LVM Online Disk Replacement (LVM OLR)

Abstract

The LVM Online Disk Replacement (OLR) feature provides new methods for replacing or isolating path components or LVM disks within an active volume group:

- Using new commands options, LVM OLR enables the system administrator to follow a simpler procedure for replacing disks in an active volume group. The procedure does not require deactivating the volume group, modifying the volume group configuration or moving any user data.

- LVM OLR can also be easily employed to isolate troublesome paths or disks to allow running diagnostics against them.

In the past, replacing a disk in an active volume group usually required moving any logical volumes from the disk and then removing the disk from the volume group configuration beforehand. Now with the introduction of LVM OLR this is unnecessary. Replacing a disk typically only requires invoking a command to inform LVM to stop using the disk, and then performing the steps necessary to replace it.

There was formerly no command to prevent LVM from accessing any device in an active volume group. Now LVM OLR can be used to force LVM to relinquish control of a path or all the paths to a device so that diagnostics can be run against them.

Overview

Attached Devices

An attached device is one that can be freely accessed by LVM at any time. LVM typically attaches all the disks belonging to a volume group when the volume group is activated or re-activated via the *vgchange(1M)* command. All the devices in an active volume group should be presumed to be *attached* by LVM regardless of their status, unless they have been explicitly *detached*.

Detached Devices

Since *attached* devices can potentially be accessed by LVM at any time, they must first be *detached* before they can be safely replaced or before diagnostics can be run on them. A *detached* device is still part of the volume group but it is not accessed by LVM. Detaching a device directs LVM to conclude any LVM operations pending to the device and prepare for the device to be replaced. Formerly, without LVM OLR, devices were *detached* only when the volume group was deactivated, consequently a volume group had to be deactivated before servicing a disk or path.

LVM OLR

LVM OLR provides new options to the *pvchange(1M)* command to allow detaching and reattaching a specific device while the volume group remains active. Once a device has been *detached*, site specific diagnostics can be run, and/or the device can be replaced

safely without interference or complaints from LVM. *Detached* devices remain that way until they are attached again using *pvchange(1M)*, or until the volume group is re-activated using *vgchange(1M)*. Detached devices are automatically attached the next time the volume group is activated or when the system is rebooted.

Data Availability Considerations

Clearly, detaching a failed disk has no effect on data availability, since the data on the disk was lost earlier when the disk failed. Detaching a failed disk and replacing it can only improve the availability of the user data.

However, when considering detaching a partially functional disk or one that is performing poorly, it is important to keep in mind that the disk likely contains some available user data and that detaching the disk will make *any available copies of the user data on the disk unavailable*. If the disk contained the only available copy of any user data, user applications will no longer be able to access the data.

Application/Filesystem Considerations

When replacing a disk using the new LVM OLR procedure, it is not necessary to stop the applications or remove the file system using the logical volume residing on the disk *if the data in logical volumes on the disk is mirrored and available on other disks in the volume group.* After the disk is replaced and reattached, LVM will automatically synchronize the data from the other copies in the volume group.

However, if the data on the disk that is being replaced is not mirrored, or if the mirrored data has been compromised due to simultaneous mirror failures, then the user data is already lost, or will be lost when the disk is replaced. Any applications using these logical volumes must be halted, and any file systems residing on them must be unmounted. The user data must then be restored manually from a backup after replacing the disk.

Be prepared with a comprehensive backup strategy that includes recent LVM configuration backups

Although disk failures are rare, it is important to be prepared with recent backup copies of the LVM configuration, and the data in the volume groups. Refer to the LVM whitepaper ***When Good Disks Go Bad: Dealing with Disk Failures under LVM*** for more information about disk replacement strategies and methods.

Commands

Pvchange(1M)

Detaching a path

pvchange –a n *dev*

The *pvchange(1M)* command provides a new ***–a n*** option that allows detaching the device associated with a specific path to a disk. Detaching the active path to a multiported device causes LVM to begin using an alternate path to the disk, if one is available and configured. If there are alternate paths to the disk that are still attached, the disk may still be in use by LVM even when the specified path is not.

Detaching a physical volume

pvchange –a N *dev*

The *pvchange(1M)* command provides a new ***–a N*** option to allow detaching all the paths to a physical volume. Detaching the last path to a physical volume detaches the associated physical volume from the volume group. Detaching the paths using *pvchange(1M)* ***–a N*** is equivalent to using *pvchange(1M)* ***–a n*** repeatedly to detach all the paths to a PV. When a PV is detached it is unavailable to the volume group. The administrator may then run any diagnostic command on the disk or alternate paths to it without any concern that it will affect the volume group.

Attaching a path

pvchange –a y *dev*

The *pvchange(1M)* command provides a new ***–a y*** option to attach the device associated with a specific path to a disk. Attaching a path to a physical volume informs LVM to examine the disk and resume using it if possible. If a disk has multiple paths each path must be individually attached to make them all available.

Vgcfgrestore(1M)

Initializing a replaced disk

vgcfgrestore –*n vg dev*

Executing *vgcfgrestore(1M)* is necessary after a disk is replaced, prior to reattaching it to the volume group, so that LVM recognizes that the disk is a replacement for the disk that was there before. The disk must be detached prior to running *vgcfgrestore(1M)*.

Procedures

Before using LVM OLR: A Cautionary Note

Before resorting to disabling or replacing a disk or path, it is important to be certain of the status of all the disks in the volume group, and to carefully consider the alternatives. Detaching the wrong path or disk can have unintended consequences. Refer to the LVM whitepaper ***When Good Disks Go Bad: Dealing with Disk Failures under LVM*** for a description of different disk failure scenarios and how best to handle them.

Isolating a troublesome device in an active volume group

There are circumstances where it is desirable to have LVM simply stop using a given device or path. For instance, if a disk containing *mirrored* data is performing badly and it is not convenient (or possible) to hot-replace the device, it might make sense to isolate the disk or path until the next maintenance interval when it can be serviced.

The *pvchange(1M)* command can be used to isolate the path or disk:

pvchange -a n *path* *(to detach the path only)*

or

pvchange -a N *path* *(to detach the disk referenced by the path. i.e. detach all paths to the disk)*

The command returns once the path or disk is detached.

Detaching a path or disk makes it unavailable to the volume group. If the last path to a PV or the entire PV is detached, the PV will not be available to the volume group. Diagnostics can be safely run on a detached device. The path or disk will remain unused by LVM until it is re-attached or the volume group is reactivated.

LVM Online Disk Replacement Procedure

Devices that are hot-replaceable can be replaced online using the LVM OLR feature without deactivating the volume group or changing the volume group configuration.

1. Prepare to replace a failed disk:

Determine if there are unmirrored logical volumes on the disk or any mirrored logical volumes that may have been compromised due to simultaneous failures. Halt any applications and unmount any file systems using these logical volumes. See the section entitled: *When replacing a disk requires halting applications and restoring data from a backup* for more information.

Halting applications and unmounting file systems is only necessary when the data is not mirrored or the mirrors have been compromised. In these circumstances, the data must be restored from backup (at a later step). Halting any activity first, prevents the applications or file systems from writing inconsistent data over the newly restored replacement disk.

2. Detach the device using a new *pvchange(1M)* command option:

pvchange -a n *path*

or

pvchange -a N *path* *(to detach **all paths to the physical volume**)*

Pvchange(1M) returns once the path or physical volume is detached. Detaching a path or disk makes it unavailable to the volume group. If the last path to a PV or the entire PV is indicated to be detached, the PV will not be available to the volume group.

3. Manually replace or repair the disk or path, and/or run site specific diagnostics as necessary to determine any problems with the device.

4. When replacing a boot disk on an IPF system, setup the disk partitions (otherwise skip to step 5):

idisk -wf *partitionfile disk*

(Note the ***disk*** refers to the disk device and ***partitionfile*** is a text Partition Description File, see the *idisk(1M)* manpage for the details)

5. When replacing boot or non-boot disks on any system, initialize the LVM information on the disk so it is identified as a replacement disk when it is later attached:

vgcfgrestore –*n vg path*

(Note that for IPF boot disks the ***path*** refers to partition 2 of the ***disk***)

6. Attach ***each*** path to the device using a new *pvchange(1M)* command option:

pvchange -a y *path*

After processing the *pvchange(1M)* command, LVM will resume using the device if possible. If it is desirable to attach all the paths in the entire volume group at once, this step can be skipped, step 7 would suffice to bring all the replaced devices online again.

7. Attach all the VG devices using *vgchange(1M)* to resume automatically recovering any other unattached disks in the volume group:

vgchange -a *mode* *vg*

(The ***mode*** is whatever mode that the ***vg*** is already activated in: i.e. **y**, **s** or **e**)

Re-activating the volume group using *vgchange(1M)* attaches all the paths for all the disks in the volume group and resumes automatically recovering any unattached failed disks in the volume group. *Vgchange(1M) should only be done after all work has been completed on all the disks and paths in the VG, and it is desirable to attach them all.*

8. If replacing a boot disk on any system, initialize the boot data

mkboot –e –l *disk* (–e and –l options are for IPF boot disks only)

lvlnboot –v –R *vg*

Verify the *lvlnboot(1M)* output indicates the proper root, boot, swap and dump logical volumes. If not, use the *lvlnboot(1M)* command to explicitly set them.

9. If there were unmirrored logical volumes or any compromised mirrored logical volumes on the disk, restore the data from backup, mount the file systems, and restart any applications that were interrupted during step 1. This step is not necessary when all the logical volumes on the replaced disk are mirrored and none are compromised.

If there are mirrored logical volumes residing on the disk, LVM will automatically synchronize the data on the disk with the available mirror copies on other disks in the volume group.

Notes:

- **Disks cannot be replaced in read-only volume groups** since the volume group must be activated in some writeable mode to allow LVM to synchronize the data on the replaced disk.
- **Disks that are detached are not spared.** For instance, if there is a spare PV associated with the VG LVM will not automatically replace the detached disk with the spare.
- **Adding a new path to a detached disk using vgextend(1M)** will cause the disk to be attached and accessed by LVM again via the new path.
- **There are special considerations for disks that are part of Serviceguard clusters**, see the section entitled *Replacing an LVM disk in a Serviceguard Cluster volume group*.

Replacing an LVM disk in a Serviceguard cluster volume group

Replacing LVM disks in volume groups in a Serviceguard cluster follows a similar procedure to the one described earlier except that if the volume group is shared, the commands to attach and detach the disks and paths must be executed individually on each of the cluster nodes sharing the volume group.

Special care is required when performing a Serviceguard rolling upgrade. Disks that are part of a shared volume group on a Serviceguard cluster should not be replaced online until after all the nodes on the cluster have the LVM Online Disk Replacement patches installed. The LVM OLR feature provides new cluster broadcast messages that are sent from the cluster server to the client nodes when disks are reattached so that all the nodes recognize that the replaced PV has stale user data on it. Any nodes not updated with these patches will crash when the replaced disk is recovered by the cluster server. This is necessary to prevent the potential data corruption that could result due to the nodes in the cluster having inconsistent maps of stale or fresh data on the replaced disk. See the section entitled *Determining whether LVM OLR is available on the system* for a description of how to ensure the LVM OLR is present on each cluster node.

When replacing a disk requires halting applications and restoring data from a backup

Replacing a disk requires halting applications and restoring data from a backup under the following circumstances:

- If the disk being replaced contains any unmirrored logical volumes
- If any data in the mirrored logical volumes on the disk being replaced has been compromised due to simultaneous disk failures

If any of the logical volumes on the disk being replaced are not mirrored, the disk being replaced held the only copy of the data and by definition there are no mirror copies elsewhere. So replacing the disk requires halting applications and filesystems using the disk and restoring the data from backup.

If all the logical volumes on the disk being replaced are mirrored and none have been compromised due to simultaneous disk failures, it is not necessary to halt applications using the logical volumes and there is no need to restore any data to the logical volumes from a backup. Once the disk is replaced and attached to the volume group again, LVM will automatically resynchronize the data on the disk.

The data in a mirrored logical volume is ***compromised*** when no remaining ***available*** disk in the volume group contains a ***non-stale*** copy of the data. Prior to replacing a disk it is important to ensure that for each extent in each logical volume on the disk being replaced there is a non-stale copy on a different available disk. The *lvdisplay(1M) -v* command can be employed to display each of the logical volumes and the state of all the extents within them. If there is any extent that does not have a non-stale copy on some other available disk, halt any applications using the data and restore the data from backup after replacing the disk.

For example, given a logical volume is mirrored across two disks: ***A*** and ***B***. If disk ***A*** contains some stale extents, and disk ***B*** fails, the logical volume is ***compromised***. The cause was a partial failure of disk ***A*** (perhaps a few I/O requests timed-out) coincident with a complete failure of ***B***, before ***A*** could be re-synced from ***B***. Displaying the logical volume shows that although disk ***A*** is available, it contains some stale data. Although disk ***B*** may appear to have a non-stale copy of all the data, that disk is down so the copy there is unavailable. Then if disk ***B*** is replaced, the data on the replacement disk cannot be automatically synced with the copy from disk ***A*** because the stale data there is unavailable. Under these circumstances applications using the logical volume must be halted prior to replacing the disk, and after the disk is replaced, it must be restored from backup.

It is important to note that simultaneous disk failures are a rare event provided that the disks and I/O hardware in the volume group are properly isolated and promptly replaced at the first sign of failure. Essentially a mirrored volume that has only a single available copy of any of its data has already weathered one or more failures and is susceptible to being compromised by just one subsequent disk failure.

Determining whether LVM OLR is available on the system

The feature is available in LVM patches for HP-UX 11.11 and 11.23. Both command and kernel components are required to enable the feature:

- For 11.11: Patches PHKL_31216 and PHCO_30698 (or superseding patches)
- For 11.23: Patches PHKL_32095 and PHCO_31709 (or superseding patches)

LVM OLR will also be available in the next HP-UX release after 11.23 (11i V2).

The *pvchange(1M)* command can be employed to determine whether LVM OLR is available: just attempt to reattach any LVM device:

pvchange -a y *path*

If the feature is *not* available on the system, the command will fail with a message indicating that the option is illegal or that the HP-UX kernel does not provide the feature. See LVM OLR messages 1 and 2 in the LVM OLR Messages section for a complete description of these messages. When LVM OLR is properly installed, the command may succeed or fail, but will never display these messages.

LVM OLR Messages

pvchange(1M)

1. Usage: pvchange [-A Autobackup] [-s] | {[-S Autoswitch] [-x Extensibility] [-a Availability] [-t IOTimeout] [-z SparePV]} PhysicalVolumePath

   "a": Illegal option.

   This error is reported when ***pvchange(1M) –a y*** is attempted on systems that do not have LVM OLR available.

2. The HP-UX kernel running on this system does not provide this feature. Install the appropriate kernel patch to enable it.

   This error is reported when the command patch is installed and the kernel patch is not. Both the LVM command and kernel components are required to enable the LVM OLR feature. If this message occurs, install the appropriate kernel patch to enable the LVM OLR feature.

3. Warning: Detaching a physical volume reduces the availability of data within the logical volumes residing on that disk. Prior to detaching a physical volume or the last available path to it, verify that there are alternate copies of the data available on other disks in the volume group. If necessary, use pvchange(1M) to reverse this operation.

   This warning is advisory and generated whenever a path or PV is detached to emphasize the potential harm that could be done.

4. Unable to detach the path or physical volume via the pathname provided. Either use pvchange(1M) -a N to detach the PV using an attached path or detach each path to the PV individually using pvchange(1M) –a n

   This error is reported when the path specified is not part of any volume group, or because the path has not been successfully attached before. If this error occurred while detaching a path (using the ***–a n*** option) reissue the command using a path that belongs to the volume group. If the error occurred while detaching a PV (using the ***–a N*** option), specify a different path to the PV that is attached. If detaching a PV (using ***–a N***) fails, the same result can be accomplished by individually detaching each path to the PV with ***pvchange(1M)*** (using ***–a n***).

vgcfgrestore(1M)

1. Cannot restore Physical Volume /dev/rdsk/c7t2d0 Detach the PV or deactivate the VG, before restoring the PV.

   This error is reported when *vgcfgrestore(1M)* is used to initialize a disk already belonging to an active volume group. If this error occurs, detach the PV or deactivate the VG before attempting to restore the PV. If there is reason to believe that the data on a given disk is corrupted, the disk can be detached and marked using

vgcfgrestore(1M) then attached again (without replacing the disk). This causes LVM to reinitialize the disk and synchronize any *mirrored* user data mapped there.

Syslog Messages

1. LVM: VG 64 0x010000: Data in one or more logical volumes on PV 188 0x072000 was lost when the disk was replaced. This occurred because the disk contained the only copy of the data. Prior to using these logical volumes, restore the data from backup.

   This warning is reported when LVM cannot synchronize the data on a replaced disk automatically, and the data must instead be restored manually by the administrator from a back-up. This warning typically occurs when LVM discovers an *unmirrored* logical volume residing on the disk that was just replaced. When all the data on a disk is *mirrored* on other disks and a copy is available, LVM will automatically synchronize the data on the replaced disk from the mirror copies.

2. LVM: VG 64 0x010000: PVLink 188 0x072000 Detached.

   This message is advisory and generated for each detached path. The major/minor number pair listed in the error message can be mapped to a path by performing a long format listing of the */dev/rdsk* directory:

   ```
   #ll /dev/rdsk/ | grep 72000
   crw-r----- 1 bin    sys    188 0x072000 Jan 16 2004 c7t2d0
   ```

   For this example, the major number 188 and minor number 0x072000 refers to /dev/rdsk/c7t2d0

For More Information

To learn more about LVM and HP-UX system administration, refer to the following documents on the HP documentation website (http://docs.hp.com):

- Managing Systems and Workgroups
  http://docs.hp.com/en/5990-8172/index.html

- When Good Disks Go Bad: Dealing with Disk Failures under LVM
  http://docs.hp.com/en/5991-1236/When_Good_Disks_Go_Bad.pdf

Call to action

HP welcomes your input. Please give us comments about this white paper, or suggestions for LVM or related documentation, through our technical documentation feedback website:

http://docs.hp.com/en/feedback.html