



IBM @server p5
Introduction to the
Virtual I/O Server
Whitepaper

September 2, 2004

Jaya Srikrishnan
IBM Corporation
jaya@us.ibm.com

TABLE OF CONTENTS

Virtual I/O	3
Introduction.....	3
Abstract.....	3
What is virtual I/O?	4
Dedicated vs shared	4
vSCSI.....	5
Virtual LAN and SEA	6
Why do you need virtual I/O?.....	7
When would you not use virtual I/O?.....	10
How do you use Virtual I/O?.....	11
Configuring Virtual I/O Server(s)	11
Configuring vSCSI	12
Configuring SEA.....	12
Configuring and installing clients	12
Mirroring/Multi-pathing at the client	13
Mirroring/Multi-pathing at the server	13
Backup/restore of Virtual I/O Server	14
Updating the Virtual I/O Server	15
Backup/restore of data on virtual disks	15

Virtual I/O

Introduction

IBM introduced a new era in UNIX® and Linux® operating environment computing with the introduction of the IBM @server® p5 server family. These systems revolutionize information technology economics with lightning-quick POWER5™ processors and new IBM Virtualization Engine™ system technologies options that included:

- Processing resources in @server p5 systems can be fine tuned to have more “virtual servers” or dynamic logical partitions (LPARs) than processors. Partitions can be allocated in units as small as 1/10th of a processor and may be incremented in units of 1/100th of a processor.
- This fine tuning capability allows consolidation of multiple independent workloads, resulting in the equivalent of a “server farm in a box”.
- Businesses get the benefit of high systems utilization and easier administration of consolidated systems, helping lower their total cost of ownership (TCO).
- Partitions can be assigned to a shared processor pool, providing automatic, non-disruptive balancing of processing power and improved service levels as processing needs change.
- The virtual I/O option includes virtual SCSI for sharing Fibre Channel and SCSI adapters and the attached disk drives and virtual networking to enable the sharing of Ethernet adapters, providing greater flexibility to configure cost-effective systems.

This paper describes the practical uses of virtual I/O and provides hints and tips for exploiting this new capability. Virtual I/O is included in the Advanced POWER Virtualization feature for @server p5 systems. The Advanced POWER Virtualization feature requires AIX 5L™ V5.3, SUSE LINUX Enterprise Server 9 for POWER™ (SLES 9) or Red Hat Enterprise Linux AS 3 for POWER, Update 3 (RHEL AS 3).

Abstract

Virtual I/O is an optional feature that enables partitions to share I/O resources like Ethernet adapters and SCSI or Fibre Channel disks. The operating system running in the partition thinks it has access to a disk or a network, when in fact that access is mediated by a Virtual I/O Server.

Due to the numbers of partitions that can be defined, allocating adapters and disks for each partition’s exclusive use can require a very large number of I/O slots and devices. These may or may not be fully utilized. Sharing these resources among multiple partitions can reduce the total cost of computing by increasing the utilization of these resources.

What is virtual I/O?

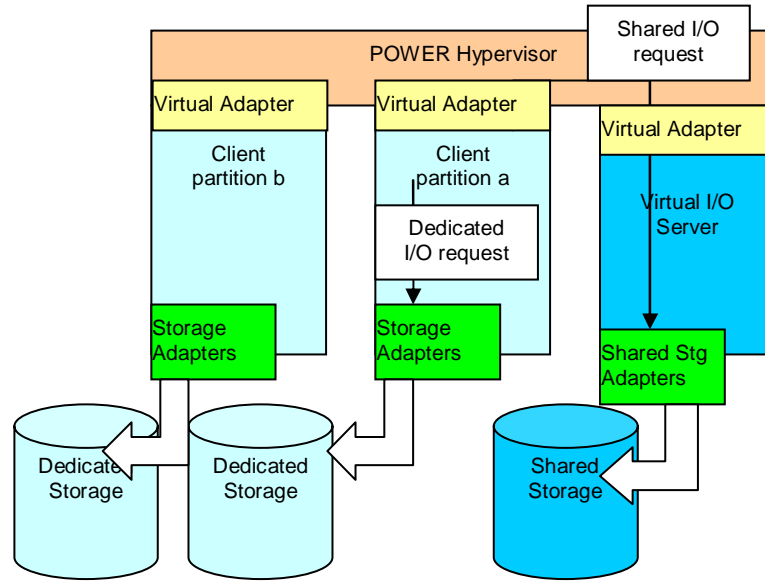
To an operating system (OS), virtual I/O is practically indistinguishable from physical I/O. The lowest levels of the operating system know that virtual I/O is different, but to the rest of the operating system and its applications, virtual I/O looks exactly like physical I/O. Logical volumes in AIX 5L are one example of virtual I/O devices. In the context of this paper, virtual devices and adapters are entities that appear to be exclusively owned by an operating system, when in fact they are shared among a number of operating systems. Virtual disks and virtual Ethernet adapters enable multiple operating systems, each running in its own LPAR, to share physical disks and Ethernet adapters.

Dedicated versus shared

An operating system (AIX 5L or Linux) running in an LPAR can have its own dedicated adapters, and these will operate just as they have done before. These dedicated adapters are allocated to the partition via the Hardware Management Console (HMC) and cannot be used by any other partition. Dynamic LPAR operations can be used to add, remove or move these adapters and their associated devices between LPARs, but only one partition can use them at a time.

Virtual I/O Server is special POWER5 partition that provides the ability to dedicate I/O adapters and devices to a virtual server, enabling the allocation and management of I/O devices across multiple partitions. The POWER Hypervisor™ does not own any physical I/O devices, and it does not provide virtual interfaces to them. Virtual I/O devices are owned by the Virtual I/O Server, which provides access to the real hardware that the virtual device is based on. Shared devices and adapters can be used simultaneously by more than one partition. The sharing is mediated by a Virtual I/O Server that owns the physical resources. To each partition, this shared resource appears as a virtual device or a virtual adapter.

The following picture shows both dedicated and shared I/O. A shared I/O request goes out via the Hypervisor to a Virtual I/O Server which multiplexes requests from different clients over its adapters. The example shown here is for storage but shared networking I/O works similarly.



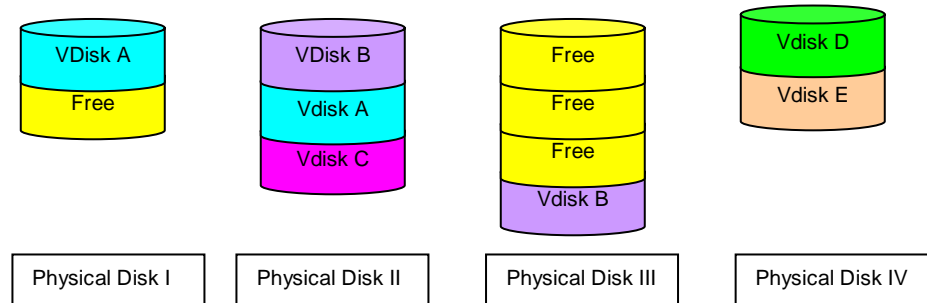
vSCSI

There are two types of virtual I/O provided on @server p5 systems. The first of these is called virtual SCSI or vSCSI. vSCSI adapters provide access to virtual disks as well as physical logical unit numbers (LUNs) or disk drives. To provide vSCSI support, an operating system is assigned a vSCSI client adapter. This client adapter is paired with a vSCSI server adapter that is owned by a Virtual I/O Server.

A virtual disk is a disk that does not map one-to-one with a physical disk drive. Like a logical volume, a virtual disk can be smaller or larger than a physical disk drive. The physical disk drive is owned by a Virtual I/O Server and can be a parallel SCSI or Fibre Channel-attached disk. It can be a LUN on an advanced function controller like the IBM TotalStorage® Enterprise Storage Server®. It can also be a virtualized disk on a SAN Virtualization Controller.

To the operating system, a virtual disk looks as just like a physical disk or hdisk. File systems and/or data backup and restore utilities can be used against these disks. The value of a virtual disk is that it can be sized to match the requirements of the operating system. For example, a boot disk does not have to be a complete physical disk drive. Since most boot images are relatively small, a virtual disk can be created for each partition's operating system, reducing the number of physical disk drives necessary for the operation of the partitions.

The following figure illustrates how virtual disks can be subsets of a physical disk drive (vdisks C,D, E) or can include sectors that are distributed over a number of physical disks (vdisks A and B)

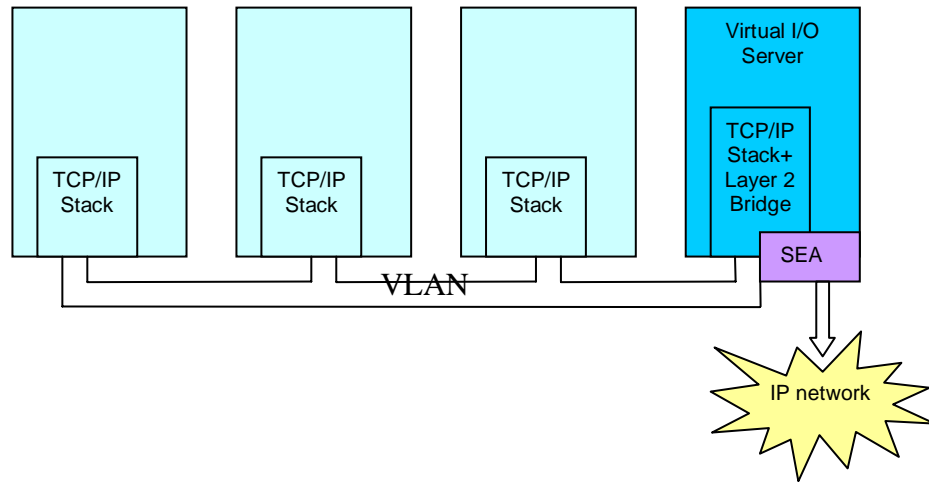


A physical disk is one that maps one-one with a disk drive. Again, the drive can be attached via parallel SCSI or Fibre Channel and can be on an advanced storage controller or SAN Virtualization Controller. Disks on these types of advanced function devices are already virtualized, so using them “as-is”, instead of virtualizing them further, can be an advantage. However, creating virtual disks on these LUNs can simplify backup, restore and management as there are fewer entities to deal with.

Virtual LAN and Shared Ethernet Adapters

@server p5 servers also provide virtual LANs and Shared Ethernet Adapters (SEA). A virtual LAN is an internal LAN that connects a set of partitions. These LANs exist only in the memory of the server and are implemented via the POWER Hypervisor™. They provide fast communication between partitions and look like Ethernet LANs to the operating system in the partition. If these LANs need to connect to external physical LANs, an Ethernet adapter is needed to connect to the virtual LAN.

To facilitate connecting a number of virtual LANs to external networks, Shared Ethernet Adapters have also been provided. A Shared Ethernet Adapter allows multiple virtual LANs to connect to the external network via the same physical Ethernet Adapter. The Shared Ethernet Adapter is implemented as a Layer-2 bridge that connects the internal virtual LAN to the external network. Link aggregation is supported on these Shared Ethernet Adapters.



Why do you need virtual I/O?

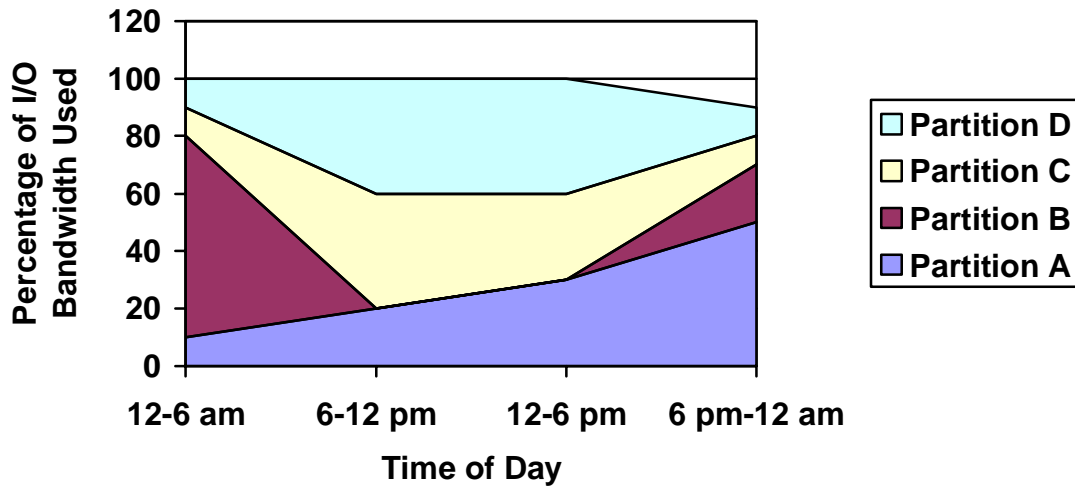
Sharing I/O devices and adapters allows multiple partitions to co-exist on a @server p5 system with fewer disk drives, adapters, cables and other infrastructure. As mentioned earlier, each partition does not have to have a physical slot with a dedicated adapter and disk drive for its boot image. Virtual disks can be sized to the needs of the operating system without wasting any space. Shared Ethernet Adapters allow multiple partitions to connect to external LANs with fewer physical Ethernet adapters. This reduces the need for slots and adapters as well as cables and switches in the external LAN.

By reducing the number of physical disk drives and networking infrastructure, virtual I/O simplifies the management of these entities. @server p5 systems are designed to support up to 254 logical partitions. If each partition could only have dedicated I/O adapters and devices for all its I/O needs, the size and complexity of the I/O needed for a single @server system would grow exponentially with the number of partitions. Virtual I/O provides a way to make do with fewer physical entities, making this problem disappear. Also, virtual I/O is created and configured through fewer interfaces, simplifying management.

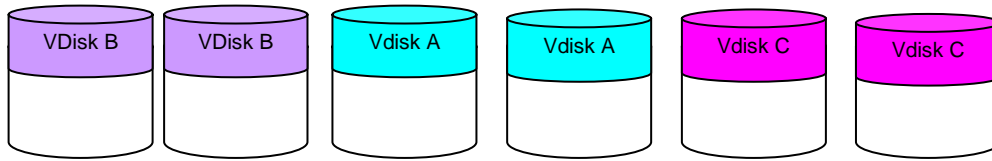
Operating system workloads fluctuate over time, with peaks and valleys. This leads to variations in the utilization of I/O. Sometimes a partition needs a lot of bandwidth and at other times, it needs far less. Typically, in the absence of virtual I/O, each partition is configured with enough I/O capacity for the higher bandwidth requirement, leaving a lot of time when the total bandwidth is under-utilized. Sharing resources where the peaks and valleys are complementary can result in a significant reduction in the total amount of resource needed to handle the aggregated workload for all the partitions that have these complementary peaks and valleys.

In the following chart, each partition's workload varies over the course of the day, using different percentages of the total I/O bandwidth available. However, at no time does the

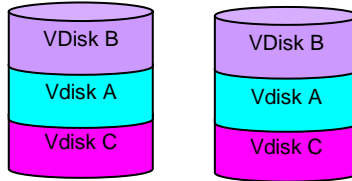
total bandwidth needed exceed 100% o so they can co-exist successfully and increase the utilization on the I/O adapters



Dynamic LPAR can also be used to handle such varying bandwidth requirements, but dynamic LPAR does not take effect immediately as the adapter has to be varied off from one partition and varied on to the other. Also, it implies that the first partition does not need access to the adapter at all during the period the second partition needs it. Sharing adapters allows for instantaneous and simultaneous access by multiple partitions. Finally, many adapters and I/O devices need to be duplicated for availability. So a partition needs two Ethernet adapters even if its utilization of one adapter is not very high. This also results in a lot of unused bandwidth. Similarly, mirroring of disks can lead to under-utilization of disk capacity on both disks. Virtual disks and Shared Ethernet Adapters allow for less redundancy overall, as a number of adapters can share a failover adapter. Also, disk capacity can be better utilized, even if mirroring is needed.



Mirrored physical disks

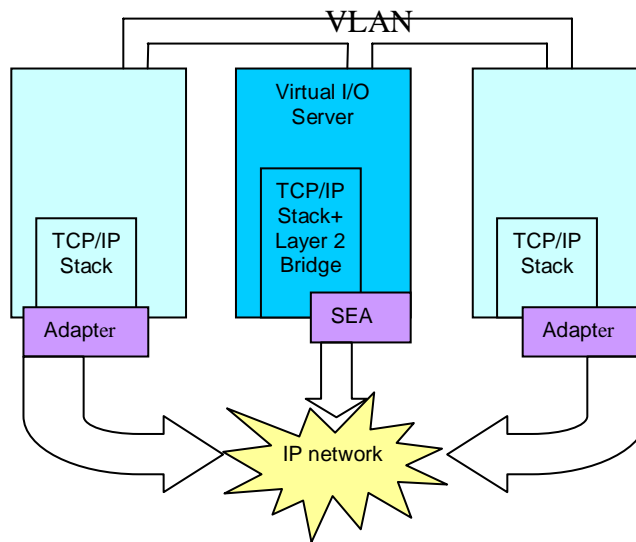


Mirrored virtual disks

Where is Virtual I/O useful?

The total bandwidth needed at any point in time cannot exceed the bandwidth of the shared adapter. Therefore, using it in cases where a partition needs the full capacity of a disk or adapter all the time may not be optimal. Consider using virtual I/O for adapters that are not fully utilized all the time, or where partitions need only a fraction of the capacity of a disk drive. Additionally, virtual I/O trades off the additional processing power used by a Virtual I/O Server for the reduction in physical I/O adapters and devices.

Virtual I/O can also be used to share failover adapters. A partition can have a dedicated adapter for exclusive use and share a failover adapter with a number of other partitions that have similar dedicated adapters. This reduces both the slot and the adapter requirements for availability.



In the diagram above, the Shared Ethernet Adapter is used as a failover adapter for either of the other adapters and instead of needing four adapters total, one can make do with three. As the number of partitions grows, this type of shared failover adapter can reduce the total number of adapters required.

Sharing I/O between partitions that have complementary patterns of usage is ideal. When one partition needs an adapter for certain times of the day or week and another partition needs the same type of adapter at other times of the day or week, virtual I/O can be used to reduce the number of adapters that are needed.

The capacity of disk drives is steadily increasing over time, leading to a lot of unused capacity when operating systems only need a small amount of disk space. Virtual disks are ideal candidates for usage in these scenarios. A single physical drive can be carved into a number of virtual disks that can be allocated to different partitions. Boot images are only one example of such usage. This is especially useful in the case of SCSI disks. It is expensive to allocate a single SCSI disk to a partition when it needs only a small percentage of the disk. Add mirroring requirements to such usage and the amount of unused capacity doubles. Virtual disks can increase the utilization on SCSI drives considerably. Fibre Channel disks are typically attached via advanced function controllers that provide virtualization of physical disk drives so virtualizing these disks is less optimal.

When would you not use virtual I/O?

The mediation of the sharing by the Virtual I/O Server adds latency to the I/O path. When a partition needs high-performance, low latency access to its I/O, virtual I/O is not a good fit. Similarly it is not a good fit for those scenarios where a partition is using the capacity of the adapter and/or the device to its maximum the majority of the time. Sharing does not provide any benefit in these cases and should not be used.

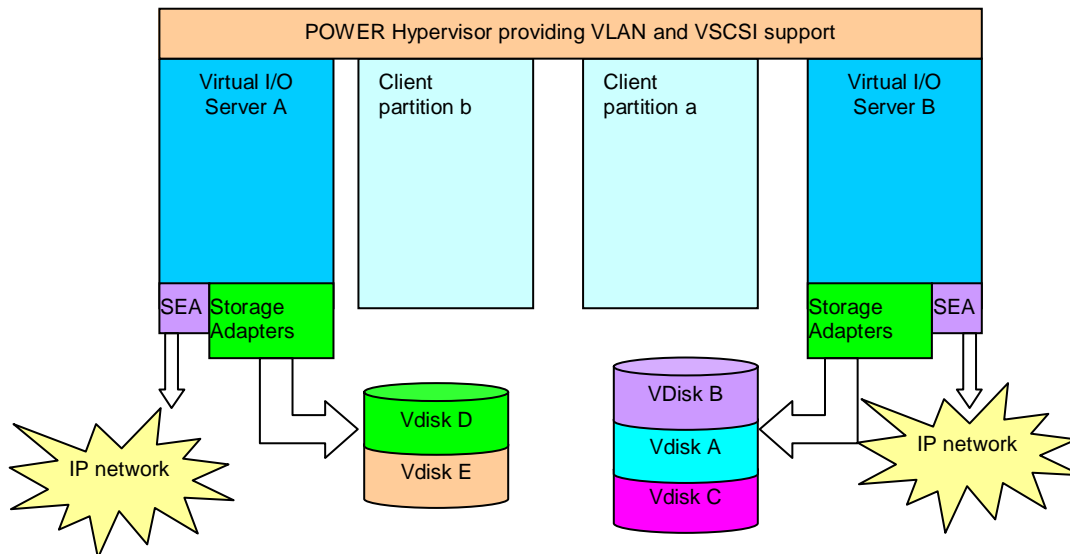
How do you use Virtual I/O?

Configuring Virtual I/O Server(s)

In order to use virtual I/O, at least one Virtual I/O Server needs to be created. This is a partition that is defined using the HMC, very much like any other partition. However, it is a special type of partition that runs a kernel designed to provide I/O sharing and virtualization. Applications cannot run in a Virtual I/O Server. It is solely intended for the purpose of providing virtual I/O.

Multiple Virtual I/O Servers can be defined for availability as well as isolation. If sets of operating systems need to be kept separate from each other, each set should have its own Virtual I/O Server. Virtual I/O Servers can be configured as pairs of redundant partitions or as individual partitions serving different sets of client operating systems.

In the following diagram, there are two Virtual I/O Servers, each with a set of storage and Ethernet adapters. The Ethernet adapters are configured as Shared Ethernet Adapters. Client partitions a and b can have their own dedicated I/O adapters and devices which are not shown here. But each client partition can also access the SEAs and the virtual disks owned by the Virtual I/O Servers. Virtual I/O Server A exports two virtual disks - one each for clients a (Vdisk D) and b (Vdisk E). Virtual I/O Server B exports 3 virtual disks: Vdisk B and Vdisk A for client a and Vdisk C for client b. Client partitions cannot share virtual disks at this time. The SEAs can connect to the same or different IP networks. The POWER Hypervisor provides the communications between the partitions.



Virtual I/O Servers are the only partitions that can have vSCSI server adapters and Shared Ethernet Adapters assigned to them. When they are defined, each Virtual I/O Server should be assigned one vSCSI server adapter for each client vSCSI adapter that will be connected to it. It should also be assigned storage (parallel SCSI or Fibre Channel) adapters that provide access to the disks that it will use to virtualize storage. The physical Ethernet adapters that will be shared via the SEA facility should also be assigned to the Virtual I/O Server. At a minimum, a Virtual I/O Server needs one storage

and one Ethernet adapter. The Ethernet adapter is also used for communication with the HMC.

Once a Virtual I/O Server is defined, its kernel needs to be installed. The Virtual I/O Server image is shipped on a special CD-ROM and can be installed from the HMC or from a CD-ROM drive. The same medium is used for as many Virtual I/O Servers as are needed on the @server p5 system.

After the installation is completed and the Virtual I/O Server boots up, the rest of the configuration is done by logging on to the Virtual I/O Server and using a special shell and commands that are provided with it.

Configuring vSCSI

The physical disks that have been allocated to a Virtual I/O Server need to be divided up into virtual disks and allocated to specific vSCSI adapters. Alternatively, entire physical disk drives can be allocated to a specific vSCSI adapter. Virtual disks are created from volume groups, or sets of virtual disks. Commands are provided to define volume groups, assign disks to those volume groups, create virtual disks in these volume groups, and assign them to vSCSI server adapters.

Configuring SEA

Internal virtual LANs are defined via the HMC. As each partition is created, its participation in one or more virtual LANs is defined. A Virtual I/O Server needs to be a member of each virtual LAN that will be sharing Ethernet adapters owned by it. This sets up the communication between the client partitions and the Virtual I/O Server.

Once this internal LAN set-up is complete, the Virtual I/O Server's command line interface can be used to configure one or more physical Ethernet adapters as SEA for these internal LANs. An Ethernet adapter can be shared among multiple internal LANs and can also be used for communication with the Virtual I/O server. With the appropriate configuration, this adapter can also be used for communications with the HMC. SEA can also use an aggregated link as the physical adapter to be shared.

Configuring and installing clients

Client partitions are installed and configured independent of a Virtual I/O Server (i.e. they can be defined before or after a Virtual I/O Server is configured) . They are defined using the HMC and each is assigned one or more vSCSI client adapters as well as being assigned to one or more internal LANs. vSCSI client and server adapters are defined in pairs. Dynamic LPAR operations can be used to create and break these paired connections. Install client partition operating systems as usual via network installs or CD-ROM installs. Once they are up and operational, the vSCSI client adapters will initiate communication with the vSCSI server adapters and the connection will be established.

There is no defined order in which clients and Virtual I/O Servers have to be booted up. The boot process can be started at the same time, or at different times. When a client comes up, it will attempt to connect to the Virtual I/O Server. If the Virtual I/O Server is

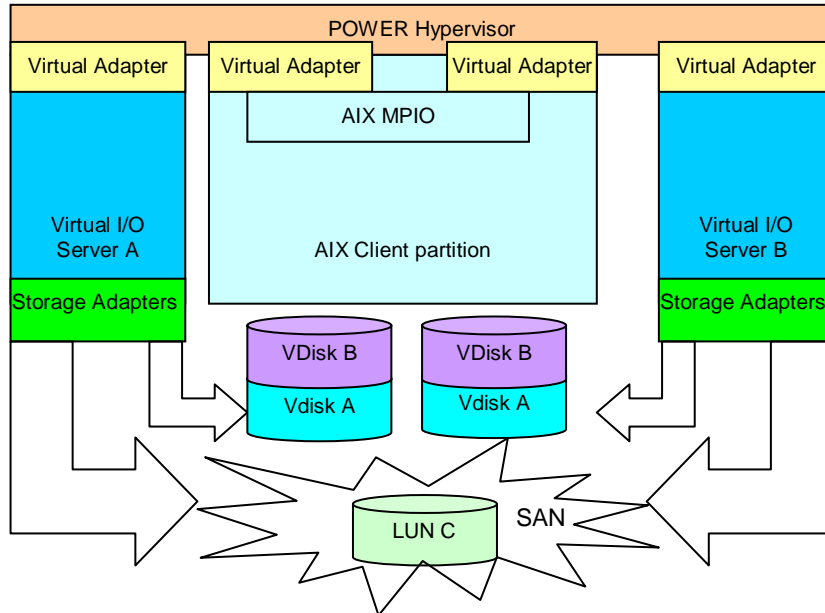
not up within a reasonable amount of time, the connection attempt will fail and the vSCSI adapter will report an error. Obviously, if the boot device for the operating system is on a vSCSI adapter, the client will not come up if the Virtual I/O Server does not come up within this period.

If such an error occurs, the client can be re-booted after the Virtual I/O Server is up. If it is a non-boot device, then doing discovery and I/O configuration at the client (e.g. using `cfgmgr` on an AIX 5L client) after the Virtual I/O Server is up will cause the vSCSI client adapter to connect.

Mirroring/Multi-pathing at the client

The client operating system can mirror virtual disks by itself. These virtual disks can be provided by different Virtual I/O Servers. This kind of mirroring can provide high availability in the case of a Virtual I/O Server failure or reboot. Software mirroring is available on both AIX 5L and Linux OS.

Similarly, in the case of physical disks that are accessed via a storage adapter on the Virtual I/O Server, multi-pathing can be used from the client to provide redundant paths to the disk. This facility is currently only available on AIX 5L clients, not on Linux clients. However, Linux clients can make use of multi-pathing in the Virtual I/O Server to protect against adapter failure. Using multi-pathing from the client provides the most highly available and automated availability solution.

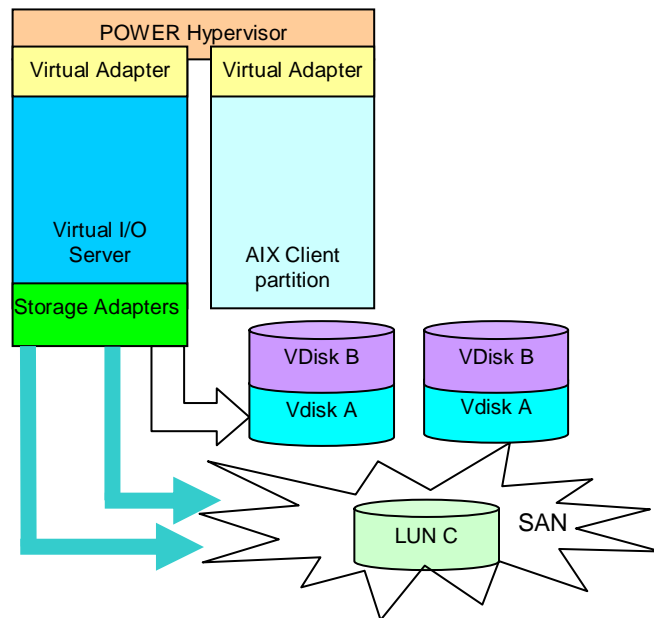


Mirroring/Multi-pathing at the server

Virtual disks can be mirrored via a Virtual I/O Server also, providing protection against disk or adapter failure. This mirroring is completely transparent to the client operating system, which is not aware of this protection.

In the case of physical disks, multi-pathing can also be configured in a Virtual I/O Server. Multi-pathing drivers such as IBM's SDD (Subsystem Device Driver) or EMC's PowerPath can also be installed in a Virtual I/O Server to provide additional function to these advanced function controllers.

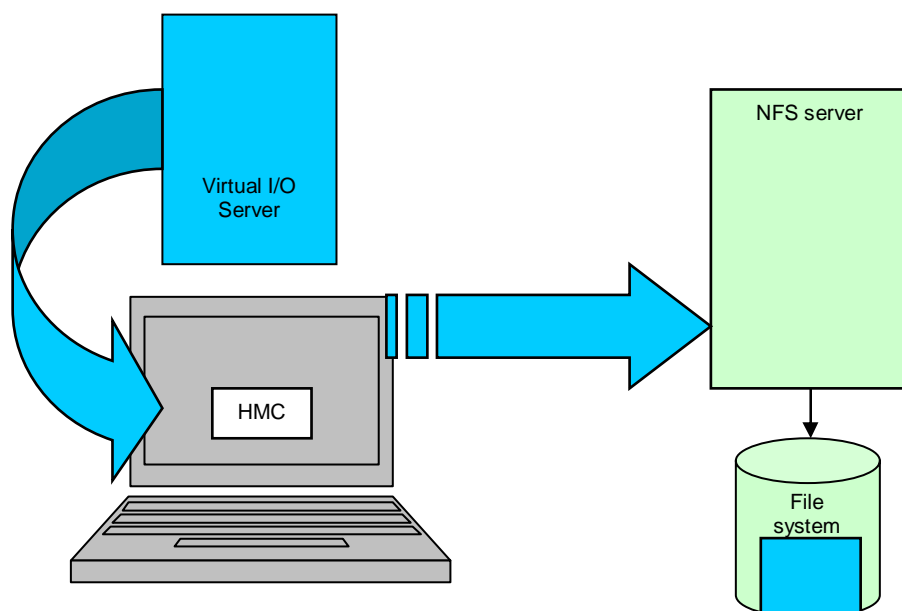
When using mirroring or multi-pathing at the server, the two paths should have as few entities (adapters, switches, etc.) in common as possible to increase the redundancy and reduce single-points-of-failure.



Backup/restore of Virtual I/O Server

Backup copies of a Virtual I/O Server must be made at regular intervals to protect against failure of the boot image in the Virtual I/O Server. It is recommended that this drive be mirrored or defined on an advanced function controller that provides mirroring or RAID. However, as a further protection, the complete partition (kernel plus I/O configuration) can be backed up onto CD-ROM via the HMC or on to a NFS server that is accessible from the HMC. It is recommended that this backup is done before and after any configuration changes and prior to installing updates to a Virtual I/O Server. It is also recommended that once an update is successful, a backup is made.

These backup copies can be restored via the HMC in case of failure.

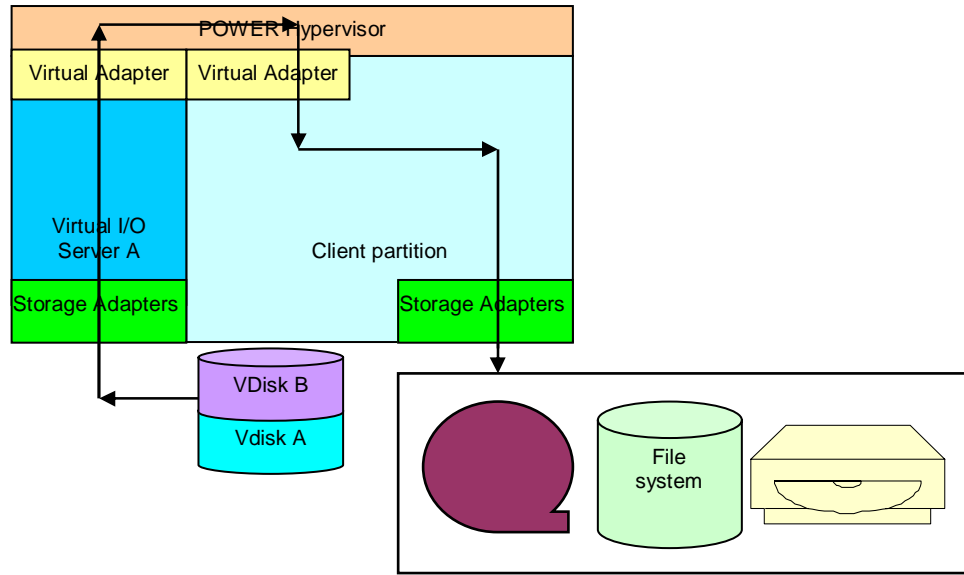


Updating the Virtual I/O Server

Periodically, IBM will provide updates to the Virtual I/O Server image. Fixes will only be provided on the latest level of the Virtual I/O Server. Some of these updates will require the Virtual I/O Server to be re-booted. If access to virtual I/O is needed during such times, it is recommended that the I/O be accessible from another Virtual I/O Server (using either mirroring or multi-pathing from the client) via a secondary set of vSCSI client adapters.

Backup/restore of data on virtual disks

Data on virtual disks can be backed up and restored as usual from the client operating system. File system or database utilities can be used to accomplish these tasks because the virtual disks look and behave exactly like physical disk drives.





© Copyright IBM Corporation 2004

IBM Corporation
Marketing Communications
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America
September 2004
All Rights Reserved

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries. The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

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

Information concerning non-IBM products was obtained from the suppliers of these products. Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM.

IBM, the IBM logo, the e-business logo, @server, AIX, AIX 5L, Enterprise Storage Server, Hypervisor, IBM Virtualization Engine, POWER, POWER Hypervisor, POWER5 and TotalStorage are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both. See <http://www.ibm.com/legal/copytrade.shtml>.

Linux is a registered trademark of Linus Torvalds in the United States, other countries or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

The IBM home page on the Internet can be found at <http://www.ibm.com>.

The IBM UNIX systems home page on the Internet can be found at <http://www.ibm.com/servers/eserver/pseries>.