Sun Cluster 3.0 U1 概念

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto,, CA 94303-4900
U.S.A. 650-960-1300

部件号码 816-1956-10
2001 年 8 月, Revision A

Adobe PostScript™

目录

前言

Sun™ Cluster 3.0 U1 概念包含关于 SunPlex™ 系统的概念与参考信息。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或售前指南。在阅读此文档前，您应该已经确定了系统需求并购买了相应的设备和软件。

要理解本书中讲述的概念，应该具备 Solaris™ 操作系统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。

印刷惯例的含义

字体或符号	含义	示例
`AaBbCc123`	命令、文件和目录的名称；计算机屏幕输出	编辑您的 `.login` 文件。 使用 `ls -a` 列出全部文件。 `% You have mail`。
`AaBbCc123`	您输入的内容，与计算机屏幕输出相对照。	`% su` 口令：

字体或符号	含义	示例
AaBbCc123	书名、新的词汇或术语、要强调的词	请阅读用户指南中的第六章。 这些被称为 *class* 选项。 您必须是超级用户才能执行此操作。
	命令行变量；用一个实际的名称或值替换	要删除文件，请输入 `rm` *filename*。

Shell 提示符

Shell	提示符
C shell	*machine_name*%
C shell 超级用户	*machine_name*#
Bourne shell 和 Korn shell	$
Bourne shell 和 Korn shell 超级用户	#

相关文档

主题	标题	部件号
安装	《*Sun Cluster 3.0 U1* 安装指南》	816-1961-10
硬件	《*Sun Cluster 3.0 U1 Hardware Guide*》	806-7070
数据服务	《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》	806-7071

主题	标题	部件号
API 开发	《*Sun Cluster 3.0 U1 Data Services Developer's Guide*》	806-7072
管理	《*Sun Cluster 3.0 U1* 系统管理指南》	816-1968-10
错误消息和问题分解	《*Sun Cluster 3.0 U1 Error Messages Manual*》	806-7076
发行说明	《*Sun Cluster 3.0 U1* 发行说明》	816-1973-10

订购 Sun 文档

Fatbrain.com 是一家 Internet 上的专业书店，供应 Sun Microsystems, Inc. 的精选产品文档。 要获取文档列表及了解 如何订购，请访问 Fatbrain.com 站点的 Sun 文档中心：

`http://www1.fatbrain.com/documentation/sun`

联机访问 Sun 文档

`docs.sun.com`SM 网站使您能够在 Web 上访问 Sun 技术文档。在下面的站点，您可以浏览 `docs.sun.com` 分类文档或搜索特定的书名或主题：

`http://docs.sun.com`

获取帮助

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下面的信息：

- 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 8）
- Sun Cluster 软件的发行版本号（例如，Sun Cluster 3.0）

使用下面的命令收集系统上每个节点的有关信息，以提供给服务供应商：

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`	显示处理器的有关信息
`showrev --p`	报告已安装了哪些修补程序
`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

也请提供 `/var/adm/messages` 文件的内容。

简介与概述

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

《Sun Cluster 3.0 U1 概念》提供了 SunPlex 文档的主要读者所需的概念信息。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 产品当前不包括的应用程序开发故障转移和可伸缩服务的应用程序开发者

本书连同 SunPlex 文档集的其他部分，共同介绍了 SunPlex 系统的全貌。

本章

- 介绍 SunPlex 系统并作了高度概括
- 描述了 SunPlex 读者的几个观点
- 标识了在使用 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用于完成这些任务的过程的文档对应起来

SunPlex 系统简介

SunPlex 系统能将 Solaris 操作环境扩展为群集操作系统。群集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）的单一客户机视图。

每个群集节点都是运行其自己的进程的一个独立服务器。这些进程可以彼此通信，协同一致地向用户提供应用程序、系统资源和数据，此群集在网络客户看来无异就是一个单一的系统。

与传统的单一服务器系统相比，群集具有几大优势。其中包括对故障转移和可伸缩服务的支持、支持模块化增长的能力，以及优越于传统硬件容错系统的低进入价位。

SunPlex 系统的目标是：

- 减少或消除由软件或硬件故障引起的系统停机时间
- 无论通常会引起单服务器系统停机的故障属于什么类型，都能确保数据和应用程序对最终用户的可用性
- 通过向群集添加节点，使服务性能随着处理器的添加而扩展，从而增大应用程序吞吐量
- 不必关掉整个群集便可执行维护，提供增强的系统可用性。

高可用性与容错性

根据设计，SunPlex 系统是一种高可用性 (HA) 系统（即对数据和应用程序提供几乎不间断的访问的系统）。

相比之下，容错硬件系统虽然也能提供对数据和应用程序的持续访问，但它需使用专用硬件，故而其成本更为昂贵。另外，容错系统通常不能解决软件故障。

SunPlex 系统通过硬件与软件的结合实现了高可用性。冗余的群集互连、存储器和公共网络防止了单点故障的发生。群集软件不间断地监视成员节点的运行状况并能阻止故障节点加入群集，从而可防止数据遭到破坏。同时，群集监视服务和相关的系统资源，并在出现故障时进行故障转移或重新启动服务。

有关高可用性的问题及解答，请参见第63页的「关于高可用性的常见问题」。

SunPlex 系统中的故障转移和可伸缩性

SunPlex 系统使您可以执行故障转移或可伸缩服务。一般来说，故障转移服务只提供高可用性（冗余），而可伸缩服务除了提供高可用性之外，还具有更优的性能。单一群集既可以支持故障转移服务，也可以支持可伸缩服务。

故障转移服务

故障转移就是群集自动将服务从一个故障主节点重新定位到指定的辅助节点的过程。通过故障转移功能，Sun Cluster 软件可提供高可用性。

当故障转移发生时，客户可能会看到一个短暂的服务中断，并可能需要在故障转移结束后重新连接。不过，客户并不知道提供该服务的物理服务器发生了更改。

可伸缩服务

当当故障恢复采用冗余技术时，可伸缩性提供持续的响应时间或吞吐量，而不用去关心负荷。可伸缩服务利用群集中的多个节点来同时运行一个应用程序，从而增强了性能。在可伸缩配置中，群集中的每一个节点都可以提供数据和处理客户请求。

有关故障转移和可伸缩服务的详细信息，请参见第48页的「数据服务」。

关于 SunPlex 系统的三个观点

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

在硬件维护人员看来，SunPlex 系统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20页的「群集节点」
- 第22页的「多主机磁盘」
- 第23页的「局部磁盘」
- 第24页的「群集互连」
- 第25页的「公共网络接口」
- 第25页的「客户机系统」
- 第26页的「管理控制台」
- 第25页的「控制台访问设备」
- 第26页的「群集对拓扑」
- 第28页的「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 3.0 U1 Hardware Guide*

系统管理员观点

在系统管理员看来，SunPlex 系统就像是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：

- 与 Solaris 软件集成在一起的专用群集软件，负责监视群集节点之间的连通性
- 专用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
- 专用群集软件能使所有的节点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
- 专用群集软件可使每个节点都能访问所有的文件，而且这些文件看起来都像是从本地连接到该节点一样。

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
- 修改配置以改善性能

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33页的「管理界面」
- 第33页的「高可用性框架」
- 第36页的「全局设备」
- 第37页的「磁盘设备组」
- 第39页的「全局名称空间」
- 第41页的「群集文件系统」
- 第43页的「定额和定额设备」
- 第47页的「卷管理器」
- 第48页的「数据服务」

- 第59页的「资源、资源组和资源类型」
- 第61页的「公共网络管理 (PNM) 和网络适配器故障转移 (NAFO)」
- 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 3.0 U1* 安装指南
- *Sun Cluster 3.0 U1* 系统管理指南
- *Sun Cluster 3.0 U1 Error Messages Manual*

应用程序编程人员观点

SunPlex 系统为诸如 Oracle、NFS、DNS、iPlanet Web Server、Apache Web Server 和 Netscape Directory Server 的应用程序提供 数据服务。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 软件提供启动、停止和监视这些应用程序的配置文件和管理方法。如果您需要创建新的故障转移或可伸缩服务，可以使用 SunPlex 应用程序编程接口 (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，以使其应用程序能够在群集上作为数据服务运行。

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，由此确定能否让其作为故障转移或可伸缩数据服务来运行。
- Sun Cluster API、DSET API 和“通用”数据服务。编程人员需要确定他们最适合使用哪种工具来编写程序或脚本来配置用于群集环境的应用程序。

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48页的「数据服务」
- 第59页的「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 3.0 U1 Data Services Developer's Guide*
- *Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*

SunPlex 系统任务

所有的 SunPlex 系统任务都需要某些概念背景。下表提供了这些任务和介绍任务步骤的文档的概括性的视图。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 任务对应关系：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 3.0 U1 Hardware Guide*》
在群集上安装 Solaris 软件	《*Sun Cluster 3.0 U1* 安装指南》
安装 Sun™ 管理中心软件	《*Sun Cluster 3.0 U1* 安装指南》
安装并配置 Sun Cluster 软件	《*Sun Cluster 3.0 U1* 安装指南》
安装并配置卷管理软件	《*Sun Cluster 3.0 U1* 安装指南》 您的卷管理文档
安装和配置 Sun Cluster 数据服务	《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》
维护群集硬件	《*Sun Cluster 3.0 U1 Hardware Guide*》
管理 Sun Cluster软件	《*Sun Cluster 3.0 U1* 系统管理指南》
管理卷管理软件	《*Sun Cluster 3.0 U1* 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

表 1–1 任务对应关系：将用户任务与文档对应起来 续下

要完成的任务	需要使用的文档
问题鉴定与建议的用户操作	《*Sun Cluster 3.0 U1 Error Messages Manual*》
创建新的数据服务	《*Sun Cluster 3.0 U1 Data Services Developer's Guide*》

关键概念 – 硬件服务供应商

本章说明与 SunPlex 系统配置的硬件组件相关的关键概念。

SunPlex 系统硬件组件

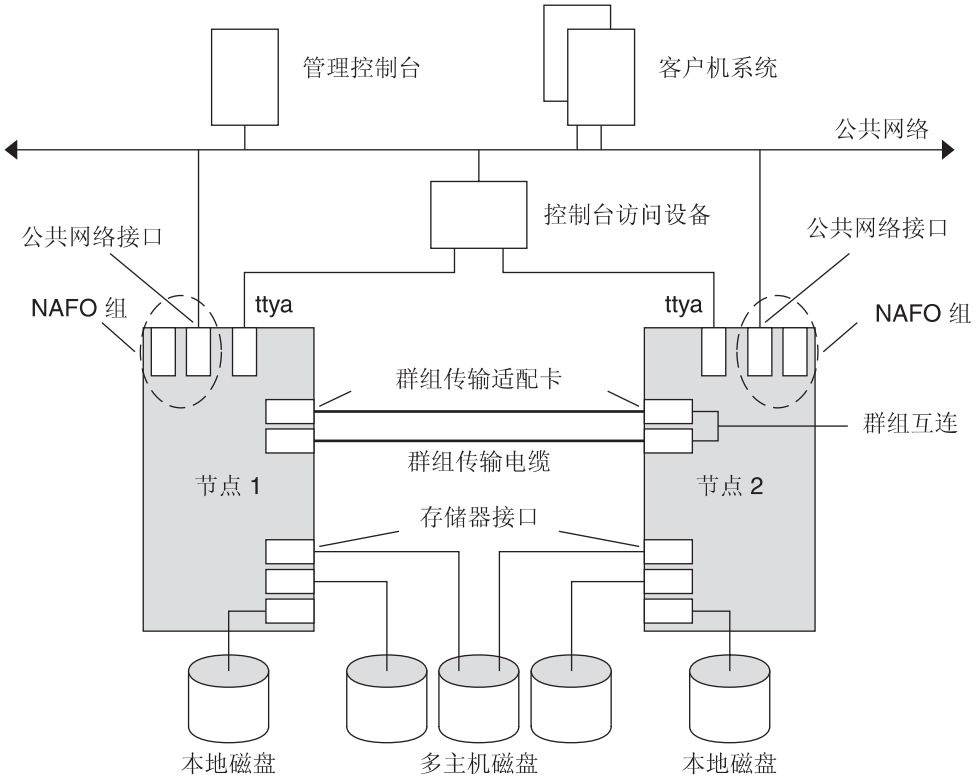
本章中的信息主要面向硬件服务供应商。在服务供应商安装、配置或维修群集硬件之前，这些概念可帮助他们理解硬件部件之间的关系。群集系统管理员可能也会发现这些信息很有用，它们可用作安装、配置和管理群集软件的背景信息。

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（多个节点之间可共享的磁盘）
- 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26页的「Sun Cluster 拓扑」中所述。

下图是一个群集配置样例:



图表 2–1　双节点群集配置样例

群集节点

群集节点是同时运行 Solaris 操作系统和 Sun Cluster 软件的机器，它要么是群集的当前成员（群集成员），要么是潜在成员。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，请参见第26页的「Sun Cluster 拓扑」。

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。

另外，并行数据库配置中的各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22页的「多主机磁盘」 和第 3 章。

群集中的所有节点都会归组到一个共用的名称下，即用于访问和管理群集的"群集名称"下。

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。

群集成员通过物理位置上相互独立的一个或多个网络与群集中的其他节点通信。这组物理位置上相互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会分清哪些是在本地运行的资源以及哪些是在其他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故障转移。因为存在故障转移的可能性，所以每个节点都必须具有足够的额外能力，能够承担那些它身为它们的备份或辅助节点的节点的工作量。

各个节点引导自己的根 (/) 文件系统。

群集成员的软件组件

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
- 卷管理软件（Solstice DiskSuite™ 或 VERITAS Volume Manager）

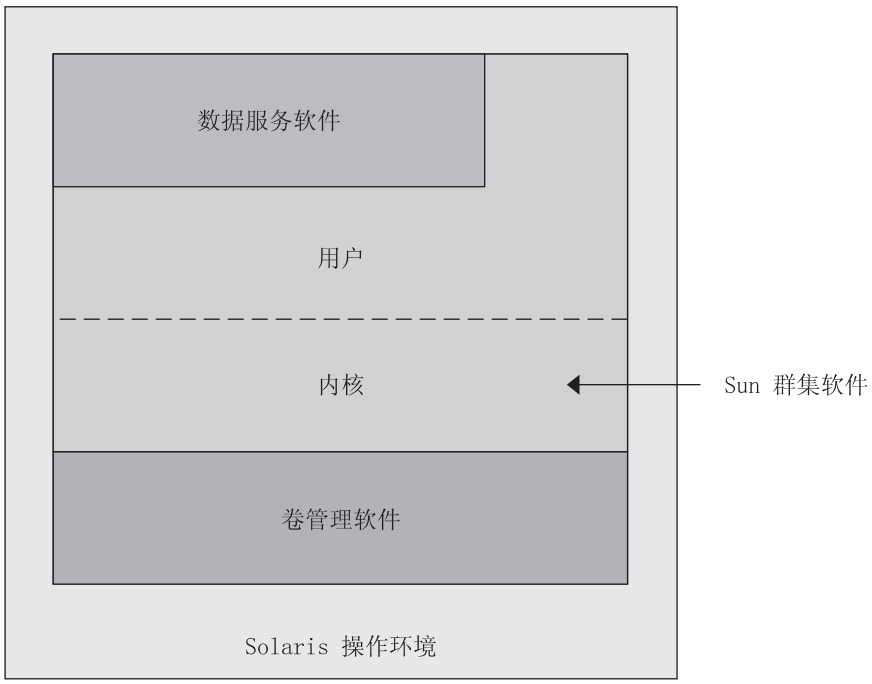
  在使用硬件独立磁盘冗余阵列 (RAID) 的一个 Oracle Parallel Server(OPS) 配置中有一个例外。该配置不需诸如 Solstice DiskSuite 或 VERITAS Volume Manager软件卷管理器来管理 Oracle 数据。

有关如何安装 Solaris 操作系统、Sun Cluster 和卷管理软件的信息，请参见《*Sun Cluster 3.0 U1* 安装指南》。

有关如何安装和配置数据服务的信息，请参见《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》。

有关上述软件组件的概念信息，请参见第 3 章。

下图展示了用于建立 Sun Cluster 软件环境的软件组件之间的逻辑关系视图。



图表 2–2 Sun Cluster 软件组件的逻辑关系

有关群集成员的问题及解答，请参见第 4 章。

多主机磁盘

Sun Cluster 需要多主机磁盘存储器：它们是可以同时连接到多个节点的磁盘。在 Sun Cluster 环境中，多主机存储器使磁盘具有高可用性。

多主机磁盘具有以下特点:

- 它们能够容忍单个节点出现故障。
- 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，则请求会切换到与同一磁盘直接连接的另一节点。
- 对多主机磁盘的访问，要么是通过“控制”磁盘的主节点进行全局访问，要么是通过本地路径进行的直接并行访问。当前唯一能使用直接并行访问的应用程序是 OPS。

卷管理器为镜像或 RAID-5 配置提供多主机磁盘的数据冗余。当前，Sun Cluster 支持 Solstice DiskSuite 和 VERITAS Volume Manager 用作卷管理器和 Sun StorEdge™ A3x00 存储单元中的 RDAC RAID-5 硬件控制器。

使用磁盘镜像和磁盘条带化，既可防止节点故障，又可防止单个磁盘故障。

有关多主机存储器的问题及解答，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，服务器节点通过将此服务器连接到特定 SCSI 总线的 SCSI 主机适配器线路，来控制 SCSI 总线活动。 该 SCSI 主机适配器线路称作 *SCSI initiator*。它启动此 SCSI 总线的全部总线活动。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 总线上的设备。）因为所有 SCSI 主机适配器的缺省 SCSI 地址均为 7，所以多启动器环境中的缺省硬件配置会导致冲突。

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，将其他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 地址当前未使用，并且以后永远也不会使用。将来不使用地址的一个示例是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。

可以通过设置 `scsi-initiator-id` Open Boot PROM (OBP) 特性来更改这些主机适配器的选定 SCSI 地址。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 3.0 U1 Hardware Guide*》中的每一磁盘群组所对应的章中，都提供了有关为每个 SCSI 主机适配器设置唯一 `scsi-initiator-id` 的说明。

局部磁盘

局部磁盘是仅连接到一个节点的磁盘。因此它们无法防止节点故障（不具备高可用性）。不过，包括局部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安装点下，可以使其他节点也能使用它们。如果当前安装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，这样在磁盘发生故障时所有节点仍能访问这些文件系统，但是卷管理器不能在发生节点故障时采取保护措施。

有关全局设备的详细信息，请参见第36页的「全局设备」 一节。

可拆卸介质

群集中支持诸如磁带驱动器和 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设备。有关安装和配置可拆卸介质的详细信息，请参考《*Sun Cluster 3.0 U1 Hardware Guide*》。

有关全局设备的详细信息，请参见第36页的「全局设备」一节。

群集互连

群集互连是用于在群集节点间传输群集专用通信和数据服务通信的设备的物理配置。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。

所有节点必须由群集互连通过至少两个物理位置上独立的冗余网络或路径连接起来，以避免单节点故障。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上的网络适配卡。其名称由设备名称和紧随的物理单元号码构成，例如 qfe2。某些适配器只有一个物理网络连接，但其他适配器（如 qfe）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。

  具有多个接口的网卡在整个卡出现故障时会成为单故障点。为了获得最高可用性，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卡。

- 结点 – 驻留在群集节点外的开关。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。可以设置节点，使之包含多个公共网络接口卡，将一个卡配置为活动卡，其他卡作为备份卡。称为“公共网络管理”(PNM) 的 Sun Cluster 软件的子系统监视着活动接口。如果活动适配器出现故障，则调用 Network Adapter Failover (NAFO) 软件进行故障转移，将接口切换至一个备份适配器。

对公共网络接口进行群集化配置时，不需使用任何特殊硬件。

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他服务器。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他服务。

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，请使用和群集硬件一起购买的终端集中器，前者指Sun Enterprise E10000 server 服务器上的 System Service Processor (SSP)或者可从每一节点访问 `ttya` 的另一种设备。

Sun 只提供了一个支持的终端集中器作为选件使用。终端集中器通过使用 TCP/IP 网络实现对各个节点上的 `/dev/console` 的访问。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) 为 Sun Enterprise E10000 server 提供控制台访问。SSP 是以太网上的机器，经配置支持 Sun Enterprise E10000 server。SSP 是 Sun Enterprise E10000 server服务器的管理控制台。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。

其他控制台访问方法包括其他终端集中器、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™ 键盘和监视器或其他串行端口设备（如果硬件服务供应商支持这些设备）。

管理控制台

可以使用专用 SPARCstation™ 系统（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 Cluster Control Panel (CCP) 和 Sun Management Center 产品的 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，请参见《*Sun Cluster 3.0 U1* 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，则必须有能力从管理控制台登录到 System Service Processor (SSP)，并能使用 `netcon(1M)` 命令进行连接。

通常不需使用监视器便可配置节点。这样，您从管理控制台（该控制台连接到终端集中器，然后从终端集中器连接到节点的串行端口）通过 `telnet` 会话访问节点的控制台。（如果使用 Sun Enterprise E10000 server，则从 System Service Processor 进行连接。）有关详细信息，请参见第25页的「控制台访问设备」。

Sun Cluster 不需要专用的管理控制台，但是使用它有以下好处：

- 在同一机器上通过组合控制台和管理工具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。

Sun Cluster 拓扑

拓扑是群集节点与群集中所用存储平台的连接方案。

Sun Cluster 支持下列拓扑：

- 群集对
- N+1（星型）

下面两节分别介绍两种拓扑。

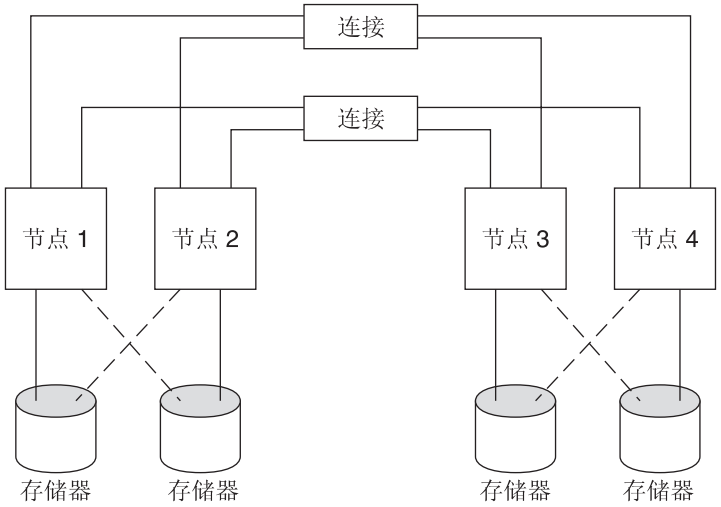
群集对拓扑

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，故障转移只会在每对节点间进行。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故障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。

下图图示说明了群集对配置。

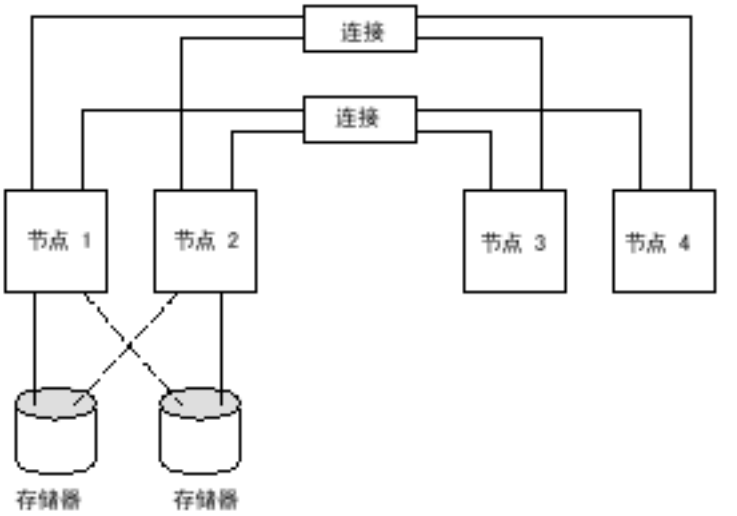


图表 2–3　群集对拓扑

Pair+M 拓扑

pair+M 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—这组节点之间本身没有进行直接连接。此配置中的所有节点依然是使用卷管理器来配置的。

下图图士说明了一个 pair+M 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。



图表 2–4　Pair+M 拓扑

N+1（星型）拓扑

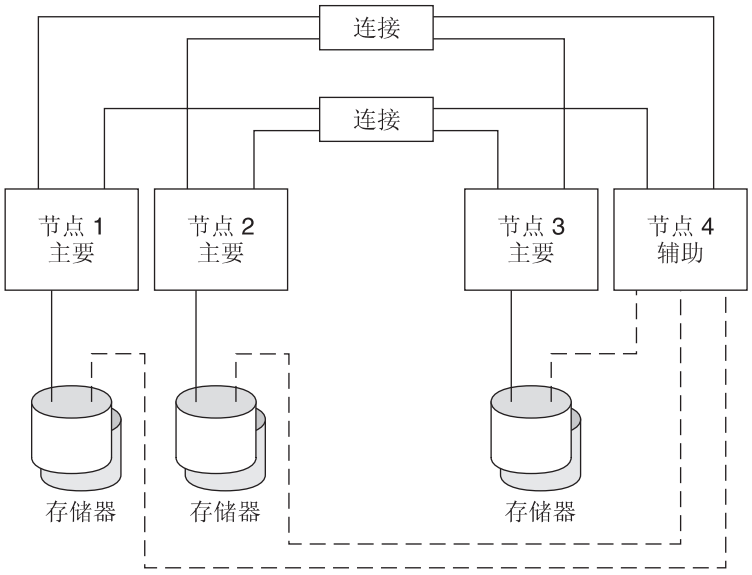
N+1 拓扑包括几个主节点和一个辅助节点。主节点和辅助节点的配置不必完全相同。一般由主节点提供应用程序服务。辅助节点在等待主节点出现故障时需处于空闲状态。

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 则会进行故障转移，将资源切换至辅助节点，这些资源将在辅助节点继续发挥作用，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。

下图图示说明了 N+1 配置。



图表 2–5　N+1 拓扑

第 3 章

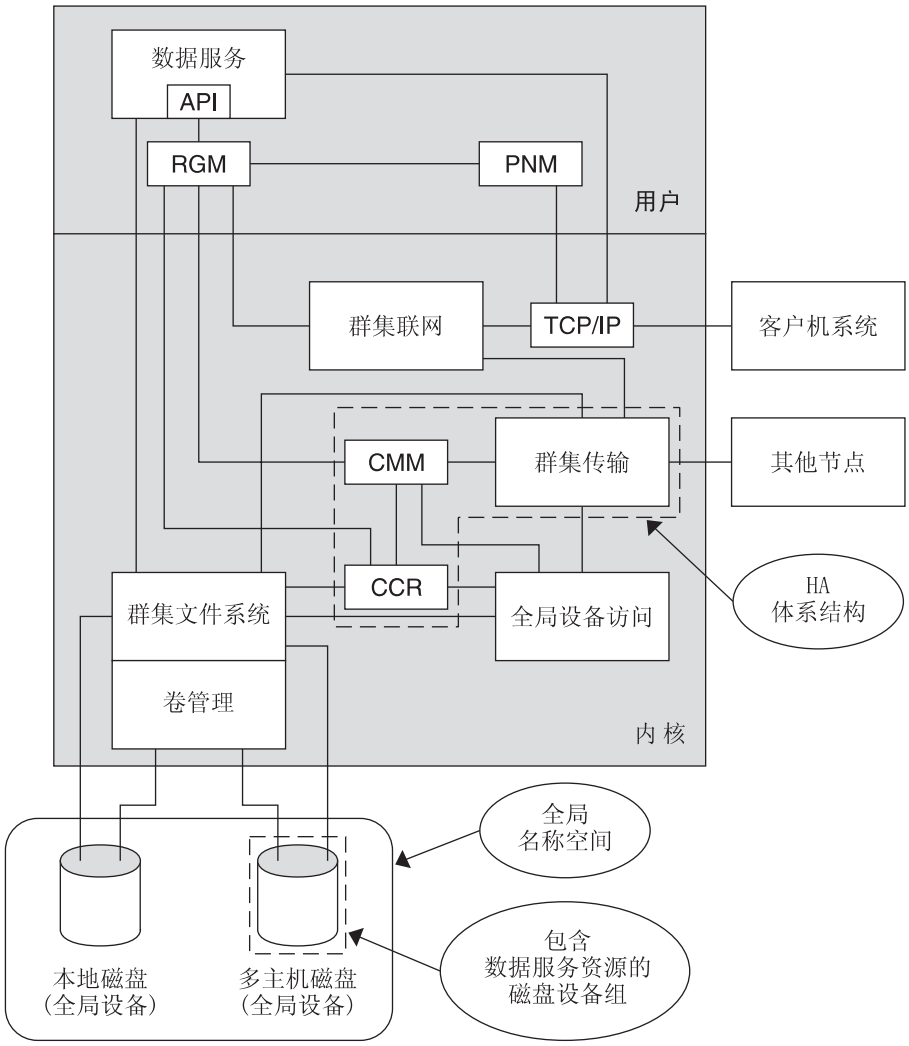
重要概念 – 管理和应用程序开发

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。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。

下图从较高的层次上显示了群集管理概念与群集体系结构之间的对应关系。



图表 3–1　Sun Cluster 软件体系结构

管理界面

您可以从若干用户界面中选择如何安装、配置和管理 SunPlex 系统。还可以通过命令行接口来完成系统管理任务。在命令行接口的上部有一些实用程序，用于简化选定的配置任务。SunPlex系统还有一个模块，作为 Sun Management Center（可向某些群集任务提供 GUI）的一部分来运行。关于管理接口的完整说明，可参见《*Sun Cluster 3.0 U1* 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。您是否将群集节点与任何外部时间源同步对于群集操作并不重要。SunPlex 系统使用网络时间协议 (NTP) 在节点间保持时钟同步。

通常，系统时钟一秒钟的时间改变不会有任何问题。然而，如果要在活动的群集中运行 `date(1)`、`rdate(1M)` 或 `xntpdate(1M)`（以交互方式或在 `cron` 脚本内）命令，就会强制执行远远超过一秒钟的时间改变，以使系统时钟与时间源同步。这一强制改变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，您可以改变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)` 来安装 Sun Cluster 软件时，其中有一步是配置群集的 NTP。Sun Cluster 软件提供了一个模板文件 `ntp.cluster`（请参见已安装的群集节点上的 `/etc/inet/ntp.cluster`），它在所有群集节点间建立一种对等的关系，其中有一个节点是“首选”节点。节点由它们的专用主机名标识，时间同步会在群集互连系统中进行。关于如何配置群集的 NTP 的说明，可从《*Sun Cluster 3.0 U1* 安装指南》中获得。

或者您也可以在群集之外设置一个或多个 NTP 服务器，并更改 `ntp.conf` 文件来体现此配置。

正常运行时，绝不需要调整群集的时间。然而，如果安装 Solaris 操作环境时时间设置不正确，现在想更改时间，可在《*Sun Cluster 3.0 U1* 系统管理指南》中找到操作步骤。

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 SunPlex 组件故障种类（硬件和软件）和高可用性框架内置的恢复种类。

表 3–1 SunPlex 故障检测和恢复级别

发生故障的群集组件	软件恢复	硬件恢复
数据服务	HA API，HA 框架	无
公共网络适配器	网络适配器故障转移 (NAFO)	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
被镜像的多主机磁盘	卷管理（Solstice DiskSuite 和 VERITAS Volume Manager）	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
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经恢复就立即可用。已恢复的框架资源不必等待其他所有框架资源完全恢复。

可用性最高的框架资源的恢复对于使用它的应用程序（数据服务）来说是透明的。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。只要存在从另一节点到磁盘的替换路径，单个节点的故障对使用连接到此节点的文件、设备和磁盘卷的其他节点上的程序来说就是完全透明的。多主机磁盘的使用就是一个样例，这些磁盘具有连接多个节点的端口。

群集成员监视器

群集成员监视器 (CMM) 是一个分布式代理程序集，每个群集成员有一个代理程序。这些代理程序通过群集互连交换信息，来实现以下功能：

- 在所有节点上保持一致的成员视图（定额）
- 使用注册的回叫来驱动同步重新配置，以响应成员更改
- 处理群集划分（群集分割，失忆）

- 保证所有群集成员间的充分连通性

与 Sun Cluster 软件以前的发行版不同，CMM 完全运行在内核中。

群集成员

CMM 的主要功能是针对在任一给定时间加入群集的节点集合建立一个群集范围内的协议。这种约束称为群集成员。

为确定群集成员并最终保证数据的完整性，CMM：

- 说明群集成员的更改，如某个节点加入或脱离群集
- 保证“故障”节点脱离群集
- 保证“故障”节点在修复前不进入群集
- 防止群集将自身划分为一些节点子集

有关群集如何防止自身划分为多个独立群集的详细信息，请参见 第43页的「定额和定额设备」。

群集成员监视器重新配置

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他节点连通性的信息。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。

故障快速防护机制

如果 CMM 检测到节点上存在严重的问题，它会要求群集框架来强制关闭该节点（应急状态）并将其从群集成员中删除。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。

- 如果节点脱离群集后试图在没有定额的情况下启动新的群集，它会被“隔离”，从而无法访问共享磁盘。有关使用故障快速防护的详细信息，请参见第46页的「故障防护」。
- 如果一个或多个群集特定的守护程序出现故障（`clexecd`、`rpc.pmfd`、`rgmd` 或 `rpc.ed`），CMM 会检测到该故障，节点将处于应急状态。

当群集守护程序中止而导致应急状态时，该节点的控制台上将显示类似以下内容的信息。

```
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
```

应急状态过后，节点可能重新引导，试图重新连接群集或停留在 OpenBoot PROM (OBP) 提示符下。所采取的措施取决于 OBP 中 `auto-boot?` 参数的设置。

群集配置库 (CCR)

群集配置库 (CCR) 是专用的群集范围的数据库，用于保存与群集的配置和状态有关的信息。CCR 是一个分布式数据库。每个节点上都有该数据库的完整副本。CCR 可确保所有节点看到的群集世界都一样。要避免损坏数据，各个节点都需要知道群集资源的当前状态。

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。

小心：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 来保证群集只有在定额建立后才能运行。CCR 负责验证数据在整个群集范围内的一致性，需要时执行恢复，并协助进行数据更新。

全局设备

SunPlex 系统使用全局设备提供群集范围内的高可用性访问，这种访问可以是对群集中的任一设备，自任意节点，而不用考虑设备的物理连接位置。如果一个节点在提供对某个全局设备的访问时出现故障，则 Sun Cluster 软件会自动找到指向该设备的另一路径，并将访问重定向到此路径。Sun Cluster 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。

群集会给群集中的每个磁盘、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，请参见第39页的「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。此驱动程序可自动给群集内的每个设备（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 驱动程序会分配一个全局名称（如 d10）供节点使用，这样就为每个节点提供了到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，请参见相应的手册页。

磁盘设备组

在 SunPlex 系统中，所有多主机磁盘都必须由 Sun Cluster 软件进行控制。首先在多主机磁盘上创建卷管理器磁盘组（Solstice DiskSuite 磁盘集或 VERITAS Volume Manager 磁盘组）。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 软件会将每一个磁盘都注册为磁盘设备组。

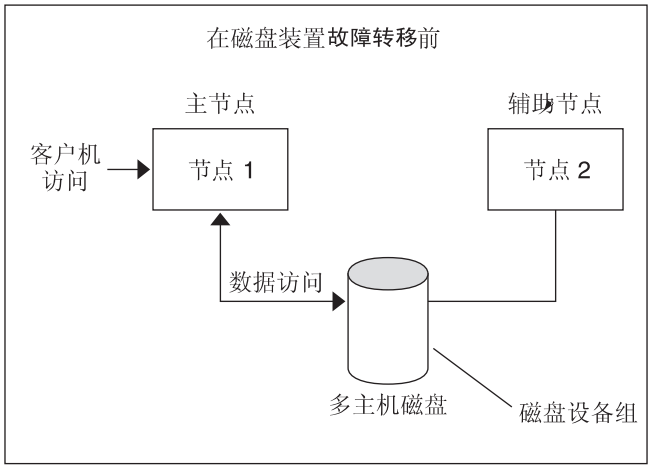
注册提供了有关各个节点连接到卷管理器磁盘组的路径的 SunPlex 系统信息。此时，在群集范围内可以对卷管理器磁盘组进行全局访问。如果多个节点均可写入（控制）磁盘设备组，存储在该磁盘设备组中的数据将具有高度可用性。此高度可用的磁盘设备组可用于存储群集文件系统。

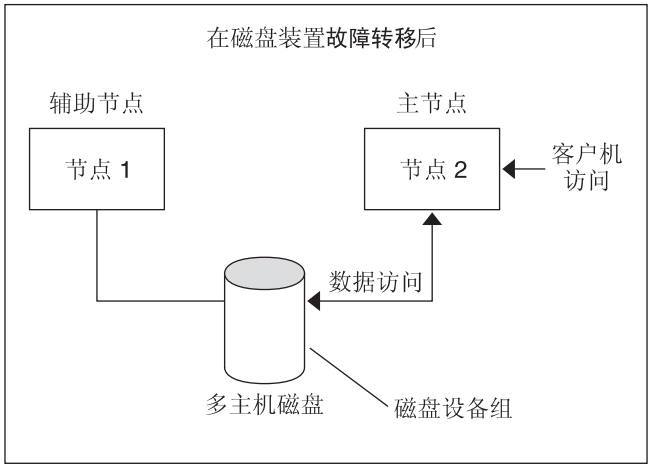
注意：磁盘设备组独立于资源组。一个节点可以控制资源组（代表一组数据服务进程），而另一个节点可以控制正被数据服务访问的磁盘组。不过最好的做法是，让存储特定应用程序数据的磁盘设备组和包含此应用程序资源（应用程序守护程序）的资源组保持在同一节点上。有关磁盘设备组和资源组之间关系的详细信息，请参见《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》中包含概述性内容的一章。

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。

磁盘设备组故障转移

因为磁盘群组连接着一个以上的节点，所以群组中所有磁盘设备组在当前控制它的节点出现故障时，都可以通过备用路径访问。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。





图表 3–2　磁盘设备组故障转移

全局名称空间

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（及所有其他群集设备，如 CD-ROM 和磁带），并提供指向多主机磁盘的多条故障转移路径。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。

正常情况下，卷管理器名称空间的驻留位置是：对于 Solstice DiskSuite，在 `/dev/md/`*`diskset`*`/dsk`（和 `rdsk`）目录下；对于 VxVM，在 `/dev/vx/dsk/`*`disk-group`* 和 `/`

dev/vx/rdsk/*disk-group* 目录下。这些名称空间分别由整个群集中引入的各 Solstice DiskSuite 磁盘集和各 VxVM 磁盘组的目录组成。每一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每个设备节点都被替换为指向 /global/.devices/node@*nodeID* 系统文件的某个设备节点的符号链接，其中 *nodeID* 是一个整数，代表群集中的节点。Sun Cluster 软件还会在卷管理器设备的标准位置上将其显示为符号链接。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，几乎不需对设备管理模型进行改动。
- 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@*ID* /dev/dsk/c0t0d0s0
DID 名称	/dev/did/dsk/d0s0	/global/.devices/node@*ID* /dev/did/dsk/d0s0
Solstice DiskSuite	/dev/md/*diskset*/dsk/d0	/global/.devices/node@*ID* /dev/md/*diskset*/dsk/d0
VERITAS Volume Manager	/dev/vx/dsk/*disk-group* /v0	/global/.devices/node@*ID* /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。也可以运行 scgdevs(1M) 命令来生成全局名称空间。

群集文件系统

群集文件系统是一个节点上的内核和某个与磁盘进行了物理连接的节点上运行的基础文件系统及卷管理器之间的代理。

群集文件系统依赖于和一个或多个节点进行了物理连接的全局设备（磁盘、磁带、CD-ROM）。全局设备可从群集中任何节点上通过同一个文件名称（如 /dev/global/）访问，而不用管此节点与存储设备是否有物理连接。可以像常规设备那样使用全局设备，也就是说，可在该设备上面可以用 newfs 和/或 mkfs 命令创建文件系统。

对于全局设备上的文件系统，可以使用 mount –g 进行全局安装，也可使用 mount 进行本地安装。

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。

群集文件系统安装在所有的群集成员上。不能在群集成员的子集上安装群集文件系统。

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，所有多主机磁盘都放在磁盘设备组中，这些组可以是 Solstice DiskSuite 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，群集文件系统中的本地文件系统（存储在节点的本地磁盘上的文件系统）不具有高可用性。

与一般的文件系统一样，您可以通过两种方式安装群集文件系统：

- 手动 — 使用 mount 命令和 –g 或 –o global 安装选项来从命令行安装群集文件系统，例如：

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

- 自动 — 在 /etc/vfstab 文件中用 global 安装选项创建一个条目，以便在引导时安装群集文件系统。接着就可以在所有节点上的 /global 目录下创建一个安装点。目录 /global 为推荐位置，不是必需的。下面是 /etc/vfstab 文件中一个群集文件系统的样例行：

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

注意：Sun Cluster 软件不强制使用群集文件系统的命名策略，所以可以通过在同一目录下（如 /global/*disk-device-group*）为所有群集文件系统创建一个安装点，以便于进行管理。有关详细信息，请参见《*Sun Cluster 3.0 U1* 安装指南》和《*Sun Cluster 3.0 U1* 系统管理指南》。

群集文件系统的特性

群集文件系统具有以下特性：

- 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。
- 相关协议用于保留 UNIX 文件访问的语义，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 移动并用，可以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咨询文件锁定功能。通过使用群集文件系统文件上的咨询文件锁定功能，运行在多个群集节点上的应用程序可以实现对数据的同步访问。节点脱离群集后，或应用程序在锁定期间出现故障后，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，也可保证对数据的不间断访问。
- 群集文件系统是独立于基础文件系统和卷管理软件的。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。

Syncdir 安装选项

syncdir 安装选项可用于将 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。如果您指定 syncdir，则保证写入的数据符合 POSIX 标准。如果不指定，则会遇到 NFS 文件系统中通常会发生的问题。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。指定了 syncdir（和 POSIX 行为），在写入操作过程中系统就能发现空间不足的问题。不指定 syncdir 很少会出现问题，所以我们建议最好不要指定该选项，这样做有助于提高系统性能。

有关全局设备和群集文件系统的常见问题，请参见第64页的「文件系统常见问题」。

定额和定额设备

由于群集节点能共享数据和资源，所以千万不要将群集分割为同时处于活动状态的独立分区。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。如果在节点间失去群集互连并且将群集划分为若干子群集，每个分区都认为自己是唯一分区，此时即会发生群集分割。这是由群集节点之间的通信问题引起的。在群集关闭后又重新启动的情况下会发生失忆，此时的群集数据的版本要早于群集关闭时记录的信息的版本。如果磁盘上存储有多个版本的框架数据，而在尚未获得最新的版本的情况下启动新的群集，则可能发生这种情况。

可以通过以下方法来避免群集分割和失忆的发生：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 定额，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非常有效。在双节点群集中，多数为二。如果这样的群集分为两个分区，则需要使用外部选票才能使其中一个分区获得定额。此外部选票由 定额设备提供。定额设备可以是这两个节点所共享的任一磁盘。用作定额设备的磁盘可以包含用户数据。

表格 3–3 说明 Sun Cluster 软件如何使用定额来避免群集分割和失忆。

表 3–3 群集定额及群集分割和失忆问题

分区类型	定额解决方案
群集分割	仅允许获得多数选票的分区（子群集）作为群集（其中最多仅能有一个拥有多数选票的分区）
失忆	在群集引导时，保证至少有一个节点是最新的群集成员之一（因而有最新的配置数据）

定额算法动态执行：当群集事件触发计算时，计算的结果可以随群集生存期的不同而改变。

定额选票计数

群集节点和定额设备都会获得选票以形成定额。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定额选票计数。节点的选票数可以是零，例如当正在安装节点或管理员将节点置于维护状态时便是如此。

定额设备获取定额选票计数基于设备的节点连接数。在设置定额设备时，它需获取一个最大选票数 *N*-1，其中 *N* 是有非零选票数的节点数，并且这些节点有到定额设备的端口。例如，连接到两个选票数非零的节点的定额设备有其中一个的定额数（二减一）。

您要在群集安装期间，或以后通过使用在《*Sun Cluster 3.0 U1* 系统管理指南》中描述的过程来配置定额设备。

注意：仅在当前连接的节点中至少有一个是群集成员时，定额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，定额设备才对选票数起作用。

定额配置

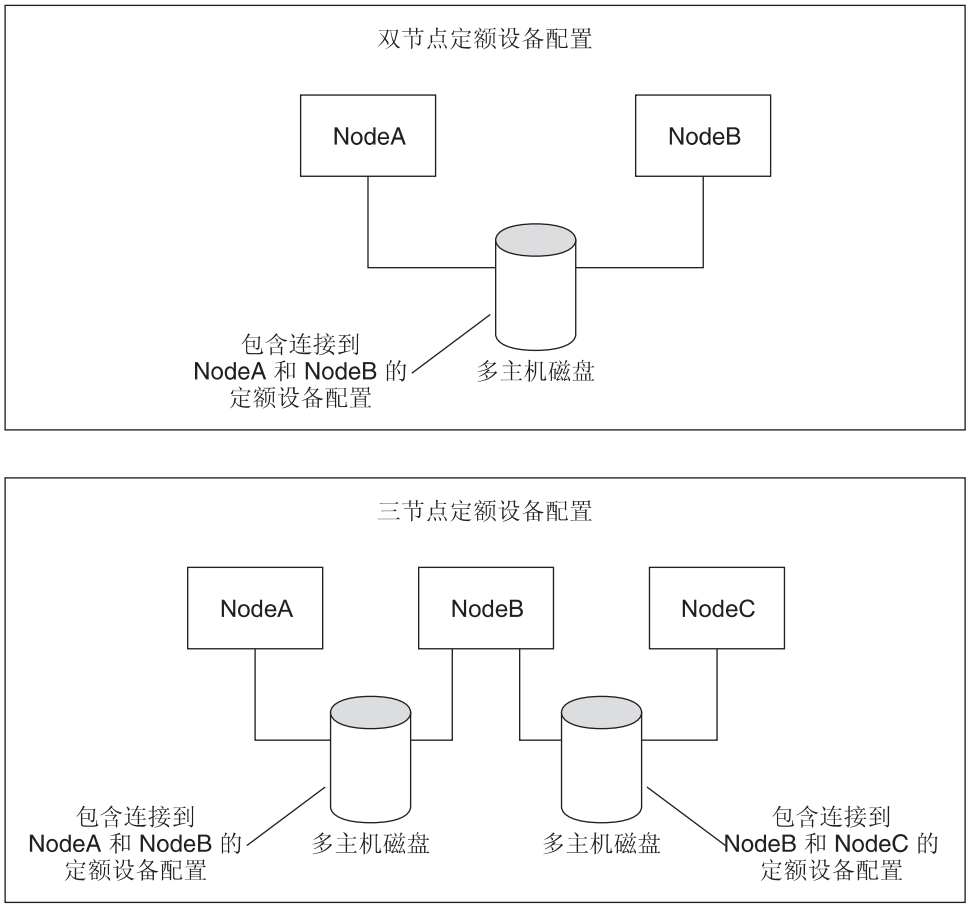
定额配置依赖于群集中节点的数目：

- 双节点群集 – 要形成双节点群集，需要两个定额选票。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定额设备。然而，在双节点群集中，必须配置一个定额设备，以确保在一个节点发生故障时另一单个节点可以继续工作。
- 多于两个节点的群集 – 您应在共享对磁盘存储器群组访问的每对节点间指定一个定额设备。例如，假定您有一个由三个节点组成的如 图表 3–3 中所示的群集。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。总共会有五个定额选票，其中三个来自节点，两个来自节点共享的定额设备。一个群集需要获得多数定额选票（即三个）才能生成。

Sun Cluster 软件不要求也不强制在共享对磁盘存储器群组访问的每对节点间指定一个定额设备。但是，对于 N+1 配置降级为一个双节点群集并且紧接着对两个磁盘群组都有访问权的节点也发生故障的情况，它可以提供所需要的定额选票。如果您在每对节点之间配置了定额设备，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，请参见图表 3–3。



图表 3–3　定额设备配置示例

定额准则

在设置定额设备时，请使用下列准则：

- 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定额设备。在共享群组内添加一个磁盘作为定额设备以确保在任何节点发生故障时，其他节点可以维持定额并可以控制共享群组上的磁盘设备组。
- 必须将定额设备连接到至少两个节点上。
- 定额设备可以是用作双端口定额设备的任何 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定额设备。有关详细信息，请参见《*Sun Cluster 3.0 U1* 安装指南》中有关计划的章节。
- 您可以使用包含用户数据的磁盘作为定额设备。

提示：要防止个别定额设备出现故障，请在节点集之间配置一个以上的定额设备。使用来自不同群组的磁盘，并在每个节点集之间配置奇数的定额设备。

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故障转移能力。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，在极短时间的中断后重又恢复对磁盘设备组的访问，。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，群集无法通知该节点释放那些将它作为主节点的设备。因而，需要一种方法来使群集中存活的成员能够接替发生故障的成员来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，可将故障节点与多主机磁盘“隔离”，使其无法访问那些磁盘。

SCSI-2 磁盘保留支持一种保留形式，它或者授予客户机对连接到磁盘的所有节点的访问权限（当没有保留上时），或者只授予客户机对某一单个节点（即拥有该保留的节点）的访问权限。

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阻止另一个节点访问共享磁盘。当发生此故障防护时，通常将防护的节点处于应急状态，并在其控制台上显示“保留冲突”的消息。

发生保留冲突是因为在节点已被检测到不再是群集成员后，一个 SCSI 保留被置于在此节点与其他节点间共享的所有磁盘上。防护节点可能不会意识到它正在被防护，并且如果它试图访问这些共享磁盘之中的一个，它会检测到保留和应急状态。

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，这种机制被称为故障快速防护。

属于群集成员的节点对它们可以访问的磁盘（包括定额磁盘）持续启用一个特定 ioctl，MHIOCENFAILFAST。该 ioctl 是对磁盘驱动程序的指令，如果某磁盘被其它节点保留而无法由该节点访问，该指令使该节点能够将自身处于应急状态。

MHIOCENFAILFAST ioctl 指示驱动程序检查节点对磁盘发出的每个读写操作所返回的错误，查看是否返回 `Reservation_Conflict` 错误代码。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，保留不是永久性的－它们无法免于节点重新引导。对于具有持久性组保留 (PGR) 的 SCSI-3 磁盘，保留信息存储在磁盘上，并在多次节点重新引导后仍保持有效。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定额的分区的一部分，它将被另一节点强行从该群集中删除。属于可获取定额的分区一部分的另一节点将保留放置在共享磁盘上，当不具备定额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。

进入应急状态之后，节点可能重新引导，试图重新连接群集；也可能停留在 OpenBoot PROM (OBP) 提示符状态下。所采取的措施取决于 OBP 中 `auto-boot?` 参数的设置。

卷管理器

SunPlex 系统使用卷管理软件通过镜像和热备份磁盘来增加数据的可用性，并处理磁盘故障和更换。

SunPlex 系统没有它自己的内部卷管理器组件，而依赖于下面的卷管理器：

- Solstice DiskSuite
- VERITAS Volume Manager

群集中的卷管理软件提供对如下功能的支持：

- 节点故障的故障转移处理
- 来自不同节点的多路径支持
- 对磁盘设备组的远程透明访问

一旦卷管理对象受控于群集，它们就成为磁盘设备组。有关卷管理器的信息，请参见卷管理器软件文档。

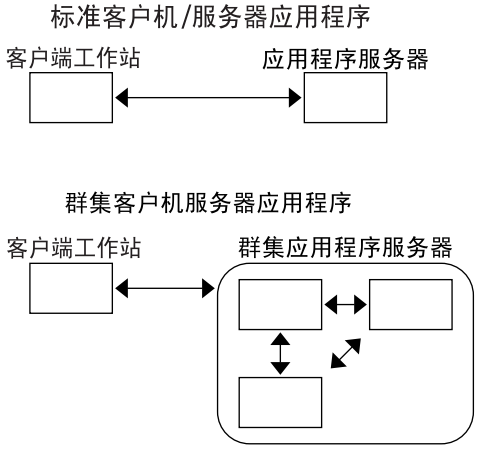
注意：在规划磁盘集或磁盘组时一个重要的考虑事项就是要了解它们的关联磁盘设备组是如何在群集内与应用程序资源（数据）相联系的。关于这些问题的讨论，请参考《Sun Cluster 3.0 U1 安装指南》和《Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide》。

数据服务

术语数据服务是用来描述诸如 Oracle 或 iPlanet Web Server 之类的第三方应用程序，该应用程序已配置为在群集上运行，而不是在一个单独的服务器上运行。 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

图表 3–4 将运行于单个应用程序服务器（单服务器模型）之上的应用程序与运行于群集（群集服务器模型）之上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。



图表 3–4　标准与群集客户机/服务器配置

在单服务器模型中，可以对应用程序进行配置，以便通过公共网络接口（主机名）来访问该服务器。 主机名与物理服务器有关。

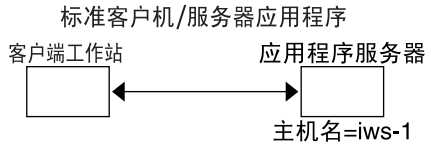
在群集服务器模型中，公共网络接口为逻辑主机名或共享地址。术语网络资源是指逻辑主机名和共享地址。

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，请参见每项数据服务的安装与配置信息。

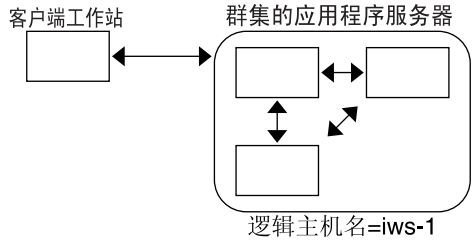
网络资源与具体的物理服务器无关－它可以在物理服务器之间进行移植。

最初，网络资源仅与一个节点，即主节点有关联。如果主节点出现故障，网络资源及应用程序资源将切换到另一群集节点（辅助节点）。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。

图表 3–5 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。该应用程序已配置为使用这一逻辑主机名，而不使用与特定服务器相关的主机名。



对群集的客户机-务器应用程序进行故障切换



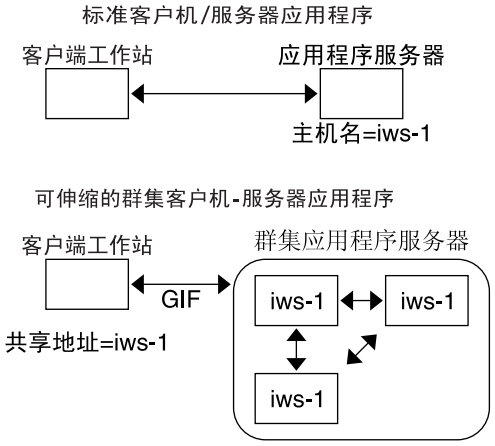
图表 3–5　固定主机名与逻辑主机名

共享地址最初也仅与一个节点相关联。这个节点被称为全局接口节点 (GIN)。共享地址被用作与群集之间的单一网络接口,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”是指可以通过添加额外的群集节点，为应用程序增添 CPU 能量，从而增强性能。

如果 GIN 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 GIN）。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。

图表 3–6 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。与逻辑主机名用于故障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。



图表 3–6　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故障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。

Resource Group Manager (RGM)

RGM 将数据服务（应用程序）作为资源控制，资源是由资源类型实现所管理的。这些实现或者由 Sun 提供，或者由具有一个普通数据服务模板、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 可用于资源和资源组。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。 第59页的「资源和资源组状态与设置」 中提供适用于资源和资源组的各种状态与设置的完整说明。

故障转移数据服务

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。故障转移服务利用了故障转移资源组，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。

对于故障转移数据服务，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（故障转移），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：利用可伸缩资源组来保存应用程序资源，利用故障转移资源组来保存可伸缩服务所依赖的网络资源（ 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故障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。

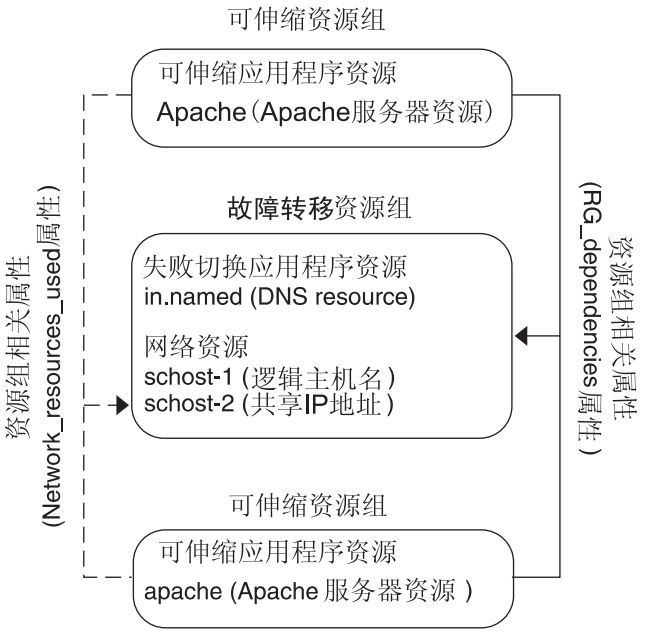
服务请求通过一个单独的网络接口（全局接口）进入群集，并依据由负载平衡策略设置的几个预定义算法之一来将这些请求分发到节点。群集可以使用负载平衡策略来平衡几个节点间的服务负载。注意：在不同节点上可以有多个全局接口以其它共享地址为主机。

对于可伸缩服务来说，应用程序实例在几个节点上同时运行。如果拥有全局接口的节点出现故障，全局接口将切换到其他节点。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。

图表 3–7 显示了故障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。故障转移资源组包括高可用性的 DNS 应用程序资源，以及由高可用的 DNS 和 Apache Web 服务器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，资源组在可伸缩和故障转移资源组（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）。



图表 3–7　故障转移与可伸缩资源组示例

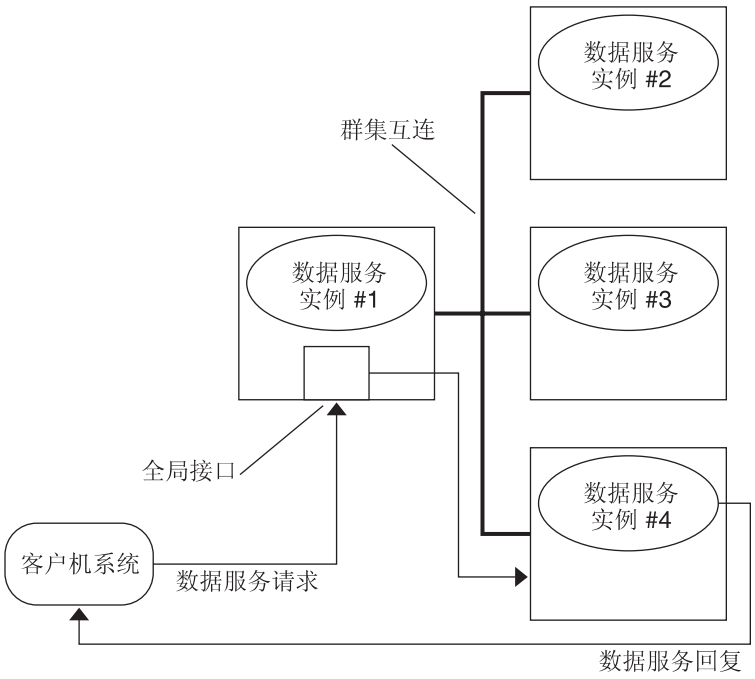
可伸缩服务体系结构

群集联网的主要目标是为数据服务提供可伸缩性。可伸缩性意味着随着提供给服务的负载的增加，在新的节点被添加到群集并运行新的服务器实例的同时，数据服务面对这种增加的工作负载能保持一个不变的响应时间。我们将这样的服务称为可伸缩数据服务。Web 服务是可伸缩数据服务的一个很好的示例。通常，可伸缩数据服务由几个实例组成，每一个示例运行在群集的不同节点上。这些实例在一起，作为来自该服务远程客户机的基准点的一个单独的服务，并实现该服务的功能。例如，我们可能会有一个由几个 `httpd` 守护程序组成的 Web 服务，并且这些守护程序在不同的节点上运行。任何 `httpd` 守护程序都服务于一个客户请求。服务于请求的守护程序依赖于负载平衡策略。对客户机的回复看起来是来自该服务，而不是服务于请求的特定守护程序，从而保留单个服务的外观。

可伸缩服务由以下功能组成：

- 对可伸缩服务的联网基础结构支持
- 负载平衡
- 对联网和数据服务的支持（使用 Resource Group Manager）

下图描绘了可伸缩服务的体系结构。



图表 3–8　可伸缩服务体系结构

当前不作为全局接口主机的节点（代理节点）与它们的回送接口共享地址。进入到全局接口的软件包被分发到基于可配置负载平衡策略的其它群集节点上。可能的负载平衡策略在下一步说明。

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他实例。

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，普通粘滞和通配粘滞。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。相对于正在监听单个端口的服务器实例，可以把该客户机说成是“粘滞的”。假定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，相对于在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”。

例如在一个电子商务站点上，顾客使用端口 80 上的普通 HTTP 购买了许多商品并放入到购物车中，但是要切换到端口 443 上的 SSL，以发送安全数据，使用信用卡付款。

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。

  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。这种加权策略保证了来自客户机的一定比例的流量直接到达特定的节点，此策略不对个别请求寻址。

- 粘滞的。在此策略中，端口的设置在配置应用程序资源时就已经知道了。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。

故障返回设置

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。故障返回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（故障返回）或仍保留当前的主节点。使用 故障返回资源组特性设置指定需要的选项。

在某些情况下，如果以资源组为主机的原始节点正在重复地发生故障并重新引导，设置故障返回可能会导致资源组减少可用性。

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，确定其是否完好。故障监视器检验应用程序守护程序是否在运行并且客户机正在接受服务。基于由探测返回的信息，可以启动一些预定义的操作，比如重新启动守护程序或引起故障转移。

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故障转移或可伸缩服务的方式运行。如果您想使之作为一个故障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，使之作为一个故障转移或可伸缩服务运行。

有一套标准可用于确定一个应用程序是否可以成为故障转移服务。特定的标准在 SunPlex 文档中进行了描述，该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，请阅读第52页的「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，这样的服务是由一个或多个服务器实例组

成的。每个实例运行在群集的不同节点上。同一服务的两个或更多实例不能在相同的节点上运行。

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，我们讲“外部的”是为了区分存储与内存内的状态，而讲“逻辑的”是因为存储看起来象单独的实体，尽管它本身可能是复制的。此外，这种逻辑数据存储有这样的特性，不论何时任何服务器实例更新该存储，其他实例会立即看到该更新。

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，假定一项服务将新数据写入外部日志文件，或修改在适当位置的现有数据。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，从而获取期望的同步。

关于这种存储类型的另一个示例是像高可用 Oracle 或 Oracle Parallel Server 那样的后端数据库。请注意，这种类型的后端数据库服务器使用数据库查询或更新事务提供内置的同步，因此多个服务器实例不需要实现它们自己的同步。

Sun 的 IMAP 服务器是当前并不体现为可伸缩服务的一个服务示例。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：

- 作为 SunPlex 系统部件提供的数据服务
- 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 3.0 U1 Data Services Installation and Configuration Guide*》介绍了如何安装和配置与 SunPlex 系统一起提供的数据服务。《*Sun Cluster 3.0 U1 Data Services Developer's Guide*》介绍了如何装备其它应用程序以使它们在 Sun Cluster框架下高度可用。

Sun Cluster API 使应用程序开发者能够开发可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故障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，从而使它作为一种故障转移或可伸缩的服务运行。

使用群集互连进行数据服务通信

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。对内部通信（如文件系统数据或可伸缩服务数据），消息是以循环方式在所有可用的互连间条带化的。

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`，请使用此名称在到节点 1 的群集互连上进行通信。 使用此名称打开的 TCP 套接字通过该群集互连进行路由，并且在发生网络故障时可以透明地重新路由。

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) 获取实际的名称。

对于应用程序级别的群集互连使用，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，而内部群集通信则在所有的互连中条带化。

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他群集通信的带宽。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故障转移。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` 的逻辑成对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。

注意，在群集互连上接受连接并为安全起见验证 IP 地址的应用程序必须检查 从 `gethostbyname` 中返回的所有 IP 地址，而不应只检查第一个 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：诸如 Apache Web Server 或 iPlanetWeb Server 之类的应用程序利用它们所依赖的网络地址（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`，而 Sun Cluster HA for Apache是资源类型 `SUNW.apache`。

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` 资源类型用于同步化资源和资源所依赖的磁盘设备组的启动。它可确保在数据服务启动时，到群集文件系统安装点、全局设备和设备组名称的路径可用。

RGM 管理的资源被放入称作资源组的组中，因此它们可以作为一个单元管理。如果故障转移或切换在资源组上启动，那么资源组就作为单元移植。

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，请参见《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》。

资源和资源组状态与设置

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。

  如果出于某种原因，要使所有群集节点都无法获得该资源，则需禁用资源。禁用的资源在一般情况下不能使用。

- 联机或脱机 – 这些是动态状态，适用于资源和资源组。

  这些状态通过切换或故障转移过程中的群集重新配置步骤，随着群集的转换而变化。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。

  故障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。可伸缩资源或资源组可以在一些节点上处于联机状态，而在另一些节点上处于脱机状态。在切换或故障转移过程中，资源组和其中的资源从一个节点脱机，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。

资源与资源组特性

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。其他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 3.0 U1 Data Services Installation and Configuration Guide*》的附录。

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 3.0 U1 Data Services Installation and Configuration Guide*》中有关数据服务的单独的章节中说明。

公共网络管理 (PNM) 和网络适配器故障转移 (NAFO)

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

Sun Cluster 公共网络管理 (PNM) 软件提供了基本的机制来监视公共网络适配器，并在检测到故障时将 IP 地址从一个适配器故障转移到另一个适配器。每个群集节点有它自己的 PNM 配置，该配置可以与其他群集节点上的不同。

公共网络适配器被编入到 *Network Adapter Failover* 组（NAFO 组）。每个 NAFO 组有一个或多个公共网络适配器。而在任何时候只有一个适配器对给定的 NAFO 组是活动的，在同一组中的更多适配器作为备份适配器使用，活动适配器上的 PNM 守护程序一旦检测到故障，在适配器故障转移期间就使用这些备份适配器。故障转移使与活动适配器相关联的 IP 地址被转移到备份适配器上，从而维持该节点的公共网络连通性。由于故障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，仅在故障转移期间有短暂的瞬时延迟。

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

NAFO 组为逻辑主机名和共享地址资源提供了构件。<command>`scrgadm(1M)` 命令在必要时自动为您创建 NAFO 组。您也可以独立于逻辑主机名和共享地址资源来创建 NAFO 组，以监视群集节点的公共网络连通性。节点上相同的 NAFO 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的详细信息，请参见《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》。

注意：NAFO 机制的设计着意于检测和屏蔽适配器故障。该设计并不旨在使用 `ifconfig(1M)` 从管理员那里恢复，以删除一个逻辑（或共享的）IP 地址。Sun Cluster 软件将逻辑和共享 IP 地址视为由 RGM 管理的资源。对于管理员来说，添加或删除 IP 地址的正确方法是使用 `scrgadm(1M)` 修改包含资源的资源组。

PNM 故障检测和故障转移过程

PNM 有规律地检查活动适配器的包计数，并假定运行良好的适配器的包计数会因通过适配器的正常网络流量而变化。如果一段时间包计数没有变化，那么 PNM 就进入一个 ping 序列，它加强了该通过活动适配器的流量。PNM 在每个序列结束时检查包计数的任何变化，并且如果在 ping 序列重复几次后包计数仍保持不变，就宣告适配器出现故障。这些事件触发了备份适配器的故障转移，只要有一个备份适配器可用，就转移到它。

输入和输出包计数由 PNM 监视，因此当两者之一或两者在一段时间内保持不变，则启动 ping 序列。

ping 序列由对 ALL_ROUTER 多址广播地址 (224.0.0.2)、ALL_HOST 多址广播地址 (224.0.0.1) 和本地子网广播地址的 ping 组成。

Ping 是以最低成本优先的方式构建的，因此如果有一个较低成本的 ping 可以成功运行，就不会运行较高成本的 ping。而且，ping 只作为在适配器上产生流量的一种方法使用。它们的退出状态不会影响对适配器功能或故障的判定。

在这一算法中有四个可以微调的参
数：inactive_time、ping_timeout、repeat_test 和 slow_network。这些参数在故障检测的速度和正确性之间提供了一种可调整的平衡。有关参数及如何更改它们的详细信息，请参见《*Sun Cluster 3.0 U1* 系统管理指南》中关于更改公共网络参数的步骤。

在 NAFO 组的活动适配器上检测到故障后，如果没有备份适配器可用，该组就被宣告关闭，同时继续对其所有备份适配器的测试。然而，如果有备份适配器可用，就会进行故障转移，切换到该适配器。当故障活动适配器被关闭并且不可查明时，逻辑地址和它们相关的标志被“转移”到备份适配器上。

一旦 IP 地址的故障转移成功完成之后，就会发送未经请求的 ARP 广播。从而保持与远程客户机的连通性。

常见问题

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

关于高可用性的常见问题

- 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?

  通过一个称为故障转移的进程，群集框架提供高可用性的环境。故障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。

- 故障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，故障转移数据服务和可伸缩数据服务。

  故障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。其他节点上可能运行其他应用程序，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故障转移，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节点 (GIN)。群集中可以有多个 GIN。每个 GIN 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。一个 GIN 具有可接收特定应用程序所有请求的全局接口，并将这些请求分发给运行应用程序服务器的多个节点。如果 GIN 发生故障，则全局接口将故障转移到一个仍正常工作的节点。

在任何一个运行着该应用程序的节点发生故障时，该应用程序在其他节点上继续运行，只是性能有所下降，直到故障节点返回该群集为止。

文件系统常见问题

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

  不可以，不要进行回送安装。

- 可否将群集文件系统用于不在 **Resource Group Manager**控制之下的应用程序?

  可以。然而，没有 RGM 的控制，当运行应用程序的节点发生故障时，需手动重新启动应用程序。

- 所有的群集文件系统都必须在 **/global** 目录下有一个安装点吗?

  并非必须。然而，将群集文件系统置于相同的安装点之下，比如 `/global`，可以使这些文件系统可以得到更好的组织和管理。

- 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，当多个节点访问一个文件，进行访问读、写、文件锁定、异步 I/O 时。
  4. 群集文件系统在某一服务器发生故障时支持无缝故障转移。NFS 支持多服务器，但只有只读文件系统有可能进行故障转移。
  5. 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。
  6. 如果您更改了文件的特性（例如，使用 `chmod(1M)`），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>@<node ID>** 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。如果某节点发生故障，其它节点无法访问这个节点的名称空间。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。

卷管理常见问题

- 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。这样的配置可以容忍单独磁盘故障。

- 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?

  这种配置由管理本地磁盘的 Solstice DiskSuite 软件和管理多主机磁盘的 VERITAS Volume Manager 提供支持。不支持其它任何组合方式。

数据服务常见问题

- 什么样的 **SunPlex** 数据服务是可用的?

  支持的数据服务列表包含在《*Sun Cluster 3.0 U1* 发行说明》中。

- **SunPlex** 数据服务支持哪些应用程序版本?

  支持的应用程序版本列表包含在《*Sun Cluster 3.0 U1* 发行说明》中。

- 我可以记下自己的数据服务吗?

  可以。有关详细信息，请参见《*Sun Cluster 3.0 U1 Data Services Developer's Guide*》和 Data Service Development Library API 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，无论在哪里，只要文档要求在 `Failover` 模式资源组中使用 `LogicalHostname` 资源，`SharedAddress` 资源或 `LogicalHostname` 资源就都可以交替地使用。`SharedAddress` 资源的使用会造成一些额外的开销，因为群集联网软件已为 `SharedAddress` 而配置，而不是为 `LogicalHostname` 而配置。

  使用 `SharedAddress` 的优点是这样一种情形，您正在配置可伸缩和故障转移两种数据服务，并想让客户能够使用相同的主机名访问这两种服务。在这种情形下，`SharedAddress` 资源与故障转移应用程序资源一起包含在一个资源组中，而可伸缩服务资源则包含在另一资源组中，并被配置为使用 `SharedAddress`。此时，可伸缩服务和故障转移服务可以使用在 `SharedAddress` 中配置的同一组主机名/地址。

公共网络常见问题

- **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。

- 在故障转移中 **MAC** 地址起什么作用?

  当故障转移发生时，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（节点故障转移到的新的物理适配器的地址）和旧的 IP 地址。当网络上的另一台机器接收这些软件包之一时，它从其 ARP 高速缓存中清除掉旧的 MAC-IP 映射并使用新的映射。

- **SunPlex** 系统是否支持在 **OpenBoot PROM** 中为主机适配器设置 **local-mac-address?=true**?

  不支持，不支持此变量。

- **NAFO** 在活动和备份适配器之间进行切换时会出现多长时间的延迟?

  延迟可能持续几分钟。这是因为 NAFO 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未经请求的 ARP。因此，只有在路由器上这个 IP 地址的 ARP 高速缓存条目超时后，它才可能使用失效的 MAC 地址。第二个延迟的原因可能是这两个 NAFO 适配器均与以太网交换器相连

接。完成 NAFO 切换后，其中一个 NAFO 适配器为不可查明，而另一个适配器为可查明。这时，以太网交换器必须禁用一个端口，并启用另一个端口，这可能需要一些时间。此外，对于以太网，交换器和新启用的适配器之间将进行速度协商，这也需用一些时间。最后，在完成切换之后，NAFO 还需对新启用的适配器进行最低限度的安全检查，以验证一切均正常运行。

群集成员常见问题

- 所有的群集成员都需要有相同的 **root** 口令吗?

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

- 节点引导的次序有重要意义吗?

  多数情况下并不重要。但是，引导次序对防止失忆很重要（关于失忆的详细信息，请参考第43页的「定额和定额设备」)。例如，如果节点 2 是定额设备的属主而节点 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。

  关于备份和恢复过程的其他信息，请参见《*Sun Cluster 3.0 U1* 系统管理指南》。

群集存储器常见问题

- 多主机存储器为什么具有高可用性?

  多主机存储器具有高可用性，是因为它可以在单磁盘丢失时因镜像（或者由于基于硬件的 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。

群集互连常见问题

- **SunPlex** 系统支持什么样的群集互连?

  目前，SunPlex 系统支持以太网（100BASE-T 快速以太网和 1000BASE-SXGb）群集互连。

- **“**电缆**”**与传输**“**路径**”**有什么不同?

  群集传输电缆配置为采用传输适配器和交换器。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。群集拓扑管理器采用可用的电缆，在节点之间构建端对端的传输路径。电缆不直接与传输路径相对应。

  管理员可静态地“启用”和“禁用”电缆。电缆可处于一种“状态”（启用或禁用），但并非一种“状况”。如果电缆禁用，它就象未经配置一样。禁用的电缆不可用作传输路径。没有对它们进行探测，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”由拓扑管理器确定。路径可处于“联机”或“脱机”状况。使用 `scstat(1M)` 可以查看传输路径的状况。

  考虑下面的例子：用四条电缆来连接双节点群集。

```
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
```

客户机系统常见问题

- 使用群集时是否需要考虑任何特殊的客户机需要或限制?

  客户机系统正如它们连接到其他任何服务器那样，也连接到该群集。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方软件或执行其他配置更改，以使客户可以连接到该数据服务应用程序。有关客户端配置需求的详细信息，

请参见《*Sun Cluster 3.0 U1 Data Services Installation and Configuration Guide*》中的各章节。

管理控制台常见问题

- **SunPlex** 系统是否需要管理控制台?

  需要。

- 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

- 管理控制台需要位于群集**"**附近**"**（比如在同一房间内）吗?

  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。

终端集中器与系统服务处理器常见问题

- **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；与之不同，后续版本不依赖于终端集中器。

- 我知道大多数 **SunPlex** 服务器都使用终端集中器，而 **E10000** 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) 是控制台，它或者使用以太网端口，或者使用 `jtag` 端口。对于 Sun Enterprise E10000 server，总是将 SSP 用于控制台。

- 使用终端集中器有什么益处?

  使用终端集中器提供从网络上任何地方的远程工作站对每个节点的控制台级访问，包括当节点是在 OpenBoot PROM(OBP) 时。

- 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他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?

  可以。注意需要重置的端口号并进行如下操作：

```
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，请参考《*Sun Cluster 3.0 U1 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。您将无法连接到节点控制台，直到集中器恢复工作。

- 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，而不连接到用于其他客户访问的网络。您可以通过限制对该特定网络的访问来控制安全性。

术语汇编

该词汇表用于 SunPlex 3.0 文档。

A

管理控制台 用于运行群集管理软件的工作站。

失忆 因群集配置数据失效而关闭后群集重新启动的一种状况。例如，在一个仅有节点 1 可操作的由两个节点构成的群集中，如果在节点 1 上发生群集配置更改，则节点 2 的 CCR 就会失效。如果群集关闭然后在节点 2 上重新启动，就会因节点 2 的 CCR 失效而出现失忆状况。

自动故障返回 主节点在出现故障后又作为一个群集成员重新启动，然后将资源组或设备组返回到其主节点的一种进程。

B

备份组 请参见“网络适配器故障转移组”。

C

检查点 由一个主节点向一个辅助节点发出的通知，用来使它们之间的软件保持同步。另请参见“主节点”和“辅助节点”。

群集 两个或更多互相连接的节点或域，它们共享一个群集文件系统，并且一同进行配置，来运行故障转移、并行或可伸缩资源。

群集配置库 (CCR) 一个可用性高的、复制的数据存储，供 Sun Cluster 软件长期存储群集配置信息使用。

群集文件系统 一种群集服务，在群集范围内提供对本地现有的文件系统的高可用性访问权限。

群集互连 硬件联网基础设施包括电缆、群集传输联结和群集传输适配器。Sun Cluster 和数据服务使用这些基础设施来进行群集间的通信。

群集成员 当前群集体的活动成员。该成员能够与其他群集成员共享资源，并同时向群集的其他成员和群集的客户机提供服务。另请参见“群集节点”。

群集成员监视器 **(CMM)** 一种用来维护一个一致的群集成员名单的软件。群集软件的其他部分使用该成员信息来确定将高可用性服务放在何处。CCM 可确保非群集成员不会破坏数据或将破坏的数据或不一致的数据传给客户机。

群集节点 已配置成一个群集成员的节点。群集节点可以是当前成员，也可以不是。另请参见“群集成员”。

群集传输适配器 驻留在一个节点上并将该节点连接到一个群集互连的网络适配器。另请参见“群集互连”。

群集传输电缆 连接到端点的网络连接。即群集传输适配器和群集传输结点之间或两个群集传输适配器之间的连接。另请参见“群集互连”。

群集传输结点 一种硬件开关，用作群集互连的一部分。另请参见“群集互连”。

配置 在同一节点上存在的特性。在群集配置过程中运用此概念来提高性能。

D

数据服务 一种应用程序，可以让它在资源组管理器 (RGM) 的控制下作为一个高可用性的资源来运行。

缺省主 其中的故障转移资源类型已处于联机状态的缺省群集成员。

设备组 用户定义的设备资源组，比如磁盘，可以从群集 HA 配置中的不同节点予以控制。该组可以包括磁盘、Solstice DiskSuite 磁盘集和 VERITAS Volume Manager磁盘组的设备资源。

设备 **ID** 用来标识设备的一种机制，通过 Solaris 提供。`devid_get(3DEVID)` 手册页中对设备 ID 进行了描述。

Sun Cluster DID 驱动程序使用设备 ID 来确定不同群集节点上的 Solaris 逻辑名称之间的相关性。DID 驱动程序探测每个设备的 ID。如果该设备 ID 与群集中某个地方的另一设备匹配，则会给予两个设备相同的 DID 名称。如果以前在群集中没有出现过该设备 ID，则会分配一个新的 DID 名称。另请参见“Solaris 逻辑名称”和“DID 驱动程序”。

DID 驱动程序 — 由 Sun Cluster 软件实现的一个驱动程序，用来在群集间提供一个一致的设备名称空间。另请参见“DID 名称”。

DID 名称 — 用来标识 SunPlex 系统中的全局设备。它是一个与 Solaris 逻辑名称具有一对一或一对多关系的群集标识符。其格式为 d *X*s*Y*，其中 *X* 是一个整数，*Y* 是片名称。另请参见“Solaris 逻辑名称”。

磁盘设备组 — 参见“设备组”。

分布式锁定管理器 **(DLM)** — 共享磁盘 Oracle Parallel Server (OPS) 环境中使用的锁定软件。DLM 使不同节点上运行的 Oracle 进程能够同步数据库访问。DLM 在设计上具有高可用性。如果某个进程或节点崩溃，其余的节点不必关闭或重新启动。执行 DLM 的快速重新配置，可以从故障中恢复。

磁盘集 — 参见“设备组”。

磁盘组 — 参见“设备组”。

E

端点 — 群集传输适配器上的物理端口或群集传输结点。

事件 — 受管对象的状态、主控、严重性或描述的改变。

F

故障返回 — 参见“自动故障返回”。

故障快速防护 — 在发现不正确的操作造成破坏之前，有序地关闭或移除群集中的一个故障节点。

故障转移 — 出现故障后，当前主节点的一个资源组或设备组自动重新定位到一个新的主节点。

故障转移资源　一种资源，其每一项资源每次只能由一个节点正确控制。另请参见“单实例资源”和“可伸缩资源”。

故障监视器　故障守护程序和用来探测数据服务的各种部分并采取措施的程序。另请参见“资源监视器”。

G

普通资源类型　数据服务模板。普通资源类型可用来使一个简单应用程序成为一个故障转移数据服务（在一个节点上停止，在另一节点上开始）。此类型不需要按 SunPlexAPI 编程。

普通资源　一个应用程序守护程序及其子进程，在资源组管理器的控制下用作一个普通资源类型的一部分。

全局设备　从所有群集成员都可以访问的一个设备，比如磁盘、CD-ROM 和磁带。

全局设备名称空间　包含用于全局设备的逻辑、群集范围名称的名称空间。Solaris 环境中的本地设备在 `/dev/dsk`、 `/dev/rdsk` 和 `/dev/rmt` 目录中定义。全局设备名称空间在 `/dev/global/dsk`、 `/dev/global/rdsk` 和 `/dev/global/rmt` 目录中定义全局设备。

全局接口　物理上主机共享地址的全局网络接口。另请参见“共享地址”。

全局接口节点　托管全局接口的节点。

全局资源　在 Sun Cluster 软件的内核级别提供的高可用性资源。全局资源可包括磁盘（HA 设备组）、群集文件系统和全局联网。

H

HA 数据服务　参见“数据服务”。

心跳　在所有可用的群集互连传输路径中定期发送的消息。在指定的时间间隔或数次重试后仍缺少心跳，可能会触发传输通信向另一路径的内部故障转移。一个群集成员的所有路径均失败会导致 CMM 重新评估群集定额。

I

实例	参见“资源调用”。

L

负载平衡	仅适用于可伸缩服务。在群集的节点间分布应用程序负载的过程，旨在使客户机的请求能够及时得到满足。有关详细信息，请参见第52页的「可伸缩数据服务」。
负载平衡策略	仅适用于可伸缩服务。在节点间分配应用程序请求负载所用的首选方式。有关详细信息，请参见第52页的「可伸缩数据服务」。
本地磁盘	在物理上专用于一个给定群集节点的磁盘。
逻辑主机	Sun Cluster 2.0（最低）的概念，包括一个应用程序、磁盘集或驻留应用程序的磁盘组，以及用来访问群集的网络地址。在 SunPlex系统中已不存在这种概念。有关此概念如何在 SunPlex 系统中实现的说明，请参见第37页的「磁盘设备组」和第59页的「资源、资源组和资源类型」。
逻辑主机名资源	包含一个表示网络地址的逻辑主机名的集合的资源。每次只能有一个节点控制逻辑主机名资源。另请参见“逻辑主机”。
逻辑网络接口	在 Internet 体系结构中，一个主机可以有一个或多个 IP 地址。Sun Cluster 软件配置更多的逻辑网络接口来在几个逻辑网络接口和一个物理网络接口之间建立映射。每个逻辑网络接口各有一个 IP 地址。这种映射可以使单个物理网络接口能够响应多个 IP 地址。这种映射也可以使 IP 地址能够在发生接管或切换的时候从一个群集成员移到其他群集成员，而不需要额外的硬件接口。

M

主 **(master)**	参见“主节点”。
元设备状态数据库复制（复制）	存储在磁盘上的一个数据库，记录所有元设备的配置和状态以及错误情况。这些信息对 Solstice DiskSuite 磁盘集的正常运行非常重要，并且它是复制的。
多重地址主机	位于多个公共网络上的主机。
多主机磁盘	物理上连接到多个节点的磁盘。

N

网络适配器故障转移 **(NAFO)** 组
一个网络适配器或同一子网中的同一节点上多个网络适配器，经配置后，能够在一个适配器出现故障时相互备份。

网络地址资源
请参见“网络资源”。

网络资源
包含一个或多个逻辑主机名或共享地址的资源。另请参见“逻辑主机名资源”和“共享地址资源”。

节点
能够作为 SunPlex 系统的一部分的物理主机或域（在 Sun Enterprise E10000 server 中）。又称“主机”。

非群集模式
通过使用 –x 引导选项引导一个群集成员而达到的结果状态。在这种状态下，节点已不再是群集成员，但仍是一个群集节点。另请参见“群集成员”和“群集节点”。

P

并行资源类型
一种资源类型（如并行数据库），已经设置为在群集环境中运行，从而可以同时由多个（两个或更多）节点控制。

并行服务实例
在个别节点上运行的并行资源类型实例。

潜在主
参见“潜在主节点”。

潜在主节点
能够在主节点失败时控制一个故障转移资源类型的群集成员。另请参见“缺省主”。

主节点
资源组和设备组当前在其上处于联机状态的节点。换言之，主节点就是当前托管或实现与资源相关的服务的节点。另请参见“辅助节点”。

主要主机名
主要公共网络上的节点名称。这通常是在 `/etc/nodename` 中指定的节点名称。另请参见“辅助主机名”。

专用主机名
用以通过群集互连来与节点通信的主机名别名。

公共网络管理 **(PNM)**
一种软件，通过故障监视和故障转移来避免因单个网络适配器或电缆故障而失去节点可用性。PNM 故障转移使用称为网络适配器故障转移组的若干组网络适配器，来提供群集节点和公共网络之间的冗余连接。故障监视和故障转移功能协同作用，用来确保资源的可用性。另请参见“网络适配器故障转移组”。

Q

定额设备
两个或多个节点共享的磁盘，用以在选票中达到一个定额，使群集能够运行。只有达到了定额票数，群集才能运行。当群集分成若干单独的节点组时，定额设备用来确定哪些节点组构成新的群集。

R

资源
资源类型的实例。同一类型可能包含许多资源，每一项资源都有自己的名称和特性值组，从而使基础应用程序的许多实例可以在群集上运行。

资源组
由 RGM 按单元管理的资源集合。RGM 管理的每一资源都必须在资源组中配置。通常，相关的和相互依赖的资源会归为一组。

资源组管理器 (RGM)
一种软件工具，通过在选定的群集节点上自动打开和关闭群集资源，来使这些资源具有高可用性和高可伸缩性。在发生硬件或软件故障或重新启动时，RGM 按预先配置的策略工作。

资源组状态
在任何给定的节点上的资源组状态。

资源调用
一种资源类型在节点上运行的实例。它是一个抽象概念，表示在节点上启动的一个资源。

资源管理 API (RMAPI)
SunPlex 系统中的应用程序编程接口，使应用程序在群集环境中具有高可用性。

资源资源监视器
资源类型实现中的一个可选部分，它定期对资源进行故障探测，以确定它们的运行是否正常以及性能如何。

资源状态
给定节点上 Resource Group Manager 资源的状态。

资源状况
故障监视器报告的资源情况。

资源类型
数据设备、LogicalHostname 或 SharedAddress 群集对象的唯一名称。数据服务资源类型既可以是故障转移类型，也可以是可伸缩性类型。另请参见“数据服务”、“故障转移资源”和“可伸缩资源”。

资源类型特性
一个关键值对，由 RGM 存储为资源类型的一部分，用来描述并管理给定类型的资源。

S

可伸缩相关接口 **(SCI)**	用作群集互连的一种高速互连硬件。
可伸缩资源	在多个节点上运行的资源（每个节点上的一个实例），它使用群集互连，因而对服务的远程客户机来说，它看起来像单一服务。
可伸缩服务	实现的一种数据服务，在多个节点上同时运行。
辅助节点	在主节点发生故障时可用于主控磁盘设备组和资源组的群集成员。另请参见“主节点”。
辅助主机名	用来在访问辅助公共网络节点的名称。另请参见“主要主机名”。
共享地址资源	群集节点上运行的所有可伸缩服务均可绑定的一个网络地址，用来使服务在这些节点上可进行伸缩。群集可拥有多个共享地址，一种服务可以绑定到多个共享地址。
单一实例资源	该资源在群集间至多只有它一个处于活动状态。
Solaris 逻辑名称	通常用来管理 Solaris 设备的名称。对磁盘来说，这通常看起来有些类似 `/dev/rdsk/c0t2d0s2` 的形式。这些 Solaris 逻辑设备名称中，每个名称都有一个基本的 Solaris 物理设备名称。另请参见“DID 名称”和“Solaris 物理名称”。
Solaris 物理名称	Solaris 中的设备驱动程序为设备选定的名称。这在 Solaris 机器上显示为 `/devices` 目录树下的路径。例如，典型 SCSI 磁盘的 Solaris 物理名称类似于： `/devices/sbus@1f,0/SUNW,fas@e,8800000/sd@6,0:c,raw` 另请参见“Solaris 逻辑名称”。
Solstice DiskSuite	SunPlex 系统使用的卷管理器。另请参见“卷管理器”。
群集分割	一个群集分成多个分区的情况，每个分区在形成时不知道任何其他子群集的存在。
切换返回	参见“故障返回”。

切换

资源组或设备组从群集中的一个主（节点）向另一个主（若资源组已为多个主节点配置，则是向多个主）的有序传输。切换是由管理员通过使用 `scswitch(1M)` 命令启动的。

系统服务处理器 **(SSP)**

在 Enterprise 10000 配置中群集外部的一种设备，专门用来与群集成员通信。

T

接管

参见“故障转移”。

终端集中器

在非 Enterprise 10000 配置中群集外部的一种设备，专门用来与群集成员通信。

V

VERITAS Volume Manager

SunPlex 系统使用的卷管理器。另请参见“卷管理器”。

未使用的卷管理器

通过磁盘条带化、链接、镜像和元设备或卷的动态增长来提供数据可靠性的一种软件产品。