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。

Chapter 2 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SOCC：很少用了，withdrawn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