AIX higher availability using SAN services

Skill Level: Introductory

Julie Craft (jcraft@us.ibm.com)
Architect
IBM

Sanket Rathi (sanrathi@in.ibm.com)
AIX Storage Device Driver Systems Programmer
IBM

Anbazhagan Mani (manbazha@in.ibm.com)
AIX Storage Device Driver Systems Programmer
IBM Software Labs, India

Chris Schwendiman (schwendi@us.ibm.com)
AIX Storage Device Driver Systems Programmer
IBM

Jim Pafumi (jpafumi@us.ibm.com)
Senior engineer
IBM

Gero Schmidt (GEROSCH@de.ibm.com)
Senior engineer
IBM

Nick Ham (nsham@us.ibm.com)
Software engineer
IBM

01 Sep 2009

Updated 10 Nov 2010

Learn the scenarios in which remapping, copying, and reuse of SAN disks is allowed

© Copyright IBM Corporation 2009, 2010. All rights reserved.

and supported. More easily switch AIX® environments from one system to another and help achieve higher availability and reduced down time. These scenarios also allow for fast deployment of new systems using cloning.

The support of these scenarios in which remapping, copying, and reuse of SAN disks is allowed and supported has never been officially documented. There have been some documents and IBM® Redbooks® that have claimed support for specific scenarios, but they do not list the specific steps or restrictions.

The scenarios detailed here guide systems administrators through the steps taken to achieve the specific environment desired. They also attempt to explain why the setup must be followed to achieve the desired results. If the steps are not followed, in some cases the system may not boot.

This article will also document further scenarios as they become supported.

Introduction

IBM® System p® systems are designed to offer the highest stand-alone availability in the industry. Enterprises must occasionally restructure their infrastructure to meet new IT requirements and handle scheduled outages (such as power outages). Today, even the smallest of IBM System p systems run logical partitions and there is a need to move the logical partitions to other available systems to avoid application down time.

IBM System p continue to introduce innovative technologies to handle such scenarios and reduce down time. Live Partition Mobility allows moving logical partitions around such that previously disruptive operations on the machine can be performed when it bests suits you. Live Partition Mobility helps to meet increasingly stringent service-level agreements (SLAs) because it allows you to proactively move running partitions and applications from one server to another. Live Partition Mobility requires a specific hardware and microcode configuration that is currently only available on POWER6®-based systems.

But, some customers using POWER5™-based systems want a solution to allow them to move their AIX environments from one System p server to another. This article explains scenarios in which a switch over of the SAN Disks (including the operating system) from a storage subsystem can be performed when using a PowerVM™ Virtual I/O server environment. This article also documents the use of 'flash-copy' services to create a backup disk to be used for system recovery.

It is very important to note while Live Partition Mobility supports active movement of logical partitions, the scenarios explained here support movement or backup of the AIX environment only after shutdown and hence involves some downtime. Care

© Copyright IBM Corporation 2009, 2010. All rights reserved.

should be taken in environments that cannot afford any downtime. Technologies such as IBM High Availability Cluster Multiprocessing (HACMP) should be considered in these situations to handle unplanned outages and to improve system availability by providing on demand failover.

PowerVM virtualization technology (previously called Advanced Power Virtualization) is a combination of hardware and software that supports and manages the virtual environments on POWER® systems. It contains major tools such as Virtual I/O Server (VIOS) to simplify and optimize your IT infrastructure.

Scenario 1. Switching AIX LPARs with SAN boot devices and VIOS

Figure 1. Test scenario setup

© Copyright IBM Corporation 2009, 2010. All rights reserved.

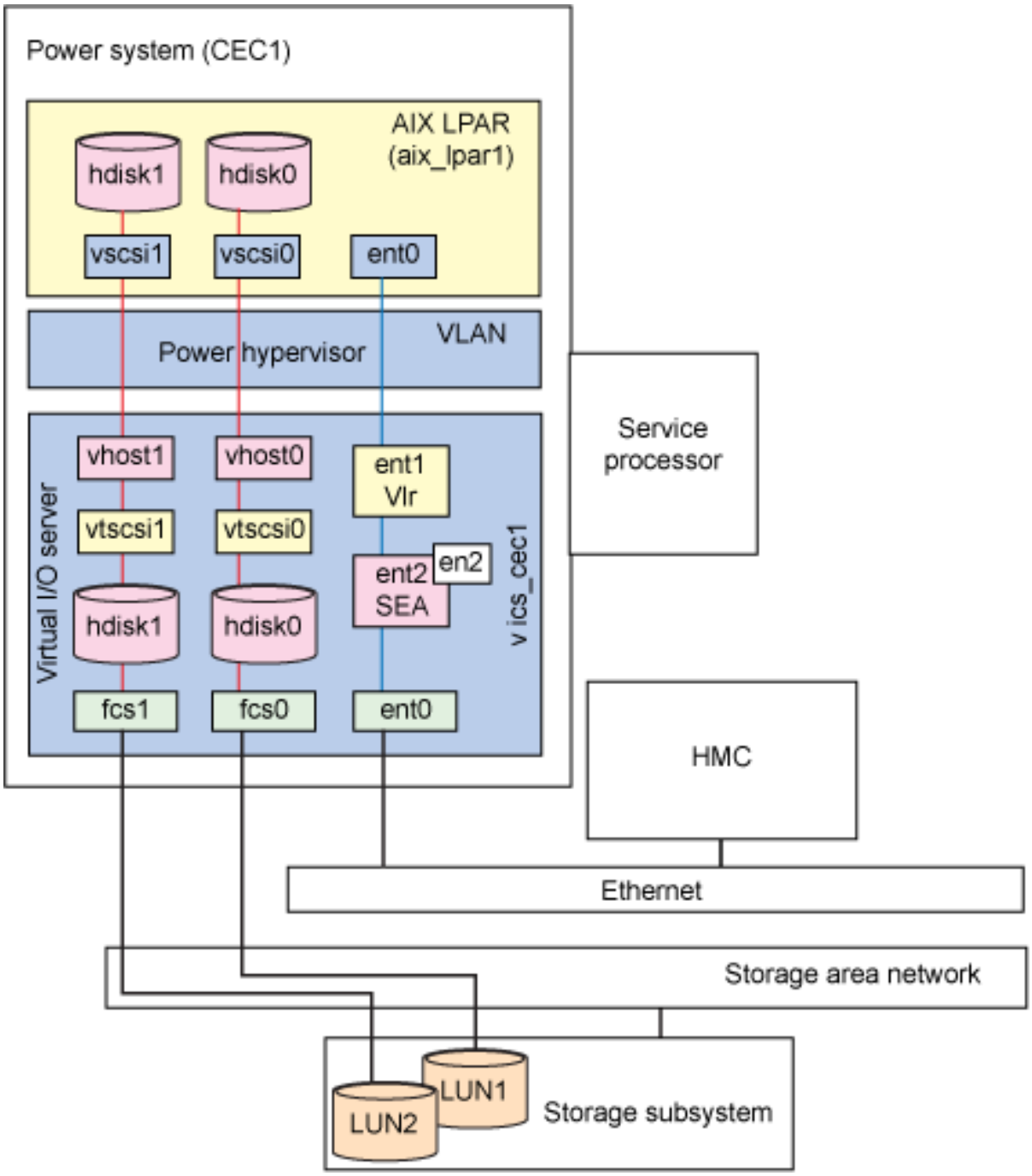


Figure 1 explains the test scenario setup using a Power5 system (CEC1).

- On the Virtual I/O Server (VIOS) partition, vios_cec1, we have two fibre channel adapters.
- Two LUNs from SAN storage are mapped to the two fibre channel adapters on the VIOS partition.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

- LUN1 is mapped to fcs0 and LUN2 is mapped to fcs1.
- On VIOS, LUN1 is available as hdisk0, and LUN2 is available as hdisk1.
- Two virtual SCSI server adapters (vhost0 and vhost1) are created on the VIOS lpar.
- hdisk0 has been assigned as the backing device for vhost0 and hdisk1 as the backing device for vhost1. As a result, vtscsi0 and vtscsi1 is available on VIOS.

Please note this is only a sample configuration. A realistic configuration is nearly 80 LUNs per adapter.

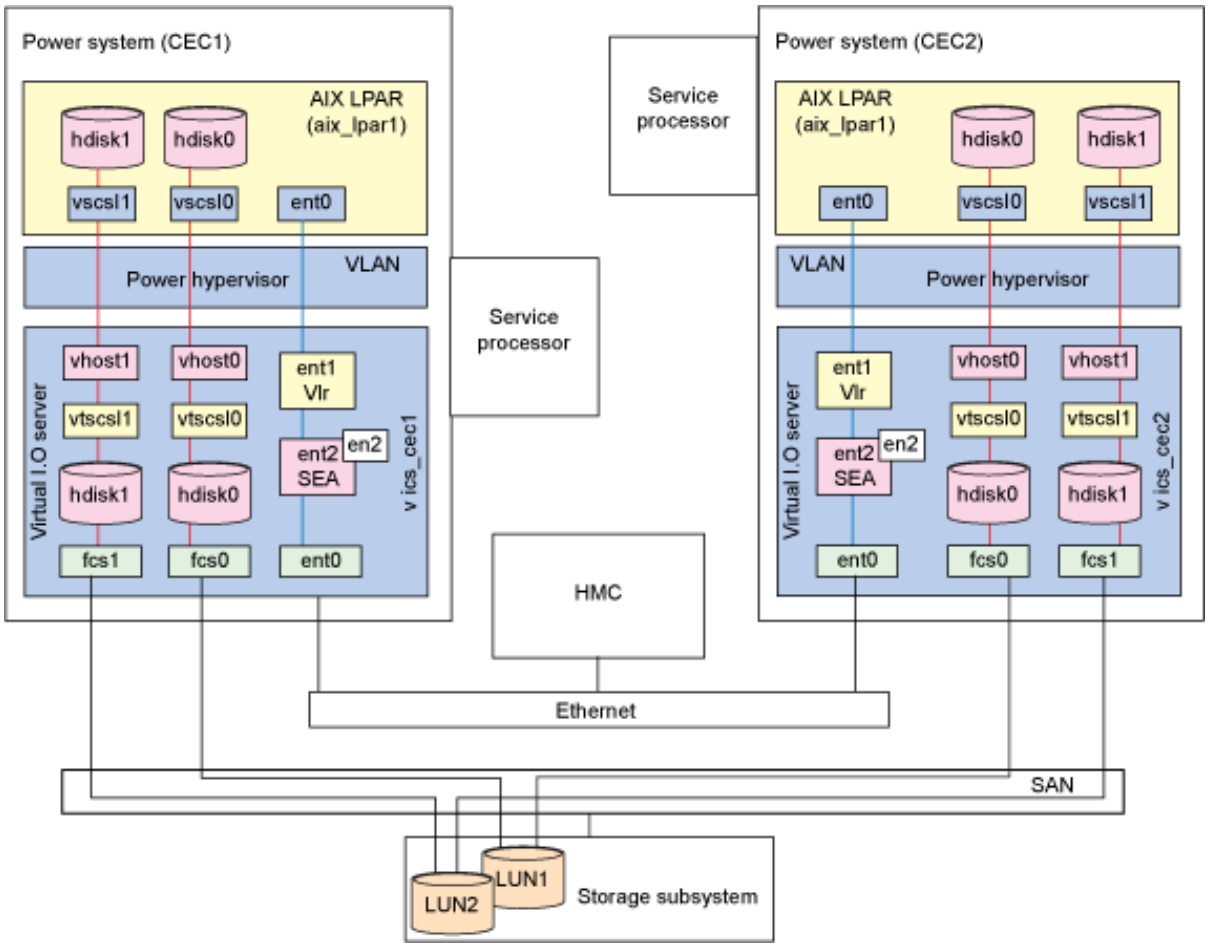
An AIX partition is created and it has the two hdisks exported using VIOS and is available as hdisk0 and hdisk1. Networking is provided through the Shared Ethernet adapter. The physical ethernet adapter (ent0) is used as a trunk adapter to provide bridging and Virtual Ethernet adapter (ent1) is used to create Shared Ethernet adapter (ent2). Interface en2 is configured on top of ent2 and a valid IP address is assigned to en2. On the AIX partition, an interface (not shown in the figure above) en0 can be created on top of the Virtual Ethernet adapter ent0 and a valid IP address can be assigned to it.

A situation could occur where a shutdown of the POWER system CEC1 is required, which would disrupt application availability. It is desirable to switch the AIX partition to another system that is available and hence the application running on the AIX partition can be made available to end users.

As shown in Figure 2, a system (CEC2) is connected to the same network and has a VIOS partition (vios_cec2) that has access to the same SAN disks that are accessed by the VIOS partition (vios_cec1). In this scenario, it is possible to shutdown the AIX partition on CEC1 and boot the AIX partition on CEC2. The whole process can be achieved within a few minutes, thereby enabling simple high availability. HACMP has support statements for specific VSCSI configurations. You may want to access those support statements.

Figure 2. Switching LPAR using SAN boot device

© Copyright IBM Corporation 2009, 2010. All rights reserved.



Prerequisites and restrictions

The following prerequisites must be met before switching your AIX environment (including rootvg and other volume groups) from one CEC to another:

- Ensure that the system processor type and modes on both the CECs are compatible. For example, both POWER5 and POWER6 can operate on 64-bit mode and are therefore compatible. If the AIX environment is using the 64-bit kernel on the original system, then the target system will need 64-bit processors.
- Both the CECs are connected to the same the network or subnet. This is important because when you switch the rootvg over to another CEC, the network interface (IP address, etc) is retained and is restored when AIX is booted (after the switchover is complete). In the case of inappropriate configuration, problems can occur such as a boot failure, hang, or loss of system access.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

- The slot numbers of the virtual ethernet and virtual SCSI client adapters for the AIX client partition (on the partition profile) must match on both CECs. The virtual SCSI client and virtual SCSI server adapter mappings have to be the same on both CECs.
- Virtual I/O server versions on the two CECs have to be the same.
- Running all software on Virtual I/O servers at the same levels, for instance, SDDPCM or Powerpath.
- All the disks that are visible on the AIX client partition have to be virtual disks (exported from VIOS using virtual SCSI). All the disks that are exported to the AIX client partition from VIOS have to be SAN disks. No internal disks should be exported to the client partition. No logical volume-based or file system backing devices should be exported to the client partition.
- The SAN disks that are exported to the AIX client partition must be available to both the VIOS partitions on both the CECs. Any SAN zoning, LUN masking, or mapping that needs to be done on the storage subsystem needs to be done to make the same LUNs available to both the VIOS partitions. The VIOS itself can be installed on either the internal disk or SAN disk.
- The attribute reserve_policy for all the SAN disks on VIOS should be set to **no_reserve**. If 'no_reserve' is not set, then VIOS on the original CEC should be shut down before the switch over is done.
- On the Virtual I/O server, a Shared Ethernet Adapter has to be created so that layer-2 bridging is available for the virtual ethernet assigned to the AIX client partition.
- Ensure VLAN configuration is done appropriately for VIOS partitions on both the CECs.
- System clock on both the CECs should be set to same date/time.

Switching AIX environment from one CEC to another CEC

To switch an AIX environment from one CEC to another, do the following:

1. Make sure all the prerequisites are met. Do not perform a switch over if any of the prerequisites are not met.
2. On CEC1, shut down all applications on the AIX partition. This ensures that the data volume groups and the file systems are consistent.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

3. Shut down the AIX partition.

4. On CEC2, ensure the VIOS partition is up and running.

5. On CEC2, boot the AIX partition.

6. When AIX partition boots, the firmware may not recognize the boot disk. The boot device information is stored in NVRAM, so if the new system boots from a disk not installed using conventional installation methods on that system, the user needs to interact with the firmware System Management Services (SMS) menus in order to select the boot device. This can be accomplished through the SMS menus, which can be accessed from a virtual terminal on the HMC.

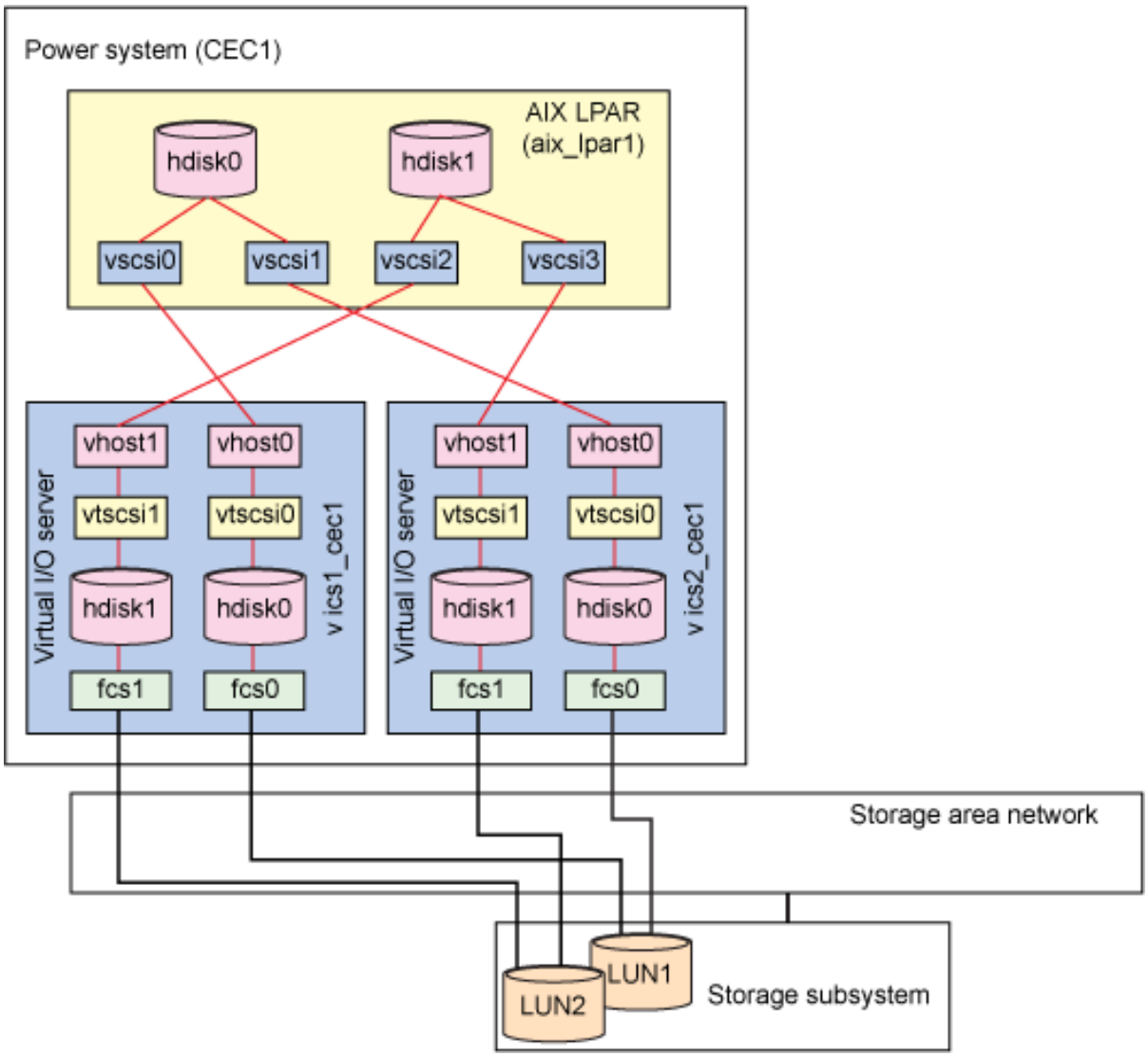
As with any production environment, we recommend that your system setup, including the switchover process, be thoroughly tested and documented before it is applied in production.

Scenario 2. Switching AIX LPARs with redundant VIOS

Figure 3 illustrates a switchover of an AIX environment using a single VIOS partition. However, many customers run two VIOS partitions to provide redundancy and fail-over to the AIX client partition. Multi-path I/O (MPIO) helps provide increased availability of virtual SCSI resources by providing redundant paths to the same resource (disk). For example, Figure 3 highlights a MPIO-based configuration for virtual disks (hdisk0 and hdisk1) on the AIX partition (aix_lpar1).

Figure 3. Multi-path I/O for AIX client partitions using redundant VIOS

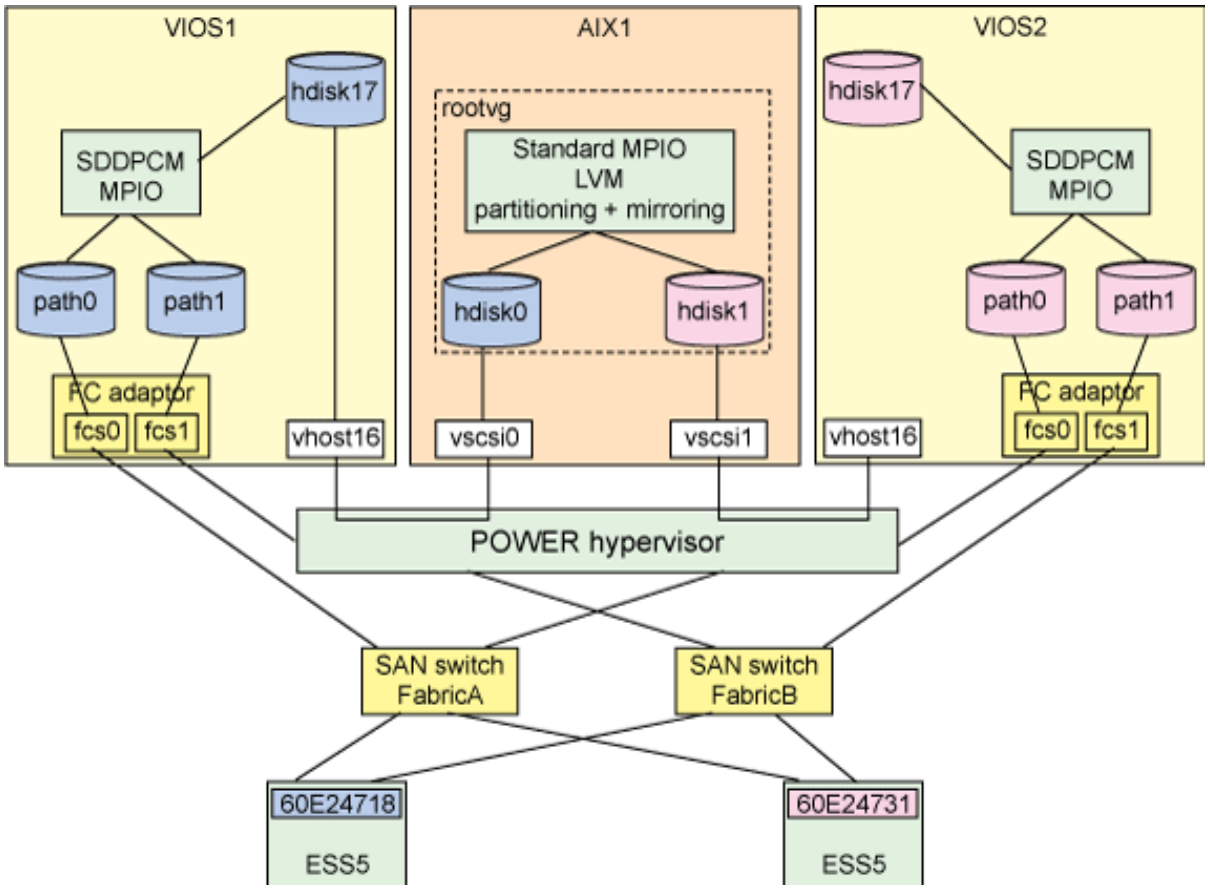
© Copyright IBM Corporation 2009, 2010. All rights reserved.



The switchover process explained in this article works fine in this scenario and has been tested under lab conditions. You have to ensure this exact setup (including two VIOS partitions with matching virtual slot numbers) is available on another CEC before the switchover is performed. There could be additional MPIO configuration on the VIOS partition using Subsystem Device Driver Path Control Module (SDDPCM). SDDPCM is a loadable path control module designed to support the multi-path configuration environment in the IBM TotalStorage® family. Such MPIO configuration on the Virtual I/O server logical partition should not impact switching over the AIX environment. All the prerequisites and restrictions that are listed in the previous sections are applicable for the MPIO scenario. Best practices would not recommend MPIO and mirroring on the same system. Typically, mirroring is used on virtual clients when LVs are exported, rather than LUNs.

Figure 4. Multi-path I/O for AIX client partitions using redundant VIOS and SDDPCM MPIO on VIOS

© Copyright IBM Corporation 2009, 2010. All rights reserved.



Scenario 3. Using FlashCopy for recovery

This scenario involves creating a copy of the AIX operating system environment to be used at a later time for recovery.

When using SAN-based boot devices, some environments require that a backup of AIX operating environment be performed. While there are many ways (for example, mksysb) to achieve backup of the AIX rootvg, a faster way to back up and restore rootvg using SAN devices is explained below.

Figure 5. Backing up rootvg

© Copyright IBM Corporation 2009, 2010. All rights reserved.

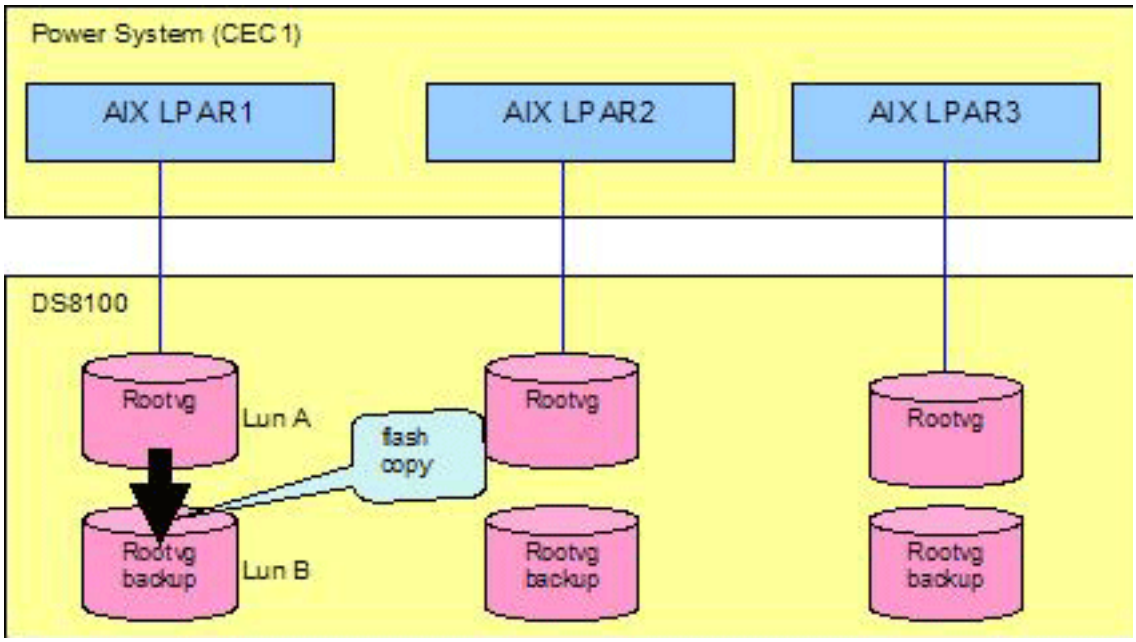
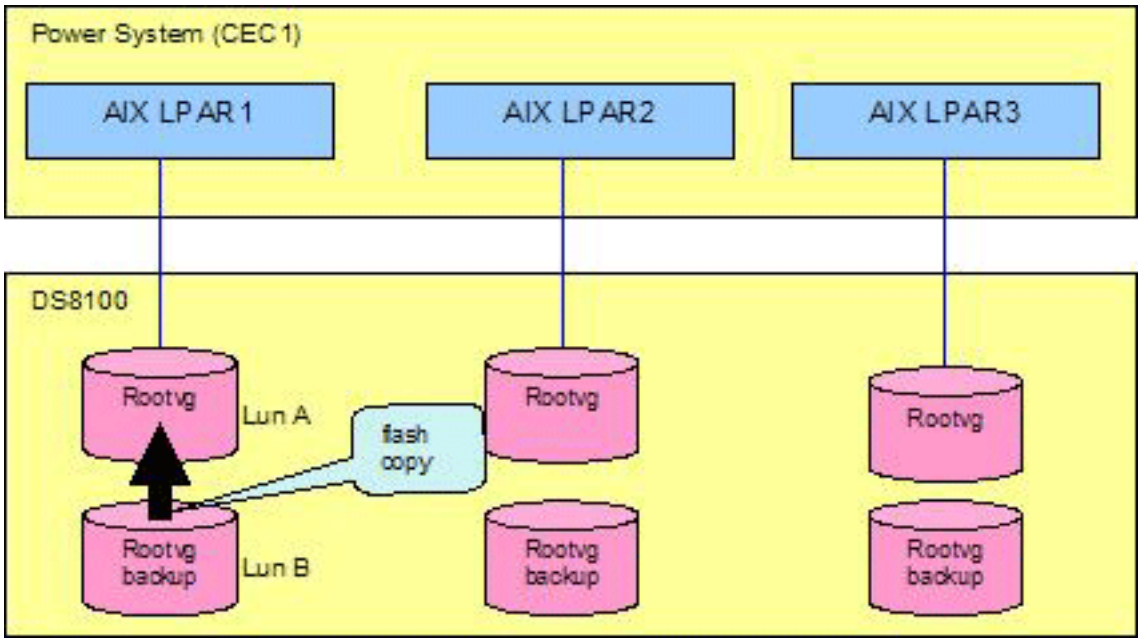


Figure 6. Restoring backed up rootvg



Scenario 4 – Booting from a Binary Copy of rootvg created by SAN services

This scenario is meant as a more generic statement of support of using SAN services, such as PPRC, global mirror, or flash copy. In these cases, the LUNs contain a binary copy of the rootvg but may or may not have the same UDID (Unique Device IDentifier). An update was made to the configuration code (IZ82061, IZ82080 - 5.3 TL 11, IZ81634, IZ81633 - 5.3 TL 12, IZ83828, IZ79602 – 6.1 TL3, IZ76109,

© Copyright IBM Corporation 2009, 2010. All rights reserved.

IZ77394 – 6.1 TL4, IZ83974, IZ78806 – 6.1 TL5, IZ76505, IZ73903 – 6.1 TL6) to allow booting a system with the same PVID but a different UDID. This will allow support for a PPRC or flash copy.

For example, in Scenario 3, if LUN A were to go bad, LUN B could be mapped to the LPAR and booted. The existing scenario may be easier to support in a customer environment, since the primary and backup LUNs are always the same, but now the option to use the backup copy is available.

It should be noted that some of the same issues and restrictions still exist when switching LPARs (or CECs) such that the hardware should be the same and connections should remain the same, especially for network connections.

Backup and recovery of AIX Environment using SAN devices

1. Shut down the AIX LPAR to ensure that the file systems are quiesced.
2. Use FlashCopy to copy the rootvg LUN to a backup LUN. It is not necessary to wait for the FlashCopy to complete. FlashCopy works in the background.
3. Reboot the AIX LPAR using the HMC.
4. In case there is a need to restore the backed-up rootvg, shut down the AIX LPAR. Use FlashCopy to copy the backed-up rootvg LUN (LUN B) to the original LUN (LUN A). The copied disk is used for system recovery; therefore, removing or reconfiguring network or device configuration information is not necessary.
5. Boot the AIX LPAR.

The Virtual I/O server is not required in this scenario. However, AIX client LPARs using virtual disks backed by LINs from the Virtual I/O server can also follow these backup and recovery procedures.

Issues with physical devices

You may have been told that switching SAN-backed storage from one system to another is not supported. The reason for this was because when the disk was booted from the new system, there were no guarantees that the configuration would be exactly the same and therefore some of the devices could become unavailable (put into the 'defined' state) and new ones created. This would lead to, at best, ghost devices, or, at worst, a system that would not boot.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

AIX development is currently working on a more flexible device configuration design to deal with more of the physical device issues, but until that time, these scenarios will not be supported with physical devices, except when dealing with the same system, as documented in Scenario 3.

Potential issues with migrating rootvg from physical to physical or virtual environment

Movement of rootvg disks from a physical environment to either another physical or virtual environment is not recommended due to the following:

1. Firmware doesn't recognize the boot disk. The boot device information is stored in NVRAM, so if the new system is to boot from a disk that was not installed using conventional installation methods on that system, then the user will need to interact with the firmware SMS menus in order to select the boot device. This can be accomplished through the SMS menus, which can be accessed from a vterm on the HMC.

2. The console device may be lost and need to be reselected. If no one is there to respond to the console selection prompt on boot, then after some period of time the system will continue booting without a console. To the user this may "look" as though it is hung on boot.

3. The rootvg must contain all of the required support for the new system. This includes device support for any new devices on the new system, as well as support for the model itself. If the level of AIX in the rootvg does not support the system or devices, problems may manifest themselves as boot failures or missing device support.

4. The new system must be capable of running the AIX as installed in the rootvg. For example, if the rootvg is using the 64-bit kernel, then the new system needs 64-bit processors.

5. Device names may change. The set of bus and adapter devices that are discovered in the new system most likely will not match the devices in the rootvg's ODM database. This will result in new device instances being created in ODM and the devices from the original system listed as "defined". For example, where the original system might have buses pci0 and pci1 and SCSI adapter scsi0, the new system will list them in the defined state with new devices pci2, pci3, and scsi1 for the PCI buses and SCSI adapter. This is true even when the new system appears to be identical with the original system. The disk devices, including those in the rootvg as well as external SAN disks, usually will be assigned the same names as on the original system. However, if a disk does not have unique

© Copyright IBM Corporation 2009, 2010. All rights reserved.

identifier support and does not have a PVID assigned, then it may be given a new name. Almost all other devices will be given new names.

6. There may be possible error log inconsistencies. Error log entries with a timestamp prior to moving the rootvg may now appear incorrect. This may be in part due to device name changes as previously described, as well as the fact that failing hardware is no longer present.

7. Inappropriate system-specific configuration information may be applied to the new system, for example, TCP/IP hostnames and IP addresses. The route attributes of the inet0 device will be applied to the new system, and if inappropriate, could manifest itself as a boot failure, hang, or loss of system access.

8. Network Interface device configuration will not be applied to the new system. Network interfaces are associated with specific network adapters. If the adapter names change, as previously described, then a new network interface will also be created with default configuration settings. This could manifest itself as a boot failure, hang, or loss of system access.

9. There may be disk reservation conflicts. When the rootvg is booted in another system, it is now a different host than it was previously. This may prevent it from accessing some of the disks it could access in the old LPAR. If the rootvg is affected by this, the new system will not boot. But this should be something that can be fixed using defined procedures for the disk subsystem.

10. iSCSI configuration problems - If iSCSI disks are being used, they may not be recognized in the new system. If accessed using an iSCSI TOE adapter, the new iSCSI TOE adapter instance may need to be configured using information from the prior adapter. If accessed using software iSCSI, then network configuration issues may need to be resolved first.

11. iSCSI-specific boot problems - If the rootvg resides on an iSCSI disk, the new system may fail to boot even after selecting the correct boot device using the firmware SMS menus as previously described. This is due to the same configuration problems as previously described. But since the system does not boot, the configuration problems cannot be resolved. The solution here is to boot into maintenance mode to correct the configuration problems.

12. There may be additional problems besides those enumerated above. These may or may not be resolvable.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

Conclusion

This article explains some simple approaches to achieve higher availability for AIX logical partitions by:

- Switching over SAN boot devices using virtualized devices.
- Using FlashCopy to back up and recover the AIX environment.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

Resources

Learn

- Using Virtual I/O server: Manage the Virtual I/O Server and client logical partitions using the Hardware Management Console (HMC) and the Virtual I/O Server command-line interface.
- IBM Power Systems PowerVM
- PowerVM Virtualization on IBM System p: Introduction and Configuration Fourth Edition provides an introduction to PowerVM virtualization technologies on IBM System p servers.
- IBM System p Advanced POWER Virtualization (PowerVM) Best Practices provides best practices for planning, installing, maintaining, and operating the functions available using the Advanced POWER Virtualization feature on IBM System p5 servers.
- discusses how Live Partition Mobility can help technical professionals, enterprise architects, and system administrators.
- System p and AIX information center is your source for technical information about AIX and System p.

Discuss

- Participate in the AIX and UNIX forums:
  - AIX Forum
  - AIX Forum for developers
  - Cluster Systems Management
  - IBM Support Assistant Forum
  - Performance Tools Forum
  - Virtualization Forum
  - More AIX and UNIX Forums

About the authors

Julie Craft

© Copyright IBM Corporation 2009, 2010. All rights reserved.

Julie Craft is an architect in AIX product development. Her specific areas of expertise include AIX installation, maintenance, and systems management.

Sanket Rathi

Sanket Rathi is a developer in the AIX storage device driver team. His specific areas of expertise include Fibre Channel, SCSI, MPIO, and Virtual I/O Server.

Anbazhagan Mani

Anbazhagan Mani is an Advisory Software Engineer in the AIX product development team at India. His specific areas of expertise include AIX systems management and solutions development.

Chris Schwendiman

Chris Schwendiman is an architect in AIX product development. His specific areas of expertise include the Object Data Manager (ODM) and AIX device configuuration.

Jim Pafumi

Jim Pafumi is a senior engineer. His area of experties is PowerVM cross functional interfaces.

© Copyright IBM Corporation 2009, 2010. All rights reserved.

Gero Schmidt

Gero Schmidt is an IT Specialist in the IBM ATS technical sales support organization in Germany. During his eight years of experience with IBM storage products, he participated in various beta test programs for ESS 800 and especially in the product rollout and beta test program for the DS6000/DS8000 series.

Nick Ham

Nick is responsible for development of storage technologies on AIX. His specific areas of expertise concentrate on storage device drivers and multi-pathing software.

© Copyright IBM Corporation 2009, 2010. All rights reserved.