Sun Cluster 3.*x* 基础

因为专注，所以
专业！

黄雨

2008.09.08

内容纲要

- 集群/群集的基本概念
- Sun Cluster的硬件和软件环境
- Sun Cluster的集群控制台
- Sun Cluster的拓扑与仲裁机制
- Sun Cluster的安装准备工作

集群/群集概述

- 定义
- 分类
- 标配
- 目标
- HA与Scalable

1.1 集群的定义

- 多台独立的服务器组成一台逻辑上的主机，对外提供统一的服务
- 常见的集群软件
  - HP　　MC/Service-Guard
  - IBM　　HACMP
  - SUN　　Sun Cluster
  - Fujitsu　Prime Cluster（PCL）
  - EMC　　AutoStart
  - Symantec VERITAS Cluster Server（VCS）
  - RedHat　RHCS
  - Oracle　RAC
  - ……

1.2 集群的分类

- 高性能集群（High performance cluster，HPC）
  - 多个节点共同完成一个任务，多用于科学运算，比如天气预报、环境监控等数据量大、计算复杂的环境中，在商用环境中很少使用，比如3DMAX
- 高可用性集群（High availability cluster，HAC）
  - 利用集群中冗余的系统，当主系统出故障时由备机接管相应的应用，如HACMP/VCS
- 负载均衡集群（Load balance cluster， LBC）
  - 可伸缩性集群（Scalable Cluster）
  - 多台主机分担来自所有用户的并行的小的工作，比如Oracle的RAC，WebLogic Cluster

1.3 集群的标准配备

- **独立的服务器节点**
  - 每个节点拥有自己的（独立的，非共享的）操作系统
- **专用的互连硬件**
  - 用于在同一个集群内进行专用的数据通信
- **多端口存储**
  - 一个集群内至少有两个节点和存储有物理连接，提供至少两条对存储的访问路径（分别经过两个节点，一个节点一条路径），为在集群中运行的应用提供数据存储服务

1.4 集群的基本目标

- HA和Scalability
  - 集群的目标是为集群中运行的应用提供高可用性（HA，high-availability）和可伸缩性（scalability）服务
- 支持多种应用
  - 无集群意识（cluster-unaware）的应用
  - 有集群意识（cluster-ware）的应用

1.5 HA和Scalable

- HA的定义
  - 高达5个9的可用性（99.999%）
  - 一年的宕机时间不超过5分钟
  - 单台硬件设备无法实现HA
- Scalable的定义
  - 应用同时运行在多个节点
  - 某节点出故障时，该节点承担的负荷自动转移到其他节点（缩），恢复后集群会自动会为该节点分配负荷（伸）
  - 与HA并不矛盾

Sun Cluster环境

- 硬件环境
- 软件环境
- 应用类型
- Sun Cluster的软件框架
- 全局命名、全局设备、全局文件系统

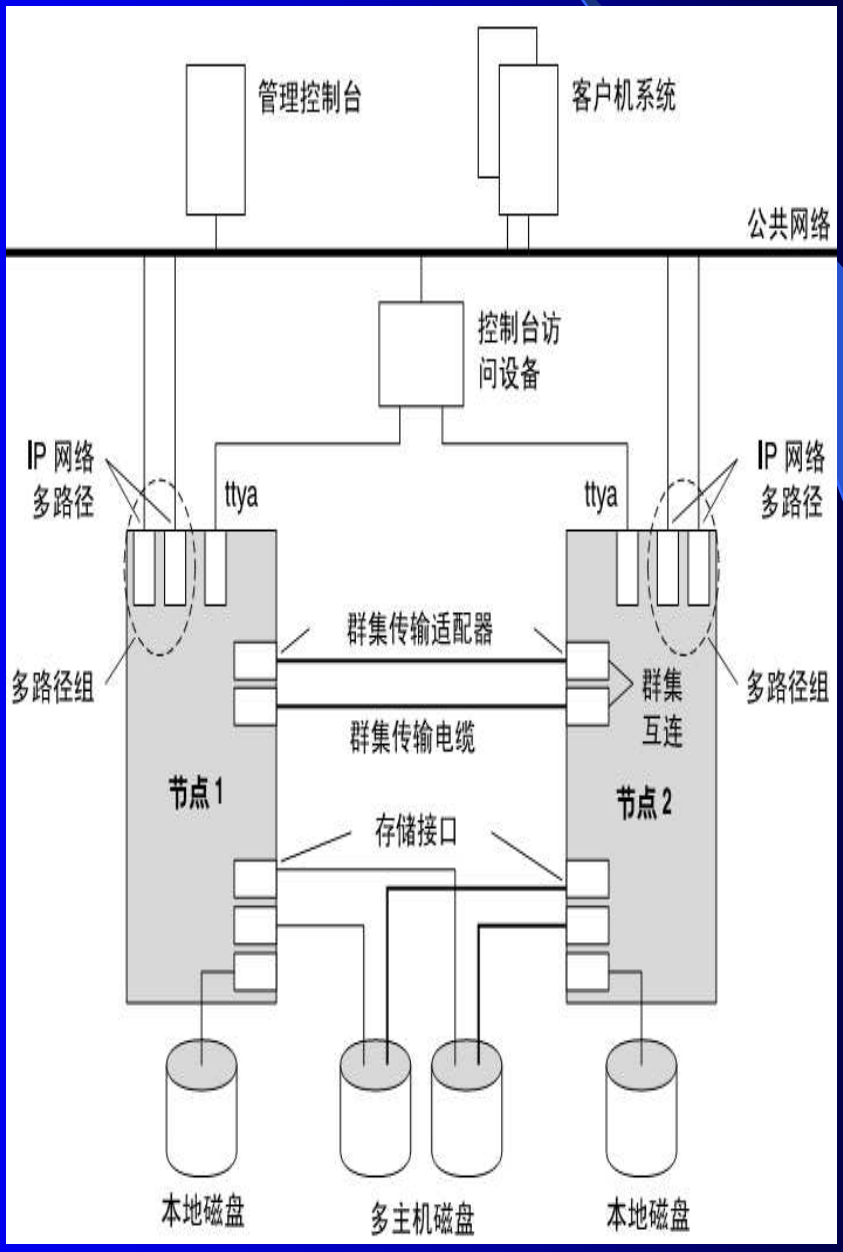
2.1 Sun Cluster 3.*x*

- Sun Cluster是SUN公司的集群软件产品
- 少有的可支持小型机和PC Server的产品
  - 分为for SPARC和for X86版本
- 免使用许可证的优秀产品
- **专门针对Sun Cluster的认证**
  - **Sun集群3.2软件认证系统管理员(CX-310-345)**
  - **链接见备注区**

2.2 Sun Cluster的特点

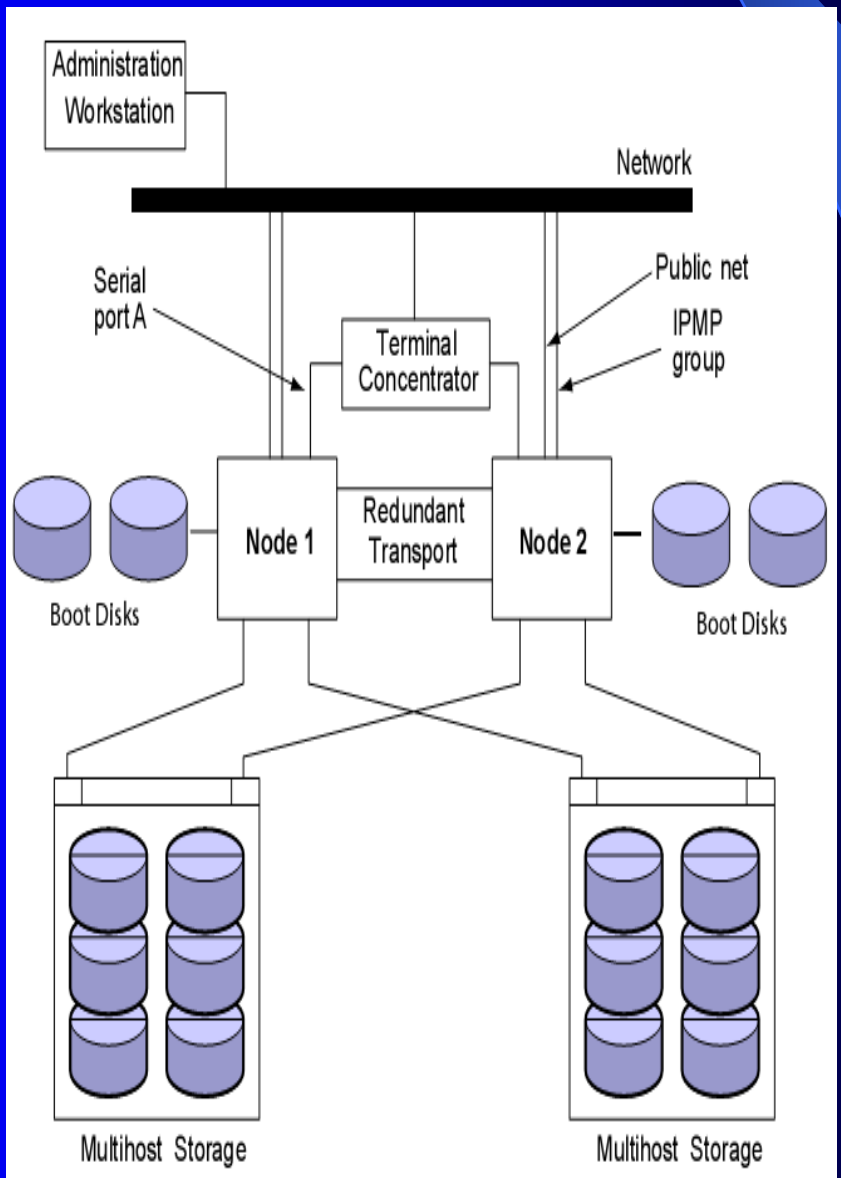
- Sun Cluster 软件最新版本为3.2
  - 支持2～16个节点
  - 全局设备
  - 全局文件系统
  - 集群框架服务直接嵌入内核中，更稳定
  - 内嵌多种已为各种应用定制好的数据服务代理
  - 通过内置的负载均衡（全局接口），使部分事先定制的应用能够实现可伸缩性服务
  - 详见备注

2.3 Sun Cluster硬件环境



2.3.a 双节点集群（拓扑）

- 解释见备注



2.3.2 集群所支持的硬件平台

- Sun Cluster集群环境支持大量的Sun硬件平台，从机架式的服务器（Netra T1 M100），到大型企业级服务器，包括Sun Fire 15K等
- Sun Cluster集群环境同样支持大量异构环境，即在一个集群内的节点可以是不同类型的服务器。这取决于网络和存储主机适配器（storage host adapter），而不是服务器本身。

2.3.3 集群传输接口（1）

- 一个集群内的所有节点都通过集群专用传输连接到一起（集群专用互连），群集专用互连必须是冗余的链路（双链路），作用如下：
  - 集群范围内监视和恢复
  - 全局数据存取（这个操作对应用而言是透明的）
  - 为有集群意识的应用（cluster-aware applications）提供特定的传输（比如Oracle Paraller Server）
- 集群专用互联至少需要两个独立的“专用网络”，在特定的环境中，甚至可以定义更多的链路用于专用互联，比如在进行全局数据访问时，流量可以以类似条带的方式分布在所有的专用互联链路上。

2.3.3 集群传输接口（2）

- 双节点集群通常使用交叉电缆（也可选用交换机，有点浪费）；当集群内的节点数多于两个时，必须使用交换机进行互联。
- 以下几种类型的硬件可用于集群传输互联（心跳）：
  - 以太网（100Mb或Gigabit），绝大多数是使用这种类型的集群互联
  - 用于RSM（remote shared memory，远程内存共享）应用的SCI（Scalabel coherent interface，可扩展性一致接口）
  - Sun Fire 专用的互联硬件，用于RSM应用，支持的型号有Sun Fire 3800-6800，15K，25K

2.3.4 公共网络接口

- 每个节点必须有公共网络接口用于传输数据，公共网络接口由Solaris IP多路径软件（IPMP，IP Multipathing）控制。强烈建议每个节点都至少有两个接口（构成一个IPMP组）连接到每个子网。
- 在集群中的绝大多数应用都有这样的需求：集群内可能运行该应用的节点必须处在同一个子网内。
- Sun Cluster集群内的节点可以连接到多个子网，但不能作为路由器使用

2.3.5 集群磁盘存储

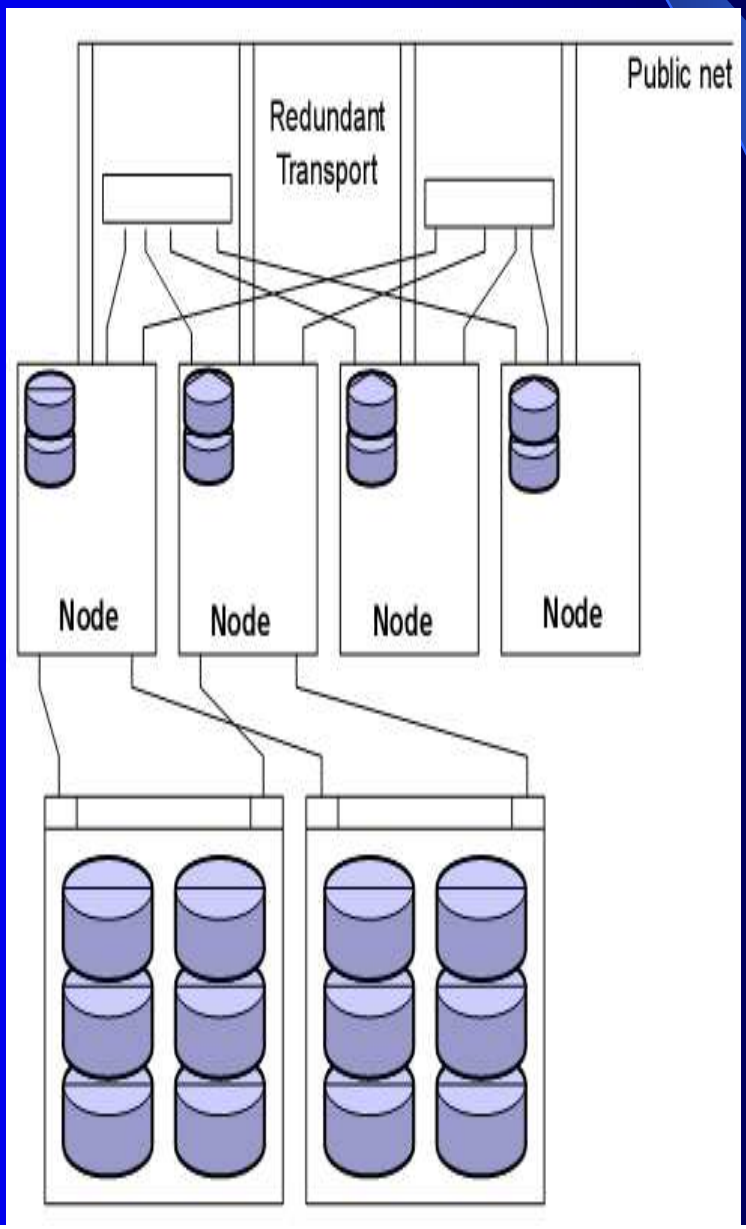
- Sun Cluster集群的硬件环境中支持几种型号Sun的存储设备，这些存储设备必须支持多主机连接。Sun StorEdge T3仅能连接一台主机，因此要通过专用的hub或交换机扩展端口才可以在Sun Cluster集群环境中使用。
- 绝大部分存储支持两台主机物理连接，部分存储支持多台节点物理连接到存储
- 通过VxVM（VERITAS卷管理器）或SVM（Solaris卷管理器），可以对存储进行跨控制器的镜像

2.3.6 启动盘

- 每个节点的启动盘都必须是本地硬盘，而不能是从多端口存储阵列上映射过来的盘。推荐使用两个本地硬盘，并通过VxVM或SVM做镜像，然后优先选择从其中一块盘启动。

2.3.b 多节点集群（拓扑）

- 解释见备注



2.3.c 对硬件环境的要求

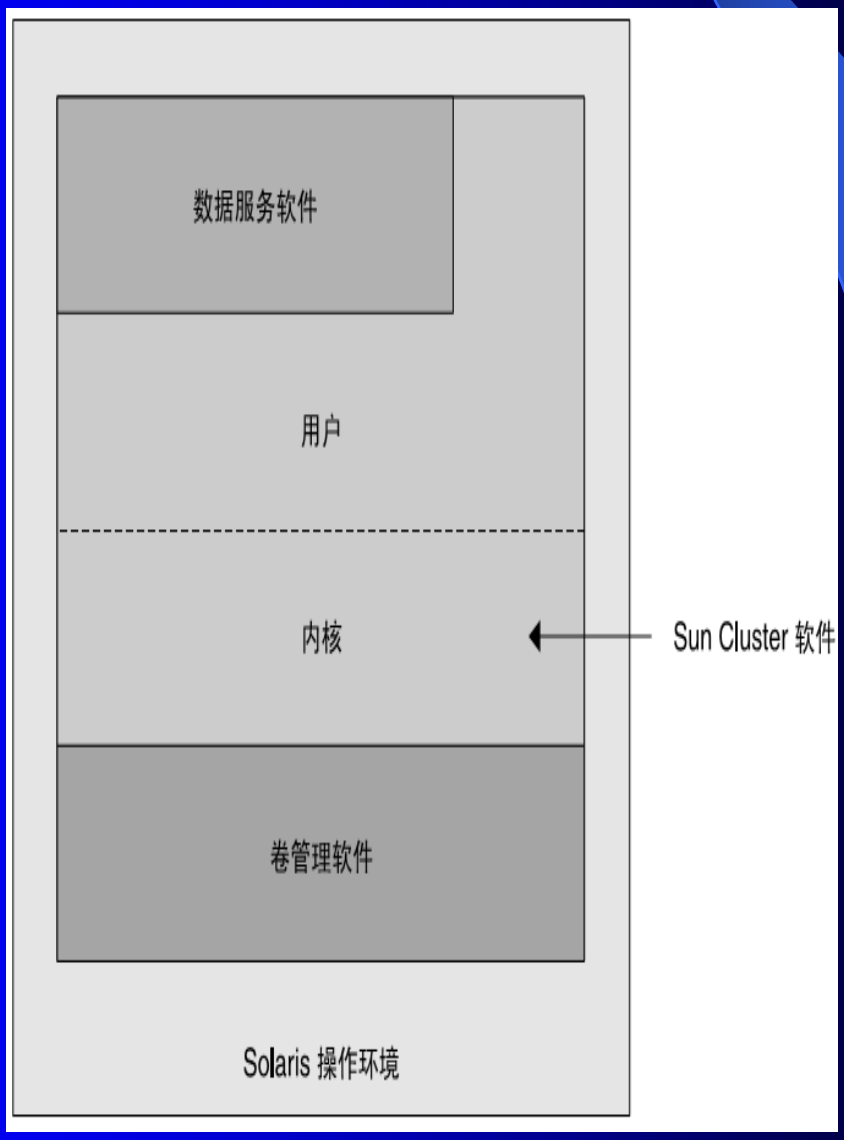
- 要构成Sun Cluster集群环境，必须满足以下硬件要求：
  - 必须：冗余服务器节点
  - 必须：冗余传输（心跳网卡）
  - 必须：冗余存储阵列
  - 必须：跨数据控制器进行软件镜像
  - 推荐：每子网实现冗余公共网络接口
  - 推荐：冗余启动盘
  - 可选：硬件RAID存储阵列
- 解释见备注

2.3.d 基于域的集群

- Sun Cluster能够部署在基于域技术的主机，比如：
  - Sun Fire 15K/25K
  - Sun Fire 3800-6800
  - Sun Enterprise 10000
- 基于域部署集群，远不如来自独立服务器的集群可靠。
- 解释见备注

2.4 Sun Cluster软件环境

- 解释见备注



2.5 应用类型

- Sun Cluster软件环境支持高可用性和可伸缩性应用
  - 无集群意思的应用
  - 有集群意思的应用

2.5.1 无集群意识的应用

- 集群内的绝大多数应用属于无集群意识的应用（cluster-unware applications），此种应用可分为两种类型：
  - 失效切换应用（主备模式，Failover applications）
  - 可伸缩性应用（负载均衡，Scalable applications）
- 不管是哪种类型，都含有以下要素：
  - 集群资源组管理器（RGM，resource group manager），负责掌控所有的资源起停操作。这些起停操作绝对不能由传统的（Solaris的）运行控制脚本来实现。
  - 通过特定的应用的数据服务代理，把应用和Sun Cluster胶合在一起，使应用能够在集群环境中正确的工作。包括在集群内正确的起停应用的方法，应用特定的故障检测器等。

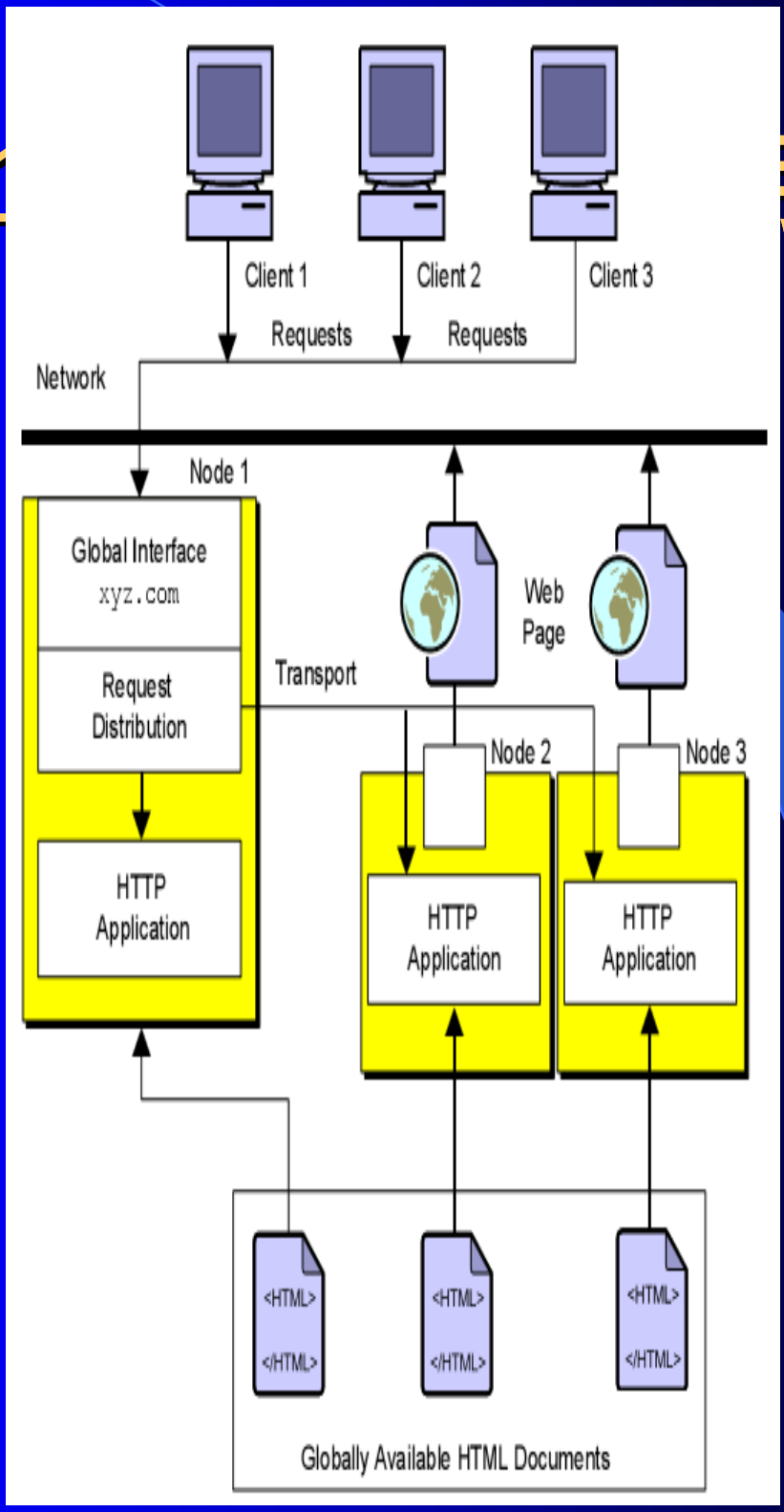
2.5.1a Failover应用

- Failover是集群中最容易实现的模式。Failover应用同一时间只在一个节点上运行；通过（在同一个节点或另一个节点上）自动重启服务来实现高可用性。
- 通常由两个节点构成Failover应用，对外提供一个应用专用的ip地址；当执行失效切换时，这个ip地址总是随应用从一个节点切到另外一个节点；对于客户机而言，相当于一台逻辑主机在为它提供服务，而不会意识到服务是在哪个节点上运行也不会知道此服务是由集群提供的。
- 在同一个资源组（resource group）内的多个应用可以共享一个IP地址，这种情况下，这些应用必须/只能同时在一个节点上运行（不推荐这种方式）

2.5.1b Scalable应用(1)

- 可伸缩性应用指的是在一个集群内同时运行多个实例（一个节点一个实例），通过全局接口的方法，仅对外提供一个ip地址并实现负载均衡，使其看起来就像一个单一的服务一样。
- 可伸缩性应用也是现成定制好的（off-the-shelf），不是所有应用都可以配置成可伸缩性应用。写数据时没有任何锁机制的应用，应当以failover模式运行，而不是配置成可伸缩性应用。
- 比如apache服务和Sun ONE Web Server应用服务

2.5.1 (2)


Client 1
Client 2
Client 3
Requests
Requests
Network
Node 1
Global Interface
xyz.com
Request
Distribution
HTTP
Application
Transport
Web
Page
Node 2
Node 3
HTTP
Application
HTTP
Application
<HTML>
</HTML>
<HTML>
</HTML>
<HTML>
</HTML>
Globally Available HTML Documents

- 解释见备注

2.5.2 有集群意识的应用

- 有集群意识的应用是指那些在软件中内置了集群功能的应用，有集群意识的应用和无集群意识的应用的主要区别点在于：
  - 运行在不同节点上的应用的多个实例能够互相意识到各自的存在，并且通过专用传输网络（private transport）进行信息交换。
  - 无需Sun Cluster软件框架中的RGM来起停这些应用。因为这些应用是有集群意识的，它们能够采用自带的脚本来启动，或手工启动。
  - 有集群意识的应用不需要通过外部应用IP地址（application ip address）把它们逻辑上编成一组

2.5.2a 并行数据库应用

- 并行数据库应用是一种特性的集群应用。数据库服务器的多个实例在集群中运行，掌控对同一个数据库的不同查询，甚至能够对大型查询提供并发查询能力。并行数据库应用的例子有：ORACLE 8i的Parallel Server（OPS），Oracle 9i/10g的Real Application Cluster（RAC）。

2.5.2b 其他的远程内存共享（RSM）应用

- 如果组成Sun Cluster的硬件设备包含SCI或Sun Fire专用的链路互联设备时，这些应用能够利用API（application programming interface）调用RSM，从在一个节点上运行的实例映射数据到在另外一个节点上运行的实例的地址空间。对于有集群意识的应用而言，这是一种非常有效的方式，通过专用互连共享大量的数据。
- Sun Cluster 3.1只支持ORACLE RAC在集群环境中使用RSM。

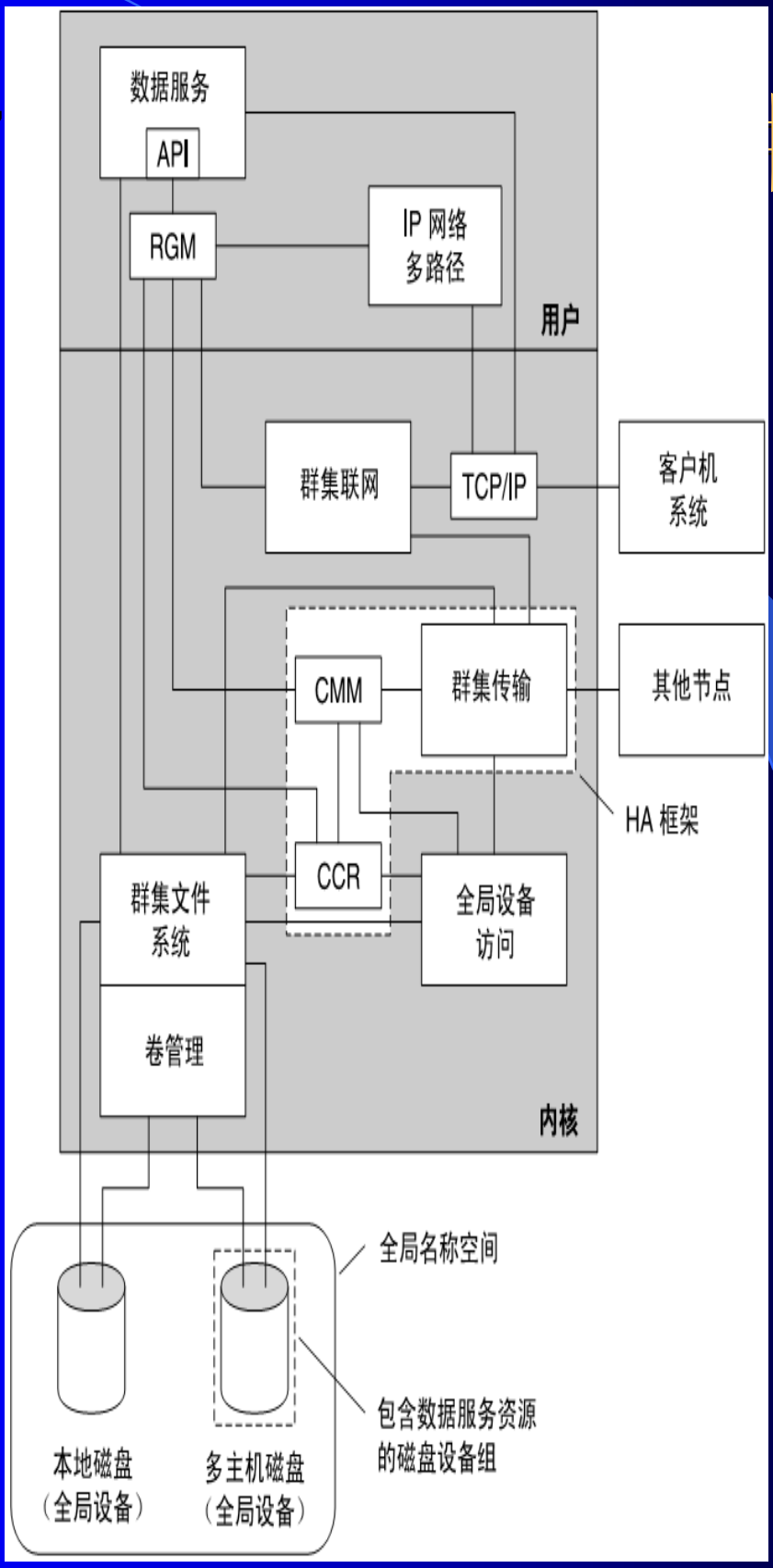
2.6 Sun Cluster软件的数据服务支持

- Sun Cluster提供了大量的数据服务的代理
- 数据服务代理实现了无集群意识的软件的高度可用性，无论配置成failover方式还是scalable方式。
- 这些代理包括：nfs，apache，dns，samba，dhcp…

2.7 Sun Cluster的HA框架

- Sun Cluster软件框架是在软件层面给位于集群中的节点提供常规集群服务，而不关注在集群中运行的是哪种应用。Sun Cluster软件框架通过一系列守护进程和内核模块来实现。
- Sun Cluster软件环境的一个优点是它的框架是在内核中实现，具有速度快，常驻内存，及稳定等特点。
- Sun Cluster软件框架提供的核心服务包括（但不限于）：
  - CMM，节点故障监测和集群成员监测
  - 网络故障监测（公共网络和集群互连网络）
  - CCR，集群配置库

2.7 体


数据服务
API
RGM
IP 网络
多路径
用户
群集联网
TCP/IP
客户机
系统
CMM
群集传输
其他节点
HA 框架
CCR
群集文件
系统
全局设备
访问
卷管理
内核
全局名称空间
包含数据服务资源
的磁盘设备组
本地磁盘
（全局设备）
多主机磁盘
（全局设备）

2.7.1 节点故障监测和集群成员监测（CMM）

- 集群成员监测器（CMM，cluster membership monitor）在每个节点上常驻内核（kernel-resident），监测主要的集群状态变化，比如一个或多个节点间的通信丢失。CMM依赖于传输内核模块（transport kernel module）生成心跳，通过传输介质（transport medium，心跳网卡）传送给集群中的其他节点。如果在一个预定义的时间周期内没有收到来自其他任何节点的心跳，则认为发生失效事件，并开始进行集群重配置操作，重新协商集群的成员资格。

2.7.2 网络故障监测

- Sun Cluster软件框架会监测公共网络接口和集群传输接口（cluster transport interface）可能发生的故障。
- **公共网络接口管理**
  - Sun Cluster软件环境需要使用IPMP（Solaris OS的标配工具之一），用来处理节点的接口故障。Sun Cluster软件添加了一个监测层，用以监测一个节点上所有的网络故障，并把有能力进行失效切换的应用切换到另外一个节点。
- **集群传输监测**
  - 每个节点都监测集群传输接口。如果监测到某节点上一个活跃的集群互连接口变成不起作用，所有的节点会把互连通信转移到可用的传输接口。这种故障对于Solaris Cluster集群中的软件应用是透明的。

2.7.2 集群配置库-1（CCR）

- 常规集群配置信息保存在全局配置文件中（通常被称为做集群配置库，cluster configuration repository，CCR）。所有节点的CCR必须保持一致，CCR的一个很重要的作用是能够让每个节点意识到自己的潜在角色是作为一个指定的备用系统。
- 不要手工去修改任何与CCR有关的文件。这些文件中包含有一个配置版本号信息，此信息对集群软件的正常工作起到非常重要的作用。但执行管理任务的命令执行或集群状态变化时，CCR的信息会自动改变。

2.7.2 集群配置库-2（CCR）

- CCR包括有以下几种类型的信息：
  - 集群和节点的名字
  - 集群传输（心跳互连）配置
  - VERITAS磁盘组或Solaris卷管理器的磁盘集的名字
  - 每个磁盘组的管理节点的名字
  - 数据服务的操作参数值（timeouts）
  - 到数据服务回叫方法的路径（path to data service callback methods）
  - 磁盘ID（DID）设备配置
  - 集群的当前状态
- 每当有错误或恢复状况发生，或集群的常规状态发生变化时，例如一个节点离开或加入集群时，就会对CCR发生存取操作（CCR is accessed）

2.8 全局命名、设备和文件系统服务

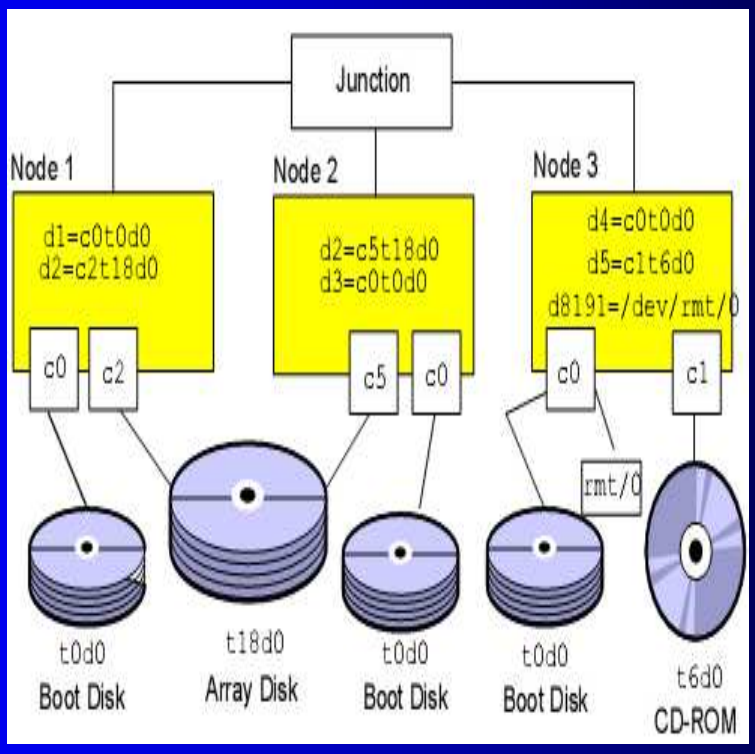
- Sun Cluster软件框架提供全局存储服务，这些特性不仅可以让可伸缩性应用能够在集群中运行，而且，能够为运行在非直连存储的节点上的failover服务提供更加灵活的环境。
- 必须能够理解这些服务之间的区别和联系：
  - 全局命名（DID设备）//global naming（DID devices）
  - 全局设备
  - 全局文件系统

2.8.1.a 全局命名/磁盘ID设备（DID）

- DID特性为集群中的每个磁盘驱动器、光驱、磁带机提供了一个唯一的设备名字。
- 多端口磁盘（存储上的磁盘）在不同的节点通常会有不同的逻辑名（因为控制器编号不同），DID特性会为多端口磁盘分配一个集群内唯一的DID实例号。
- 而在每个节点本地的磁盘，即使使用同样的逻辑名（比如都叫c0t0d0），将会被分配一个不同的唯一DID实例号。

2.8.1.b 全局命名/磁盘ID设备（DID）

- 本地命名和全局命名比较
- 下图显示了来自存储的磁盘在节点1和节点2上的逻辑名不同



2.8.1.c 全局命名/磁盘ID设备（DID）

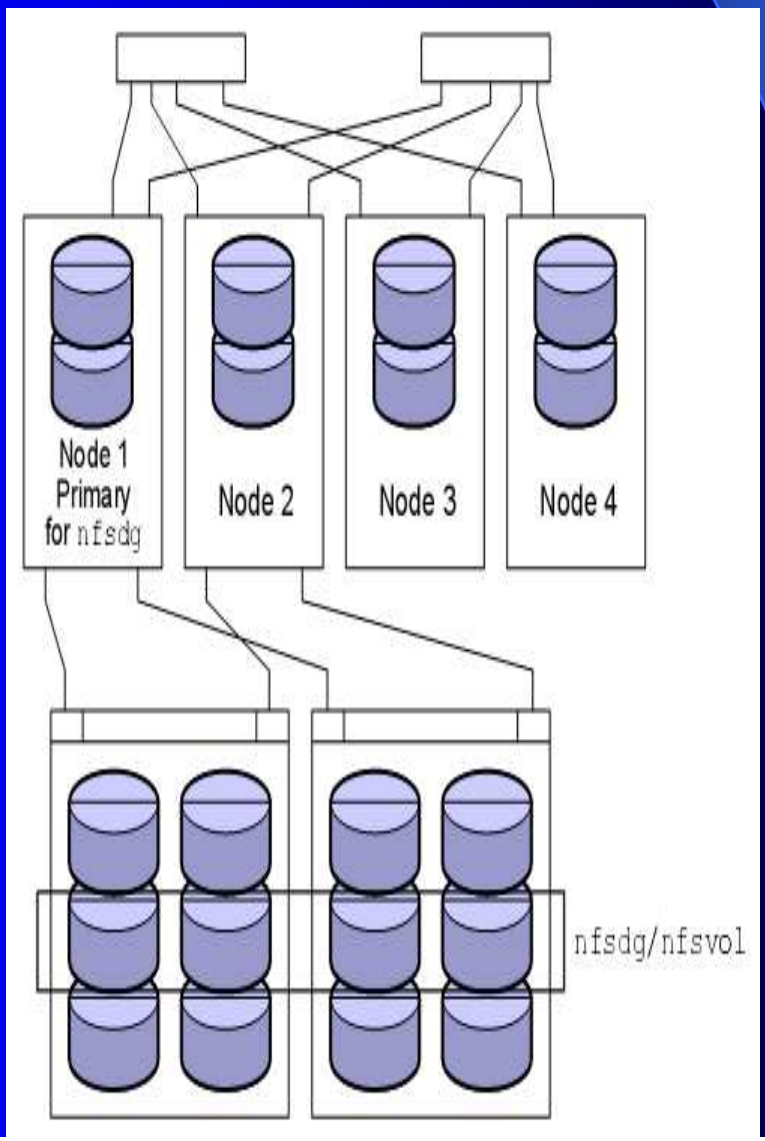
- 通常会为每个磁盘创建8个Solaris磁盘分区设备文件，位于/dev/did/dsk和/dev/did/rdsk目录中。比如：/dev/did/dsk/d2s3和/dev/did/rdsk/d2s3。
- 需注意的是，DID仅仅作为一种全局命名方案（global naming scheme），而不是全局访问方案（global access scheme）。
- DID可以作为Solaris卷管理器创建的卷的成员（components），可以作为集群仲裁设备；但不能作为VxVM卷的成员。

2.8.3.a 全局设备

- 全局设备特性能够让所有的节点同时访问所有的存储设备，即使某些节点并没有和存储有物理连接。包括私有的DID磁盘设备，光驱和磁带，包括VxVM卷和SVM元设备（Solaris卷管理器的metadevices）
- 全局设备最常用在使用VxVM和SVM设备的集群中，卷管理软件根本不会意识到集群中已经部署了全局设备服务。
- Sun Cluster软件框架负责管理全局设备组的主节点（primary node）的失效切换。所有节点都使用同样的设备路径，但对于特定的设备而言实际上只有主节点通过存储介质和磁盘设备交流。所有的其他节点利用集群互连通过与主节点通信实现对设备的访问。

2.8.3.b 全局设备

- 下图显示了所有的节点都能够同时访问设备/dev/vx/rdsk/nfsdg/nfsvol，当节点1失效时，节点2将成为主节点。



2.8.3.c 全局设备

- 请注意全局命名（DID设备）和全局设备的区别，DID实现了对来自存储的磁盘的统一命名（仅是一个名字而已），不能实现全局访问。全局设备则是指来自存储的物理设备，任意时刻，只能有一个主节点负责掌管某个特定的存储磁盘，并通过Sun Cluster软件框架使其变成全局设备供其他节点访问（比如Oracle的裸设备读写数据）。

2.9.1.a 全局设备的设备文件

- Sun Cluster集群中的每个节点上都有一个特殊的文件系统，**专门用来存放全局设备的设备文件**，挂接点为/global/.devices/node@nodeID，nodeID是一个集群内具有唯一性的按节点数递增的整数，每个整数代表集群中的一个节点。这个文件系统通常被称为全局设备存放空间，每个节点都必须为此文件系统留出一个专用的分区，就像为metadb留出一个专用的分区一样，通常是在启动硬盘上预留此分区。
- 集群内所有的/global/.devices/node@nodeID文件系统对集群内所有的节点都是可见的。换而言之，这些文件系统都是全局文件系统
- 解释见备注

2.9.2 全局文件系统

- Sun Cluster的全局文件系统特性使文件系统能够被集群中所有的节点同时访问，而不管实际的物理连接情况。
- 全局文件系统的这个能力不依赖于磁盘上的实际的文件系统是如何实现的。无论是UNIX文件系统（ufs），VERITAS文件系统（vxfs），还是hsfs，都可以支持。
- Sun Cluster通过使用“global”挂接选项使文件系统变成全局可用。通常在/etc/vfstab文件中定义，也可以在使用mount命令挂接文件系统时指定global选项
- mount -o global,logging /dev/md/dg-nfs/dsk/d100 /global/nfs
- 在/etc/vfstab中的写法：
- /dev/md/dg-nfs/dsk/d100 /dev/md/dg-nfs/rdsk/d100 /global/dg-nfs/data ufs 2 yes global,logging

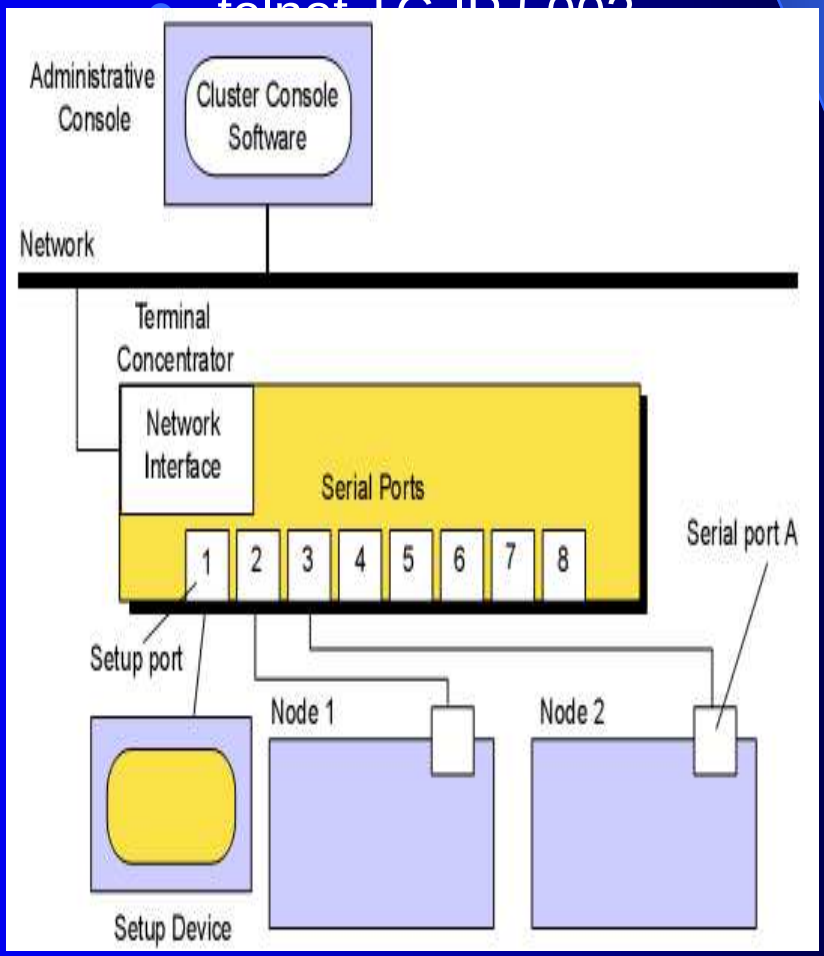
- 解释见备注

控制台访问

- 控制台访问设备
  - Serial Port（TTYA）
  - TC
  - SSP
  - SC
- 集群控制台

3.1 Terminal Concentrator

- TC通常也叫NTS（网络终端服务器，network terminal server）
- 以下命令将连接到与TC的串口2连接的节点的控制台
  - telnet TC_IP 5002



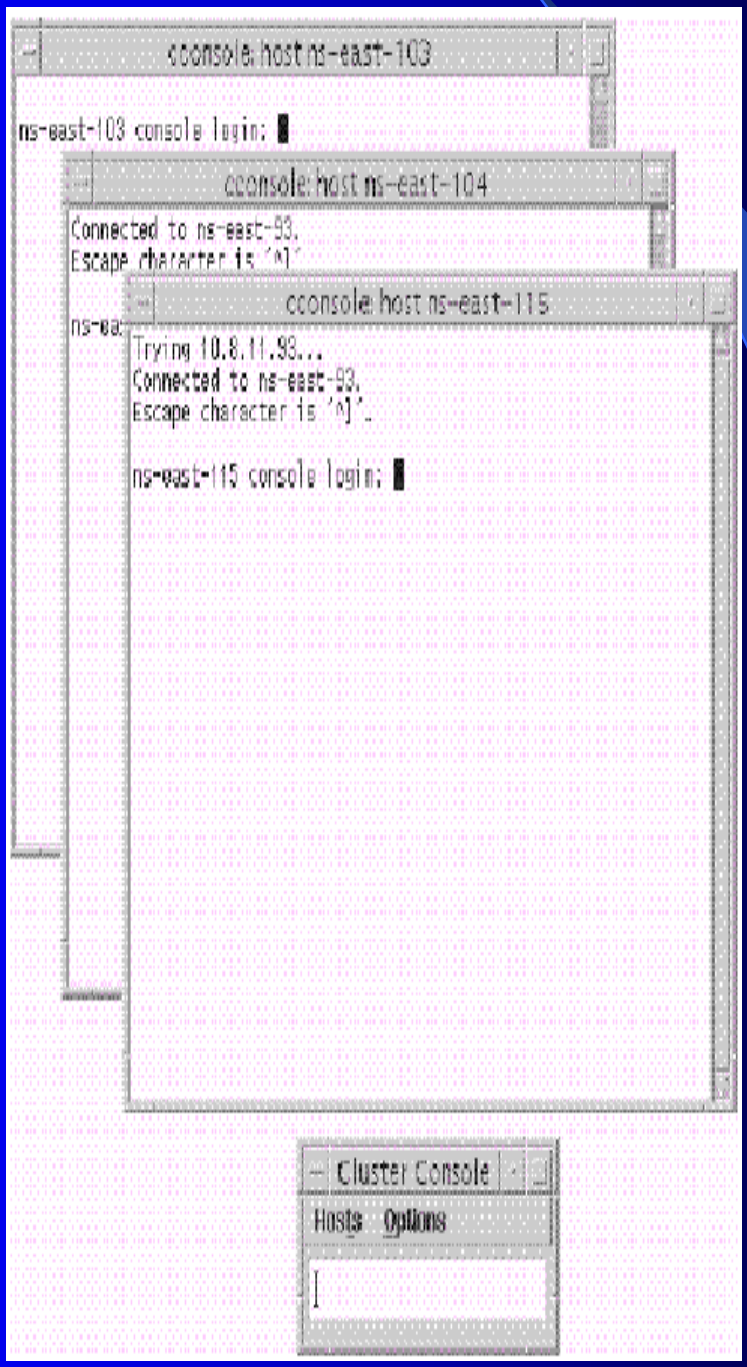
3.2 SSP & SP

- 基于域的服务器中的域（节点）没有办法通过串口访问其控制台。但可以访问主系统支持处理器（SSP，system support processor，Sun Enterprise 10000支持），或者通过系统控制台（SC，system controller，Sun Fire系列服务器支持），然后通过虚拟控制台协议（virtual console protocol）来访问域中的节点的控制台
- 手工登录到SSP，用户名为ssp，然后使用netcon命令连接到适当的域中
- Sun Fire E15K/25K，以域管理员身份手工登录SC（不同的域，管理员帐号可能不同），然后调用console命令来登录到恰当的域
- Sun Fire 3800-6800，SC以类似TC的仿真（在TCP端口5001,5002等监听）运行，通过这些端口来访问域控制台或域的shell（可能还需要额外的密码挑战）

3.3 Sun cluster的管理控制台

- Sun Cluster带有集中管理软件，用于管理的控制台软件包为：SUNWccon，只能安装在Solaris系统上
- 集中控制台能够为访问集群中的节点提供一个统一的访问入口
- 提供3种访问工具
  - Cconsole
    - 通过TC或其他的远程控制台访问方法来访问节点控制台
  - Crlogin
    - 执行rlogin命令来访问节点
  - Ctelnet
    - 执行telnet命令来访问节点

3.4 cconsole界面



3.5 集群控制台的命令界面

- 可以手工输入命令连接到集群节点或整个集群
- `# /opt/SUNWcluster/bin/cconsole node1 &`
- `# /opt/SUNWcluster/bin/ctelnet my-cluster &`
- `# /opt/SUNWcluster/bin/crlogin node3 &`

- 集群控制台的集群控制面板提供对三种工具的统一访问通道，从命令行启动集群控制面板的命令：
- # /opt/SUNWcluster/bin/ccp [clustername] &

3.6 相关的配置文件

- 集群管理控制台提供的集群管理工具相关的配置文件为：/etc/serialports（仅用于cconsole）和/etc/clusters文件。
- /etc/clusters文件中的配置条目为：
  - cluster-name node-1-name node-2-name
- 在/etc/clusters文件中定义了节点后，需要在/etc/serialports文件中定义到每个节点的控制台的远程路径。典型的配置示例：
- 节点的名字 TC的名字 TCP端口
- node-1-name sc-tc 5002
- node-2-name sc-tc 5003
- 其他的serialports配置见备注

Sun Cluster的拓扑和仲裁

- 要求条件
  - 启动设备
  - 服务器硬件
- 拓扑
  - Clusterd pairs
  - Pair+N
  - N+1
  - N*N
- 仲裁机制（quorum）
  - Failure fencing
  - Amnesia prevention

4.1 Sun cluster的特定要求

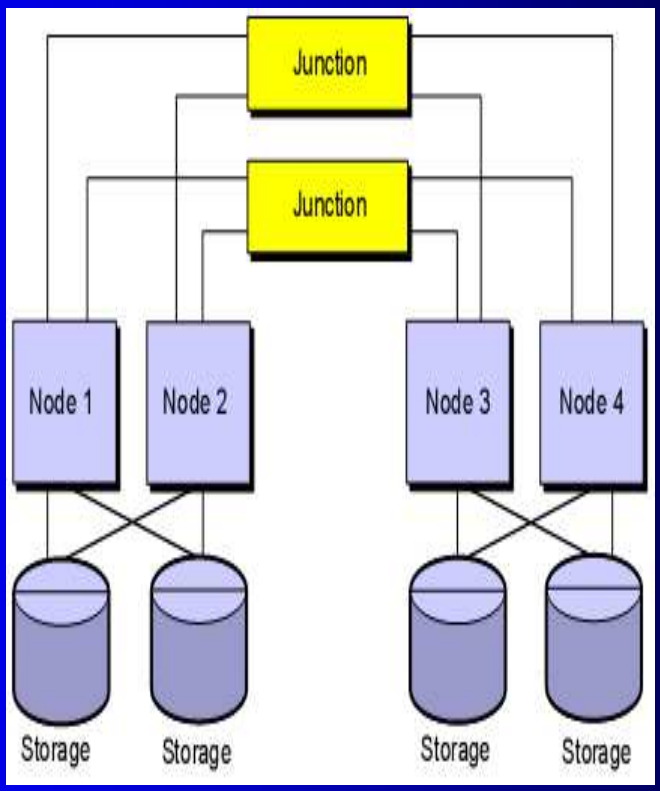
- 启动盘必须有至少100MB的/globaldevices文件系统
- 启动盘必须至少有10M的分区用于存放metadb
- Solaris至少必须选择End User软件簇
- 至少512M内存
- 至少750M交换空间

4.2 Sun Cluster 集群拓扑

- SPARC集群节点和数据存储设备的连接可采用多种拓扑结构。下面是典型的集群拓扑：
  - 成对的集群拓扑（群集对拓扑）
  - 成对+N拓扑
  - N+1拓扑
  - 多端口（超过两个节点）的N*N可伸缩性拓扑
- X86集群仅支持双节点拓扑

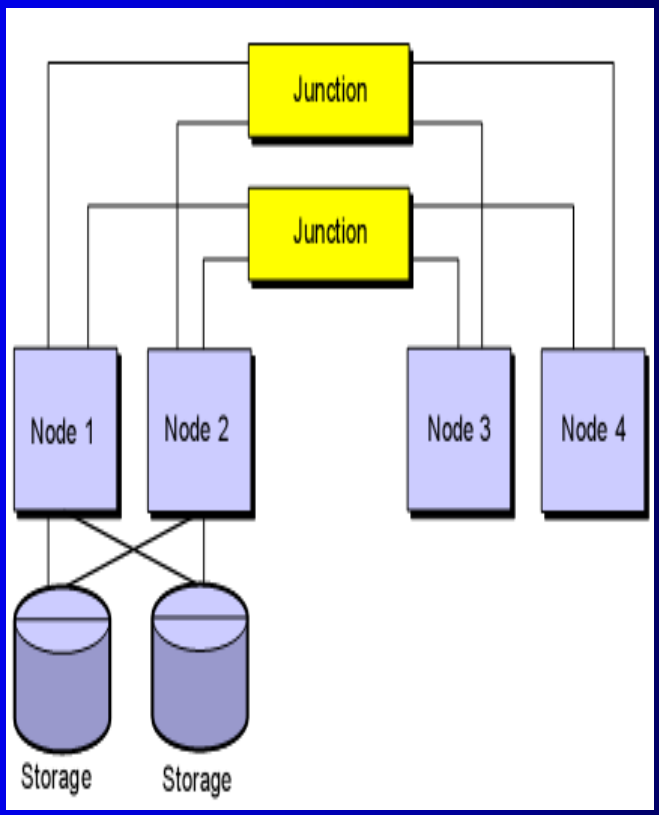
4.2.1 SPARC 成对集群

- 节点总数永远是偶数
- 每对节点内共享存储
- 所有节点必须集群互连
- 适用于failover
- 建议应用在同一对节点内运行



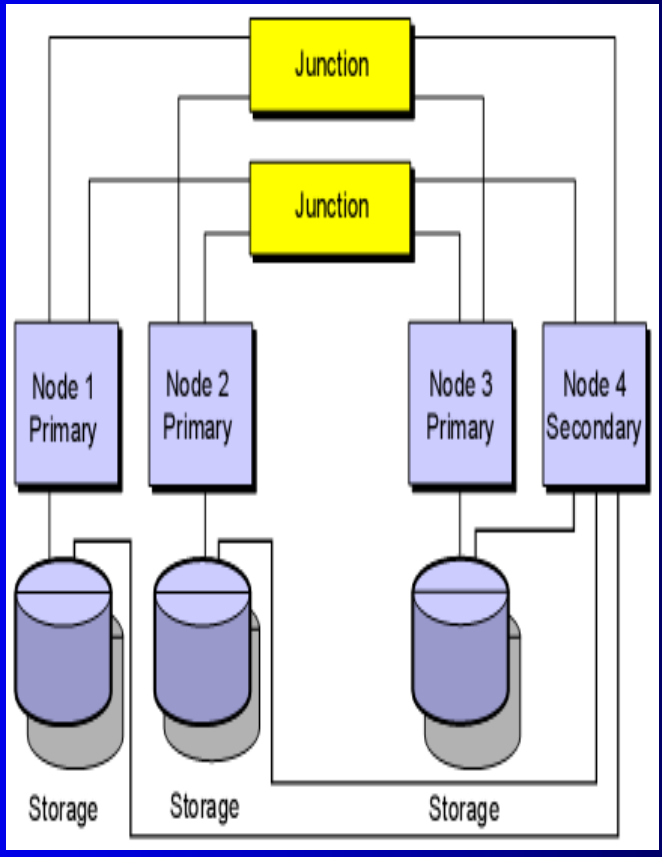
4.2.2 SPARC 成对+N

- 仅有一对节点与存储有物理连接
- 通过全局设备和全局文件系统，其他节点可访问存储
- 更适用于可伸缩性数据服务
- 集群互连必须有足够的带宽（可添加更多的互连网卡或使用千兆卡），以满足其他节点通过集群互连访问存储的数据流量的要求



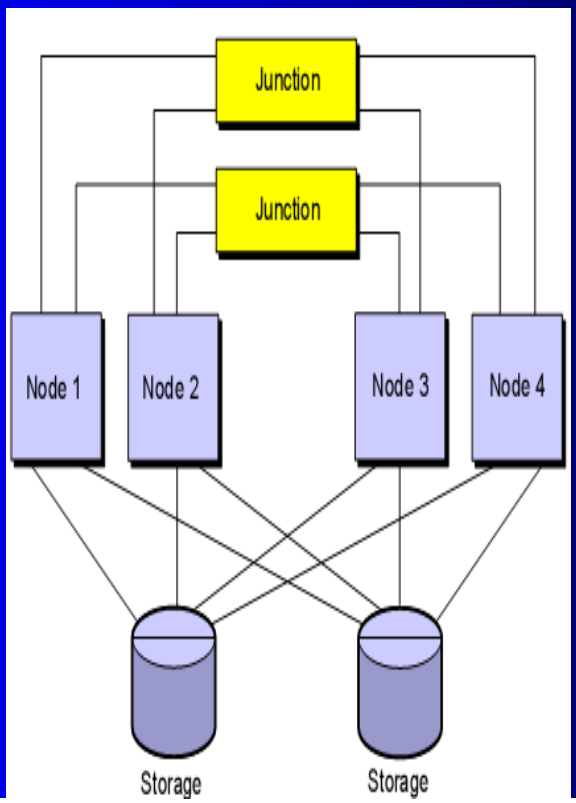
4.2.3 SPARC N+1

- 在N+1的拓扑结构中，一个节点作为集群中所有其他节点的备机（备用系统）。所有存储设备的第二条路径都连接到备用系统/次节点（redundant system/secondary node），在正常情况下，备用系统的负荷仅为运行系统所需的工作开销。



4.2.4 SPARC N*N

- N*N的可伸缩性拓扑，每个存储设备上连接的节点数超过两个。
- 可以在多个节点上运行Oracle RAC
- 通常而言，对于无集群意识的应用，共享存储中的每个特定的磁盘组或磁盘集，在一个时刻依然仅支持来自一个节点的物理流量。但，多个节点（超过两个）和存储设备有物理连接，可以增加集群配置的灵活性和可靠性。

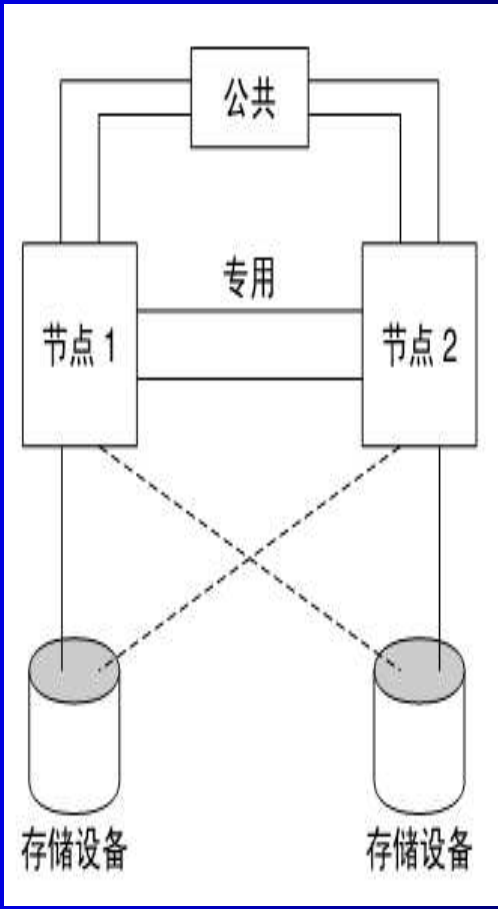


4.2.5 SPARC 无存储拓扑

- 多于两个节点的集群甚至可以不需要有共享存储设备，这种配置适用于那些纯粹基于计算且不需要任何数据存储的应用。
- 仅有双节点的集群必须要有共享存储，因为双节点继续需要一个仲裁设备才能完成投票选举操作。

4.2.6 X86 双节点拓扑

- 由基于X86的系统组成的Sun Cluster集群仅支持两个节点，共享存储设备必须物理连接到这两个节点上。



4.3 仲裁/法定概述

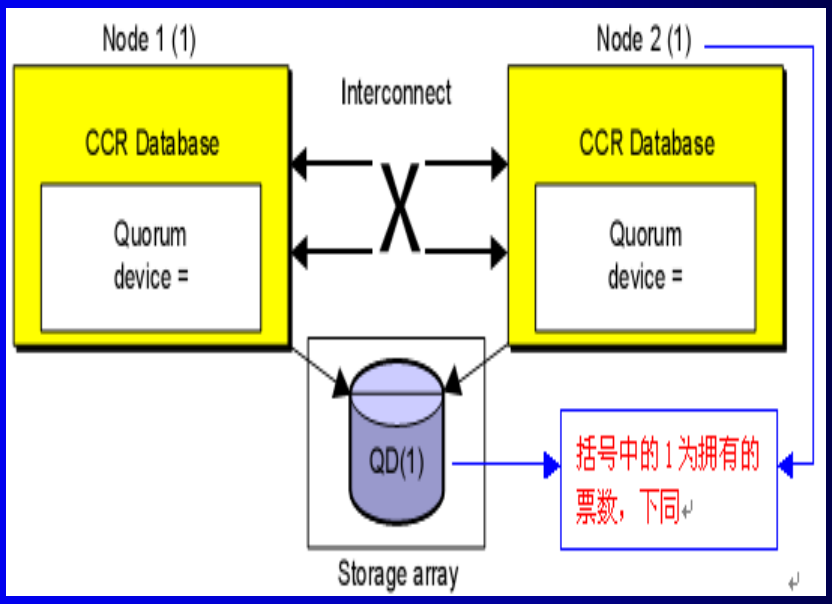
- 为了维持集群的稳定性，Sun Cluster软件框架采取了一种称为投票系统（voting system）的机制。
  - 每个节点都被明确分配了一张选票
  - 指定特定的磁盘（可多个）作为仲裁设备（quorum devices），并给予选票
  - 采用多数票原则，任何节点的票数必须超过所有选票数的50%才能够形成一个集群或继续呆在集群中

4.3.1 脑分裂与失忆

- 正是基于上述规则，因此仅具有两个节点的集群，必须拥有额外的仲裁磁盘选票；否则的话，如果仅拥有来自节点的选票，那么在双节点中任一节点发生故障时，就无法在集群内达到多数同意的目的。
- 仲裁选举和仲裁设备解决了以下两个主要问题：
  - Failure fencing（故障防护，防止脑分裂/集群分割）
  - Amnesia prevention（防止失忆）
- 这是两个截然不同的问题，Sun Cluster 3.x软件用同样的一个仲裁机制就成功的解决了这两个问题；而其他的厂家的cluster软件通常要采用两个不同的机制来分别解决这两个问题，这会使集群的管理显得更加复杂。

4.3.2 failure fencing

- 如果节点间的互连通信中断，不论是互连链路故障，还是因为一个节点崩溃导致，每个节点都必须假定另外一个依然正常工作。这就是所谓的“脑分裂（split-brain）”操作。绝对不允许出现两个独立的集群共存的情况，这有可能会造成数据收到破坏。两个节点都会努力争取获得另外一张仲裁选票以尝试建立一个集群。
- 只会且只能有一个节点会在竞争中胜出，从而保留集群资格，失败的节点将会退出集群。



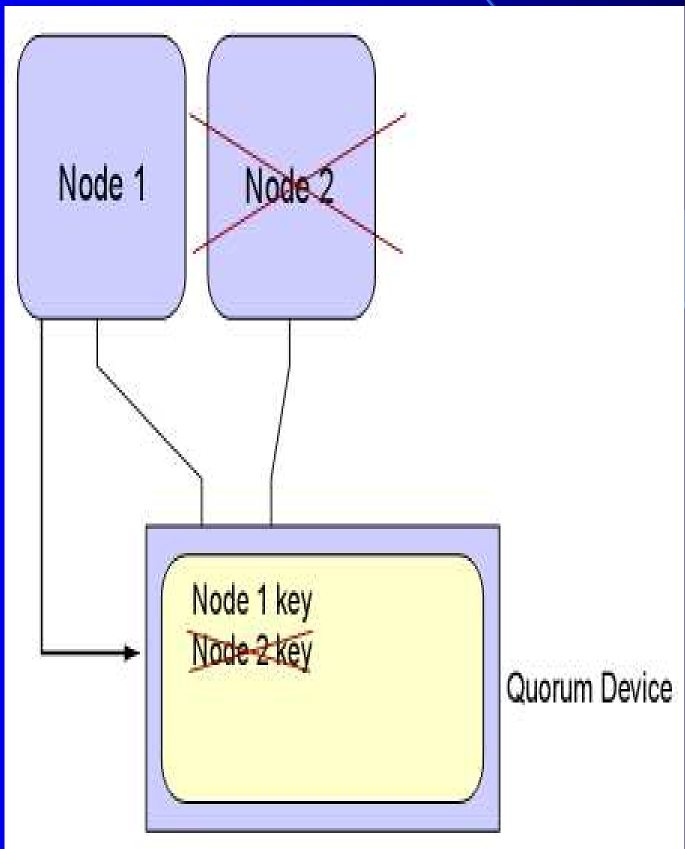
4.3.3 amnesia（1）

- “集群失忆”场景是指：一个或多个节点（在集群中首先启动的节点）能够用一份稳定版的集群配置的拷贝来形成集群。试想一下以下场景：
  - 在一个双节点集群中（节点1，节点2），节点2由于崩溃或维护的目的处在停机状态
  - 在节点1上，对集群配置进行了修改
  - 节点1关闭
  - 启动节点2，形成一个新的集群
- 这这种情况下，节点2并不知道节点1已经对集群配置进行了修改，……

4.3.3 amnesia（2）

- Sun Cluster软件的仲裁使用持久保留来防止节点启动形成集群。节点2将无法使用仲裁设备来完成选票计数。因此节点2将会一直等待直到其他节点（节点1）启动才能达到仲裁选举所需的票数。
- 持续保留/预约（persistent reservation）这种方法需要把预约信息写在仲裁设备上：
  - ● 即使和设备连接的所有节点都重置（reset），信息依然保留
  - ● 即使仲裁设备自身上电或下电（powed on and off），信息依然保留
- 很显然，这涉及到往仲裁盘上写入一些特定类型的信息。这些信息叫做保留密钥（reservation key）：
  - 每个节点都被指定一个唯一的64位的保留密钥值
  - 和仲裁设备有物理连接的每个节点都会把自己的保留密钥写到仲裁设备上

4.3.3 amnesia（3）



在节点1加入集群后，它能够通过集群传输监测到节点2，并把节点2的保留密钥重新添加回到仲裁设备，于是所有要素重新得到均衡。[illegible] 因此，保留密钥（节点2）只能由集群中的另外一个节点（节点1）添加回到仲裁设备中。此节点（节点1）的保留密钥已经存在于该仲裁设备上。

- 不论是什么原因导致节点2离开集群，节点1都会删除 [illegible] 在仲裁设备上的保留密钥。[illegible] 仲裁设备的（persistent reservations and reservation keys）[illegible] 的最后一步，当节点2首先被启动，会试启动加入集群时，也无法从仲裁选票中获得所需的票数，节点2必须等待节点1启动。

4.4 仲裁设备的规则

- 仲裁设备的常见规则：
  - 在双节点集群中，两个节点都必须能够使用仲裁设备
  - 仲裁设备的信息保留在CCR数据库中，属于全局维护信息
  - 仲裁设备可以存放用户数据
  - 仲裁设备所能够贡献的最大和最佳选票数应该是节点选票数减一（N-1）
    - 如果仲裁设备数等于或超过节点数，万一发生太多个仲裁设备失效的情况，集群将无法启动，即使所有的节点都是正常的。很明显，这种情况令人无法接受。
  - 多于两个节点的集群可以不需要仲裁设备，但为获得更高的集群可用性，依然建议采用仲裁设备
  - 在Sun Cluster软件安装完毕后，可以手工配置仲裁设备
  - 仲裁设备使用的是DID设备，并且只有和它直连的节点才能使用。

4.5 仲裁算法

- 当一个集群在运行时，它必须能够清楚的知道以下事情：
  - 所有可能的仲裁选票数（节点数+在集群中定义的来自磁盘的仲裁选票数）
  - 所有当前的仲裁选票数（集群中当前启动的节点数+能够被这些节点物理访问的磁盘仲裁选票数）
  - 所有所需的仲裁选票数（必须达到所有可能的仲裁选票数的一半以上，即>50%）
- 对选票异常事件，集群软件采用以下处理方式：
  - 如果节点在启动时无法找到所需的选票数，将停滞等待其他节点加入，以获得期待的选票
  - 已在集群中启动的节点，但无法继续找到所需的选票数，将发生kernel panics。

4.

图 3.9 是一个典型的成对+2 的仲裁磁盘配置示意，总共用了 3 块仲裁盘

在成对集群配置中，节点总数总是以偶数出现，每对中的节点通常都对外提供 failover 数据服务，或者做为同一对中另外一个节点的 failover 数据服务的备机。

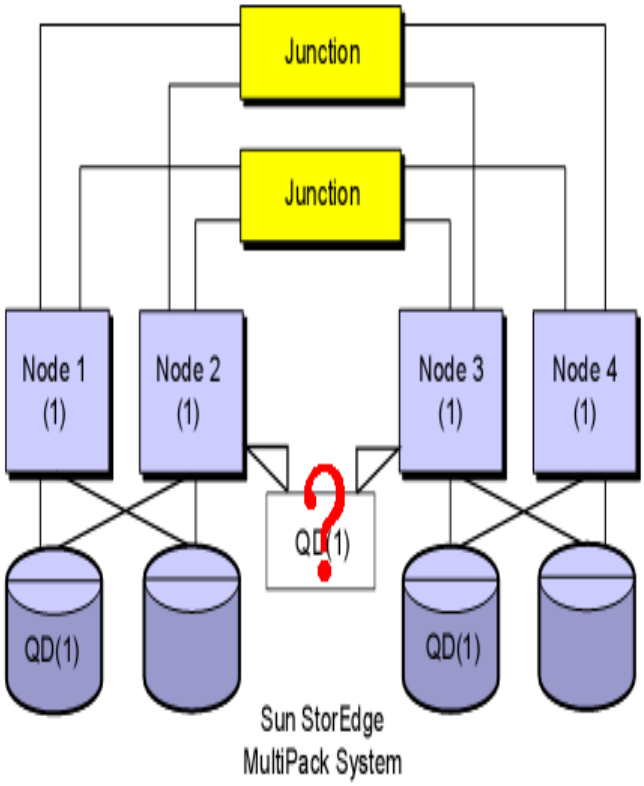


图 3.8　成对集群的仲裁设备

图 3.8 中，有许多种可能会导致脑分裂的发生，在发生脑分裂时，有可能导致集群的运行终止。如果在成对集群配置中没有“另外的（extra）”仲裁设备，将会发生以下情况：

- 所有可能的选票数为 6 张
- 要完成仲裁（获得多数票），需要 4 张选票
- 如果两个仲裁设备都失效，集群依然可以启动。节点会一直等待直到所有的节点都启动
- 如果节点 1 和节点 2 都失效，节点 3 和节点 4 就无法获得足够的票数，就没法继续运行
  - 可在节点 2 和节点 3 之间放置一个象征意义上的仲裁设备来解决这个问题。这可使用 Sun StorEdge MultiPack desktop array 来实现
  - 当一对节点完全失效时，7 张选票中依然有四张选票可用
  - 集群操作依然可以继续运作

安装前的准备工作

- 操作系统的环境收集
- 明确集群的拓扑结构
- 选择所需的仲裁设备
- 检查集群的互连配置
- 确定公共网络的接口

5.1 系统环境收集

- 01. 记录各节点的内存
  - 运行/usr/sbin/prtconf（快捷方式为/etc/prtconf）
  - # /usr/sbin/prtconf | grep Memory
  - Memory size: 16384 Megabytes
- 02. 查看是否有/globaldevices文件系统，至少100MB的容量，运行df –h命令查看
- 03. 查看是否有足够的SWAP硬盘空间，至少750MB，运行swap -l查看
- 04. （可选，通常是必须的）查看本地硬盘是否预留有专用的metadb分区，format查看

5.2 明确拓扑结构

- 01. 记录准备采用的集群拓扑配置
  - 表：集群的拓扑配置

节点的总数	
存储阵列的数量	
存储阵列的类型	

- 02. 检查集群中的存储阵列是否已经按照设计的拓扑进行连接，如果不是，请重新连接。

5.3 确定仲裁设备

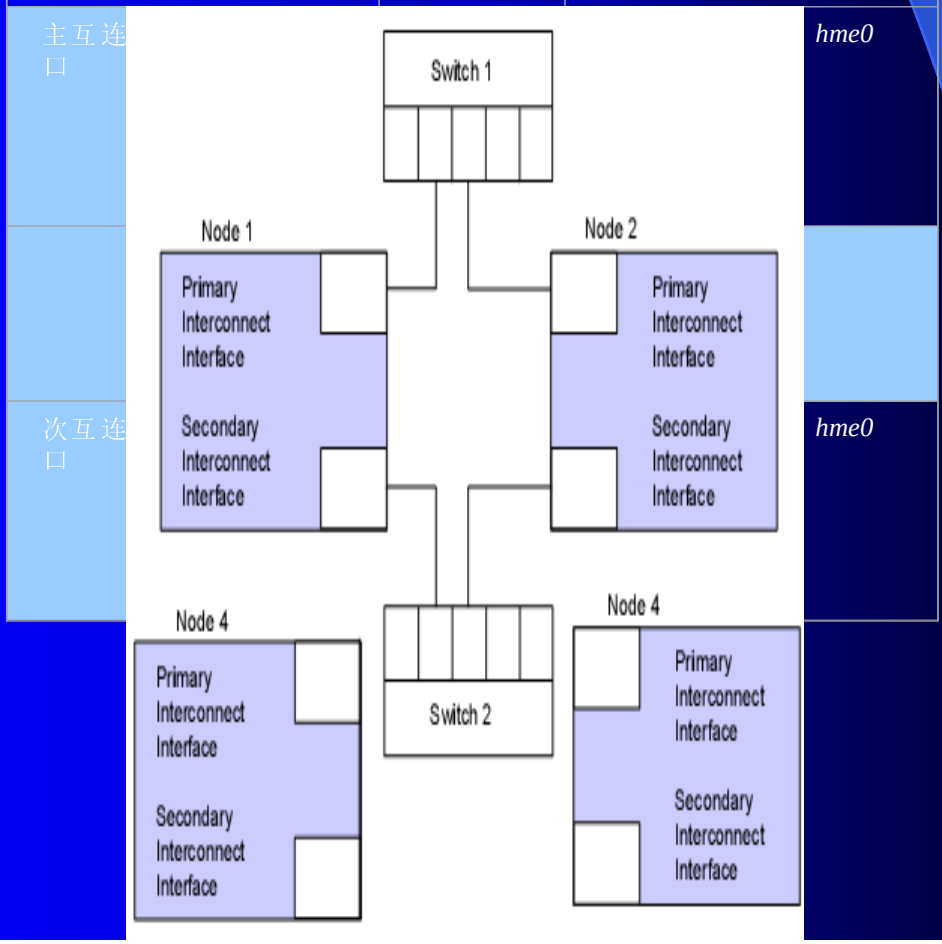
- 01.记录所需的仲裁设备的数量，在集群软件安装完毕后，需要配置仲裁设备
  - 仲裁盘的数量：_____
- 02.记录仲裁盘的逻辑路径（仲裁盘肯定来自集群中的某个存储阵列），用format命令查看。
  - 仲裁盘的路径：_____
- 03. 用Control-D退出format

5.4 确定集群互连

- 基于交换机的集群互连
- 点对点集群互连
  - 把用于集群互连的接口的逻辑名（hme0，qfe1，...）填入下表，可根据需要调整图中的节点数。

	节点1		节点2
主互连口			
次互连口			

  - 检查每个集群互连接口是否都连到正确的交换机上



5.5 确定网络接口

确定每个节点连接到公共网络的网卡的逻辑名，如有可能，尽量选择两个网卡，用来组成 IPMP 组

14. 把选定的公共网络接口的逻辑名（hme2，qfe3，fjgi0，...）填入表 3.4 中。

表 3.4　潜在的 IPMP 以太网接口的逻辑名

主机系统	主 IPMP 接口	备 IPMP 接口
节点 1		
节点 2		
...		

15. 检查每个节点的 IPMP 接口是否都已正确的连接到公共网络

后续…

- 安装和创建Sun 集群
- Sun Cluster的配置和使用
- ……

- 谢谢！