AIXCHINA

HACMP for AIX
学习笔记

www.aixchina.com

Aix HACMP

IBM Certification Study Guide Test 167

前言

Test167 已经被 Test 187 替代，认证考试的书也更新，新旧资料大约有 70%的内容是相同的。英文功底深厚的朋友请直接看 187 的学习指南，不用浪费时间看这本整理的 167 学习笔记。

第 2 章 群集计划

2.1 群集节点

2.1.1 配置选择

每个节点最少 32M 内存，1GB 硬盘。

2.1.2 群集节点要求

需要考虑处理器能力能否满足应用的要求，业务预期增长，I/O 槽是否充足。
节点对其联上的每个网络可有多达 7 个的 Standby 网卡。
Sharevg 做镜像时，要考虑一台机器连接硬盘的两块 I/O 卡要在的不同总线。

2.2 群集网络

分为 ICP/IP 和非 TCP/IP 两大类

2.2.1 TCP/IP 网络

2.2.1.1 支持类型

Generic IP，ATM，Ethernet，FCS，FDDI，Sp switch 私有（Private 网络），SLIP，SOCC，Token-Ring。

HACMP 每个群集最多支持支持 32 个网络每个节点最多支持 24 个网卡。

2.2.1.2 各种类型的特性

ATM：点到点，和 FCS、SP Switch 都不支持硬件地址切换。

SLIP：一般不用，太慢

SOCC：很少用了，withdrawn

IP 地址接管：只有 SP Switch 可以用 ifconfig alias 在一块卡上实现，其它都需要两块卡。

2.2.2 非 TCP/IP 网络

HACMP 可以不用非 TCPIP 网络仍可工作，但建议采用，以区分网络（TCPIP）故障，还是节点故障（心跳线）

2.2.2.1 支持类型

Serial (RS232)
Target –mode SCSI
Target-mode SSA
在 HA 的配置中，这三种 Network Type 都是 Serial。

2.2.2.2 特性

Seral： 双机时，只要一个串口，多机时，每节点要二个串口构成环；
S7X 无串口，因此要订多口异步卡；
SP 的节点，多个串口只有一个可用于 HACMP；

TM SCSI： 只有 SCSI-2Diff 和 SCSI-2 Diff F/W 以后的卡支持；
SCSI/SE 和 SCSI-21SE 不支持；
建议一个群集中不要超过 4 个 target mode SCSI 网络；

TMSSA： 用 6215/6219 Enhanced RAID-5 以后的卡，支持 Multi-Initiator 特性，微码高于 1801

2.3 群集磁盘

2.3.1 SSA Disk

分 2 种

- 7131 SSA Multi-Storage Tower Model 405
- 7133 SSA Disk Subsystem 010,500,020,600,D40,T40

所有的 7133 都有可热插拔的冗余电源，风扇，线也是热插拔的。
7131，7133 的硬盘都是热插拔，7131：2-5 个，7133：4-16 个。

2.3.1.1 磁盘容量

9.1G 18.2G Ultra star 盘 Buffer 4096KB 160MBPS

2.3.1.2 卡 6214 6215 6230 6216 6217 6218 6218

6215 4-N PCI Enhanced RAID-5
6219 4-M MCA Enhanced RAID-5
最多 8 块，每 loop 可混用；
单个系统中最多 4 块卡，6217、6218 也支持 RAID-5，但每 loop 最多 1 块，不可用于 HACMP。只有 6215、6219 支持 TMSSA。早期的 6214（MAX2）到 6216（MAX8）都不支持 RAID-5

2.3.1.3 SSA LOOP 规则

- 每个 LOOP 必须连结在一块卡的某对接口上（A1＆A2 或 B1＆B2）
- 一块卡上的 2 对口不能在同一个 LOOP 卡。
- 一个 LOOP 最多 48 个设备。
- 一个系统中最多 2 块卡在同一个 LOOP 上。

6215/6219（可混用）的规则： 每个 LOOP 中最多几块取决于下列条件

- 8 块卡：当 loop 中没有一块 Disk 参于阵列，且不使用快写操作

- 2 块卡：loop 中 Disk 可参于阵列，但不使用快写操作，这点意味着，当节点多于 2 个时，只能采用 Mirror 保护
- 1 块卡：loop 中 Disk 参于阵列，且使用快写，不能用于 HACMP

陈列中的盘应在同一个 loop 中。
IBM7190-100 SCSI TOSSA converter 一个 loop 中最多 4 个。

2.3.1.4 SSA 优点

每个环路可以有 MAX 127 设备，当前支持到 96 设备，自动配置不用设地址，Concurrent Access to disks，双全工，不用终结器，不需仲裁，热插拔，25m 铜线，10km 光纤，双路到设备，20MB×2×2×2=160MB A1=20×2
A2=20×2
B1=20×2
B2=20×2

2.3.2 SCSI 盘 7135

2.3.2.1 容量

135G for RAID0 108G for RAID5

2.3.2.2 数量

因线缆长度限制，HACMP 支持一条 SCSI 总线上 2 台 7135。

2.3.2.3 支持的卡

SCSI-2 Diff 以上的卡

- MCA

SCSI-2 Diff CTL 8it
SCSI-2 Diff Adp 16bit
Enhanced SCST-2 Diff F/W Adp 不支持 7135-110 支持 7135-210

- PCI

SCSI-2 Diff Adp
Diff Ultra SCSI Adp 不支持 7135-110 支持 7135-210

2.3.2.4 优劣

- 支持 RAID0、RAID1、RAID3(Model 110 only)、RAID5

每个 LUN 可有自己的 RAID Level

- 多个 LUN

虽然只占用 1 个 SCSI ID，但阵列支持 6 个 LUN，在 AIX 看来都分别对应一个 hdisk ，可分给不同的 vg，不同的系统。

- 冗余电源，风扇，在线维护，可选双控制卡。

2.4 资源计划

资源类型有：VG、Disks、FS、FS to be NFS mounted/exported、IP、APP

2.4.1 资源组选项

分三类资源

☞ Cascading Resource Groups

要点：Inactive Takeover 为真时，第一个启动的节点接管资源，随后加入的如有更高优先级则接管。避免开机时，不必要的接管。

Inactive Takover 为假时，第一个启动的节点不接管资源（除非有最高级别）随后加入的如有更高优先级则接管。

☞ Rotating Resource Groups

先加入的节点就得到资源，除非节点故障或人工要求接管，否则不发生接管

☞ Concurrent Resource Groups

这类资源不会发生接管，因为节点都可以访问到它们。

资源一般指裸磁盘，有裸逻辑卷的 vg，应用服务程序。

2.4.2 共享的 LVM 内容

2.4.1.1 无共同访问的磁盘配置，

Hot-Standby 热备份 Cassading 2 次接管，A 坏 B 接管，A 好后接管回

Rotating Standby 轮转备份 Rotating A 坏 B 接管，A 好不再次发生接管停顿

Mutual Takeover 互为备份 Cascading 充份利用设备，但也有 2 次接管过程

Third-Party Takeover 第三方接管 Cascading 避免性能问题

2.4.2.2 共同访问的磁盘配置

7135 支持 4 个节点，7133 支持 8 个节点

这种情况下，IP 地址仍应设为 Cascading Resource

2.4.3 IP 地址接管

2.4.3.1 网络拓朴

Single Network： 网络存在单点失败

Dual Network ：

Point-to-Point Network：

2.4.3.2 网络

2 个要素：

☞ 网络名：同一个物理网络用同一个网络名

☞ 网络属性：

public 公有：联结 2-32 个节点，允许 client 访问

private 私有 :提供节点通讯，不允许 client 访问，但 ATM 和 SP Switch 允许 client 访问

serial：心跳

2.4.3.3 网卡

☞ Adapter Label：在/etc/hosts 中一个 label 对应一个 IP 地址

☞ Adapter Function：
Service Adapter
Boot Adapter 和 Service Adapter 同一子网，SP 中用 ifconfig alias 定义于 CSS0 网络
Standby Adapter 和 Service Adapter 在不同子网，0-7 块每个系统

2.4.3.4 硬件地址交换

IP takeover 后，通过硬件地址交换，将 IP 地址和新网卡相联，不用专门去刷新 ARP Cache，适用于 Ethernet，Token-Ring60 秒，FDDI 120 秒。不适用于 SP、ATM，如果不采用，要
将 Client 的 IP 地址加入 PING_CLIENT_LIST 变量；
或，该变量在/usr/sbin/cluster/etc/clinfo.rc 文件中，将可刷新 ARPcache。

2.4.4 NFS Exports and NFS Mounts

File systems to Export，export 本地的 FS，让 client 和其它节点可以 mount
File systems to NFS mount ，远程的 FS，多个节点都可以 mount 该 FS，当前拥有它的 Node 失败后，接管的 Node 同时接管该 FS，其它节点远程 mount 它。

2.5 应用计划

应用程序也是要接管的资源，通过定义 Scripts 来完成这些程序的 Stop 和 Start。

2.5.1 性能要求

出现节点失败接管时，应用的性能如何？

2.5.2 应用 Startup 和 Shutdown 程序

注意 HACMP 不会同步应用的 Scripts，所以要手工同步保证路径的一致，权限一致。

2.5.3 Licensing

有些应用和处理器有关，因此需要 2 个或多个 licence
Floating Licenses：通过 License server 授权
Node-locked：和节点有关

2.5.4 接管后多个应用是否会冲突

2.5.5 关键/非关键优先权

有时需要停止非关键应用，保证关键应用。

2.6 客户化计划

2.6.1 事件客户化

不能增加群集事体的新定义，但可用 HACMP 的 SMIT 增加“ 监听某事件并做何处理 ”的定义。

2.6.1.1 特殊应用的要求

有些应用可能要求在节点失败接管事件发生时，监听到并重置计数器，解除锁定等操作。

2.6.1.2 Event Notification

诸如 network down 和 network up 的事件可通过 mail 告知。

2.6.1.3 预防性事件错误纠正

如一个用户 logging off，系统可多次偿试 umount 他的 FS，以保证成功.

2.6.2 Error Notification

通过 odmadd<filename>文件内容如下：将 Err Notification obj 加入 ODM，对某些错误做出反应，如：

errnotify:

en_name = “ Failuresample ”

en_persistenceflg =0

en_class = “ H ”

en_type = “ PERM ”

en_rclass = “ disk ”

en_method = “ errpt –a –l $1|mail –s ‘Disk Error’root ”

2.6.2.1 单点失败硬件内容恢复

SP 机器中，节点的 Switch adapter 失败，就相当于节点失败，因为只有一块 Switch 卡，除非 Switch network 没有用到。

这类 Error Label，如 HPS_FAULT9_ER，HPS_FAULT3_ER，除用传统方法加入 ODM，还可用 Smit hacmp > RAS Support > Error

Notification > Add a Notify Method 加入，这样单 Errlabel 记入错误日志时，就有相应处理。

2.6.2.2 Notification

上述加入 ODM 的操作，HACMP 不能同步到另一台机器，须手工加入各机。

2.6.2.3 应用失败

可通过 errlogger < message > 记入错误日志，并加入相应处理到 ODM 中。

2.7 用户 ID 计划

2.7.1 群集用户和组 ID

管理员要保证各机的/etc/passwd 和/etc/security/*的文件一致，可用 rdist 或 rcp 同步，SP 用 PCP 或 Super 同步。C-SPOC(Cluster Single Point of Control)群集可自动同步（除/etc/security/ passwd）。

2.7.2 群集密码

如果未采用 NIS 或 DCE，即使是 C-SPOC 命令，也需要手工拷贝/etc/security/passwd 文件到各机。

2.7.3 用户 Home 目录计划

节点失败时，要保证用户的 Home 目录持续可用。

2.7.3.1 Home Dir 放在 Shared Volumes

局限性，一个时刻，Home Dir 只对一台机器有效可用。

2.7.3.2 NFS-Mounted Home Dir

用户的 Home Dir 可以同时 mount 到多台机器，但有风险，包含 Home Dir 的机器失败后，大家都访问不到。

2.7.3.3 NFS-Mounted Home Dir on Shared Volumes

能解决上述问题，当主机失败时，备机先 Break 它 mount 的主机 NFS 文件锁，再 umount NFS，取到 Shared Volumes，mount Shared FS，再给用户提供服务

第 3 章 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……

第 4 章 HACMP 安装和群集定义

4.1 安装 HACMP

预先要求：A1X4.3.2 以上，/usr 下空间够，HAview 要求 nv6000

4.1.1 首次安装软件包

Cluster.base：所有 server node 都要安装此基本内容
包括：*.base.client.lib, *.base.Client.rte, *.client .utils *.server.utils,*.server.diag, *.server.events,*.server rte
Cluster.cspoc 单点控制相关命令和环境，所有 Server node 要安装，包括 *.rte,*.cmds,*.dsh
Cluster.adt：开发例子，头文件包括 Server 和 client 装在开发机上
Cluster.man.en_US.data：man 的信息
Cluster.msg.en_US：msg 的信息
Cluster.vsm：用于图形管理程序 xhacmpm 的标记和图块
Cluster haview：装在 Netview 机器上，不是装在群集节点机上
Cluster.man.en_US.haview.data：用于 haview 的 man
Cluster.msg.en_US.haview：用于 haview 的 msg
Cluster.taskguides：方便创建 Shared vg 的工具
Cluster.clvm：包括 Concurrent Resource Manager 选项
Cluster.hc：应用的心跳守护进程 Oracle 并行 Server 用到它
Bos.rte.lvm.usr 4.3.2.0：要先装此包，才能用 clvm 每个节点(Server node)装好后，重启

4.1.2 从早期版本升级

备份自己的 script 和配置文件，做 mksysb，如现有版本未 commit，先 commit 再升级，用 HACMP Snapshot 保存现有配置，不要在群集中，混合运行不同的 HACMP 版本。
4.1.2.1 A1X4.3 以前，将 HACMP 从 4.1.0-4.2.2 升级到 V4.3
先升级 AIX 到 4.3.2
再在其中一个节点安装 HACMP4.3，重复直到 over
检查升级后配置
只含 client 的升级，如从 HAMP4.1 升上去，要先删去原有的 Server 部分，否则会有问题
4.1.2.2 A1X4.3.2，将 HACMP 从 4.2.2 升到 4.3

4.2 定义群集拓扑

通过 SMIT 输入到 HACMP ODM 中，SMIT HACMP。

4.2.1 定义群集

Cluster ID 1~99999 Cluster Name Up to 31 char

4.2.2 定义节点

Node Name: 群集中节点名按 ASCII 序排列，出于心跳目的，系统认为相邻的节点名代表相邻的节点，最高和最低节点名也相邻，初始配置后，也可以加或改变节点名，看 HACMP 管理资料。

4.2.3 定义卡：

- 卡名 Adapter IP Label：对应于每个 IP 地址的 ASCII text 描述。通过 Add an Adapter 加入。名字中不要用“ - ”。因 Clverify 检验时会更花时间。
- IP/Adapter ：对有 IP 地址的 Adapter，如按 IP Label 在 DNS 或/etc/hosts 中查不到，则要在此给出 IP 地址。对 RS232 给出/dev/ttyN，对 tm scsi 给出/dev/tmscsiN，对 tmssa 给出/dev/tmssaN.im 或/dev/tmssaN.tm
- 硬件 Network TYPE：如 Serial(RS232)，TmSCSI，TmSSA，Ethernet 等。
- 网名 Network Name：每个物理网络给一个唯一网名。
- 分类 Network Attribute：public：如 Ethernet、TokenRing、FDDI、SLIP
  private：如 SOCC、ATM、SP Switch
  serial：如 RS232，tmSCSI，tmSSA

用途 Adapter Function：Service，standby，boot

注意：serial 和 private 类没有 standby 之说，但 ATM 例外支持 standby。

注意：SP 机器中，Ethernet 可做为 service 卡，但不能做 IPAT。IPAT 通过 ifconfig alias 对 css0 网设定

注意：ATM 中，卡用途应指为 SVC-S 以表明用于 HACMP Server

- 硬件地址 Adapter Hardware Address：适用于 Ethernet、Token Ring、FDDI，当定义 Service adapter，且它有 boot address，并且用到硬件地址切换时，可指定其 Hardware Address，其它情况不用。
- 节点名 Node Name：除了 service 卡会共享于几个 Node 之间，其它卡都可给于节点名。

当 IPAT 有用到时，/etc/inittab 中和 IP 相关的入口，如 rctcpip，rcnfs 的运行级别改为 a，则这些服务 boot 时不会启动，而随 HA 启动。

4.2.4 配置网络模块 Network Modules:

网络模块用于维护各自网络的连通性，当一定时间收不到心跳，可以判定是网络失败。可做配置的内容就是检测的敏感性。

4.2.5 同步群集的定义：

群集中定义做的任何修改，都要同步。

如果是初次安装 HA，cluster manager 没有在本地（发同步命令的机器）运行，所有在默认配置目录中的 ODM 数据都被拷贝到其它节点，但如果其它节点此时有运行 Cluster manager 则同步操作不能执行。

两个同步选项：

Ignore Cluster Verification Error：Yes/NO

No：如果 Verification 有错，则不同步，可看 errlog

Emulate or Actual: 是模拟还果真正同步。

4.3 定义资源：

资源包括：Disks、VGs、FSs、Network Address、App servers

多个资源构成资源组，和节点的关系有：cascading、concurrent、rotating

4.3.1 配置资源组 Add a Resource Group

4.3.1.1 为资源组配置资源。

注意：配置带 NFS mount point 的 cascading resource group 要将 IPAT 也加入资源组，还要将 Filesystems Mounted Before IP configured - Ture.

注意：配置带 IPAT 的 cascading 资源，每个节点的资源组不要超过 N+1 个，N 是 standby 卡的数目。（网络中拥有 stby 卡的节点数目）。

选项： Service IP lable：被接管的 IP。

Filesystems：HACMP 会自动处理和该 FS 相关的 VGs 和 RAWDISK PVIDs

FS 一致性检查：fsck(默认)logredo(快速恢复)

FS 恢复方法：串行(默认)并行（更快但不能用于 shared nested 嵌套的 FS）

FS to Export：是上述 FS 的子集。

FS to NFS mount：资源链中的所有节点（不含当前拥有节点）都会偿试去远程 Mount

最后：同步群集资源。

4.3.1.2 配置 run-time 参数：

debug 级别：high：所有 cluster manager 行动被 logged。

low：只有 error 被 logged。

是否使用 NIS 或 DNS。

4.3.1.3 定义 APP Server

一般对应于一个 script，注意路径名字，权限的一致。

4.3.1.4 同步群集资源

第一次同步，所有节点要在它的 boot 地址上，否则/etc/rc.net 不会修改未同步的节点，将不能加入群集。

4.4 初始测试

4.4.1 clverify 检查

/usr/sbin/cluster/diag/clverify 包括软件和群集检查；
软件：等价于 lppcheck –v；
群集：拓朴和配置检查　　smit clverify；

4.4.2 初次启动

判断 HA 是否启动：用 ps –e|grep clstr，lssrc –g cluster 或用 netstat –i 看 IP 是否在 boot address。
启动： smit clstart：start → now，
broadcast msg → true，
start cluster lock services → false，
start cluster info daemon → True
在每个节点都启动后，用 clstat 看群集状态。

4.4.3 检查 takeover 和 reintegration

在一个节点 smit clstop graceful with takeover；
在同一个节点 smit clstart。

4.5 群集快照 Snapshot

保存群集的配置，但 APP Server 的 Script 不会保存；
这些信息都是 TXT，可方便于问题分析；
注意：节点的 HACMP 版本应相同。

4.5.1 恢复一个快照

恢复配置，在任意一个节点恢复配置，可完成所有节点恢复，如果所有节点的 cluster 未激活，恢复将只涉及默认配置目录即可(推荐)
如果本地节点 Cluster Service 激活，则恢复是一次动态重配置过程，还将包括激活配置目录。但须要 2 次：一次恢复群集拓朴定义，一次恢复群集资源定义。

第 5 章 群集客户化

5.1 事件客户化

在/usr/sbin/cluster/events 下有预定义的事件；

可以客户化的内容包括：加、改、删事件，事件之前和之后处理事件通知，事件恢复和重试。

5.1.1 预定义群集事件

5.1.1.1 节点事件

包括：node_up　　归还

本机　　其它机

node_up_local　　node_up_remote

acquire_service_addr　　release_takeover_addr

acquire_takeover_addr　　stop_server 停下属于其它机的应用

get_disk_vg_fs

|　　release_vg_fs

node_up_complete　　cl_deactivate_nfs

|　　node_up_remote_complete

node_up_local_complete

start_server 起本机 server

node_down　　接管

node_down_local　　node_down_remote

stop_server 停本机 server　　acquire_takeover_addr

release_takeover_addr　　get_disk_vg_fs

release_vg_fs　　node_down_complete

release_seruice_addr　　node_down_remote_complete

node_clown_local_complete

node—remote_compete

start_server 起远程 server

5.1.1.2 网络事件

network_down：分为 Local（个别机器 down）和 Global（整个网络 down）两种

network_down_complete

network_up

network_up_complete

5.1.1.3 网卡事件

swap_adapter：切换 service 和 standby 网卡

swap_adapter_complete：证实 Local ARP 也刷新了
fail_standby：Standby 卡坏，或发生了 IPAT，standby 已变为 service
foin_standby：Standby 卡重新可用

5.1.1.4 群集状态事件

config_too_long：节点在重配置状况超出 6 分钟
reconfig_topology_start：开始动态重配置
reconfig_topology_complete：
reconfig_resource_acquire：受动态配置影响的资源被节点请求访问，或释放
reconfig_resource_release：
reconfig_resource_complete：

5.1.2 pre-and post-Event Processing

smit hacmp→cluster Configuration→Resource→Cluster Events→Change show Cluster Events 定义某个命令或 script 在某事件之前或之后执行

5.1.3 Event Notification

在 5.1.2 节的 smit 界面中，有 Notify 命令(指定命令或 script)，可用于通知管理员某事件正要发生，或已经发生

5.1.4 Event Recovery and Retry

可以指定事件的命令或 script 运行失败时，做 recovery 操作，如 retry 非 0，则再次运行该事件的命令或 script，如 umount fs 不成功，因为有进程在访问它，可通过 recovery 过程杀进程再偿试 umount。

5.1.5 客户化事件处理注意事项

要在 notify,recovery,and pre-or post-event 的 script 文件中指定 sh，如：#!/bin/sh；
如果 node_down 事件的 force 选项打开，上述 4 个处理将不发生；
同步 HA 配置不会自动拷贝 script，要手工做。各节点的 script 内容允许不同，但名字、位置、权限位要相同。

5.1.6 事件模拟

为测试配置，HA 可在各节点运行一个事件模拟工具，以提供各事件发生模拟，输出在 EMU_OUTPUT 环境变量中指定，默认是/tmp/emuhaemp.out。

5.2 Error Notification

AIX 的 Error Notification 工具可定义复杂的 error 检测条件，监听系统错误日志，当有满足条件的错误发生时，notify method 做出反应。例如当一块 scsi 卡坏，HA 和 LVM 都不会做出反应，这时 Notify method 做出反应关机，让另一节点接管 shared disk。

5.3 网络模块，拓朴和组服务

网络模块的增、删一般极少用到，常用的是修改失败检测率，分为三档：快、正常、慢，对 Ethernet 而言分别是 4 次/秒，2 次/秒，1 次/秒，如果你的网络很忙，可选慢，以免心跳信号被阻塞，引起错误的接管。

拓朴和服务，Charge/show Topology and Group Service Configure 一般只改 log 的长度。

5.4 NFS 相关内容

5.4.1 创建共享 VG

因为 NFS Client 对 NFS 的访问有用到 Shared VG 的 Major Number，所以在创建 Shared VG 时，要指定唯一的 MajorNum，在其它节点 importvg 该 VG 时，也要用该 Major Num。

5.4.2 Export NFS

HACMP 的默认 Script 没有用到/etc/exports 文件，是通过 cl_export_fs 工具中，调用 exportfs 带-i 参数指定文件系统，该文件系统在 HACMP ODM 中指定。用户可以修改 cl_export_fs。

5.4.3 NFS Mounfing

Client 须要创建一个 NFS Mount Point。

5.4.4 Cascading 接管交叉 mount 的 NFS

5.4.4.1 Server-to-server NFS 交叉 mount

A /afs 在 A 机本地 mount
/afs 在 A 机 nfs-exported
A 机 NFS-mounted NodeB:/bfs

B /bfs 在 B 机本地 mount

/vda 在 B 机 nfs-exported

B 机 NFS-mounted NodeA:/afs

当 A 机失败，B 机用 cl-nfskill 工具关闭打开的 NodeA：/afs 中的文件，umount NodeA：/afs，mount /afs 本地 mount，re-export 给客户，takeover 以后

B /bfs 本地 mounted

/bfs nfs-exported

/afs 本地 mounted

/afs nfs-exported

注意 5.4.1 的要求

节点名和 TCP/IP 卡 label 要一致，否则把节点名做为 NFS hostname 做 mount 会出错。

可以这样处理：1、nodename=service adapter label

2、在/etc/hosts 中为 service adapter label 加 alias nodename

5.4.5 Network Lockmanager 和 NFS 交叉 mount

例如，节点 A mount 一个 FS，并且 export，节点 B 做为其 client mount 该 fs，节点 B 上有应用访问了 fs，并用了 flock 上锁。当节点 A 失败时，节点 B 应做 umount fs，mount fs 本地，再 export，但由于上锁，umount 将失败。在 cl_deactivate_nfs 中加命令清除这些锁，注意会将所有 nfs 的锁都清除。

第 6 章 群集测试

6.1 节点检查

在对 HA 做检查之前，先做节点检查。

6.1.1 设备状态

ding -a cleanup the VPD
errpt | more or errpt –a –j *XXXXXX*
lsdev –C | more all in available state
lsattr -El ascsi0 SCSI 地址是否唯一
cat < /dev/tmscsi#.tm cat /etc/hosts>/dev/tmscsi#.im
做此操作时，HA 要 stop，每个方向做 2 次。
stty</dev/tty# (两机同时)检查心跳线(Serial RS232)

6.1.2 系统参数

date （是否用 daylight save 复时制）
lslicense
smit chgsys 检查高水平线
sysdumpdev –l 和 sysdumpdev –e 确认 dumpspace 正确
lslv hd7 主 dump 足够大，Dump dev 不镜像.
crontab –l
由 HA 启的 APP，没有自动启，more /etc/inittab

6.1.3 进程状态

lsps -a 看 paging space
ps -ef | more
vmstat 2 5 是否 runqueue<5 且 cpu 不忙

6.1.4 网络状态

ifconfig lo0，ifconfig en0，ifconfig en1 网卡配置
/usr/lpp/ssp/css/ifconfig css0 SP switch 卡配置
netstat -i 或 netstat -in 网络配置

netstat -V ent0|more 看替代的 MAC 地址
netstat -m | more 被拒绝的多吗？mbuf
netstat -r 或 netstat -rAn 看路由
no -a | more 看 ipforwarding 和 ipsendredirects
ping 检查所有的 IP 地址
arp –a
lssrc -g tcpip
/etc/hosts 中没有错误，特别是结尾
more /etc/resolv.conf DNS 是否有用，且正确
ps -ef | grep ypbind 和 lssrc –g yp NIS 状态
exportfs 是否有非 HA 控制的 NFS export
Snmpinfo -m dump –o /usr/sbin/cluster/hacmp.defs/address
看 HA 有关的 snmp 信息

6.1.5 LVM 状态

lsvg lsvg -o quorum，auto-varyon，sharedVG 状态
lsvg –l 不存在 stale 分区
df -k 文件系统 mounted，rootvg 下文件系统有空间
lspv PVid 是否设好，且无 ghost 盘
more /etc/filesystems lsfs

6.1.6 群集状态

lssrc -g cluster，lssrc –g lock 看群集的 daemon
/usr/sbin/cluster/clstat 群集和网络接口状态
tail –f /tmp/hacmp.out 看 log
more /usr/sbin/cluster/history/cluster.mmdd 当前时间
tail –f /var/adm/cluster.log
more /tmp/cm.log
odmget HACMPcluster 看节点名
/usr/sbin/cluster/diag/clconfig –v '-tr'检查群集配置
/usr/sbin/cluster/utilities/cllscf 看群集配置
snmpinfo -m dump –o /usr/sbin/cluster/hacmp.defs clstrmgr，看 clstrmgr 版本

6.2 模拟错误

Node F：失败节点，Node T：接管节点
注意观察/tmp/hacmp.out
HACMP 的运行参数 debug 级别设为高
应用服务 Script 包括 set-x 和 echo

6.2.1 卡坏

6.2.1.1 以太网或令牌环网接口错误

这三点都要求的：

Node F：errclear0 清除 errlog

所有节点启动 HA

监视 NodeT 的 cluster log 文件

Node T 不参于下述活动：

ifconfig en0 down　　en0 是 Node F 的 Service adapter，不能是 SP 的管理卡。

Standby 卡将接管 service IP，MAC 地址也接管，原 Service 接口则做为 standby 接口。若 ifconfig en1 down，则又接管回原卡。

6.2.1.2 以太和令版环卡和线坏

拔掉 Service 卡的线，Standby 卡接管其 IP 和 MAC 地址，连回线，原 service 卡接口将变为 standby 接口，拔掉原 standby 卡线，则回复原样，再连回线缆。

6.2.1.3 Switch 卡坏

每个 SP 节点有一块 switch 卡

让 NodeF 作为 Eprimary

(1)Switch 卡坏，在 errlog 中会有 HPS_FAULT9_ER 或 HPS_FAULT3_ER 一类的错误，因此可手工在 errlog 中加入该入口；

(2)如果 network_down 事件被客户化也被监听，ifconfig css0 down

(3)fence out NodeF 从控制工作站　Efence NodeF

上述三种办法会让 NodeT 接管 NodeF 的资源，但 Eprimary 身份将给节点名按字符序最低的节点，NodeB 低于 NodeA。

用 netstat –i 和 ping，lsvg –o，vi 一个测试文件，ps -U<appwid>Eprimang 来检查接管。

在 NodeF 上重启 HACMP，将接管回资源，但 Eprimary 身份不动。

6.2.1.4 7133 卡坏

从卡上拔掉所有的线

errpt –a|more

lsvg –l NodeFvg 有些 lvcopy 会 stale

df –k

lsps –a 看所有 fs、ps 仍可访问，接回线

df-k

lsps –a

6.2.2 节点失败/重加入

6.2.2.1 AIX 崩溃

SMP 的机器，先用 mpcfg -cf11 1 让它快速 reboot；

用 cat/etc/hosts > /dev/kmem 让 NodeF 出 888 错；

NodeT 接管，用 netstat –i， ping，lsvg –o，vi，ps –u <appuid>等检查；

重启 NodeF，HA，重加入后接管回资源，检查同上。

6.2.2.2 CUP 坏

直接断电 NodeF 来模拟。

6.2.2.3 TCP/IP 子系统坏

(1)sh /etc/tpc.clean 停下 TCP/IP

(2)将 sb_max 和 the wall 用 no 改为一个大数，ping NodeT 也会引起 TCP/IP 问题，改之前记录原值加到/etc/rc.net 后，TCP/IP 的失败，只会引起 network_down 事件，一般情况下系统不会有什么反应，除非做过客户化。可看 hacmp.out。节点心跳仍在，所以不是节点失败，用 startsrc –g tcpip 重启，network_up。

6.2.3 网络失败

拔掉所有网线(除了 SP 的管理网卡)，这时 network_down 发生默认不会有行动，除非客户自行定义，可检查定义的动作是否发生。

6.2.4 盘坏

6.2.4.1 镜像 rootvg 盘 hdisk0,hdisk1 坏了一个

bootlist -m normal –o 应包含 hdisk0 和 hdisk1

打开机箱，开机状态下拔去 hdisk0 电源，或热插拔盘直接拔出，errpt –a|more，检查 fs、ps、df、lsps –a 应都可用。

smit clstop，shutdown –F 重启应可成功，关机，接上电源重启。钥匙都在 Normal 档。

lsvg –l rootvg 不存在 stale lv。

6.2.4.2 7135 盘坏

(1)盘是热插拔的，直接拔去。

(2)smit raidiant；RAIDant Disk Array Manager→Chamge/show Drive Status→sdect the appropriate hdisk→select the appropriate physical disk→F4 to select a Drive Status of 83(Faildrive)

7135 的黄灯亮，或用 smit，List all SCSI RAID Arrays 看。

用 df、lsps –a 检查 fs、ps 都 OK。

如有热交换的 RAID5，等其自动重构完，再插回原盘。

如无热交换的 RAID5，插回原盘后 F4 to select status 84 Replace Drv

如 RAID1，syncvg NodeFvg

则 lsvg –l sharedvg 无 stale(用 lslv lvname 也可以，但是只看一个 lv) 黄灯也是灭的。

6.2.4.3 镜像 7133 盘坏，

7133 是热插拔盘，直接拔去，看 fs、ps 是否仍可用。

插回后，做 syncvg NodeFvg

6.2.5 应用失败

依赖于客户的定义

第 7 章 群集故障解决

7.1 HACMP log 文件

/var/adm/cluster.log 包含事件的启停记录
/tmp/hacmp.out verbose 模式下，包含 scripts 的每个命令和参数
system error log
/usr/sbin/cluster/history/cluster.mmdd 记录每天的事件
/tmp/cm.log 记录 clstrmgr 活动，重启 HA 就清空旧记录
/tmp/cspoc.log 记录单点控制命令的信息
/tmp/dms_logs.out Deadman Switch 信息
/tmp/emuhacmp.out 事件模拟输出，来自各个节点，记录 script 中的命令，但实际并未执行

7.2 config_too_long

当一个事件引起的 Script 执行时间超过 360 秒(可能是正常的，也可能是挂起)/tmp/hacmp.out 中有记录。

reconfig too long；sth may be wrong

分析 /tmp/hacmp.out，/var/adm/cluster.log，排除故障后，用 clruncmd 或 SMIT Cluster Recovery Aids 恢复 cluster 运行。

7.3 Deadman switch

当 Clstrmgr 因某个原因，一定时间内没有处理计时器，超时，则 Deadman switch 停止节点，让其它节点接管资源。

可能是 clstrmgr 挂起，或应用的优先级大于 clstrmgr，或以下 4 个性能问题。

7.3.1 调整 I/O Pacing

smit chgsys high-water mark → 33
low-water mark → 24

7.3.2 提高 syncd 频率

将/sbin/re.boot 中 syncd 频率由 60 秒 1 次→30/20/10 秒 1 次，可分摊 IO，避免太集中。

7.3.3 增加通信子系统的可用内存

thewall

7.3.4 改变失败检测率

和网络类型相关，slow 有不同的值，在 7.3.1，7.3.2 不解决问题的情况下，用 SMIT Chg /show a Cluster Network Modwle 将 Failure Detection Rate→slow

7.4 节点 Isolation 和 Partitioned 群集

当一个节点的所有网络连接都断了，但节点仍活着，叫做 Node Isolation，此 Cluster 叫做 partitioned Cluster，此时节点会去接管不该接管的资源，将产生不可预期的问题，所以 HA 中采用 Non-TCP/IP 来传递心跳，以避免过于依赖 TCP/IP 产生此问题。

7.5 DGSP 消息

Diagnostic Group Shutdown Partition

当 partitioned Cluster 出现，接管节点已开始工作，此时如 Isolation Node 重新加入，网络上会有 2 个 IP(相同)，Disk 会 Dual-Attached，为避免问题，较小的(或字母序低的，B 低于 A)Partitioned Cluster 的每个节点会收到 DGSP 消息，立即停机。

7.6 User ID 问题

注意，不同节点，同一用户的 ID 相同，组 ID 亦然。
NFS FS 的权限问题亦然。

7.7 问题解决策略

保存 log 文件，特别的/tmp/hacmp.out 和/tmp/cm.loy 避免丢失；
重视问题，不要太相信用户的描述(基于应用的)；
按步就班，不要凭感觉；
思路开放，不要太多假设，按测试下结论；
孤立问题；
简单到复杂；
一次只改一处；
小事引起大问题　检查 plug connector cable；
做好记录；

第 8 章 群集管理

8.1 监视群集

用/usr/sbin/cluster/clstat 检查群集状态，同时观察/tmp/hacmp.out，用 SMIT Show Cluster Services Screen 观察守护进程状态。

8.1.1 Clstat

Clstat 是 Clinfo 的客户程序，所以 Clinfo 要先在各节点运行。
/usr/sbin/cluster/etc/clhosts 要做好配置，包含各节点。
可运行于字符终端和 X-window，在 X 环境中，用-a 标志可出字符界面。

8.1.2 HAView

HAView 挂在 Netview 之中，要求 SNMP，各节点 SMUX peer daemon 要运行。

8.1.3 log 文件

8.1.3.1 /var/adm/cluster.log
错误、事件。
8.1.3.2 /tmp/hacmp.out
配置和起动脚本的输出是 cluster.log 的延伸，默认是 Debug Level run-time para→high，verbose 状态。
8.1.3.3 /usr/sbin/cluster/history/cluster.mmdd
细节的描述，用于错误处理。
8.1.3.4 System Error Log
8.1.3.5 /tmp/cm.log
clstrmgr 的活动，IBM 人员使用。
8.1.3.6 /tmp/cspoc.log
C-SPOC 命令的活动。
8.1.3.7 /tmp/emuhacmp.out
事件模拟的输出，也可用环境变量 EMUL OUTPUT 指定别的文件。
8.1.3.8 HACMP/ES 用到的三个 log
/var/ha/log/grpsvcs.<filename> grpsvcs 守护进程的活动
/var/ha/log/topsvcs.<filename> topsvcs 守护进程的活动
/var/ha/log/grpglsm grpglsm 守护进程的活动

8.2 在节点或客户机启动和停止 HACMP

8.2.1 HACMP 守护进程

8.2.1.1 clstrmgr 所有节点
管理心跳，监视节点，接口，执行相应的 scripts。
8.2.1.2 clsmuxpd 要求 snmpd，所有节点
维护群集的状态信息。
8.2.1.3 cllockd 可选
用于 Concurrent access 配置，但不是必要的。
8.2.1.4 clinfo 任意节点或客户机均可
是 clstat 所必需的，它会执行/usr/sbin/cluster/etc/clinfo.rc

With RSCT CRIS(system Cluster Technology)on HACMP/ES V4.3
8.2.1.5 Cluster Topology Services daemon (topsvcsd)
所有 HACMP/ES 节点要运行，监视网卡状态。
8.2.1.6 Cluster Event Management daemon（emsvcsd）
一个域中所有节点，监视系统资源。
8.2.1.7 Cluster Group Services daemon(grpsvcsd)
所有 HACMP/ES 节点要运行，管理所有群集操作要求的 distributed 协议。
8.2.1.8 Cluster Globalized Server Daemon daemon（grpglsmd）
是 grpsvcsd 的客户进程，让 switch 卡可被各节点访问。

8.2.2 在节点启动 HA

用 SMIT Start cluster Services 建立和执行/usr/sbin/cluster/etc/rc.cluster 脚本，该脚本设置好环境变量，调用/usr/sbin/cluster/utilities/clstart 脚本，该脚本通过 startsrc 命令启各 HA 守护进程，子系统、组。

也可用 C-SPOC 工具，/usr/sbin/cluster/utilities/cl_rc.cluster。在某个节点或客户端串行启动各节点的 HA。（调用各节点 clstart cluster）
8.2.2.1 自动启 HA
通过 SMIT Start Cluster Services 或 rc.cluster –R 可在 inittab 中加入
hacmp:2:wait:/usr/sbin/cluster/etc/rc.cluster-boot > /dev/console 2 >&1
注意：不利于节点出问题后，先解决再加入 HA 的处理。
8.2.2.2 IPAT enable
rc.cluster 调用/etc/rc.net 做网络配置。

8.2.3 在节点停止 HA

通过 SMIT 建立和执行/usr/sbin/cluster/etc/clstop 脚本，该脚本通过 stopsrc 来停止 HA 的守护进程。

或用 C-SPOC 工具，cl_clstop 调用各节点 clstop 串行停止 HA。

8.2.3.1 何时停止

软硬件维护，重配置时。

8.2.3.2 停止类型

Graceful：HA 停止守护进程，释放资源，别的节点不接管。

Graceful with Takeover：同上，但别的节点接管资源。

Forced :只停守护进程,不释放资源 :如 service IP,应用不停,fs 不会 unmout,sharevg 不会 varyoff，别的节点也不会来接管。HACMP/ES 不支持这种类型，只有传统 HA 有支持。

8.2.3.3 非正常守护进程中止

如果 SRC 检测到 daemon 异常中止，将用/usr/sbin/cluster/utilities/clexit.rc 停止系统，避免不可预期问题，不能用 kill-9 停止 clstrmgr，会引起系统停止。

8.2.4 在客户机启动停止 HA 服务(clinfo)

可用 rc.cluster 或 startsrc –s clinfo 启动

用 stopsrc –s clinfo 停止

8.2.4.1 clhosts 文件

客户端运行 clinfo，要先配好/usr/sbin/cluster/etc/clhosts 要包含所有节点的主机名或 IP(boot/service/standby)，Clinfo 将和这些节点的 clsmuxpd 通讯。

8.3 更换故障部件

8.3.1 节点

注意：内存足够，单 CPU 应用在 SMP 上可能更慢，槽位够吗，卡的位置，应用 lic 和 CPU ID 有关吗？

8.3.2 卡

先查线、hub 的问题，SSA 线可在线更换，SCSI 卡注意堵头和 ID。

8.3.3 硬盘

略

8.4 共享 LVM 内容的修改

建 VG，扩展、减少、改变、删除 VG，Importing，镜像，不镜像，同步 VG；

建 LV，扩展、减少、改变、拷贝、删除 LV 或 LV 的拷贝；
建 FS，扩展、改变、删除 FS；
要在其它节点做 exportvg 和 importvg 的操作，来同步 VG DA 和 ODM 有 4 种途径

8.4.1 手工改

比如删 VG 必须手工在每个节点做(exportvg)。
步骤：
停 HA（或和 SVG 有关的应用），改动 LVM
umount fs
vargoffvg
在另外节点，
exportvg
importvg from one of its disks，注意 Vgmajor number
改 VG→not auto varyon
mount fs
test fs
umount fs
varyoffvg
在其它节点重复，重启 HA

8.4.2 Lazy 修改

HA 可以自动完成 export 和 import VG 的工作，当 HA varyon 一个 VG 时，会比较 VGDA 中的时间片和节点/usr/sbin/cluster/etc/vg 文件中记录的时间片是否一致，如不一致，HA 会在激活 VG 之前先做 export 和 import，这会使接管动作多花几分钟，称之为 Lazy Update。
注意：
(1)如 SVG 中第一个盘被换过，Lazy Update 的 import 会出错，因为 ODM 中该盘的 PVID 和实际对不上，importvg 是自动从第一个盘中取 VGDA。
(2)SCSI RAID 支持多个 LUN 的也有类似问题，因每个 LUN 对节点而言，就是一块盘。

8.4.3 C-SPOC

可以管理用户和组，维护 SLVM，控制 HA。
可避免手工在各节点执行命令引起的人为问题。在一个节点执行的命令，同时也在其它节点执行，比如建一个用户，用 C-SPOC 命令可同时在其它节点也加上。
C-SPOC 可通过 SMIT→HA→Cluste System Management 调用，smit cl_lvm；smit cl_conlvm 共同访问 vg。

8.4.4 Task Guide

8.4.4.1 要求

图形终端，已配好的 HA。

8.4.4.2 启动

命令行/usr/sbin/cluster/tguides/bin/cl_ccvg 或 smit hacmp→cluster sys Management → cluster LVM→Taskguide for Creating a Shared VG

8.5 修改群集资源组

资源为便于管理配置，被分成组，可控制资源组的节点又有优先级，这些定义的修改包括加、删、改资源组。

8.5.1 加删改资源组

不须要停止 HA，就可以做加删改的操作。

8.5.2 同步群集资源

smit 菜单 synchronize Cluster Resource

客户化的信息如 pre、post、notify、recovery 事件脚本的名字会被同步，因此各节点应有这些脚本，内容可以不同。

如本地节点 cluster manager 是不活动的，只 copy ODM 在 DCD(Default Configuration Dir)中的内容到远程节点。

如本地节点 clmanagr 是活动的，动态重配置事件开始，不仅 DCD、ACD（Active）中的内容也被 Copy，且守护进程会刷新，在 log 文件中，有 reconfig resource-release、reconfig resource-acquire、reconfig resource-complete 三个事件被记录。

如本地节点 Clmanager 未激活，但其它节点有激活，同步失败。

8.5.3 DARE 资源迁移

Dynamic Reconfiguration Resource Migration utility 可以方便地控制一个资源组从一个节点到另一个节点，或禁止要求一个资源组，而不影响节点，HA 的其它正常活动。

8.5.3.1 资源迁移的类型

Sticky 资源迁移，让一个资源组习惯性地属于某一节点，除非节点 Down，高优先级节点加入，资源组不会属于该节点。一般用于 cascading 类配置，不用于 rotating 配置。

Non-Sticky 资源迁移：让资源组临时性属于某一节点，适用于 rotating 配置，或 Cascading Inactive Takeover=false 时用。

8.5.3.2 Locations 位置

可以用节点名填入 Location 域，说明 Sticky 和 Non-Sticky 的归属节点。也可用关键字 Default 和 stop(DARE 工具支持)来填 Location。

Default：清除以前的 stickiness 设置，让资源组按默认的 Cascading 和 rotating 行为习惯活动。如空着 Location 域，就是 default。

当一个 Cascading 资源组，INACTIVE_TAKEOVER→fales，而主节点没开机，该资源组不会被激活，当用 Default 迁移可让活动的最高优先级的节点拿到该资源组。

Stop :让一个资源组 stop，且不会被任何失败和重加入事件导致活动(要带 sticky)。

8.5.3.3 cldare 命令

cldare：同时执行所有的迁移请求，有一个不行，则所有迁移都不会发生。

cldare -m <resgroup name>：[Location |default stop nodename][sticky]

-M 代表迁移，nodename 是 up 的且在 resgroup 的 nodelist 中，可重复多个资源组。

一般来讲 stop 和 sticky 同用，避免 reboot 时启动资源。

但 stop 和 non-sticky 用于停下 INACTIVE-TAKEOVER→false 的 cascading 资源组。

8.5.3.4 clfindres

可查找一个资源组的状态。

8.5.3.5 清除 sticky 标志(记在 HA OIM 中)

当 HA 是 UP 时，可用 Default 关键字，或 non-sticky 迁移

当 HA 是 Down 时，用 cldare –v(不校验) –M <resgrp name>：stop

8.6 安装 PTF

apply AIX PTF 可以让 B 机 TakeoverA 机应用，先装 A，再将 B，关键应用不会停止。

1、smit clstop with takeover
2、Apply PTF
3、/usr/sbin/cluster/diay/clverify 和其它 Test
4、reboot node，如果涉及 cluster.base.client.lib 而应用到 Cluster Lock Manager 或 Clinfo API function，要重新 relink 应用。
5、Restart HA:smit clstart 接管回资源
6、在 B 机重复 1-5

注意：(1)群集节点要求保持 AIX maintenance Level 一致，但 update 时例外；
(2)群集节点要求 HA maintenance level 一致，且 update 时要求 Down 掉 HA

8.7 备份策略

8.7.1 Split-Mirror 备份

由于备份要求数据保持 offline，不能应用一边要修改，一边做备份，而有些应用又不允许长时间 offline 做备份，可用下法：

1、lsvg –l VGNAME 找到 lv 的名字和 fs 相关
2、Stop 应用 umount fs 注意:至少有一次镜像的 lv 才行
3、splitlvcopy –y newlv fslv 将原 lv 分成 2 个 lv
4、mount fs start 应用

5、用 newlv 建 fs(如何建?在/etc/file systems 中加一个 fs→newlv),并 mount it read-only

6、备份 newlv 的 fs

7、删除 newlv

8、用 mk lvcopy 加 fslv 一个镜像，并同步恢复原状。

8.7.2 用事件安排定时备份

利用 crontab 做备份，当发生接管时，利用 node_down_remote 事件的 post-event 脚本修改 crontab,将接管资源的备份也包含进去,用 node_up_remote 事件脚本改回 crontab，备份脚本支持 split 备份。

8.8 用户管理

用 C-spoc 创建的用户可保证有相同的 uid 和 gid(在整个群集)

8.8.1 列出所有群集节点用户

(1)lsuser (RSH)在每个节点执行；

(2)cl_lsuser(C-SPOC)会自动执行 lsuser 在每个节点，如用了 NIS，有些 user 信息不会显示；C-SPOC 也有 SMIT 入口

8.8.2 建新用户

(1)mkuser 相关文件 /usr/lib/security/mkuser.default

(2)cl-mkuser

8.8.3 改用户

(1)chuser

(2)cl_chuser 当采用 NIS 时，不要用 cl_chuser

8.8.4 删用户

用户的帐号删除了，但用户的 home 目录保留，只有 root 和同组用户可访问。

(1)rmuser

(2)cl_rmuser

8.8.5 组管理

和用户管理类似。

8.8.6 C-SPOC log

/tmp/cspoc.log 只记录在本机运行 cspoc 命令情况。

第 9 章 和 SP 有关

9.1 HACWS

重要作用：壹台 CWS 故障 SP 仍可工作；
CWS 做维护，SP 不受影响，方便备份。

9.1.1 硬件

(1)cws spdata fs 要在 shared disk 上，两台 cws 都可访问到。
(2)SP 有多个 frame，应给 cws 配 8-port 卡，已保证串口够用。
(3)cws 到 frame 的 supervisor 卡是串行直连线，主备 cws 连同一 frame 的 tty 口应相同，如：都用 tty3 来和 frame A。(节点 supervisor 卡连到菊花链串行线的哪个接口决定其是 No1 还是 No3，而该菊花链只连到电源的 J8(奇数节点)或 J7(偶节点和 switch)，用菊花链上标有 PDU 的接口。菊花链不连 cws 和 frame 的 supervisor)
frameA 上要有 Y-cable（串行口）
(4)cws 只要一块网卡/每 cws

9.1.2 软件

AIX、PSSP、HA 都要用相同版本查手册，如 A1X4.3.2 PSSP3.1HA4.3

9.1.3 配置备份 cws

主 CWS 象无 HACWS 时一样配置。
备 CWS 的 kerberos 配置 如主 cws 是 authentication server，则备 cws 配成，secondary anthenticaton server。如主 cws 是 authertication client of some other server，则备 cws 也是 client。

9.1.4 安装软件

除 HA 外，还有 ssp.hacws

9.1.5 HACWS 配置

(1)用/usr/sbin/hacws/spw-apps –d 停下 spmgr、splagd、hardmon、sysctld、supfilesrv、sp_sonfigd 等守护进程

(2)配心跳线，网卡要用 boot 地址启动，/spdata fs→共享盘
(3)定义 cluster ID，NAME，node，卡应有 boot 和 service 地址，同步
(4)ssp.hacws 带了一些 script，可做为 HA 的应用服务器
建议资源组定义如下：
资源组名：hacws_group1
Node Relationship：rotating
participating Node Name：[主节点名，备节点名]
Service IP lable：至少主 cws 主机名
Filesystem：/spdata
VG：包含/spdata 的 vg
Appserver:

9.1.6 设置和测试 HACWS

先让主备 CWS 可按其 hostname 寻址，再在主 CWS 上运行：
/usr/sbin/hacws/install_hacws -p 主 cws 名 -b 备 cws 名 –s 设 cws 节点名
定义事件脚本：/usr/sbin/hacws/spcw_addevents
校验：/usr/sbin/hacws/hacws_verify
检查线缆：/usr/sbin/hacws/spcw_verify_cabling
依次重启主备 cws，依次启动 HA，smit clstart
检查 hacmp.out
grep “SPCW_APPS COMPLETE” /tmp/hacmp.out

9.2 Kerberos 安全

基于 DES 算法加密
Identification 实体告诉别人自己是谁
Authentication 认证某个实体
Authorization 检查一个认证的实体是否有授权
Client 和 Server 的通信分三步
Get a ticket-granting ticker(有生命期) 用到 pwd，避免以后再用到 pwd
Get a service ticket
Get the work done on the service provider

9.2.1 在 HA4.3 上配置 kerberos

HA4.3 可用 cl_setup_kerberos 定义 kerberos principals 给群集中所有的 IP labels，这样远程的 kerberized 命令可以工作。

SP 上用 setup_authent 做 SP 相关的 kerberos 设置，IP labels 来自于 SDR，因 SDR 不允许一个接口多个 Iplabels，对有 IPAT 的情况，HA 的 kerberos 设置要做 2 次，或用 setup_authent 直接设，或用 setup-server 间接调用 setup_authent。

手工也可设，用 kadmin。

必须的 principal 有 rcmd，godm2 个，格式 godm.had1_stby@IBM.com。

将 principals 加入 kerberos DB 后，还要将它们从 DB 中解出，加到各节点的 /etc/krb_srvtab 文件中。现在可以在 root 的。Klogin 和/etc/krb.realms 中用新的 principals。

9.3 VSDs-RVSDs

9.3.1 Virtual Shared Disk

在 pssp 的 ssp.csd.vsd 文件集中。

VSD 只支持裸 lv，不支持 fs。VSD 没有上锁机制，要通过 Oracle Parallel Server 来完成。一个节点可以自身没有 HD，没有 VSD 进程，但仍可成为其它节点的 VSD 的 client。VSD 可以使用各种 IP 网络，出于性能考虑，建议用 Switch 网络。VSD 对访问本地 HD 的影响很小。

VSD 结构图

三种数据来源

VSD 状态图

buddy buffer

略

9.3.2 Recoverable Virtual Shared Disk

物理上提供对 disk 的 2 个以上节点的连接。

RVSD 功能图

RVSD 相关守护进程功能

略

9.4 SP switch 做 HACMP 网络

早期 HPS 或 HiPS 又称为 TB2switch HACMP/ES V4.3 PSSP3.1 以后不再支持。有缺点，一个节点的变化影响整个网络其它节点。

SP switch 又称为 TB3 switch，克服上述缺点。

9.4.1 switch 基本内容

switch 网络是点到点网络，配成 private 网络。要支持 IPAT，就要用 ARP。用 IPAT，还要用 ifconfig alias 将 service 和 boot 地址赋于 CSS0 卡。SDR 中有 switch 卡的 base IP 地址，不要用于 HA 网络，会引起 PSSP 混乱。所有 SP switch 网络地址的掩码和 CSS0 base IP 地址的掩码一致。

9.4.2 Eprimary 管理

Eprimary 节点管理 SPS 的交通路由事件，有 backup 节点。

HACMP/ES4.3 以前(不含 4.3)有 primary 节点管理功能，如果升级到 V4.3，switch 也升为 spswitch，这些管理已经内置，HA 不再管理。

用 odmget –q'name=EPRIMARY' HACMPSP2 看是否有管理 Eprimary。

升级前用/usr/es/sbin/cluster/events/utils/cl_HPS_Eprimary unmanage 将其置为不管理。

9.4.3 Switch 失败

Switch 卡是单点失败，应认为节点失败。参考 2.6.2.1 的定义，关掉节点，让接管发生。如该节点是 Eprimary 节点，且是 Spswitch，SP 软件会选一个新的 Eprimary 节点，不须要 HA 软件的介入。

第 10 章 HA Classic vs HA/ES vs HANFS

10.1 HACMP for AIX classic

群集管理通过 NIM(网络接口模块)管理节点卡和网络状态。

10.2 HACMP for AIX /Enhanced Scalability

HACMP/ES 不再通过 NIM，而是通过一个原来为 SP 开发的技术，现在从 PSSP 中分离出来的 RSCT 技术来实现。该包原来叫 ssp.ha 文件集，现在集成在 HACMP/ES 中，RISC System Cluster Techonlogy。

RSCT 包含以下内容（为什么用 HACMP/ES 的原因？）：

(1)Event Manager；

(2)Group Service；

(3)Topology Service；

(4)安全性上，可选择标准的，用/.rhosts 来满足同步要求；增强的，用 kerberos 来认证同步时的远程命令。

10.3 HA for Network File System

只支持 cascading 资源组，HANFS 只支持 2 个节点，基于 Classic HA 结构，在 NFS 功能上能处理双重请求和保证单机失败接管时锁状态的正确性。HANFS 和 HACMP Classic 不能混用于一个 cluster 中，但可在同一物理网络上。

10.4 相同和不同

结构上是相同的，都需要 Cluster Manager 控制节点，追踪状态，激发事件，不同之处在于追踪状态的技术不同。HACMP/ES V4.3 可支持 SP 和非 SP 机器，但如用在 SP 上，要求 PSSP3.1，而 PSSP3.1 又要求 Spswitch，当然 SP 也可用传统的 HACMP，这样对 PSSP 就没有要求。

10.5 如何选择

如果是 HPS Switch，只能用不高于 PSSP2.4，因此只能用 HACMP classic 或 HACMP/ES V4.2.2 及以下。无 switch 或新的 spswitch 可选功能更强的 HA。用于 NFS, 则用 HANFS。

`