AIXCHINA

HACMP for AIX
学习笔记

www.aixchina.com

Aix HACMP

IBM Certification Study Guide Test 167

Chapter 3 群集软硬件准备

3.1 节点设置

3.1.1 卡槽设置

参考下列三本资料
PCI Adapter Placement Reference Guide SA38—0538
Adapters,Devices,and Cable Information for Micro Channel Bus Systems SA38—0533
Adapters,Devices and Cable Information for Multiple Bus systems SA38—0516

3.1.2 Rootvg 镜像

通常的文件系统和磁盘的镜像基于 LVM，不能保护到操作系统的失败或引导盘的失败。

OS 的失败一般是内存不能正常 page IO，渐渐地失效，而不是瞬间突然失败。

虽然 HACMP 环境不强求一定做 rootvg mirror，但有做则可减少系统的失效时间，毕竟节点接管须一定时间。

Rootvg Mirror 步骤

1、镜像所有的 rootvg 中的 FS，同 LVM 中的操作

2、建附加的 blv，以下详解

3、修改 bootlist，包含所有的 boot device

dump device 不能镜像，因为 AIX4.3.3 以前，LVM 不支持对 Dump device 的镜像写。Dump 有以下三态：1、Not available 未发生 Dump；2、Available and not corrupted 发生 Dump 且可用；3、Available and corrupted 一般是由于 Dump 出了问题，可能是由于镜像了 Dump Device

建议为 Dump Device 建一个未镜像的 lv，不采用作为 paging space 的 hd6。

在 A1X4.2.1 中，有 mirrorvg 和 unmirrorvg 命令，mirroring 做了下述 2，3，4 的步骤，其中 mirrorvg 不会将 Dump Dev 做镜像，但如果 Dump Dev 和 paging Dev 是同一个 lv，则会做镜像。

3.1.2.1 过程

```
1、extendvg rootvg hdisk1                  #假设 rootvg 原在 hdisk0 上
2、chvg -Qn rootvg                         #Disable Quorum
3、mklvcopy   hd1    2    hdisk1           #/home
   mklvcopy   hd2    2    hdisk1           #/usr
   mklvcopy   hd3    2    hdisk1           #/tmp
   mklvcopy   hd4    2    hdisk1           #/
   mklvcopy   hd5    2    hdisk1           #blv
```

mklvcopy hd6 2 hdisk1 #paging space
mklvcopy hd8 2 hdisk1 #/fs log
mklvcopy hd9rar 2 hdisk1 #/var
如有多个 paging space，不论在不在 rootvg 中，建议都 mirror。
hd5 如超过一个 lp，必须让镜像的 hd5 的 PP 连续。
lslv – m hd5 如不连续，删去，重镜像用 mklvcopy –m 参数。

4、syncvg –v rooting

5、bosboot –a -d /dev/ hdisk?
这里？用 lslv –l hd5 确定 PV 头的第一块 disk

6、bootlist -m normal hdisk0 hdisk1
虽然做了定义，但不是所有情况下，只要 hdisk0 失败，就会从 hdisk1 引导，有时要通过介质引导，进入 maintenance，修改 bootlist 去掉 hdisk0 才行，有些型号，固件中有选 bootlist 功能。

7、Shutdown -Fr 让 Quorum OFF 工作

3.1.2.2 须要特定的 APAR
eg V4R3 need 1X72550，用 instfix –I -K <aparnum>

3.1.3 HACMP 对 AIX 和其它 LPP 的要求

eg:HACMP4.3 for AIX 要求 AIX4.3.2 SP 还要 PSSP2.2

3.1.4 AIX 参数设置和 HACMP 有关

3.1.4.1 I/O Pacing
高低水平线(默认是 0，即不做控制)
当大的 I/O 作业发生时，交互作业响应受影响，甚至影响到心跳信号的定时传播，引起误接管，设定高低水平线对性能有轻微影响，当对一个达到高水平线文件写时，会被控制进入等待，直到 I/O 完成许多后，回到低水平线时才继续写，以平衡其它应用的 I/O。这个值对每个系统都不同，可以试着从 33/24 开始试一试。

3.1.4.2 NO 值的设置
一般要增加每个节点的 the wall 值，netstat –m 观察 mbuf
在/etc/rc.net 中，no -o the wall=5120(kbit)；用-d 默认大小，实存的一半。

3.1.4.3 /etc/hosts 或 NS 的/etc/resolv.conf 设置
别少了 boot 地址和/etc/hosts 中的
127.0.0.1 loopback localhost

3.1.4.4 cron 和 NIS 的问题
HACMP 的节点如使用了 NIS，因/etc/passwd map 到 NIS Server，且在/etc/inittab 中 cron 在 rcnfs(包含 ypbind)之前启动，其相关用户信息在 NIS Server 中，还取不到，因此会有问题可如下处理：
方法 1：如果 cron 可以在 HACMP 之后启动，则将 cron 在 inittab 中的运行级别由 2(系统 boot 时)改为-a，并从 rcnfs 之前调到之后。
方法 2：在 rcnfs 之后加一个 shell 来刷新 cron。

3.1.4.5 /.rhosts 文件

配置 HACMP 时，在每个节点的/.rhosts 中加上所有的 service 和 boot 地址，用于中心配置的/usr/sbin/cluster/utilities/clruncmd 命令和/usr/sbin/cluster/godm 后台进程才可执行。其它象 rcmd、rsh 和 C-SPOC 命令都需要/.rhosts，配置完后，出于安全考虑可以删去这些 Service、boot 地址。

SP with HACMP ES(Enhanced Security)不用/.rhosts 文件来控制。

3.2 网络连接和测试

3.2.1 TCP/IP 网络

3.2.1.1 线缆

距离的限制；

2 台 Hub 以避免 Hub 的单点故障，H1 坏，出现 Swap-adapter 事件，Standby 网卡接管网络，H2 坏出现 fail-standby 事件，可自行定义 notify method 通知管理员。

3.2.1.2 IP 地址和子网

Standby 网卡用于 Cluster Manager 的内部通信和 Service 网卡不能在同一子网，即使同在一个物理网上，但子网掩码要相同。

3.2.1.3 测试

netstat：网卡状态，通信路径

ping：点到点连接

ifconfig：网卡的 IP，子网掩码，广播地址

/tmp/hacmp.out：/etc/rc.net 是否出错？

boot IP→Start HACMP→Service IP？

lssrc –g tcpip：inetd running？

lssrc –g portmap：portmapper running？

arp：是否有相同 IP、hardware address？

3.2.2 非 TCP/IP 网络

3.2.2.1 线缆

RS232：用 null-modem 线，超过 60M 要考虑用光纤转换；

TMSCSI：要求 Diff 的 SCSI 卡，两端终结，卡上电阻要去掉；

TMSSA：要求 SSA Multi-Initiator RAID 卡，微码>1801，HACMP4.2.2 with IX75718

3.2.2.2 RS232

用 smit tty 加一个 RS232 tty 并禁止 login

3.2.2.3 TMSCSI

要将一块 Diff SCSI 卡的 TARGET MODE 设成 enabled，首先要将它的子设备(其上的 HD)设成 Defined，用命令 rmder –l hdiskx，才能改成功，其后 reboot 或 cfgmgr 会出现 tmscsix 设备，替带该卡的 SCSI ID。

3.2.2.4 TMSSA

要将每个节点的节点号由默认值 0 改为各不相同的值，

用 chdev –l ssar –a node number = #修改；
看是否改好用 lsattr –El ssar；
再运行 cfgmgr，配置 tmssa#，用 lsdev –c |grep tmssa 检查。

3.2.2.5 测试

RS232：cat < /dev / ttyx 另一边 cat /etc/environment > /dev/ttyy
TMSSA：cat< /dev/tmssax.tm 另一边 cat /etc/enviroment > /dev/tmssay.im x.y 代表 node number
TMSCSI：cat < /dev/tmscsix.tm 另一边 cat /etc/environment > /dev/tmscsiy.im

3.3 群集磁盘设置

3.3.1 SSA

3.3.1.1 线缆

参考 2.3.1.3，最多 3 个连续的 dummy 盘，SSA 线 25M，但用光纤拓展可达 2.4km。

3.3.1.2 AIX 配置

如线缆连接正确，且 SSA 软件安装正确，boot 时会自动生成下列设备：
1.SSA Adapter Router (ssar)
它是个概念性的东西，总在 Defined 状态。
lsdev -c |grep ssar
2. SSA Adapter
lsdev -c |grep ssa
ssa0 Available 00–07 location SSA Enhanced Adapter
3. SSA Disk
SSA 的物理盘总对应一个 pdisk，而 hdisk 则是逻辑的，可能是一个阵列磁盘组，或未参于任何阵列的单独的盘。阵列管理工具可将一个盘指定为热交换盘，阵列后选盘：candidate disk 不是热交换盘，区别：前者未加入某个阵列，后者已加入。
logic disk：hdisk0，hdisk1……
支持字符设备和块设备：rhdisk0,hdisk0
支持 I/O CTL 子函数 for non service and diag functions
接受读写，可以成为 vg 的成员
lsdev -Cc disk | grep SSA
physical disk：pdisk0,pdiskl
错误日志会记录其错误
pdisk 是字符设备，无对应块设备
支持 I/O CTL 子函数 for service and ding function
不可读写子函数调用
lsdev – Cc pdisk | grep SSA
Diagnostics 关于 SSA 的 diag 工具
diag→Task selection→SSA Service Aids
set service mode：确定某块盘在 loop 中的位置，且可删除它
Link Verification：确定连接的状态

Config Verification：确定 pdisk 和 hdisk 的关系
Format disk：
Certifg disk：测试某磁盘是否可读
Display/downlwwd：看盘的微码和下载新的微码
避免从多台连接到 SSA 的主机上同时运行 diag SSA

3.3.1.3 微码升级

包括卡微码和盘微码

3.3.1.4 配置 RAID

早期的 SSA 本身只支持 RAID5，RAID0，1 由 LVM 支持。6230 以后的 SSA 卡可以支持到 RAID01。

smit ssaraid

3.3.2 SCSI

3.3.2.1 线缆

SCSI 卡见 2.3.2.2/3.3.2.3

RAID：一台 7135 支持 2 X15 块 SE 磁盘，HACMP 环境中支持 2 台 7135。7135 可配两块 Controller，都占用一个 SCSIID，应连在不同的 SCSI bus 上。避免 SCSI adapter、cable、controller 的单点失败

3.3.2.2 连接 RAID 子系统

7135-110 可在 8bit 或 16bit 的 bus 上
7135-210 只连在 16bit bus 上，且要 Enhanced SCSI-2 Diff F/WADP
7135-110 8bit bus 总长最长 19m
7135-110 16bit bus 总长最长 25m
7135-210 16bit bus 总长最长 25m

3.3.2.3 卡 SCSI ID 和 Termination 电阻处理

SCSI-2 Diff Controller 卡（4—2）用于 8bit 设备连接；
需要去掉 u8,u26 两个电阻，避免 bus 中途终结；
注意：Apply change to DB only →Yes 用 Smit chgscsi 修改卡 SCSI ID(0-7)避免冲突；
或 chdev –l scsil –a id=6 –p 越大越优先；
改后 reboot。

SCSI-2 Diff F/W Adapter/A 和 Enhanced SCSI –2 Diff F/W Adapter/A 通常用于 16bit 设备，也可用于 8bit 设备，需要去掉 RN1，RN2，RN3 三个电阻；
Smit chgscsi 改 External SCSI ID 避免冲突，Apychg to DB only→Yes；
chdev –l ascsil –a id=6 –p；
改后 reboot。

3.4 共享的 LVM 配置

一般来讲，在一个节点做完 vg、lv、FS 定义后，在其它节点 import。
Non-concurrent access 一般采用于 fs，而 concurrent access 采用 row lv。

3.4.1 创建共享的 VG

3.4.1.1 创建 Non –concurrent VG

Smit mkvg

VG name：在 Cluster 中是唯一的

自动激活 vg：No，不可在启机时就激活 vg

创建后就激活：Yes，可以继续建 LV 等

VG MAJORNUMBER：如不使用 NFS，可用系统默认值，如使用 NFS，应先观察其它节点，用 lvlstmajor，找到未使用的值。

3.4.1.2 创建 Concurrent VG

在 SSA 设备上建 Concurrent VG,要在每个节点上对 Ssar 指定各不相同的 node-number(非零值)。如何指定：

A、如 concurrent 资源组的 SSA 磁盘 fencing，则同步资源时，自动指定；

B、否则：chdev –l ssar –a node number = #(非 0，且各不相同)

在 7133 上建的 concurrent VG，也可以 varyon 成 non – concurrent 方式，而且在 vg 上建 lv 时，必须要 varyon 成 non-concurrent

smit mkvg

vg name：指定唯一的　　创建后 Varyonvg:no

PP size：default　　VG MAJORNUMBER：default

PV names：选择 hdisk　　创建 concurrent capablevg：Yes

重启时自动 vanyonvg：no　　自动 varnon 到 concurrent：no

在 7135 上建 Concurrent VG 选项同 3.4.1.1 创建 Non – concurrent VG，且 Creat VG concurrent capable 设成 no。

同样在 vg 上建 lv 时，也要 Varyon 成 non – concurrent 方式。

3.4.2 创建共享的 LV 和 FS

Smit crjfs 建 fs，就会同时建 lv，但建好后要对 lv 和 loglv 改名。

重启时自动 mount fs ：no

Start Disk A ccouting：no

对 lv 和 loglv 改名

1、用 lsvg –l vgname 找类型为 jfs 和 jfslog 的 lv

2、用 smit chlv 改名

3、/etc /filesystems，确认 fs 对应的 Dev 已改为新 lv 名，所有 fs 的 log 已由原 loglv 改为新 loglv

增加 lv 的拷贝数。RAID 设备上的 lv 不必做。

1、smit mklvcopy 先加 jfslog lv，再加 jfslv，以免空间不足。

2、lsvg –l vgname 比较 pp 和 lp 数，拷贝应在不同硬盘上。

3、lspv –l hdiskx 了解 lp 和 pp 是否相等，判断是否同一 lp 的 PP 在同一盘上。

测试文件系统

1、fsck /fsname

2、mount /fsname

3、umount /fsnme

3.4.3 镜像原则

只对非 RAID 设备做，RAID 设备不要做 mirror。
拷贝应在不同的盘上，这些盘由不同的卡控制，卡在不同的 drawer 中。

3.4.4 Importing to Other Nodes

3.4.4.1 varyoffvg
在源节点完成上述操作后 varyoff vg
3.4.4.2 Importvg
在目的节点 smit importvg
vgname：和源节点一致
PV name：vg 中的一个 PV，注意同一 PV 在不同节点的 hdiskx 可能不同
import 后激活：Yes
vg MAJORNUMBER：不用 NFS，则用 default 值，用 NFS 要和源节点相同
3.4.4.3 修改 vg 重启时自动激活 no
import vg 时，该项自动设为 Yes，应手工改为 no
smit chvg 还要设 A QUORUM of disks required to keep the vg online 何值，详见 3.4.5
3.4.4.4 在目的节点上 varyoffvg

3.4.5 Quorum 合法票数/人数/PV 数

Quorum 对 RAID 无意义。
Quorum 是 LVM 的特性，用来决定 varyonvg 是否能执行，和一个或多个 PV 失败时，vg 能否继续 varyon。
Vg 中的 PV 都有一个 VGDA(描述 vg 的 PVs 和 LVs，lp→pp 的映射)和一个 VGSA(PV 和 PP 的状态，还可以判断 mirror copy 的状态是否同步)。
3.4.5.1 Quorum at vary on
当半数以上 VGDA、VGSA OK，varyon 成功(不包含 50%)
3.4.5.2 Quorum after vargon
当 write 一个 PV 失败时，别的 PVS 上的 VGSA 都被通知到该 PV 失败，只要可用的 VGDA，VGSA 仍多于半数，该 VG 不会被 varyoff，当然失败盘的数据不可访问，除非这些数据有镜像且在好盘上。注意 Quorum 和 Mirror 没有一点关系。
3.4.5.3 Disable or Enable Quorum
Enable 是 Default 值
At：当 Quorum 超过半数，vg 可以 Varyon。
After：当 Quorum 仍超过半数，vg 不会被 varyoff。
Disable：chvg –Qn vgname
At：PVs 必须是 100%OK，vg 才可以 Varyon。
After：只要有一个 PV OK，vg 就不会被 varyoff，不能发现连续多次坏盘的情况。
可以用 –f 强制 varyonvg 一个有坏盘且 disable Quorum 的 vg，disable Quorum 是为了防止数据乱，但你要对数据分布十分了解，强烈反对在 HACMP Script 中用-f 参数

Quorum 对 Non – concurrent Access 的意义

Quorum 在一定程度上，防止数据混乱或提高可用性，因有利有弊，对 Non-Concurrent 没有多少实际意义。

Quorum 在 Concurrent Access 时必需被 enable，Disable 后导致数据混乱。

3.4.6 Task Guide

配置 HACMP 的一个图形工具，可减少错误，减化工作。

命令行启动：/usr/sbin/cluster/tguides/bin/cl-ccvg or smit:smit hacmp→Cluster System Management→Cluster Logical Volume Manager→Taskguide for Creating a Shared Volume Group

可完成：initial and sharing nodes，disks，concurrent access，vgname，ppsize，cluster setting……