案系統均必須具有一個位於 /global 目錄下的裝載點？**

  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。

- **使用叢集檔案系統和匯出 NFS 檔案系統之間的差異是什麼？**

  有多處的差異：

  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，當某個檔案正在被多個節點存取進行讀取、寫入、檔案鎖定和非同步輸入/輸出。
  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屬性 (例如，使用 `chmod(1M)`)，此變更會立即反映在所有節點上。對於匯出式 NFS 檔案系統，此動作要花費較長時間。

- **檔案系統 /global/.devices/node@<nodeID> 出現在我的叢集節點上。是否可以使用此檔案系統，來儲存我希望其成為具有高可用性和整體性的資料？**

  這些系統檔會儲存整體裝置的名稱空間。它們不供一般使用。當它們爲整體時，從不以整體方式存取，每一節點只存取自己的整體裝置的名稱空間。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 FAQ

- **是否需要鏡像所有磁碟裝置？**

  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。

- **是否可以對本機磁碟 (開機磁碟) 使用一個容體管理程式，而對多重主機磁碟使用其他容體管理程式？**

  SPARC: 此配置受管理本機磁碟的 Solaris Volume Manager 軟體以及管理多重主機磁碟的 VERITAS Volume Manager 支援。但並不支援其他組合。

  x86: 否，此配置不受支援，因爲在基於 x86 的叢集中僅支援 Solaris Volume Manager。

資料服務 FAQ

- **哪些 SunPlex 資料服務可用？**

  「*Sun Cluster 3.1 9/04* **版本說明** *(***適用於** *Solaris* **作業系統***)*」的「支援的產品」中包含受支援資料服務的清單。

- **哪些應用程式版本受 SunPlex 資料服務的支援？**

  「*Sun Cluster 3.1 9/04* **版本說明** *(***適用於** *Solaris* **作業系統***)*」的「支援的產品」中包含受支援應用程式版本的清單。

- **是否可以寫入自己的資料服務？**

  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「資料服務開發程式庫參考」，以取得更多資訊。

- **在建立網路資源時，我是否應指定數字 IP 位址或主機名稱？**

  指定網路資源，最好是使用 UNIX 主機名稱，而非數字型 IP 位址。

- **建立網路資源時，使用邏輯主機名稱 (LogicalHostname 資源) 與共用位址 (SharedAddress 資源) 之間的差異是什麼？**

  除了 Sun Cluster HA for NFS 的情況外，說明文件提到在 `Failover` 模式資源群組中使用 `LogicalHostname` 資源時，可能會交替使用 `SharedAddress` 資源或 `LogicalHostname` 資源。使用 `SharedAddress` 資源需要一些額外的負擔，因爲叢集網路軟體是針對 `SharedAddress` 而非 `LogicalHostname` 配置的。

  使用 `SharedAddress` 的優點，是當您同時配置可延伸和故障轉移資料服務，而且要用戶端能夠使用相同的主機名稱來存取這兩種服務。在此情形下，`SharedAddress` 資源以及故障轉移應用程式資源是包含在一個資源群組中，而可延伸的服務資源是包含在另一個資源群組中，並配置爲使用 `SharedAddress`。於是，可延伸和故障轉移服務均可使用 `SharedAddress` 資源中配置的同一組主機名稱/位址。

公用網路 FAQ

- **SunPlex 系統支援哪些公用網路配接卡？**

  目前，SunPlex 系統支援 Ethernet (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。

- **在故障轉移中 MAC 位址的角色為何？**

  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。

- **SunPlex 系統是否支援設定 `local-mac-address?=true`？**

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

  您可以在基於 SPARC 的叢集中，於 OpenBoot PROM `ok` 提示符號後，使用 `eeprom`(1M) 來設定 `local-mac-address?`；或者在基於 x86 的叢集中，於 BIOS 啓動後使用您選擇執行的 SCSI 公用程式來設定。

- **當 IP Network Multipathing 在配接卡之間執行切換保護移轉時，我可以預期多長時間的延遲？**

  延遲可以達數分鐘。這是因爲在完成 IP Network Multipathing 切換保護移轉後，牽涉到送出免費的 ARP。然而，並不保證用戶端和叢集間的路由器將使用免費的 ARP。因此，在路由器上此 IP 位址的 ARP 快取項目逾時之前，它可能一直使用舊的 MAC 位址。

- **偵測到網路配接卡故障的速度有多快？**

  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 FAQ

- **所有的叢集成員是否需要相同的 root 密碼？**

  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。

- **節點啓動的順序是否相當重要？**

  在大多數情況下並不會有影響。然而，啓動順序對於防止 Amnesia 很重要 (請參考第 44 頁的 「關於故障隔離」，以取得關於 Amnesia 的詳細資訊)。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。這樣可以防止您意外啓動具有過時叢集配置資訊的節點。

- **是否需要在叢集節點中鏡像本機磁碟？**

  可以。雖然這種鏡像並非必要，但鏡像叢集節點的磁碟可以排除非鏡像磁碟故障而導致節點當機的情況。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。

- **叢集成員備份的問題有哪些？**

  您可以對叢集使用多種備份方法。其中一種方法是令某個節點連接磁帶機/磁帶庫作爲備份節點。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「備份與修復叢集」，以取得關於如何備份與修復資料的其他資訊。

- **節點何時正常到足以作為次要節點？**

  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。

叢集儲存體 FAQ

- **什麼原因使多重主機儲存體具備高可用性？**

  多重主機儲存體具備高可用性，是因爲有了鏡像 (或硬體式的 RAID-5 控制器) 而可以承受單一磁碟的遺失。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。

叢集交互連接 FAQ

- **SunPlex 系統支援哪些叢集交互連接？**

  目前，SunPlex 系統在基於 SPARC 和 x86 的叢集中，支援乙太網路 (100BASE-T Fast Ethernet 與 1000BASE-SX Gb) 叢集交互連接。SunPlex 系統則僅在基於 SPARC 的叢集中支援 SCI 網路介面叢集交互連接。

- **"電纜"和傳輸"路徑有何不同？"**

  叢集傳輸電纜是使用傳輸配接卡和切換器來配置的。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。由於電纜不是探測式的，所以無法得知它們的狀態。電纜的狀況可以使用 `scconf -p` 來檢視。

  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。路徑的狀態可以是「線上」或「離線」。傳輸路徑的狀態可以使用 `scstat (1M)` 來檢視。

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
```

有兩個可能的傳輸路徑可由這四條電纜形成。

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

用戶端系統 FAQ

- **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？**

  用戶端系統連接至叢集，與連接至任何其他伺服器相同。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*」中的個別章節，以取得有關用戶端配置需求的其他資訊。

管理主控台 FAQ

- **SunPlex 系統是否需要管理主控台？**

  可以。

- **管理主控台是否必須專屬於叢集，還是可用於其他作業？**

  SunPlex 系統不需要專用的管理主控台，但是使用專用主控台可以有以下優點：

  - 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

- **管理主控台位置是否需要「靠近」叢集本身，例如在同一房間中？**

  請洽詢您的硬體服務供應商。供應商可能會要求主控台位置要靠近叢集本身。將主控台置於同一房間中，並無技術上的原因。

- **一部管理主控台在符合所有距離要求的前提下，是否可以服務多個叢集？**

  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。

終端機集線器與系統服務處理器 FAQ

- **SunPlex 系統是否需要終端機集線器？**

  所有以 Sun Cluster 3.0 開始的軟體版次均不需要終端機集線器來執行。與 Sun Cluster 2.2 產品 (需要終端機集線器以實施故障隔離) 不同，以後的產品並不依靠終端機集線器。

- **我發現大部分 SunPlex 伺服器使用終端機集線器，但是 Sun Enterprise E10000 server 不使用。原因為何？**

  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。其主控台是串列埠。Sun Enterprise E10000 server 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 E10000 server，您一定要將 SSP 用於主控台。

- **使用終端機集線器有哪些優勢？**

  使用終端機集線器可爲網路上任意位置的遠端工作站之各節點提供主控台層級的存取，包括當節點在基於 SPARC 的節點上處於 OpenBoot PROM (OBP) 中時，或者在基於 x86 的節點上作爲啓動子系統時。

- **如果我使用的終端機集線器不受 Sun 支援，那麼我需要知道哪些內容，才能使我要使用的終端機集線器合乎標準？**

  Sun 支援的終端機集線器與其他主控台裝置的主要差異，是 Sun 終端機集線器具有特殊的韌體可以防止終端機集線器在開機時送出中斷。請注意，如果您的主控台裝置會送出中斷，或可能會被解釋爲中斷的信號，它將會關閉節點。

- **我是否可以釋放 SUN 所支援的終端機集線器上已鎖定的通訊埠，而不需重新啓動它？**

  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
  #
  ```

  請參考下列手冊，以取得關於如何配置與管理 Sun 所支援之終端機集線器的更多資訊。

  - 「*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.x Hardware Administration Manual for Solaris OS*」中的「Installing and Configuring the Terminal Concentrator」

- **假如終端機集線器本身發生故障，該怎麼辦？必須常備另一台終端機集線器嗎？**

  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。

- **如果我真的使用終端機集線器，它的安全性如何？**

  一般而言，終端機集線器是連接至系統管理員所使用的小型網路，不是連接到其他用戶端存取的網路。您可以藉由限制該特定網路的存取權來控制安全性。

- **SPARC: 我如何藉由磁帶機或磁碟機使用動態重新配置？**
  - 判斷磁碟機或磁帶機是否爲作用中裝置群組的一部分。如果磁碟機不是作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作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。如果磁碟機是作用中裝置群組的一部分，請移至第 72 頁的 「SPARC: 磁碟與磁帶機的 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。

Caution – 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

路

隔

電

對

磁

管

網

整

儲

應

檔

叢

關

驅

邏