Sun Cluster 概念指南（适用于 Solaris OS）

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

文件号码: 817–6384–10
2004 年 4 月，修订版 A

目录

序

《*Sun*™ *Cluster* **概念指南（适用于** *Solaris OS*）》包含有关基于 SPARC™ 和 x86 系统的 SunPlex™ 系统的概念和参考信息。

注意 – 在本文档中，术语“x86”是指 Intel 32 位微处理器芯片系列和 AMD 制造的兼容微处理器芯片系列。

SunPlex 系统包括组成 Sun 群集解决方案的所有硬件和软件组件。

此文档面向接受过 Sun Cluster 软件知识培训并且经验丰富的系统管理员。 不要将此文档作为规划指南或售前指南。 在阅读本文档之前，您应当已经确定了系统要求并购买了相应的设备及软件。

要理解本书中讲述的概念，应该具备 Solaris™ 操作环境的相关知识和有关与 SunPlex 系统一起使用的卷管理器软件的专业经验。

注意 – Sun Cluster 软件运行在 SPARC 和 x86 两种平台上。本文档中的信息适用于两种平台，除非在特定的章、节、说明、标有项目符号的项、图、表或示例中另外说明。

印刷惯例

下表描述了本书中使用的印刷惯例。

表 P–1 印刷惯例

字体或符号	含义	示例
`AaBbCc123`	命令、文件和目录的名称以及计算机屏幕输出	编辑 `.login` 文件。 使用 `ls -a` 列出所有文件。 `machine_name% you have mail.`
`AaBbCc123`	您键入的内容，与计算机屏幕输出的内容相对照	`machine_name%` **`su`** `Password:`
AaBbCc123	命令行占位符：用实际名称或实际值替换	要删除文件，键入 **`rm`** *filename*。
AaBbCc123	书名、新术语或要强调的术语	请参见《**用户指南**》第 6 章。 这些称为**类**选项。 执行此操作者，必须是 *root* **用户**。

命令示例中的 shell 提示符

下表显示了 C shell、Bourne shell 和 Korn shell 的缺省系统提示符和超级用户提示符。

表 P–2 shell 提示符

shell	提示符
C shell 提示符	`machine_name%`
C shell 超级用户提示符	`machine_name#`
Bourne shell 和 Korn shell 提示符	`$`
Bourne shell 和 Korn shell 超级用户提示符	`#`

相关文档

有关相关的 Sun Cluster 主题的信息，可从下表列出的文档中获得。 所有 Sun Cluster 文档均存放在 `http://docs.sun.com` 网页中。

主题	文档
概念	《*Sun Cluster* **概念指南（适用于** *Solaris OS*）》
概述	《*Sun Cluster* **概述（适用于** *Solaris OS*）》
词汇表	*Sun Java Enterprise System 2003Q4 Glossary*
硬件管理	*Sun Cluster Hardware Administration Manual for Solaris OS* 各种版本的硬件管理指南
软件安装	《*Sun Cluster* **软件安装指南（适用于** *Solaris OS*）》
数据服务管理	《*Sun Cluster* **数据服务规划和管理指南（适用于** *Solaris OS*）》 各种版本的数据服务指南
数据服务开发	《*Sun Cluster* **数据服务开发者指南（适用于** *Solaris OS*）》
系统管理	《*Sun Cluster* **系统管理指南（适用于** *Solaris OS*）》
错误消息	*Sun Cluster Error Messages Guide for Solaris OS*
命令和功能参考	*Sun Cluster Reference Manual for Solaris OS*

有关 Sun Cluster 文档的完整列表，请参见位于 `http://docs.sun.com` 站点的 Sun Cluster 软件版本的发行说明。

联机访问 Sun 文档

可以通过 docs.sun.com^{SM} 网站联机访问 Sun 技术文档。 您可以浏览 docs.sun.com 档案或查找某个具体的书名或主题。 URL 是 `http://docs.sun.com`。

订购 Sun 文档

Sun Microsystems 提供一些印刷的产品文档。 有关文档列表以及如何订购它们，请参见 `http://docs.sun.com` 中的“购买印刷的文档”。

获得帮助

如果您在安装或使用 SunPlex 系统时有任何问题，请与您的服务供应商联系并提供以下信息：

- 您的姓名和电子邮件地址（如果有）
- 您的公司名称、地址和电话号码
- 系统的型号和序列号
- 操作环境的发行版本号（例如，Solaris 9）
- Sun Cluster 软件的发行版本号（例如，3.1 4/04）

使用以下命令获得有关系统中每个节点的信息以提供给您的服务供应商。

命令	功能
`prtconf -v`	显示系统内存的大小并报告有关外围设备的信息
`psrinfo -v`	显示有关处理器的信息
`showrev -p`	报告已安装了哪些修补程序
SPARC: `prtdiag -v`	显示系统诊断信息
`scinstall -pv`	显示 Sun Cluster 软件发行版本和软件包版本信息
`scstat`	提供群集状况的快照
`scconf -p`	列出群集配置信息
`scrgadm -p`	显示有关安装的资源、资源组和资源类型的信息

还请提供 `/var/adm/messages` 文件的内容。

第 1 章

简介与概述

SunPlex 系统是集成的硬件和 Sun Cluster 软件解决方案，用于创建高可用性和可伸缩的服务。

《Sun Cluster 概念指南（适用于 Solaris OS）》为 SunPlex 文档的主要读者提供了所需的概念信息。 这些读者包括

- 安装和维护群集硬件的服务供应商
- 安装、配置和管理 Sun Cluster 软件的系统管理员
- 为 Sun Cluster 产品当前所未包括的应用程序开发失效转移和可伸缩服务的应用程序开发者

本书和 SunPlex 文档集中的其它文档一起对 SunPlex 系统进行了全面的介绍。

本章

- 介绍 SunPlex 系统并作了简要的概述
- 介绍 SunPlex 读者的几种观点
- 明确在使用 SunPlex 系统之前需要理解的一些关键概念
- 将关键概念与包括过程和相关信息的 SunPlex 文档对应起来
- 将群集相关的任务与包含完成这些任务所遵照的步骤的文档对应起来

SunPlex 系统介绍

SunPlex 系统使 Solaris 操作环境扩展为群集操作系统。 群集（丛）是一种松散耦合的计算节点集合，提供网络服务或应用程序（包括数据库、Web 服务和文件服务）的单一客户视图。

每个群集节点都是运行其自己的进程的一个独立服务器。 这些进程彼此进行通信，对网络客户机来说就像是形成了一个单一系统，协同起来向用户提供应用程序、系统资源和数据。

与传统的单一服务器系统相比，群集有几个优点。 这些优点包括对失效转移和可伸缩服务的支持、适应模块化增长的容量，以及优越于传统硬件容错系统的低进入价。

SunPlex 系统旨在：

- 减少或消除由软件或硬件故障引起的系统停机时间
- 确保数据和应用程序对最终用户的可用性，而不管通常引起单服务器系统停机的故障属于什么类型
- 通过向群集添加节点，使服务随着处理器的添加而伸缩，从而增大应用程序吞吐量
- 提供增强的系统可用性，使您能够不必关掉整个群集就可执行维护

有关容错和高可用性的详细信息，请参阅《*Sun Cluster* **概述（适用于** *Solaris OS*）》中的“Sun Cluster 使应用程序获得高可用性”。

有关高可用性的问题及解答，请参阅第 67 页 “高可用性 FAQ”。

SunPlex 系统的三种观点

本部分说明关于 SunPlex 系统的三种不同观点和与每种观点相关的关键概念和文档。 这些观点来自：

- 硬件安装和维护人员
- 系统管理员
- 应用程序编程人员

硬件安装和维护观点

对于硬件维护的专业人员来说，SunPlex 系统看起来就像是一个包括服务器、网络和存储器在内的现成的硬件集合。 这些部件用电缆连接起来，使每个部件都有一个备份，因而不存在单点故障。

关键概念 – 硬件

硬件维护人员需要理解下面的群集概念。

- 群集硬件配置和电缆连接
- 安装与维护（添加、拆卸与更换）：
  - 网络接口组件（适配器、结点、电缆）
  - 磁盘接口卡
  - 磁盘阵列

    - 磁盘驱动器
    - 管理控制台和控制台访问设备
- 设置管理控制台和控制台访问设备

硬件概念参考建议

下面几节包含与前面的关键概念相关的材料：

- 第 16 页 “群集节点”
- 第 18 页 “多主机磁盘”
- 第 19 页 “本地磁盘”
- 第 19 页 “群集互连”
- 第 20 页 “公共网络接口”
- 第 20 页 “客户机系统”
- 第 21 页 “管理控制台”
- 第 20 页 “控制台访问设备”
- 第 22 页 “SPARC: 群集对拓扑”
- 第 23 页 “SPARC: N+1（星型）拓扑”

相关的 SunPlex 文档

下面的 SunPlex 文档包含与硬件服务概念相关的过程和信息：

- *Sun Cluster Hardware Collection*

系统管理员观点

对于系统管理员来说，SunPlex 系统看起来就像用电缆连接起来共享存储设备的一个服务器（节点）集合。 系统管理员将看到：

- 专用的群集软件与 Solaris 软件集成在一起来监视群集节点之间的连通性
- 专用的软件监视用户应用软件程序在群集节点上的运行状况
- 卷管理软件设置和管理磁盘
- 专用的群集软件使所有的节点可以访问所有的存储设备，甚至包括那些并未直接连接到磁盘的设备
- 专用的群集软件使文件可以显示在每个节点上，就好像已将这些文件本地连接到该节点上

关键概念 – 系统管理

系统管理员需要理解下面的概念和进程：

- 硬件和软件组件之间的相互作用
- 安装和配置群集的一般流程包括：

- 安装 Solaris 操作环境
- 安装和配置 Sun Cluster 软件
- 安装和配置卷管理器
- 安装和配置应用程序软件，使其为群集做好准备
- 安装和配置 Sun Cluster 数据服务软件

- 添加、拆除、更换及维护群集硬件和软件组件的群集管理过程
- 修改配置以提高性能

系统管理员概念参考建议

下面几节包含与前面的关键概念相关的材料：

- 第 27 页 “管理界面”
- 第 28 页 “高可用性框架”
- 第 30 页 “全局设备”
- 第 31 页 “磁盘设备组”
- 第 33 页 “全局名称空间”
- 第 34 页 “群集文件系统”
- 第 39 页 “仲裁和仲裁设备”
- 第 43 页 “数据服务”
- 第 52 页 “资源、资源组和资源类型”
- 第 62 页 “公共网络适配器和 IP Network Multipathing ”
- 第 4 章

相关的 SunPlex 文档 – 系统管理员

下面的 SunPlex 文档包含与系统管理概念相关的过程和信息：

- 《*Sun Cluster* **软件安装指南**》
- 《*Sun Cluster* **系统管理指南**》
- *Sun Cluster Error Messages Guide*
- *Sun Cluster Release Notes*
- *Sun Cluster Release Notes Supplement*

应用程序编程人员观点

SunPlex 系统为诸如以下的应用程序提供了**数据服务**，例如 Oracle（基于 SPARC 的系统）、NFS、DNS、Sun™ Java System Web Server（以前的 Sun Java System Web Server）、Apache Web Server（基于 SPARC 的系统）和 Sun Java System Directory Server（以前的 Sun Java System Directory Server）。 数据服务是通过配置现成的应用程序，使之在 Sun Cluster 软件的控制下运行来创建的。 Sun Cluster 软件提供了用于启动、停止和监视该应用程序的配置文件和管理方法。 如果需要创建新的失效转移或可伸缩服务，可以使用 SunPlex 应用程序设计接口 (API) 和数据服务启用技术 API (DSET API) 来开发所需的配置文件和管理方法（它们可以启用相应的应用程序，使其作为数据服务在群集上运行）。

关键概念 – 应用程序编程人员

应用程序编程人员需要理解下面的内容：

- 应用程序的特性，由此确定其能否作为失效转移服务或可伸缩数据服务来运行。
- Sun Cluster API、DSET API 和“普通”数据服务。 编程人员需要确定哪个工具最适合用来编写程序或脚本，以便配置应用程序，使之适合于在群集环境下运行。

应用程序编程人员概念参考建议

下面几节包含与前面的关键概念相关的材料：

- 第 43 页 “数据服务”
- 第 52 页 “资源、资源组和资源类型”
- 第 4 章

相关的 SunPlex 文档 – 应用程序编程人员

下面的 SunPlex 文档包含与应用程序编程人员概念相关的过程和信息：

- 《*Sun Cluster* **数据服务开发者指南**》
- 《*Sun Cluster* **数据服务规划和管理指南**》

SunPlex 系统任务

所有的 SunPlex 系统任务都需要某些概念背景。 下表介绍了任务的高层视图以及描述任务步骤的文档。 本书中的概念部分讲述概念与这些任务的对应关系。

表 1–1 任务表：将用户任务与文档对应起来

要完成的任务	需要使用的文档
安装群集硬件	*Sun Cluster Hardware Collection*
在群集上安装 Solaris 软件	《*Sun Cluster* **软件安装指南**》
SPARC: 安装 Sun™ Management Center 软件	《*Sun Cluster* **软件安装指南**》
安装和配置 Sun Cluster 软件	《*Sun Cluster* **软件安装指南**》
安装并配置卷管理软件	《*Sun Cluster* **软件安装指南**》 您的卷管理文档

表 1–1 任务表：将用户任务与文档对应起来　（续）

要完成的任务	需要使用的文档
安装和配置 Sun Cluster 数据服务	《*Sun Cluster* **数据服务规划和管理指南**》
维护群集硬件	*Sun Cluster Hardware Collection*
管理 Sun Cluster 软件	《*Sun Cluster* **系统管理指南**》
管理卷管理软件	《*Sun Cluster* **系统管理指南**》和您的卷管理文档
管理应用程序软件	您的应用程序文档
问题鉴定与建议的用户操作	*Sun Cluster Error Messages Guide*
创建新的数据服务	《*Sun Cluster* **数据服务开发者指南**》

第 2 章

关键概念 – 硬件服务供应商

本章说明与 SunPlex 系统配置中的硬件组件相关的关键概念。 包括下列主题：

SunPlex 系统硬件和软件组件

本章中的信息主要面向硬件服务供应商。 在服务供应商安装、配置或维护群集硬件之前，这些概念可帮助他们理解各硬件部件之间的关系。 群集系统管理员可能也会发现这些信息很有用，它们可用作安装、配置和管理群集软件的背景信息。

群集由下列硬件部件组成：

- 具有本地磁盘的群集节点（不共享）
- 多主机存储器（节点间的共享磁盘）
- 可拆卸介质（磁带和 CD-ROM）
- 群集互连
- 公共网络接口
- 群集系统
- 管理控制台
- 控制台访问设备

SunPlex 系统使您能将这些组件组合成多种配置，如第 21 页 “SPARC: Sun Cluster 拓扑示例”中所述。

有关双节点群集配置样例的说明，请参阅《*Sun Cluster* **概述（适用于** *Solaris OS*）》中的“硬件环境”。

群集节点

群集节点是同时运行 Solaris 操作环境和 Sun Cluster 软件的机器，它或者是群集的当前成员（**群集成员**），或者是潜在成员。

SPARC: Sun Cluster 软件使您可在一个群集中部署两到八个节点。 有关支持的节点配置，请参阅第 21 页 “SPARC: Sun Cluster 拓扑示例”。

x86: Sun Cluster 软件使您可在一个群集中部署两个节点。 有关支持的节点配置，请参阅第 25 页 “x86: Sun Cluster 拓扑示例”。

群集节点一般连接着一个或多个多主机磁盘。 未连接到多主机磁盘的节点使用群集文件系统来访问多主机磁盘。 例如，可伸缩服务配置允许节点为请求提供服务，而节点不必直接连接到多主机磁盘。

此外，并行数据库配置中的节点共享对所有磁盘的并行访问。 有关并行数据库配置的详细信息，请参见第 18 页 “多主机磁盘”和第 3 章。

群集中的所有节点都会归组到一个通用名称（即群集名称）下，将通过该名称来对群集进行访问和管理。

公共网络适配器将节点连接到公共网络，为客户机提供对群集的访问。

群集成员通过一个或多个物理上独立的网络与群集上的其它节点进行通信。 这组物理上独立的网络称作**群集互连**。

群集中的每一节点都会知道另一节点的加入或离开。 另外，群集中的每一节点还都会意识到本地运行的资源和在其它群集节点上运行的资源。

同一群集中的节点应具备相似的处理能力、内存和 I/O 容量，以便能够在性能不显著下降的情况下实现失效转移。 因为存在失效转移的可能性，所以每个节点都必须具有足够的额外能力，能够承担它们所备份或辅助的所有节点的工作量。

各个节点引导自己的根 (/) 文件系统。

群集硬件成员的软件组件

要起到群集成员的作用，必须安装下列软件：

- Solaris 操作环境
- Sun Cluster 软件

- 数据服务应用程序
- 卷管理（Solaris Volume Manager ™或 VERITAS Volume Manager）

  使用硬件独立磁盘冗余阵列 (RAID) 的配置是一个例外。此配置可能不需要软件卷管理器，如 Solaris Volume Manager或 VERITAS Volume Manager。

有关如何安装 Solaris 操作环境、Sun Cluster 和卷管理软件的信息，请参阅《*Sun Cluster* **软件安装指南**》。

有关如何安装和配置数据服务的信息，请参阅《*Sun Cluster* **数据服务规划和管理指南**》。

有关上述软件组件的概念信息，请参见第 3 章。

下图展示了用于共同创建 Sun Cluster 软件环境的软件组件的高层次视图。

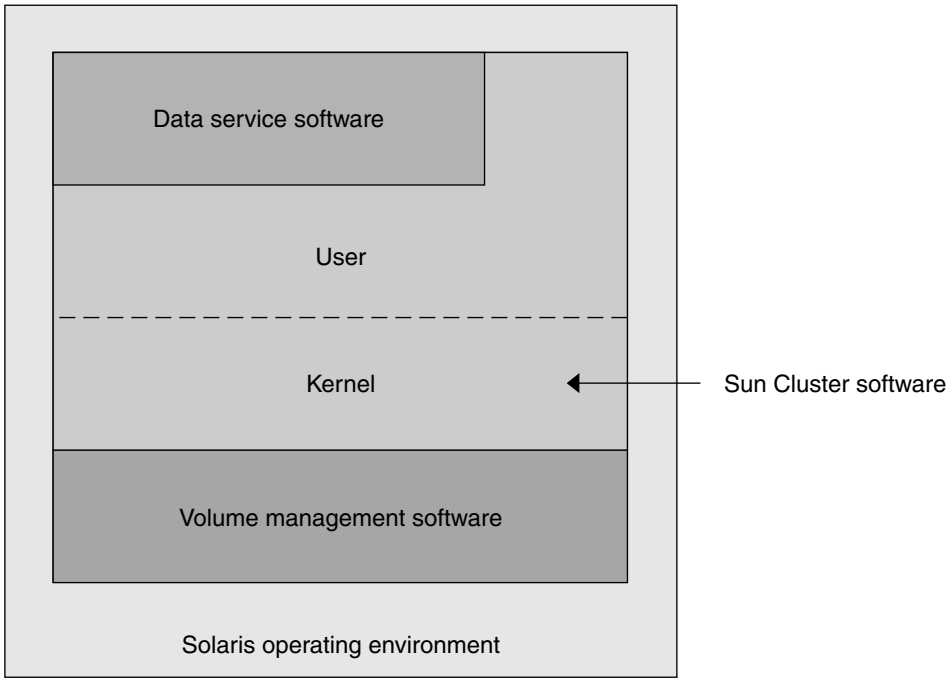


图 2–1 Sun Cluster 软件组件的高层关系

有关群集成员的问题及解答，请参见第 4 章。

多主机磁盘

同时可以连接到多个节点的磁盘就是多主机磁盘。 在 Sun Cluster 环境中，多主机存储使磁盘具有高可用性。 Sun Cluster 需要双节点群集的多主机存储来建立仲裁。 大于三个节点的群集不需要多主机存储。

多主机磁盘具有以下特点。

- 它们能够容许单个节点出现故障。
- 它们存储应用程序数据，并能存储应用程序二进制文件和配置文件。
- 它们可在节点出现故障时提供保护措施。 如果客户机请求通过一个节点访问数据，但该节点出现故障，则请求会切换至另一节点，该节点与客户机所要访问的那些磁盘之间也存在直接的连接。
- 对多主机磁盘的访问要么是通过“主控”磁盘的主节点进行的全局访问，要么是通过局部路径进行的直接并行访问。 当前唯一使用直接并行访问的应用程序是 Oracle Parallel Server。

卷管理器为镜像的或 RAID-5 配置提供多主机磁盘数据冗余。 目前，Sun Cluster 支持 Solaris Volume Manager™ 和 VERITAS Volume Manager，它作为卷管理器只能在基于 SPARC 的群集中使用，而作为 RDAC RAID-5 硬件控制器则可以在若干硬件 RAID 平台上使用。

通过将多主机磁盘与磁盘镜像和磁盘条带化结合使用，既可防止节点故障，又可防止单个磁盘故障。

有关多主机存储的问题及解答，请参见第 4 章。

多启动器 SCSI

本节中的内容只适于用作多主机磁盘的 SCSI 存储设备，而不适于光纤通道存储器。

在独立服务器中，服务器节点通过将该服务器与特定的 SCSI 总线相连的 SCSI 主机适配器电路来控制 SCSI 总线的活动。 此 SCSI 主机适配器电路称作 *SCSI* **启动器**。 该电路启动了此 SCSI 总线的所有活动。 在 Sun 系统中，SCSI 主机适配器的缺省 SCSI 地址是 7。

通过使用多主机磁盘，群集配置共享多个服务器节点间的存储器。 当群集存储器由单端或差分 SCSI 设备组成时，这样的配置称作多启动器 SCSI。 正如此术语的字面含义那样，SCSI 总线上存在多个 SCSI 启动器。

SCSI 规格要求 SCSI 总线上的每个设备都具有唯一的 SCSI 地址。（主机适配器也是 SCSI 总线上的一个设备。） 由于所有 SCSI 主机适配器均缺省设置为 7，因此多启动器环境中的缺省硬件配置会引起冲突。

要解决这一冲突，请在每个 SCSI 总线上将一个 SCSI 主机适配器的 SCSI 地址保留为 7，将其它主机适配器设置到未使用的 SCSI 地址。 正确的规划要求这些“未使用”的 SCSI 地址中包括当前未使用的 SCSI 地址，又包括以后也不会使用的 SCSI 地址。 将来也不会使用地址示例如下：通过在空驱动器插槽中安装新驱动器来增加存储器。

在大多数配置中，第二个主机适配器的可用 SCSI 地址是 6。

您可以使用以下工具之一设置 `scsi-initiator-id` 特性，来为这些主机适配器更改选定的 SCSI 地址：

- `eeprom`(1M)
- 基于 SPARC 系统上的 OpenBoot PROM
- 在基于 x86 的系统上 BIOS 引导之后，您选择运行的 SCSI 公用程序

可对某个节点就此特性进行全局设置，或对每个主机适配器逐个进行设置。 在 *Sun Cluster Hardware Collection* 中，可在每一磁盘群组所对应的章中找到有关为各 SCSI 主机适配器设置唯一 `scsi-initiator-id` 的说明。

本地磁盘

本地磁盘是仅连接到单个节点的磁盘。 因此它们无法防止节点故障（不具备高可用性）。 不过，包括本地磁盘在内的所有磁盘都包含在全局命名空间中，并且配置为**全局设备**。 因此，从所有群集节点都可看到这些磁盘。

通过将本地磁盘上的文件系统放在全局装载点下，可以使其它节点也能使用它们。 如果当前装载了这些全局文件系统之一的节点出现故障，所有节点都将无法访问该文件系统。 可使用卷管理器来对这些磁盘进行镜像，这样磁盘故障就不会导致这些文件系统变得不可访问，但是卷管理器不能防止节点故障的发生。

有关全局设备的详细信息，请参见第 30 页 “全局设备”部分。

可拆卸介质

群集中支持诸如磁带机和 CD-ROM 驱动器的可拆卸介质。 通常，这些设备的安装、配置和维修方式与在非群集环境中相同。 这些设备在 Sun Cluster 中配置为全局设备，因此从群集中的任何节点都可访问每一台设备。 有关安装和配置可拆卸介质的详细信息，请参阅 *Sun Cluster Hardware Collection*。

有关全局设备的详细信息，请参见第 30 页 “全局设备”部分。

群集互连

群集互连是设备的物理配置，可以使用这些设备在群集节点之间传送群集专用通信和数据服务通信。 因为群集专用通信中大量使用群集互连，所以会限制性能。

只有群集节点可以连接到群集互连。 Sun Cluster 安全模型假定只有群集节点具有对群集互连的物理访问权。

必须至少通过两个物理上独立的冗余网络或路径，将所有的节点通过群集互连连接起来，以避免出现单故障点。 您可以在任意两个节点间部署若干物理上独立的网络（二至六个）。 该群集互连由三个硬件组件构成： 适配器、结点和电缆。

下面的列表对这些硬件组件逐一进行说明。

- 适配器 – 驻留在每个群集节点中的网络接口卡。 适配器的名称由设备名称和紧跟其后的物理单元编号组成，例如 qfe2。某些适配器只具有一个物理网络连接，但其它适配器（如 qfe 卡）则具有多个物理连接。 某些适配器还同时包含网络接口和存储器接口。

  具有多个接口的网络适配器在整个适配器出现故障时会成为单故障点。 为了获得最高的可用性，请在规划群集时确保两个节点间的唯一路径不会依赖一个网络适配器。
- 结点 – 驻留在群集节点外部的转接器。 它们实现通路和切换功能，使您可将两个以上的节点连接到一起。 双节点群集中不需结点，因为两个节点可通过连接到各自冗余适配器上的冗余物理电缆直接连接。 超过两个节点的群集配置通常需要结点。
- 电缆 – 两个网络适配器之间或适配器和结点之间的物理连接。

有关群集互连的问题及解答，请参见第 4 章。

公共网络接口

客户机通过公共网络接口与群集相连。 每个网络适配卡可连接一个或多个公共网络，这取决于卡上是否具有多个硬件接口。 可以将节点设置为包含配置的多个公共网络接口卡，从而激活多个接口卡，并作为相互之间的失效转移备份。 如果其中一个适配器发生故障，则会调用 IP Network Multipathing 软件，以便进行失效转移，从故障接口切换到组中的另一个适配器。

进行群集化时，不用为公共网络接口考虑任何特殊的硬件。

有关公共网络的问题及解答，请参见第 4 章。

客户机系统

客户机系统包括通过公共网络访问群集的工作站或其它服务器。 客户端程序使用群集中运行的服务器端应用程序提供的数据或其它服务。

客户机系统不具备高可用性。 群集中的数据和应用程序具备高可用性。

有关客户机系统的问题及解答，请参见第 4 章。

控制台访问设备

您必须能对所有群集节点进行控制台访问。 要获得控制台访问权，请使用随群集硬件一起购买的终端集中器、Sun Enterprise E10000™ 服务器（用于基于 SPARC 的群集）上的系统服务处理器 (SSP)、Sun Fire™ 服务器（也用于基于 SPARC 的群集）上的系统控制器或其它可以访问每个节点上的 `ttya` 的设备。

Sun 只提供了一个支持的终端集中器作为选件使用。 该终端集中器通过使用 TCP/IP 网络来访问各个节点上的 `/dev/console`。 这样就可从网络上的任一远程工作站对每一节点进行控制台级别的访问。

系统服务处理器 (SSP) 提供了对 Sun Enterprise E10000 server 的控制台访问。 SSP 是以太网上经配置可支持 Sun Enterprise E10000 server 的机器。 SSP 是 Sun Enterprise E10000 server 的管理控制台。 使用 Sun Enterprise E10000 Network Console 功能，网络上的任何工作站都可打开主机控制台会话。

其它控制台访问方法包括其它终端集中器、从另一节点进行的 `tip(1)` 串行端口访问和哑终端。 可以使用 Sun™ 键盘和监视器或其它串行端口设备（如果硬件服务供应商支持这些设备）。

管理控制台

您可以使用专用 UltraSPARC® 工作站或 Sun Fire™ V65x 服务器（称为**管理控制台**）来管理活动群集。 通常在管理控制台上安装并运行的管理工具软件有群集控制面板 (CCP) 和用于 Sun Management Center™ 产品（仅限与基于 SPARC 的群集一起使用）的 Sun Cluster 模块。 使用 CCP 下的 `cconsole` 可使您能同时连接到多个节点控制台。 有关使用 CCP 的详细信息，请参阅《*Sun Cluster* **系统管理指南**》。

管理控制台并不是一个群集节点。 您可以使用管理控制台通过公共网络或选择性地通过基于网络的终端集中器来远程访问群集节点。 如果群集由 Sun Enterprise E10000 平台组成，则必须能够从管理控制台登录到系统服务处理器 (SSP)，并能使用 `netcon(1M)` 命令进行连接。

配置节点时通常不配置监视器。 然后，您可以通过 `telnet` 会话从管理控制台访问节点的控制台。管理控制台连接到终端集中器，终端集中器又连接到节点的串行端口。 （如果使用 Sun Enterprise E10000 server，则从系统服务处理器进行连接。） 有关详细信息，请参阅第 20 页 “控制台访问设备”。

Sun Cluster 不要求专用的管理控制台，但如果使用，则具有以下好处：

- 通过在同一机器上给控制台和管理工具分组来启用集中化的群集管理
- 可能会使硬件服务供应商更快地解决问题

有关管理控制台的问题及解答，请参见第 4 章。

SPARC: Sun Cluster 拓扑示例

拓扑是群集节点与群集中所用存储平台的连接方案。 Sun Cluster 支持遵循以下原则的任何拓扑。

- 不管您实现的存储配置是什么样的，由基于 SPARC 的系统组成的 Sun Cluster 在群集中均最多支持八个节点。
- 共享的存储设备可以连接的节点数与存储设备支持的节点数一样多。

- 共享的存储设备不需要连接群集中的所有节点。 但是，这些存储设备必须至少连接两个节点。

Sun Cluster 不要求使用特定的拓扑配置群集。 以下拓扑词汇用来说明群集的连接方案。 这些拓扑都是典型的连接方案。

- 群集对
- Pair+N
- N+1（星型）
- N*N （可伸缩）

以下部分包括每种拓扑的样例图。

SPARC: 群集对拓扑

群集对拓扑是在单一群集管理框架下运行的两对或更多对节点。 在此配置中，只会在一对节点间进行失效转移。 但是，所有节点都通过群集互连连接在一起，并且在 Sun Cluster 软件控制下运行。 您可以使用此拓扑在一对节点上运行并行数据库应用程序，在另一对节点上运行失效转移或可伸缩应用程序。

通过使用群集文件系统，还可部署两对节点的配置，在此配置中，即使所有节点都未直接连接到存储应用程序数据的磁盘，两个以上的节点仍可运行可伸缩服务或并行数据库。

下图所示为群集对配置。

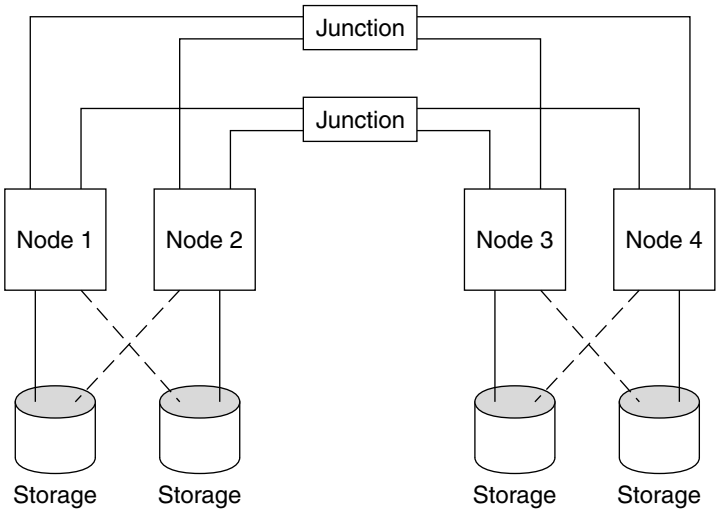


图 2–2 SPARC: 群集对拓扑

SPARC: Pair+N 拓扑

Pair+N 拓扑包括一对直接连接到共享存储器的节点和一组附加的使用群集互连来访问共享存储器的节点（这组节点内部并未直接相连）。

下图展示了一个 Pair+N 拓扑，其四个节点中有两个（节点 3 和 4）使用群集互连来访问该存储器。 可以扩展此配置，以包含那些对共享存储器没有直接访问权的节点。

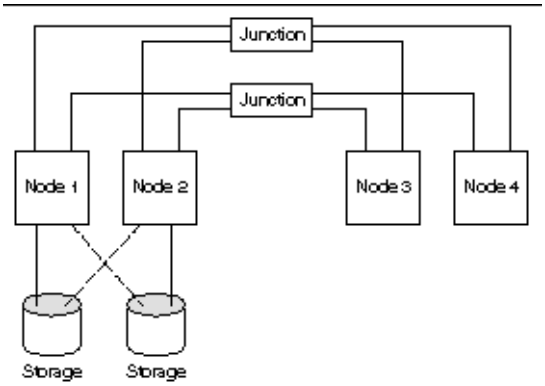


图 2–3 SPARC: Pair+N 拓扑

SPARC: N+1（星型）拓扑

N+1 拓扑包括几个主节点和一个辅助节点。 主节点和辅助节点的配置不必完全相同。 一般由主节点提供应用程序服务。 在主节点出现故障之前，辅助节点不必处于空闲状态。

辅助节点是配置中与所有多主机存储器有物理连接的唯一节点。

如果主节点出现故障，Sun Cluster 则会进行失效转移，将资源切换至辅助节点，这些资源将在辅助节点继续工作，直到切换回（自动或手动）主节点。

如果一个主节点出现故障，辅助节点必须具备足够的 CPU 能力处理负载。

下图所示为 N+1 配置。

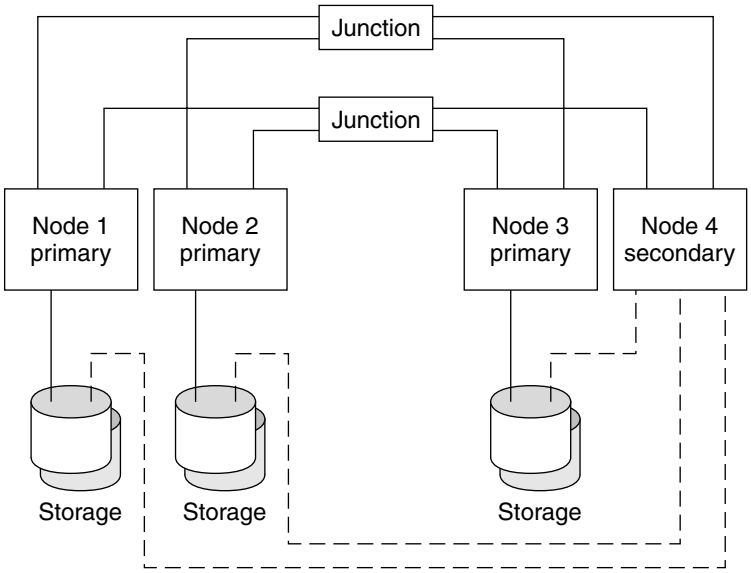


图 2–4 SPARC: N+1 拓扑

SPARC: N*N（可伸缩）拓扑

N*N 拓扑允许群集中的每个共享存储设备连接到群集中的任意节点。 此拓扑允许高可用应用程序进行失效转移，在不降低服务质量的情况下，从一个节点切换到另一个节点。 当发生失效转移时，新节点可以通过本地路径（而不是专用互连）来访问存储设备。

下图所示为 N*N 配置。

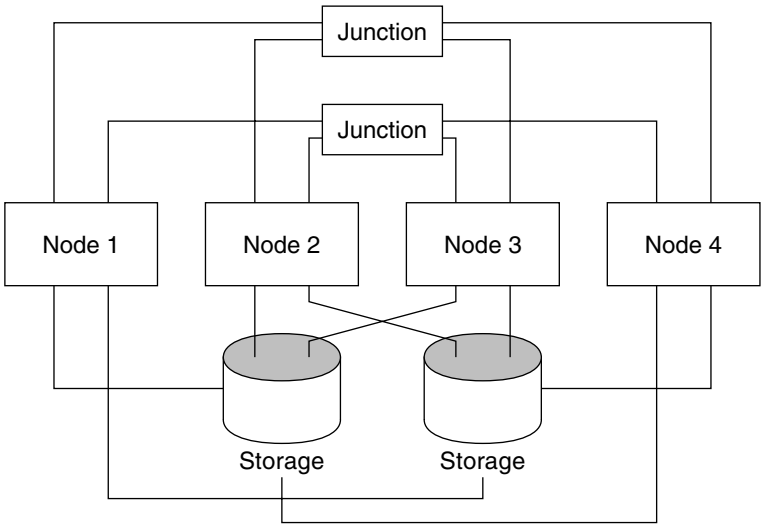


图 2–5 SPARC: N*N 拓扑

x86: Sun Cluster 拓扑示例

拓扑是群集节点与群集中所使用的存储平台的连接方案。 Sun Cluster 支持遵循以下原则的任何拓扑。

- 由基于 x86 的系统组成的 Sun Cluster 在群集中支持两个节点。
- 共享存储设备必须连接到这两个节点上。

Sun Cluster 不要求使用特定的拓扑配置群集。 对以下群集对拓扑（仅可用于由基于 x86 的节点组成的群集）进行了说明以提供词汇表来讨论群集的连接方案。 此拓扑是典型的连接方案。

以下部分包括拓扑的样例图。

x86: 群集对拓扑

群集对拓扑是在单一群集管理框架下运行的两个节点。 在此配置中，只会在一对节点间进行失效转移。 但是，所有节点都通过群集互连连接在一起，并且在 Sun Cluster 软件控制下运行。 您可以使用此拓扑在一对节点上运行并行数据库或者运行失效转移或可伸缩应用程序。

下图所示为群集对配置。

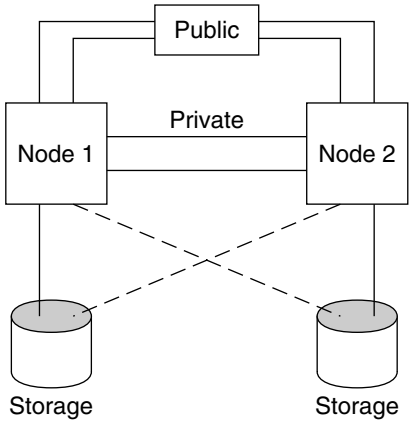


图 2–6 x86: 群集对拓扑

第 3 章

关键概念 – 管理和应用程序开发

本章说明与 SunPlex 系统的软件组件相关的关键概念。 包括下列主题：

群集管理和应用程序开发

此信息主要面向使用 SunPlex API 和 SDK 的系统管理员和应用程序开发者。 群集系统管理员可以将此信息作为安装、配置和管理群集软件的背景知识。 应用程序开发者可以通过此信息来了解他们工作的群集环境。

管理界面

可以选择如何通过若干用户界面来安装、配置和管理 SunPlex 系统。 通过 SunPlex Manager 图形用户界面 (GUI) 或文档化的命令行界面都可以完成系统管理任务。 在命令行界面的顶部是若干实用程序（如 `scinstall` 和 `scsetup`），可以简化选定的安装

和配置任务。 SunPlex 系统还有一个模块，作为 Sun Management Center（可向特定群集任务提供 GUI）的一部分来运行。 此模块只能在基于 SPARC 的群集中使用。 有关管理接口的完整说明，请参阅《*Sun Cluster* **系统管理指南**》中包括介绍性内容的一章。

群集时间

群集中的所有节点之间的时间必须同步。 对于群集操作来说，群集节点与任何外部时间源是否同步并不重要。 SunPlex 系统使用网络时间协议 (NTP) 在节点间保持时钟同步。

通常，系统时钟一秒种的时间改变不会造成问题。 然而，如果要对活动的群集运行 `date(1)`、`rdate(1M)` 或 `xntpdate(1M)`（以交互方式或在 `cron` 脚本内）命令，就可对系统时间强制进行远大于一秒种的更改，以使系统时钟与时间源同步。 这一强制更改会给文件修改时间戳记带来问题，或造成 NTP 服务混乱。

在每个群集节点上安装 Solaris 操作环境时，您有机会更改节点的缺省时间和日期设置。一般情况下，可以接受出厂缺省设置。

使用 `scinstall(1M)` 安装 Sun Cluster 软件时，装载过程中有一步是为群集配置 NTP。 Sun Cluster 软件提供了一个模板文件 `ntp.cluster`（请参阅已安装的群集节点上的 `/etc/inet/ntp.cluster`），可以使用一个节点作为“首选”节点，在所有群集节点之间建立对等关系。 节点由它们的专用主机名标识，并且在群集互连中实现时间同步。 《*Sun Cluster* **软件安装指南**》中对如何为群集配置 NTP 进行了说明。

另外，您可以在群集外部设置一个或多个 NTP 服务器，并更改 `ntp.conf` 文件来反映此配置。

正常运行时，绝不需要调整群集的时间。 但是，如果安装 Solaris 操作环境时时间设置得不正确，而现在想进行更改，则可参阅《*Sun Cluster* **系统管理指南**》中的相关步骤进行操作。

高可用性框架

SunPlex 系统使用户和数据间的“路径”上的所有组件都具有高度的可用性，这些组件包括网络接口、应用程序本身、文件系统和多主机磁盘。 一般情况下，如果一个群集组件可从系统中的任何单一（软件或硬件）故障中恢复，则它是高度可用的。

下表显示了各种 SunPlex 组件故障（包括硬件故障和软件故障），以及高可用性框架中内置的各种恢复。

表 3–1 SunPlex 故障检测和恢复级别

有故障的群集组件	软件恢复	硬件恢复
数据服务	HA API，HA 框架	不适用

表 3–1 SunPlex 故障检测和恢复级别　（续）

有故障的群集组件	软件恢复	硬件恢复
公共网络适配器	IP Network Multipathing	多个公共网络适配卡
群集文件系统	主要和辅助复制	多主机磁盘
镜像多主机磁盘	卷管理（Solaris Volume Manager 和 VERITAS Volume Manager，只能在基于 SPARC 的群集中使用）	硬件 RAID-5（例如 Sun StorEdge™ A3x00）
全局设备	主要和辅助复制	到设备、群集传输结点的多个路径
专用网	HA 传输软件	多个专用硬件独立网络
节点	CMM，故障快速防护驱动程序	多个节点

Sun Cluster 软件的高可用性框架可迅速检测到节点故障，并在群集中的其余节点上为该框架资源创建一个新的等效服务器。 不会出现所有框架资源都不可用的情况。 在恢复期间，未受崩溃节点影响的框架资源是完全可用的。 而且，故障节点的框架资源一恢复就立即可用。 已恢复的框架资源不必等待其它所有框架资源都完全恢复。

大多数高度可用的框架资源都透明恢复到使用该资源的应用程序（数据服务）。 框架资源访问的语义在整个节点故障期间得到全面的保护。 应用程序根本不知道框架资源已转移到另一节点。 单个节点的故障对使用连接到此节点的文件、设备和磁盘卷的其它节点上的程序来说，是完全透明的。 多主机磁盘的使用就是一个例证，这些磁盘具有连接多个节点的端口。

群集成员监视器

为确保数据免遭破坏，所有节点必须在群集成员上达成一致协议。 需要时，CMM 将协调群集服务（应用程序）的群集重新配置，以作为对故障的响应。

CMM 会从群集传输层接收到关于与其它节点连通性的信息。 CMM 使用群集互连在重新配置期间交换状态信息。

检测到群集成员有更改后，CMM 执行群集的同步配置，这时群集资源可能会按群集的新的成员关系被重新分配。

与 Sun Cluster 软件以前的发行版不同，CMM 是完全在内核中运行的。

有关群集如何保护自身不被划分为多个独立群集的详细信息，请参阅第 39 页 “仲裁和仲裁设备”。

故障快速防护机制

如果 CMM 检测到某个节点发生了严重问题，它将调用群集框架来强制关闭（应急）该节点并从群集成员中删除该节点。 实现这种功能的机制称为**故障快速防护**。 故障快速防护会使节点以两种方式关闭。

- 如果节点脱离群集，然后尝试启动新的群集，而不进行仲裁，则会被“隔离”，不能访问共享的磁盘。 有关使用故障快速防护的详细信息，请参见第 42 页 “故障防护”。
- 如果一个或多个特定于群集的守护程序出现故障（`clexecd`、`rpc.pmfd`、`rgmd` 或 `rpc.ed`），CMM 会检测到该故障，而节点将处于应急状态。 群集守护程序出现故障造成节点进入应急状态，一条类似于以下内容的消息将显示在该节点的控制台上。

```
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
```

出现应急状态之后，节点可能重新引导并尝试重新加入群集；或者，如果群集是由基于 SPARC 的系统组成的，则停留在 OpenBoot™ PROM (OBP) 提示符处。 所采取的操作取决于 `auto-boot?` 参数的设置。 可以在 OpenBoot PROM `ok` 提示符处使用`eeprom`(1M) 设置 `auto-boot?`。

群集配置系统信息库 (CCR)

CCR 使用两阶段提交算法进行更新： 更新必须在所有群集成员上均成功完成，否则更新将被转返。 CCR 使用群集互连来应用分布式更新。

Caution – 尽管 CCR 是由文本文件组成的，但也绝不要手动编辑 CCR 文件。 每个文件都包含一个校验和记录来保证节点间的一致性。 手动更新 CCR 文件可能会导致某个节点或整个群集不能工作。

CCR 靠 CMM 来保证群集只有在仲裁建立后才能运行。 CCR 负责跨群集验证数据的一致性，需要时执行恢复，并为数据更新提供工具。

全局设备

SunPlex 系统使用**全局设备**实现群集范围内的高可用性访问，使您可以从任一节点对群集中的任一设备进行访问，而不用考虑设备的实际连接位置。 在通常情况下，如果某个节点在提供对全局设备的访问时出现故障，则 Sun Cluster 软件会自动找到该设备的其它路径并将访问重定向到该路径。 SunPlex 全局设备包括磁盘、CD-ROM 和磁带。 不过，磁盘是唯一支持的多端口全局设备。 这意味着 CD-ROM 和磁带设置目前还不是高可用性的设备。 每个服务器上的本地磁盘也不是多端口的，因而也不是高可用性设备。

群集自动为群集中的每个磁盘、CD-ROM 和磁带设备分配唯一的 ID。 这种分配使得从群集中任何节点到每个设备的访问都保持一致性。 全局设备名称空间保存在 `/dev/global` 目录下。 有关详细信息，请参阅第 33 页 “全局名称空间”。

多端口全局设备可为一个设备提供多个路径。 至于多主机磁盘，因为这些磁盘是以一个以上节点作为主机的磁盘设备组的一部分，所以它们是高可用性设备。

设备 ID (DID)

Sun Cluster 软件通过一种称为设备 ID (DID) 伪驱动程序的结构来管理全局设备。 可使用此驱动程序自动为群集内的每个设备（包括多主机磁盘、磁带驱动器和 CD-ROM）分配唯一的 ID。

设备 ID (DID) 伪驱动程序是群集的全局设备访问功能的基本构成部分。 DID 驱动程序探测群集中的所有节点并建立唯一磁盘设备列表，给每个磁盘设备分配唯一的主/次编号，这些编号在群集中所有节点上都是一致的。 执行对全局设备的访问时使用的是 DID 驱动程序所分配的唯一设备 ID，而非传统的 Solaris 设备 ID（如某一磁盘的标识 `c0t0d0`）。

这一措施可确保任何访问磁盘的应用程序（如卷管理器或使用原始设备的应用程序）都能在群集上使用一致的路径。 此一致性对多主机磁盘尤为重要，因为每个设备的本地主/次编号在各节点上都可能不相同，因而也就改变了 Solaris 设备命名惯例。 例如，节点 1 可能将一个多主机磁盘看作 `c1t2d0`，而节点 2 可能会完全不同，将同一磁盘看作是 `c3t2d0`。 DID 驱动程序指定一个全局名称（如 d10），而节点则将使用此名称，以便为每个节点提供对多主机磁盘的一致映射。

您可以通过 `scdidadm(1M)` 和 `scgdevs(1M)` 更新和管理设备 ID。 有关详细信息，请参阅相应的手册页。

磁盘设备组

在 SunPlex 系统中，所有多主机磁盘必须受 Sun Cluster 软件的控制。 首先，在多主机磁盘上创建卷管理器磁盘组 — Solaris Volume Manager 磁盘集或 VERITAS Volume Manager 磁盘组（只能在基于 SPARC 的系统中使用）。 然后将卷管理器磁盘组注册为**磁盘设备组**。 磁盘设备组是一种全局设备。 此外，Sun Cluster 软件还为群集中的每个磁盘和磁带设备创建一个原始磁盘设备组。 然而，这些群集设备组将一直处于脱机状态，直到您将其作为全局设备访问为止。

注册为 SunPlex 系统提供了有关哪个节点具有到哪个卷管理器磁盘组的路径的信息。 此时，在群集范围内可以对卷管理器磁盘组进行全局访问。 如果有多个节点可以写入（控制）磁盘设备组，存储在该磁盘设备组中的数据将具有高度可用性。 这个高度可用的磁盘设备组可用于存储群集文件系统。

注意 – 磁盘设备组独立于资源组。 一个节点可以控制资源组（代表一组数据服务进程），而另一个节点可以控制正被数据服务访问的磁盘组。 不过最好的做法是，让存储特定应用程序数据的磁盘设备组和包含此应用程序资源（应用程序守护程序）的资源组保持在同一节点上。 有关磁盘设备组和资源组之间关系的详细信息，请参阅《*Sun Cluster* **数据服务规划和管理指南**》中包含概述性内容的那一章。

通过磁盘设备组，卷管理器磁盘组成为“全局”组，因为它为基础磁盘提供了多路径支持。 物理连接到多主机磁盘的每个群集节点都提供了一条到磁盘设备组的路径。

磁盘设备组失效转移

因为磁盘群组连接着多个节点，所以在当前控制磁盘设备组的那个节点出现故障时，磁盘群组中的所有磁盘设备组都可以通过备用路径访问得到。 控制设备组的节点出现故障不会影响对此设备组的访问，但在执行恢复和一致性检查时除外。 在这段时间，所有请求都被阻挡（对应用程序是透明的），直到系统使该设备组可用为止。

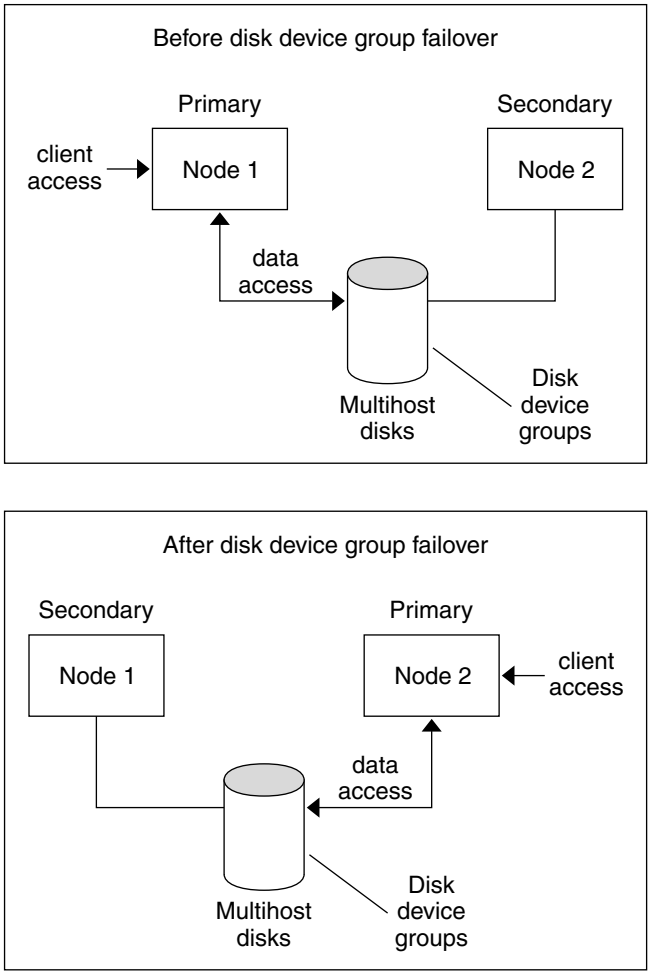


图 3–1 磁盘设备组失效转移

多端口磁盘设备组

本部分介绍了使您能够在多端口磁盘配置中协调性能和可用性的磁盘设备组特性。 Sun Cluster 软件提供了两个用于进行多端口磁盘配置的特性： `preferenced` 和 `numsecondaries` 。 您可以使用 `preferenced` 特性控制发生失效转移时节点尝试取得控制的顺序。 使用 `numsecondaries` 特性设置设备组所需的辅助节点数目。

当主节点出现故障，而又没有适当的辅助节点能够升级为主节点时，则认为高可用服务停止。 如果服务发生失效转移，且 `preferenced` 特性为 `true`，则节点将按照节点列表中的顺序选择一个辅助节点。 设置的节点列表定义了节点尝试取得主控制的顺序，或从空闲节点变为辅助节点的顺序。 您可以使用`scsetup`(1M) 实用程序动态更改设备服务的首选项。 与相应的服务供应商关联的首选项（例如，全局文件系统）将成为该设备服务的首选项。

在正常操作过程中，主节点将对辅助节点进行节点检查。 在多端口磁盘配置中，对每个辅助节点的检查会导致群集性能下降并会额外占用内存。 实现空闲节点支持可以减小节点检查造成的性能下降和内存的额外占用。 缺省情况下，磁盘设备组具有一个主节点和一个辅助节点。 其余的可用供应商节点将以空闲状态联机。 如果发生了失效转移，辅助节点将成为主节点，而节点列表中优先级最高的节点将成为辅助节点。

所需辅助节点的数目可以设置为一到设备组中非主供应商有效节点的数目之间的任意整数。

注意 – 如果正在使用 Solaris 卷管理器，则必须先创建磁盘设备组，然后将`numsecondaries` 特性设置为缺省值以外的数目。

设备服务缺省的所需辅助节点数为一。 除非有效的非主供应商数目小于所需数目，否则由复本框架维护的实际辅助供应商数目就是所需数目。 如果正在配置中增加或删除节点，您可能希望更改 `numsecondaries` 特性并重新检查节点列表。 维护节点列表和所需辅助节点数目可以防止配置的辅助节点数目和框架实际允许的数目之间发生冲突。对于 Solaris Volume Manager 设备组，请使用`metaset`(1M) 命令；或者，如果使用的是 Veritas Volume Manager，请将用于 VxVM 磁盘设备组的`scconf`(1M) 命令与 `preferenced` 和 `numsecondaries` 特性设置一起使用来管理在配置中添加和删除节点。 有关更改磁盘设备组特性的过程信息，请参阅《*Sun Cluster* **系统管理指南（适用于** *Solaris OS* **）**》中的"管理群集文件系统概述"。

全局名称空间

用于启用全局设备的 Sun Cluster 软件机制是**全局名称空间**。 全局名称空间包括 `/dev/global/` 分层结构和卷管理器名称空间。 全局名称空间可以反映多主机磁盘和本地磁盘（及所有其它群集设备，如 CD-ROM 和磁带），并提供指向多主机磁盘的多条失效转移路径。 物理连接到多主机磁盘的每个节点都为群集中的任何节点提供了到存储器的路径。

对于 Solaris Volume Manager，卷管理器名称空间通常位于 `/dev/md/`*diskset*`/dsk`（和 `rdsk`）目录中。 对于 Veritas VxVM，卷管理器名称空间位于 `/dev/vx/dsk/` *disk-group* 和 `/dev/vx/rdsk/` *disk-group* 目录中。 这些名称空间分别由整个群集中引入的各 Solaris Volume Manager磁盘集和各 VxVM 磁盘组的目录组成。 每一个这样的目录中都有此磁盘集或磁盘组中每个元设备或卷的设备节点。

在 SunPlex 系统中，本地卷管理器名称空间中的各个设备节点用 `/global/.devices/node@` *nodeID* 文件系统中某设备节点的符号链接来替换，其中 *nodeID* 是一个整数，用来在群集中代表节点。 在卷管理器设备的标准位置上，Sun Cluster 软件还继续用符号链接来表示这些卷管理器设备。 全局名称空间和标准卷管理器名称空间两者在任何群集节点上都可以找到。

全局名称空间的优点有：

- 每个节点可保持相当的独立性，不需要对设备管理模型做什么改动。
- 可以有选择地使设备变成全局设备。
- 第三方链接产生器可继续工作。
- 只要给出本地设备名称，就会有一个简单的映射用以获得其全局名称。

本地和全局名称空间示例

下表显示的是一个多主机磁盘 `c0t0d0s0` 的本地名称空间和全局名称空间之间的映射关系。

表 3–2 本地和全局名称空间映射

组件/路径	本地节点名称空间	全局名称空间
Solaris 逻辑名称	`/dev/dsk/c0t0d0s0`	`/global/.devices/node@`*nodeID* `/dev/dsk/c0t0d0s0`
DID 名称	`/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`

全局名称空间在安装时自动生成，并在每次重新配置后重新引导时自动更新。 您也可以通过运行 `scgdevs(1M)` 命令来生成全局名称空间。

群集文件系统

群集文件系统具有以下特征：

- 文件访问位置是透明的。 一个进程可打开位于系统中任何位置的文件，而且所有节点上的进程都可以使用同样的路径名定位文件。

注意 – 在群集文件系统读取文件时，它不会更新这些文件的访问时间。

- 使用了一致的协议，以确保 UNIX 文件访问在语义上的一致，即使从多个节点并行访问文件时也是如此。
- 大规模高速缓存与零复制批量 I/O 移动一起使用，以便有效地移动文件数据。
- 通过使用 `fcntl(2)` 接口，群集文件系统提供高度可用的报告文件锁定功能。 在多群集节点中运行的应用程序可以使用锁定在群集文件系统文件中的报告文件同步访问数据。 节点脱离群集后，或应用程序在锁定操作期间出现故障后，文件锁定会立即恢复。
- 即使出现故障也可以确保对数据的不间断访问。 只要到磁盘的路径仍然有效，应用程序就不会受到故障的影响。 对于原始磁盘访问和所有文件系统操作，也可保证。
- 群集文件系统与基础文件系统和卷管理软件无关。 群集文件系统可使任何支持的磁盘上的文件系统具有全局性。

在全局设备上，您可以使用 `mount -g` 进行全局装载或使用 `mount` 进行本地装载。

通过相同的文件名称（例如 `/global/foo`），程序可以从群集中的任何节点访问群集文件系统中的文件。

群集文件系统装载在所有的群集成员上。 不可以在群集成员的子集上装载群集文件系统。

群集文件系统不是特殊的文件系统类型。 也就是说，客户机看到的是基础文件系统（如 UFS）。

使用群集文件系统

在 SunPlex 系统中，所有多主机磁盘都存放在磁盘设备组中，这些组可以是 Solaris Volume Manager 磁盘集、VxVM 磁盘组或不受基于软件的卷管理器控制的独立磁盘。

要使群集文件系统具有高可用性，基础磁盘存储器必须连接到一个以上的节点。 因此，成为群集文件系统的本地文件系统（存储在节点的本地磁盘上的文件系统）并不具有高可用性。

与一般的文件系统相同，您可以通过以下两种方式装载群集文件系统：

- **手动** — 使用 `mount` 命令和 `-g` 或 `-oglobal` 装载选项从命令行装载群集文件系统，例如：

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

- **自动** — 使用 `global` 装载选项在 `/etc/vfstab` 文件中创建一个条目，以便在引导时装载群集文件系统。 接着就可以在所有节点上的 `/global` 目录下创建一个装载点。 推荐（但不要求）使用目录 `/global`。 下面是 `/etc/vfstab` 文件中一个群集文件系统的样例行：

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

注意 – Sun Cluster 软件并不强制要求在群集文件系统中使用一种命名策略，所以您可以在同一目录下（如 `/global/`*disk-device-group*）为所有群集文件系统创建一个装载点，从而简化管理。 有关详细信息，请参阅《*Sun Cluster* **软件安装指南**》和《*Sun Cluster* **系统管理指南**》。

`HAStoragePlus` 资源类型

`HAStoragePlus` 资源类型设计的目的是使诸如 UFS 和 VxFS 之类的非全局文件系统配置具有高可用性。 使用 `HAStoragePlus` 可将本地文件系统集成到 Sun Cluster 环境中，并使该文件系统具有高可用性。 `HAStoragePlus` 提供了诸如校验、装载和强制卸载之类附加的文件系统功能，使得 Sun Cluster 能对本地文件系统进行失效转移。 为了进行失效转移，本地文件系统必须驻留在启用了相似性切换功能的全局磁盘组中。

有关如何使用 `HAStoragePlus` 资源类型的信息，请参见 *Data Services Installation and Configuration Guide* 中的各数据服务章节，或第 14 章“Administering Data Services Resource”中的“Enabling Highly Available Local File Systems”。

也可以使用 `HAStoragePlus` 使资源的启动和这些资源所依赖的磁盘设备组的启动同步。 有关详细信息，请参见第 52 页 “资源、资源组和资源类型”。

Syncdir 装载选项

`syncdir` 装载选项可用于将 UFS 用作基础文件系统的群集文件系统。 不过，如果不指定 `syncdir`，性能会有明显提高。 如果指定 `syncdir`，可保证写入操作与 POSIX 兼容。 如果不指定，您会看到与 NFS 文件系统一样的行为。 例如，在某些情况下，如果不指定 `syncdir`，就只能在关闭一个文件后才发现空间不足。 有了 `syncdir`（和 POSIX 行为），空间不足的情况应该在写入操作期间就已发现了。 如果不指定 `syncdir`，很少会出现问题。所以我们建议您不指定 `syncdir`，以便获得高性能。

如果使用的是基于 SPARC 的群集，Veritas VxFS 没有与 UFS 的 `syncdir` 装载选项等价的装载选项。 未指定 `syncdir` 装载选项时，VxFS 的行为与 UFS 的行为相同。

有关全局设备和群集文件系统的常见问题，请参见第 68 页 “文件系统 FAQ”。

磁盘路径监视

Sun Cluster 软件的当前版本支持磁盘路径监视 (DPM)。 本部分介绍了有关 DPM、DPM 守护程序和用来监视磁盘路径的管理工具的概念信息。 有关如何监视、取消监视和检查磁盘路径状况的过程信息，请参考《*Sun Cluster* **系统管理指南（适用于** *Solaris OS*）》。

注意 – 运行 Sun Cluster 3.1 4/04 软件 之前发行的版本的节点不支持 DPM。 进行滚动升级时，请不要使用 DPM 命令。 所有节点均升级后，必须使这些节点处于联机状态以便使用 DPM 命令。

概述

通过监视辅助磁盘路径可用性，DPM 总体上提高了失效转移和切换的可靠性。 在切换资源之前，使用 `scdpm` 命令验证该资源所使用的磁盘路径的可用性。 由 `scdpm` 命令提供的选项使您能够监视单个节点或群集中所有节点的磁盘路径。 有关命令行选项的详细信息，请参见 `scdpm`(1M) 手册页。

DPM 组件是通过 `SUNWscu` 软件包安装的。 `SUNWscu` 软件包是通过 Sun Cluster 标准安装过程安装的。 有关安装界面的详细信息，请参见 `scinstall` (1M) 手册页。 下表说明 DPM 组件的缺省安装位置。

位置	组件
守护程序	`/usr/cluster/lib/sc/scdpmd`
命令行界面	`/usr/cluster/bin/scdpm`
共享库	`/user/cluster/lib/libscdpm.so`
守护程序状况文件（运行时创建）	`/var/run/cluster/scdpm.status`

每个节点上都运行一个多线程 DPM 守护程序。 当节点引导时，`rc.d` 脚本将启动 DPM 守护程序 (`scdpmd`)。 出现问题时，守护程序将由 `pmfd` 管理并自动重启。 下面的列表说明了 `scdpmd` 在初始启动过程中运行的方式。

注意 – 启动时，每个磁盘路径的状况都被初始化为 `UNKNOWN`。

1. DPM 守护程序从早期的状况文件或 CCR 数据库中收集磁盘路径和节点名称信息。 有关 CCR 的详细信息，请参见第 30 页 “群集配置系统信息库 (CCR)”。 启动 DPM 守护程序后，您可以强制该守护程序从指定文件名读取受监视的磁盘列表。
2. DPM 守护程序对通信接口进行初始化，以回应来自守护程序外部的组件（如命令行界面）的请求。
3. 每隔 10 分钟，DPM 守护程序会使用 `scsi_inquiry` 命令对监视列表中的各个磁盘路径执行一次强制回应操作。 每项均被锁定，以防止对正在进行修改的内容进行通信接口访问。
4. DPM 守护程序将通知 Sun Cluster Event Framework，并通过 UNIX `syslogd` (1M) 机制记录该路径的新状况。

注意 – pmfd (1M) 将报告与守护程序相关的所有错误。 如果成功，API 的所有函数返回 0，否则返回 -1。

DPM 守护程序监视借助多路径驱动程序（如 MPxIO、HDLM 和 PowerPath）而可视的逻辑路径的可用性。 由这些驱动程序管理的单独物理路径不受到监视，因为多路径驱动程序掩盖了 DPM 守护程序中的失败。

监视磁盘路径

本部分介绍了用来监视群集中磁盘路径的两种方法。 第一种方法由 scdpm 命令提供。 使用此命令监视、取消监视或显示群集中磁盘路径的状况。 此命令也用于打印故障磁盘列表和监视文件的磁盘路径。

监视群集中磁盘路径的第二种方法是由 SunPlex Manager 图形用户界面 (GUI) 提供的。 SunPlex Manager 提供了群集中受监视磁盘路径的拓扑视图。 该视图每 10 分钟更新一次，以提供有关失败的强制回应数目的信息。 使用 SunPlex Manager GUI 提供的信息和 scdpm(1M) 命令来管理磁盘路径。 有关 SunPlex Manager 的信息，请参阅《*Sun Cluster* **系统管理指南（适用于** *Solaris OS*）》中的"管理 Sun Cluster 图形用户界面"。

使用 *scdpm* 命令监视磁盘路径

scdpm(1M) 命令提供了使您可以执行以下任务的 DPM 管理命令：

- 监视新的磁盘路径
- 取消监视磁盘路径
- 再次读取 CCR 数据库中的配置数据
- 读取磁盘以从指定的文件监视或取消监视
- 报告群集中一个或所有磁盘路径的状况
- 打印通过节点可以访问的所有磁盘路径

使用 scdpm(1M) 命令和任何活动节点的磁盘路径变量对群集执行 DPM 管理任务。 磁盘路径变量通常由节点名称和磁盘名称组成。 如果不需要节点名称，它将缺省为 all（如果未指定任何节点名称）。 下表说明了磁盘路径的命名惯例。

注意 – 全局磁盘路径名称在整个群集中是一致的，因此，建议使用全局磁盘路径名称。 UNIX 磁盘路径名称在整个群集中不一致。 在各个群集节点上的一个磁盘的 UNIX 磁盘路径可能会有所不同。 一个节点上的磁盘路径可能是 c1t0d0，而另一个节点上的磁盘路径则可能是 c2t0d0。 如果使用的是 UNIX 磁盘路径名称，请在发出 DPM 命令前使用 scdidadm -L 命令将 UNIX 磁盘路径名称映射为全局磁盘路径名称。 请参阅 scdidadm(1M) 手册页。

表 3–3 磁盘路径名称样例

名称类型	磁盘路径名样例	说明
全局磁盘路径	`schost-1:/dev/did/dsk/d1`	`schost-1` 节点上的磁盘路径 `d1`
	`all:d1`	群集中所有节点上的磁盘路径 `d1`
UNIX 磁盘路径	`schost-1:/dev/rdsk/c0t0d0s0`	`schost-1` 节点上的磁盘路径 `c0t0d0s0`
	`schost-1:all`	`schost-1` 节点上的所有磁盘路径
所有磁盘路径	`all:all`	群集中所有节点上的所有磁盘路径

使用 *SunPlex Manager* 监视磁盘路径

SunPlex Manager 使您能够执行以下基本的 DPM 管理任务：

- 监视磁盘路径
- 取消监视磁盘路径
- 查看群集中所有磁盘路径的状况。

有关如何使用 SunPlex Manager 执行磁盘路径管理的过程信息，请参见 SunPlex Manager 联机帮助。

仲裁和仲裁设备

由于群集节点共享数据和资源，因此切勿将群集分割为同时活动的多个独立分区是很重要的。 CMM 保证任何时候最多只有一个群集是有效的，即使已对群集互连进行了分区。

群集分区会引起两类问题： 群集分割和失忆。 节点间的群集互连丢失后，群集划分为多个子群集，每个子群集都认为自己是唯一的分区时，就会发生群集分割。 这是由群集节点之间的通信问题引起的。 失忆在群集关闭后又重新启动时发生，此时的群集数据比关闭时旧。 如果框架数据有多个版本存储在磁盘上，而新的群集体在尚未获得最新的版本时启动，则可能发生这种情况。

群集分割和失忆可以通过以下方法避免：赋予每个节点一个选票，并规定只有获得多数选票才能成为有效群集。 获得多数选票的分区拥有**仲裁**，因此允许其运行。 只要群集中有两个以上的节点，这种多数选票机制就能运行良好。 在双节点群集中，多数为二。 如果此类群集被划分，则每个分区都需要一个外部选票才能获得仲裁。 此外部选票由**仲裁设备**提供。 两个节点之间共享的任何磁盘都可作为仲裁设备。 用作仲裁设备的磁盘可以包含用户数据。

仲裁算法自动执行： 当群集事件触发其计算时，计算的结果可以随群集生存期的不同而改变。

仲裁选票计数

群集节点和仲裁设备都会投票以形成仲裁。 缺省情形下，群集节点在引导并成为群集成员时获取其中一个的仲裁选票计数。 节点的选票数可以是零，例如当正在安装节点时，或当管理员将节点置于维护状态时。

仲裁设备获取仲裁选票计数基于设备的节点连接数。 当设置仲裁设备时，它获得了一个最大选票计数 *N*-1，其中 *N* 是仲裁设备连接的投票数。 例如，连接到两个选票数非零的节点的仲裁设备有其中一个的仲裁数（二减一）。

在群集安装期间（或以后通过使用《*Sun Cluster* **系统管理指南**》中描述的过程）配置仲裁设备。

注意 – 仅在当前连接的节点中至少有一个是群集成员时，仲裁设备才对选票数起作用。 同时，在群集引导期间，仅在当前连接的至少一个节点正在引导，并且在关闭时它是最近刚刚引导的群集成员的情况下，仲裁设备才对选票数起作用。

仲裁配置

仲裁配置依赖于群集中节点的数目：

- **双节点群集 –** 双节点群集需要两个仲裁选票才能形成。 这两个选票可以来自于两个群集节点，或者只来自一个节点和一个仲裁设备。 然而，在双节点群集中，必须配置一个仲裁设备，以确保在一个节点发生故障时另外那个节点可以继续工作。
- **多于两个节点的群集** – 您应在对磁盘存储器群组进行共享访问的每对节点中指定一个仲裁设备。 例如，假定您拥有一个类似于图 3–2 中所示的三节点群集。 在此图中，nodeA 和 nodeB 共享对同一磁盘群组的访问权，而 nodeB 和 nodeC 共享对另一磁盘群组的访问权。 总共会有五个仲裁选票，其中三个来自节点，两个来自节点共享的仲裁设备。 一个群集需要多数仲裁选票才能形成。

  Sun Cluster 软件不要求也不强迫在对磁盘存储器群组进行共享访问的每对节点中指定一个仲裁设备。 然而，在 N+1 配置退化为双节点的群集、并且可访问两个磁盘群组的节点也发生故障时，此软件可以提供所需的仲裁选票。 如果您在每对节点之间配置了仲裁设备，则其余的节点仍可作为一个群集来运行。

  有关这些配置的示例，请参见图 3–2。

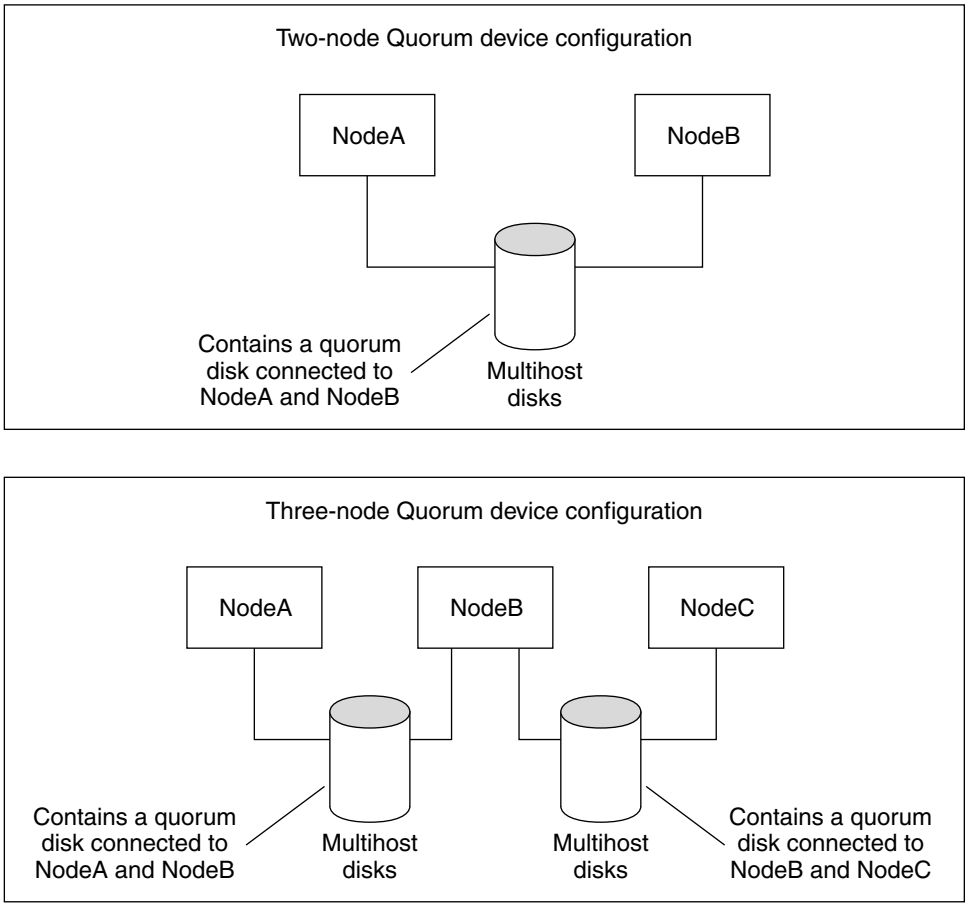


图 3–2 仲裁设备配置示例

仲裁原则

在设置仲裁设备时，请遵循下列原则：

- 在连接到同一个共享的磁盘存储器群组的所有节点间建立仲裁设备。 在共享群组内增加一个磁盘作为仲裁设备以确保在任何节点发生故障时，其它节点可以维持仲裁并可以控制共享群组上的磁盘设备组。
- 必须将仲裁设备连接到至少两个节点上。
- 仲裁设备可以是任何用作双端口仲裁设备的 SCSI-2 或 SCSI-3 磁盘。 连接到超过两个节点的磁盘必须支持 SCSI-3 持久性组保留 (PGR)，而不论磁盘是否用作仲裁设备。 有关详细信息，请参阅《*Sun Cluster* **软件安装指南**》中有关规划的章节。
- 您可以使用包含用户数据的磁盘作为仲裁设备。

故障防护

群集的一个主要问题是引起群集分区的故障（称作**群集分割**）。 当此故障发生时，并不是所有节点都可以通信，所以个别节点或节点子集可能会尝试组成个体或群集子集。 每个子集或分区都可能以为它对多主机磁盘具有唯一访问权和所有权。 多个节点试图写入磁盘会导致数据损坏。

故障防护通过以物理方式防止对磁盘的访问，限制了节点对多主机磁盘的访问。 当节点脱离群集时（它或是发生故障，或是分区），故障防护确保了该节点不再能访问磁盘。只有当前成员节点有权访问磁盘，以保持数据的完整性。

磁盘设备服务为使用多主机磁盘的服务提供了失效转移能力。 在当前担当磁盘设备组主节点（属主）的群集成员发生故障或变得无法访问时，一个新的主节点会被选中，使得对磁盘设备组的访问得以继续，而只有微小的中断。 在此过程中，旧的主节点必须放弃对设备的访问，然后新的主节点才能启动。 然而，当一个成员从群集断开并变得无法访问时，群集无法通知那个节点释放那些将该节点作为主节点的设备。 因而，您需要一种方法来使幸存的成员能够从失败的成员那里控制并访问全局设备。

SunPlex 系统使用 SCSI 磁盘保留来实现故障防护。 使用 SCSI 保留可以将发生故障的节点从多主机磁盘中“隔离”出来，从而防止发生故障的节点访问这些磁盘。

SCSI-2 磁盘保留支持一种保留形式，它或者给所有连接到磁盘的节点都授予访问权（当没有进行任何保留时），或者限制对单个节点（即拥有该保留的节点）的访问权。

当群集成员检测到另一个节点不再通过群集互连进行通信时，它启动故障防护措施来避免另一个节点访问共享磁盘。 如果出现这种故障防护，则在节点的控制台上显示带有“保留冲突”的隔离节点应急消息是很正常的。

发生保留冲突的原因是：在某个节点已被检测为不再是群集成员后，又将一个 SCSI 保留置于在此节点与其它节点所共享的所有磁盘上。 防护节点可能不会意识到它正处于防护状态；如果它试图访问这些共享磁盘之中的一个，它会检测到该保留并进入应急状态。

故障防护的故障快速防护机制

群集框架通过一种机制确保故障节点无法重新引导并开始写入共享存储器，这种机制称为**故障快速防护**。

属于群集成员的节点对它们可以访问的磁盘（包括仲裁磁盘）持续启用一个特定 ioctl：MHIOCENFAILFAST。 该 ioctl 是对磁盘驱动程序的指令，它能使节点在以下情况下自身进入应急状态：某磁盘由于被其它节点保留而无法让该节点进行访问。

MHIOCENFAILFAST ioctl 使驱动程序检查节点发布给磁盘的每个读写操作返回的错误，以查找 `Reservation_Conflict` 错误代码。 该 ioctl 定期在后台向磁盘发出一个测试操作，检查是否出现 `Reservation_Conflict`。 如果系统返回 `Reservation_Conflict` 消息，前台和后台控制流路径均进入应急状态。

对于 SCSI-2 磁盘，保留不是持久的 — 节点重新引导之后，保留信息不再存留。 对于带有持久性组保留 (PGR) 的 SCSI-3 磁盘，保留信息存储在磁盘上，并且在节点重新引导之后这些信息仍然存留。 无论使用 SCSI-2 磁盘还是 SCSI-3 磁盘，故障快速防护机制的工作方式都是一样的。

如果某节点与群集中其它节点失去连接，并且它不属于可获取仲裁的分区的一部分，它将被另一节点强行从该群集中删除。 属于可获取仲裁的分区一部分的另一节点将保留放置在共享磁盘上，当不具备仲裁的节点试图访问共享磁盘时，它将接到保留冲突消息，并在故障快速防护机制的作用下进入应急状态。

出现应急状态之后，节点可能重新引导并尝试重新加入群集；或者，如果群集是由基于 SPARC 的系统组成的，则停留在 OpenBoot™ PROM (OBP) 提示符处。 所采取的操作取决于 `auto-boot?` 参数的设置。 您可以在基于 SPARC 的群集中的 OpenBoot PROM `ok` 提示符处使用`eeprom`(1M) 来设置 `auto-boot?`，也可以在基于 x86 的群集中，在 BIOS 引导之后选择运行 SCSI 实用程序来设置 `auto-boot?`。

数据服务

术语**数据服务**指被配置为在群集而不是在单一服务器上运行的第三方应用程序，如 Sun Java System Web Server（以前的 Sun Java System Web Server），以及基于 SPARC 的群集上的 Oracle。 数据服务包括应用程序、专用的 Sun Cluster 配置文件，以及控制下列应用程序操作的 Sun Cluster 管理方法。

- 启动
- 停止
- 监视并采取纠正措施
- 有关数据服务类型的信息，请参阅《*Sun Cluster* **概述（适用于** *Solaris OS*）》中的“数据服务”。

图 3–3 将运行在单个应用程序服务器（单服务器模型）上的应用程序与在群集（群集服务器模型）上运行的同一应用程序进行比较。 注意：从用户的观点来看，这两种配置没有任何区别，只是群集的应用程序可能运行速度更快、可用性更高而已。

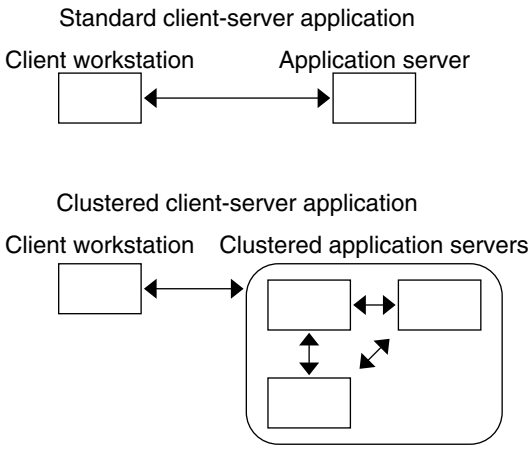


图 3–3 标准客户机/服务器配置与群集客户机/服务器配置

在单服务器模型中，您可以将应用程序配置为通过特定的公共网络接口（一个主机名）访问服务器。 此主机名与该物理服务器相关联。

在群集服务器模型中，公共网络接口是一个**逻辑主机名**或一个**共享地址**。 术语**网络资源**用来指逻辑主机名和共享地址。

某些数据服务要求指定逻辑主机名或共享地址作为网络接口 — 它们不可互换。 其它数据服务允许您指定逻辑主机名或共享地址。 有关必须指定的接口类型的详细信息，请参阅每项数据服务的安装与配置信息。

网络资源与具体的物理服务器无关 — 它可以在物理服务器之间进行移植。

网络资源初始与一个节点（**主节点**）相关联。 如果主节点发生故障，网络资源和应用程序资源将失效转移至另一个群集节点（辅助节点）上。 进行网络资源故障切换后，经过短暂延迟，应用程序资源就可以在辅助节点上继续正常运行。

图 3–4 将单服务器模型与群集服务器模型进行比较。 注意：在群集服务器模型中，网络资源（即本例中的逻辑主机名）可以在两个或多个群集节点之间移动。 应用程序被配置为使用此逻辑主机名代替与特定服务器相关联的主机名。

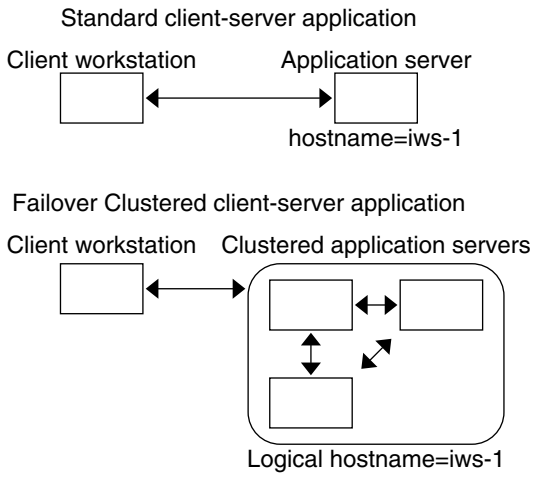


图 3–4 固定主机名与逻辑主机名

共享地址初始与一个节点相关联。 这个节点被称为全局接口节点。 共享地址被用作群集的单个网络接口。 这就是**全局接口**。

逻辑主机名模型与可伸缩服务模型之间的区别在于，在后一种模型中，每个节点在其回送接口上还配置有共享地址。 这种配置使同一数据服务的多个实例可以同时在几个节点上使用。 术语"可伸缩服务"表示通过添加附加群集节点，您可以为应用程序增加 CPU 处理能力，从而提升性能

如果全局接口节点出现故障，将在另一个也在运行该应用程序的实例的节点上启用共享地址（因此这个节点就成为新的全局接口节点）。 或者，可以将共享地址故障切换到之前未运行该应用程序的另一个群集节点上。

图 3–5 将单服务器配置与可伸缩群集服务配置进行比较。 注意：在可伸缩服务配置中，共享地址出现在所有节点上。 与逻辑主机名用于失效转移数据服务的方式类似，应用程序已配置为使用这个共享地址，而不使用与特定服务器相关联的主机名。

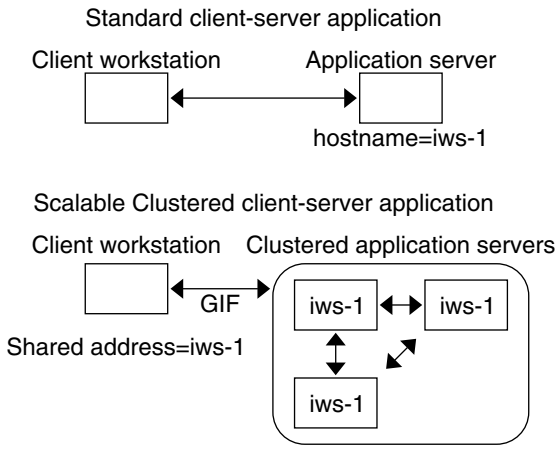


图 3–5 固定主机名与共享地址

数据服务方法

Sun Cluster 软件提供了一套服务管理方法。 这些方法在Resource Group Manager (RGM) 的控制下运行，RGM 使用它们来启动、停止和监视群集节点上的应用程序。 这些方法连同群集框架软件和多主机磁盘一起，使应用程序能够实现失效转移或可伸缩的数据服务。

RGM 也管理群集中的资源，包括应用程序实例和网络资源（逻辑主机名和共享地址）。

除 Sun Cluster 软件提供的方法之外，SunPlex 系统还提供一个 API 和多种数据服务开发工具。 这些工具使应用程序编程人员能够开发所需要的数据服务方法，以使其它应用程序作为高度可用的数据服务与 Sun Cluster 软件一起运行。

失效转移数据服务

如果正在运行数据服务的节点（主节点）发生故障，那么该服务会被移植到另一个工作节点而无需用户干预。 失效转移服务利用了**失效转移资源组**，它是一个用于应用程序实例资源和网络资源（**逻辑主机名**）的容器。 逻辑主机名是一些可以配置到节点上的 IP 地址，然后自动在原始节点解除配置，并配置到另一节点上。

对于失效转移数据服务，应用程序实例仅在一个单独的节点上运行。 如果故障监视器检测到一个故障，它或者试图在同一节点上重新启动该实例，或者在另一个节点上启动实例（失效转移），这取决于该数据服务是如何配置的。

可伸缩数据服务

可伸缩数据服务对多个节点上的活动实例有潜能。 可伸缩服务使用两个资源组： **可伸缩资源组**，包含应用程序资源；失效转移资源组，包含可伸缩服务依赖的网络资源（**共享地址**）。 可伸缩资源组可以在多个节点上联机，因此服务的多个实例可以立刻运行。 以共享地址为主机的失效转移资源组每次只在一个节点上联机。 以可伸缩服务做主机的所有节点使用相同的共享地址来主持该服务。

服务请求通过一个单独的网络接口（全局接口）进入群集，并依据由**负载平衡策略**设置的几个预定义算法之一来将这些请求分发到节点。 群集可以使用负载平衡策略来平衡几个节点间的服务负载。 注意：在包含其它共享地址的不同节点上可以有多个全局接口。

对于可伸缩服务，应用程序实例在多个节点上同时运行。 如果拥有全局接口的节点出现故障，全局接口将切换到其它节点。 如果一个正在运行的应用程序实例发生故障，则该实例尝试在同一节点上重新启动。

如果应用程序实例不能在同一节点上重新启动，而另一个未使用的节点被配置运行该服务，那么该服务会切换到这个未使用的节点。 否则，它继续运行在那些剩余节点上，并且很可能会降低服务吞吐量。

注意 – 每个应用程序实例的 TCP 状态与该实例一起保存在此节点上，而不是在全局接口节点上。 因此，全局接口节点上的故障不影响连接。

图 3–6 显示了失效转移和可伸缩资源组的一个示例，以及在它们之间存在的对于可伸缩服务的依赖性。 此示例显示了三个资源组。 失效转移资源组包括高度可用的 DNS 应用程序资源，以及由高度可用的 DNS 和 Apache Web Server（只能在基于 SPARC 的群集中使用）共同使用的网络资源。 可伸缩资源组仅包括 Apache Web 服务器应用程序实例。 注意，资源组在可伸缩和失效转移资源组（实线）之间存在依赖性，而所有的 Apache 应用程序资源都依赖于网络资源 `schost-2`，这是一个共享地址（虚线）。

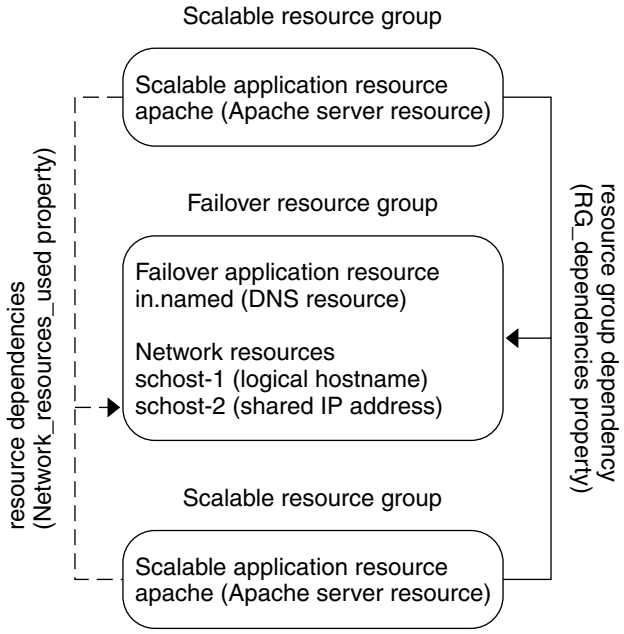


图 3–6 SPARC: 失效转移和可伸缩资源组示例

负载平衡策略

负载平衡在响应时间和吞吐量上同时提高了可伸缩服务的性能。

可伸缩数据服务有两类： **纯粹服务**和**粘滞服务**。 纯粹服务就是它的任何实例都可以对客户机的请求作出响应的服务。 粘滞服务是客户机发送请求到相同实例的那种服务。 那些请求不被重定向到其它实例。

纯粹服务使用加权的负载平衡策略。 在这种负载平衡策略下，客户机请求按缺省方式被均衡地分配到群集内的服务器实例之上。 例如，在一个三节点群集中，让我们来假设每个节点的加权为 1。每个节点将代表该服务对所有客户机请求的 1/3 进行服务。 加权可以在任何时候由管理员通过 `scrgadm(1M)` 命令界面或通过 SunPlex Manager GUI 进行修改。

粘滞服务有两种风格：**普通粘滞服务**和**通配粘滞服务**。 粘滞服务允许多个 TCP 连接上并行的应用程序级会话来共享状态内内存（应用程序会话状态）。

普通粘滞服务允许客户机在多个并行的 TCP 连接之间共享状态。 相对于监听单个端口的服务器实例来说，可以说客户机是“粘滞的”。 假设实例保持打开状态并可访问，并且在服务处于联机状态时负载平衡策略不更改，将保证该客户机的所有服务请求都传给相同的服务器实例。

例如，客户机的 Web 浏览器连接到使用三种不同 TCP 连接，连接到端口为 80 的共享 IP 地址，但连接在服务时正在它们之间交换已缓存的会话信息。

粘滞策略的普遍化延伸到在相同实例场景后面交换会话信息的多个可伸缩服务。 当这些服务在相同实例场景后面交换会话信息时，相对于不同端口上监听的相同节点上的多个服务器实例来说，可以说客户机是“粘滞的”。

例如，在电子商务站点中的顾客通过端口 80 使用普通的 HTTP 在他的购物车中装满商品，但要切换到端口 443 通过 SSL 来发送安全数据，以便通过信用卡支付购物车中的商品。

通配粘滞服务使用动态分配的端口号，但仍期望客户机请求去往相同的节点。 相对于相同的 IP 地址来说，客户机就是端口上的“粘滞通配”。

被动模式 FTP 是这一策略的一个好例子。 客户机连接到端口 21 上的 FTP 服务器，并接着被服务器通知须连接回动态端口范围中的收听器端口服务器。 此 IP 地址的所有请求都被转发到服务器通过控制信息通知客户的相同节点上。

请注意，对于每个粘滞策略，加权的负载平衡策略都是缺省生效的，从而使客户的最初请求被定向到由负载平衡程序指定的实例。 在客户机已经为正运行着实例的节点建立一种亲密关系之后，只要该节点可访问并且负载平衡策略未更改，以后的请求就会定向到此实例。

关于特定的负载平衡策略的补充详细信息在下面进行讨论。

- 加权的。 根据指定的加权值在各种节点间分配负载。 此策略是使用 `Load_balancing_weights` 特性的 `LB_WEIGHTED` 值设置的。 如果一个节点的加权未明显地设置，则会使用此节点的缺省加权值 1。

  这种加权策略将来自客户机的一定比例的流量重定向到特定的节点。 已知 X=加权，A=所有活动节点的总加权，活动的节点可以获得定向到该活动节点的新连接总数的 X/A 近似值（如果连接总数足够大）。 此策略不对个别请求寻址。

  注意，这一策略不是循环共享的。 循环共享策略总是会使来自客户机的每个请求到达不同的节点：第一个请求到达节点 1，第二个请求到达节点 2，以此类推。
- 粘滞的。 在此策略中，端口集在配置应用程序资源时是已知的。 此策略是使用 `Load_balancing_policy` 资源特性的 `LB_STICKY` 值设置的。
- 粘滞通配符。 此策略是普通“粘滞”策略的超集。 对于由 IP 地址识别的可伸缩服务，端口由服务器来分配（并且事先不知道）。 端口可能会变化。 此策略是使用 `Load_balancing_policy` 资源特性的 `LB_STICKY_WILD` 值设置的。

恢复设置

资源组在一个节点出现故障时转移到另一个节点。 出现这种情况时，原始的辅助节点成为新的主节点。 恢复设置指定了在原始主节点恢复联机状态时将进行的操作。 选项包括使原始的主节点重新恢复为主节点（恢复）或仍保留当前的主节点。 使用恢复资源组特性设置指定需要的选项。

在某些实例中，如果托管资源组的原始节点重复发生故障和重新引导，则设置恢复可能会降低资源组的可用性。

数据服务故障监视器

每个 SunPlex 数据服务都提供一个故障监视器来定期探测数据服务，确定其是否运作正常。 故障监视器将验证应用程序守护程序是否正在运行，还将验证客户机是否正接受服务。 根据探测返回的信息，可以启动预定义的操作，例如重新启动守护程序或导致故障切换。

开发新的数据服务

Sun 提供了配置文件和管理方法模板，使您能够在一个群集中让各种应用程序以失效转移或可伸缩服务的方式运行。 如果您想使之作为一个失效转移或可伸缩服务来运行的应用程序不是 Sun 当前提供的，则可以使用一个 API 或 DSDL API 来配置该应用程序，使之作为一个失效转移或可伸缩服务运行。

存在一组标准，用于确定应用程序是否可以成为失效转移服务。 特定的标准在 SunPlex 文档中进行了说明，这些文档说明您的应用程序可使用的 API。

这里，我们提出一些准则来帮助您了解您的服务是否可受益于可伸缩数据服务体系结构。 有关可伸缩服务的更多基本信息，请查阅第 47 页 “可伸缩数据服务”。

满足下列准则的新服务可以利用可伸缩服务。 如果现有的服务不完全符合这些准则，则可能需要重写一些部分，使服务符合这些准则。

可伸缩数据服务具有以下特点。 首先，此类服务由一个或多个服务器**实例**组成。 每个实例运行在群集的不同节点上。 在同一节点上不能运行两个或多个相同服务的实例。

其次，如果服务提供外部逻辑数据存储，那么从多个服务器实例对此存储的并行访问必须同步，以避免丢失更新信息或在数据更改时读取数据。 请注意，我们所说的“外部的”是为了区分存储于内存内的状态，而所说的“逻辑的”是因为存储看起来像单独的实体，尽管它本身可能是复制的。 此外，这种逻辑数据存储还有以下特性：不论何时只要有服务器实例更新了该存储，其它实例就将立即看到该更新。

SunPlex 系统通过它的群集文件系统和全局原始分区来提供这样一个外部存储器。 又比如，假设一项服务将新数据写入外部日志文件，或修改现有的数据。 当此服务的多个实例运行时，每个都可以访问此外部日志，并且可能会同时访问这一日志。 每个实例必须同步其对日志的访问，否则这些实例就会彼此干扰。 此服务可以通过 `fcntl(2)` 和 `lockf(3C)` 来使用普通的 Solaris 文件锁定，从而如愿实现同步。

此类存储的另一个示例是后端数据库，如高度可用的 Oracle 或用于基于 SPARC 群集的 Oracle Parallel Server/Real Application Clusters。 请注意，使用数据库查询或更新事务时，此类后端数据库服务器提供了内置同步功能，因此，多个服务器实例不需要实现各自的同步。

在当前阶段还不是可伸缩服务的一个服务示例是 Sun 的 IMAP 服务器。 该服务更新一个存储，但那个存储是专用的，并且当多个 IMAP 实例写入到这一存储时，它们因为更新没有被同步而相互覆盖。 IMAP 服务器必须被重写以使并行访问同步。

最后要注意的一点是，实例可能具备一些专用数据，这些数据未与其它实例的数据相连接。 在这种情况下，该服务不必关心自己与并行访问是否同步，因为数据是专用的，只有这个实例才可对其进行处理。 此时，您必须小心不要在群集文件系统下存储此专用数据，因为它有变为全局访问的可能性。

数据服务 API 和数据服务开发库 API

SunPlex 系统提供以下组件以使应用程序具有高可用性：

- 将数据服务作为 SunPlex 系统的一部分来提供
- 一个数据服务 API
- 一个数据服务开发库 API
- 一种“普通的”数据服务

《*Sun Cluster* **数据服务规划和管理指南**》介绍了如何安装和配置随 SunPlex 系统一起提供的数据服务。 《*Sun Cluster* **数据服务开发者指南**》介绍了如何装备其它应用程序以使它们在 Sun Cluster 框架下具有高可用性。

Sun Cluster API 使应用程序开发者能够开发用于启动和停止数据服务实例的故障监视器和脚本。 有了这些工具，应用程序就可以被装备成为一种失效转移或可伸缩的数据服务。 另外，SunPlex 系统提供一种“普通的”数据服务，这种服务可以用于快速生成应用程序所需的启动和停止方法，从而使它作为一种失效转移或可伸缩的服务运行。

为数据服务通信使用群集互连

一群集必须有节点之间的多个网络互连，构成群集互连。 群集软件使用多个互连来提高可用性并改善性能。 如果是内部通信（如文件系统数据或可伸缩服务数据），消息将以循环方式在所有可用的互连间均匀进行分配。

群集互连对应用程序也是可用的，从而在节点间进行高可用性通信。 例如，一个分布式应用程序的组件可能运行在不同的需要进行通信的节点上。 通过使用群集互连而不使用公共传输，这些连接可承受单个链接失败。

要使用群集互连来在节点间进行通信，应用程序必须使用安装群集时配置的专用主机名。 例如，如果节点 1 的专用主机名是 `clusternode1-priv`，请使用此主机名通过群集互连与节点 1 进行通信。使用此名称打开的 TCP 套接字通过群集互连路由并可以在网络发生故障的情况下透明地重新路由。

注意，由于在安装时可以配置专用主机名，所以群集互连可使用此时选择的任何名称。可以使用 `scha_privatelink_hostname_node` 变量从 `scha_cluster_get(3HA)` 处获取实际的名称。

如果是在应用程序级别上使用群集互连，则在每对节点之间使用单独的一个互连；但如果可能，不同的节点对会使用不同的互连。 例如，试想一个运行在三个基于 SPARC 的节点上的应用程序通过群集互连进行通信。 在节点 1 和 2 之间的通信可能会在接口 `hme0` 上进行，而节点 1 和 3 之间的通信可能会在接口 `qfe1` 上进行。 即，任何两个节点间的应用程序通信仅限于单个互连，而内部群集通信则均匀地分配到了所有的互连上。

注意，应用程序共享与内部群集通信的互连，所以对该应用程序可用的带宽取决于用于其它群集通信的带宽。 如果出现故障，内部通信会在仍正常运行的互连上循环，而失败的互连上的应用程序连接可切换到一个正常互连上。

两种类型的地址支持群集互连，且专用主机名上的 gethostbyname(3N) 通常会返回两个 IP 地址。 第一个地址称为**逻辑成对地址**，第二个地址称为**逻辑单节点地址**。

每对节点各分配了一个逻辑成对地址。 此小型逻辑网络支持连接失效转移。 每个节点还分配了一个固定的单节点地址。 即，clusternode1-priv 的逻辑成对地址因节点而异，而 clusternode1-priv 的逻辑单节点地址在各个节点上相同。 但是，一个节点对它自身来说并没有成对地址，所以节点 1 上的 gethostbyname (clusternode1-priv) 仅返回逻辑单节点地址。

请注意，通过群集互连接受连接、然后出于安全原因检验 IP 地址的应用程序，必须检查 gethostbyname 返回的所有 IP 地址，而不仅仅检查第一个 IP 地址。

如果需要使 IP 地址在所有点上的应用程序中保持一 致，请配置应用程序，使单节点地址同时绑定到客户端和服务器端，从而使所有的连接看起来是通过单节点地址出入。

资源、资源组和资源类型

数据服务利用了几种类型的**资源**： Sun Java System Web Server（以前的 Sun Java System Web Server）或 Apache Web Server 等应用程序使用这些应用程序依赖的网络地址（逻辑主机名和共享地址）。 应用程序和网络资源组成由 RGM 管理的一个基本单元。

数据服务是资源类型。 例如，Sun Cluster HA for Oracle 是资源类型 SUNW.oracle-server，而 Sun Cluster HA for Apache 是资源类型 SUNW.apache。

注意 – 资源类型 SUNW.oracle-server 仅在基于 SPARC 的群集中使用。

资源就是群集范围内定义的**资源类型**的实例化。 有数种已定义的资源类型。

网络资源或者是 SUNW.LogicalHostname 资源类型，或者是 SUNW.SharedAddress 资源类型。 这两种资源类型已由 Sun Cluster 软件预注册。

SUNW.HAStorage 和 HAStoragePlus 资源类型用于使资源的启动和这些资源所依赖的磁盘设备组的启动同步。 它可确保在数据服务启动时，到群集文件系统装载点、全局设备和设备组名称的路径可用。 有关详细信息，请参见 *Data Services Installation and Configuration Guide* 中的“Synchronizing the Startups Between Resource Groups and Disk Device Groups”。 （HAStoragePlus 资源类型在 Sun Cluster 3.0 5/02 中就已可用并且增加了另一功能，从而使本地文件系统具有高可用性。 有关此功能的详细信息，请参阅第 36 页 “HAStoragePlus 资源类型”。）

RGM 所管理的资源被放入一个称作**资源组**的组中，这样就可将它们作为一个单元来进行管理。 如果对资源组启动失效转移或切换，那么该资源组就将作为单元移植。

注意 – 当您使一个包含应用程序资源的资源组联机时，应用程序便启动。 数据服务启动方法会一直等待，直到应用程序在成功退出前启动并运行。 决定何时应用程序启动并运行的方法，与数据服务故障监视器决定数据服务是否正在服务于客户机所采用的方法相同。 有关此过程的详细信息，请参阅《*Sun Cluster* **数据服务规划和管理指南**》。

Resource Group Manager (RGM)

RGM 将数据服务（应用程序）作为资源进行控制，而资源是由**资源类型**实现所管理的。 这些实现可以由 Sun 提供，或者由拥有普通数据服务模板、数据服务开发库 API (DSDL API) 或资源管理 API (RMAPI) 的开发者创建。 群集管理员在称为**资源组**的容器中创建和管理资源。 RGM 根据群集成员关系的变化停止和启动所选节点上的资源组。

RGM 对**资源**和**资源组**进行操作。 RGM 操作导致资源和资源组在联机和脱机状态之间进行转换。 有关可应用于资源和资源组的状态和设置的完整说明，请参阅第 53 页 “资源和资源组状态和设置”一节。 有关如何启动 RGM 控制下的资源管理项目的信息，请参见第 52 页 “资源、资源组和资源类型”。

资源和资源组状态和设置

管理员在资源和资源组上应用静态设置。 这些设置只能通过管理员操作来进行更改。 RGM 在各种动态“状态”之间移动资源组。 下表列出了这些设置和状态。

- 管理或取消管理 – 这些是群集范围的设置，仅适用于资源组。 资源组由 RGM 进行管理。 `scrgadm(1M)` 命令可用于指示 RGM 对资源组进行管理或取消其管理。 这些设置不会随群集的重新配置而更改。

  首次创建资源组后，它是不受管理的。 必须先对资源组进行管理，放入该资源组的资源才能起作用。

  在一些数据服务（例如可伸缩 Web 服务器）中，必须在启动网络资源之前以及停止网络资源之后进行工作。 通过初始化 (INIT) 和结束 (FINI) 数据服务方法来进行此项工作。 只有在资源所在的资源组处于管理状态时才可运行 INIT 方法。

  如果某资源组从取消管理状态变成管理状态，任何已注册的组 INIT 方法均可对组中资源运行。

  如果资源组从管理状态变成取消管理状态，要求对所有已注册的 FINI 方法执行清除。

  INIT 和 FINI 方法最常用于可伸缩服务的网络资源，但它们也可用于进行应用程序没有完成的任何初始化或清除工作。

- 启用或禁用 – 这些是群集范围的设置，适用于资源。 `scrgadm(1M)` 命令可用于启用或禁用资源。 这些设置不会随群集的重新配置而更改。

  资源的正常设置应为：处于启用状态，并正在系统中运行。

  如果由于某些原因，要使资源在所有群集节点上无法使用，请禁用资源。 禁用的资源在一般情况下不能使用。

- 联机或脱机 – 这些是应用于资源和资源组的动态状态。

  在切换移转或失效转移过程中，这些状态会随着由于重新配置群集而发生的群集转换而改变。 它们还可通过管理员操作来进行更改。 `scswitch(1M)` 可用于更改资源或资源组的联机或脱机状态。

  失效转移资源或资源组在任何时候都只能在一个节点上处于联机状态。 可伸缩资源和资源组可以在某些节点上联机，而在其它节点上脱机。 在切换转移或失效转移过程中，资源组和资源组中的资源将在一个节点上脱机，然后在另一个节点上联机。

  如果资源组脱机，则其中的所有资源均脱机。 如果资源组联机，则其启用的所有资源均联机。

  资源组可包含若干个资源，各资源之间存在依赖性。 这些依赖性要求资源以特定的顺序联机和脱机。 对于各个资源来说，用于使资源联机和脱机的各种方法可能需要花费不同的时间。 由于资源依赖性以及启动和停止时间的差异，在一个群集的重新配置过程中，单个资源组中的各个资源可能处于不同的联机和脱机状态。

资源和资源组特性

您可以为您的 SunPlex 数据服务配置资源和资源组的特性值。 标准特性对于所有数据服务都是通用的。 扩展特性是每个服务的特定特性。 一些标准和扩展特性已配置为缺省值，因此您不必去修改它们。 其它特性作为创建和配置资源进程的一部分，需要进行设置。 每个数据服务的文档都指定了哪些资源特性可以进行设置，以及如何设置这些特性。

标准特性用于配置那些通常独立于任何特定数据服务的资源和资源组特性。 标准特性集的说明可见于《*Sun Cluster* **数据服务规划和管理指南**》的附录。

RGM 扩展特性提供应用程序二进制文件和配置文件的位置等信息。 当您配置数据服务时，就修改了扩展特性。 扩展特性集在《*Sun Cluster* **数据服务规划和管理指南**》中专门介绍数据服务的章节中进行了说明。

数据服务项目配置

可以将数据服务配置为在使用 RGM 联机时，在 Solaris 项目名称下启动。 该配置将 RGM 管理的资源或资源组与 Solaris 项目 ID 相关联。 从资源或资源组到项目 ID 的映射使您能够使用 Solaris 环境中提供的高级控制来管理群集中的工作量和消耗量。

注意 – 只有运行具有 Solaris 9 的 Sun Cluster 软件的当前版本时，才能执行此配置。

通过在群集环境中使用 Solaris 管理功能，可以确保大多数重要的应用程序在与其它应用程序共享一个节点时具有优先权。 如果采用统一服务，或者由于应用程序发生失效转移，则多个应用程序可能共享一个节点。 使用此处介绍的管理功能，可以防止其它优先级较低的应用程序过多地消耗系统供应（如 CPU 时间），从而提高重要应用程序的可用性。

注意 – 有关此功能的 Solaris 文档介绍了 CPU 时间、进程、任务和类似的组件（如资源）。 同时，Sun Cluster 文档采用术语“资源”来说明由 RGM 控制的实体。 以下部分将采用术语“资源”指代由 RGM 控制的 Sun Cluster 实体，并采用术语“供应”指代 CPU 时间、进程和任务。

本部分介绍了将数据服务配置为在指定的 Solaris 9 `project`(4) 中启动的进程的概念说明。 本部分还介绍了若干失效转移情况，以及规划使用 Solaris 环境提供的管理功能的建议。 有关管理功能的详细概念文档和过程文档，请参见 *Solaris 9 System Administrator Collection* 中的 *System Administration Guide: Resource Management and Network Services*。

当在群集中将资源和资源组配置为使用 Solaris 管理功能时，请考虑使用以下高级进程：

1. 将应用程序配置为资源的一部分。
2. 将资源配置为资源组的一部分。
3. 启用资源组中的资源。
4. 使资源组处于管理状态。
5. 为资源组创建一个 Solaris 项目。
6. 配置标准特性，使资源组名称与步骤 5 中创建的项目相关联。
7. 使资源组联机。

要配置标准的 `Resource_project_name` 或 `RG_project_name` 特性，以便将 Solaris 项目 ID 与资源或资源组相关联，请使用 `-y` 选项和`scrgadm`(1M) 命令。 将特性值设置为资源或资源组。 有关特性定义，请参阅《*Sun Cluster* **数据服务规划和管理指南（适用于** *Solaris OS*）》中的“标准特性”。 有关特性说明，请参阅`r_properties`(5) 和`rg_properties`(5)。

指定的项目名称必须存在于项目数据库 (`/etc/project`) 中，而 root 用户必须被配置为此命名项目的成员。 有关项目名称数据库的概念信息，请参见 *Solaris 9 System Administrator Collection* 中 *System Administration Guide: Resource Management and Network Services* 中的“Projects and Tasks”。 有关项目文件语法的说明，请参见 `project`(4)。

当 RGM 将资源或资源组联机时，它将启动项目名称下的相关进程。

注意 – 用户可以随时将资源或资源组与项目相关联。 但是，必须使用 RGM 将资源或资源组脱机，然后再联机，新的项目名称才有效。

通过启动项目名称下的资源和资源组，您可以配置以下功能，以便在群集中管理系统供应。

- 扩展记帐 – 提供一个用于以任务或进程为基础记录消耗的灵活方式。 扩展记帐用于检查历史使用情况，以及估定将来工作量的容量要求。

- 控制 – 提供一个用于系统供应约束的机制。 可以防止进程、任务和项目消耗大量的指定系统供应。
- 公平共享调度 (FSS) – 可以根据工作量的重要性，控制可用 CPU 时间在工作量之间的分配。 工作量重要性是由分配给每个工作量的 CPU 时间的份额数表示的。 有关将 FSS 设置为缺省调度程序的命令行描述，请参见 `dispadmin` (1M)。 有关详细信息，请参阅`priocntl`(1)、`ps`(1) 和 `FSS`(7)。
- 池 – 可以根据应用程序的要求使用交互式应用程序的分区。 池可以用于对支持许多不同软件应用程序的服务器进行分区。 使用池可以使每个应用程序的响应更加容易预测。

确定项目配置的要求

在 Sun Cluster 环境中配置数据服务以使用由 Solaris 提供的控制之前，必须确定要如何在切换或失效转移过程中控制和跟踪资源。 配置新项目之前请考虑标识群集中的相关性。 例如，资源和资源组依赖于磁盘设备组。 使用通过 `scrgadm` (1M) 配置的 `nodelist`、`failback` 、`maximum_primaries` 和 `desired_primaries` 资源组特性来标识资源组的节点列表优先级。 有关资源组和磁盘设备组之间的节点列表依赖性的简要讨论，请参阅《*Sun Cluster* **数据服务规划和管理指南（适用于** *Solaris OS*）》中的“资源组和磁盘设备组之间的关系”。 有关特性的详细说明，请参见 `rg_properties` (5)。

使用通过 `scrgadm`(1M) 和 `scsetup`(1M) 配置的 `preferenced` 和 `failback` 特性来确定磁盘设备组节点列表的优先级。 有关过程信息，请参阅《*Sun Cluster* **系统管理指南（适用于** *Solaris OS*）》中的“管理磁盘设备组”的“如何更改磁盘设备特性”。 有关节点配置以及失效转移和可伸缩数据服务方式的概念信息，请参见第 15 页 “SunPlex 系统硬件和软件组件”。

如果要对所有的群集节点进行同样的配置，则会对主节点和辅助节点强制执行同样的使用限制。 对于所有节点配置文件中的所有应用程序，其项目配置参数不必相互一致。 与应用程序相关联的所有项目必须至少可以在该应用程序的所有潜在主控主机上由项目数据库访问。 假设应用程序 1 由 *phys-schost-1* 控制，但是可以切换或失效转移到 *phys-schost-2* 或 *phys-schost-3*。 那么必须在这三个节点 (*phys-schost-1* 、*phys-schost-2* 和 *phys-schost-3*) 上都能访问与应用程序 1 关联的项目。

注意 – 项目数据库信息可以是本地 `/etc/project` 数据库文件，也可以存储在 NIS 映射或 LDAP 目录服务中。

Solaris 环境允许对使用参数进行灵活配置，而且 Sun Cluster 的限制很少。 配置选项取决于站点的需求。 在对系统进行配置之前，请考虑以下部分中的一般原则。

设置每个进程的虚拟内存限制

设置 `process.max-address-space` 控制，以便限制以每个进程为基础的虚拟内存。 有关设置 `process.max-address-space` 值的详细信息，请参见 `rctladm`(1M)。

当对 Sun Cluster 使用管理控制时，请正确配置内存限制，以防止不必要的应用程序失效转移和应用程序的“乒乓”效果。 通常情况下：

- 请勿将内存限制设置得过低。

  当应用程序到达内存限制时，可能会发生失效转移。 由于数据库应用程序到达虚拟内存限制时可能出现意外后果，因此此原则对于数据库应用程序尤为重要。
- 请勿对主节点和辅助节点设置相同的内存限制。

  如果设置了相同的内存限制，当应用程序到达内存限制并进行失效转移，切换到具有相同内存限制的辅助节点时，可能导致乒乓效果。 请将辅助节点的内存限制设置得略高一些。 内存限制差别有助于防止乒乓情况的出现，并可以使系统管理员有时间根据需要对参数进行调整。
- 请务必为负载平衡使用资源管理内存限制。

  例如，使用内存限制可以防止错误的应用程序消耗过多的交换空间。

失效转移情况

可以对管理参数进行配置，以便项目配置 (`/etc/project`) 在正常的群集操作中和切换或失效转移情况下进行分配。

以下部分是示例情况。

- 前两部分，即“具有两个应用程序的双节点群集“和“具有三个应用程序的双节点群集“，显示了整个节点的失效转移情况。
- “仅限资源组的失效转移“部分说明了仅一个应用程序的失效转移操作。

在群集环境中，应用程序配置为资源的一部分，而资源配置为资源组 (RG) 的一部分。当发生失效转移时，资源组及其关联的应用程序将进行失效转移，切换到另一个节点。以下示例中没有明确显示资源。 假设每个资源仅有一个应用程序。

注意 – 按照 RGM 中设置的首选节点列表顺序进行失效转移。

以下示例的约束包括：

- 应用程序 1 (App-1) 配置于资源组 RG-1 中。
- 应用程序 2 (App-2) 配置于资源组 RG-2 中。
- 应用程序 3 (App-3) 配置于资源组 RG-3 中。

虽然分配的份额数相同，但是在失效转移后，分配给每个应用程序的 CPU 时间的比例将发生变化。 此比例取决于节点上运行的应用程序数，以及分配给每个活动的应用程序的份额数。

在以上情况下，假设采用了以下配置。

- 所有应用程序均在通用项目下配置。
- 每个资源都仅有一个应用程序。
- 这些应用程序是节点上仅有的活动进程。
- 群集中每个节点上的项目数据库配置相同。

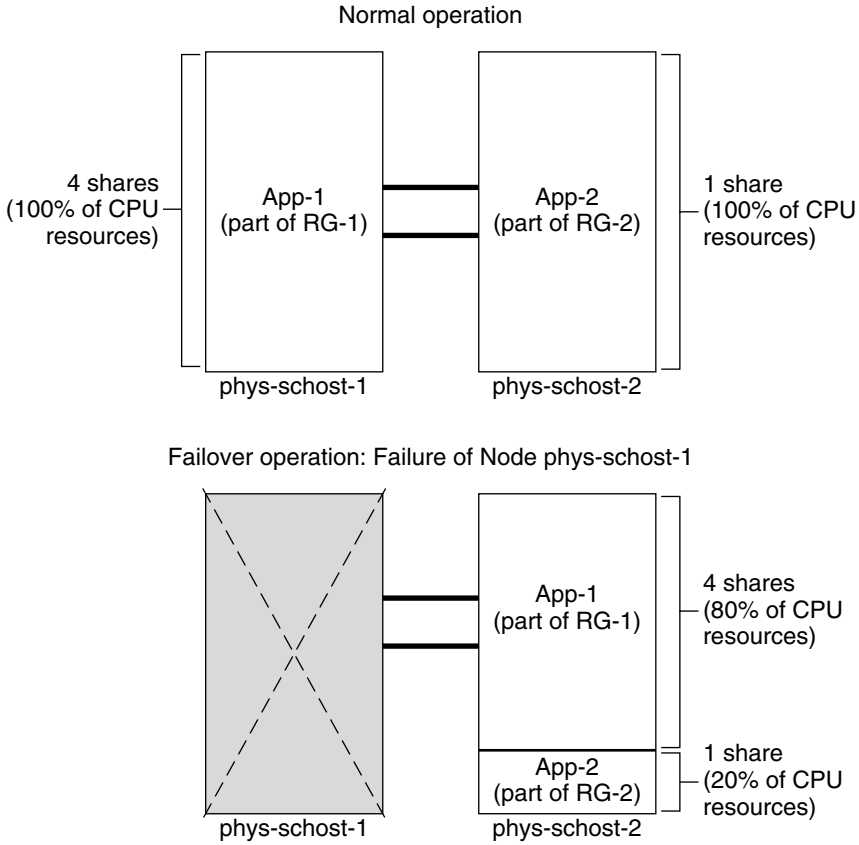
具有两个应用程序的双节点群集

您可以在双节点群集上配置两个应用程序，以确保每个物理主机（*phys-schost-1* 和 *phys-schost-2*）都充当一个应用程序的缺省主节点。 每个物理主机都充当另一个物理主机的辅助节点。 两个节点上的项目数据库文件中必须表示与应用程序 1 和应用程序 2 关联的所有项目。 当群集正常运行时，每个应用程序均运行在各自的缺省主控主机上，在其中管理设备为其分配了所有的 CPU 时间。

发生失效转移或切换后，两个应用程序均运行在一个节点上，在该节点上，应用程序会分配到配置文件中所指定的相应份额。 例如，`/etc/project` 文件中的相应项指定应用程序 1 分配到 4 份份额，应用程序 2 分配到 1 份份额。

```
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)
```

下图说明了此配置的正常操作和失效转移操作。 分配的份额数不变。 但是，根据分配给每个请求 CPU 时间的进程的份额数不同，每个应用程序可用的 CPU 时间比例可能发生变化。



具有三个应用程序的双节点群集

在具有三个应用程序的双节点群集上，您可以将一个物理主机 (*phys-schost-1*) 配置为一个应用程序的缺省主节点，而将第二个物理主机 (*phys-schost-2*) 配置为其余两个应用程序的缺省主节点。 假设在每个节点上采用以下示例项目数据库文件。 当发生失效转移或切换时，该项目数据库文件不发生变化。

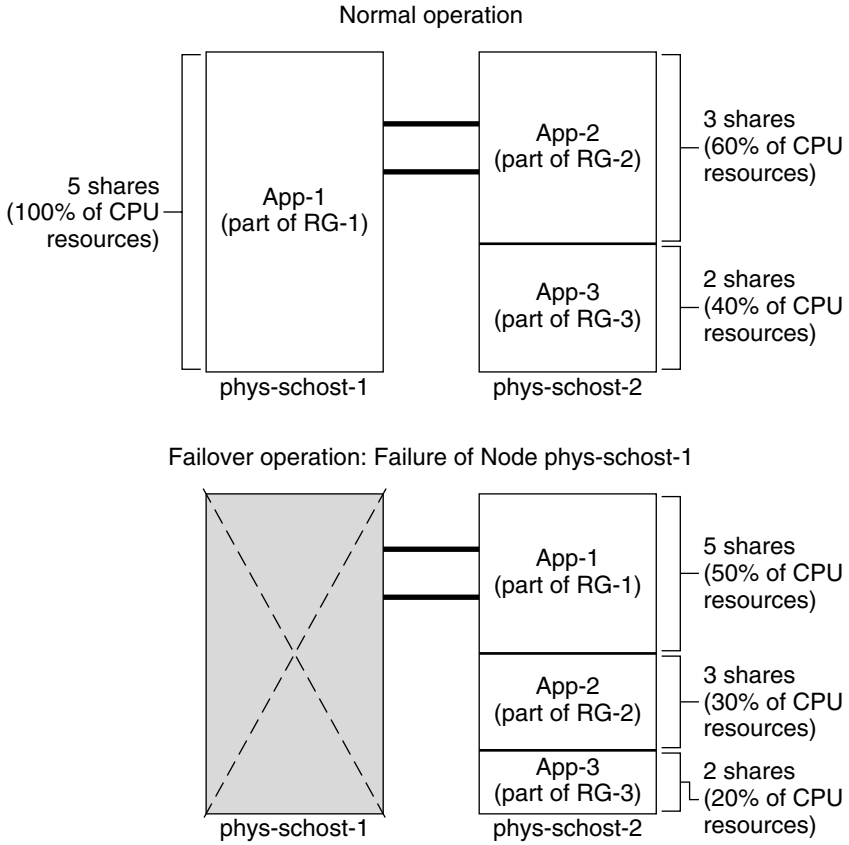
```
Prj_1:103:project for App-1:root::project.cpu-shares=(privileged,5,none)
Prj_2:104:project for App_2:root::project.cpu-shares=(privileged,3,none)
Prj_3:105:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

当群集正常运行时，应用程序 1 在其缺省主控主机（*phys-schost-1*）上分配到 5 份份额。 此份额数相当于 100% 的 CPU 时间，因为应用程序1 是该节点上唯一一个请求 CPU 时间的应用程序。 应用程序 2 和应用程序 3 分别在缺省主控主机（*phys-schost-2* ）上分配到 3 份和 2 份份额。 正常操作过程中，应用程序 2 将分配到 60% 的 CPU 时间，而应用程序 3 将分配到 40% 的 CPU 时间。

如果发生了失效转移或切换，且应用程序 1 切换到 *phys-schost-2*，则三个应用程序的份额都相同。 但是，CPU 资源的比例将根据项目数据库文件重新进行分配。

- 应用程序 1 拥有 5 份份额，分配到 50% 的 CPU。
- 应用程序 2 拥有 3 份份额，分配到 30% 的 CPU。
- 应用程序 3 拥有 2 份份额，分配到 20% 的 CPU。

下图说明了此配置的正常操作和失效转移操作。



仅限资源组的失效转移

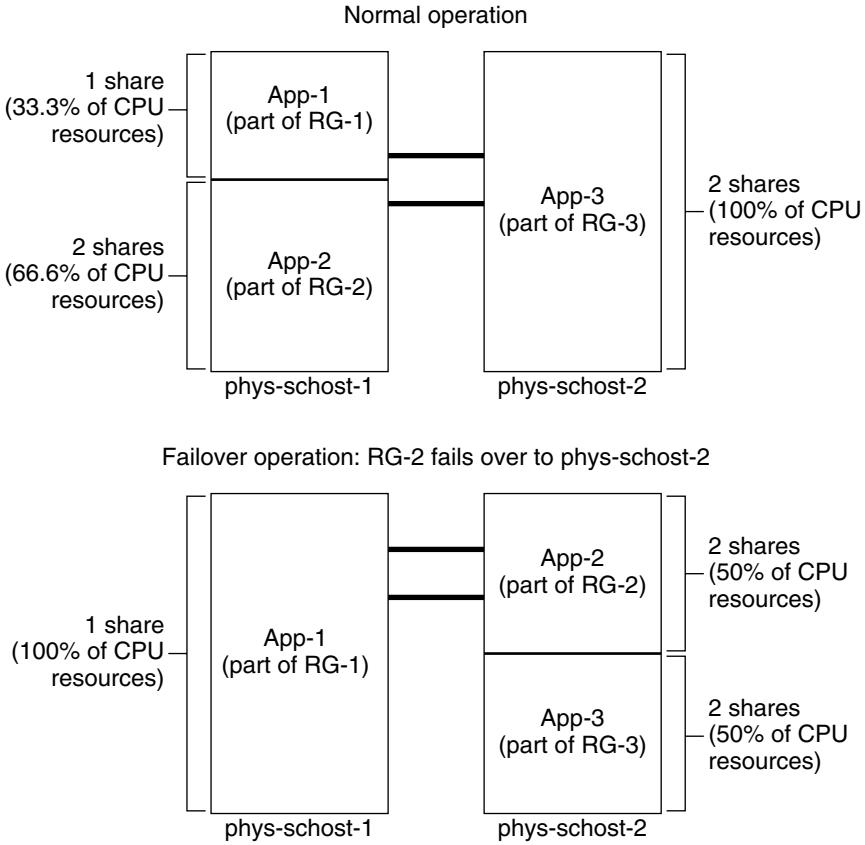
在多个资源组具有相同的缺省主控主机的配置中，资源组（及其关联的应用程序）可以进行失效转移或切换到辅助节点。 同时，缺省主控主机运行于群集中。

注意 – 失效转移过程中，发生失效转移的应用程序分配到的资源与辅助节点上的配置文件所指定的资源相同。 在此示例中，主节点和辅助节点上的项目数据库文件具有相同的配置。

例如，此样例配置文件指定应用程序 1 分配到 1 份份额，应用程序 2 分配到 2 份份额，应用程序 3 分配到 2 份份额。

```
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)
```

下图说明了此配置的正常操作和失效转移操作，其中包含应用程序 2 的 RG-2 进行失效转移，切换到 *phys-schost-2*。 请注意，分配的份额数不变。 但是，根据分配给每个请求 CPU 时间的应用程序的份额数不同，每个应用程序可用的 CPU 时间比例可能发生变化。



公共网络适配器和 IP Network Multipathing

客户机通过公共网络向群集提出数据请求。 每个群集节点通过一对公共网络适配器至少连接到一个公共网络。

Sun Cluster 中的 Solaris 网际协议 (IP) 网络多路径软件提供了一个基本机制，用于监视公共网络适配器，以及监视检测到故障时一个适配器到另一个适配器的失效转移 IP 地址。 每个群集节点有它自己的 IP Network Multipathing 配置，该配置可以与其它群集节点不同。

公共网络适配器组织为 *IP* **多路径组**（多路径组）。 每个多路径组具有一个或多个公共网络适配器。 多路径组中的每个适配器都可能处于活动状态，也可以配置备用接口，这些接口只有在发生失效转移时才会激活。 `in.mpathd` 多路径守护程序使用一个测试 IP 地址检测故障和检修。 如果多路径守护程序在某个适配器上检测到故障，则发生失效转移。 所有的网络访问均进行失效转移，从发生故障的适配器切换到多路径组中的另一个

功能适配器，从而维护该节点的公共网络连通性。 如果配置了备用接口，则守护程序会选择该备用接口。 否则，`in.mpathd` 将选择具有最小 IP 地址数目的接口。 由于失效转移发生在适配器接口级，像 TCP 这样的更高级别的连接则不受影响，仅在失效转移期间有短暂的瞬时延迟。 一旦 IP 地址的失效转移成功完成之后，就会发送未经请求的 ARP 广播。 通过这种方法保持了与远程客户机的连通性。

注意 – 由于 TCP 的拥塞恢复特性，TCP 端点可以在成功的失效转移后经受更长的延迟，同时一些段可能会在失效转移期间丢失，激活了 TCP 中的拥塞控制机制。

多路径组为逻辑主机名和共享地址资源提供了构件。 您也可以独立于逻辑主机名和共享地址资源来创建多路径组，以监视群集节点的公共网络连通性。 节点上相同的多路径组可以拥有任意数目的逻辑主机名或共享地址资源。 有关逻辑主机名和共享地址资源的详细信息，请参阅《*Sun Cluster* **数据服务规划和管理指南**》。

注意 – IP Network Multipathing 机制的设计着意于检测和屏蔽适配器故障。 该设计并非要通过让管理员使用 `ifconfig(1M)` 删除一个逻辑（或共享）IP 地址而得以恢复。 Sun Cluster 软件将逻辑 IP 地址和共享 IP 地址视为由 RGM 管理的资源。 管理员添加或删除 IP 地址的正确方法是使用 `scrgadm`(1M) 来修改包含资源的资源组。

有关 IP 网络多路径的 Solaris 实现的详细信息，请参见群集中安装的 Solaris 操作环境的相应文档。

操作环境发行版	有关说明，请转到...
Solaris 8 操作环境	*IP Network Multipathing Administration Guide*
Solaris 9 操作环境	*System Administration Guide: IP Services* 中的 “IP Network Multipathing Topics”

SPARC: 动态重新配置支持

Sun Cluster 3.1 4/04 对动态重新配置 (DR) 软件功能的支持正在进一步开发过程中。 本部分说明了 Sun Cluster 3.1 4/04 对 DR 功能的支持所涉及的一些概念和注意事项。

请注意：相关文档中适用于 Solaris DR 功能的所有要求、步骤和限制同样适用于 Sun Cluster DR 支持（唯一的区别是操作环境静态操作）。 因此，在使用 Sun Cluster 软件的 DR 功能**之前**，请查阅 Solaris DR 功能的有关文档。 您特别要注意那些在执行 DR 分离操作时将影响非网络 IO 设备的问题。 可以从 `http://docs.sun.com` 下载 *Sun Enterprise 10000 Dynamic Reconfiguration User Guide* 和 *Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual*（包含在 *Solaris 8 on Sun Hardware Collection* 或 *Solaris 9 on Sun Hardware Collection* 中）。

SPARC: 动态重新配置一般描述

DR 功能允许在运行的系统中进行各项操作，如删除系统硬件。 DR 进程的设计旨在确保系统操作的连续性，而不必使系统停机或中断群集的使用。

DR 操作在板级别进行。 因此，DR 操作会影响板上的所有组件。 每块板可以包含多个组件，如 CPU、内存以及用于磁盘驱动器、磁带机和网络连接的外部接口。

删除包含活动组件的板将导致系统错误。 删除板之前，DR 子系统对其它子系统（如 Sun Cluster）进行查询，以确定是否使用该板中的组件。 如果 DR 子系统发现板正在使用中，则不执行 DR 删除板操作。 因此，发布 DR 删除板操作始终是安全的，因为 DR 子系统能够拒绝对包含活动组件的板的操作。

DR 增加板操作也始终是安全的。 系统自动将新增加到板上的 CPU 和内存投入使用。 不过，系统管理员必须手动配置群集，然后才可随意使用新添加的板上的组件。

注意 – DR 子系统包含若干个级别。 如果较低的级别报错，则较高的级别同样也会报错。 但是，较低的级别报告具体错误，而较高的级别报告“未知错误”。 系统管理员应该忽略较高的级别所报告的“未知错误”。

下面各节说明了对于不同设备类型的 DR 考虑事项。

SPARC: 有关 CPU 设备的 DR 群集的注意事项

因为存在 CPU 设备，所以 Sun Cluster 软件将不拒绝 DR 删除板操作。

当 DR 增加板操作成功后，增加的板上的 CPU 设备会自动并入系统操作中。

SPARC: 有关内存的 DR 群集注意事项

基于 DR 目的，有两种内存需要加以考虑。 这两种内存仅在用法上有所不同。 对于这两种内存而言，实际的硬件是相同的。

操作系统所用的内存称作内核内存箱。 Sun Cluster 软件不支持对包含内核内存箱的板执行删除操作，并将拒绝执行这样的操作。 当 DR 删除板操作针对除内核内存箱以外的内存时，Sun Cluster 将不拒绝此操作。

当针对内存的 DR 增加板操作成功后，增加的板上的内存将自动并入系统操作。

SPARC: 有关磁盘和磁带机的 DR 群集注意事项

Sun Cluster 拒绝在主节点中的活动驱动器上进行 DR 删除板操作。 可以对主节点中非活动状态的驱动器和辅助节点的任何驱动器执行 DR 删除板操作。 DR 操作之后，对群集数据的访问象以前一样继续。

注意 – Sun Cluster 拒绝进行影响仲裁设备可用性的 DR 操作。 有关定额设备的考虑事项以及对其执行 DR 操作的过程，请参阅第 65 页 “SPARC: 仲裁设备的 DR 群集注意事项”。

有关如何执行这些操作的详细说明，请参阅《*Sun Cluster* **系统管理指南**》。

SPARC: 仲裁设备的 DR 群集注意事项

如果 DR 删除板操作针对的板包含一个配置为仲裁设备的接口，则 Sun Cluster 将拒绝执行此操作，并标识出可能受此操作影响的仲裁设备。 只有将仲裁设备进行处理使之不再是仲裁设备之后，您才能对其执行 DR 删除板操作。

有关如何执行这些操作的详细说明，请参阅《*Sun Cluster* **系统管理指南**》。

SPARC: 群集互连接口的 DR 群集注意事项

如果 DR 删除板操作针对的板包含一个活动的群集互连接口，则 Sun Cluster 将拒绝执行此操作，并标识出可能受此操作影响的接口。 在 DR 操作成功之前，必须使用 Sun Cluster 管理工具来禁用活动的接口。（另请参见下面的注意事项）。

有关如何执行这些操作的详细说明，请参阅《*Sun Cluster* **系统管理指南**》。

Caution – Sun Cluster 要求每个节点与群集中的其它节点之间至少有一个有效路径。 如果某个专用互连接口支持到任何群集节点的最后一条路径，则请勿禁用它。

SPARC: 公共网络接口的 DR 群集注意事项

如果 DR 删除板操作针对的板包含一个活动的公共网络接口，则 Sun Cluster 将拒绝执行此操作，并标识出可能受此操作影响的接口。 删除包含活动的网络接口的板之前，必须先使用 `if_mpadm`(1M) 命令将该接口上的所有通信切换到多路径组中的另一个功能接口上。

Caution – 如果在已禁用的网络适配器上执行 DR 删除操作时，其余网络适配器发生故障，可用性将受到影响。 另一个适配器在执行 DR 操作期间无法进行失效转移。

有关如何在公共网络接口上执行 DR 删除操作的详细说明，请参阅《*Sun Cluster* **系统管理指南**》。

第 4 章

常见问题

INDEXTERM-336

本章包含关于 SunPlex 系统的最常见问题的解答。 问题是按主题编排的。

高可用性 FAQ

- **到底什么是高可用系统？**

  SunPlex 系统将高可用性 (HA) 定义为群集使应用程序保持活动状态并运行（即使发生通常会使服务器系统不可用的故障）的能力。

- **群集是通过什么样的进程提供高可用性的？**

  通过一个称为失效转移的进程，群集框架提供高可用性的环境。 失效转移就是一系统列由群集执行的步骤，它将数据服务资源从一个故障节点转移到群集上另一个可操作节点。

- **失效转移与可伸缩数据服务间有什么不同？**

  有两种高可用性数据服务类型：失效转移数据服务和可伸缩数据服务。

  失效转移数据服务每次只能在群集中的一个主节点上运行应用程序。 其它节点上可能运行其它应用程序，但每个应用程序只能运行在单一节点上。 如果主节点发生故障，正在故障节点上运行的应用程序进行失效转移，切换到另一个节点并继续运行。

  可伸缩服务将一个应用程序扩展到多个节点之上来创建一个单独的逻辑服务。 可伸缩服务平衡它们在其上运行的整个群集中的节点和服务器的数目。

  对于每个应用程序，一个节点具有一个至群集的物理接口。 这个节点被称作全局接口 (GIF) 节点。 群集中可以有多个 GIF 节点。 每个 GIF 节点都有一个或多个逻辑接口，可伸缩服务可使用这些接口。 这些逻辑接口被称作**全局接口**。 每个 GIF 节点都具有一个全局接口，用来接收针对特定应用程序的所有请求。GIN 还会将这些请求

分发给运行应用程序服务器的多个节点上。 如果 GIF 发生故障，则全局接口将失效转移到一个仍正常工作的节点。

如果某个正在运行应用程序的节点发生故障，该应用程序将在其它节点上继续运行，只是性能有所下降，直到该故障节点返回该群集为止。

文件系统 FAQ

- **可否将一个或多个群集节点作为高度可用的 NFS 服务器运行，而将其它群集节点当作客户机？**

  不可以，不要进行回送装载。

- **可否将群集文件系统用于不受 Resource Group Manager 控制的应用程序？**

  是的。 然而，由于不受 RGM 的控制，当运行这些应用程序的节点发生故障时，需手动重新启动这些应用程序。

- **所有群集文件系统是否都必须在 /global 目录下具有一个装载点？**

  不是。但是，将多个群集文件系统放置在同一装载点下（如 `/global`）能够对这些文件系统进行更好的组织和管理。

- **使用群集文件系统和导出 NFS 文件系统有哪些不同？**

  有以下几点不同：

  1. 群集文件系统支持全局设备。 NFS 不支持对设备的远程访问。
  2. 群集文件系统有一个全局名称空间。 只需要一个定位命令。 使用 NFS 时，必须在每个节点上定位文件系统。
  3. 与 NFS 相比，群集文件系统从高速缓存访问文件的情况更多。 例如，多个节点同时访问一个文件，以执行读、写、文件锁定、异步 I/O 等操作。
  4. 群集文件系统是为了利用能够提供远程 DMA 和零拷贝功能的快速群集互连而建立的。
  5. 如果您更改了群集文件系统中某个文件的特性（例如，使用 `chmod(1M)`），所做的更改会立即反映到所有的节点上。 使用导出的 NFS 文件系统，这可能会花费更长的时间。

- **文件系统 /global/.devices/node@<nodeID> 出现在我的群集节点上。 可否使用这个文件系统来存储要作为高度可用数据和全局数据的那些数据？**

  这些文件系统存储全局设备名称空间。 它们不可以通用。 如果是全局文件系统，不能以全局的方式对其进行访问，每个节点只能访问自己的全局设备名称空间。 如果节点处于关闭状态，则其它节点无法访问处于关闭状态的该节点的名称空间。 这些文件系统不具备高可用性。 它们不适合用于存储需全局访问或高度可用的数据。

卷管理 FAQ

- **需要镜像所有磁盘设备吗？**

  必须镜像被视为具有高可用性的磁盘设备，或者使用 RAID-5 硬件。 所有数据服务应该要么使用高可用磁盘设备，要么使用定位到高可用磁盘设备上的群集文件系统。这样的配置可以容许单独磁盘故障。

- **可否将一个卷管理器用于本地磁盘（引导磁盘），而将另一个卷管理器用于多主机磁盘？**

  SPARC: 此配置支持 Solaris Volume Manager 软件管理本地磁盘，支持 VERITAS Volume Manager 管理多主机磁盘。 不支持其它任何组合方式。

  x86: 不，不支持这种配置，在基于 x86 的群集中只支持 Solaris Volume Manager。

数据服务 FAQ

- **可以获得哪些 SunPlex 数据服务？**

  Sun Cluster Release Notes 中列出了所支持的数据服务。

- **SunPlex 数据服务支持哪些应用程序版本？**

  Sun Cluster Release Notes 中列出了所支持的应用程序版本。

- **我可以记下自己的数据服务吗？**

  是的。 有关详细信息，请参阅《*Sun Cluster* **数据服务开发者指南**》和 Data Service Development Library API 所附带的 Data Service Enabling Technologies 文档。

- **创建网络资源时，我应该指定数字 IP 地址还是主机名？**

  指定网络资源的首选方法是使用 UNIX 主机名，而非使用数字 IP 地址。

- **创建网络资源时，使用逻辑主机名（一个 LogicalHostname 资源）与使用共享地址（一个 SharedAddress 资源）有什么不同？**

  除了 Sun Cluster HA for NFS 之外，只要文档要求在 `Failover` 模式资源组中使用 `LogicalHostname` 资源，`SharedAddress` 资源或 `LogicalHostname` 资源就可以交替地使用。 使用 `SharedAddress` 资源会造成一些额外的开销，因为群集联网软件是为 `SharedAddress` 而配置的，而不是为 `LogicalHostname` 而配置的。

  使用 `SharedAddress` 的优点在以下情况下就可体现出来：您要配置可伸缩和失效转移两种数据服务，并想让客户能够使用相同的主机名访问这两种服务。 在这种情况下，`SharedAddress` 资源以及失效转移应用程序资源包括在一个资源组中，而可伸缩服务资源包括在一个独立的资源组中并被配置为使用 `SharedAddress`。 然后，可伸缩服务和失效转移服务都可以使用在 `SharedAddress` 资源中配置的同一主机名/地址集。

公共网络 FAQ

- **SunPlex 系统支持哪些公共网络适配器？**

  目前，SunPlex 系统支持以太网（10/100BASE-T 和 1000BASE-SX Gb）公共网络适配器。 因为新的接口可能会在将来得到支持，所以请向 Sun 销售代表咨询以获取最当前信息。

- **在失效转移中 MAC 地址起什么作用？**

  当失效转移发生时，生成新的地址解析协议 (ARP) 软件包并进行广播。 这些 ARP 软件包包含新的 MAC 地址（节点失效转移到的新的物理适配器的地址）和旧的 IP 地址。 网络中的其它计算机收到其中一个软件包之后，将刷新 ARP 高速缓存中的旧 MAC-IP 映射，然后使用新的映射。

- **SunPlex 系统是否支持设置 `local-mac-address?=true`？**

  是的。 事实上，IP 网络多路径要求必须将 `local-mac-address?` 设置为 `true`。

  您可以在基于 SPARC 的群集中的 OpenBoot PROM `ok` 提示符处使用 `eeprom`(1M) 来设置 `local-mac-address?`，也可以在基于 x86 的群集中，在 BIOS 引导之后选择运行 SCSI 实用程序来设置 `local-mac-address?`。

- **当 IP Network Multipathing在适配器之间执行切换时，将会有多久的延迟？**

  延迟可能持续几分钟。 这是因为 IP Network Multipathing切换完成后，还需要发送一个未经请求的 ARP。 但是，不保证客户机与群集之间的路由器将使用该未经请求的 ARP。 因此，直到路由器的此 IP 地址的 ARP 高速缓存项目超时，才有可能使用无效 MAC 地址。

- **检测网络适配器的故障的速度有多快？**

  缺省的故障检测时间是 10 秒钟。 算法尽量与故障检测时间相符，但实际的检测时间取决于网络负载。

群集成员 FAQ

- **所有的群集成员都需要有相同的 root 用户口令吗？**

  不要求让每个群集成员使用相同的 root 用户口令。 但是，您可以通过在所有的节点上使用相同的 root 用户口令来简化该群集的管理。

- **节点引导的顺序有重要意义吗？**

  在大多数情况下没有意义。但是，引导顺序对于防止失忆是很重要的（有关失忆的详细信息，请参考第 39 页“仲裁和仲裁设备”）。 例如，如果节点 2 是仲裁设备的属主而节点 1 停机，并且您此时将节点 2 停机，那么您在启动节点 1 之前必须先启动节点 2。 这可避免意外使用过时的群集配置信息启动节点。

- **是否需要在群集节点中镜像本地磁盘？**

是的。 尽管这一镜像并不是一种要求，但是镜像群集节点磁盘可防止非镜像磁盘故障使节点停机。 镜像群集节点本地磁盘的缺点是，将耗费更多的系统管理开销。

- **群集成员的备份是指什么？**

  您可以对一个群集使用多种备份方法。 一种方法是将一个节点作为备份节点，连接一个磁带机/库。 然后使用群集文件系统来备份数据。 不要将此节点连接到共享磁盘上。

  有关备份和恢复过程的其它信息，请参阅《*Sun Cluster* **系统管理指南**》。

- **节点何时可以作为辅助节点使用？**

  重新引导后，当节点显示登录提示时，节点就可以作为辅助节点使用了。

群集存储器 FAQ

- **多主机存储器为什么具有高可用性？**

  多主机存储器之所以具有高可用性，是因为它在丢失单个磁盘的数据的情况下仍能借助镜像（或者基于硬件的 RAID-5 控制器）而幸免于难。 因为多主机存储器设备有不止一个主机连接，所以它也可以经受它所连接的单一节点的丢失。 此外，从每个节点到附加存储器的冗余路径为主机总线适配器、电缆或磁盘控制器的故障提供了容错。

群集互连 FAQ

- **SunPlex 系统支持什么样的群集互连？**

  目前，SunPlex 系统在基于 SPARC 和基于 x86 的群集中支持以太网（100BASE-T 快速以太网和 1000BASE-SX Gb）群集互连。 SunPlex 系统只在基于 SPARC 的群集中支持 SCI 网络接口群集互连。

- **“电缆”与传输“路径”有什么不同？**

  群集传输电缆是使用传输适配器和交换机配置的。 电缆在组件对组件的基础上将适配器与交换器连接在一起。 群集拓扑管理器使用可用的电缆来建立节点间的端到端传输路径。 电缆不直接与传输路径相对应。

  电缆可由管理员静态“启用”和“禁用”。 电缆可处于一种“状态”（启用或禁用），但并非一种“状况”。 如果电缆处于禁用状态，就如同电缆没有进行配置一样。 禁用的电缆不可用作传输路径。 不对它们进行探测，因此不可能知道它们的状况。 使用 `scconf -p` 可以查看电缆的状态。

  传输路径由群集拓扑管理器动态建立。 传输路径的“状况”由拓扑管理器来确定。 路径可处于“联机”或“脱机”状况。 可以使用 `scstat(1M)` 查看传输路径的状况。

以下面的群集为例，该群集有两个节点，通过四条电缆进行连接。

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

这四条电缆可能形成两条传输路径。

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

客户机系统 FAQ

- **使用群集时是否需要考虑任何特殊的客户机需要或限制？**

  就像连接到任何其它服务器上一样，客户机系统可连接到群集。 在某些情况下，根据具体的数据服务应用程序，您可能需要安装客户端软件或执行其它配置更改，以使客户机可以连接到该数据服务应用程序。 有关客户端配置要求的详细信息，请参阅《*Sun Cluster* **数据服务规划和管理指南**》中的相关章节。

管理控制台 FAQ

- **SunPlex 系统是否需要管理控制台？**

  是的。

- **管理控制台必须专用于该群集吗？它可以用于其它任务吗？**

  SunPlex 系统不需要专用的管理控制台，但使用它有以下优点：

  - 通过在同一机器上给控制台和管理工具分组来启用集中化的群集管理
  - 可能会使硬件服务供应商更快地解决问题

- **是否需要管理控制台位于群集自身的"旁边"（例如，在同一个房间中）？**

  请向硬件服务供应商咨询。 供应商可能会要求控制台位于群集的近旁。 使控制台处在同一房间内没有技术上的原因。

- **是否只要所有距离要求也首先得到满足，管理控制台就可以服务于多个群集？**

  是的。 可以从一个单独的管理控制台控制多个群集。 也可以在群集间共享一个单独的终端集中器。

终端集中器和系统服务处理器 FAQ

- **SunPlex 系统需要终端集中器吗？**

  Sun Cluster 3.0 之后的所有软件发行版本均不需要终端集中器来运行。 Sun Cluster 2.2 要求一个终端集中器来进行故障防护；后续版本与之不同，不再依赖于终端集中器。

- **我知道大多数 SunPlex 服务器都使用终端集中器，而 Sun Enterprise E10000 server 却不使用。 为什么呢？**

  对于大多数服务器来讲，终端集中器实际上是一个串行到以太网的转换器。 其控制台端口是一个串行端口。 Sun Enterprise E10000 server 没有串行控制台。 系统服务处理器 (SSP) 就是其控制台，它或者使用以太网端口，或者使用 `jtag` 端口。 对于 Sun Enterprise E10000 server，请始终将 SSP 用作控制台。

- **使用终端集中器有什么益处？**

  使用终端集中器提供了对每个节点的控制台级别的访问，可以从网络上任意位置的远程工作站访问节点，节点可以位于基于 SPARC 节点的 OpenBoot PROM (OBP) 上，也可以位于基于 x86 的节点上的引导子系统上。

- **如果使用 Sun 不支持的终端集中器，需要了解哪些信息来确定我要使用的终端集中器是否符合要求？**

  Sun 所支持的终端集中器与其它控制台设备之间的主要差别，是 Sun 终端集中器有特殊的固件来防止终端集中器在控制台引导时向控制台发送中断。 注意，如果您有一个控制台设备，可以发送中断或发送可能被解释为发给控制台的中断的信号，那么该控制台设备将关闭该节点。

- **是否可以不重新引导而释放一个 Sun 所支持的终端集中器上的锁定端口？**

  是的。 记下需要重置的端口号并键入以下命令：

  ```
  telnet tc
  Enter Annex port name or number: cli
  annex: su -
  annex# admin
  admin : reset port_number
  admin : quit
  annex# hangup
  #
  ```

  有关配置和管理 Sun 所支持的终端集中器的详细信息，请参阅《*Sun Cluster* **系统管理指南**》。

- **终端集中器本身发生故障怎么办？ 我必须有备用终端集中器吗？**

  不必。 如果终端集中器发生故障，您不会丢失任何群集可用性。 但在集中器恢复工作之前，您将无法连接到节点控制台。

- **使用终端集中器时，其安全性如何？**

  通常，终端集中器连接到系统管理员使用的一个小型网络，而不连接到用于其它客户访问的网络。 您可以通过限制对该特定网络的访问来控制安全性。

- **如何使用磁带机或磁盘驱动器进行动态重新配置？**

- 确定磁盘驱动器或磁带机是否是活动设备组的一部分。 如果该驱动器不是活动设备组的组成部分，您就可以对其执行 DR 删除操作。
- 如果 DR 删除板操作将影响活动的磁盘驱动器或磁带机，则系统将拒绝执行该操作并且标识出可能会受该操作影响的驱动器。 如果驱动器是活动设备组的组成部分，请转到第 64 页 “SPARC: 有关磁盘和磁带机的 DR 群集注意事项”。
- 确定驱动器是主节点的组件还是辅助节点的组件。 如果驱动器是辅助节点的组件，您就可以对其执行 DR 删除操作。
- 如果驱动器是主节点的组件，您就必须先将主节点和辅助节点对调，然后才对该设备执行 DR 删除操作。

Caution – 如果当前的主节点在您正对辅助节点执行 DR 操作时出现故障，则会影响群集的可用性。 在提供辅助节点之前，该主节点将无法进行失效转移。

索引

L

M

N

O

P

R

S

U

V

板

保

系

项

选

引

应

硬

终

仲

主

专

装

资

组