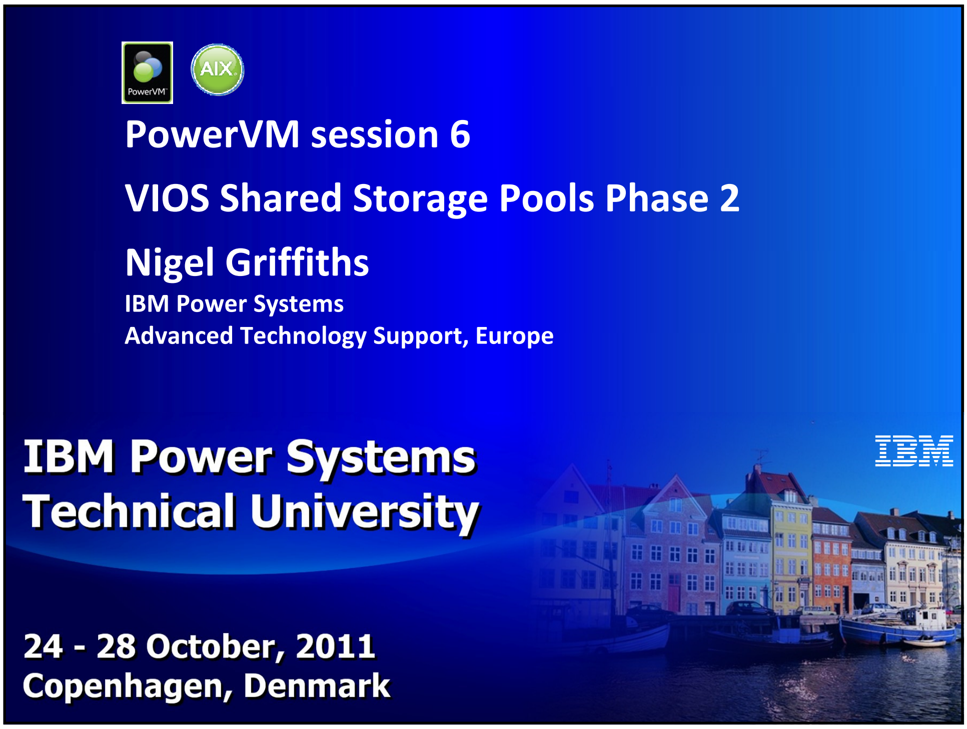


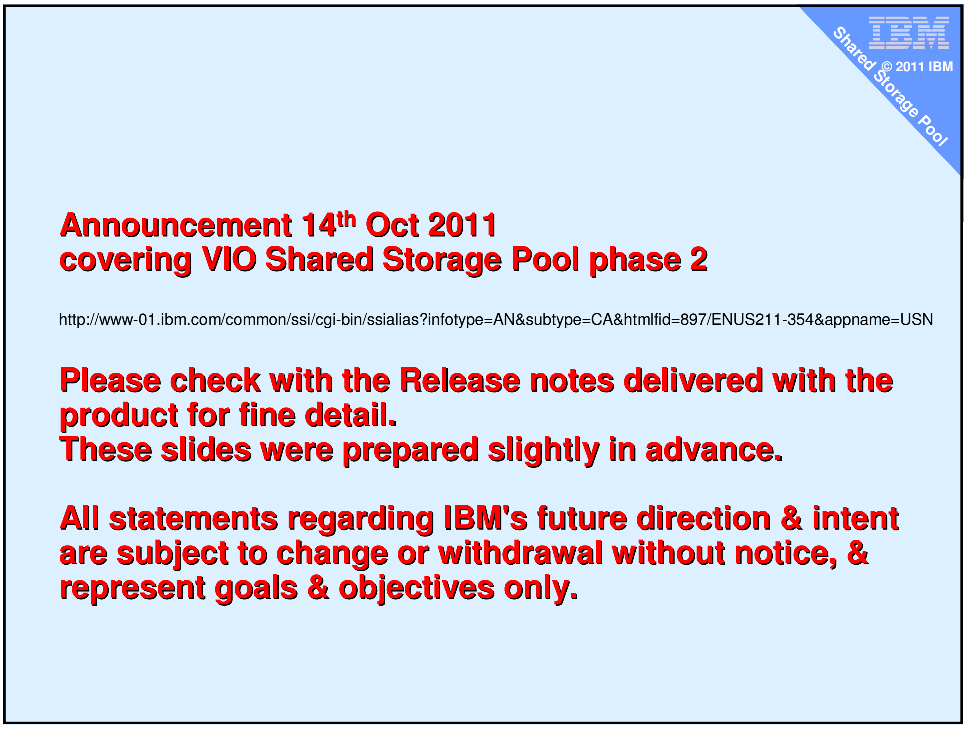
PowerVM session 6

VIOS Shared Storage Pools Phase 2

Nigel Griffiths
IBM Power Systems
Advanced Technology Support, Europe

IBM Power Systems Technical University

24 - 28 October, 2011
Copenhagen, Denmark



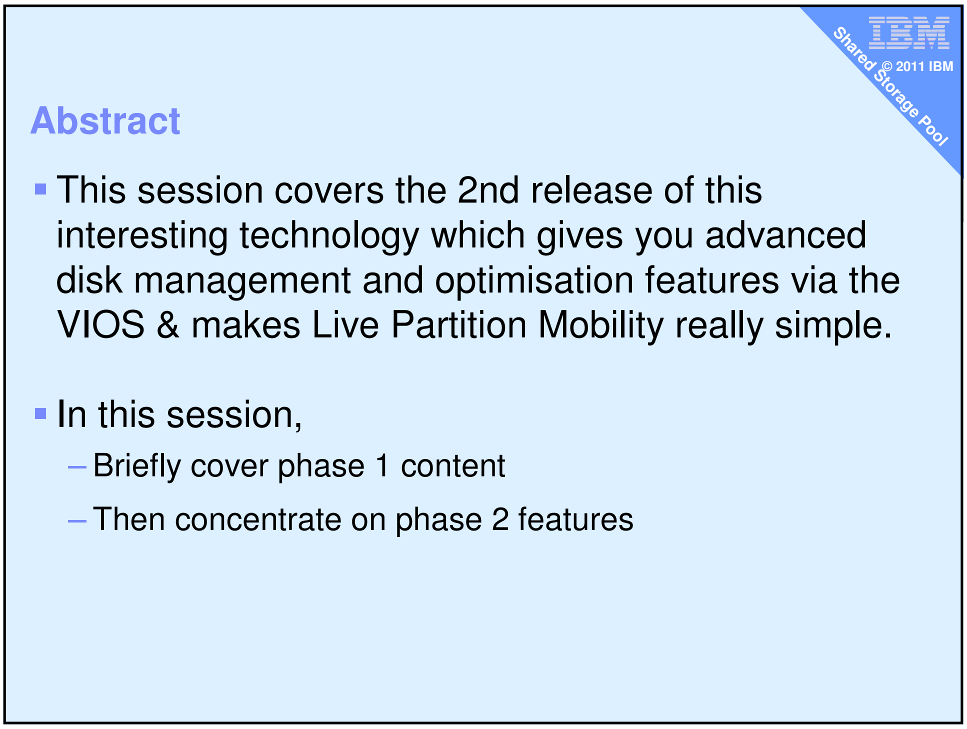
Shared Storage Pool

Announcement 14th Oct 2011
covering VIO Shared Storage Pool phase 2

http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=AN&subtype=CA&htmlfid=897/ENUS211-354&appname=USN

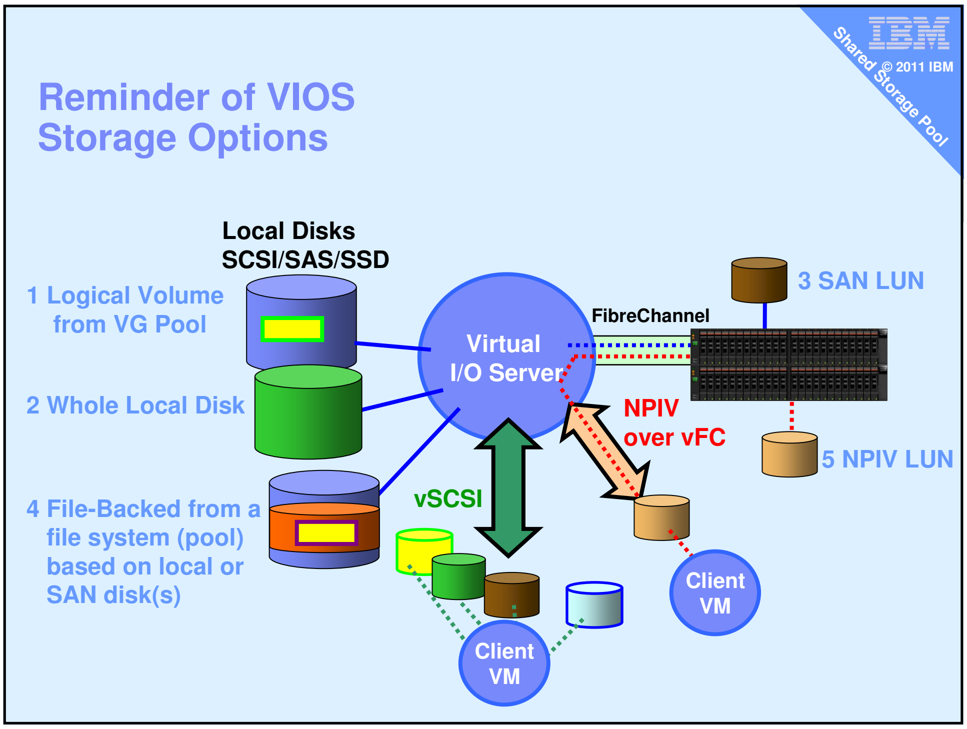
Please check with the Release notes delivered with the product for fine detail.
These slides were prepared slightly in advance.

All statements regarding IBM's future direction & intent are subject to change or withdrawal without notice, & represent goals & objectives only.



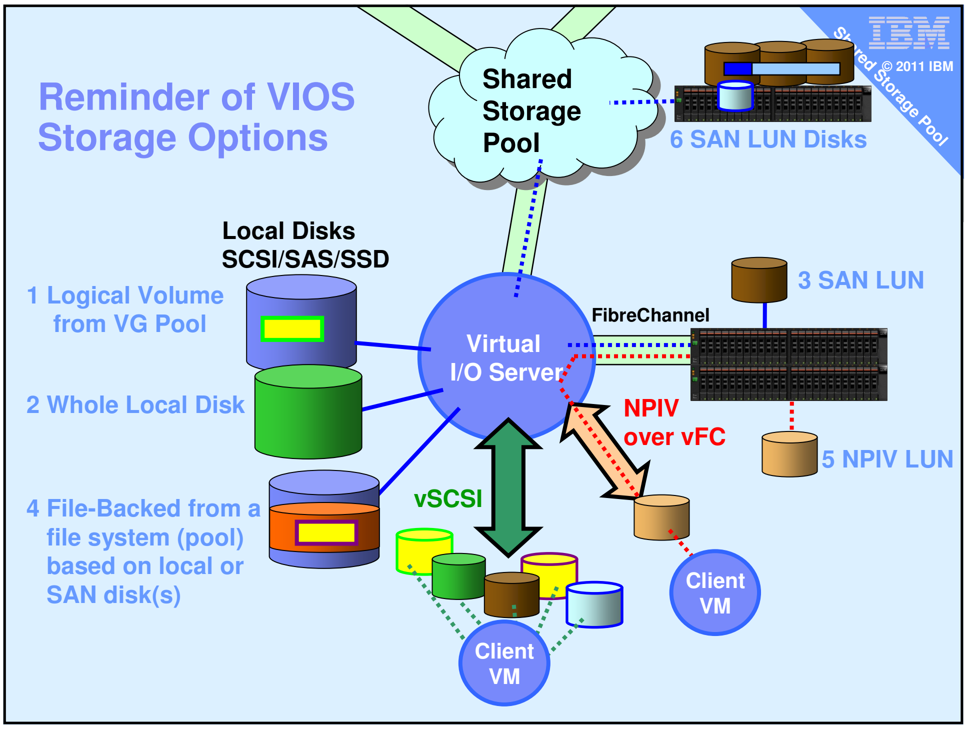
Abstract

- This session covers the 2nd release of this interesting technology which gives you advanced disk management and optimisation features via the VIOS & makes Live Partition Mobility really simple.
- In this session,
  - Briefly cover phase 1 content
  - Then concentrate on phase 2 features



Reminder of VIOS Storage Options

Reminder of VIOS Storage Options



Shared Storage Pool

6 SAN LUN Disks

Local Disks SCSI/SAS/SSD

1 Logical Volume from VG Pool

2 Whole Local Disk

4 File-Backed from a file system (pool) based on local or SAN disk(s)

Virtual I/O Server

FibreChannel

3 SAN LUN

NPIV over vFC

5 NPIV LUN

vSCSI

Client VM

Client VM

Is vSCSI LUN or NPIV dead?

No, absolutely not

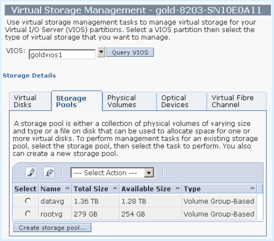
Customers continue to have ALL 6 options

Note:
All come with VIOS at no extra cost, just upgrade your VIOS

Note:
Shared Storage Pools comes with PowerVM Standard & Enterprise
Not PowerVM Express

Why add SSP? Nigel's Opinion here

- Fibre-Channel LUN & NPIV is complex
  1. SAN switch, SAN disk subsystem – weird GUI !!
  2. Typical lead time: 4 minutes, 4 hours, 4 days, 4 weeks!
  3. With rapidly changing needs with mandatory responsiveness it is simply not good enough
  4. Many smaller computer rooms have no dedicated SAN guy
  5. LPM hard work as most people don't pre-Zone the target so have to Zone before the move = complexity, slow, error prone
- Shared Storage Pool
  - Allocate pool LUNs to the VIOS(s) + one VIOS cmd to allocate space to a VM
  - Also via cfgassist (VIOS's smitty) or Pool space allocation function via HMC →



Virtual Storage Management - gold-8203-SN10E0A11

Use virtual storage management tasks to manage virtual storage for your Virtual I/O Server (VIOS) partitions. Select a VIOS partition then select the type of virtual storage that you want to manage.

VIOS: goldvios1

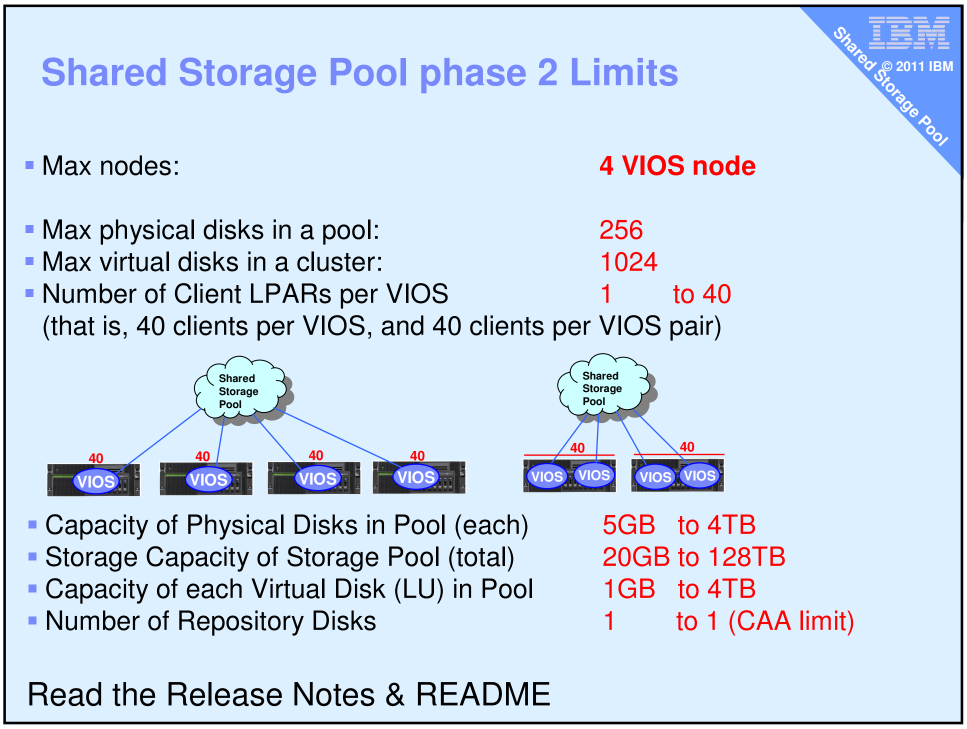
Storage Details

A storage pool is either a collection of physical volumes of varying size and type or a file on disk that can be used to allocate space for one or more virtual disks. To perform management tasks for an existing storage pool, select the storage pool, then select the task to perform. You also can create a new storage pool.

Select	Name	Total Size	Available Size	Type
○	datavg	1.36 TB	1.28 TB	Volume Group-Based
○	rootvg	279 GB	254 GB	Volume Group-Based

Shared Storage Pool phase 2 Requirements

- Platforms: **POWER6** & **POWER7** only (includes **Power Blades**)
- VIOS Storage Pool (minimums):
  - Direct fibre-channel attached LUNs:
  - **1 for repository ~1 GB and**
  - **1 or more for data, 1 GB → in practice lots more**
- Pool Storage Redundancy: Repository & pool storage must be **RAIDed**
- VIOS **name resolution** to resolve hostnames
- Nigel's recommendation no skinny Virtual I/O Server(s):
  - **Minimum CPU: 1**
  - **Minimum Memory: 4GB**



Shared Storage Pool phase 2 Limits

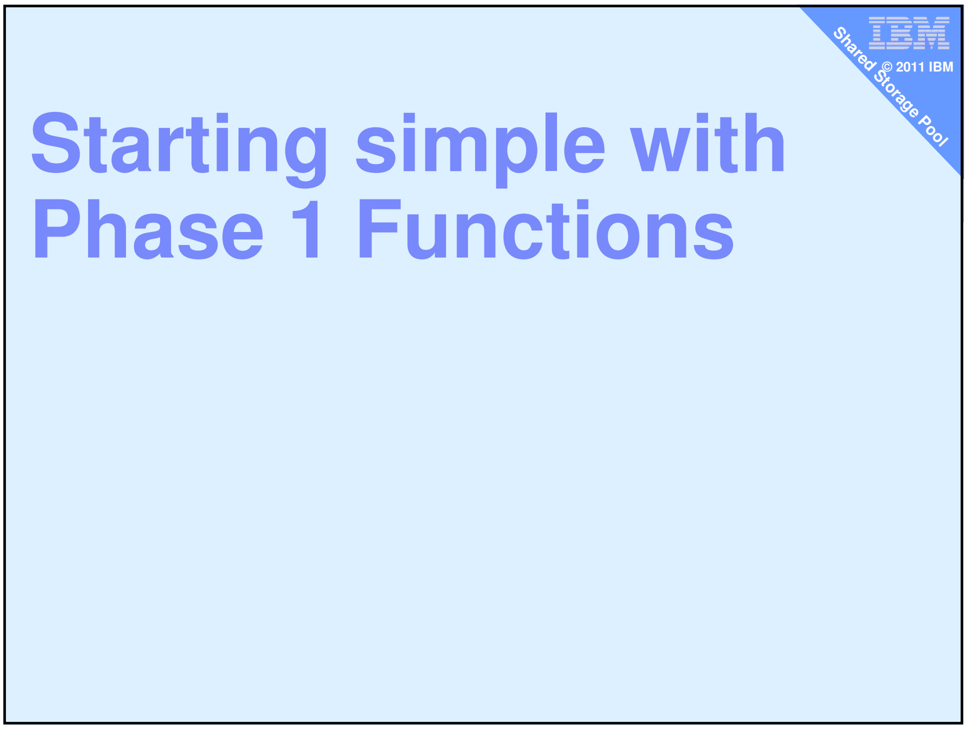
Max nodes:	**4 VIOS node**
Max physical disks in a pool:	256
Max virtual disks in a cluster:	1024
Number of Client LPARs per VIOS	1 to 40

(that is, 40 clients per VIOS, and 40 clients per VIOS pair)

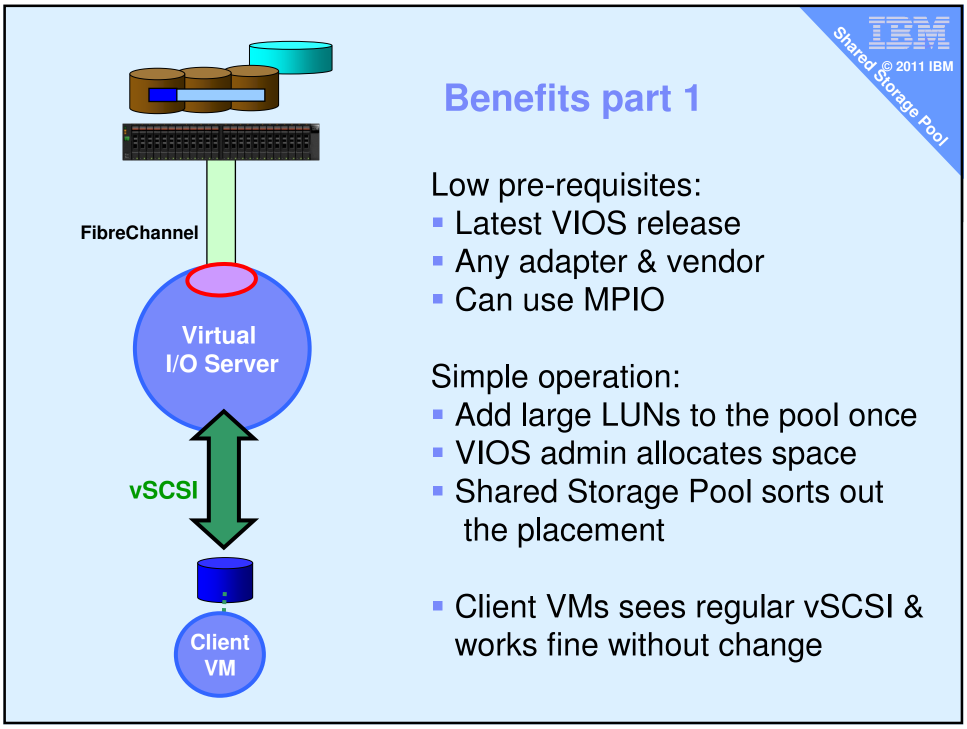
Shared Storage Pool
40
40
40
40
VIOS
VIOS
VIOS
VIOS
Shared Storage Pool
40
40
VIOS
VIOS
VIOS
VIOS

Capacity of Physical Disks in Pool (each)	5GB to 4TB
Storage Capacity of Storage Pool (total)	20GB 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



Starting simple with Phase 1 Functions



Benefits part 1

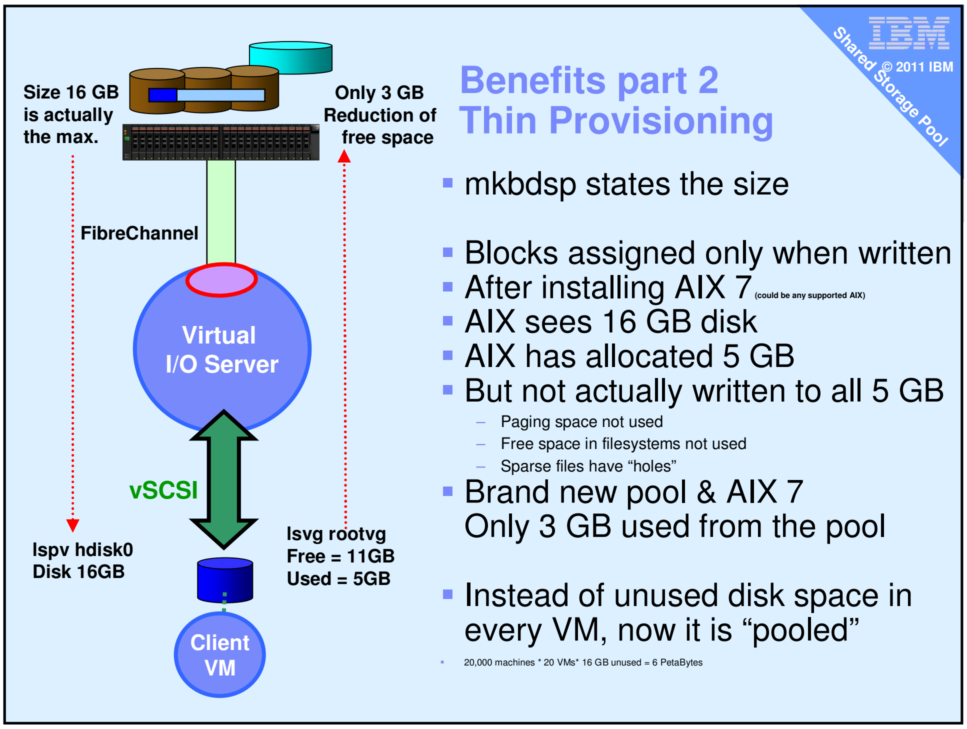
Low pre-requisites:

- Latest VIOS release
- Any adapter & vendor
- Can use MPIO

Simple operation:

- Add large LUNs to the pool once
- VIOS admin allocates space
- Shared Storage Pool sorts out the placement

- Client VMs sees regular vSCSI & works fine without change



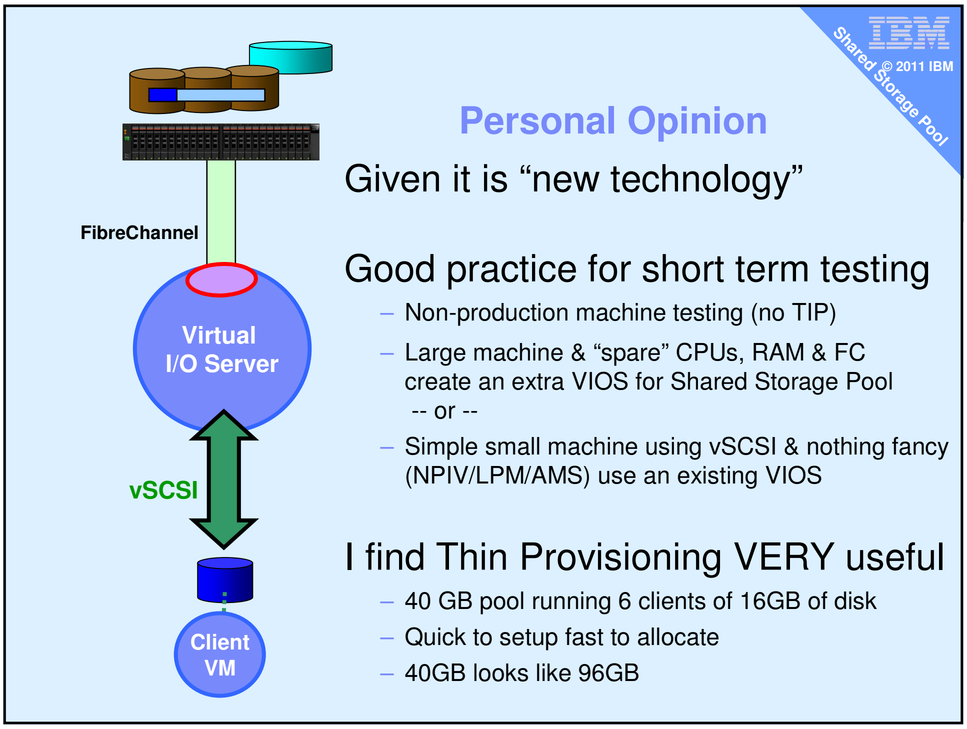
Benefits part 2
Thin Provisioning

- mkbdsp states the size

- Blocks assigned only when written
- After installing AIX 7 (could be any supported AIX)
- AIX sees 16 GB disk
- AIX has allocated 5 GB
- 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 "pooled"

- 20,000 machines * 20 VMs* 16 GB unused = 6 PetaBytes



Personal Opinion

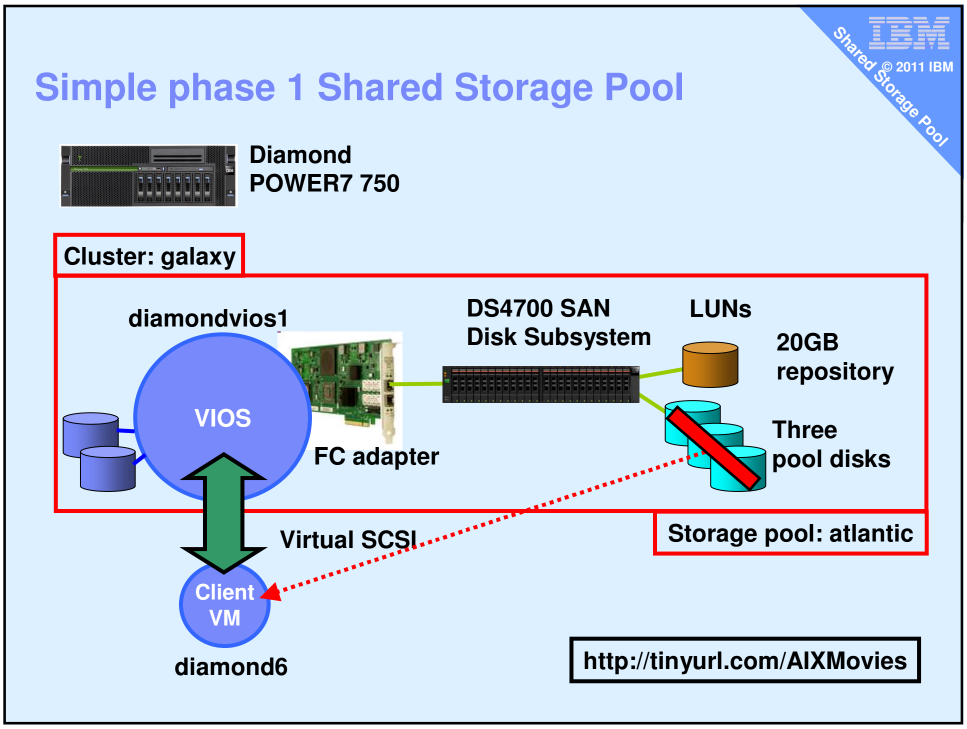
Given it is “new technology”

Good practice for short term testing

- Non-production machine testing (no TIP)
- Large machine & “spare” CPUs, RAM & FC create an extra VIOS for Shared Storage Pool
  -- or --
- Simple small machine using vSCSI & nothing fancy (NPIV/LPM/AMS) use an existing VIOS

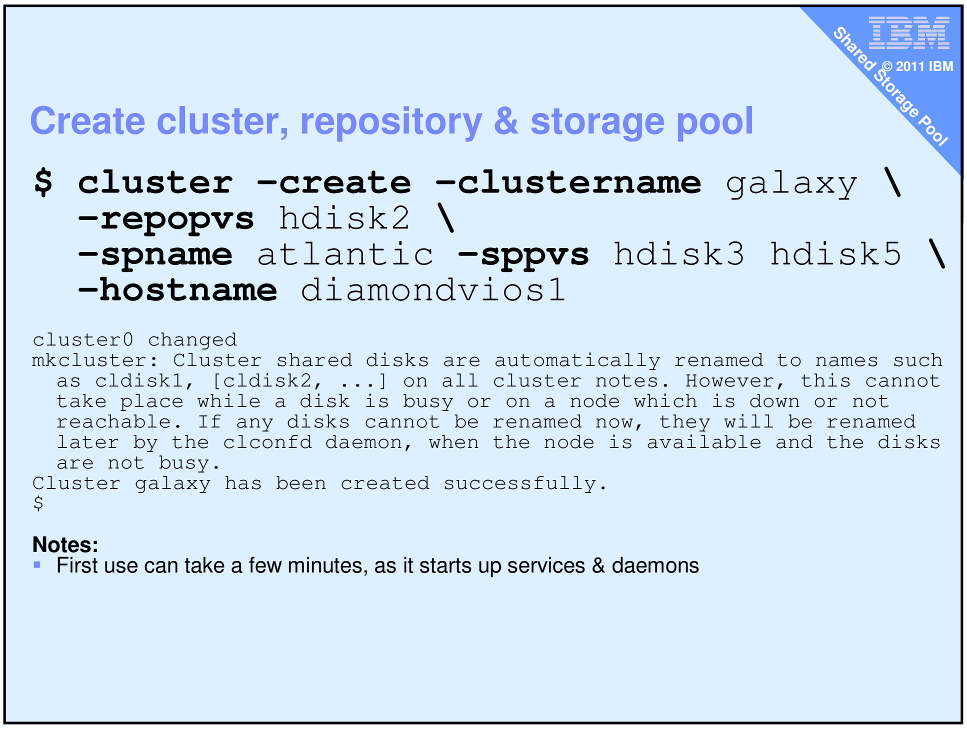
I find Thin Provisioning VERY useful

- 40 GB pool running 6 clients of 16GB of disk
- Quick to setup fast to allocate
- 40GB looks like 96GB



Simple phase 1 Shared Storage Pool

http://tinyurl.com/AIXMovies



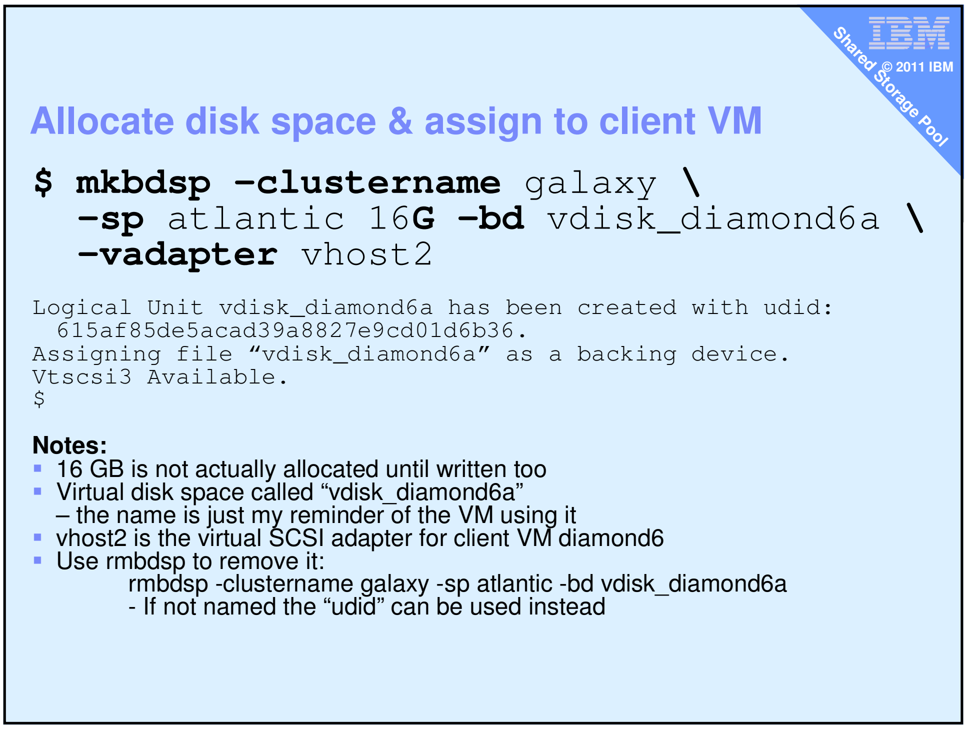
Create cluster, repository & storage pool

```
$ cluster -create -clustername galaxy \
  -repopvs hdisk2 \
  -spname atlantic -sppvs hdisk3 hdisk5 \
  -hostname diamondvios1
```

```
cluster0 changed
mkcluster: Cluster shared disks are automatically renamed to names such
  as cldisk1, [cldisk2, ...] on all cluster notes. However, this cannot
  take place while a disk is busy or on a node which is down or not
  reachable. If any disks cannot be renamed now, they will be renamed
  later by the clconfd daemon, when the node is available and the disks
  are not busy.
Cluster galaxy has been created successfully.
$
```

Notes:

- First use can take a few minutes, as it starts up services & daemons



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:
  615af85de5acad39a8827e9cd01d6b36.
Assigning file “vdisk_diamond6a” as a backing device.
Vtscsi3 Available.
$
```

Notes:

- 16 GB is not actually allocated until written too
- Virtual disk space called “vdisk_diamond6a”
  – the name is just my reminder of the VM using it
- vhost2 is the virtual SCSI adapter for client VM diamond6
- Use rmbdsp to remove it:
  rmbdsp -clustername galaxy -sp atlantic -bd vdisk_diamond6a
  - If not named the “udid” can be used instead

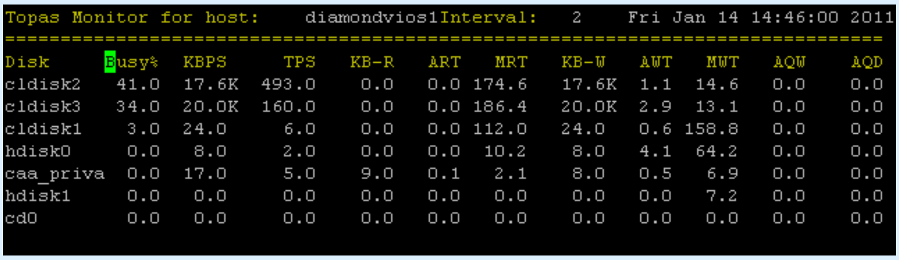
Monitoring Disk Use

```
$ lssp -clustername galaxy -sp atlantic -bd
Lu(Disk) Name                           Size(MB        Lu Udid
vdisk_diamond6a                         16384          615af85de5acad39a8827e9cd01d6b36
vdisk_diamond8a                         16384          917c0ccd290c69c0f1c56bd9c06c4306
vdisk_diamond5a                         8192           f14421c104b217d8c4afdc93571b8adf
vdisk_diamond5b                         8192           ebecd7a45e3ea665fe38895ee400b87c
vdisk_diamond3a                         10240          afcec802224193a83eb0f6a22de19b8d
$ lssp -clustername galaxy
Pool                 Size(mb)  Free(mb)  LUs       Type   PoolID
atlantic             47552     17945     5         CLPOOL 9523836302054477476
$ lspv -size
NAME          PVID                                SIZE(megabytes)
hdisk0        00f60271506a4a40                    140013
hdisk1        00f60271652513ca                    140013
caa_private0  00f6027150d1b7fa                    20480
cldisk1       none                                16384
cldisk3       none                                16384
cldisk2       none                                15158
```

47522 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

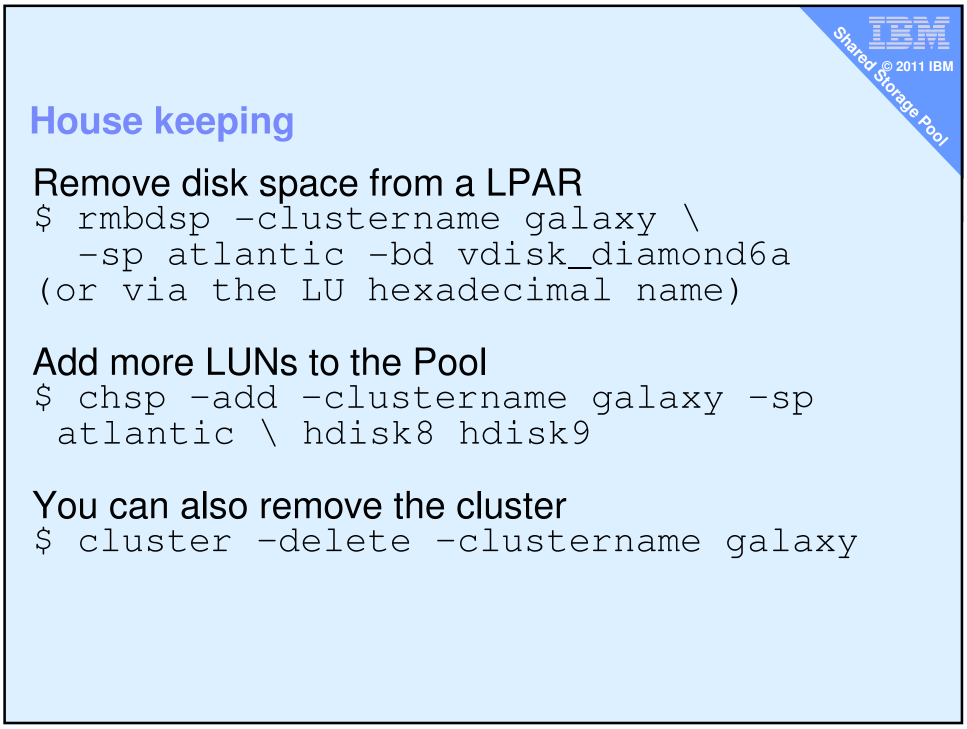
Monitoring: topas on VIOS then “D”



```
Topas Monitor for host:    diamondvios1Interval:   2    Fri Jan 14 14:46:00 2011
===============================================================================
Disk      Busy%  KBPS    TPS   KB-R   ART   MRT  KB-W   AWT   MWT   AQW    AQD
cldisk2    41.0 17.6K  493.0    0.0   0.0 174.6 17.6K   1.1  14.6   0.0    0.0
cldisk3    34.0 20.0K  160.0    0.0   0.0 186.4 20.0K   2.9  13.1   0.0    0.0
cldisk1     3.0  24.0    6.0    0.0   0.0 112.0  24.0   0.6 158.8   0.0    0.0
hdisk0      0.0   8.0    2.0    0.0   0.0  10.2   8.0   4.1  64.2   0.0    0.0
caa_priva   0.0  17.0    5.0    9.0   0.1   2.1   8.0   0.5   6.9   0.0    0.0
hdisk1      0.0   0.0    0.0    0.0   0.0   0.0   0.0   0.0   7.2   0.0    0.0
cd0         0.0   0.0    0.0    0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0
```

One client VM running: yes >/dev/tmp/x

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



House keeping

Remove disk space from a LPAR

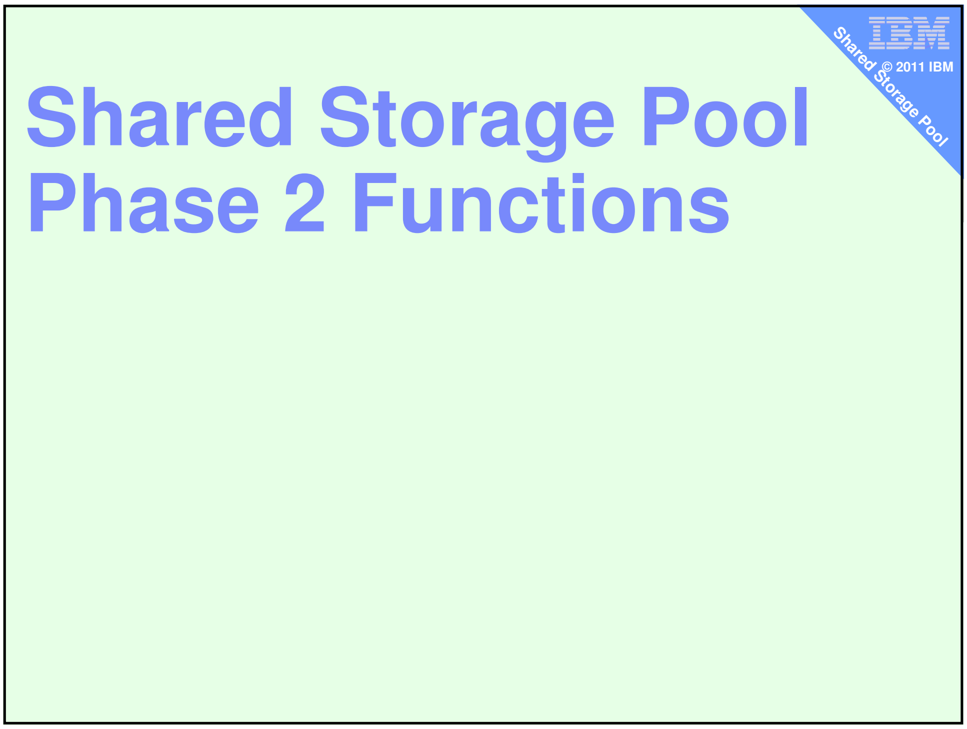
```
$ rmbdsp -clustername galaxy \
  -sp atlantic -bd vdisk_diamond6a
(or via the LU hexadecimal name)
```

Add more LUNs to the Pool

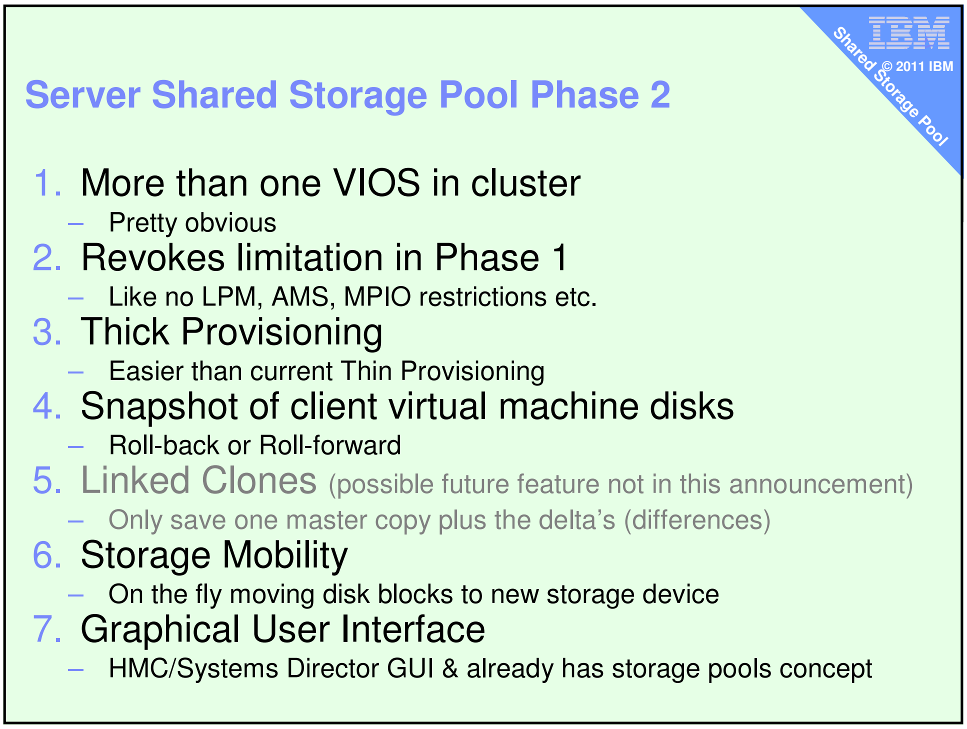
```
$ chsp -add -clustername galaxy -sp
 atlantic \ hdisk8 hdisk9
```

You can also remove the cluster

```
$ cluster -delete -clustername galaxy
```

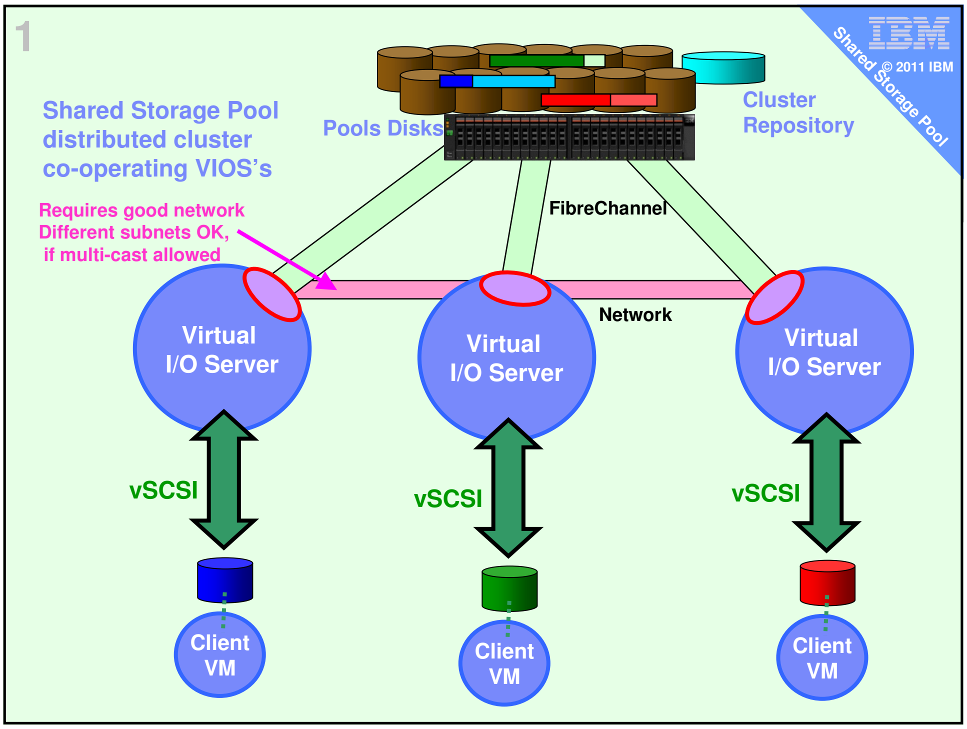


Shared Storage Pool Phase 2 Functions



Server Shared Storage Pool Phase 2

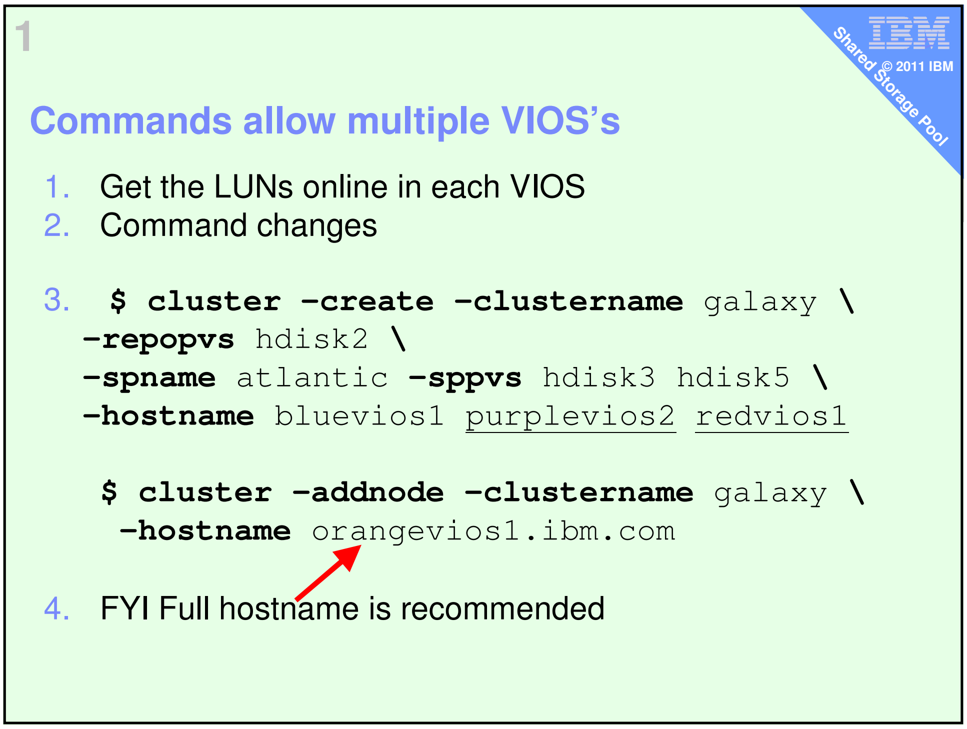
1. More than one VIOS in cluster
   - Pretty obvious
2. Revokes limitation in Phase 1
   - Like no LPM, AMS, MPIO restrictions etc.
3. Thick Provisioning
   - Easier than current Thin Provisioning
4. Snapshot of client virtual machine disks
   - Roll-back or Roll-forward
5. Linked Clones (possible future feature not in this announcement)
   - Only save one master copy plus the delta's (differences)
6. Storage Mobility
   - On the fly moving disk blocks to new storage device
7. Graphical User Interface
   - HMC/Systems Director GUI & already has storage pools concept



1

Shared Storage Pool distributed cluster co-operating VIOS's

Requires good network
Different subnets OK,
if multi-cast allowed



1

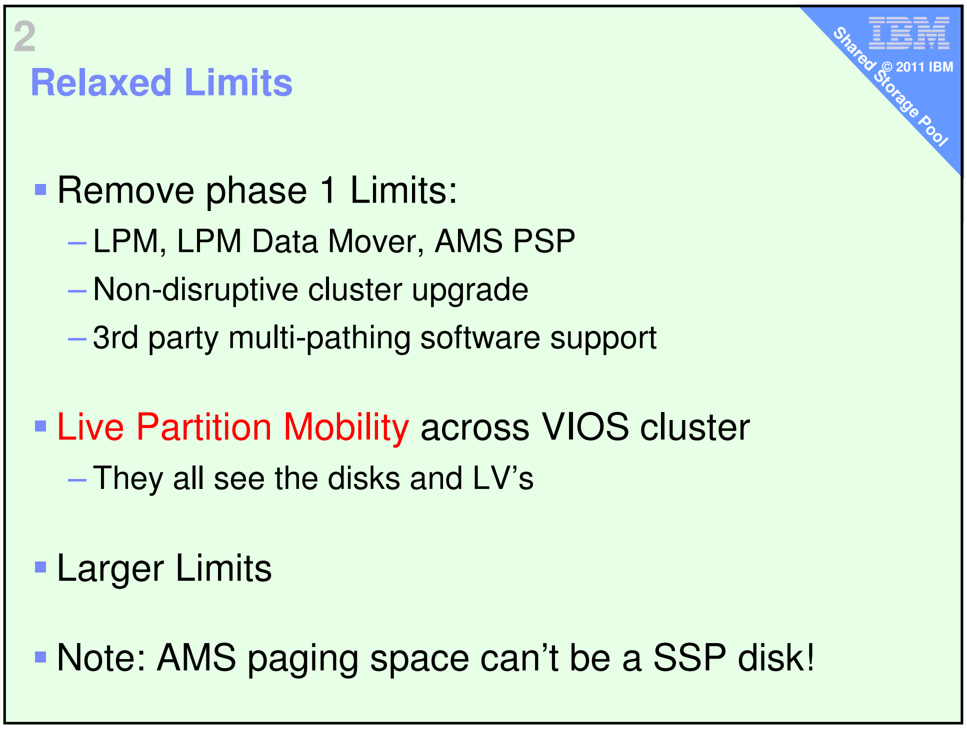
Commands allow multiple VIOS's

1. Get the LUNs online in each VIOS
2. Command changes
3.

```
 $ cluster -create -clustername galaxy \
-repopvs hdisk2 \
-spname atlantic -sppvs hdisk3 hdisk5 \
-hostname bluevios1 purplevios2 redvios1

  $ cluster -addnode -clustername galaxy \
   -hostname orangevios1.ibm.com
```

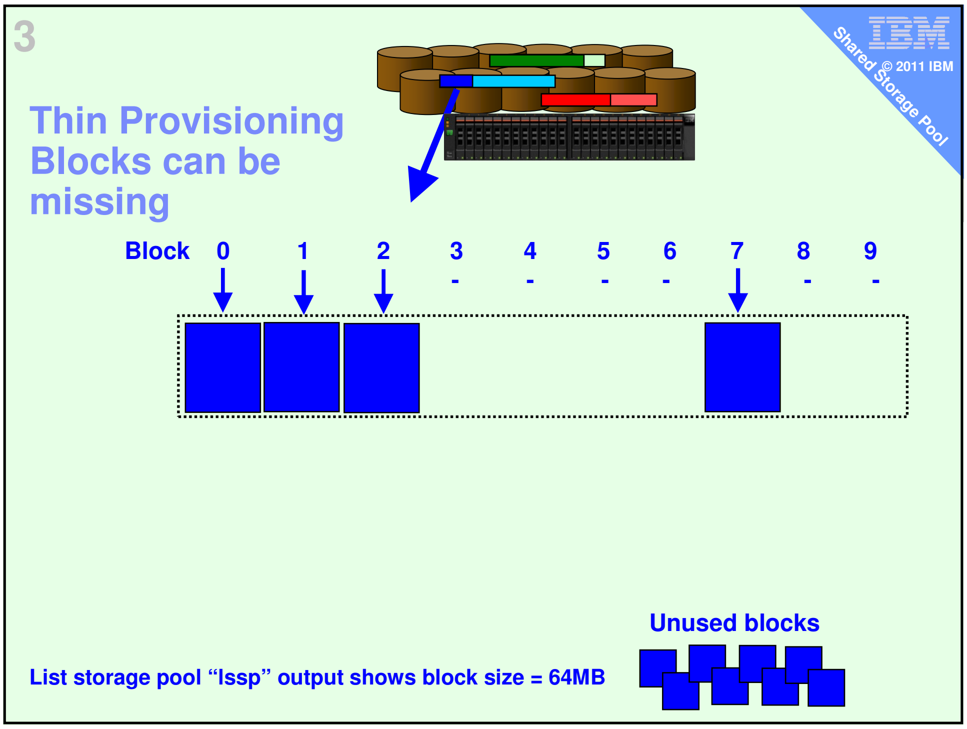
4. FYI Full hostname is recommended



2

Relaxed Limits

- Remove phase 1 Limits:
  - LPM, LPM Data Mover, AMS PSP
  - Non-disruptive cluster upgrade
  - 3rd party multi-pathing software support
- Live Partition Mobility across VIOS cluster
  - They all see the disks and LV's
- Larger Limits
- Note: AMS paging space can't be a SSP disk!



3

Shared Storage Pool

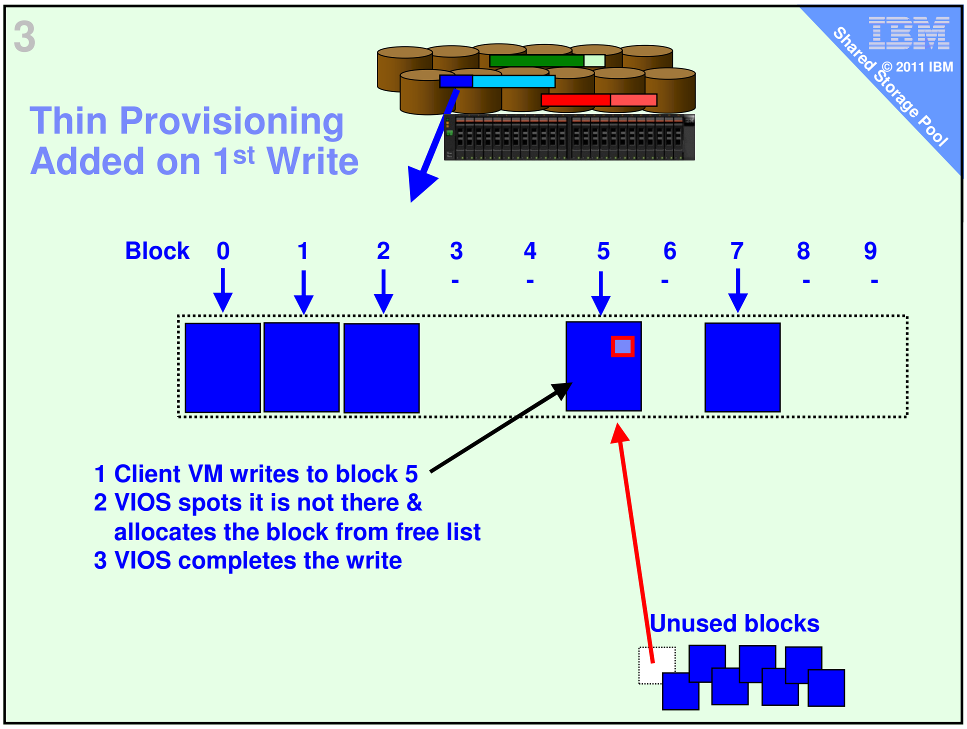
© 2011 IBM

Thin Provisioning Blocks can be missing

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

Unused blocks

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



3

Shared Storage Pool

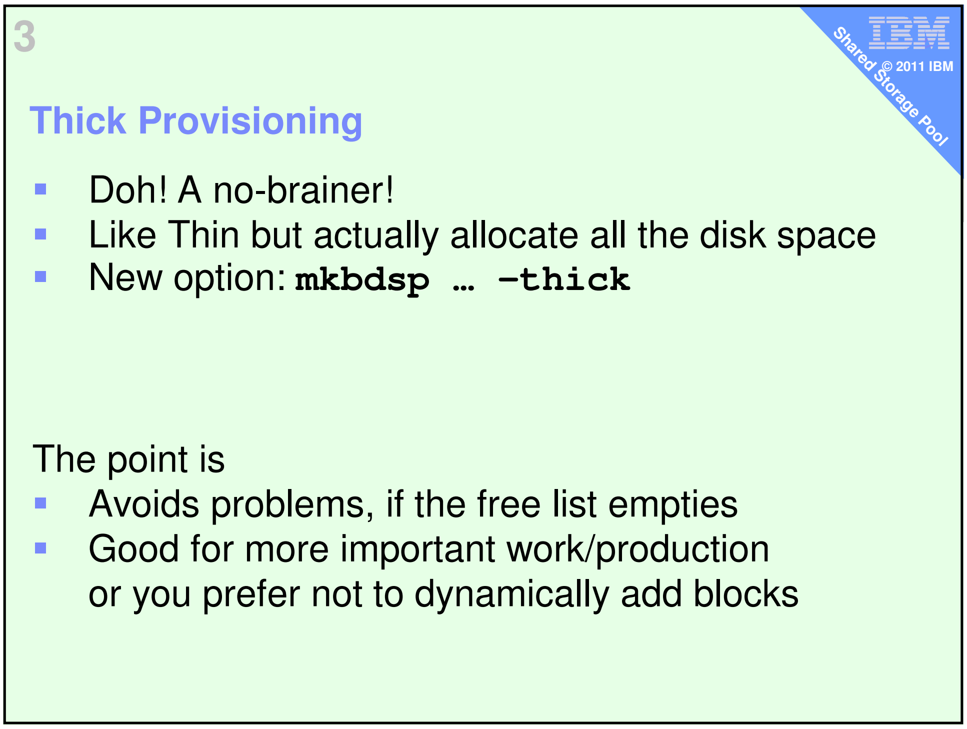
© 2011 IBM

Thin Provisioning Added on 1st Write

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

1 Client VM writes to block 5
2 VIOS spots it is not there &
 allocates the block from free list
3 VIOS completes the write

Unused blocks



3

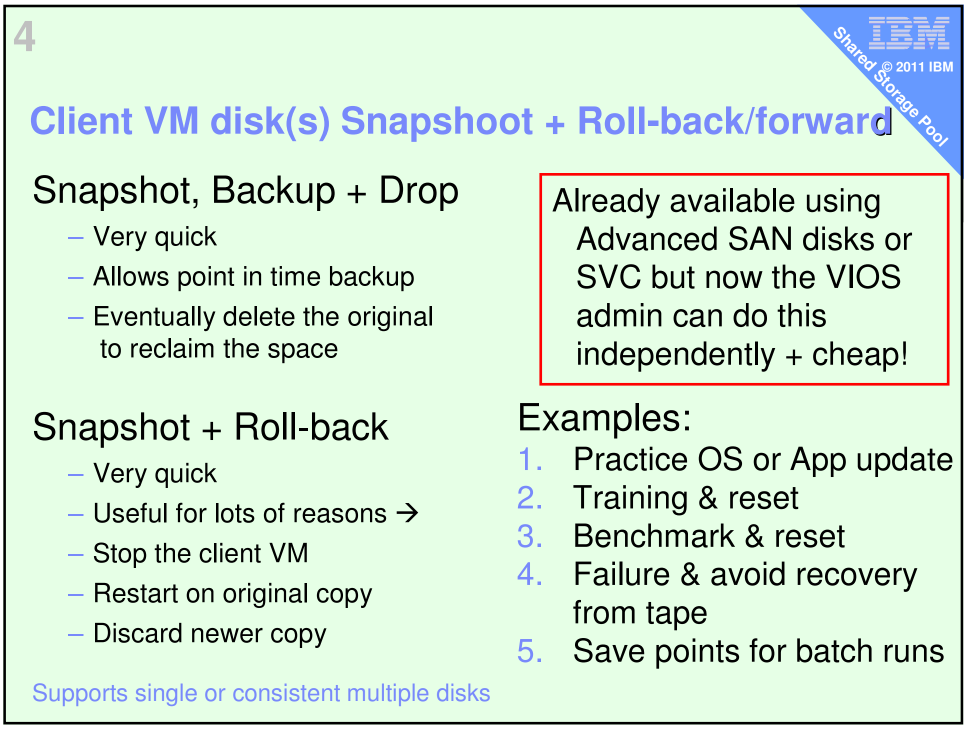
Shared Storage Pool © 2011 IBM

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



4

Shared Storage Pool © 2011 IBM

Client VM disk(s) Snapshoot + Roll-back/forward

Snapshot, Backup + Drop

- Very quick
- Allows point in time backup
- Eventually delete the original to reclaim the space

Snapshot + Roll-back

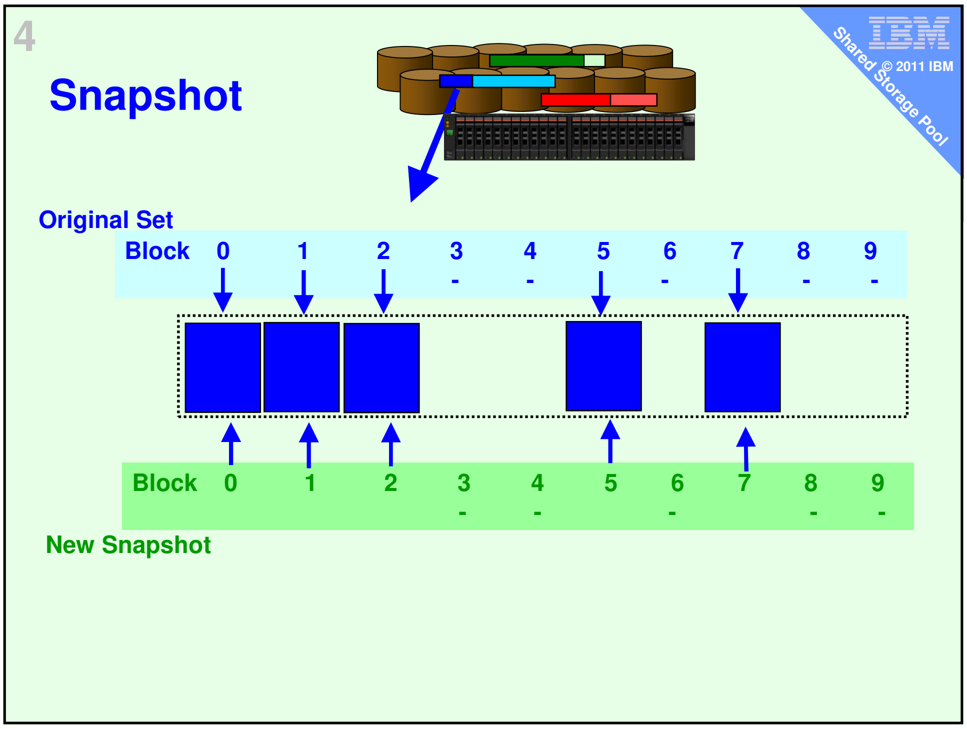
- Very quick
- Useful for lots of reasons →
- Stop the client VM
- Restart on original copy
- Discard newer copy

Supports single or consistent multiple disks

Already available using Advanced SAN disks or SVC but now the VIOS admin can do this independently + cheap!

Examples:

1. Practice OS or App update
2. Training & reset
3. Benchmark & reset
4. Failure & avoid recovery from tape
5. Save points for batch runs



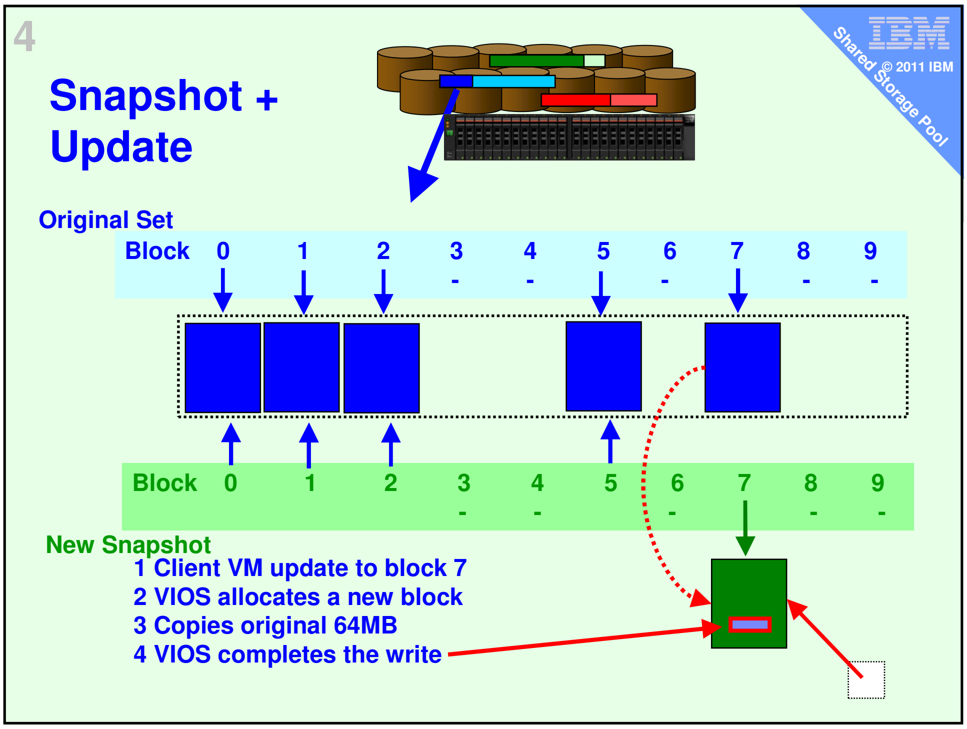
Snapshot

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 Snapshot



Snapshot + Update

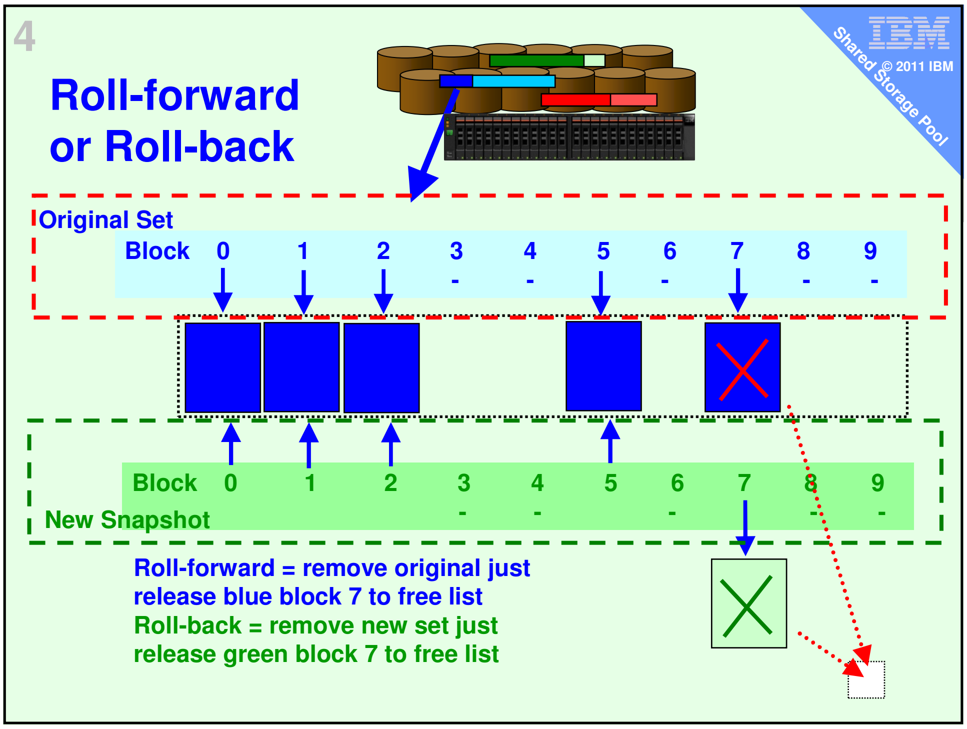
Original Set

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

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

New Snapshot

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



4

Shared Storage Pool © 2011 IBM

Roll-forward or Roll-back

Original Set

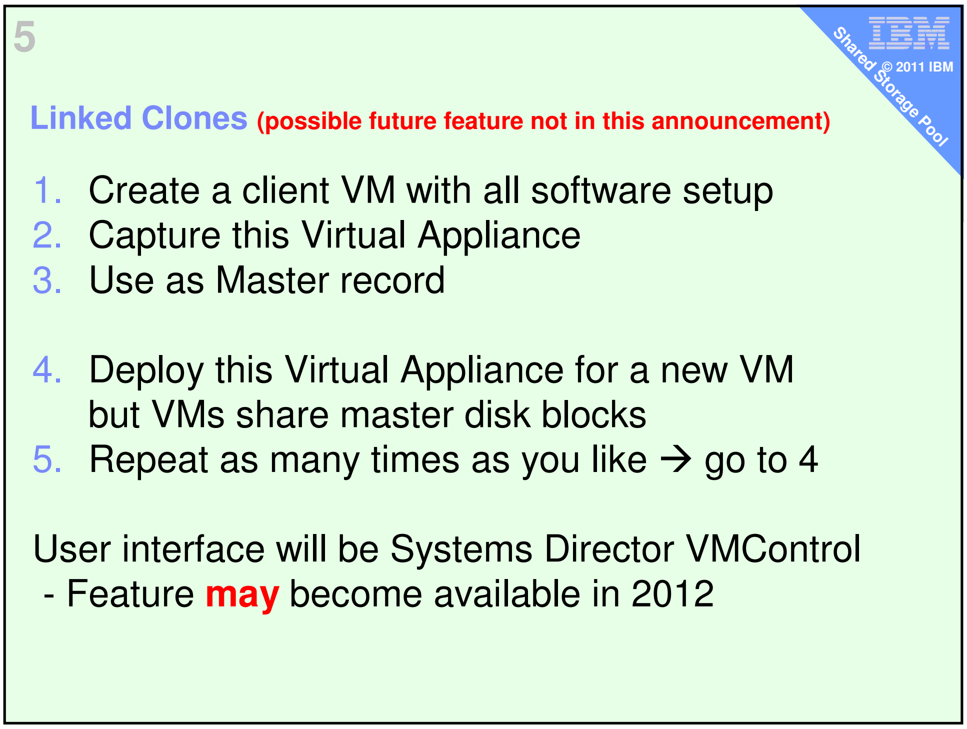
Block	0	1	2	3	4	5	6	7	8	9
				-	-		-		-	-

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

New Snapshot

Roll-forward = remove original just release blue block 7 to free list

Roll-back = remove new set just release green block 7 to free list



5

Shared Storage Pool © 2011 IBM

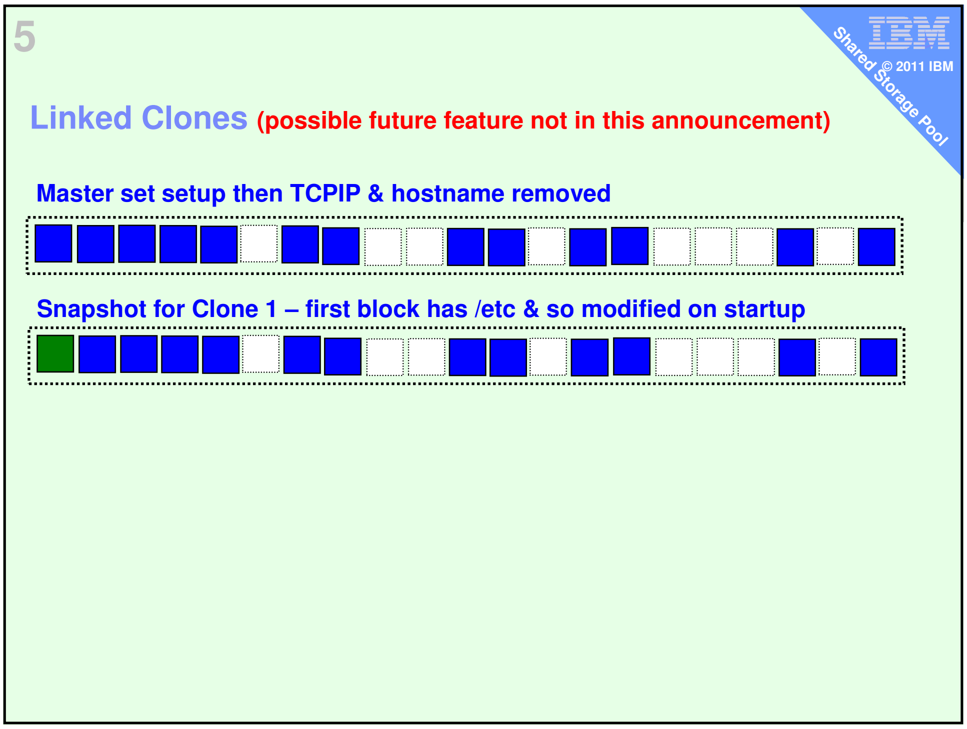
Linked Clones (possible future feature not in this announcement)

1. Create a client VM with all software setup
2. Capture this Virtual Appliance
3. Use as Master record

4. Deploy this Virtual Appliance for a new VM but VMs share master disk blocks
5. Repeat as many times as you like → go to 4

User interface will be Systems Director VMControl

- Feature **may** become available in 2012



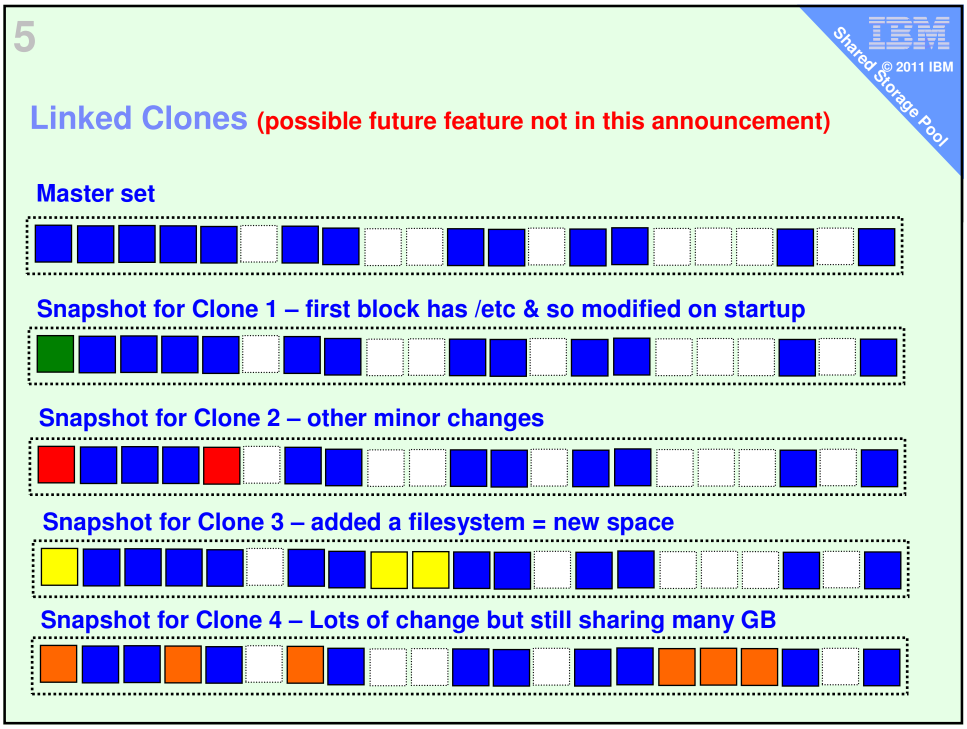
5

Shared Storage Pool © 2011 IBM

Linked Clones (possible future feature not in this announcement)

Master set setup then TCPIP & hostname removed

Snapshot for Clone 1 – first block has /etc & so modified on startup



5

Shared Storage Pool © 2011 IBM

Linked Clones (possible future feature not in this announcement)

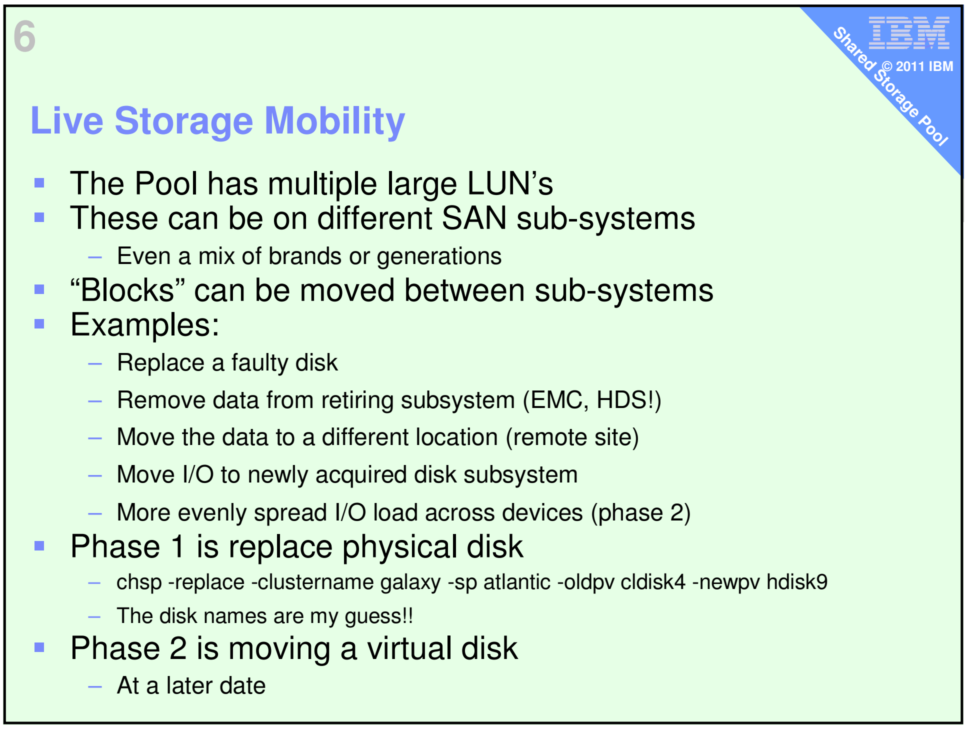
Master set

Snapshot for Clone 1 – first block has /etc & so modified on startup

Snapshot for Clone 2 – other minor changes

Snapshot for Clone 3 – added a filesystem = new space

Snapshot for Clone 4 – Lots of change but still sharing many GB



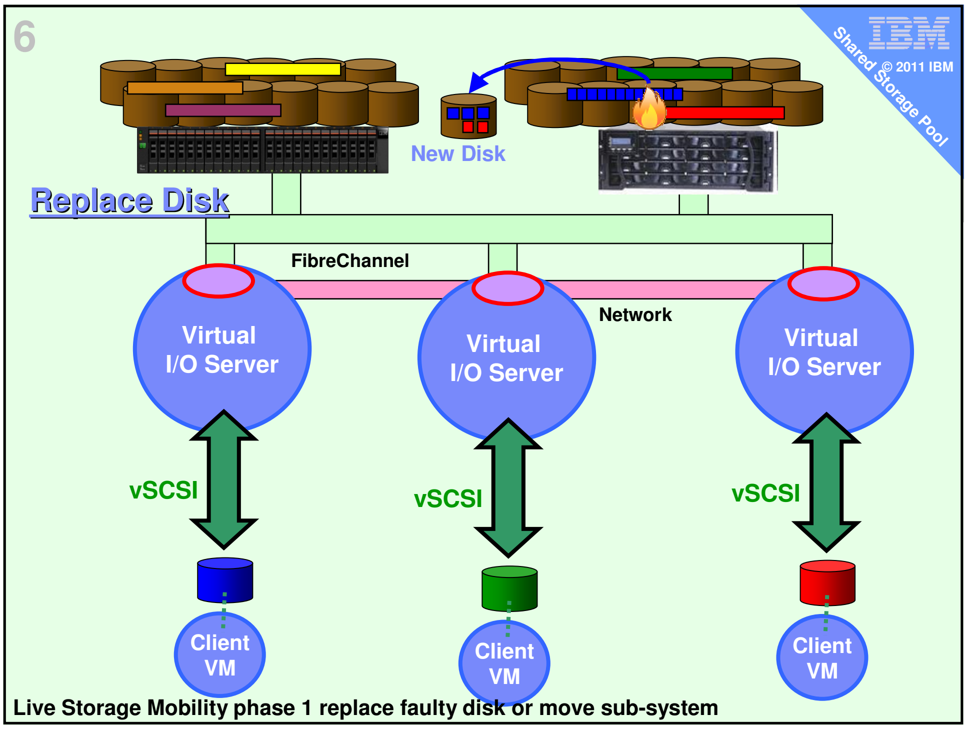
6

Shared Storage Pool

© 2011 IBM

Live Storage Mobility

- The Pool has multiple large LUN’s
- These can be on different SAN sub-systems
  - Even a mix of brands or generations
- “Blocks” can be moved between sub-systems
- Examples:
  - Replace a faulty disk
  - Remove data from retiring subsystem (EMC, HDS!)
  - Move the data to a different location (remote site)
  - Move I/O to newly acquired disk subsystem
  - More evenly spread I/O load across devices (phase 2)
- Phase 1 is replace physical disk
  - chsp -replace -clustername galaxy -sp atlantic -oldpv cldisk4 -newpv hdisk9
  - The disk names are my guess!!
- Phase 2 is moving a virtual disk
  - At a later date



6

Shared Storage Pool

© 2011 IBM

Replace Disk

New Disk

FibreChannel

Network

Virtual I/O Server

Virtual I/O Server

Virtual I/O Server

vSCSI

vSCSI

vSCSI

Client VM

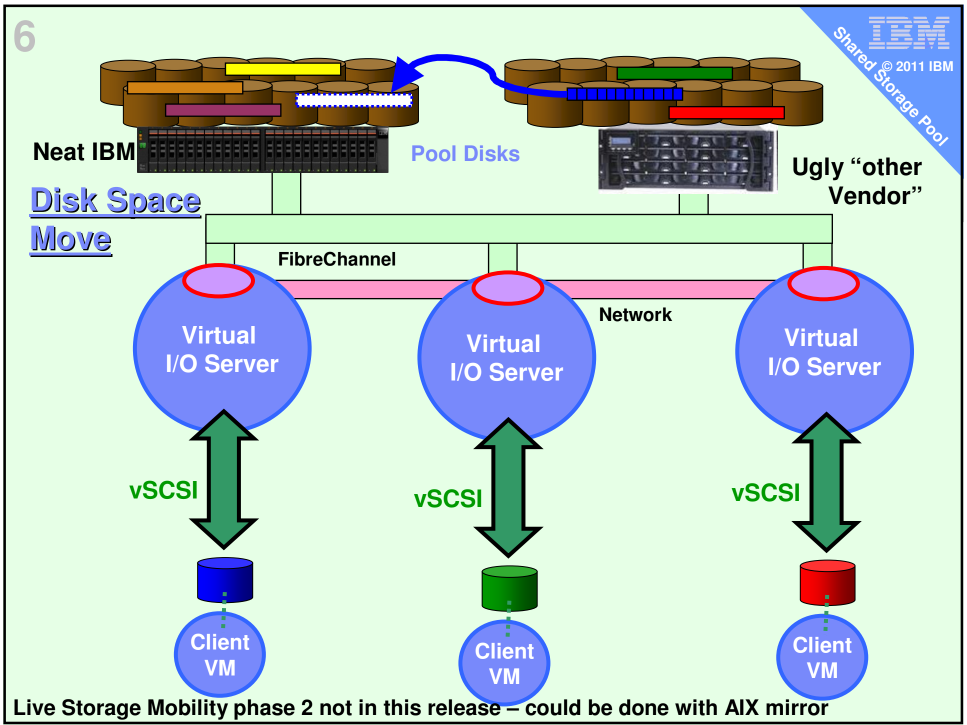
Client VM

Client VM

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

6

Disk Space Move



Neat IBM

Pool Disks

Ugly "other Vendor"

FibreChannel

Network

Virtual I/O Server

Virtual I/O Server

Virtual I/O Server

vSCSI

vSCSI

vSCSI

Client VM

Client VM

Client VM

Live Storage Mobility phase 2 not in this release – could be done with AIX mirror

GUI

7

- Systems Director
  - Adds a GUI & it will appear as a further Storage Pool type
  - VMControl Image Management
    - Provisioning in seconds
  - Cluster Management
    - LPM cluster load balancing with no SAN team help
- Shipped with a Systems Director upgrade
  - Not part of the VIOS package
- Some setup functions will be command line only

Shared Storage Pool

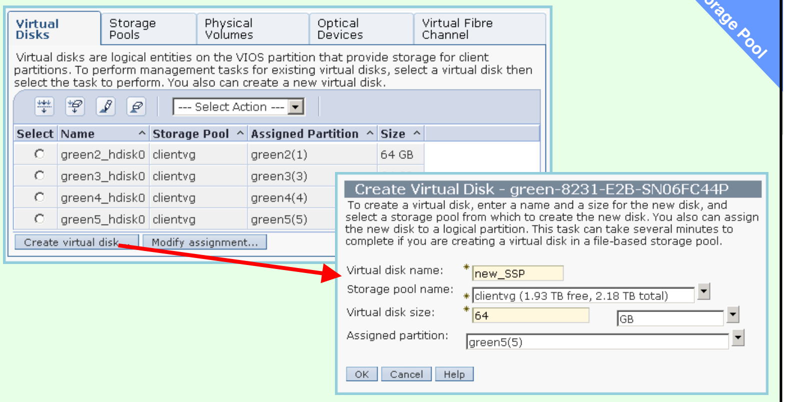
© 2011 IBM

GUI

7

- HMC

HMC Virtual Storage Management



Virtual Disks | Storage Pools | Physical Volumes | Optical Devices | Virtual Fibre Channel

Virtual disks are logical entities on the VIOS partition that provide storage for client partitions. To perform management tasks for existing virtual disks, select a virtual disk then select the task to perform. You also can create a new virtual disk.

--- Select Action ---

Select	Name	Storage Pool	Assigned Partition	Size
○	green2_hdisk0	clientvg	green2(1)	64 GB
○	green3_hdisk0	clientvg	green3(3)	
○	green4_hdisk0	clientvg	green4(4)	
○	green5_hdisk0	clientvg	green5(5)	

Create virtual disk... | Modify assignment...

Create Virtual Disk - green-8231-E2B-SN06FC44P

To create a virtual disk, enter a name and a size for the new disk, and select a storage pool from which to create the new disk. You also can assign the new disk to a logical partition. This task can take several minutes to complete if you are creating a virtual disk in a file-based storage pool.

Virtual disk name: * new_SSP

Storage pool name: * clientvg (1.93 TB free, 2.18 TB total)

Virtual disk size: * 64 GB

Assigned partition: green5(5)

OK | Cancel | Help

- Shipped with HMC upgrade
  - Not part of the VIOS package
- Some setup functions will be command line only

Shared Storage Pool

© 2011 IBM

What if you loose 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

  viosbr -backup -clustername clusterName -file FileName
      [-frequency daily|weekly|monthly [-numfiles fileCount]]
- View

  viosbr -view -file FileName -clustername clusterName
      [-type devType][-detail | -mapping]
- Restore

  viosbr -restore -file FileName -skipcluster

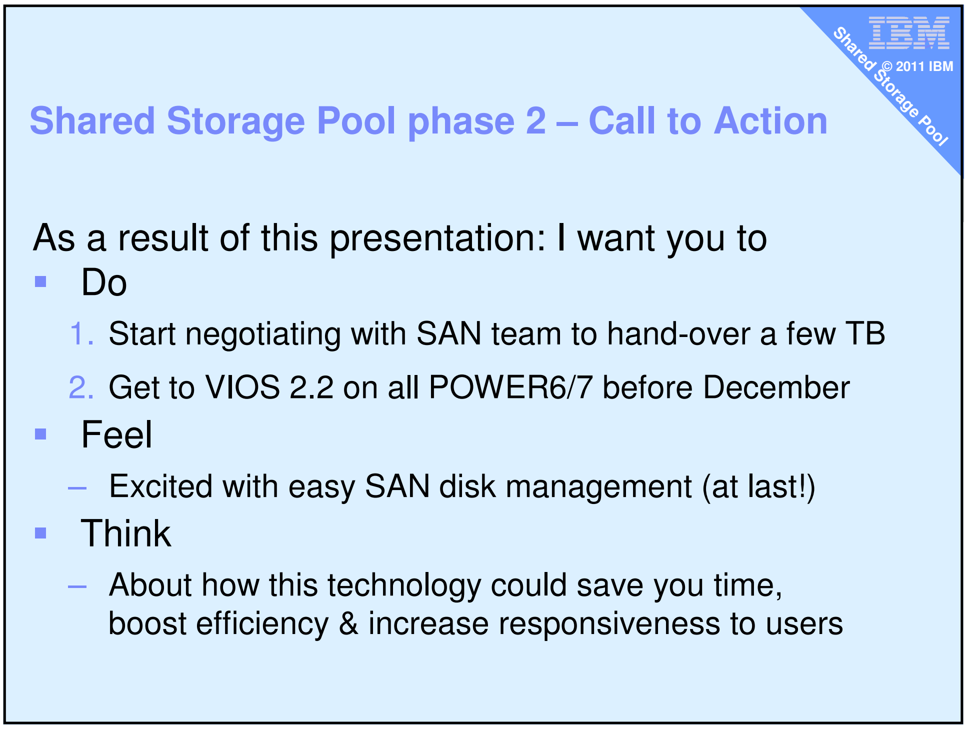
  viosbr -restore -clustername clusterName -file FileName -subfile NodeFile
      [-validate | -inter | -force][-type devType][-skipcluster]

  viosbr -restore -clustername clusterName -file FileName -repopvs list_of_disks
      [-validate | -inter | -force][-type devType][-currentdb]

  viosbr -restore -clustername clusterName -file FileName
      -subfile NodeFile –xmlvtds

  viosbr -recoverdb -clustername clusterName [-file FileName]

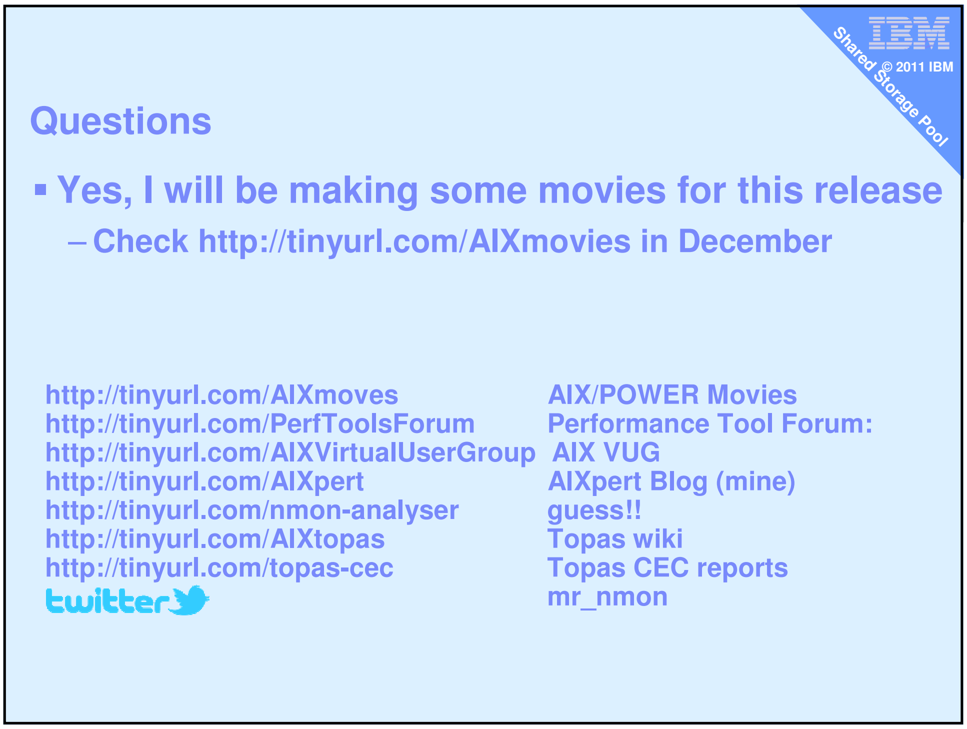
  viosbr -migrate -file FileName



Shared Storage Pool phase 2 – Call to Action

As a result of this presentation: I want you to

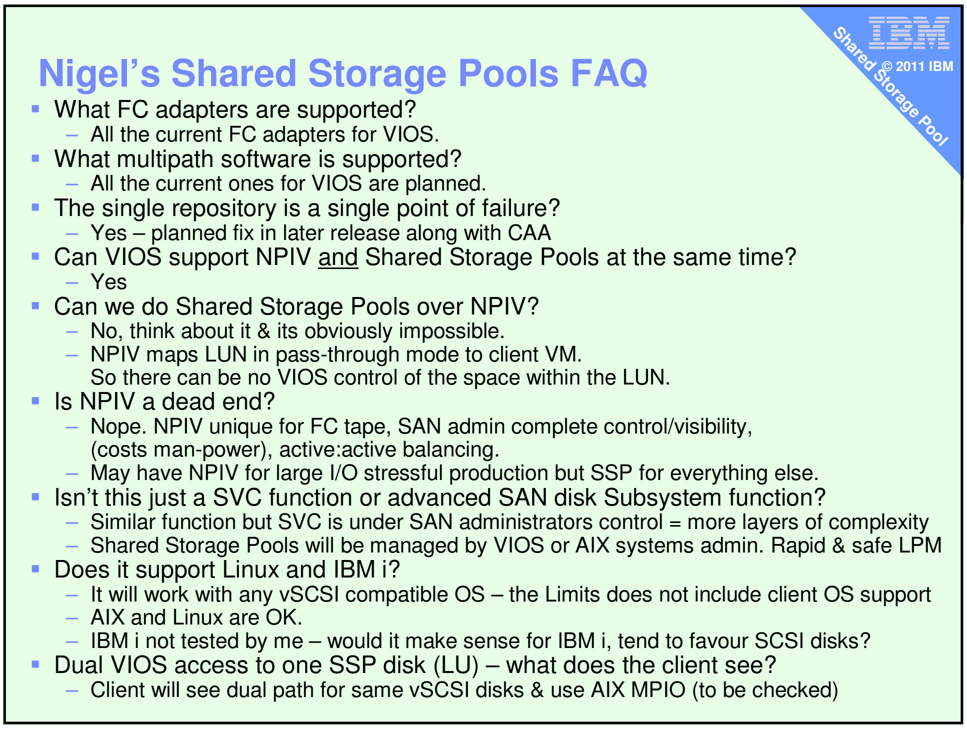
- Do
  1. Start negotiating with SAN team to hand-over a few TB
  2. Get to VIOS 2.2 on all POWER6/7 before December
- Feel
  - Excited with easy SAN disk management (at last!)
- Think
  - About how this technology could save you time, boost efficiency & increase responsiveness to users



Questions

- **Yes, I will be making some movies for this release**
  - **Check http://tinyurl.com/AIXmovies in December**

http://tinyurl.com/AIXmoves	AIX/POWER Movies
http://tinyurl.com/PerfToolsForum	Performance Tool Forum:
http://tinyurl.com/AIXVirtualUserGroup	AIX VUG
http://tinyurl.com/AIXpert	AIXpert Blog (mine)
http://tinyurl.com/nmon-analyser	guess!!
http://tinyurl.com/AIXtopas	Topas wiki
http://tinyurl.com/topas-cec	Topas CEC reports
twitter	mr_nmon



Shared Storage Pool

© 2011 IBM

Nigel’s Shared Storage Pools FAQ

- What FC adapters are supported?
  - All the current FC adapters for VIOS.
- What multipath software is supported?
  - All the current ones for VIOS are planned.
- The single repository is a single point of failure?
  - Yes – planned fix in later release along with CAA
- Can VIOS support NPIV and Shared Storage Pools at the same time?
  - Yes
- Can we do Shared Storage Pools over NPIV?
  - No, think about it & its obviously impossible.
  - NPIV maps LUN in pass-through mode to client VM.
    So there can be no VIOS control of the space within the LUN.
- Is NPIV a dead end?
  - Nope. NPIV unique for FC tape, SAN admin complete control/visibility, (costs man-power), active:active balancing.
  - May have NPIV for large I/O stressful production but SSP for everything else.
- Isn’t this just a SVC function or advanced SAN disk Subsystem function?
  - Similar function but SVC is under SAN administrators control = more layers of complexity
  - Shared Storage Pools will be managed by VIOS or AIX systems admin. Rapid & safe LPM
- Does it support Linux and IBM i?
  - It will work with any vSCSI compatible OS – the Limits does not include client OS support
  - AIX and Linux are OK.
  - IBM i not tested by me – would it make sense for IBM i, tend to favour SCSI disks?
- Dual VIOS access to one SSP disk (LU) – what does the client see?
  - Client will see dual path for same vSCSI disks & use AIX MPIO (to be checked)