VIOS Shared Storage Pools
Phase 2 – from December 2011
Presentation Version 17

Nigel Griffiths
IBM Power Systems
Advanced Technology Support, Europe

© 2012 IBM Corporation

VIOS Shared Storage Pool phase 2
- Announced 14th Oct 2012
- Available 16th Dec 2012
- Please read the Release notes:
  http://www-01.ibm.com/support/docview.wss?rs=0&uid=isg400000876
- Then read VIOS 2.2 User Guide:
  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hb1/p7hb1.pdf
- Technical Hands-On Movies
  – http://tinyurl.com/AlXmovies
Reminder of VIOS Storage Options

- 1 Logical Volume from VG Pool
- 2 Whole Local Disk
- 3 SAN LUN
- 4 File-Backed from a file system (pool) based on local or SAN disk(s)
- 5 NPIV LUN
- 6 SAN LUN Disks

Is vSCSI LUN or NPIV dead?

No, absolutely not

Customers continue to use all 6 options

Some people have the idea NPIV is somehow a strategic direction with IBM - this is not true IBM equally supports all 6 disk options.

Are the SAN guys spreading this rumour?
- they like the extra control of NPIV
- don’t care if it means 10 times the server setup work
How is it paid for?

Shared Storage Pools is a feature of PowerVM Standard & Enterprise

How is it installed?

Shared Storage Pool is a VIOS feature so just upgrade to VIOS 2.2.1.3 = FP25+sp1 December 2011 service pack

Note: This VIOS is AIX 6.1 TL7 based
NIM server needs to be AIX 6.1 TL7 or AIX 7.1 TL1

Note: VIOS 2.2.1.4 is now available

Why SSP? Nigel’s Opinion here

- Fibre-Channel LUN & NPIV is complex
  - SAN switch, SAN disk subsystem = hard work & weird GUI !!
  - Typical LUN lead time: 4 minutes, 4 hours, 4 days, 4 weeks?
  - With rapidly changing needs with mandatory responsiveness it is simply not good enough!
  - Many smaller computer rooms have no dedicated SAN guy
  - LPM hard work as most people don’t pre-Zone the target so have to Zone before the move = complexity, slow, error prone
  - LPM = zero outage for Hardware & Firmware upgrades

- Shared Storage Pool
  1. Allocate LUNs to the Virtual I/O Servers once
  2. One VIOS command to allocate space to a VM
     - Or use: cfgassist (VIOS’s smitty)
     - Or use: HMC Virtual Storage Management GUI
  3. LPM any time you like
Shared Storage Pool phase 2  1 of 3

Requirements

- Platforms: POWER6 & POWER7 only (includes Power Blades)
- VIOS Storage Pool (minimums):
  - Direct fibre-channel attached LUNs:
    - 1 for repository ~10 GB (NEW INFO: 1GB is enough) &
    - 1 or more for data, 10 GB → in practice lots more [like 1TB+]
- Pool Storage Redundancy: Repository & pool storage must be RAIDed
- VIOS name resolution to resolve hostnames
- Virtual I/O Server(s):
  - Minimum CPU: 1 (shared, uncapped is good)
  - Minimum Memory: 4 GB
  - Nigel’s recommendation: Please, no skinny VIOS
- To further VIOS upgrades
  - Client VM’s must be stopped, plus cluster down (use clstartstop)

Read the Release Notes
http://www-01.ibm.com/support/docview.wss?rs=0&uid=isg400000876

Shared Storage Pool phase 2  2 of 3

Limits

- Max nodes: 4 VIOS nodes
- Max physical disks in a pool: 256
- Max virtual disks (LUs) in a cluster: 1024
- Number of Client LPARs per VIOS: 1 to 40
  (that is, 40 clients per VIOS, or 40 clients per VIOS pair)
- Capacity of Physical Disks in Pool (each): 5GB to 4TB
- Storage Capacity of Storage Pool (total): 10GB to 128TB
- Capacity of each Virtual Disk (LU) in Pool: 1GB to 4TB
- Number of Repository Disks: 1 to 1 (CAA limit)

Read the Release Notes & README
Shared Storage Pool phase 2  3 of 3

Restrictions

Network
- Reliable & only IPv4
- No changes to hostname or IP address
- VIOS can’t use VLAN tagging
- DNS should use local /etc/hosts first
- Hostnames must be fully qualified
- Forward & reverse lookup must work
- Recommended to synchronise clocks
- Restoring VIOS from viosbr – get the networks configured 1st

Storage
- No interruption
- Can’t resize a LUN
- SSP may take more CPU
- No SCSI reservations (Reserve/Release)
- HA SAN solutions used to mitigate outages
- SANCOM not supported
- SEA must use default threaded mode
- If cluster or pool name >16 characters add APAR IV11852m13
- Don’t use vSCSI adapter “Any client partition can connect”
- AMS or Suspend/Resume can’t use SSP for Paging Space

Terms

Shared Storage Pool phase 2 = SSP2

1 SSP2 cluster = set of co-operating Virtual I/O Servers
Currently a VIOS can only be in one cluster.

Here we show two clusters
Shared Storage Pool phase 2 = SSP2

Terms

1 SSP2 cluster = set of co-operating Virtual I/O Servers
Currently a VIOS can only be in one cluster

2 SSP2 pool = set of LUNs
Data LUNs + a special Repository LUN = cluster meta-data
Currently a cluster can only have one pool

3 LUNs actualised on SAN disks connected to whole cluster
Any SAN disks supported by VIOS

2 SSP2 pool = set of LUNs
Data LUNs + a special Repository LUN for cluster meta-data
Currently a cluster can only have one pool

1 SSP2 cluster = set of co-operating Virtual I/O Servers
Currently a VIOS can only be in one cluster
Terms

Shared Storage Pool phase 2 = SSP2

3 LUNs actualised on SAN disks connected to whole cluster
Any SAN disks supported by VIOS

2 SSP2 pool = set of LUNs
Data LUNs + a special Repository LUN for cluster meta-data
Currently a cluster can only have one pool

1 SSP2 cluster = set of co-operating Virtual I/O Servers
Currently a VIOS can only be in one cluster

4 SSP2 space allocated to virtual machines = LU
LU = Logical Unit = 64 MB blocks, evenly distributed across all LUNs

5 Live Partition Mobility available by default
Whole SSP2 cluster has LU online
Preparation

- All the Cluster VIOSs need the LUNs online
  Make sure they are available = Zoned

- **BEFORE** you start the cluster for all LUNs
  chdev -dev <device name> -attr reserve_policy=no_reserve

- Don’t forget this for extra disks that you add later

- Forgetting this = a real mess to make the
  low level disk attribute change

Cluster Management

Section Contents

- Create Cluster and Pool
- Find out information about the SSP
- Add node
- House Keeping
Cluster full hostnames mandatory + DNS

1. Need full DNS working or /etc/hosts
2. $ hostname
   redvios1.ibm.com
   $ host redvios1.ibm.com - must work
   $ host 9.123.456.99 - must work

VIOS
oem_setup_env
smitty tcpip
Example →

Cluster create on 1st node

Create cluster on one VIOS (here called bluevios1)

$ cluster -create -clustername galaxy \
   -repopvs hdisk2 \
   -spname atlantic -sppvs hdisk3 hdisk5 \
   -hostname bluevios1.ibm.com

... Cluster galaxy has been created successfully.

It will take a minute or two, then output Cluster created
You will find a bunch of new daemons running.

If it complains the disks are "in use" check.
If certain they are correct, wipe the disk content with:
   # cleandisk -- hdiskX
   # cleandisk -- hdiskX
It may ask you to confirm y/n ?
Then on that 1st node - add other nodes

On the first VIOS running the cluster

```bash
$ cluster -addnode -clustername galaxy \        
   -hostname redvios1.ibm.com
```

Partition redvios1.aixncc.uk.ibm.com has been added to the galaxy cluster

```bash
$
```

Add other node(s) as necessary.

---

List cluster & cluster nodes

```bash
$ cluster -list
Cluster Name   Cluster ID
galaxy         68c06102fc5311e093c8f6027171fc64
$
```

```bash
$ cluster -status -clustername galaxy
Cluster Name   State
galaxy         OK
Node Name        MTM           Partition Num  State  Pool State  
diamondvios1     8233-E8B02100271P         2  OK     OK
diamondvios2     8233-E8B02100271P         1  OK     OK
redvios1         8203-E4A0310E0A41         1  OK     OK
$
```

Example of a 3 node cluster
House keeping

You can remove a node from the cluster
- LPM any important client Virtual machines elsewhere
- Stop remaining VMs
- Remove the client VMs
- Remove their allocated virtual disks
  then

  $ cluster -rmnode -clustername galaxy \n     -hostname redviosl.ibm.com

You can also remove the cluster completely
- Once all disk space unassigned & nodes removed

  $ cluster -delete -clustername galaxy

Pool Disk Space Management

Content
- Allocate pool disk space and give to a VM
  - Ditto as two commands
- Removing the disk space
- Monitoring the pool
Allocate disk space & assign to client VM

$ mkbdsp -clustername galaxy \ 
  -sp atlantic 16G -bd vdisk_diamond6a \ 
  -vadapter vhost2

Logical Unit vdisk_diamond6a has been created with udid: 615af85de5ac9a8827e9cd01d6b36. 
Assigning file “vdisk_diamond6a” as a backing device. 
Vtscsi3 Available.

Notes:
▪ 16 GB is not actually allocated until written too 
▪ vdisk_diamond6a is just a name = reminder of the VM using it 
▪ vhost2 is the virtual SCSI adapter for client VM diamond6

Same but two steps

1) Create Logical Unit  Note: no –vadapter option
$ mkbdsp -clustername galaxy -sp atlantic 10G -bd LU42
Lu Name:LU42
Lu Udid:374a609cb072e4015d558f290b9f0bd

List the pool contents
$ lssp -clustername galaxy -sp atlantic -bd
Lu Name  Size(mb) ProvisionType Lu Udid 
LU42 10240  THIN 374a609cb072e4015d558f290b9f0bd ...

2) Example of two ways using “-bd LU42” or “–luuid <hexidecimal>” 
▪ -bd only works if LU42 is unique 
▪ Note: below no Size argument (or it creates another LU with same name!)
$ mkbdsp -clustername galaxy -sp atlantic -bd LU42 -vadapter vhost2
Assigning file “LU42” as a backing device. 
VTD:vtscsi1 
- or - -
$ mkbdsp -clustername galaxy -sp atlantic 
  --luuid 374a609cb072e4015d558f290b9f0bd -vadapter vhost2
Assigning file “374a609cb072e4015d558f290b9f0bd” as a backing device. 
VTD:vtscsi1
Dual path via Two VIOSs

1. Setup virtual SCSI adapter pairs as normal
   - client VM virtual SCSI adapter A <-> VIOS C
   - client VM virtual SCSI adapter B <-> VIOS D

2. On VIOS C: use "lsmap -all" to map slot to vhostN
   - mkbdsp -clustername galaxy -sp atlantic 16G -bd vdisk_red6a -vadapter vhostN

3. On VIOS D: use lsmap -all to map slot to vhostM
   - mkbdsp -clustername galaxy -sp atlantic -bd vdisk_red6a -vadapter vhostM

4. On the client VM
   - $ lspath
     Enabled hdisk0 vscsi0
     Enabled hdisk0 vscsi1

5. LPM still available – dual VIOS to dual VIOS

Removing an LU (Logical Unit)

Assuming it is NOT used !

On the VIOS remove disk space
rmbdsp = remove backing device from storage pool

- $ rmbdsp -clustername galaxy \
  -sp atlantic -bd vdisk_diamond6a

or via the LU hexadecimal name
- $ rmbdsp -clustername galaxy -sp atlantic \
  -luuid 858152297879adfe0d75b05f586d36ee
House keeping

Add more physical LUNs to the Pool

$ chsp -add -clusternam galaxy -sp atlantic hdisk8

Remove a physical LUN from the Pool
- You can not ... with this release.
- We can replace a disk but not remove one
  - Replacement disk - equal or larger size

Experiments in Thin provisioning
= Allocating disk blocks only when they are used i.e. written
Thin Provisioning
Blocks can be missing

List storage pool "lssp" output shows block size = 64MB

1 Client VM writes to block 5
2 VIOS spots it is not there & allocates the block from free list
3 then VIOS completes the write
4 so the client is unaware of this
**Thin Provisioning**

- `mkbdsp` states the “LU” size
- Blocks assigned only when written
- After installing AIX 7 (could be any supported OS)
- AIX sees 16 GB disk
- AIX has allocated 5 GB in rootvg
- But not actually written to all 5 GB
  - Paging space not used
  - Free space in filesystems not used
  - Sparse files have “holes”
- Brand new pool & AIX 7 only 3 GB used from the pool
- Instead of unused disk space in every VM, now it is SSP “pooled”

Complete guesswork: 20,000 machines * 20 VMs * 16 GB unused = 6 Petabytes

**Thick Provisioning**

- Doh! A no-brainer!
- Like Thin but actually allocate all the disk space
- New option: `mkbdsp ... -thick`

The point is
- Avoids problems, if the free list empties
- Good for more important work/production or you prefer not to dynamically add blocks
Monitoring: topas on VIOS then “D”

One client VM running: yes >/tmp/x

Disk I/O spread across disks
Allocation unit is 64MB (was in the lssp output)

Monitoring Disk use with lssp

$ lssp -clustername galaxy -sp atlantic -bd

Lu(Disk) Name      Size(MB) ProvisionType Lu Udid
vdisk_diamond6a   16384   THIN          615af . . .
vdisk_diamond8a   16384   THIN          917c0 . . .
vdisk_diamond5a    8192   THICK         f1442 . . .
vdisk_diamond5b    8192   THICK         ebecd . . .
vdisk_diamond3a   10240   THIN          afcec . . .

$ lssp -clustername galaxy

POOL_NAME:       atlantic
POOL_SIZE:     47552
FREE_SPACE:    17945
TOTAL_LU_SIZE: 59392

47552 Pool Physical Size
17945 Pool Physical Free
29607 Pool Physical Used
Pool use 29607/47522x100=62%

59392 Allocated
Pool Over commit 59392/47522= 1.25
allocated 25% more than I have!
= Thin provisioning
Thin provisioning risks pool free space = zero Ek!
Next write needing a new SSP block, gets a disk error!
Just don’t go there – you need to be warned!

Thin Provisioning Alerts

- Set alerts to warn on free pool space getting too low %
  `alert -set -clusternamex galaxy -spnamex atlantic -value 10`

- To list the alert threshold:
  - `alert -list -clusternamex galaxy -spnamex atlantic`
  {\$} `alert -list -clusternamex galaxy -spnamex atlantic`
  - PoolName: atlantic
  - PoolID: 00000000009893EDD000000004F174D22
  - ThresholdPercent: 35

- To remove the alert:
  - `alert -unset -clusternamex galaxy -spnamex atlantic`
  - Threshold is set to 0 (zero) – it will not happen!!

- The default alert is free pool space below 35%
House keeping – Alert Reporting

- Reported on **any one of the VIOS cluster**
- padmin user: `errlog`
  - Like AIX `errpt`
  - `$ errlog | more`

```
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
0FD4CF1A 0215112612 I O VIOD_POOL  Informational Message
...
```

- `$ errlog –ls | more`

See example on the next page

- Can also be reported to high levels SM like Systems Director etc.
House keeping - Thin Provisioning Alerts

- It is vital that you get these warning messages
- Suggest on ALL VIOS
  1. Email the Pool stats every night to the admin guys (cron as root)

```
/home/padmin/.profile
lssp -clustername galaxy | /usr/bin/mailx -s "SSP stats" ops@acme.com
```
  2. Script to check and if free space is low then email or send phone TEXT message or escalate
- Possible reactions are:
  - Add a new LUN to the pool,
  - Delete allocated space = unused LU or entire VM & space
  - Drop a Snapshot
Alert bug fix

Warning:
If running the initial Phase 2 release
= ioslevel: 2.2.1.3 = Fix Pack 25, service pack 1 + fix 1

You need a further efix to get Alerts working
Details on my AIXpert blog

VIOS 2.2.1.4 is now available

Snapshot
## Snapshots on VM disks and Cloning

Snapshot available using
- Advanced SAN disks or SAN Volume Controller (SVC)

but now VIOS admin can do this too!

### Snapshot + Drop
- Very quick
- Allows point in time backup
- Later delete the original to reclaim the space

**Examples:**
- Backup VM stopped, quiesce, live

### Snapshot + Roll-back
- Very quick
- Useful for lots of reasons
- Stop the client VM
- Restart on original copy
- Discard newer copy

**Examples:**
- Practice OS or App update
- Training & reset
- Benchmark & reset
- Failure & avoid recovery from tape
- Save points for batch runs

Supports single disk or a consistent set of disks

---

## Snapshot – create, list, delete or rollback

### Snapshot Usage:

<table>
<thead>
<tr>
<th>Command</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>snapshot -create &lt;filename&gt;</td>
<td>-clustername galaxy -spname atlantic -lu LUs</td>
</tr>
<tr>
<td>snapshot -delete &lt;filename&gt;</td>
<td>-clustername galaxy -spname atlantic -lu LUs</td>
</tr>
<tr>
<td>snapshot -rollback &lt;filename&gt;</td>
<td>-clustername galaxy -spname atlantic -lu LUs</td>
</tr>
<tr>
<td>snapshot -list</td>
<td>-clustername galaxy -spname atlantic</td>
</tr>
</tbody>
</table>

**Notes:**
- Alternatively swap “-lu LU_name(s)” for “-luuid Hexadecimal”
- LUs means a space separated list disk names

---
Snapshot – create and list

Create

```sh
$ snapshot -create diamond5s.snap -clusternname galaxy -spname atlantic -lu vdisk_diamond5a
```

List

```sh
$ snapshot -list -clustername galaxy -spname atlantic
Lu Name    Size(mb)    ProvisionType  Lu Udid
vdisk_diamond5a  16384       THIN             b3f3a . . .
```

Also snapshot appear in the lssp output

```sh
$ lssp -clustername galaxy -sp atlantic -bd
Lu Name    Size(mb)    ProvisionType  Lu Udid
vdisk_diamond5a  16384       THIN             b3f3a . . .
vdisk_diamond6a  16384       THIN             4c9e9 . . .
```

Snapshot – delete or rollback

When sure you never want to rollback
Delete original & continue on the current blocks

```sh
$ snapshot -clustername galaxy -delete diamond5t.snap -spname atlantic -lu vdisk_diamond5a
```

Rollback to a snapshot
Stop the virtual machine/LPAR then

```sh
$ snapshot -clustername galaxy -rollback diamond5t.snap -spname atlantic -lu vdisk_diamond5a
```

You lose any updated you made since the last snapshot
Creating a snapshot only involved copying the meta data i.e. list of the blocks within the LU (not the block themselves)
Snapshot + Update

Original Set becomes the Snapshot

Block 0 1 2 3 4 5 6 7 8 9
- - - - - - - - -

Block 0 1 2 3 4 5 6 7 8 9
- - - - - - - - -

New working set
1 Client VM update to block 7
2 VIOS allocates a new block
3 Copies original 64MB
4 VIOS completes the write

Snapshot delete or Roll-back

Original Set

Block 0 1 2 3 4 5 6 7 8 9
- - - - - - - - -

Block 0 1 2 3 4 5 6 7 8 9
- - - - - - - - -

New working set

Delete ➔ remove Original set not in the New working set = blue block 7 to free list
Roll-back = remove New set not in the Original set = green block 7 to free list
Storage Management

- Reminder currently,
  - One pool of large LUNs - syntax suggests multiple pools later
  - Pool can be on a mix of brands or generations of disk sub-systems
  - 64 MB chunks are spread as evenly as possible across LUNs

Live Storage Mobility

Server Admin would like to:
A. Replace a faulty LUN
B. Move all blocks off one disk subsystem (retiring a disk subsystem)
C. Recover from repository failure
D. Select which disk subsystems a particular VM uses
E. Ensure mirrors are on different subsystems (even different sites)

- A and B → via replace physical disk
  - chsp -replace -clustername galaxy -sp atlantic -oldpv hdisk4 -newpv hdisk24

- C → see viosbr command (later)

- D and E → in a later SSP release
  - Multiple pools is an obvious solution here (not in the current release)
  - Could use SVC now for lower level mirror (E)
Replace Disk

Live Storage Mobility phase 1 replace faulty disk or move all from old sub-system

New LUNs

Replace Disk

New LUN
Disk Space Move

Not in this Release

Virtual I/O Server

vSCSI

Client VM

FibreChannel

Network

Virtual I/O Server

vSCSI

Client VM

Virtual I/O Server

vSCSI

Client VM

Move Blocks of one particular LU

Mirror LU Blocks between pools

Local Storage Pool A

Remote Storage Pool B

User Interface

- Command line
  - Already shown in this presentation by example
  - Some feature will remain command line only \( \rightarrow \) like: cluster -create

- cfgassist
  - This is the VIOS version of smitty
  - Menu driven interface for CLI
  - Fully covers SSP functions

- Graphical User Interface
  - HMC – now (note the don’t have any VIOS vSCSI slots in “Any node”

- System Director - **Future release**
  - Already has Storage Pools concept and features
  - SSP is just another storage pool type
  - Then may adds new unique items – like VMControl appliance deploy to SSP disk space or cloning
Hardware Management Console

- HMC from October 2011
  - for SSP support & LPM of SSP LPARs
  - Addition feature to Virtual Storage Management

- Shipped with HMC upgrade V7 R7.4 SP0+
  - Not part of the VIOS package
What if you lose the VIOS?

- Updated `viosbr` supports backup / restore of SSP config
  - Warning: this saves the config but not the data
- Backup – will perform regular backups for you
  
  ```sh
  viosbr -backup -clustername Name -file File   
  [-frequency daily|weekly|monthly [-numfiles fileCount]]
  ```
- View
  
  ```sh
  viosbr -view -file File -clustername Name [-type devType][-detail |-mapping]
  ```
- Restore
  
  ```sh
  viosbr -restore -clustername N -file F -subfile NodeFile [-validate | -inter | -force][-type devType]
  viosbr -restore -clustername N -file F -repopvs disks [-validate | -inter | -force][-type devType][-currentdb]
  viosbr -restore -clustername N -file F -subfile NodeFile --xmlvtds
  viosbr -recoverdb -clustername N [-file F]
  ```
- Can recover from
  1. Repository Disk is corrupted (see `repopvs`)
  2. One SSP VIOS is reinstalled
  3. SSP Database is corrupted
  4. Restore to old configuration on the VIOS node
    - Changes done to SSP mappings on the node after a backup
Shared Storage Pool phase 2 – Call to Action

As a result of this presentation: I want you to 
Do
1. Start negotiating with SAN team to hand-over a few TB
2. Get to VIOS 2.2.1.4 on all POWER6/7 … ASAP

Feel
– Excited with easy SAN disk management & LPM

Think
– About how this technology could save you time, boost efficiency & increase responsiveness to users

SSP2 command cheat sheet

1. chdev -dev <device name> -attr reserve_policy=no_reserve
2. cluster -create -clusternname galaxy -repovps hdisk2
   -spname atlantic -spvps hdisk3 hdisk5 -hostname blueviols1.ibm.com
3. cluster -list
4. cluster -status -clustername galaxy
5. cluster -addnode -clustername galaxy -hostname redviols1.ibm.com
6. cluster -rmnode [-f] -clustername galaxy -hostname redviols1.ibm.com
7. cluster -delete -clustername galaxy

8. lcluster -s or -d or -c or -i = CAA commands
9. chap -add -clustername galaxy -sp atlantic hdisk8 hdisk9
10. chap -replace -clustername galaxy -sp atlantic -oldpv hdisk4 -newpv hdisk24
11. mkbdsp -clustername galaxy -sp atlantic 16G
   -bd vdisk_red6a -vadapter vhost2 [-thick]
12. rmbdsp -clustername galaxy -sp atlantic -bd vdisk_red6a
13. lsap -clustername galaxy -sp atlantic -bd
14. lsap -clustername galaxy

15. alert -set -clustername galaxy -spname atlantic -value 80
16. alert -list -clustername galaxy -spname atlantic

17. errlog -ls
18. snapshot -create name -clustername galaxy -spname atlantic -lu LU42
19. snapshot -delete name -clustername galaxy -spname atlantic -lu LU42
20. snapshot -rollback name -clustername galaxy -spname atlantic -lu LU42
21. snapshot -list -clustername galaxy -spname atlantic
22. viosbr -backup -clustername galaxy -file Daily -frequency daily -numfiles 10
23. viosbr -view -file File -clustername Name ...
24. viosbr -restore -clustername Name ...