Tips & Tricks for Using LVM Effectively / Intro to VxVM

Renay Gaye
Hewlett-Packard
renay_gaye@hp.com
LVM Concepts
LVM Concepts-Physical Volume

- LVM Managed Disk
- Each PV is assigned a unique PVID

```
pvcreate /dev/rdsk/cntndn
```
LVM Concepts-Volume Group

- One or more physical volumes
- pool of physical extents

mkdir /dev/vgname
mkmar /dev/vgname/group c 64 0xnn0000
vgcreate vgname /dev/dsk/ctnndn ...
Volume Group Map

extent size (-s n)
* max extents/pv (-e n)

max. useable space/disk
LVM Device Files

/block PV device files

/raw PV device files

/raw LV device files

VG device file

Block LV device files

Raw LV device files
/etc/lvmtab

• Non-ascii file
• Records volume group/physical volume associations
• Used by many LVM commands for sanity checking

example:
/dev/vg00 (vgid)
/dev/dsk/c0t6d0 (vgid,pvid)
/dev/vg01 (vgid)
/dev/dsk/c1t2d1 (vgid,pvid)
/dev/dsk/c1t2d2 (vgid,pvid)
Volume Group Activation

• Done automatically when vg is created
• automatically at boot (/etc/lvmrc)
• required in order to access any lvols

vgchange -a y vgname
vgchange -a r vgname
LVM Concepts-Quorum

- More than 1/2 disks in a vg required to activate the volume group
- Can override:
  \[ \text{vgchange} \ -a \ y \ -q \ n \ \text{vg01} \]
- Booting without quorum:
  \[ \text{ISL> hpux} \ -lq \]
LVM Concepts-Logical Volume

```
lvcreate -L n  vgname
lvcreate -L n -C y vgname
lvcreate -l n vgname
lvcreate -L n -n name vgname
```
Placing logical volumes

lvcreate -n datalv vgdata
lvextend -L 500 /dev/vgdata/datalv
/dev/dsk/c4t2d0
Multi-pathing solutions
Multipathing Solutions with LVM

- **LVM pvlinks**
  - active/passive
  - no additional software required

- **Auto path**
  - active/active, load balancing
  - supported for XP and VA arrays
  - additional cost product

- **Powerpath**
  - active/active, load balancing
  - supported for EMC arrays
  - additional cost product
PV links

```
pvcreate /dev/rdsk/c5t5d0
pvcreate /dev/rdsk/c5t5d1
mkdir /dev/vg01
mknod /dev/vg01/group c 64 0x010000
vgcreate vg01
```

```
/dev/dsk/c5t5d0  /dev/dsk/c7t5d0
/dev/dsk/c7t5d1  /dev/dsk/c5t5d1
```

```
c5t5d0  (p)   c7t5d0  (a)
c7t5d1  (p)   c5t5d1  (a)
```
PV links-switching the order

• Temporary Change
  `pvchange -s /dev/dsk/c7t5d0`

• Permanent Change
  `vgreduce vg01 /dev/dsk/c5t5d0`
  `vgextend vg01 /dev/dsk/c5t5d0`
Auto Path-Dynamic load balancing

- Load balancing functionality
  - Supports up to 8 paths from a server to an end device
  - Provides dynamic load balancing across all paths to an end device
  - Choose from 4 load balancing policies, including “no load balancing”
  - Supports the XP and VA disk arrays
  - Load balancing supported in clustered environments
**Auto Path driver**

- A pseudo driver
- The Auto Path driver is layered between the LVM (Logical Volume Manager) and the SCSI device driver
- The driver provides the Command Line Interface
Moving Data
Moving Disks

Three Step Process

1. Remove definition of volume group
2. Move disk(s)
3. Add definition of volume group

Two commands
   vgexport(1m)
   vgimport(1m)
Volume Group Definition — Review

/etc/lvmtab
/dev/vg01
/dev/dsk/c0t5d0

name of Logical Volumes
/dev/vg01
  group 64 0x010000
  lvol1 64 0x010001
  rlvoll 64 0x010001
  databaseslv 64 0x010002
  rdatabasevlv 64 0x010002

VGRA + PVRA
  - PV ID
  - VG ID
  PE to LE map
Exporting a Volume Group

Syntax:
vgexport [-p][-v][-m file]VG
- p    Preview actions only
- v    verbose
- m    used to specify a map file for logical volume names

- Removes volume group definition from the system completely by updating /etc/lvmtab and kernel memory.
- The volume group must first be deactivated with vgchange(1m).

Example:
    vgchange -a n /dev/vg01
    vgexport  -v  -m  /etc/lvmconf/vg01.map  /dev/vg01
Importing a Volume Group

**Syntax:**

```bash
vgimport [-p][-v][-m file] VG PV [PV...]
```

- **-p** Preview actions only
- **-v** verbose
- **-m** used to specify a map file for logical volume names

**Example:**

```bash
mkdir /dev/vg01
mknod /dev/vg01/group c 64 0x010000
vgimport -v -m /etc/lvmconf/vg01.map /dev/vg01 /dev/dsk/c0t1d0
vgchange -a y /dev/vg01
vgcfgbackup vg01
```
Moving LVM Data

**Syntax:**

```
pvmove [-n lv] from_PV [to_PV]
```

**Example:**

```
pvmove -n /dev/vg01/lvol1 /dev/dsk/c0t5d0 /dev/dsk/c0t4d0
```
Renaming LVM objects
Renaming Logical Volumes

• lvol names are not stored in the LVM maps or in /etc/lvmtab
• Simply rename device files and update /etc/fstab if necessary
Renaming Volume Groups

• Volume group name is kept in /etc/lvmtab and is used as the directory name to anchor the device files for the group

• Use vgexport and vgimport to rename volume group
LVM Mirroring
Mirrored Volumes

In a mirrored logical volume ...

- Each logical volume consists of one or more Logical Extents (LEs).
- Each LE maps to two or three Physical Extents (PEs) on disk.
- The logical volume remains accessible even when one of the disks is unavailable.
Extending and Reducing Mirrors

Create a new, mirrored logical volume:

```
# lvcreate -m 1 -L 16 -n myfs1 vg01
```

Mirror an existing logical volume:

```
# lvextend -m 1 /dev/vg01/myfs1
```

Mirror an existing logical volume to a specific disk:

```
# lvextend -m 1 /dev/vg01/myfs1 /dev/dsk/c0t3d0
```

Add a second mirror:

```
# lvextend -m 2 /dev/vg01/myfs1
```

Remove a logical volume's mirrors:

```
# lvreduce -m 0 /dev/vg01/myfs1 /dev/dsk/c0t3d0
```

Check a mirrored logical volume's status:

```
# lvdisplay -v /dev/vg01/myfs1
```
## Mirrored I/O Scheduling

<table>
<thead>
<tr>
<th></th>
<th>Parallel (-d p)</th>
<th>Sequential (-d s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td>Access PV with lowest outstanding I/Os</td>
<td>Read in PV order</td>
</tr>
<tr>
<td><strong>Write</strong></td>
<td>Schedule writes simultaneously to all PVs</td>
<td>Schedule writes in PV order</td>
</tr>
</tbody>
</table>
MWC/MCR

- Writes are recorded in MWC in memory
- MCR record written to disk when a write is done to a logical track group not already recorded
- After a crash only "dirty" LTGs need be resynced
## Mirror Consistency Recovery Options

<table>
<thead>
<tr>
<th></th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
</table>
| 1 | **MWC/MCR**
   | -M y                                             | Fast recovery on a system crash                   |
|   |                                                | Runtime overhead to write MCR record             |
| 2 | **NOMWC**
   | -M n -c y                                        | No extra overhead at runtime                      |
|   |                                                | Slow recovery on a system crash                   |
| 3 | **NONE**
   | -M n -c n                                        | Application can do necessary recovery            |
|   |                                                | No system managed recovery                       |
LVM Boot Disks
LVM Boot volume structure

- LIF Directory
  - PVRA
  - BDRA
  - LIF
  - VGRA
  - PE

- Hardware path/device file of disks in volume group
- boot, root, dump, swap lvols
- Maintained with the lvlnboot command
Mirroring the LVM Boot Disk

• Initialize disk for LVM. Leave room for BDRA/LIF
  
  \texttt{pvcreate -B /dev/rdsk/cntndn}

• Add disk to root VG
  
  \texttt{vgextend vg00 /dev/rdsk/cntndn}

• Install boot files in LIF area
  
  \texttt{mkboot /dev/rdsk/cntndn}

• Change the auto file on both the primary and alternate boot disk
  
  \texttt{mkboot -a "hpux -lq" /dev/rdsk/cntndn}

• Mirror each of the lvols in the root vg
  
  \texttt{lvextend -m 1 /dev/vg00/lvoln /dev/dsk/cntndn}

• Add the mirror disk definition to /stand/bootconf
Booting When BDRA is damaged

- Boot system into maintenance mode
  
  ISL> hpux -lm

- Activate vg00
  
  vgchange -a y vg00

- Use lvlnboot to examine/repair BDRA
  
  lvlnboot -v
  lvlnboot -b /dev/vg00/lvol1
  lvlnboot -r /dev/vg00/lvol3
  lvlnboot -s /dev/vg00/lvol2
  lvlnbtoo -d /dev/vg00/lvol2

- Reboot the system
  
  reboot
Moving the Boot Disk

- Problem: /etc/lvmtab contains the old device files for the boot disk
  - Solution: boot into maintenance mode, export and reimport volume group

- Problem: The BDRA and Label files contain the old device information
  - Solution: use lvlnboot to fix
Cookbook for moving root disk

Boot from new device. Reply Y to interact with IPL

ISL> hpux -lm
# vgexport -v -m vg00.map vg00
# mkdir /dev/vg00
# mknod /dev/vg00/group c 64 0x000000
# vgimport -v -m vg00.map vg00 /dev/dsk/new_device_file
# vgchange -a y vg00
# vgcfgbackup vg00
# lvlnboot -R
# lvlnboot -v
Recovering Corrupt LVM Information
Valuable Data?

- The various LVM data structures are crucial to the continued availability of logical volumes

```bash
vgdisplay -v
```

Sanity check

```
/dev/vg01
/dev/dsk/c0t5d0
```

/etc/lvmtab

Kernel memory

Active VGs

PVRA/
VGRA

Cross reference

InterWorks 2002
THE HP TECHNICAL TRAINING CONFERENCE
Backing up LVM Structures

- Done automatically by commands that modify the configuration

Syntax:

```
vgcfgbackup [-f file] VG
```
Recovering LVM Structures

- Only restores the LVM structures. Contents of logical volumes must be restored separately.

Syntax:

```
vgcfgrestore [-n VG|-f file][-o PV]PV
```

```
vgcfgrestore -l -n VG
```

/etc/lvmconf

vg01.conf  vg00.conf
LVM Control File

- Not ascii data
  - can use strings (1) to read the ascii part
- Used primarily at boot up but also used to sanity check commands
- Can be rebuilt if lost

- All VG IDs the same – all disks in same vg
Recover or Repair
/etc/lvmtab

Syntax:

vgscan [-v] [-p]

Semi-intelligent look at every disk to classify it
• imports those that are known to be on this system
• recommends the import of "new" disks
• rebuilds /etc/lvmtab

Ways to confuse vgscan
• redundant group files
• "old" LV M structures
LVM Performance Tips
Use PVGs to Offload Busy Controllers
Use Striping to Offload Busy Drives

Without Striping

With Striping
LVM Distributed Allocation

Logical Volume

Volume Group vg01

lvcreate -l 12 -s g -D y /dev/vg01
Mirroring and Striping

Logical Volume

LE0
LE1
LE2
LE3
LE4
LE5
LE6
LE7
LE8
LE9
LE10
LE11

Volume Group vg01

PVG1

PVG2

PV1
PV2
PV3
PV4

PV1
PV2
PV3
PV4

lvcreate -l 12 -m 1 -s g -D y /dev/vg01
LVM in an MC/ServiceGuard Environment
MC/ServiceGuard Configuration

- System A
- Primary
- Mirror
- System B

Volume Group
Step 1: Configure Volume Group on First System
Step 2: Import Volume Group to Second System
VG Export and Import – s Option

```
vgexport -p -s -m <mapfilename> <vg_name>
create mapfile without removing the VG,
and save the VGID for the vgimport command
```

```
vgimport -s -m <mapfilename> <vg_name>
scan for disks that have the same VG ID as in the mapfile
```
LVM Definitions-Both Nodes

Node 1

/dev/vg01

- group 64 0x010000
- lvoll 64 0x010001
- rlvoll 64 0x010001
- database lv 64 0x010002
- rdatabase lv 64 0x010002

/etc/lvmtab

The nodes contain the following:

- /dev/vg01
- /dev/dsk/c5t5d0
- /dev/dsk/c7t5d0
- /dev/dsk/c7t5d1
- /dev/dsk/c5t5d1

Node 2

/dev/vg01

- group 64 0x010000
- lvoll 64 0x010001
- rlvoll 64 0x010001
- database lv 64 0x010002
- rdatabase lv 64 0x010002

/etc/lvmtab

The nodes contain the following:

- /dev/vg01
- /dev/dsk/c5t5d0
- /dev/dsk/c7t5d0
- /dev/dsk/c7t5d1
- /dev/dsk/c5t5d1
Cluster Volume Group

vgchange -c y vg01
Volume Group Exclusive Activation

Q. Do you have the volume group activated?
A. No I do not

```
vgchange -a e vg01
```
Marking Volume Groups as MC/ServiceGuard Volume Groups

Marking Volume Group for MC/ServiceGuard

Marking Volume Group as non-MC/ServiceGuard

Standard Volume Group Activation

Exclusive Volume Group Activation

```
vgchange -c y VGName
vgchange -c n VGName
vgchange -a y VGName
vgchange -a e VGName
```
Disk Array Data Replication Issues
Problem: Duplicate vgids

- vgsec(svols)
- vgprim(pvols)

Disk Array

- vgid-12345 c1t3d1
- vgid-12345 c1t3d2
- vgid-12345 c2t6d1
- vgid-12345 c2t6d2
Solution-vgchgid

vgchgid /dev/rdsk/c2t6d1 
/dev/rdsk/c2t6d2
mkdir /dev/vgsec
mknod /dev/vgsec/group c 64 0x020000
vgexport -pm mapfile vgprim
vgimport -m mapfile vgsec 
/dev/dsk/c2t6d1 
/dev/dsk/c2t6d2
vgchange -a y vgsec
vgcfgbackup vgsec
Problem after a reverse sync

/etc/lvmtab
/dev/vgpriv (vgid:12345)
/dev/dsk/c1t3d1
/dev/dsk/c1t3d2

vgid-56789 c1t3d1
vgid-56789 c1t3d2

vgprism(pvols)

vgid-56789 c2t6d1
vgid-56789 c2t6d2

vgsec(svols)

Disk Array
Solution

- export volume group
- reimport volume group

/etc/lvmtab

/dev/vgpriv (vgid: 56789)
/dev/dsk/c1t3d1
/dev/dsk/c1t3d2
VxVM Concepts
VxVM Intro

- Another Volume Manager Supported on HP-UX platform
- Can co-exist with LVM
- Allows flexible disk space management
- Includes vmsa gui as well as cli
- Available from HP beginning with HP-UX 11i.
  Available from Veritas for other releases
- No rootability until 11i Release 1.5 (11.20)
- Lite & Full versions
VxVM Objects - Disks

Physical Disk:
Disk initialized for VxVM use

cntndn or enclosure based name

Free disk pool

or

Disk Group - dgname

VMdisks
VxVM Concepts-Disk Group

• Similar to an LVM Volume Group
• One or more VMdisks
• Pool of space that can then be "partitioned" into volumes
• Entire group can be moved between systems using import/deport

*dgname*

dgname01  dgname02  dgname03
VxVM Concepts-Disk Group "activation"

- There is no /etc/lvmtab file
- hostid in private region is compared to hostid on system
- autoimport flag indicates group should be imported at start up
- deport removes hostid from private region but leaves disks in tact

```
/etc/vx/volboot
host1
```

```
mydg01  mydg02  mydg03
host1   host1   host1
```
VxVM Concepts-Configuration Database

- Describes all objects contained in the disk group
- Replicated on multiple disks in the disk group (4 by default)
VxVM rootdg

- Required for VxVM to run
- Does not necessarily contain the root file system
  - 11i: will not include the root file system
  - 11i 1.5: will include the root file system
  - future releases will have choice
- rootdg cannot be moved between systems
- discourage including customer data
VxVM Concepts-Volume

- Similar to an LVM logical volume
- Analogous to a "partition"
- Can contain a file system, a swap area or raw data
- Represented by device files:
  
  /dev/dsk/dgname/volname    /dev/rdsk/dgname/volname
VxVM Concepts-plex

- Data Plex: A copy of the data
- Non-mirrored volumes will have one data plex, mirrored volumes will have multiple data plexes
- A plex can span disks
- Log plex: Mirrored and raid5 volumes can have a log plex
VxVM Concepts-subdisk

- Contiguous space on a single vmdisk
- entire subdisk belongs to a single plex
- usually created when volume is created
VxVM Objects - Review

Data

\[ \text{vol01 (volume)} \]

\[ \text{vol01-01 (plex)} \]

\[ \text{vol01-02 (plex)} \]

\[ \text{dname01-01 (vmdisk)} \]

\[ \text{cntndn (physical disk)} \]

\[ \text{dname02-01} \]

\[ \text{dname02} \]

\[ \text{cntndn} \]

\[ \text{dname03-01} \]

\[ \text{dname03} \]

\[ \text{cntndn} \]
Data Plex Layouts

- Concatenated
- Striped
- raid5
Concatenated Plex Examples

- Simple volume
  - One plex
  - One subdisk
- Spanned Volume
  - One plex
  - Two subdisks
- Mirrored Volume
  - Two plexes
  - Two subdisks
Striped Plex Layout
raid5 plex Layout

d1 | d4 | d7 | p4

d2 | d5 | p3 | d10

d3 | p2 | d8 | d11

p1 | d6 | d9 | d12
Mixed volumes

- You can combine plex layouts in a single volume

- By default the striped plex will be preferred over the concatenated plex

- You can not mirror raid5 plex layout

mixed volume
Layered Volumes-Concat-pro

Normal Mirror (spanning multiple disks):

Layered Volume (spanning multiple disks):

sub-volume (layout=mirror)

sub-volume (layout=mirror)

volume with layout=concat
Creating Volumes

• vxassist (cli) or vmsa (gui)
  – you create volume
  – plex, sub-disks automatically created

• bottom up
  – manually create sub-disk(s)
  – specify sub-disk(s) to form plex
  – specify plex(es) to form volume
Dynamic Multi-Pathing

Multiple paths

Controller (SBA) c2xxxx

Controller (SBA) c7xxxx

/dev/vx/dmp/c2t3d0 /dev/rdmp//c2t3d0

/dev/dsk/c2t3d0 /dev/rdsk/c2t3d0

dev/dsk/c7t3d0 /dev/rdsk/c7t3d0

/dev/vx/dsk/userdg/myvol /dev/vx/rdsk/userdg/myvol
VxVM-LVM Comparisons

<table>
<thead>
<tr>
<th>LVM</th>
<th>VxVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Volume</td>
<td>Physical Disk</td>
</tr>
<tr>
<td>Physical Volume in a VG</td>
<td>VMdisk</td>
</tr>
<tr>
<td>Volume Group</td>
<td>Disk Group</td>
</tr>
<tr>
<td>Logical Volume</td>
<td>Volume</td>
</tr>
<tr>
<td>group of contig. extents</td>
<td>subdisk</td>
</tr>
<tr>
<td>single mirror copy</td>
<td>plex</td>
</tr>
<tr>
<td>MWC/MCR</td>
<td>dirty region log</td>
</tr>
<tr>
<td>pvra/vgra</td>
<td>private region</td>
</tr>
<tr>
<td>extents</td>
<td>public region</td>
</tr>
</tbody>
</table>
Information Sources

- H6285S Hands-On with LVM course
- H6487S Hands-On with MC/ServiceGuard course
- H5875S HP-UX System and Network Admin for Experienced UNIX System Administrators
- U1592AAE Intro to VxVM - HP Education Virtual Classroom Seminar
  
  www.hp.com/education

- Various white papers and cookbooks from the Response Center