HP WORLD 2004
Solutions and Technology Conference & Expo

HP TruCluster Recovery Techniques

Christian Klein

HP Tru64 Unix Support
Hewlett-Packard

“"Expect the best. Prepare for the worst....”

Zig Ziglar

When good clusters go bad

Things we need to know to enhance our troubleshooting abilities

- The filesystem layout in a TruCluster
- Data in a CNX partition
- Layout of a member specific boot disk
- How TruClusters boot and find cluster_root
- What information needs to be saved to restore our configuration and how we can make this easier on ourselves
- Voting
- How to make our life easier

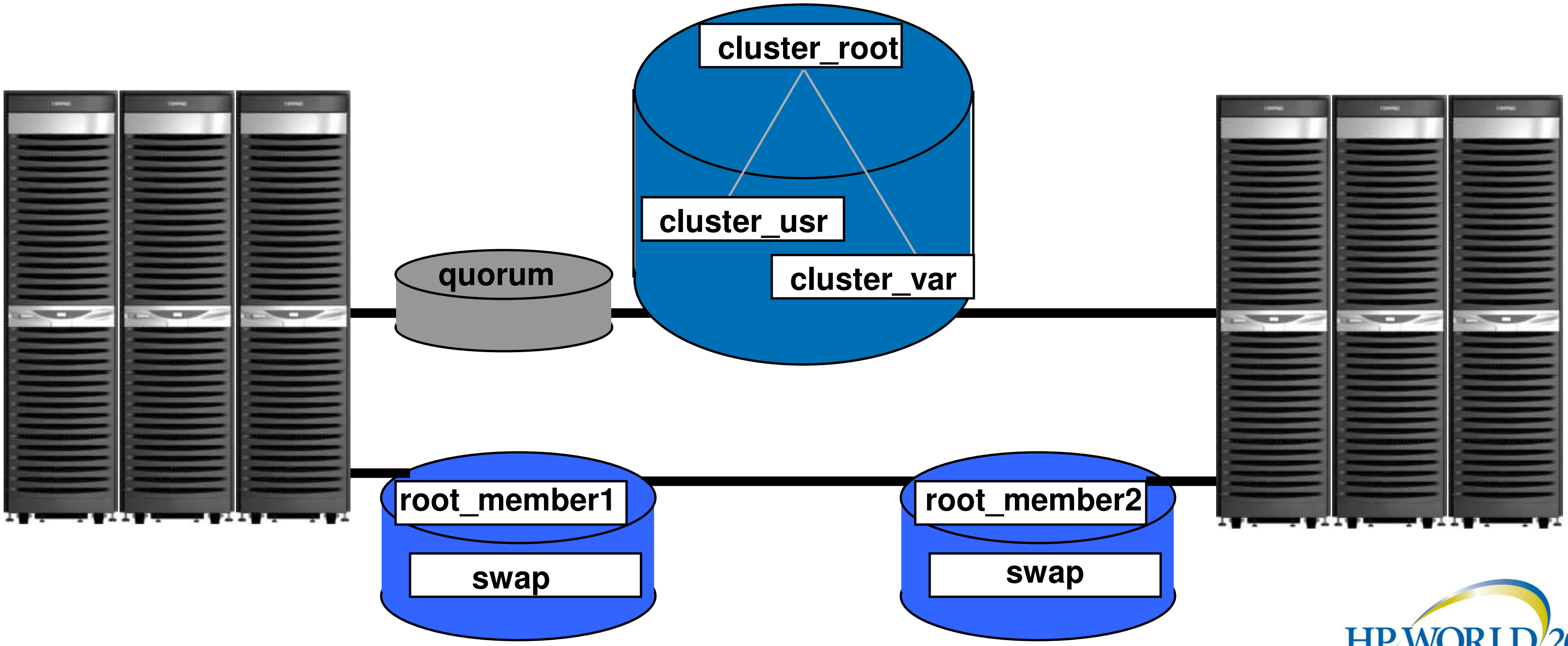
The filesystem layout in a TruCluster

The filesystem layout in a TruCluster

TruCluster V5.x Minimum Disk Configuration

A four disk minimum in a two member TruCluster

(not including the original UNIX/Emergency Repair Disk)



Data in a CNX partition

Data in a CNX partition

- CNX partitions contain the **logical** device name of the device(s) that make(s) up cluster_root

```
oscar# clu_bdmgr -d dsk10
# clu_bdmgr configuration file
# DO NOT EDIT THIS FILE
::TYP:m:CFS:/dev/disk/dsk2a:LSM:47,/dev/disk/dsk103h|priv::
```

```
oscar# clu_bdmgr -d dsk3
# clu_bdmgr configuration file
# DO NOT EDIT THIS FILE
::TYP:q:CFS:/dev/disk/dsk2a:LSM:47,/dev/disk/dsk103h|priv::
```

- m is for member disk
- q is for quorum disk

Layout of a member specific boot disk

Layout of a member specific boot disk

- Each cluster member has its own boot disk
- The “a” partition is AdvFS, the “b” partition is swap, and the “h” partition is cnx (Connection Manager)
- There is a vmunix, sysconfigtab, rc.config, and parts of the hwmgr database on the boot partition (a)

Disklabel of a member boot disk

```
# disklabel -r dsk5
# /dev/rdisk/dsk5c:
type: SCSI
disk: HSG80
label: clu_member1   ←(16 character field that can be used to identify disks)
…
8 partitions:
#            size       offset     fstype  fsize  bsize   cpg # ~Cyl values
  a:       262144            0      AdvFS                     #       0 - 77*
  b:      4355579       262144       swap                     #     77*- 1366*
  c:      4619771            0     unused      0      0       #       0 - 1366*
  d:            0            0     unused      0      0       #       0 - 0
  e:            0            0     unused      0      0       #       0 - 0
  f:            0            0     unused      0      0       #       0 - 0
  g:      2113277       393216     unused      0      0       #    116*- 741*
  h:         2048      4617723        cnx                     #   1366*- 1366*
```

How TruClusters boot and find cluster_root

How TruClusters boot and find cluster_root

- An AlphaServer boots the disk (bootdef_dev or specified)
- Loads the vmunix which recognizes the hardware
- vmunix parses **sysconfigtab**
- The **sysconfigtab** file contains the major and minor numbers of the h partitions on both the member boot disk and the quorum disk

```
clubase:
  cluster_seqdisk_major=19
  cluster_seqdisk_minor=96 ←boot disk
  cluster_qdisk_major=19
  cluster_qdisk_minor=160 ←quorum disk
  cluster_qdisk_votes=1
```

How TruClusters boot and find cluster_root

- vmunix parses the hwmgr database on the boot disk
- vmunix reads the CNX partition on the boot disk.
- The data **inside** the CNX partition is used to find cluster_root (member disk and quorum disk)
- Once cluster_root is found, the /etc/fstab is used to find what filesystems to mount.
- AdvFS filesystems are resolved to actual devices via the /etc/fdmns hierarchy.

Voting

Voting

- **Quorum** is a majority of votes (greater than 50%).
- If you have an even number of voting members, you need a voting quorum disk
- If you have an odd number of voting members, you do not need a quorum disk
- Every member can have one or zero votes so check the output of clu_quorum

Making our life easier

Making our life easier: Quorum Disk

- You can also use the **disklabel** command to look for a quorum disk. All partitions in a quorum disk are unused, except for the h partition, which has fstype cnx.

```
oscar# disklabel -r dsk3
# /dev/rdisk/dsk3c:
type: SCSI
disk: HSV110 (C)COMPA
label: Quorum Disk
flags:

…

8 partitions:
#            size       offset    fstype  fsize  bsize   cpg  # ~Cyl values
  a:       131072            0    unused      0      0        #      0 - 7
  b:       262144       131072    unused      0      0        #      8 - 23
  c:      4194304            0    unused      0      0        #      0 - 255
  d:            0            0    unused      0      0        #      0 - 0
  e:            0            0    unused      0      0        #      0 - 0
  f:            0            0    unused      0      0        #      0 - 0
  g:      1900544       393216    unused      0      0        #     24 - 139
  h:         2048      4192256       cnx                      #    255*- 255
```

Making our life easier: Storage controllers

- Each LUN on an **HSG** or **HSV** controller should have the identifier set
  - HSG80_TOP> set D62 IDENTIFIER=62
  - Use the San appliance to set the OS Identifier for the HSV110

```
judy # hwmgr -v d
 HWID: Device Name          Mfg      Model            Location
 ------------------------------------------------------------------------
…
  108: /dev/disk/dsk5c     DEC      HSG80            IDENTIFIER=57
  140: /dev/disk/dsk7c     COMPAQ   HSV110 (C)COMPAQ IDENTIFIER=31
….
```

Making our lives easier: Disklabels

- The cluster software edits the disklabel’s “label:” field for the Quorum disk and the member boot disks
- One can edit this 16 character field with “**disklabel –e dskX**” and make changes to it even when the disk is in use.
- You can be creative within the 16 character boundary

Making our lives easier: Disklabels

```
skipper# disklabel -r dsk5
# /dev/rdisk/dsk5c:
type: SCSI
disk: HSG80
label: clu_member1     ←(16 character field that can be used to identify disks)
…
8 partitions:
#            size        offset     fstype  fsize  bsize   cpg  # ~Cyl values
  a:       262144             0      AdvFS                      #       0 - 77*
  b:      4355579        262144       swap                      #     77*- 1366*
  c:      4619771             0     unused      0      0        #       0 - 1366*
  d:            0             0     unused      0      0        #       0 - 0
  e:            0             0     unused      0      0        #       0 - 0
  f:            0             0     unused      0      0        #       0 - 0
  g:      2113277        393216     unused      0      0        #    116*- 741*
  h:         2048       4617723        cnx                      #   1366*- 1366*
```

Making our lives easier: Disklabel Example

- On the following cluster_root disk,
  - The diskname is dsk2
  - cluster_root is on the “a” partition
  - cluster_usr is on the “g” partition
  - cluster_var is on the “e” partition

```
calvin# disklabel -r dsk2 | grep label
 label: dsk2_CRa_CUg_CVe
```

Making our life easier: Documenting the Configuration

- Make a spreadsheet with device names, identifiers, World Wide Ids, size in blocks, etc.   (You can use **sys_check**'s storage map or the quick script on the next slide for a start)
- Run "**sys_check –escalate**" periodically. Save off the output somewhere else
  - sys_check –escalate also saves off disklabels in /var/recovery
- Run "**volsave**" after every LSM configuration change no matter how small
  - sys_check –escalate also does a volsave into /var/recovery
- Back up the operating system to local tape (if available) using **vdump**

Example script

```
#!/usr/bin/ksh -p
for h in `hwmgr -v d | grep dsk | sed s/://g | awk '{print $1}'`
do
   echo "========================================================="
   echo " BEGIN the record for HWID # $h "
   echo "========================================================="
   echo " IDENTIFIER and WWID info "
   echo "---------------------------------------------------------"
   hwmgr -v d -id $h
   hwmgr -sh scsi -full -id  $h
   echo "========================================================="
   echo " DISKLABEL INFO "
   echo "---------------------------------------------------------"
   d=`hwmgr -sh scsi -id $h | grep dsk | awk '{print $8}'`
   disklabel -r $d
   echo "========================================================="
   echo " /etc/fdmns info "
   echo "---------------------------------------------------------"
   find /etc/fdmns -name "$d*"
   echo "========================================================="
   echo " END the record for HWID # $h "
done
```

Making our life easier: Backing up the correct data

- The following filesystems should be backed up

```
cluster_root#root on / type advfs (rw)
cluster_usr#usr on /usr type advfs (rw)
cluster_var#var on /var type advfs (rw)
root1_domain#root on /cluster/members/member1/boot_partition type advfs (rw)
root2_domain#root on /cluster/members/member2/boot_partition type advfs (rw)
```

- The member boot partitions are often **forgotten**, perhaps it's because they do not appear in /etc/fstab
- We will **not** be able to boot without them!
- Avoid keeping **user data** in OS filesystems as it will prolong backup and recovery times

Making our life easier: Emergency Repair Disk

- Keep the original UNIX disk around
- This is the disk you ran clu_create from
- It would not hurt to create another copy of this disk on shared storage if it is local to a particular member

Example of Restoring a Cluster to Totally Different Hardware

Restoring a cluster

- Boot the operating system cdrom and get a **UNIX** shell.
- Create the tape device(s)
  - /sbin/dn_setup –install_tape
- You will need disks with the **same** names as the old disks, so get out your records
- Use **dsfmgr –m** (move) and **dsfmgr –e** (exchange) to rename disks appropriately, for example:

```
# dsfmgr -m dsk4 dsk12
dsk4a=>dsk12a  dsk4b=>dsk12b  dsk4c=>dsk12c
dsk4d=>dsk12d  dsk4e=>dsk12e  dsk4f=>dsk12f
dsk4g=>dsk12g  dsk4h=>dsk12h  dsk4a=>dsk12a
dsk4b=>dsk12b  dsk4c=>dsk12c  dsk4d=>dsk12d
dsk4e=>dsk12e  dsk4f=>dsk12f  dsk4g=>dsk12g
dsk4h=>dsk12h
```

Restoring a cluster (continued)

- Label disks as necessary with the **disklabel** or **diskconfig** commands
  - Make sure that the member boot disk has an “a” partition (the default size is fine), “b” partition (that starts at the end of the a partition and goes to the end of the disk minus 2048 blocks), and that the “h” partition starts 2048 blocks from the end of the disk and is exactly 2048 blocks in size
- Remake the cluster_root filesystem
  - `mkfdmn -o /dev/disk/dskNy cluster_root`
  - `mkfset cluster_root root`

Restoring a cluster (continued)

- Mount and restore the new cluster_root

```
– mount cluster_root#root /mnt
– cd /mnt
– vrestore -x
```

- Make the new member1’s boot_partition

```
– mkfdmn -o -r /dev/disk/dskNa root1_domain
– mkfset root1_domain root
```

- Mount and restore the new member1’s boot_partition (note where we are mounting it)

```
– mount root1_domain#root
  /mnt/cluster/members/member0/boot_partition
– cd /mnt/cluster/members/member0/boot_partition
– vrestore -x
```

Restoring a cluster (continued)

- Copy the hwmgr database pieces from /var/etc (a memory filesystem when booted from cd) to cluster_root and the member boot disk

```
# cd /var/etc
# ls
cfginfo           dec_devsw_db.bak  dec_hwc_ldb.bak   dfsc.dat
dccd.bak          dec_hw_db         dec_scsi_db       dfsl.bak
dccd.dat          dec_hw_db.bak     dec_scsi_db.bak   dfsl.dat
dcdd.bak          dec_hwc_cdb       dec_unid_db
dcdd.dat          dec_hwc_cdb.bak   dec_unid_db.bak
dec_devsw_db      dec_hwc_ldb       dfsc.bak

# cp d* /mnt/etc/
# mkdir -p /mnt/var/etc
# cp d* /mnt/var/etc/
```

Restoring a cluster (continued)

- We need to write the **cnx** partition on the member boot disk, but **clu_bdmgr** is not on the OS CD, so we'll have to work around that with **chroot**

```
# chroot /mnt /sbin/sh
# dsfmgr -K            (make device special files for sure)
# dsfmgr -vVF          (make sure everything is in order)
# dsfmgr -vVF          (run it again if they are not)
```

- Now restore the CNX partition with **clu_bdmgr**

```
# clu_bdmgr -h dsk1
/cluster/members/member0/boot_partition/etc/clu_bdmgr.conf

# exit    (to exit the chroot sh)
```

Restoring a cluster (continued)

- We need to find the **minor** number of the h partition of your member boot disk and quorum disk (if you have one)

```
# file /dev/disk/dsk*h
# cd /cluster/members/member0/boot_partition/etc
# vi sysconfigtab and correct the following:
   clubase:
          cluster_seqdisk_major=19  ← always 19
          cluster_seqdisk_minor=96
          cluster_qdisk_major=19   ← always 19
          cluster_qdisk_minor=160
```

Restoring a cluster (continued)

- Please also note that if you have a **LAN interconnect** cluster, you may need to change the cluster interconnect devices in /etc/sysconfigtab. "hwmgr –show name" and "ifconfig –a" are ways to show you the adapters in the system.
- There will **not** be an interconnect related sysconfigtab change for a memory channel cluster
- You may want to change IP addresses for the members and cluster aliases in /mnt/etc/hosts at this time. Else, you should bring up this cluster on a LAN **disconnected** from the same LAN as the cluster the backups came from

Restoring a cluster (continued)

- Boot the machine with **cluster_expected_votes=0**

```
>>> boot dwhatever -fl i
```

- When prompted for the kernel to boot, enter

```
genvmunix clubase:cluster_expected_votes=0
```

- You will boot up to single user mode
- Executing “**/sbin/bcheckrc**” will mount local filesystems
- You should build a customized kernel for this platform with the “**doconfig**” command
- Executing “**/sbin/init 3**” will take you to multiuser

Restoring a cluster (continued)

- At this point you can
  - Restore another member's boot disk
    - clu_bdmgr –c dskX N   (where N is the memberid)
    - mount rootN_domain#root /mnt
    - cd /mnt
    - vrestore -x
  - Delete and add back the additional members with the **clu_delete_member** or **clu_add_member** commands
  - Delete and readd the quorum disk with the **clu_quorum** command
  - Customize the system to tailor it to the different hardware

Resources

- TruCluster 5.1b Documenation

  http://h30097.www3.hp.com/docs/pub_page/cluster51B_list.html

- IT Resource Center

  http://itrc.hp.com/

Questions?

invent

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional