AIXCHINA

HACMP for AIX
学习笔记

www.aixchina.com

Aix HACMP

IBM Certification Study Guide Test 167

Chapter 4 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。

Cluster 5 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。