Replicating data to the IBM Cloud - GLVM

Background

As more companies evolve to using a hybrid computing mode, that is some of their applications running on-premise and some running in one or more commercial data centres (or clouds). Problems can arise when these two environments want to share data as at the moment the only connection supported is over IP. While this is not such a concern if the application itself will support replication over IP, such as some databases like DB2, Oracle and SAP HANA. It is an issue for non-database applications etc.

AIX has supported IP replication of data volumes for a long time now – first there was GeoRM, which was part of HAGeo. Then in 2008 AIX supported both synchronous and asynchronous replication of logical volumes over IP via Geographic Logical Volume Manager (GLVM). This meant that a file system or raw logical volume could be replicated to a remote system with no restriction imposed by the choice of database or the application that you were using.

In this paper I will show how you can create an AIX Volume group that spans LUNs attached to your local system and LUNs attached to your LPAR in the IBM Cloud. While Geographic Mirrored Volume Groups (GMVG) can support up to 3 copies, GLVM currently only supports two sites, therefore only one site can have two copies of the mirror. This configuration is common where the customer wants to avoid moving the application to another site if they experience an outage due to a failure of one copy of the local storage. You can also have multiple servers at each site to provide greater availability, however only one server can manage the disks in the GMVG at a time.

As well as mirroring data between data centres, GLVM can be used to mirror data between a data centre and the cloud or between clouds.

It is also important to note, that while GLVM is part of AIX, and requires no extra licensing for a basic configuration, PowerHA SystemMirror Enterprise Edition is required to monitor the environment and to automate the management of both GLVM and the application(s).

> Note: Without PowerHA, AIX has no ability to monitor the state of either site and is therefore unable to control the state or mode of the GLVM daemons. Thus it is a manual process, with no checks to prevent the corruption or loss of data.

Please feel free to contact me for further information, corrections or a demonstration.

Antony (Red) Steel*
antony.steel@belisama.com.sg

* with thanks to Shawn Bodily for his advice and assistance

GLVM Concepts

At a very high level, GLVM provides a pseudo physical volume or volumes, which are treated by the AIX LVM as standard physical volumes and can be added to a volume group with actual local physical volumes. In reality each is just a local logical representations of the remote physical volume. On the remote system, where the actual physical volume resides, there is a Remote Physical Volume (RPV) Server for each replicated physical volume. On the local system, there is a device driver for each pseudo physical volume called the RPV Client. The AIX LVM manages the reads/writes for the pseudo physical volumes, while the RPV Client/Server pair manages the transfer of this data to the actual physical volume over the network.

GLVM

GLVM provides software based mirroring between two AIX Systems over an IP network to protect against loss of data from the active site. GLVM will work with any disk type supported by the AIX LVM. There is no requirement for the same type of disk subsystem at source and destination, just as the AIX LVM can mirror between two different disk subsystems locally. GLVM also has no dependency on the type of data being mirrored and supports both File systems and raw logical volumes.

The distance between the sites is limited only by the acceptable latency (for synchronous configurations) or by the size of the cache (for asynchronous configurations). For asynchronous replication, the size of the cache represents the maximum acceptable amount of data that can be lost in a disaster.

Note: GLVM is not supported for the rootvg.

To mirror your data across two sites, configure a volume group that contain both local and remote physical volumes. This is called a Geographic Mirrored Volume Group (GMVG).

Remote Physical Volume (RPV)

This is the pseudo local representation of the remote physical volume that allows the LVM to consider the physical volume at the remote site as another local, albeit slow, physical volume. The actual I/O operations are performed at the remote site.
The RPV consists of:

The RPV client

The RPV client is a pseudo device driver that runs on the active server/site, i.e. where the volume group has been activated. There is one RPV Client for each physical volume on the remote server/site and is named `hdisk#`. The LVM sees it as a disk and performs the I/Os against this device.
The RPV Client definition includes the remote server address and timeout values.

The RPV server

The RPV server is an instance of the kernel extension of the RPV device driver that runs on the node on the remote server/site, that is, on the node which has the actual

physical volume. The RPV Server receives and handles the I/O requests from the RPV client.
There is one RPV Server for each replicated physical volume and is named `rpvserver#`.

The GLVM Cache
This is a special type of logical volume of type aio_cache that is designed for use in asynchronous mode GLVM. For asynchronous mode, rather than waiting for the write to be performed on the remote physical volume, the write is recorded on the local cache, and then acknowledgement is returned to the application. At some later point in time, the I/Os recorded in the cache are played in order against the remote disk(s) and then deleted from the cache once successful (acknowledged).

Geographic Mirrored Volume Group

This is an AIX Volume Group that contains both local physical volumes and RPV Clients. See Figure 1 for a diagram of the components. You can mirror your data across two sites by configuring volume groups that contain both local physical disks and RPVs. With an RPV device driver, the LVM does not distinguish between local and remote physical volumes - it maintains mirror copies of the data across attached disks. The LVM is, for the most part, unaware that some disks are located at a remote site.
For PowerHA SystemMirror installations, the GMVGs can be added to resource groups and they will then be managed and monitored by PowerHA.

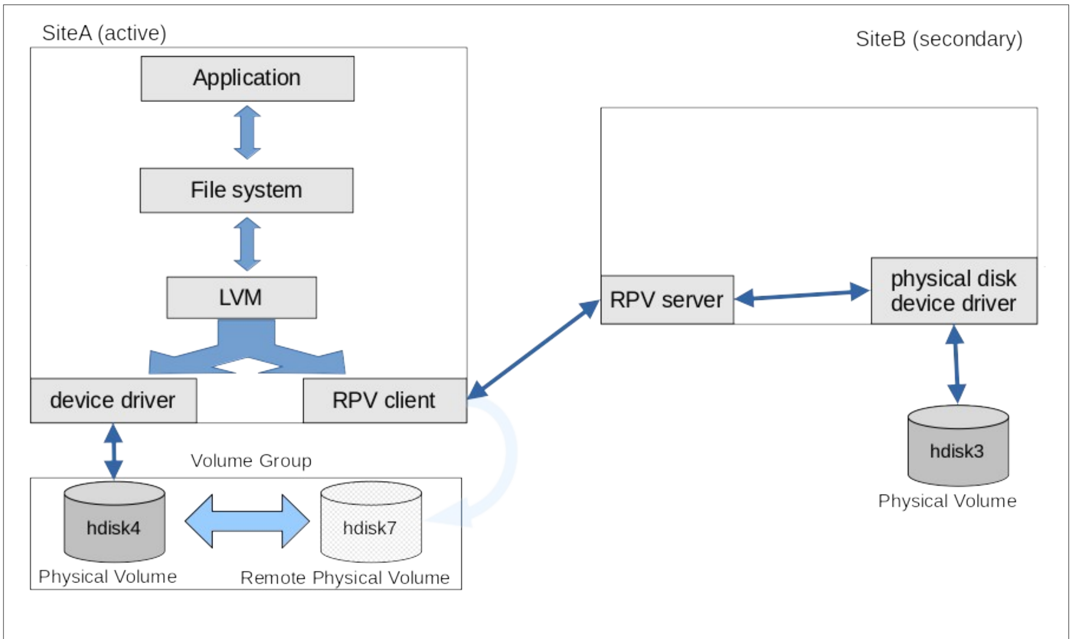


Figure 1: GLVM operation

Modes of replication

GLVM supports two modes of replication:

Synchronous

This was the first mode supported on AIX. Writes to a synchronous GMVG will be not complete until the remote copy acknowledges the successful write. This mode is typically impractical except for configurations where the two sites are typically within 100 km, depending on the latency requirements of the application.

Asynchronous

In this mode, writes are cached locally in a special logical volume in the same Volume Group and then marked as complete. Over time the changes recorded in the cache are played against the remote copy, and removed from the cache when the remote site acknowledges the change. This mode is much less sensitive to the latency, but limited by the size of the cache, remembering that the cache also represents the amount of data you can afford to loose in a disaster. This must be balanced against the cache being too small, since if the cache fills, all I/O will be suspended until space is cleared in the cache.

> Note: The size of the cache is based on what is required to manage the application’s peak workload. Tools such as `rpvstat` can be used to monitor the number of times the cache fills.

The size of the cache will also affect the amount of time taken during a move of the application from one site to the other, as the application will not be able to start until all the outstanding writes from the other site have been synchronised with the local copy.

In summary GLVM mirroring:

- Does not depend on the type of database and file system. There is no modification required of applications that use GLVM mirroring capabilities;
- Performs the data mirroring over standard TCP/IP network(s) without having a dependency on the specific data that is being mirrored;
- Is often less expensive than hardware-based mirroring solutions and does not require the same vendors storage at source and destination;
- Uses the mirroring function of the AIX LVM and operates as a layer under the LVM;
- The read preference can be configured to favour the local copy (discussed below), when available, to maximise performance; and
- During a write, the LVM and RPV device driver work together to allow the application to perform a single logical write. This results in multiple writes to the local and remote physical volumes that make up the GMVG.

Synchronous Operation

A simple configuration with a single GMVG “glvmg_vg” made up of hdisk4 at Site A and hdisk3 at Site B. When the Volume Group is active at Site A, it looks like Figure 2.

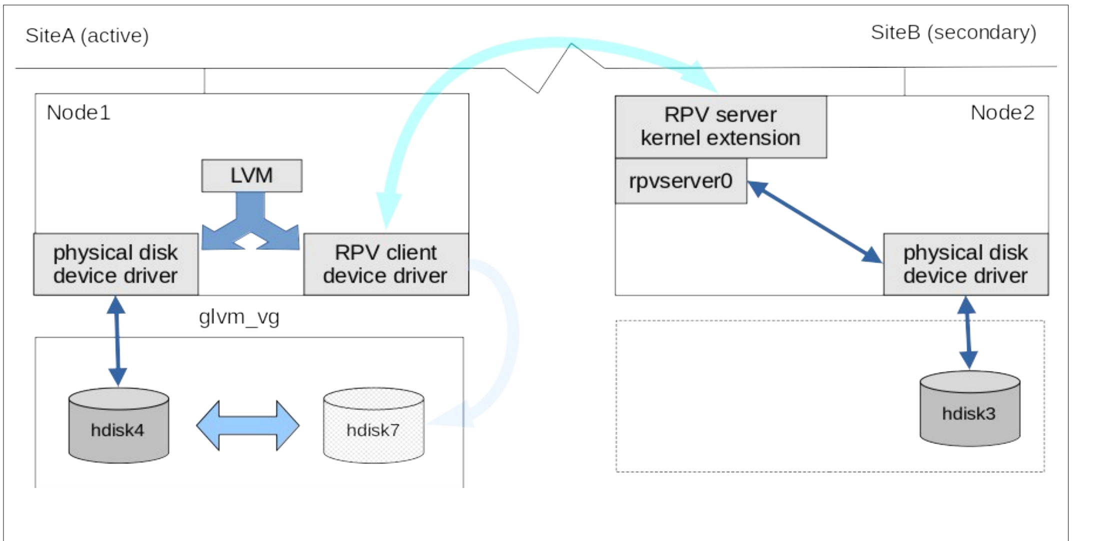


Figure 2: GMVG active at Site A

The application will write down through the LVM, which will send the write to both mirror copies, hdisk4 and the RPV client device driver, hdisk7. Once hdisk4 returns IOdone, the LVM will wait until the other mirror write is complete. The RPV Client will transfer the I/O over the network to the matching RPV Server, which will perform the same write on hdisk3 at the remote site. Once completed, the acknowledgement will be sent back via the RPV Client the LVM. Typically the local copy will have completed and acknowledgement can now return to the application.

When operation is reversed and the Volume Group is activated on Site B, see Figure 3.

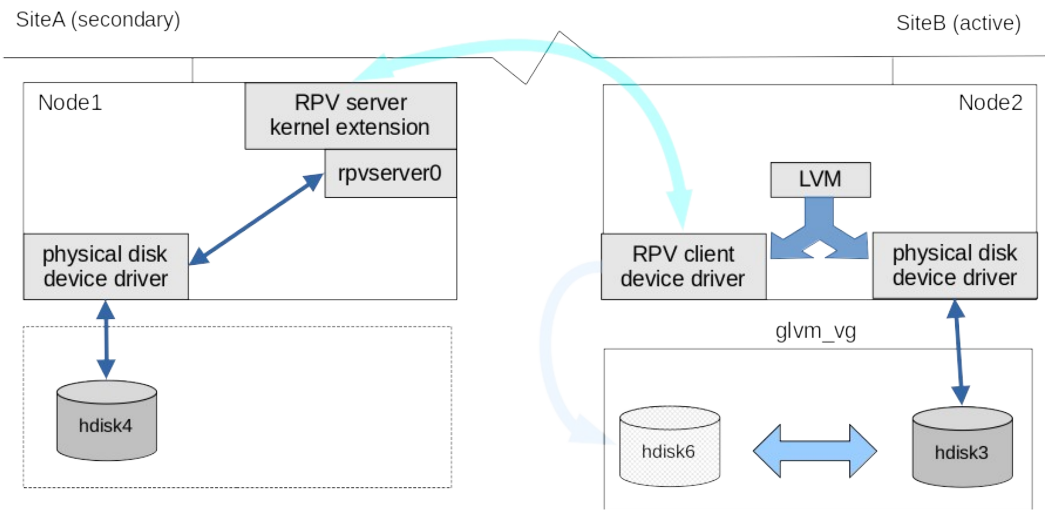


Figure 3: GMVG active at Site B

The same process as above with the GMVG active instead on Site B, so the operations reversed.

More complex scenarios

As mentioned above, you can have two copies of the Volume Group at one Site, see Figure 4.

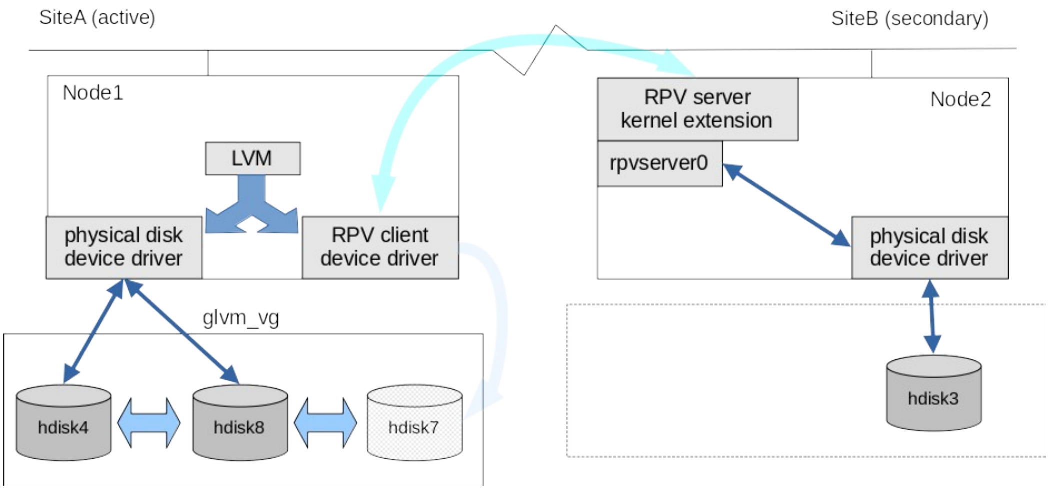


Figure 4: GMVG with two copies at Site A

If the configuration requires more than one physical volume, there will be a separate RPV Server and RPV client for each mirrored physical volume. See Figure 5.

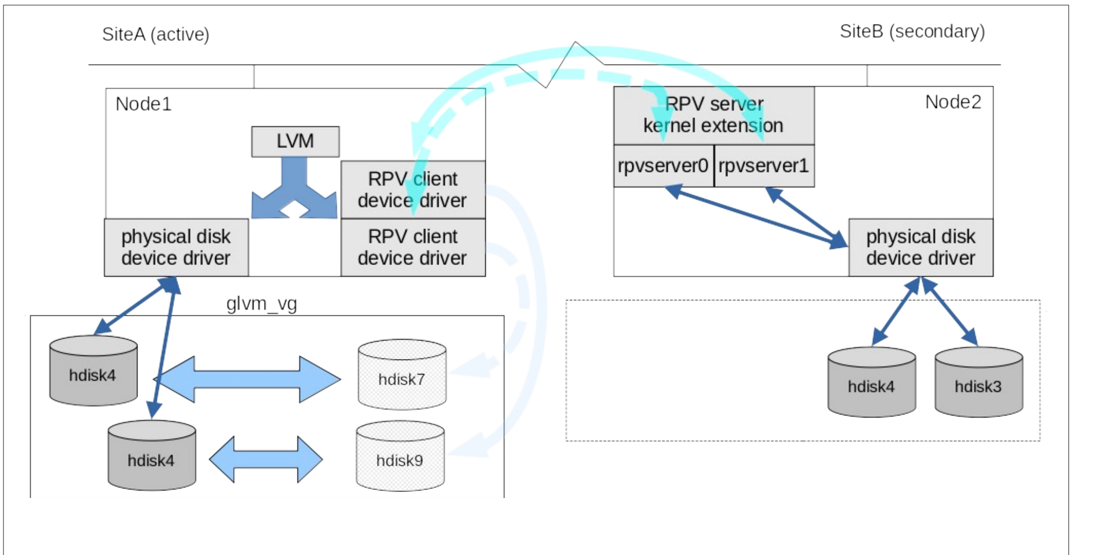


Figure 5: GMVG consisting of two mirrored physical volumes

Asynchronous mode GLVM

As discussed above, asynchronous mode GLVM allows control to be returned to the application once I/O completes to both the local physical volume and the local cache logical volume. The remote copy is then updated as bandwidth allows. This mode improves application response times, but it also increases the amount of data that can be potentially lost.

Asynchronous mode has stricter requirements and requires the use of *mirror pools*. There is also the cache logical volume of type `aio_cache`, that must be created for each mirror pool. In GLVM design, the `aio_cache` in the mirror pool at Production is the mirror pool associated with DR, as it contains the outstanding data updates for the logical volume(s) at DR and visa versa. See Figure 6 for details.

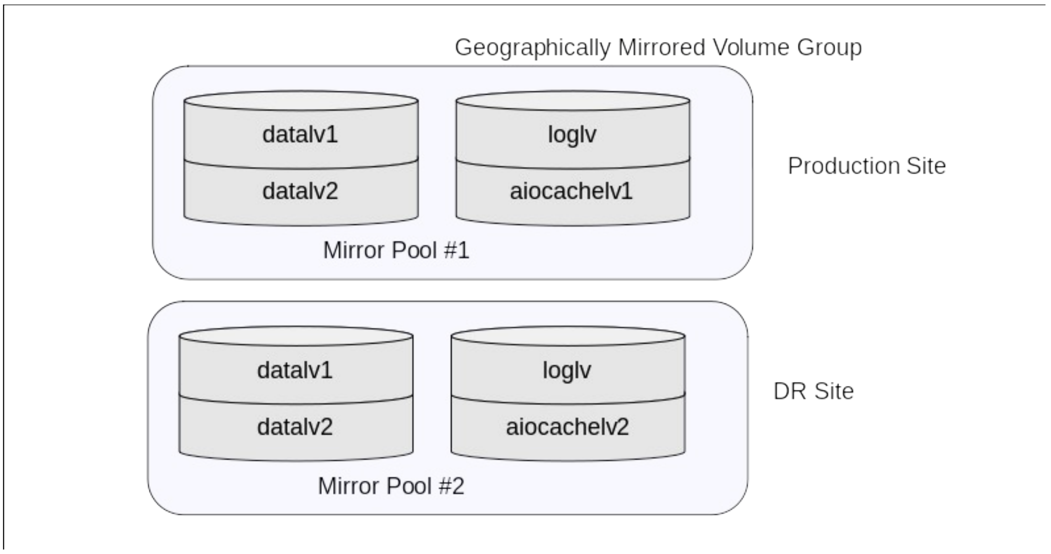


Figure 6: Asynchronous cache in each mirror pool

So above, if the GMVG is active on production, the logical volume aiocachelv1 will store the updates for the logical volumes at DR.

For local writes, the application passes the write to the LVM, the LVM passes the write to both the physical disk device driver and the RPV Client. When the physical disk I/O is complete, the LVM is updated and waits for the RPV Client to complete. Meanwhile the RPV Client has updated the cache with the write, and when that completes, updates the LVM. The LVM then returns control to the Application. See Figure 7.

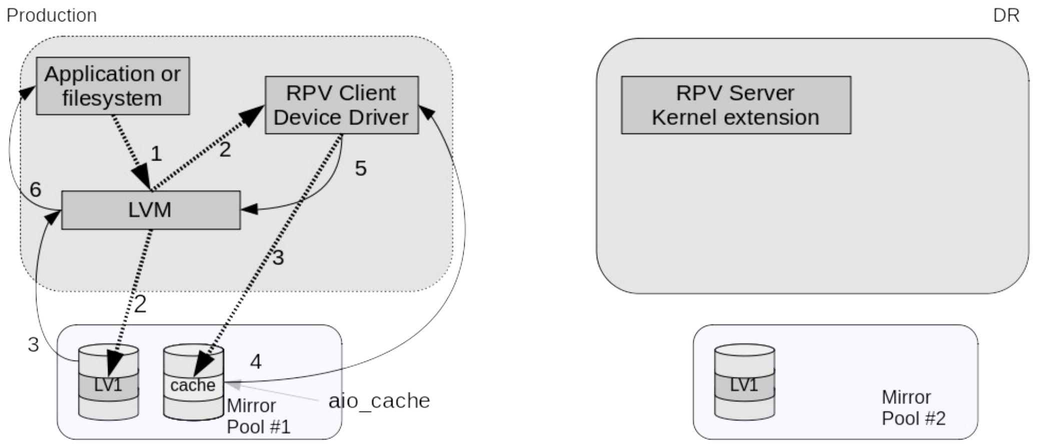


Figure 7: Asynchronous mode local write

When network bandwidth allows, the RPV Client will check for the next record in the cache, pass the I/O to the RPV Server, which will update the remote physical volume with the write, then return a completed response to the RPV Client. The RPV Client then deletes the record from the cache. See Figure 8.

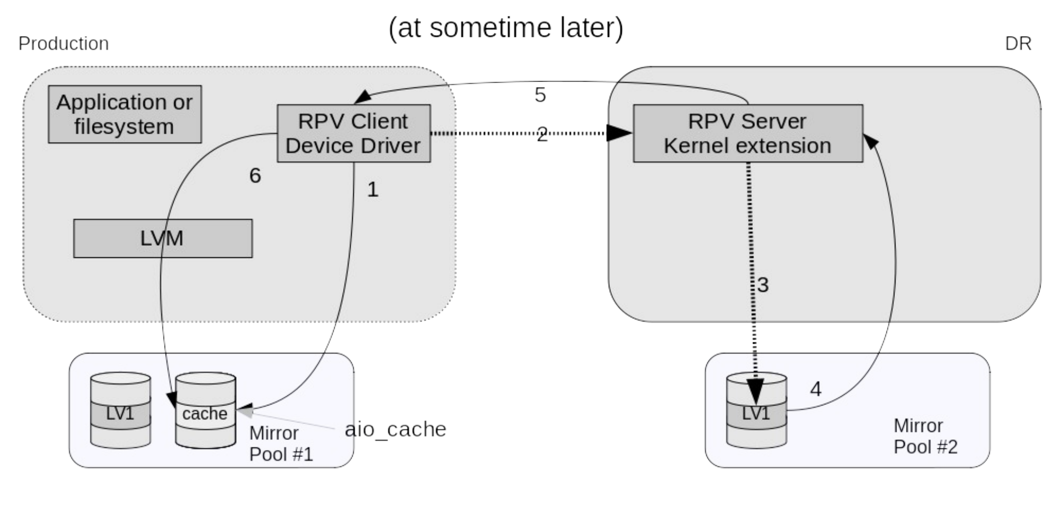
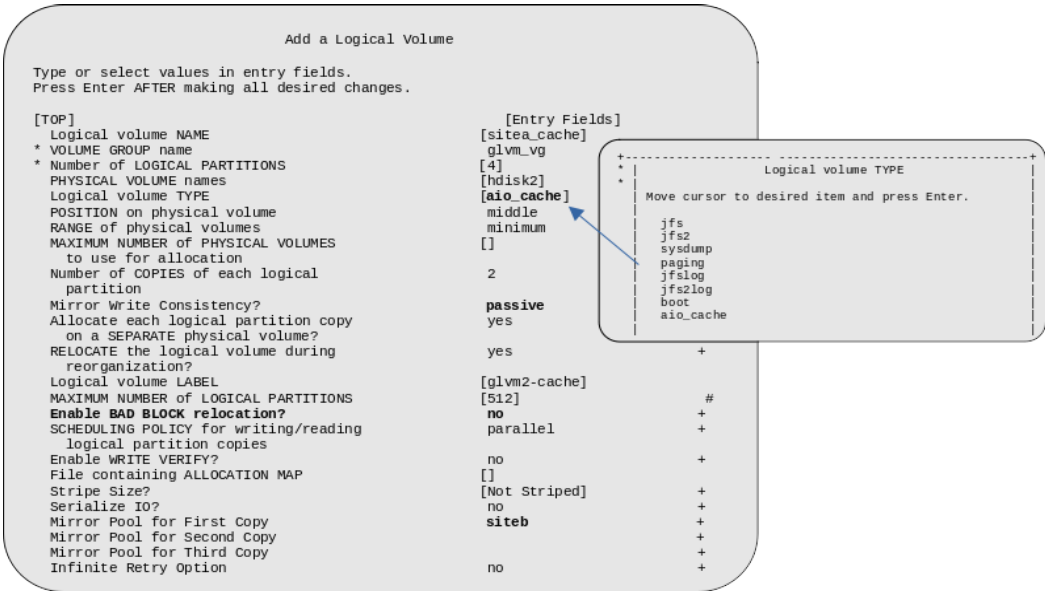


Figure 8: Asynchronous write, updating the remote physical volume

See Figure 9 for details when creating the aio_cache logical volume.



```
                              Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                  [Entry Fields]
  Logical volume NAME                               [sitea_cache]
* VOLUME GROUP name                                  glvm_vg
* Number of LOGICAL PARTITIONS                      [4]
  PHYSICAL VOLUME names                             [hdisk2]
  Logical volume TYPE                               [aio_cache]
  POSITION on physical volume                        middle
  RANGE of physical volumes                          minimum
  MAXIMUM NUMBER of PHYSICAL VOLUMES                []
    to use for allocation
  Number of COPIES of each logical                   2
    partition
  Mirror Write Consistency?                          passive
  Allocate each logical partition copy               yes
    on a SEPARATE physical volume?
  RELOCATE the logical volume during                 yes                     +
    reorganization?
  Logical volume LABEL                              [glvm2-cache]
  MAXIMUM NUMBER of LOGICAL PARTITIONS              [512]                     #
  Enable BAD BLOCK relocation?                       no                      +
  SCHEDULING POLICY for writing/reading              parallel                +
    logical partition copies
  Enable WRITE VERIFY?                               no                      +
  File containing ALLOCATION MAP                    []
  Stripe Size?                                      [Not Striped]            +
  Serialize IO?                                      no                      +
  Mirror Pool for First Copy                         siteb                   +
  Mirror Pool for Second Copy                                                +
  Mirror Pool for Third Copy                                                 +
  Infinite Retry Option                              no                      +
```

```
+-------------------- ---------------------------------+
* |                  Logical volume TYPE                  |
* |                                                       |
  | Move cursor to desired item and press Enter.          |
  |                                                       |
  |   jfs                                                 |
  |   jfs2                                                |
  |   sysdump                                             |
  |   paging                                              |
  |   jfslog                                              |
  |   jfs2log                                             |
  |   boot                                                |
  |   aio_cache                                           |
  |                                                       |
```

Figure 9: Cache for Asynchronous mode

GLVM Standalone

You can configure geographically mirrored volume groups in AIX GLVM, without having to install and configure a PowerHA SystemMirror cluster. The AIX GLVM technology provides the same geographic data mirroring functionality as GLVM for PowerHA SystemMirror Enterprise Edition, only without the automated monitoring and recovery which is provided by PowerHA SystemMirror.

Features introduced by PowerHA

These features include:

- Provides automatic detection and response to site and network failures in the geographic cluster without user intervention;
- Performs automatic site takeover and recovery and keeps mission-critical applications highly available through application fallover and monitoring;
- Allows for simplified configuration of volume groups, logical volumes and resource groups. Supports either standard or enhanced concurrent volume groups that are geographically mirrored;
- Uses up to four TCP/IP networks for remote mirroring;
- Supports concurrent access configurations, allowing all the nodes at one site to have concurrent access to the geographically mirrored volume groups. This is only supported at one site, so you cannot have concurrent access from nodes at both sites; and
- The ability to control the preferred read policy based on site.

Planning

Synchronous mode

Memory and CPU

Need to consider how much memory and CPU is required to achieve the I/O required and if compression is enabled (without NX Crypto Acceleration enabled).

Network bandwidth

Network bandwidth is a limiting factor when the amount of data to be sent over the network exceeds the network's capacity. If the network, or networks as PowerHA can support up to four, is at full capacity, network buffers and queues fill up and messages must wait to be sent. When this happens, I/O to the remote physical volume(s) will take even longer and application response times will suffers. While this might be acceptable for brief periods of peak activity or when running batch applications / non-critical interactive applications, it is typically not acceptable for most mission critical applications. Users will perceive the application as hanging, when in fact it is just waiting for remote I/Os to complete.
A network bandwidth problem can be resolved by upgrading the network or adding another network. For standalone use Etherchannel, if using PowerHA, it can support multiple networks. It is important to configure the network bandwidth to be able to handle the data throughput for the application workload at its peak, which typically means paying for higher bandwidth that may be rarely utilised.

Network latency

Network latency is the time that it takes for messages to go across the network. Even when there is plenty of network bandwidth, it still takes a finite amount of time for the bits to travel over the communication link. The speed of the network is limited by the quality of the switches and the laws of physics; the greater the distance between the sites, the greater the network latency. Even if a network is capable of transmitting data at a rate of 120 kilometres per millisecond, that still adds up over a long distance. For example, if the sites are 60 km apart, a remote physical volume I/O request must travel 60 km from the RPV client to the RPV server. After the disk gets updated, the result of the I/O request must travel 60 km from the RPV server back to the RPV client. This 120 km round trip adds about 1 millisecond to each remote physical volume I/O request, and this time can be much greater depending on the number and quality of routers or gateways traversed. Suppose that the sites are 4000 km apart, so each I/O request requires an 8000 km, adding approximately 67 milliseconds to each I/O request. The resulting application response time would in most cases be unacceptable. Synchronous mirroring is typically only practical, depending on the application, for metro distances, that is in the order of 100km or less. Greater distances would necessitate asynchronous replication.

Asynchronous mode

Memory and CPU

As with synchronous mode, need to consider how much memory and CPU is required to achieve the I/O required and if compression is enabled (without NX Crypto Acceleration enabled).

Network bandwidth

Typically a much smaller bandwidth is required for asynchronous operation as it smooths out the network use. Bandwidth has to be just be large enough to ensure that space can be kept free in the cache during peak workload.

Network latency

Asynchronous mode is ideal for configurations where there is a greater distance between the two data centres. As long as there is sufficient space in the cache, network latency will not impact application performance.

AIO cache logical volume size

How much do you expect the I/O load to exceed the network throughout during peak periods. Any backlog in transmitted data will be stored in the cache (with 1GB modified data requiring approximately 2GB of cache).

Another way of looking at the cache size is to use it as a way to limit the amount of data that would be lost in a disaster. Should you lose access to the production site and the cache logical volume, then all the updates in the cache will be lost.

Data divergence
Data divergence occurs when the GMVG is activated on one site, while there is outstanding data in the original site's cache – that is there is data that hasn't been updated at this site. In this instance the administrator has to decide if the original site can be brought online, and if so, how long will it take, or should operations continue with missing data. On returning to the original site which had data in the cache, the decision must also be made to continue with the version on the recently active site, or to revert back to the older local copy. Typically this will depend on how many updates where in the cache compared with updates made at the recently active site.

Figure 10 below shows an example of a peak in I/O against the GLVM mirrored logical volume. The second graph shows the affect of the network bandwidth with the same I/O load showing a slower draining of the I/Os.

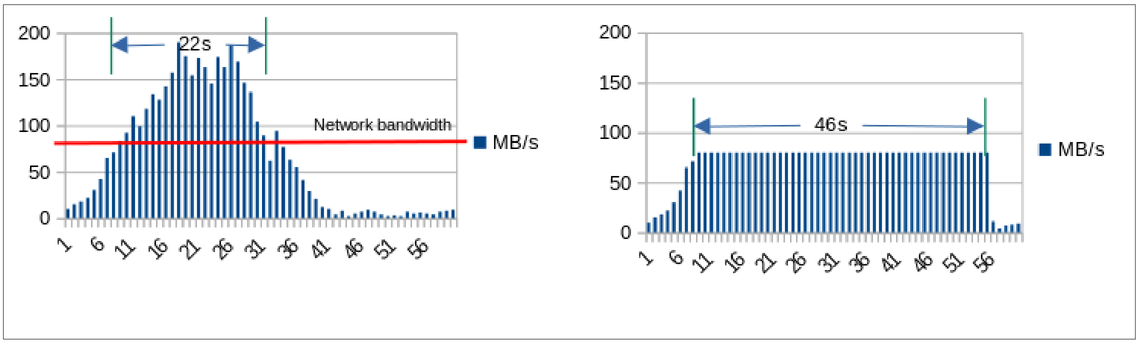


Figure 10: I/O and the impact of network bandwidth and slow drain of I/O

Another important consideration is whether to have two copies of each logical volume at the primary site. While this will mean that operations can continue at the primary site should one of the storage units fail, it does require consideration when moving back to the two copy data centre from the single copy data centre. If after recovery, operations continue at the site with one copy, there will be network traffic synchronising the two remote copies, and will not interfere with local read/write operations. However if operations move back before the copies are in sync, then reads to stale partitions will be done against the remote physical partition, competing on the network with traffic due to the synchronisation of the stale partitions.

Note: GLVM does not coalesce the writes to the two remote mirror copies, so doubles the network traffic.

However as stated above this design is not part of this demonstration, but is an important consideration if this design is chosen.

Quorum issues

In general, it is recommended to disable quorum for geographically mirrored volume groups in order to minimise the possibility of the volume group going offline should access to the remote copy be lost. Thus you will be able to keep operating in the event of an inter-site network failure or maintenance activity on the remote site.

> Note: If using PowerHA SystemMirror, it is a different discussion, since PowerHA detects quorum loss and manages the volume group.
> Disabling quorum will also require setting forced varyon for the volume group in PowerHA.

Requirements and limitations

GLVM imposes the following limitations:

- The inter-disk allocation policy for logical volumes in AIX must be set to `superstrict`. This policy ensures that there is a complete mirror copy on each set of either local or remote physical volumes. In GLVM, using the super strict policy for mirroring ensures that when you create a mirrored logical volume, you will always have a complete copy at each site;
- Up to three copies of the logical volumes can be created, with at least one mirror copy at each site. One of the sites may optionally contain a second copy. As mentioned above there will be extra considerations when moving back to the site with the two copies, as the write to each copy is sent separately over the network;
- Two sites, one local and one remote site. Site names must correspond with the site names if using PowerHA;
- The rootvg volume group cannot be geographically mirrored;
- While asynchronous mode requires configuring Mirror Pools, it is recommended for synchronous mode;
- The asynchronous GLVM volume group cannot contain active paging space logical volume and it is not recommended for synchronoous GLVM;
- You must use scalable volume groups. They can be non-concurrent or enhanced concurrent mode. The use of enhanced concurrent volume groups is required for use with PowerHA but doesn’t really provide any advantage for standalone GLVM. If you do use enhanced concurrent in standalone, there will be extra steps to active the GMVG;
- The volume group should not be configured to auto active (varyon);
- Bad block relocation should be turned off. If a bad block is detected at one site and the block is relocated, then the block maps will differ between sites. This is only required for asynchronous replication as it will impact the playing of the cached I/O against the remote physical volume(s) if the block maps differ;
- IP Security (IPsec) can be configured to secure the RPV client-server network traffic between the sites;
- 1MB of free space is required in /usr prior to installation; and
- Port 6192 TCP/UDP is open between the two servers.

AIX LVM Mirror Pools

Mirror pools are just a way to divide the physical volumes in a volume group into distinct groups or “Pools” and then control the placement of the logical partition’s mirrored copies. They were introduced in AIX 6.1.1.0 and only apply to scalable volume groups. Mirror pool names must be less than 15 characters and are unique within a volume group.

A mirror pool consists of one or more physical volumes and each physical volume can only belong to one mirror pool at a time. When defining a logical volume, each copy of the logical volume can be assigned to a specific mirror pool. This ensures that when a copy of a logical volume is assigned to a mirror pool, only partitions will be allocated from physical volumes in that pool. Prior to the introduction of mirror pools, the only way one could extend logical volumes and guarantee that partitions were allocated from the correct physical volume was to use a map file. Physical volumes can be assigned to a mirror pool with `chpv` or `exentvg`.

There cannot be more than three mirror pools in each volume group and each mirror pool must contain at least one complete copy of each logical volume that is defined in that pool.

> Note: Once mirror pools have been defined, the volume group can no longer be imported into versions of AIX prior to AIX 6.1.1.0. If using enhanced concurrent mode volume groups, all nodes in the cluster must also be greater than AIX 6.1.1.0.

Mirror pool strictness can be used to enforce tighter restrictions on the allocation of partitions in mirror pools. Mirror pool strictness can have one of the following values:

off This is the default setting and no restrictions apply to the use of the mirror pools.

on Each logical volume created in the volume group must have all copies assigned to mirror pools.

super This is specifically for GLVM and ensures that local and remote physical volumes cannot be assigned to the same mirror pool.

Mirror pool characteristics can be changed, however, any changes will not affect currently allocated partitions. This it is recommended to use the `reorgvg` command after any mirror pool changes so allocated partitions can be moved to conform to the mirror pool restrictions.

> Note: AIX LVM Mirror pools are only recommended for synchronous mode, but are required for asynchronous mode.

This mirror pools are used to ensure that:

- Each site has complete copy of each mirrored logical volume in the GMVG; and
- The cache logical volume for asynchronous GMVGs are configured and managed correctly.

Tuning options

The `rpvutil` command has the following options for tuning the operation of GLVM:

```
rpv_net_monitor=1|0
```

Setting rpv_net_monitor to 1, will turn on monitoring of the RPV network by rpvutil, so that the RPV client will detect any network failures and attempt to resume after the network recovers. The default is 0 (disabled).

`compression=1|0`

Before using compression, check that:

- Both the RPV client and the RPV server are running AIX version 7.2.5, or later, with all the latest RPV device drivers;
- Both the RPV server and the RPV client are IBM Power Systems servers with NX842 acceleration units. (See below); and
- The compression tunable parameter is enabled on both the RPV server and RPV client so that the I/O data packets are compressed when the workload is failed over between the RPV client and the RPV server.

When the compression tunable parameter is set to 1, the rpvutil command compresses the I/O data packet before it is sent from the RPV client to the RPV server by using the cryptography and compression units (NX842) on IBM Power Servers. If the I/O data packet is compressed successfully, a flag is set in the data packet. When the RPV Server receives a packet with the compressed flag set, the packet will be decompressed. If the NX842 compression unit is not available, the RPV Server will attempt software decompression of the packet.

By default this option is set to 0 (disabled).

`io_grp_latency=timeout_value` (milliseconds)

Used to set the maximum expected delay before receiving the I/O acknowledgement for a mirror pool that is configured in asynchronous mode. The default delay value is 10 ms and a lower value can be set to improve I/O performance, but may be at the cost of higher CPU consumption.

`nw_sessions=<number of sessions>` (1 to 99)

This is a new tunable (available in AIX 7.2.5.2) that controls the number of RPV sessions (sender and receiver threads) to be configured per network. This is used to increase the number of parallel RPV sessions per GLVM network, which will send more data in parallel, improve data transfer rate and more fully utilise the network bandwidth.

Setting hardware compression

Check that hardware compression is possible, for example:

```
pcha1:/:# prtconf
. . .
NX Crypto Acceleration: Capable and Enabled
. . .
```

Listing the value of compression tunable:

```
pcha1:/:# rpvutil -o compression
compression = 0
```

Turning on compression:

```
pcha1:/:# rpvutil -o compression=1
Setting compression to 1
```

Showing compression enabled:

```
pcha1:/:# rpvutil -o compression
```

```
        compression = ENABLED
```

Turning off compression:

```
        pcha1:/:# rpvutil -o compression=0
        Setting compression to 0
```

Improving IOPS with io_grp_latency

As noted above GLVM forms async groups once every 10 ms and performs write to remote site. Thus each write would wait for at least 10 ms. io_grp_latency provides the ability to control group formation time and the reduction of this time would give quicker response to applications. By creating new I/O groups faster, the application will have faster writes to the cache device. but at the possible cost of higher CPU consumption.

Recommendations

The following settings are recommended for setting and configuring up GLVM.

General recommendations:

- Issues have been found with potential deadlocks if Mirror Write Consistency is set to “active” for asynchronous GMVG. Setting of “passive” is recommended for both asynchronous and synchronous modes;
- Configure RPV level IO timeout value to avoid any issues related to network speed or IO timeouts. This value can be modified, when RPV disk is in defined state. Default value is 180 secs;
- AIX LVM allows the placement of disks in mirror pools, and then selecting read preference based on the mirror pool. A feature that was added for GLVM in PowerHA, is for physical volumes to be added to sites, and then the preferred read to be set to “siteaffinity”. This option is not available for standalone GLVM users, instead they will need to set the LVM preferred read to the local mirror pool before activating the volume group;
- Turn of quorum and have multiple networks in PowerHA or etherchannel in the standalone. Ensure that all networks follow different paths and have no shared point of failure;
- `rpvstat -n` will give details with individual network and `rpvstat -A` will give details about asynchronous IO; and
- For better performance ensure that disk driver parameters are configured correctly for the storage deployed in your environment. Refer to AIX and storage documentation for setting those tunables (e.g.: queue_depth, num_cmd_elems, etc.).

Recommendations for asynchronous mode GLVM:

- Asynchronous GLVM is ONLY supported on scalable volume group(s). These may be in enhanced concurrent mode;
- As discussed above, you can lower the timeout parameter for the RPV client to improve application response times, but balance this against latency problems. This value can be changed when the RPV client is in a defined state;
- Reducing the max_transfer size for the remote device while there is data in the AIO cache can cause remote IO failures. (`lsattr -El hdiskX  -a max_transfer`);

- In a stand-alone GLVM environment, you must ensure that all the backup disks in the secondary sites are in an active state before you bring the volume group online. During the online recovery of the volume group, if the RPV device driver detects that the RPV server is not online it updates the cache disk detailing a failed request and all subsequent I/Os will be treated as synchronous. To convert back to asynchronous mode after the problem is rectified, one must first convert the mirror pool to synchronous mode and then back to asynchronous mode using `chmp` as described below;
- When an asynchronous GMVG it brought online, it will perform a cache recovery. If previously the node halted abruptly, say with a power outage, it is possible that the cache is not empty. In this case, cache recovery may take some time, depending upon amount of data in the cache and the network speed. No application writes are allowed to complete while cache recovery in progress to handle consistency at remote site. In this case, the application users may observe a pause;
- After a site failure, asynchronous mirror state on remote site will be inactive. After integrating back with primary site, mirror pool needs to be converted to synchronous and then back to asynchronous so as to continue in asynchronous mode. (See maintenance tasks below);
- Monitor regularly whether the asynchronous mirroring state of the GLVM is active by using the `lsmp` command.
- `rpvstat -C` will give the details about IO cache Monitor and `rpvstat –G` will give details such as number of times the cache is full.
- For better performance, ensure that the disk driver parameters of the storage device that is deployed in your environment is configured correctly.

Monitoring and analysis

The following tools can be used to both plan the configuration and monitor the ongoing operation.

Tools to analyse I/O for an existing system

gmdsizing

`gmdsizing` is a command to estimage network bandwidth requirements for GLVM networks. It was originally part of HAGeo / GeoRM and is part of the samples in PowerHA installations (find in /usr/es/sbin/cluster/samples/gmdsizing/gmdsizing). It monitors disk utilisation over the specified period and produces a report to be used as an aid for determining bandwidth requirements. For details see Appendix 3.

lvmstat

`lvmstat` which reports input/output statistics for logical partitions, logical volumes and volume groups. Also reports pbuf and blocked I/O statistics and allows pbuf allocation changes to volume groups.

```
lvmstat { -l | -v } Name [ -e | -d ] [ -F ] [ -C ] [ -c Count ] [ -s ]
[ Interval [ Iterations ] ]
```

iostat

iostat command reports CPU statistics, asynchronous input/output (AIO) and input/output statistics for the entire system, adapters, TTY devices, disks CD-ROMs, tapes and file systems. Use flags -s -f to show logical and disk I/O.

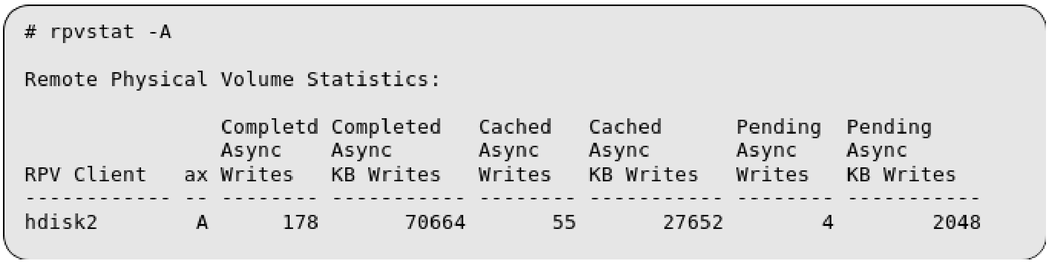
Other tools

Other monitoring tools such as nmon and topas can also be used. An example is provided below using Grafana and InfluxDB.

RPV and Cache monitoring

The rpvstat command provides detailed reporting of RPV Client statistics. For asynchronous mode GLVM, the state of the cache is critical to the operation of GLVM. If the cache does become full, then all local writes are suspended until space is cleared. The rpvstat command can be used to determine how many times this occurred. The administrator has to decide whether to increase the size of the cache (and potentially losing more data if a disaster occurs) or to increase the network bandwidth (incurring greater cost).

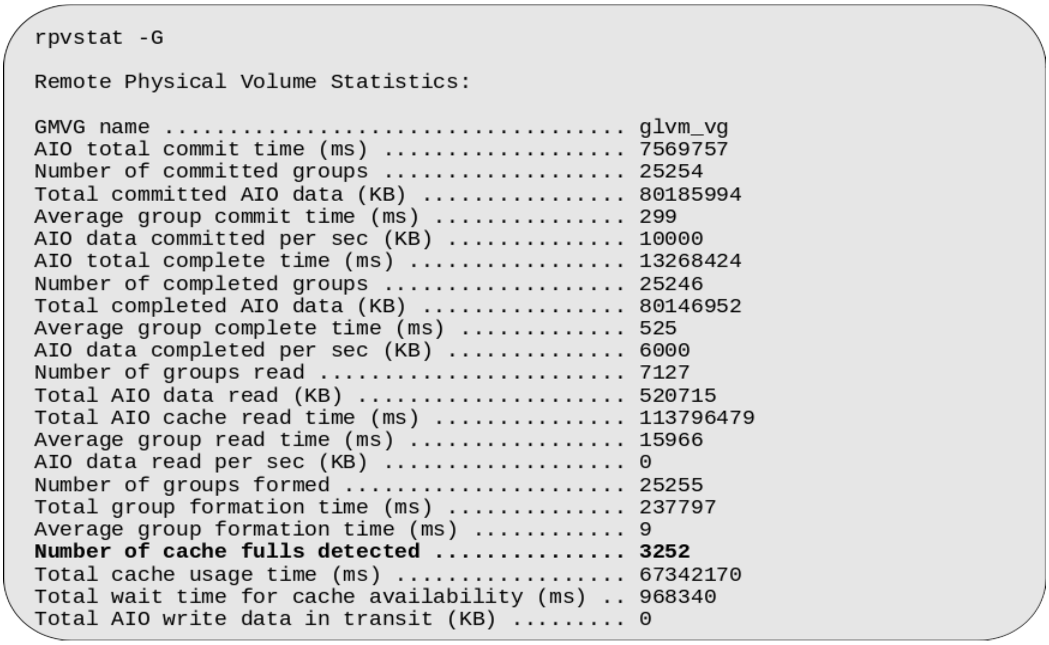
The command rpvstat -A will show the synchronous statistics. See Figure 11



```
# rpvstat -A

Remote Physical Volume Statistics:

                Completd Completed   Cached   Cached      Pending  Pending
                Async    Async       Async    Async       Async    Async
RPV Client   ax Writes   KB Writes   Writes   KB Writes   Writes   KB Writes
------------ -- -------- ----------- -------- ----------- -------- -----------
hdisk2        A      178       70664       55       27652        4        2048
```

Figure 11: rpvstat -A

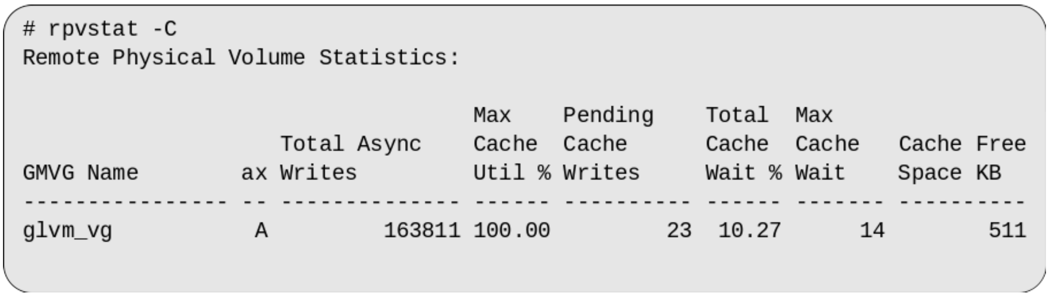
The command `rpvstat -G` will show the number of times the cache fills as well as performance statistics. See Figure 12.



```
rpvstat -G

Remote Physical Volume Statistics:

GMVG name .................................... glvm_vg
AIO total commit time (ms) ................... 7569757
Number of committed groups ................... 25254
Total committed AIO data (KB) ................ 80185994
Average group commit time (ms) ............... 299
AIO data committed per sec (KB) .............. 10000
AIO total complete time (ms) ................. 13268424
Number of completed groups ................... 25246
Total completed AIO data (KB) ................ 80146952
Average group complete time (ms) ............. 525
AIO data completed per sec (KB) .............. 6000
Number of groups read ........................ 7127
Total AIO data read (KB) ..................... 520715
Total AIO cache read time (ms) ............... 113796479
Average group read time (ms) ................. 15966
AIO data read per sec (KB) ................... 0
Number of groups formed ...................... 25255
Total group formation time (ms) .............. 237797
Average group formation time (ms) ............ 9
Number of cache fulls detected ............... 3252
Total cache usage time (ms) .................. 67342170
Total wait time for cache availability (ms) .. 968340
Total AIO write data in transit (KB) ......... 0
```

Figure 12: `rpvstat -G`

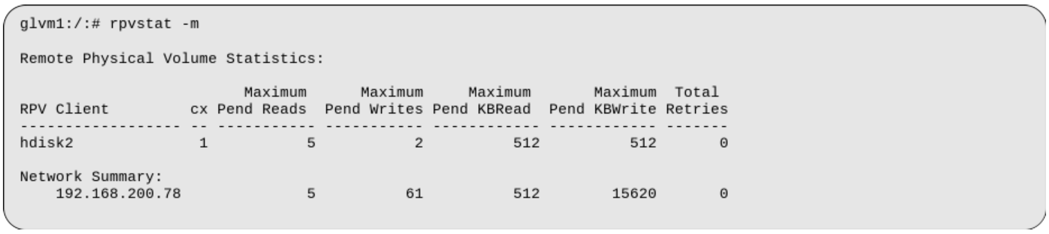
The command `rpvstat -C` will provide details around the number of writes, waits and free space in the cache. See Figure 13.



```
# rpvstat -C
Remote Physical Volume Statistics:

                                   Max    Pending    Total  Max
                    Total Async    Cache  Cache      Cache  Cache   Cache Free
GMVG Name        ax Writes         Util % Writes     Wait % Wait    Space KB
---------------- -- -------------- ------ ---------- ------ ------- ----------
glvm_vg           A         163811 100.00         23  10.27      14        511
```

Figure 13: `rpvstat -C`

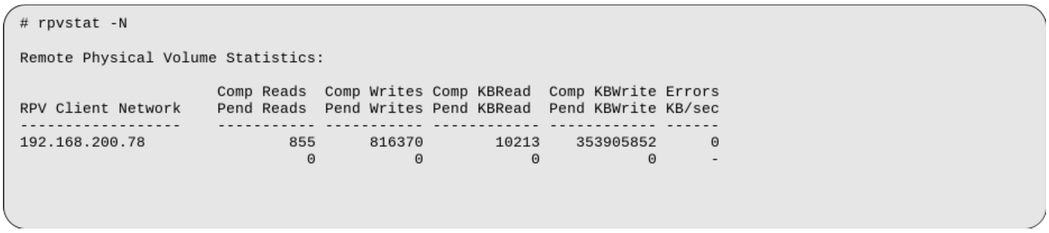
The command `rpvstat -m` will provide details around the number actual and pending reads and writes by client and totals for each network. See Figure 14.



```
glvm1:/:# rpvstat -m

Remote Physical Volume Statistics:

                         Maximum      Maximum     Maximum       Maximum  Total
RPV Client         cx Pend Reads  Pend Writes Pend KBRead  Pend KBWrite Retries
------------------ -- ----------- ----------- ------------ ------------ -------
hdisk2              1           5           2          512          512       0

Network Summary:
    192.168.200.78              5          61          512        15620       0
```

Figure 14: `rpvstat -m`

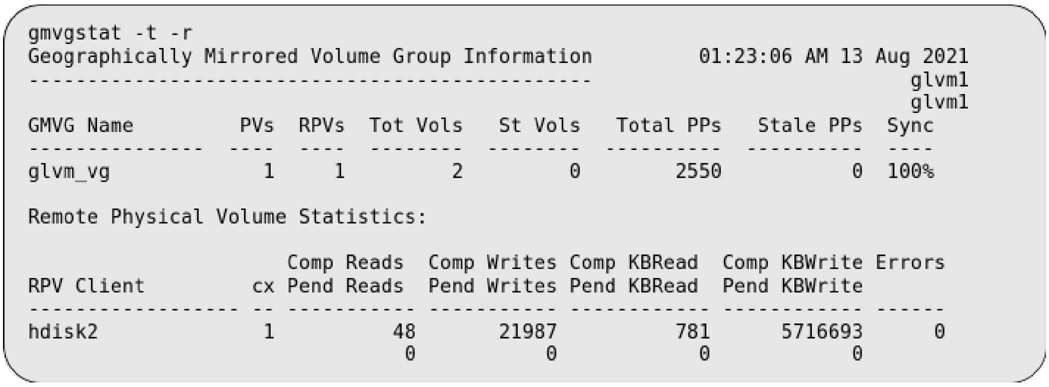
The command `rpvstat -N` will provide read and write details by network. See Figure 15.



```
# rpvstat -N

Remote Physical Volume Statistics:

                      Comp Reads  Comp Writes Comp KBRead  Comp KBWrite Errors
RPV Client Network    Pend Reads  Pend Writes Pend KBRead  Pend KBWrite KB/sec
------------------    ----------- ----------- ------------ ------------ ------
192.168.200.78                855      816370        10213    353905852      0
                                0           0            0            0      -
```

Figure 15: `rpvstat -N`

The command `gmvgstat` provides gmvg and rpv statistics. See Figure 16.



```
gmvgstat -t -r
Geographically Mirrored Volume Group Information         01:23:06 AM 13 Aug 2021
------------------------------------------------                           glvm1
                                                                           glvm1
GMVG Name         PVs  RPVs  Tot Vols   St Vols   Total PPs   Stale PPs  Sync
---------------  ----  ----  --------  --------  ----------  ----------  ----
glvm_vg             1     1         2         0        2550           0  100%

Remote Physical Volume Statistics:

                      Comp Reads  Comp Writes Comp KBRead  Comp KBWrite Errors
RPV Client         cx Pend Reads  Pend Writes Pend KBRead  Pend KBWrite
------------------ -- ----------- ----------- ------------ ------------ ------
hdisk2              1          48       21987          781      5716693      0
                                0           0            0            0
```

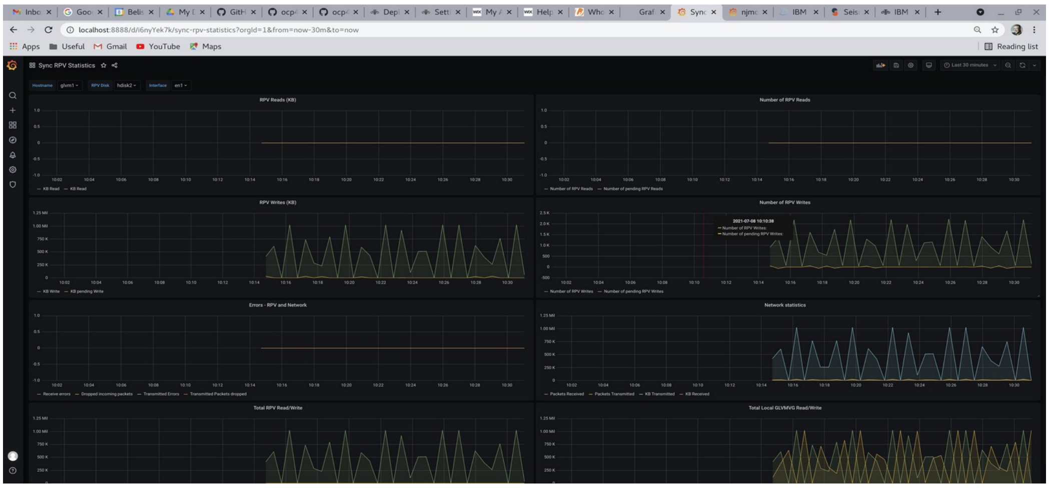
Figure 16: gmvgstat -t -r

Detailed monitoring

Following the IBM Support description for setting up Grafana / InfluxDB, the following panels can be produced. Nigel Griffiths also provides detailed steps to configure displaying nmon data using Grafana / InfluxDB.

For the following panels, I used a script to capture rpvstat data every 30 seconds and then load it in a central influxDB. I configured Grafana to display these values. Please contact me if you are interested in details or a demonstration.

Example of synchronous statistics - Figure 17.



localhost:8888/d/i6nyYek7k/sync-rpv-statistics?orgId=1&from=now-30m&to=now

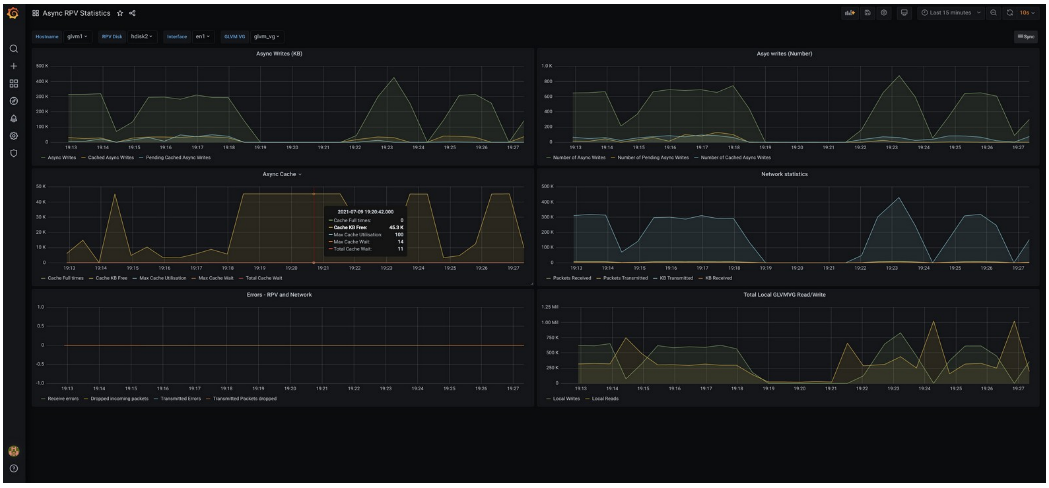
Hostname glvm1

RPV Disk hdisk2

Interface en1

Figure 17: Sync stats shown in Grafana

Example of asynchronous statistics - Figure 18.



Hostname glvm1

RPV Disk hdisk2

Interface en1

GLVM VG glvm_vg

Figure 18: Async stats shown in Grafana

Scenario - 2 sites and one PV at each

After planning network configuration / bandwidth, configuring GLVM is relatively simple. For standalone, the sites need to be configured and then the Servers / Clients configured. PowerHA handles the sites, has an option of multiple networks and the GUI includes a GLVM configuration Wizard. The GMVG is configured on one site using the RPV Client(s). The only difference for asynchronous mode is that mirror pools are compulsory and there needs to be a local `aio_cache` logical volume in each pool.

The following steps outline the steps to configure a simple GLVM configuration with 2 nodes and one disk on each node that will be used to create the GMVG. Initially GLVM will be configured to be synchronous, then changed to be asynchronous.

Setting	Site 1	Site 2
Site name	glvm1	glvm2
Address	192.168.200.138	192.168.200.78
Mirror pool	glvm1	glvm2
PVIDs	00c8d23057b60c26	00c8cf4057f2d781
Volum group	glvm_vg	
jfs log logical volume	glvmlv01	
jfs2 log logical volume	glvmlv02	
aio_cache	glvm2_cache	glvm1_cache
file system	/glvm_data	

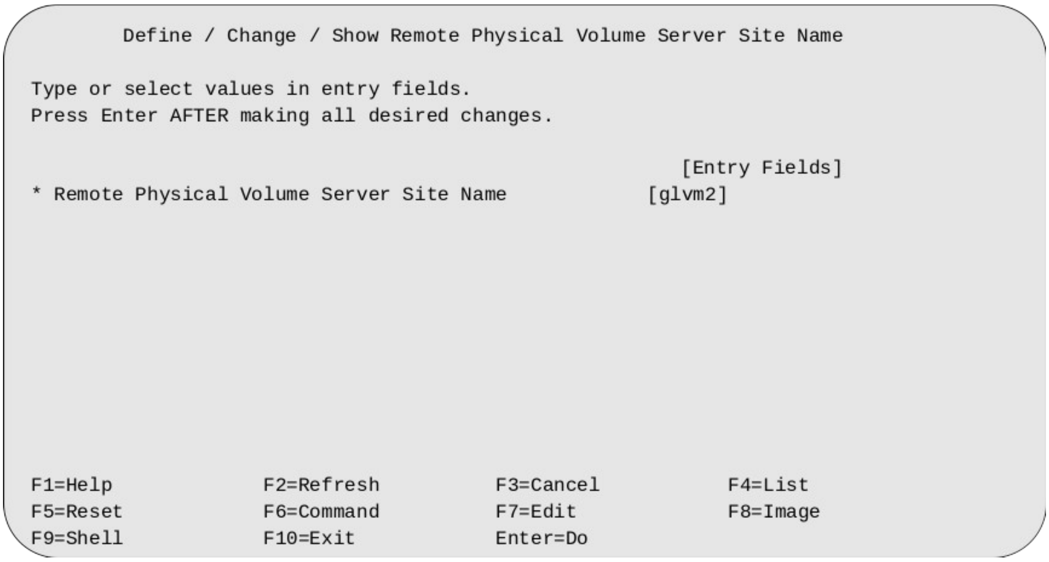
Configuring sites

This can be configure through smit or using

```
/usr/sbin/rpvsitename -a [sitename]
```

For SMIT:

smit glvm_utils > Remote Physical Volume Servers > Remote Physical Volume Server Site Name Configuration > Define / Change / Show Remote Physical Volume Server Site Name. Enter site name. See Figure 19.



```
        Define / Change / Show Remote Physical Volume Server Site Name

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Remote Physical Volume Server Site Name            [glvm2]









F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 19: Set site name

Repeat on the node on the other site.

Creating the RPV Server on glvm2

Again this can be performed using the command line:

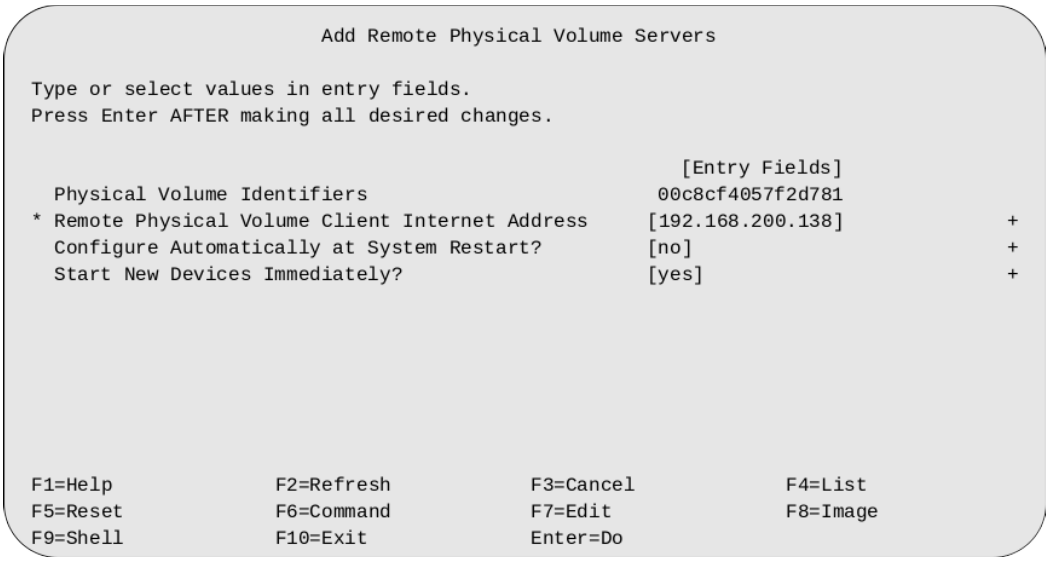
```
/usr/sbin/mkdev -c rpvserver -s rpvserver -t rpvstype \
    -a rpvs_pvid=00c8cf4057f2d781 -a client_addr='192.168.200.138'\
    -a auto_online='n'
```

Command wil respond with:

```
rpvserver0 Available
```

or via SMIT:

smit glvm_utils > Remote Physical Volume Servers > Add Remote Physical Volume Servers > Select the local physical volume (name and pvid listed) > Set “Configure Automatically at System Restart” to no and “Start New Devices Immediately” to yes – See Figure 20.



```
                      Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                    [Entry Fields]
  Physical Volume Identifiers                       00c8cf4057f2d781
* Remote Physical Volume Client Internet Address   [192.168.200.138]        +
  Configure Automatically at System Restart?       [no]                     +
  Start New Devices Immediately?                   [yes]                    +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 20: SMIT create RPV Server

SMIT will respond with rpvserver0 available.

Creating the RPV Client on glvm1

Again this can be performed using the command line:

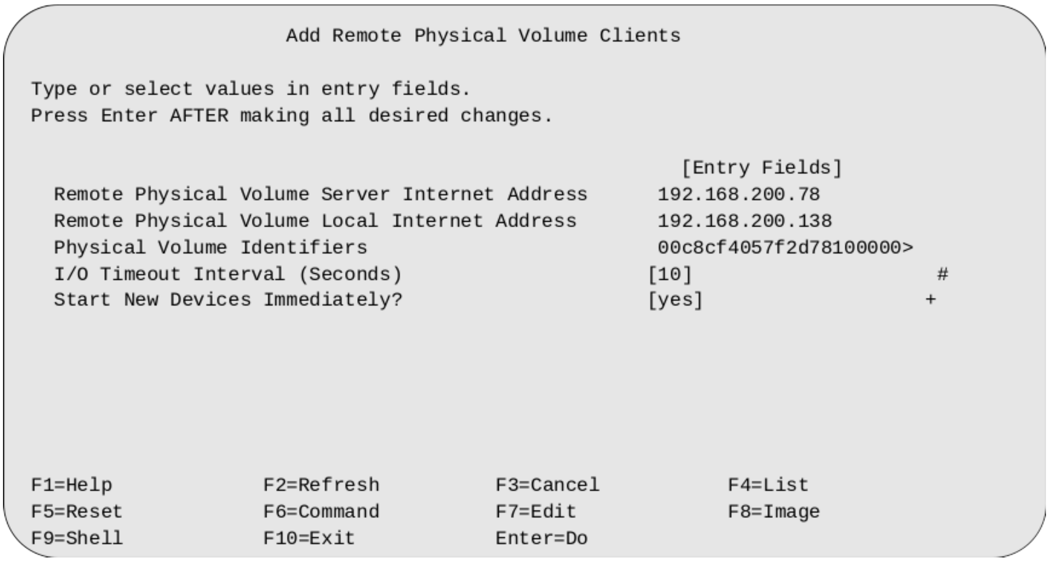
```
/usr/sbin/mkdev -c disk -s remote_disk -t rpvclient \
   -a pvid=00c8cf4057f2d781 -a server_addr='192.168.200.78' \
   -a local_addr='192.168.200.138' -a io_timeout='180'
```

Command wil respond with:

```
hdisk2 Available
```

or via SMIT:

smit glvm_utils > Remote Physical Volume Clients > Select if your mirroring network uses IPv6 > Add the RPV Server IP address > Select the local network address > Select the hdisk on the Server that this client will point to > Select timeout and start now – See Figure 21.



```
                      Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Remote Physical Volume Server Internet Address      192.168.200.78
  Remote Physical Volume Local Internet Address       192.168.200.138
  Physical Volume Identifiers                         00c8cf4057f2d78100000>
  I/O Timeout Interval (Seconds)                     [10]                    #
  Start New Devices Immediately?                     [yes]                  +





F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

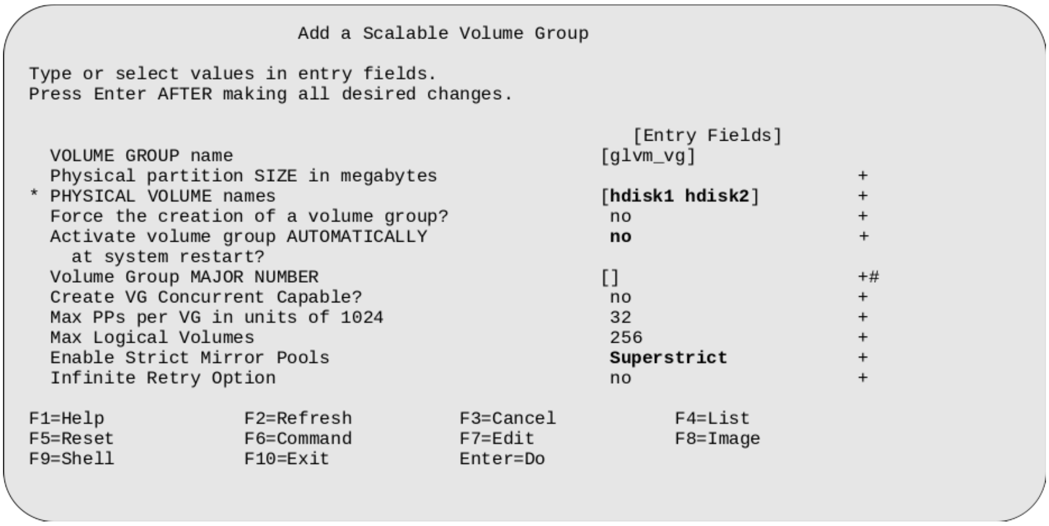
Figure 21: Create RPV Client

SMIT will respond with hdisk2 available.

Creating the GMVG

Before we configure the RPV Server(s) / Client(s) for replication from Server glvm2 to glmv1, we create the GMVG on glvm1 using the local physical disk and the active RPV Client.

First step is to create a Scalable Volume group using these two hdisks and setting Superstrict – see Figure 22.



```
                         Add a Scalable Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  VOLUME GROUP name                                  [glvm_vg]
  Physical partition SIZE in megabytes                                 +
* PHYSICAL VOLUME names                              [hdisk1 hdisk2]    +
  Force the creation of a volume group?               no               +
  Activate volume group AUTOMATICALLY                 no               +
    at system restart?
  Volume Group MAJOR NUMBER                          []                +#
  Create VG Concurrent Capable?                       no               +
  Max PPs per VG in units of 1024                     32               +
  Max Logical Volumes                                 256              +
  Enable Strict Mirror Pools                          Superstrict      +
  Infinite Retry Option                               no               +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 22: Creating GMVG

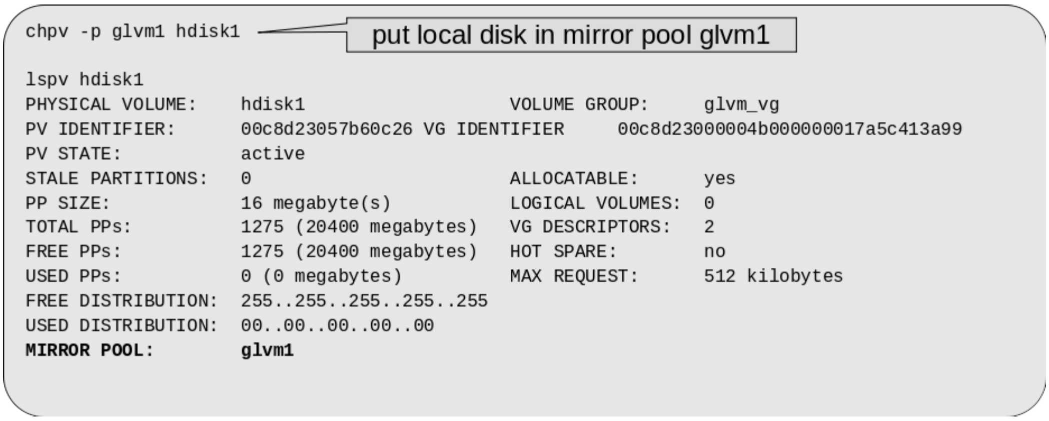
Now stop bad block relocation for the volume group:

```
chvg -b n glvm_vg
```

Add disks to mirror pools at each site

```
chpv -p glvm1 hdisk1
chpv -p glvm2 hdisk2
```

For example see Figure 23

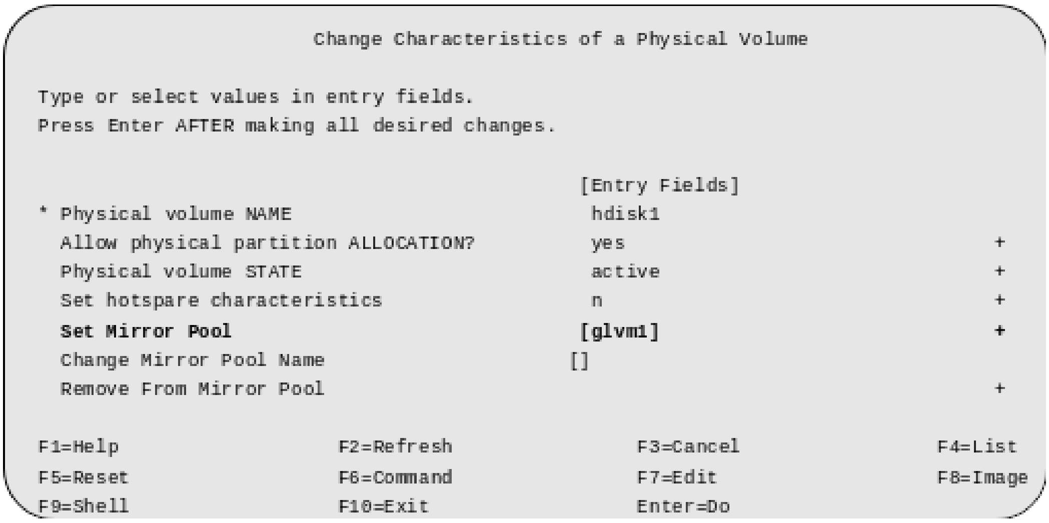


```
chpv -p glvm1 hdisk1

lspv hdisk1
PHYSICAL VOLUME:    hdisk1                   VOLUME GROUP:     glvm_vg
PV IDENTIFIER:      00c8d23057b60c26 VG IDENTIFIER     00c8d23000004b000000017a5c413a99
PV STATE:           active
STALE PARTITIONS:   0                        ALLOCATABLE:      yes
PP SIZE:            16 megabyte(s)           LOGICAL VOLUMES:  0
TOTAL PPs:          1275 (20400 megabytes)   VG DESCRIPTORS:   2
FREE PPs:           1275 (20400 megabytes)   HOT SPARE:        no
USED PPs:           0 (0 megabytes)          MAX REQUEST:      512 kilobytes
FREE DISTRIBUTION:  255..255..255..255..255
USED DISTRIBUTION:  00..00..00..00..00
MIRROR POOL:        glvm1
```

put local disk in mirror pool glvm1

Figure 23: hdisk1 in mirror pool glvm1

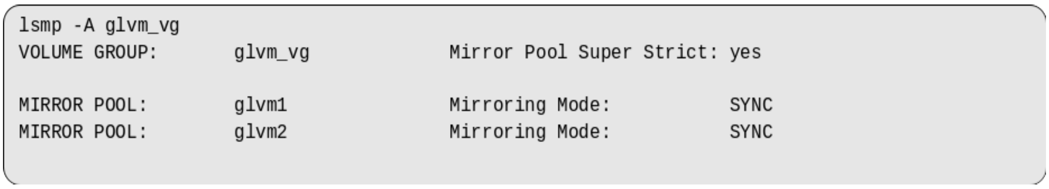
or via SMIT – Add the physical volume(s) at each site to a unique mirror pool. See Figure 24.



```
                     Change Characteristics of a Physical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                               [Entry Fields]
* Physical volume NAME                          hdisk1
  Allow physical partition ALLOCATION?          yes                                +
  Physical volume STATE                         active                             +
  Set hotspare characteristics                  n                                  +
  Set Mirror Pool                              [glvm1]                             +
  Change Mirror Pool Name                      []
  Remove From Mirror Pool                                                          +

F1=Help                 F2=Refresh              F3=Cancel               F4=List
F5=Reset                F6=Command              F7=Edit                 F8=Image
F9=Shell                F10=Exit                Enter=Do
```

Figure 24: Assign physical volume to local mirror pool

See Figure 25 for example of the resulting disk configuration.

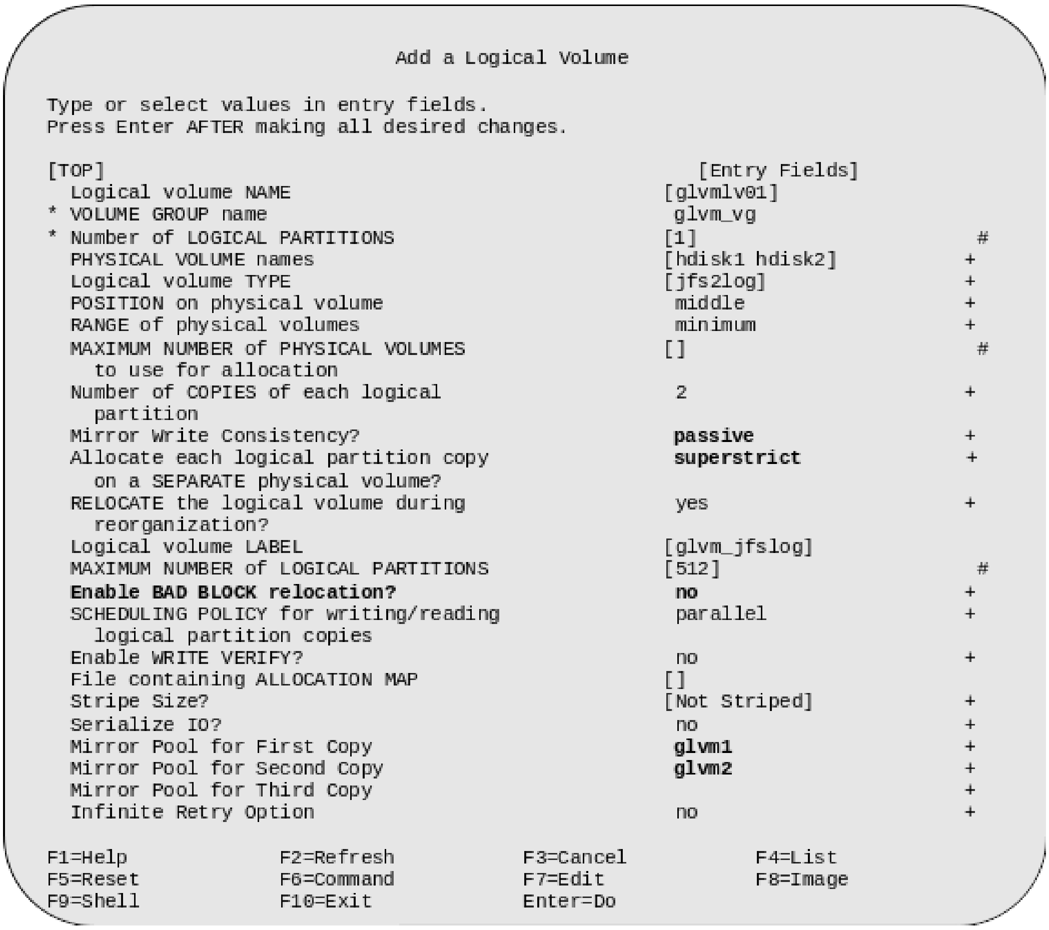


```
lsmp -A glvm_vg
VOLUME GROUP:       glvm_vg             Mirror Pool Super Strict: yes

MIRROR POOL:        glvm1               Mirroring Mode:           SYNC
MIRROR POOL:        glvm2               Mirroring Mode:           SYNC
```

Figure 25: lsmp showing Mirror Pool configuration

Creating logical volumes for synchronous replication

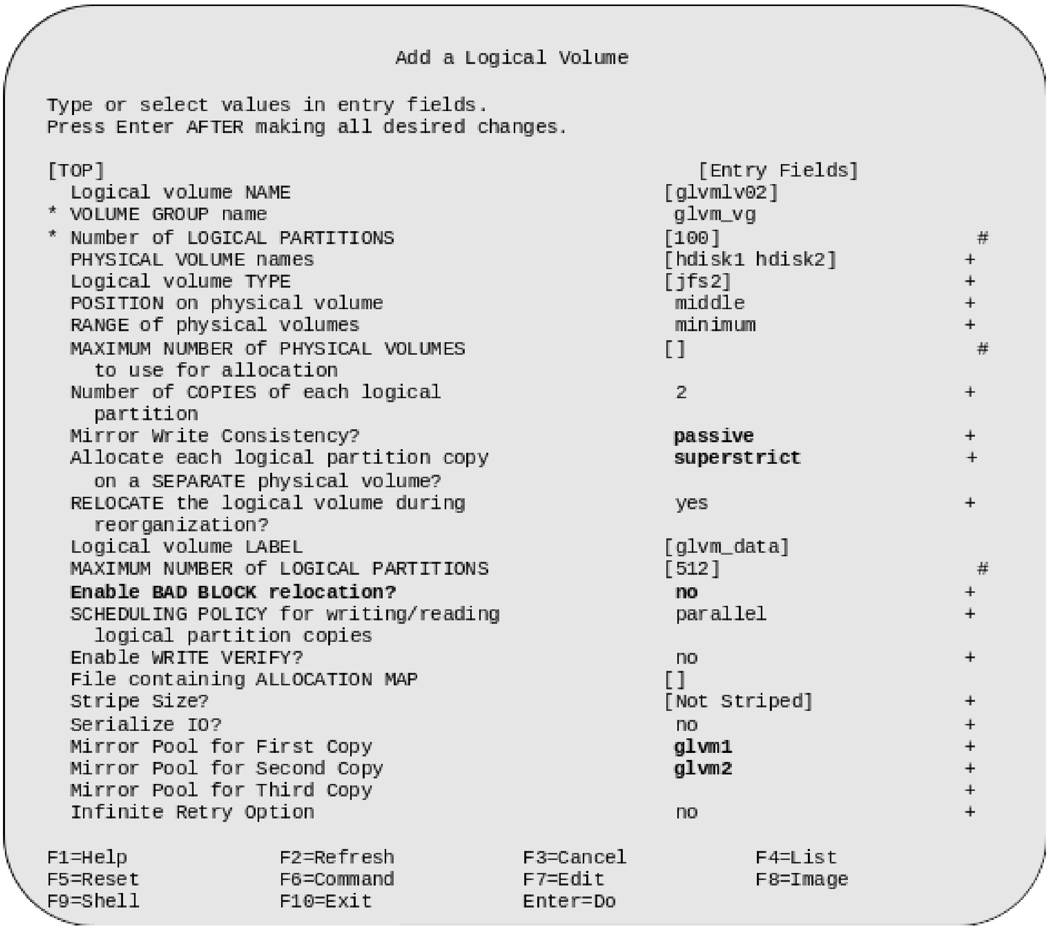
The logical volumes are now created, using both disks, superstrict, passive MWC and define the mirror pool for each copy. See Figure 26.



```
                          Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                 [glvmlv01]
* VOLUME GROUP name                                    glvm_vg
* Number of LOGICAL PARTITIONS                        [1]                 #
  PHYSICAL VOLUME names                               [hdisk1 hdisk2]    +
  Logical volume TYPE                                 [jfs2log]          +
  POSITION on physical volume                          middle            +
  RANGE of physical volumes                            minimum           +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                  []                  #
    to use for allocation
  Number of COPIES of each logical                     2                 +
    partition
  Mirror Write Consistency?                            passive           +
  Allocate each logical partition copy                 superstrict       +
    on a SEPARATE physical volume?
  RELOCATE the logical volume during                   yes               +
    reorganization?
  Logical volume LABEL                                [glvm_jfslog]
  MAXIMUM NUMBER of LOGICAL PARTITIONS                [512]               #
  Enable BAD BLOCK relocation?                         no                +
  SCHEDULING POLICY for writing/reading                parallel          +
    logical partition copies
  Enable WRITE VERIFY?                                 no                +
  File containing ALLOCATION MAP                      []
  Stripe Size?                                        [Not Striped]      +
  Serialize IO?                                        no                +
  Mirror Pool for First Copy                           glvm1             +
  Mirror Pool for Second Copy                          glvm2             +
  Mirror Pool for Third Copy                                             +
  Infinite Retry Option                                no                +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 26: Create jfs2log logical volume

As with the jfs2 log, create the logical volume for the file system. See Figure 27.



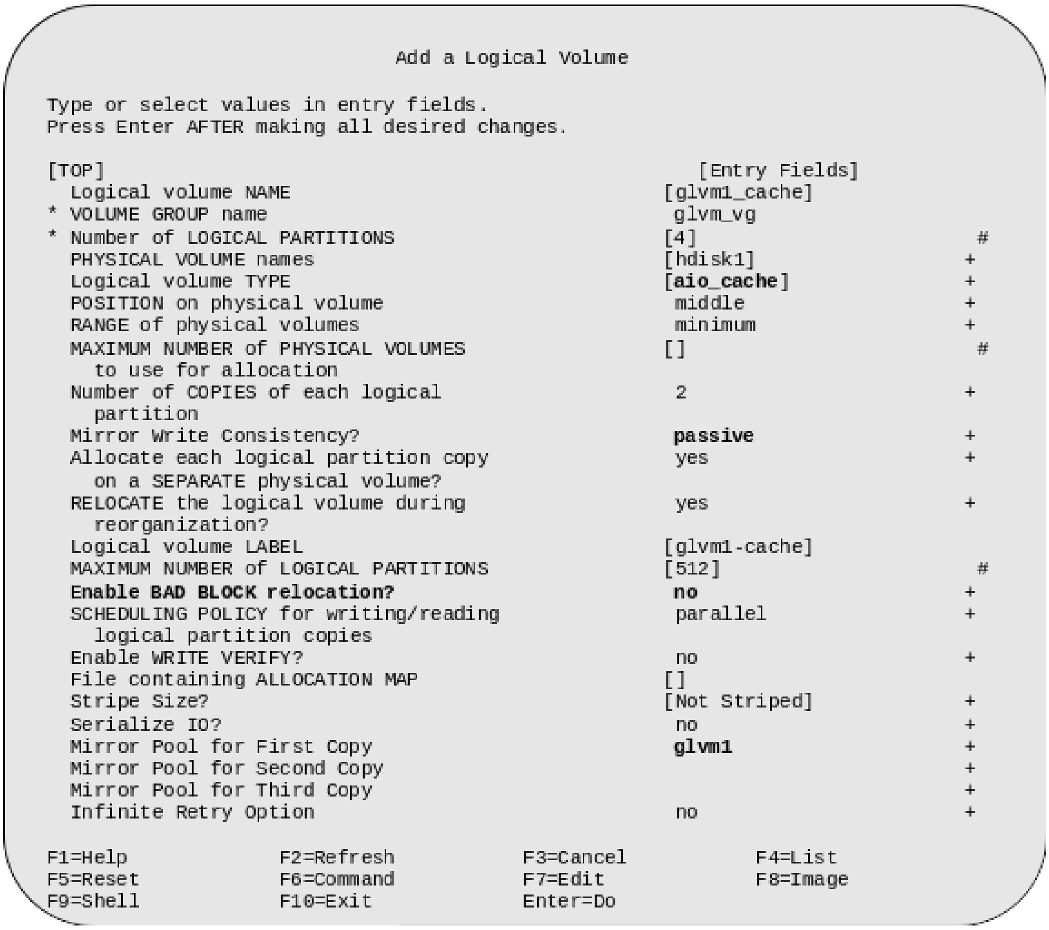
```
                          Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                [glvmlv02]
* VOLUME GROUP name                                   glvm_vg
* Number of LOGICAL PARTITIONS                       [100]                  #
  PHYSICAL VOLUME names                              [hdisk1 hdisk2]       +
  Logical volume TYPE                                [jfs2]                +
  POSITION on physical volume                         middle               +
  RANGE of physical volumes                           minimum              +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                 []                     #
    to use for allocation
  Number of COPIES of each logical                    2                    +
    partition
  Mirror Write Consistency?                           passive              +
  Allocate each logical partition copy                superstrict          +
    on a SEPARATE physical volume?
  RELOCATE the logical volume during                  yes                  +
    reorganization?
  Logical volume LABEL                               [glvm_data]
  MAXIMUM NUMBER of LOGICAL PARTITIONS               [512]                  #
  Enable BAD BLOCK relocation?                        no                   +
  SCHEDULING POLICY for writing/reading               parallel             +
    logical partition copies
  Enable WRITE VERIFY?                                no                   +
  File containing ALLOCATION MAP                     []
  Stripe Size?                                       [Not Striped]         +
  Serialize IO?                                       no                   +
  Mirror Pool for First Copy                          glvm1                +
  Mirror Pool for Second Copy                         glvm2                +
  Mirror Pool for Third Copy                                               +
  Infinite Retry Option                               no                   +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 27: Create data logical volume

Now a file system can be created using the logical volumes and GLVM will be able to replicate the data between the two sites. Create /glvm_data using the jfs log (glvmlv01) and the logical volume (glvmlv02).

Creating logical volumes for asynchronous replication

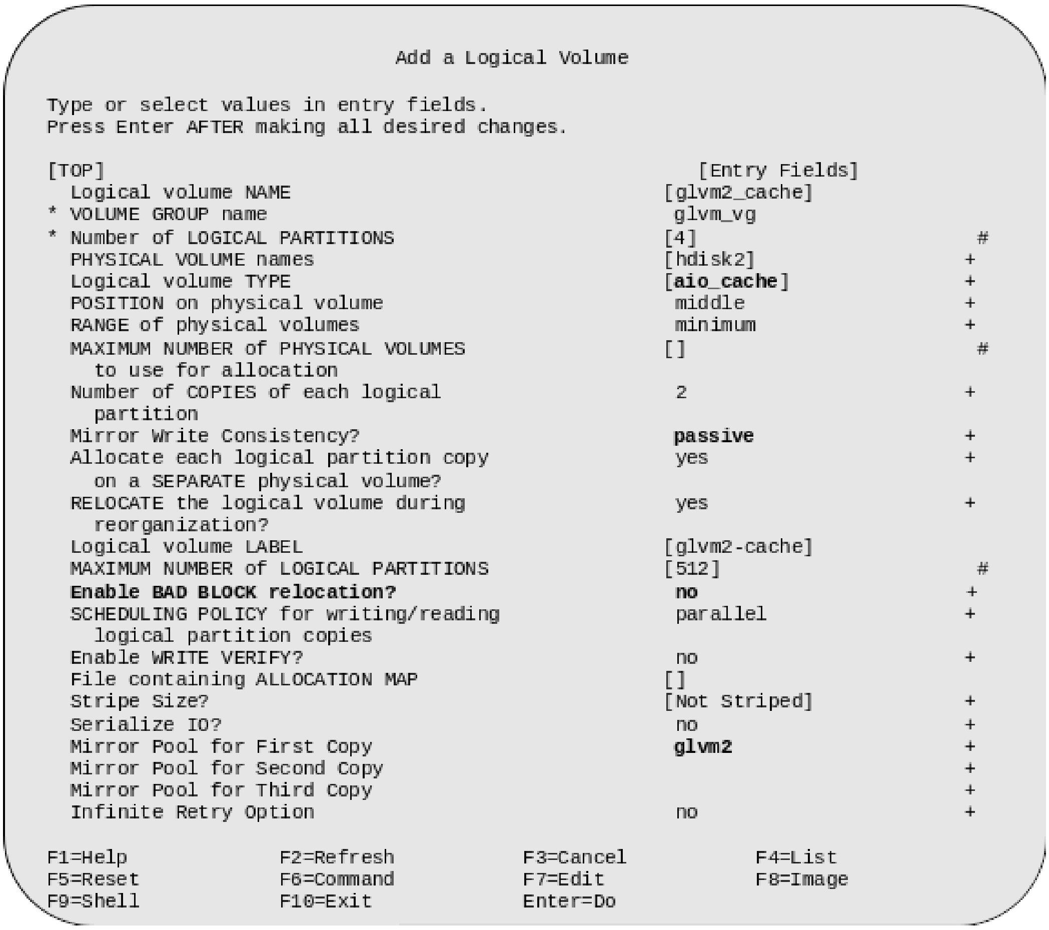
For asynchronous mode, a cache logical volume needs to be created for each pool. For site glvm1 see Figure 28.



```
                              Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                    [Entry Fields]
  Logical volume NAME                                 [glvm1_cache]
* VOLUME GROUP name                                    glvm_vg
* Number of LOGICAL PARTITIONS                        [4]                        #
  PHYSICAL VOLUME names                               [hdisk1]                  +
  Logical volume TYPE                                 [aio_cache]               +
  POSITION on physical volume                          middle                   +
  RANGE of physical volumes                            minimum                  +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                  []                         #
    to use for allocation
  Number of COPIES of each logical                     2                        +
    partition
  Mirror Write Consistency?                            passive                  +
  Allocate each logical partition copy                 yes                      +
    on a SEPARATE physical volume?
  RELOCATE the logical volume during                   yes                      +
    reorganization?
  Logical volume LABEL                                [glvm1-cache]
  MAXIMUM NUMBER of LOGICAL PARTITIONS                [512]                      #
  Enable BAD BLOCK relocation?                         no                       +
  SCHEDULING POLICY for writing/reading                parallel                 +
    logical partition copies
  Enable WRITE VERIFY?                                 no                       +
  File containing ALLOCATION MAP                      []
  Stripe Size?                                        [Not Striped]             +
  Serialize IO?                                        no                       +
  Mirror Pool for First Copy                           glvm1                    +
  Mirror Pool for Second Copy                                                   +
  Mirror Pool for Third Copy                                                    +
  Infinite Retry Option                                no                       +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 28: Creating cache logical volume for site glvm1

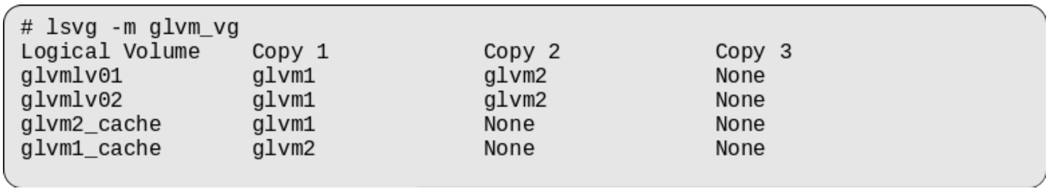
For site glvm2 see Figure 29



```
                          Add a Logical Volume

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                 [glvm2_cache]
* VOLUME GROUP name                                    glvm_vg
* Number of LOGICAL PARTITIONS                        [4]               #
  PHYSICAL VOLUME names                               [hdisk2]          +
  Logical volume TYPE                                 [aio_cache]       +
  POSITION on physical volume                          middle           +
  RANGE of physical volumes                            minimum          +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                  []                #
    to use for allocation
  Number of COPIES of each logical                     2                +
    partition
  Mirror Write Consistency?                            passive          +
  Allocate each logical partition copy                 yes              +
    on a SEPARATE physical volume?
  RELOCATE the logical volume during                   yes              +
    reorganization?
  Logical volume LABEL                                [glvm2-cache]
  MAXIMUM NUMBER of LOGICAL PARTITIONS                [512]             #
  Enable BAD BLOCK relocation?                         no               +
  SCHEDULING POLICY for writing/reading                parallel         +
    logical partition copies
  Enable WRITE VERIFY?                                 no               +
  File containing ALLOCATION MAP                      []
  Stripe Size?                                        [Not Striped]     +
  Serialize IO?                                        no               +
  Mirror Pool for First Copy                           glvm2            +
  Mirror Pool for Second Copy                                           +
  Mirror Pool for Third Copy                                            +
  Infinite Retry Option                                no               +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 29: Creating cache logical volume for site glvm2

See Figure 30 for the resulting disk configuration.



```
# lsvg -m glvm_vg
Logical Volume     Copy 1             Copy 2             Copy 3
glvmlv01           glvm1              glvm2              None
glvmlv02           glvm1              glvm2              None
glvm2_cache        glvm1              None               None
glvm1_cache        glvm2              None               None
```

Figure 30: lsvg showing the VG and mirror pool

The configuration now consists of the rpvserver created on one site and the rpvclient on the other. The next step is to stop both the RPV server and then the RPV client, and create the opposite pair following the same steps as above.

To stop the rpvclient:

```
rmdev -l hdisk2
```

To stop the rpvserver:

```
rmdev -l rpvserver0
```

Next create the RPV Server on glvm1 and then the corresponding RPV Client on glvm2. Once these are available, the GMVG glvm_vg can be imported on node glvm2 and we have completed the configuration of a synchronous geographic mirrored volume group.

To change the Mirror Pool as asynchronous

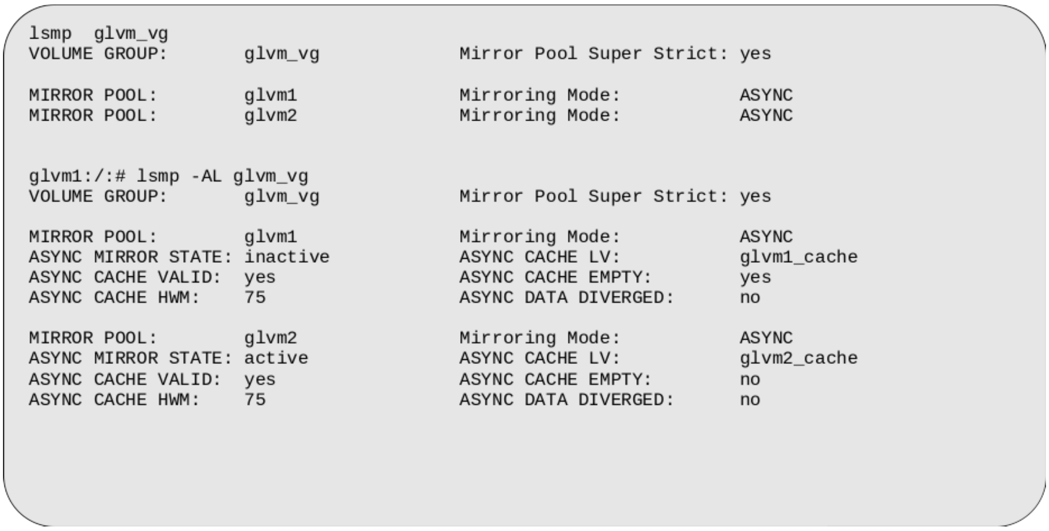
```
/usr/sbin/chmp -A  -m'glvm1' -c'glvm1_cache' -h'75' glvm_vg
/usr/sbin/chmp -A  -m'glvm2' -c'glvm2_cache' -h'75' glvm_vg
```

or via SMIT:

smit glvm_utils > Geographically Mirrored Volume Groups > Manage Geographically Mirrored Volume Groups with Mirror Pools > Configure Mirroring Properties of a Mirror Pool > Convert to Asynchronous Mirroring for a Mirror Pool > Select the mirror pool > Select the LV cache > set the high water mark for the cache (%).

Repeat for the other Mirror Pool.

Listing the status of the volume group will now show it as ASYNC – see Figure 31.



```
lsmp  glvm_vg
VOLUME GROUP:       glvm_vg           Mirror Pool Super Strict: yes

MIRROR POOL:        glvm1             Mirroring Mode:           ASYNC
MIRROR POOL:        glvm2             Mirroring Mode:           ASYNC


glvm1:/:# lsmp -AL glvm_vg
VOLUME GROUP:       glvm_vg           Mirror Pool Super Strict: yes

MIRROR POOL:        glvm1             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive          ASYNC CACHE LV:           glvm1_cache
ASYNC CACHE VALID:  yes               ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    75                ASYNC DATA DIVERGED:      no

MIRROR POOL:        glvm2             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active            ASYNC CACHE LV:           glvm2_cache
ASYNC CACHE VALID:  yes               ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    75                ASYNC DATA DIVERGED:      no
```

Figure 31: Disk showing details as ASYNC

To revert back to Synchronous mode

```
/usr/sbin/chmp -S  -m'glvm1' glvm_vg
```

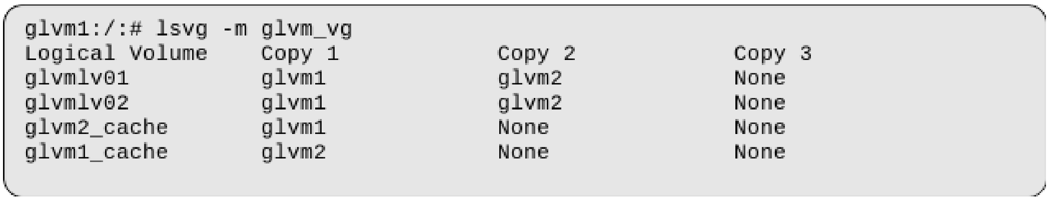
```
/usr/sbin/chmp -S  -m'glvm2' glvm_vg
```

or via SMIT:

smit glvm_utils > Geographically Mirrored Volume Groups > Manage Geographically Mirrored Volume Groups with Mirror Pools > Configure Mirroring Properties of a Mirror Pool > Convert to Synchronous Mirroring for a Mirror Pool > Select the mirror Pool Name > Confirm the details displayed.
Repeat for the other Mirror Pool.

Verification of GLVM configuration

To check the Mirror Pool configuration (`lsvg -m glvm_vg`) - See Figure 32.



```
glvm1:/:# lsvg -m glvm_vg
Logical Volume     Copy 1           Copy 2           Copy 3
glvmlv01           glvm1            glvm2            None
glvmlv02           glvm1            glvm2            None
glvm2_cache        glvm1            None             None
glvm1_cache        glvm2            None             None
```

Figure 32: Volume Group / Mirror Pool configuration

To check the GLVM configuration:

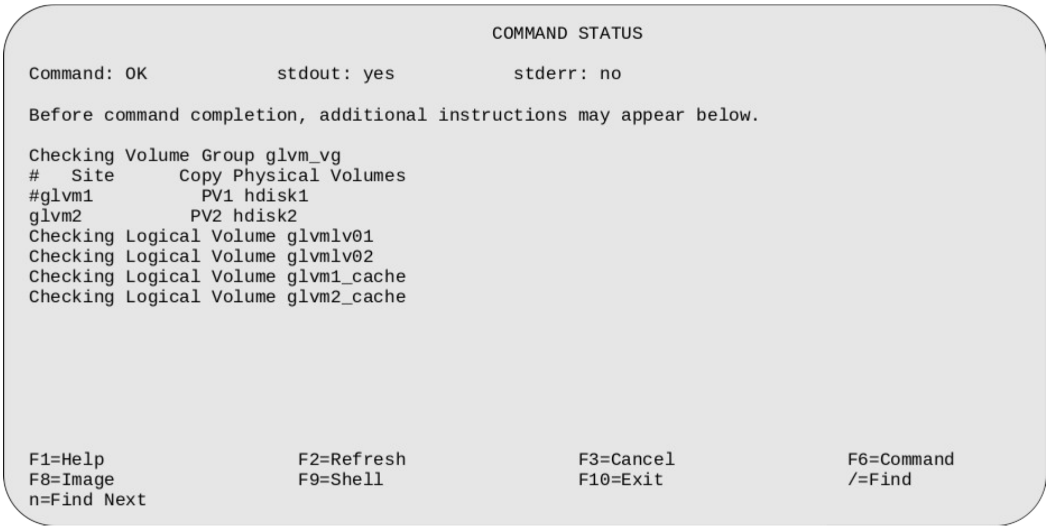
```
/usr/sbin/ckglvm -V 'glvm_vg'
```

or

```
/usr/sbin/lsglvm -c
```

or via SMIT:

smit glvm_utils > Geographically Mirrored Volume Groups > Verify Mirror Copy Site Locations for a Volume Group > Choose the Volume Group – See Figure 33.

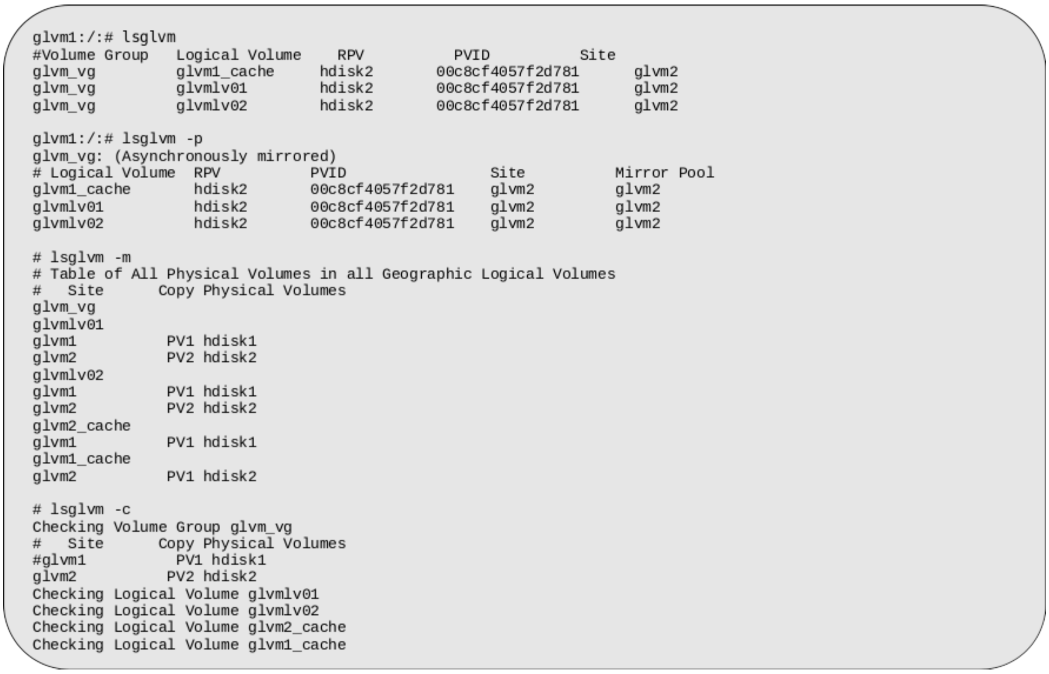


```
                                    COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Checking Volume Group glvm_vg
#   Site      Copy Physical Volumes
#glvm1          PV1 hdisk1
glvm2          PV2 hdisk2
Checking Logical Volume glvmlv01
Checking Logical Volume glvmlv02
Checking Logical Volume glvm1_cache
Checking Logical Volume glvm2_cache







F1=Help                 F2=Refresh              F3=Cancel               F6=Command
F8=Image                F9=Shell                F10=Exit                /=Find
n=Find Next
```

Figure 33: Verifying configuration

To display the configuration:

The lsglvm command has a number of useful flags – see Figure 34.



```
glvm1:/:# lsglvm
#Volume Group   Logical Volume    RPV          PVID          Site
glvm_vg         glvm1_cache     hdisk2       00c8cf4057f2d781      glvm2
glvm_vg         glvmlv01        hdisk2       00c8cf4057f2d781      glvm2
glvm_vg         glvmlv02        hdisk2       00c8cf4057f2d781      glvm2

glvm1:/:# lsglvm -p
glvm_vg: (Asynchronously mirrored)
# Logical Volume  RPV          PVID                Site          Mirror Pool
glvm1_cache       hdisk2       00c8cf4057f2d781    glvm2         glvm2
glvmlv01          hdisk2       00c8cf4057f2d781    glvm2         glvm2
glvmlv02          hdisk2       00c8cf4057f2d781    glvm2         glvm2

# lsglvm -m
# Table of All Physical Volumes in all Geographic Logical Volumes
#   Site      Copy Physical Volumes
glvm_vg
glvmlv01
glvm1          PV1 hdisk1
glvm2          PV2 hdisk2
glvmlv02
glvm1          PV1 hdisk1
glvm2          PV2 hdisk2
glvm2_cache
glvm1          PV1 hdisk1
glvm1_cache
glvm2          PV1 hdisk2

# lsglvm -c
Checking Volume Group glvm_vg
#   Site      Copy Physical Volumes
#glvm1          PV1 hdisk1
glvm2          PV2 hdisk2
Checking Logical Volume glvmlv01
Checking Logical Volume glvmlv02
Checking Logical Volume glvm2_cache
Checking Logical Volume glvm1_cache
```

Figure 34: lsglvm options

Maintenance tasks

The following sections covers some of the common GLVM maintenance tasks and assumes that neither the RPV Servers or RPV Clients are active.

Setting preferred read to local disks for standalone

As mentioned above, PowerHA has the ability to set preferred read to “siteaffinity”, so all reads use local disks if they are available. To set the preferred read for each logical volume:

```
chlv -R n <LV_Name>
```

where: n = copy number for the mirror pool (as shown in lslv Figure 35)

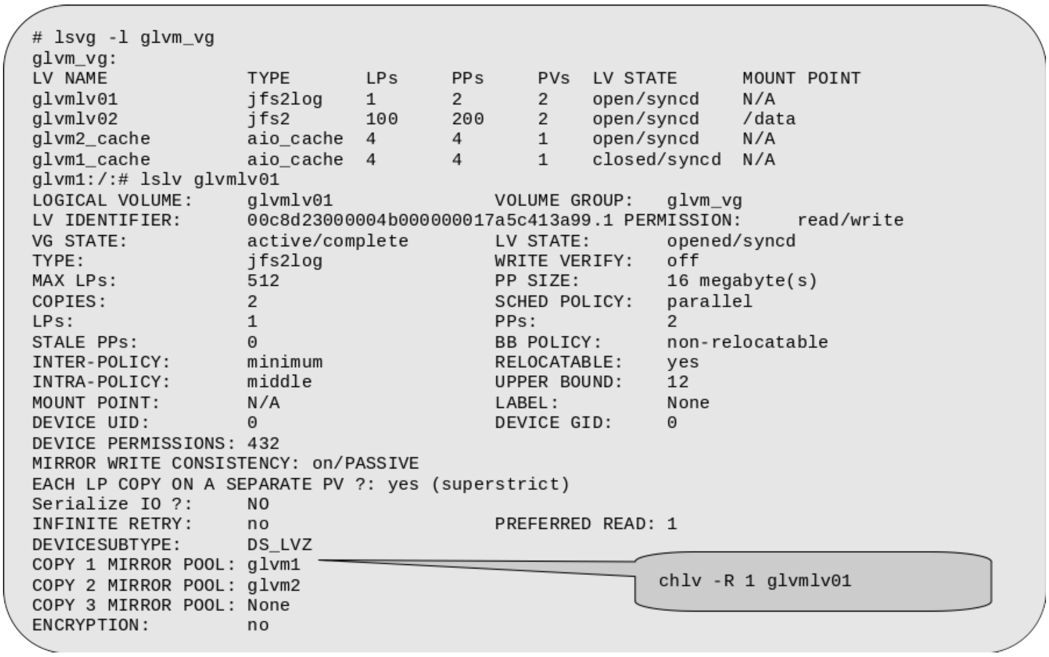
For example:

```
chlv -R 1 glvmlv01
```

See Figure 35 for details of the lslv output and the changed preferred read. To turn this feature off:

```
chlv -R 0 glvmlv02
```

> Note: This must be set when you change site, else all reads will be done from the remote logical volumes



```
# lsvg -l glvm_vg
glvm_vg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
glvmlv01            jfs2log    1       2       2    open/syncd    N/A
glvmlv02            jfs2       100     200     2    open/syncd    /data
glvm2_cache         aio_cache  4       4       1    open/syncd    N/A
glvm1_cache         aio_cache  4       4       1    closed/syncd  N/A
glvm1:/:# lslv glvmlv01
LOGICAL VOLUME:     glvmlv01               VOLUME GROUP:   glvm_vg
LV IDENTIFIER:      00c8d23000004b000000017a5c413a99.1 PERMISSION:     read/write
VG STATE:           active/complete        LV STATE:       opened/syncd
TYPE:               jfs2log                WRITE VERIFY:   off
MAX LPs:            512                    PP SIZE:        16 megabyte(s)
COPIES:             2                      SCHED POLICY:   parallel
LPs:                1                      PPs:            2
STALE PPs:          0                      BB POLICY:      non-relocatable
INTER-POLICY:       minimum                RELOCATABLE:    yes
INTRA-POLICY:       middle                 UPPER BOUND:    12
MOUNT POINT:        N/A                    LABEL:          None
DEVICE UID:         0                      DEVICE GID:     0
DEVICE PERMISSIONS: 432
MIRROR WRITE CONSISTENCY: on/PASSIVE
EACH LP COPY ON A SEPARATE PV ?: yes (superstrict)
Serialize IO ?:     NO
INFINITE RETRY:     no                     PREFERRED READ: 1
DEVICESUBTYPE:      DS_LVZ
COPY 1 MIRROR POOL: glvm1
COPY 2 MIRROR POOL: glvm2
COPY 3 MIRROR POOL: None
ENCRYPTION:         no
```

```
chlv -R 1 glvmlv01
```

Figure 35: Set preferred read

Starting the GMVG and mounting the file system(s)

The following is the recommended way of activating GLVM

1. Start the RPV Server on the remote Server
   ```
   mkdev -l rpvserver0
   ```
2. Start the local RPV Client

```
mkdev -l hdisk2
```

3. Activate the Volume Group
   ```
   varyonvg glvm_vg
   ```
4. Mount the file system
   ```
   mount /data
   ```
5. Start monitoring!

Un-mounting the file system(s) and stopping the GMVG

To stop GLVM:

1. un-mount the filesystem
   ```
   umount /data
   ```
2. Deactivate the Volume Group
   ```
   varyoffvg glvm_vg
   ```
3. Wait while any outstanding I/O is synchronised with remote
   Check activity with `rpvstat` command
4. Stop the RPV Client
   ```
   rmdev -l hdisk2
   ```
5. On the remote server, stop the RPV Server
   ```
   rmdev -l rpvserver0
   ```

Reversing the flow

To reverse the flow, follow the steps above to deactivate the GLVM on the running site and then start the RPV Server on the new remote site and RPV Client on the new active site.

Note: If you have set preferred read, you must change this to the local mirror pool.

Recovering from a failed cache LV

Should there be any problem with the Cache LV, such as read or write failues (check the error report). the aio_cache LV will be marked as invalid. This will stop GLVM writing to the cache, so it will mark all the remote partitions as stale.
In this scenario, the following action is recommended:

1. Convert the mode from asynchronous to synchronous, where glvm1 is the local Mirror Pool. The -f flag will force the conversion, even if the aio_cache is not available.
   chmp -S -f -m glvm1 glvm_vg
2. Synchronise the remote copy, by reactivating the RPV Client and resuming communications with the RPV Server
   ```
   chdev -l hdisk2 -a resume=yes
   ```
3. Get the LVM to check that the disk is no longer unavailable
   varyonvg glvm_vg
4. Resolve the problem with the aio_cache LV, or create a new Cache LV
5. Convert back to asynchronous mode, were again glvm1 is the local Mirror Pool.
   ```
   /usr/sbin/chmp -A  -m'glvm1' -c'glvm1_cache' -h'75' glvm_vg
   ```

Recovering on the same site (asynchronous mode)

It is important to note that if the active site fails and you decide to recover on the same site, rather than moving the application to the secondary site, there will be special handling of the data in the

AIO cache. When the GMVG is activated, the updates recorded in the cache will first be played against the remote LUN(s). Thus no local writes will be allowed until the local cache is drained. Thus extra time must be added to the recovery plan to allow for cache recovery.

Appendix 1 - References

Nigel Griffiths using Grafana / InfluxDB to capture and monitor nmon performance data
http://nmon.sourceforge.net/pmwiki.php?n=Site.Njmon

IBM Support steps to install InfluxDB and Grafana
https://www.ibm.com/support/pages/aix-installing-influxdb-18-and-grafana-7

Asynchronous Geographic Logical Volume Mirroring (GLVM) - Best Practices for Cloud deployments - White paper available in PowerVS seismic

Numerous Tech U presentations by Red.

Sensible advise and improvements from Shawn Bodily

The IBM documentation has a full set of documentation for GLVM (under PowerHA), but does mention the standalone configuraiton.
https://www.ibm.com/docs/en/powerha-aix/7.2?topic=edition-planning

Appendix 2 – Standalone code installed

Code to install:

```
        glvm.rpv.client              7.2.5.0
        glvm.rpv.server              7.2.5.0
        glvm.rpv.util                7.2.5.0
```

co-requisites installed:

```
        bos.msg.en_US.alt_disk_install.rte 7.2.5.0  # Alternate Disk Install Msgs ...
        bos.msg.en_US.diag.rte 7.2.5.0              # Hardware Diagnostics Message...
        bos.msg.en_US.net.ipsec 7.2.5.0             # IP Security Messages - U.S. ...
        bos.msg.en_US.net.tcp.client 7.2.5.0        # TCP/IP Messages - U.S. English
        bos.msg.en_US.rte 7.2.5.0                   # Base OS Runtime Messages - U...
        bos.msg.en_US.txt.tfs 7.2.5.0               # Text Formatting Services Msg...
        devices.msg.en_US.base.com 7.2.5.0          # Base Sys Device Software Msg...
        devices.msg.en_US.diag.rte 7.2.5.0          # Device Diagnostics Messages ...
        devices.msg.en_US.rspc.base.com 7.2.5.0     # RISC PC Software Messages - ...
        devices.msg.en_US.sys.mca.rte 7.2.5.0       # Micro Channel Bus Software M...
        glvm.rpv.client 7.2.5.0                     # Remote Physical Volume Client
        glvm.rpv.server 7.2.5.0                     # Remote Physical Volume Server
        glvm.rpv.util 7.2.5.0                       # Geographic LVM Utilities
        invscout.msg.en_US.rte 2.1.0.2              # Inventory Scout Messages - U...
        openssh.msg.en_US 8.1.102.2101              # Open Secure Shell Messages -...
        printers.msg.en_US.rte 7.2.0.0              # Printer Backend Messages - U...
        rsct.msg.en_US.basic.rte 3.2.6.0            # RSCT Basic Msgs - U.S. English
        rsct.msg.en_US.core.auditrm 3.2.6.0         # RSCT Audit Log RM Msgs - U.S...
        rsct.msg.en_US.core.errm 3.2.6.0            # RSCT Event Response RM Msgs ...
        rsct.msg.en_US.core.fsrm 3.2.6.0            # RSCT File System RM Msgs - U...
        rsct.msg.en_US.core.gui 3.2.6.0             # RSCT GUI Msgs - U.S. English
        rsct.msg.en_US.core.gui.com 3.2.6.0         # RSCT GUI JAVA Msgs - U.S. En...
        rsct.msg.en_US.core.hostrm 3.2.6.0          # RSCT Host RM Msgs - U.S. Eng...
        rsct.msg.en_US.core.lprm 3.2.6.0            # RSCT LPRM Msgs - U.S. English
        rsct.msg.en_US.core.microsensorrm 3.2.6.0   # RSCT MicorSensor RM Msgs - U...
        rsct.msg.en_US.core.rmc 3.2.6.0             # RSCT RMC Msgs - U.S. English
        rsct.msg.en_US.core.rmc.com 3.2.6.0         # RSCT RMC JAVA Msgs - U.S. En...
        rsct.msg.en_US.core.sec 3.2.6.0             # RSCT Security Msgs - U.S. En...
        rsct.msg.en_US.core.sensorrm 3.2.6.0        # RSCT Sensor RM Msgs - U.S. E...
        rsct.msg.en_US.core.sr 3.2.6.0              # RSCT Registry Msgs - U.S. En...
        rsct.msg.en_US.core.utils 3.2.6.0           # RSCT Utilities Msgs - U.S. E...
        rsct.msg.en_US.opt.storagerm 3.2.6.0        # RSCT Storage RM Msgs - U.S. ...
        xlC.msg.en_US.cpp 9.0.0.0                   # C for AIX Preprocessor Messa...
        xlC.msg.en_US.rte 16.1.0.3                  # IBM XL C++ Runtime Messages-...
        xlsmp.msg.en_US.rte 5.1.0.0                 # XL SMP Runtime Messages - U....
```

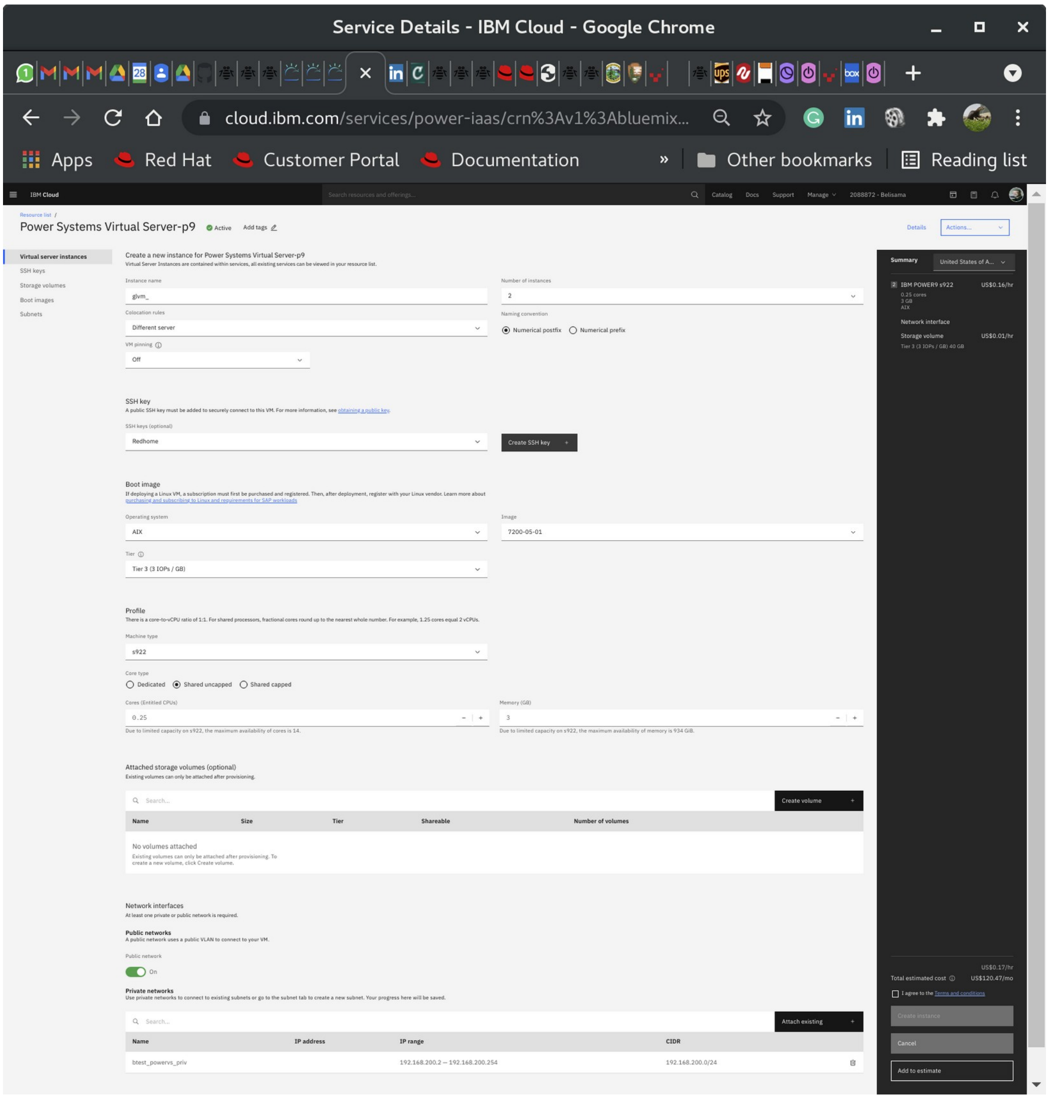
Appendix 3 – gmdsizing command

The command is found in /usr/es/sbin/cluster/samples/gmdsizing/gmdsizing

```
gmdsizing -i interval -t time {[-p pv [-p pv]...] | [-v vg [-v vg]…]} [-f
filename ] [-T] [-A] [-w] [-D char] [-V] [-h]
where:
      -i interval Interval at which disk activity is checked.
      -t time     Time period the command should measure. This defaults to
                  seconds. The minimum number of seconds is 10. The value
                  can be appended with the following letters to change the
                  unit of time:
                  d number of days
                  h number of hours
                  m number of minutes
                  s number of seconds
                  For example, to check over 5 days, you could use 5d,
                  120h, or 7200m.
      -p pv       Names of physical disks to monitor.
      -v vg       Names of volume groups to monitor.
      -f filename File in which to write the report, the default is
                  stdout.
      -T          Add time scale to the output.
      -A          Aggregate the output.
      -w          Collect data for write activities only.
      -D char     Use 'char' as delimiter in the output.
      -V          Verbose mode. Adds summary at end of the report.
      -h          Print the Help message.
```

Appendix 4 – Creation of PowerVS enviornment

The screen for the PowerVS LPAR creation:



Power Systems Virtual Server-p9 ● Active Add tags

Create a new instance for Power Systems Virtual Server-p9

Virtual Server Instances are contained within services, all existing services can be viewed in your resource list.

Instance name: glvm_

Number of instances: 2

Colocation rules: Different server

Naming convention: ● Numerical postfix ○ Numerical prefix

VM pinning: Off

SSH key

A public SSH key must be added to securely connect to this VM. For more information, see obtaining a public key.

SSH keys (optional): Redhome

Boot image

If deploying a Linux VM, a subscription must first be purchased and registered. Then, after deployment, register with your Linux vendor. Learn more about purchasing and subscribing to Linux and requirements for SAP workloads

Operating system: AIX

Image: 7200-05-01

Tier: Tier 3 (3 IOPs / GB)

Profile

There is a core-to-vCPU ratio of 1:1. For shared processors, fractional cores round up to the nearest whole number. For example, 1.25 cores equal 2 vCPUs.

Machine type: s922

Core type: ○ Dedicated ● Shared uncapped ○ Shared capped

Cores (Entitled CPUs): 0.25

Due to limited capacity on s922, the maximum availability of cores is 14.

Memory (GB): 3

Due to limited capacity on s922, the maximum availability of memory is 934 GiB.

Attached storage volumes (optional)

Existing volumes can only be attached after provisioning.

Name	Size	Tier	Shareable	Number of volumes
No volumes attached Existing volumes can only be attached after provisioning. To create a new volume, click Create volume.				

Network interfaces

At least one private or public network is required.

Public networks

A public network uses a public VLAN to connect to your VM.

Public network: On

Private networks

Use private networks to connect to existing subnets or go to the subnet tab to create a new subnet. Your progress here will be saved.

Name	IP address	IP range	CIDR
btest_powervs_priv		192.168.200.2 – 192.168.200.254	192.168.200.0/24

Summary — United States of A...

2 IBM POWER9 s922 0.25 cores 3 GB AIX	US$0.16/hr
Network interface	
Storage volume Tier 3 (3 IOPs / GB) 40 GB	US$0.01/hr
Total estimated cost	US$0.17/hr US$120.47/mo

□ I agree to the Terms and conditions

Figure 36: Creating PowerVS environment