Sun Cluster 3.0 的规划、安装、配置及管理

其它功能 王延峰 2006-09-30 发表

Sun Cluster 3.0 的规划、安装、配置及管理

一、 组网需求：
SUN服务器或者工作站两台
SUNOS
二、 组网图：
无
三、 配置步骤：

适用于SUNOS以及SUN CLUSTER3.0版本
1．Sun Cluster的基本概念：
1.1 Sun Cluster3.0支持的两种服务模式：
Failover(失败切换)：当故障发生时，系统自动将应用程序等资源（APP、IP、DATA）一个故障主节点上重新定位到指定的辅助节点，客户可能会看到一个短暂的服务中断（一般为10s）,并可能需要在失败切换结束后重新连接，但客户并不知道哪一个物理服务器向他们提供应用程序和数据。做到了应用程序的冗余。
Scalable(可伸缩)：利用集群中的多个节点来同时运行一个应用程序，每个节点都可以提供数据和处理客户请求，这样既提供了高可用性，还提供了更高的性能。
Sun Cluster3.0单一集群既可以支持失败切换应用程序，也可以支持可伸缩应用程序,而2.2只支持失败切换应用程序。
1.2双节点群集配置样例：
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(cluster member)，要么是潜在成员。Sun Cluster3.0 软件使您可在一个群集中部署两到八个节点,而2.2只支持到4个节点。
群集名：群集中的所有节点都会归组到一个集群名称下—用于访问和管理群集。如sc mail
公共网络适配器（nafo）：客户机通过公共网络接口与群集相连。每个网络适配器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卡。如果活动适配器出现故障，则调用NetworkAdapter Failover (NAFO) 软件进行失败切换，将接口切换至一个备份适配器。
私网（private networks）：群集成员通过物理上独立的一个或多个网络
（private networks）与群集中的其他节点通信，知道另一节点的加入或离开
管理控制台:可以使用专用SPARCstationTM 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SunManagement Center 产品的Sun Cluster 模块。它的使用可以使控制台和管理工具归组到同一机器上，实现集中化的群集管理。
控制台访问设备：Sun 只提供一种支持的终端集线器。您可选择使用支持的Sun 终端集线器。终端集线器通过使用TCP/IP 网络实现对每一节点上ttya 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。
1.3设备标识(DID)
Sun Cluster 通过一种叫做设备标识(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中的每个设备分配唯一的标识，包括多主机磁盘、磁带驱动器和CD-ROM。设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标识，而非传统的Solaris 设备ID（如某一磁盘的标识c0t0d0）。这一措施可保证任何使用磁盘设备的应用程序（如卷管理器或使用原始设备的应用程序）都可使用一致的路径访问设备。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 驱动程序则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。
1.4 Quorum 设备
Sun Cluster 配置使用quorum 设备维护数据和资源的完整性。如果群集暂时丢失与节点的连接，则quorum 设备阻止在群集节点试图重新连接群集时出现健忘或使人头疼的问题。通过使用scsetup(1M) 实用程序来指定quorum 设备。
规划quorum 设备时请考虑以下几点。
_ 最小值– 两个节点的群集必须至少有一个分配为quorum 设备的共享磁盘。对于其他拓扑，quorum 设备是可选的。

_ 奇数规则– 如果在直接与quorum 设备连接的两个节点的群集或一对节点中配置多个quorum 设备，则配置奇数个quorum 设备，以便这些设备有完全独立的失败通道。
_ 连接– quorum 设备不能与两个以上的节点连接。

2.准备工作：

2.1安装TC（可选）

1) 两台节点机的串口A分别接到TC的2、3号端口，管理机的串口A连接到TC的1号端口

2) 在管理机上修改文件如下：

#vi /etc/remote

hardware: dv=/dev/term/a:br#9600:el=^C^S^Q^U^D:ie=%$:oe=^D

3) 执行#tip hardwire,按下TC面板上的test键，直到Power灯闪放开

4) 在按一下TC面板上的test键（2s）

5) 管理机的屏幕上显示monitor::

用addr修改TC的 IP地址，按“~.退出”，重起TC

6) telnet到 TC,执行

annex: su

passwd: <tc ip address>

annex# admin

7) 配置串口模式：

admin: set port=1-8 type dial_in imask_7bits Y

admin: set port=2-8 mode slave ps_history_buffer 32767

admin: quit

annex#boot

2.2配置管理机：（可选）

1) 用root用户登陆管理机，修改/etc/hosts,将所有节点机的主机名和对应地址写入

2) 添加cluster console software

#pkgadd –d . SUNWccon

3) 修改/.profile文件

PATH=$PATH:/opt/SUNWcluster/bin

MANPATH=$MANPATH:/opt/SUNWcluster/man

Export PATH MANPATH

4) 使profile生效 # . /.profile

5) 编辑/etc/clusters

cluster-name node1-name node2-name

6) 编辑/etc/serialports

node1-name TC-address 5002(在TC上的端口号)

node2-name TC-address 5003

7) 执行#ccp cluster-name &，使用clogin或cconsole/ctelnet工具

2.3 修改SCSI Initiator Id

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。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 规格需要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，在第二个主机适配器的SCSI 地址改为6。

1) ok show-disks 记下控制器的路径

2) 创建一个nvramrc脚本设置scsi-initiator-id

ok nvedit

0: probe-all install-console banner

1: cd /pci@6,4000/scsi@3

2: 6 " scsi-initiator-id" integer-property

3: device-end

4: cd /pci@6,4000/scsi@2,1

5: 6 " scsi-initiator-id" integer-property

6: device-end

7: banner (Control C)

ok nvstore

ok setenv use-nvramrc? True

ok setenv auto-boot? true

ok reset-all

ok boot
2.4 在两个节点机上安装操作系统solaris（至少选用最终用户模式安装），打上推荐补丁。
/: 最小100M
swap:最小750M,是内存的2倍
/globaldevices:100M
起码保留一个100M的未用分区供卷管理软件存储卷信息使用。
2.5确认local-mac-address值为false
#eeprom |grep mac
3. Sun Cluster软件安装：
3.1 在每个节点机上编辑/.profile文件
PATH=$PATH:/usr/cluster/bin
MANPATH=$MANPATH:/usr/cluster/man:/usr/share/man
Export PATH MANPATH
3.2 在每个节点机上编辑.rhosts
+
3.3 在每个节点机上编辑/etc/default/login文件
#COnSOLE=/dev/console
3.4 在每个节点机上编辑/etc/hosts文件，将对方节点，逻辑主机名的对应ip写入
3.5 建立一个全新的cluster节点
1) 运行SunCluster_3.0/Tools/scinstall
2) Establish a new cluster
c. 输入集群名字
d. 输入集群中另一台节点的机器名
e. 不使用DES认证
f. 使用默认的集群传输私网地址
g. 接受默认的全局设备文件系统
h. 接受装完后自动重起
3.6 向集群中添加另一个节点
a． 运行SunCluster_3.0/Tools/scinstall
b． Add this machine as a node in an established cluster
c． 输入主节点的机器名
d． 接受默认的全局设备文件系统
e． 接受装完后自动重起
3.7 打上cluster的补丁
3.8 配置Quorum 设备
1) 运行scdidadm –L选择准备作为Quorum disk的磁盘号，该磁盘必须在两个节点都能访问的共享磁盘
2) 运行scsetup,输入前面选定的DID设备号
3) 在两个节点的集群中不需再添加Quorum 设备
4) 接受安装
3.9 配置网络时钟同步协议
修改每个节点机的/etc/inet/ntp.conf,将不存在的节点删除，即将以下行删除
peer clusternode3-priv
peer clusternode4-priv
peer clusternode5-priv
peer clusternode6-priv
peer clusternode7-priv
peer clusternode8-priv
此时，运行scstat –q,可以看到一共有3票；运行scdidadm –L，可以看到所有的DID设备；运行scconf –p，可以看到集群状态、节点名、网卡配置、quorum设备状态。运行scshutdown –y –g 15,以后关的机器先启为原则测试cluster时候能正常启动。
可以运行scheck检查cluster安装是否有错
4.卷管理：
4.1 使用veritas作为卷管理软件
1) 停止veritas volume manager dynamic multipathing功能，防止它和cluster功能冲突
#mkdir /dev/vx
#ln –s /dev/dsk /dev/vx/dmp
#ln –s /dev/rdsk /dev/vx/rdmp
2) 安装veritas volume manager 软件，并打上veritas的补丁
pkgadd –d . VRTSvmdev VRTSvmman VRTSvxvm
3) 修改两台节点机的vxio号为一致，并不与其他设备冲突。修改时必须把cluster软件停止
#grep vxio /etc/name_to_major

4) 封装rootdg
#vxconfigd –m disable
#vxdctl init
#vxdg init rootdg
#vxdctl add disk c0t0d0sX(未用分区) type=simple
#vxdisk –f init c0t0d0sX type=simple
#vxdg adddisk c0t0d0sX
#vxdctl enable
#rm /etc/vx/reconfig.d/state.d/install-db(该文件不删除，系统不会启动vm软件)
你也可以使用/usr/sbin/vxinstall对整个系统盘进行封装，形成rootdg,但你必须事先保留两个未用分区，一般为slice 3和slice 4。
5) 重起节点1，看vm是否正常启动
VxVM starting in boot mode…
VxVM general startup…
可以使用vxprint察看已配disk group状况
6) 在另一台节点机上安以上步骤安装veritas软件并配置rootdg,重起。
7) 在新建的dg上建立卷
#vxassist –g xxx make volname 200m layout=mirror
8) 注册disk grups
#scconf –a –D type=vxvm ,name=xxxx,nodelist=node1:node2
如果再已注册的dg上添加vol,需运行scsetup同步注册信息。
9) 使用已建vol
#newfs /dev/vx/rdsk/dg-name/volname
#mkdir /global/xxx(两节点同时做)
#vi /etv/vfstab（两节点同时做）
/dev/vx/dsk/dgname/volname /dev/vx/rdsk/dgname/bolname /global/xxx ufs 2 yes global,logging
#mount /global/xxx
5.资源配置：
5.1 配置nafo
#pnmset
输入nafo组号及改组包含的网卡名称
5.2 配置资源组
1） Sun Cluster3.0支持两种资源类型：
1.1) 数据资源类型（Data service resource）：
oracle、iplanet、netscape、apache、dns、nfs
1.2) 预注册资源类型(Preregistered Resource)：
SUNW.HAStorage、SUNW.LogicalHostname(供failover数据资源使用)、SUNW.SharedAddress（供scalable数据资源使用）
2) 配置failover 数据资源，以Sun Cluster HA for NFS为例：
2.1) 添加NFS数据资源包（两个节点机），可用pkgadd命令，也可用scinstall交互界面
2.2) 建立NFS目录
#mkdir –p /global/nfs/admin/SUNW.nfs
#mkdir –p /global/nfs/data
#chmod 777 /global/nfs/data
2.3) 编辑NFS参数文件
vi /global/nfs/admin/SUNW.nfs
share –F nfs –o –rw –d”Home Dirs” /global/nfs/data
2.4) 注册数据资源(资源必须注册后才能使用)
#scrgadm –a –t SUNW.nfs
#scrgadm –a –t SUNW.HAStorage
2.5) 建立failover资源组
#scrgadm –a –g nfs-rg –h node1,node2 –y Pathprefix=/global/nfs/admin
2.6) 往资源组中添加资源
#scrgadm –a –L –g nfs-rg –l clustername-nfs (注：clustername-nfs在两台节点机的/etc/hosts中有相应记录)
#scrgadm –a –j has-res –g nfs-rg –t SUNW.Hastorage –x ServicePaths=/global/nfs –x AffinityOn=True (AffinityOn=True: 应用切换，磁盘存贮也跟随切换)
#scrgadm –a –j nfs-res –g nfs-rg –t SUNW.nfs –y Resource_dependencies=has-res
2.7) 初始化资源组，是之生效
#scswitch –Z –g nfs-rg
2.8) 检测cluster状态

#scstat –g
#scswitch –z –h dest-node –g nfs-rg
3) 配置scalable数据资源组，以Sun Cluster Scalable Service for Apache为例
3.1) 添加Apache数据资源包（两个节点机），可用pkgadd命令，也可用scinstall交互界面
3.2)关闭apache自动启动和关闭功能
#mv /etc/rc0.d/K16apache /etc/rc0.d/k16apache
#mv /etc/rc1.d/K16apache /etc/rc1.d/k16apache
#mv /etc/rc2.d/ K16apache /etc/rc2.d/k16apache
#mv /etc/rc3.d/S16apache /etc/rc3.d/s16apache
#mv /etc/rcS.d/ K16apache /etc/rcs.d/k16apache
3.3) 在两个节点机的/etc/hosts种都加入clustername-web的相应内容
clustername-web IP_address
3.4) 编辑控制文件，建立相应的服务目录
#cp /etc/apache/httpd.conf-example /etc/apache/httpd.conf
#vi /etc/apache/httpd.conf
Server Name clustername-web (去掉原有的注释)
DocumentRoot “/global/web/htdocs”
<Directory “/global/web/htdocs”>
scriptAlias /cgi-bin/ “/global/web/cgi-bin”
<Direcotory “/global/web/cgi-bin”>
3.5) 建立html和cgi目录文件
#mkdir /global/web/htdocs
#mkdir /global/web/cgi-bin
#cp –rp /var/apache/htdocs /global/web
#cp –rp /var/apache/cgi-bin /global/web
3.6) 注册数据资源(资源必须注册后才能使用)
#scrgadm –a –t SUNW.apache
3.7) 建立资源组
#scrgadm –a –g sa-rg –h node1,node2
h. 往资源组里添加sclable资源
#scrgadm –a –S –g sa-rg –l clustername-web
#scrgadm –a –g web-rg –y Maximum_primaries=2 –y Desired_primaries=2 –y RG_dependencies=sa-rg
#scrgadm –a –j apache-res –g web-rg –t SUNW.apache –x \
Confdir_list=/etc/apache –x Bin_dir=/usr/apache/bin \
–y Scalable=TRUE –y Network_resources_used=clustername-web
3.8) 初始化资源组，是之生效
#scswitch –Z –g sa-rg
#scswitch –Z –g web-rg
3.9) 检测cluster状态
#scstat –g
3.10) 调整节点负载，默认为1：1
#scrgadm –c –j web-res –y Load_balance_weights=5@node1,2@node2
6.Sun Cluster的日常维护和管理：
6.1) 显示sun cluster 发行版本
#scinstall –pv
6.2) 显示已配置的资源类型、资源组和资源
#scrgadm –p
6.3) 检查集群组件状态及配置
#scstat –p
#scconf –p
#scrgadm –pv(v)
6.4) 关闭集群
#scshutdown –g 0 –y
此命令将每个节点都关闭到OK状态，可用boot命令启动，然后用scstat –n状态件则节点状态。
6.5) 关闭单个节点
#scswitch –s –h node2
#shutdown –g 0 –y
6.6) 将某一节点置为维护状态
#scconf –c –q globaldev=quorumdivice(dx),maintstate
6.7) 手工修改节点quoroum的投票数
a. ok> boot –x

b. #cd /etc/cluster/ccr
c. #vi infrastructure
cluster.nodes.1.name torrey
cluster.nodes.1.properties.quorum_vote 3
d. #cd /etc/cluster/ccr
e. #/usr/cluster/lib/sc/ccradm –l /etc/cluster/ccr/infrastructure –o
f. #reboot
6.8) 删资源组和disk group
a. 停资源：scswitch –n –j resourcename
b. 删资源：scrgadm –r –j resourcename
c. 删资源组：scrgadm –r –g resourcegroup
d. 删dg：vxdg destroy dgname
6.9) 删除 cluster软件
g. ok>boot –x
h. #pkgrm clusterpkgname
i. rm –r /var/cluster /usr/cluster /etc/cluster
j. vi /etc/vfstab,将原来所有的dis参数恢复，重建/global/devices
k. rm /etc/ntp.conf
l. vi /etc/nsswitch.conf,除去cluster选项
m. rm –r /dev/did
n. rm –f /devices/pseudo/did*
o. rm /etc/path_to_inst
p. reboot -- -ra
7.Sun cluster 3.0与2.2的比较
Sun cluster 3.0
支持8个节点
支持ufs,hsfs,为实现scalable出现了global概念
cluster networking share address
支持scalable,failover模式
于系统核心绑定很紧，只有network和resource group由相关demond启动管理
使用boot –x可以只起系统，不起cluster
支持solaris 8
支持千兆网卡，不支持sci card
以资源组为切换单位，资源类型有app,ip,data

sun cluster2.2
支持4个节点
支持ufs,没有global
logical host address
只支持failover模式
于系统核心绑定不紧，可以先起系统再手工起cluster
支持sci card(100M byte/s)
以逻辑机为切换单位
四、 配置关键点：
无