Planning and Sizing for Virtualization on System P

March 2008

http://www.circle4.com/papers/cmg-virt-concepts.pdf

Jaqui Lynch – jaqui.lynch@mainline.com
Mainline Information Systems

Agenda

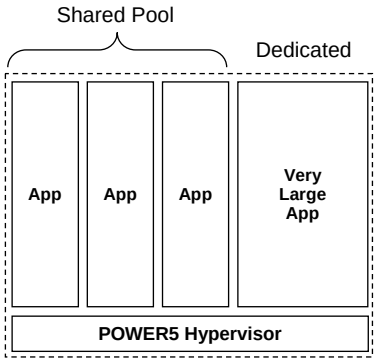
- Virtualization Options
- Pros and Cons
- Planning
- Virtual CPU
- Virtual I/O
  - Virtual Ethernet
  - Virtual SCSI
- Sizing thoughts

Virtualization Options

- Real
  - Dedicated processors/cores
  - Dedicated fibre or SCSI
  - Dedicated Ethernet
- Virtual
  - Shared processors/cores
  - Virtual ethernet
  - Shared ethernet adapter
    - Built on virtual ethernet
  - Shared SCSI
    - Can be SCSI or fibre
  - Ethernet and SCSI used a custom LPAR called a VIO server
    - Must include processor and memory resources in planning for that LPAR or LPARs

Step 1 – Investigate Virtual (Shared) CPUs

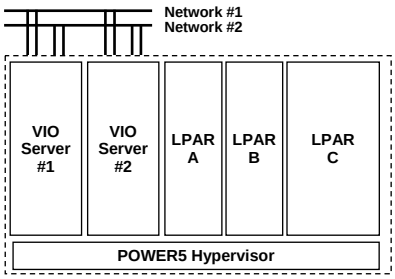
- Potential Benefits
  - Increase CPU utilization
  - Actual deployment effort is modest
- Issues/Considerations
  - High utilization LPARs will be poor donors but might benefit from use of the uncapped pool
  - Most mainframes run in exclusively this mode
  - Understand entitlement, VPs, weight, capped/uncapped, weight, reserve capacity on demand, processor folding.
  - Software licensing - use of uncapped LPARs with unnecessary VPs may impact costs
  - Review performance management tools
  - Not every application likes sharing – depends on workload characteristics



Source: IBM

Step 2 – Investigate Virtual Ethernet

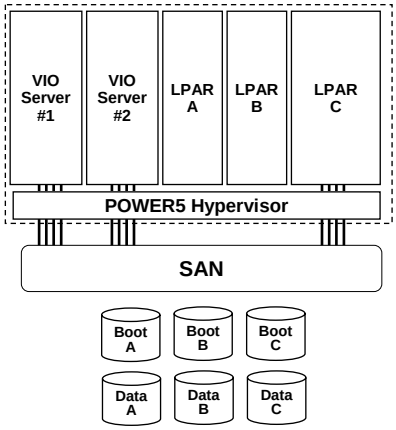
- Potential Benefits
  - Reduce the number Ethernet adapters, ports
  - Reduce cabling efforts and cables in frames
  - Reduce number of I/O drawers and/or frames
- Issues/Considerations
  - Understand Ethernet adapter/port utilization
  - Understand high availability cluster support requirements
  - Understand implications on backup architecture
  - Understand virtual I/O sizing and large send capabilities
  - Understand use of link aggregation and/or VLANS
  - Understand VIO high availability Ethernet options
  - Simplicity!!



Source: IBM

Step 3 – Investigate Virtual SCSI

- Potential Benefits
  - Reduce the number FC adapters and ports
  - Reduce cabling efforts and cables in frames
  - Reduce number of I/O drawers and frames.
- Issues/Considerations
  - Understand current SAN adapter / port utilization
  - Investigate high availability cluster support for virtual I/O
  - Understand implications on backup architecture
  - Understand virtual I/O server sizing
  - Understand availability choices such as dual VIOS, number of HBAs, O/S mirroring, etc

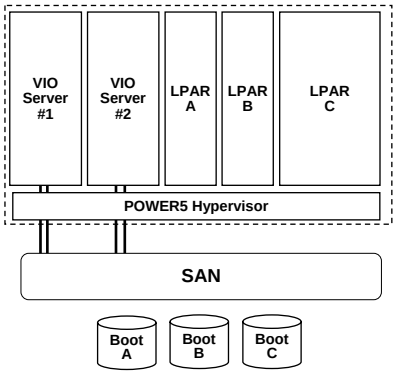


Note: Some LPARs could virtualize storage while others have direct HBA access.

Source: IBM

Step 4 – Investigate Boot from SAN

- Potential Benefits
  - Reduce the number of I/O drawers
  - Reduce number of frames
- Issues/Considerations
  - Use internal disk for VIO servers
  - Need robust, available SAN
  - Understand and size VIOS LPARs
  - Understand availability choices such as dual VIOS, multi-path I/O, O/S mirroring, etc.



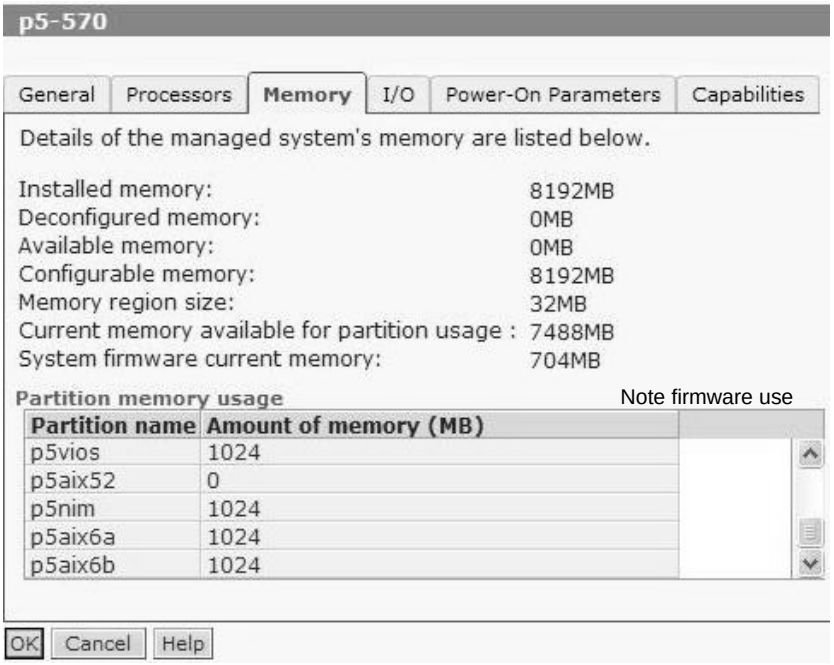
Note: LPARs could boot through the VIOS and have dedicated HBAs for data access.

Source: IBM

Planning

Memory Usage

From HMC



Note firmware use

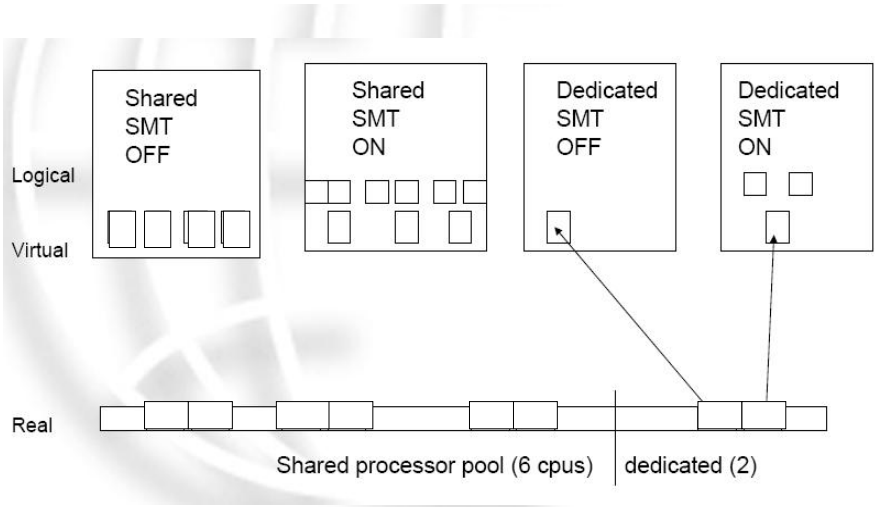
Planning for Memory

PLANNING SHEET Memory Overhead Calculation	Mem	Max Mem	Mem Ohead	Divide by 256	Round Up	New Overhead
lp1	98304	102400	1600	6.25	7	1792
lp2	16384	20480	320	1.25	2	512
lp3	16384	20480	320	1.25	2	512
lp4	24576	28672	448	1.75	2	512
NIM	4096	8192	128	0.5	1	256
VIO Server 1	4096	8192	128	0.5	1	256
VIO Server 2	4096	8192	128	0.5	1	256
Hypervisor						768
TCEs for drawers, etc?						512
IVEs (102mb per active port)					0	
Memory needed	167936	Or 164GB				**5376**
TOTAL Overhead						
TOTAL NEEDED	173312	170GB				

This gives a rough estimate
Assumes LMB size is 256 – each active IVE port adds 102MB

Don't forget memory overhead

Logical, Virtual or Real?



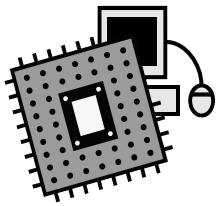
In shared world there is no one to one relationship between real and virtual processors
The dispatch unit becomes the VP

MicroPartioning
Shared processor partitions

- Micro-Partitioning allows for multiple partitions to share one physical processor
- Up to 10 partitions per physical processor
- Up to 254 partitions active at the same time
- One shared processor pool – more on the p6-570
- Dedicated processors are in the pool by default if their LPAR is powered off
- Partition's resource definition
  - Minimum, desired, and maximum values for each resource
  - Processor capacity (processor units)
  - Virtual processors
  - Capped or uncapped
    - Capacity weight
    - Uncapped can exceed entitled capacity up to number of virtual processors (VPs) or the size of the pool whichever is smaller
  - Dedicated memory
    - Minimum of 128 MB and 16 MB increments
  - Physical or virtual I/O resources
  - Some workloads hate the SPP – SAS is one

Defining Processors

- Minimum, desired, maximum
- Maximum is used for DLPAR
  - Max can be used for licensing
- Shared or dedicated
- For shared:
  - Capped
  - Uncapped
    - Variable capacity weight (0-255 – 128 is default)
    - Weight of 0 is capped
    - Weight is share based
    - Can exceed entitled capacity (desired PUs)
    - Cannot exceed desired VPs without a DR operation
  - Minimum, desired and maximum Virtual Processors
    - Max VPs can be used for licensing



Virtual Processors

- Partitions are assigned PUs (processor units)
- VPs are the whole number of concurrent operations
  - Do I want my .5 as one big processor or 5 x .1 (can run 5 threads then)?
- VPs round up from the PU by default
  - .5 Pus will be 1 VP
  - 2.25 Pus will be 3 VPs
  - You can define more and may want to
  - Basically, how many physical processors do you want to spread your allocation across?
- VPs put a cap on the partition if not used correctly
  - i.e. define .5 PU and 1 VP you can never have more than one PU even if you are uncapped
- Cannot exceed 10x entitlement
- VPs are dispatched to real processors
- Dispatch latency – minimum is 1 millisec and max is 18 millisecs
- VP Folding
- Maximum is used by DLPAR
- Use commonsense when setting max VPs!!!
- In a single LPAR VPs should never exceed Real Processors

How many VPs

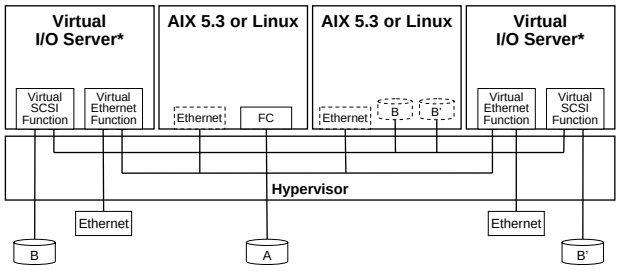
- Workload characterization
  - What is your workload like?
  - Is it lots of little multi-threaded tasks or a couple of large long running tasks?
  - 4 cores with 8 VPs
    - Each dispatch window is .5 of a processor unit
  - 4 cores with 4 VPs
    - Each dispatch window is 1 processor unit
  - Which one matches your workload the best?

Examples

- LPAR 1 - uncapped
  - Ent = 2.0
  - Max = 6.0
  - VPs = 4.0
  - Can grow to 4 processor units
  - VPs cap this
- LPAR 2 - uncapped
  - Ent = 2.0
  - Max = 6.0
  - VPs = 6.0
  - Can grow to 6 processor units
- LPAR 3 - Capped
  - Ent = 2.0
  - Max = 6.0
  - VPs = 4.0
  - Can't grow at all beyond 2 processor units

Virtual I/O Overview

Virtual I/O



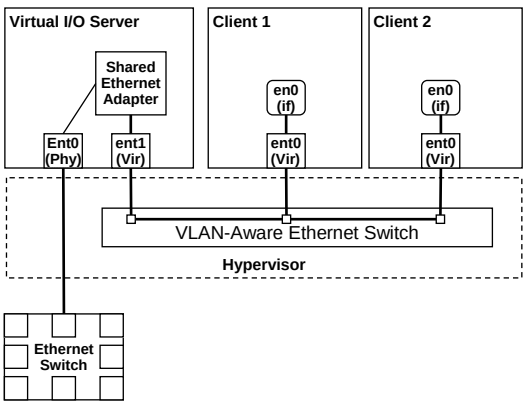
- Virtual I/O Architecture
  - Mix of virtualized and/or physical devices
  - Multiple VIO Servers* supported
- Virtual SCSI
  - Virtual SCSI, Fibre Channel, and DVD
  - Logical and physical volume virtual disks
  - Multi-path and redundancy options
- Benefits
  - Reduces adapters, I/O drawers, and ports
  - Improves speed to deployment
- Virtual Ethernet
  - VLAN and link aggregation support
  - LPAR to LPAR virtual LANs
  - High availability options

Source: IBM

Virtual Ethernet Concepts and Rules of Thumb

IBM POWER5 Virtual Ethernet

- Two basic components
  - VLAN-aware Ethernet switch in the Hypervisor
    - Comes standard with a POWER5 server.
  - Shared Ethernet Adapter
    - Part of the VIO Server
    - Acts as a bridge allowing access to and from an external networks.
    - Available via the Advanced POWER virtualization feature.



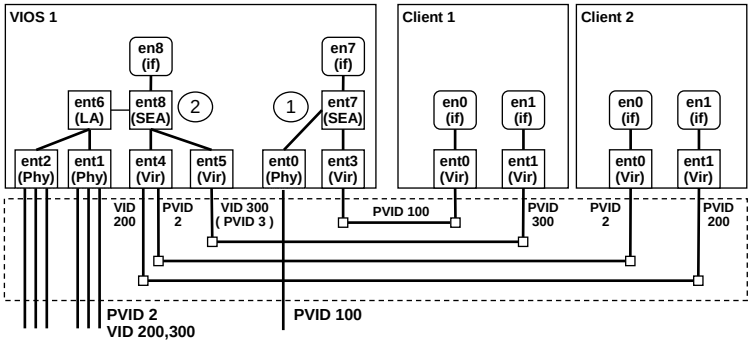
Source: IBM

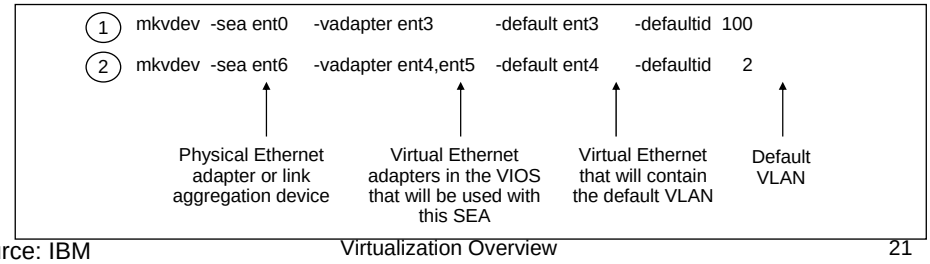
Shared Ethernet Adapter

In most cases, it is unnecessary to create more than one Virtual Ethernet adapter for a SEA.
(Think simple!)

Multiple VLANs can be added to a single SEA

LPAR only sees packets on its VLAN.





Source: IBM

Virtual Ethernet

- General Best Practices
  - Keep things simple
  - Use PVIDs and separate virtual adapters for clients rather than stacking interfaces and using VIDs.
  - Use hot-pluggable network adapters for the VIOS instead of the built-in integrated network adapters. They are easier to service.
  - Use dual VIO Servers to allow concurrent online software updates to the VIOS.
  - Configure an IP address on the SEA itself. This ensures that network connectivity to the VIOS is independent of the internal virtual network configuration. It also allows the ping feature of the SEA failover.
  - For the most demanding network traffic use dedicated network adapters.

Source: IBM

Virtual Ethernet

- Link Aggregation
  - All network adapters that form the link aggregation (not including a backup adapter) must be connected to the same network switch.
- Virtual I/O Server
  - Performance scales with entitlement, not the number of virtual processors
  - Keep the attribute tcp_pmtu_discover set to “active discovery”
  - Use SMT unless your application requires it to be turned off.
  - If the VIOS server partition will be dedicated to running virtual Ethernet only, it should be configured with threading disabled (Note: this does not refer to SMT).
  - Define all VIOS physical adapters (other than those required for booting) as desired rather than required so they can be removed or moved.
  - Define all VIOS virtual adapters as desired not required.

Source: IBM

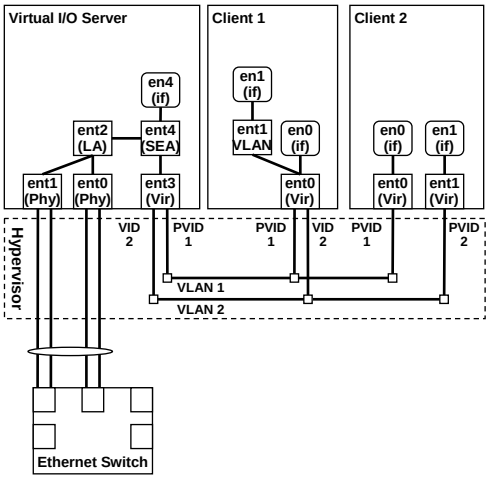
Virtual Ethernet Performance

- Performance - Rules of Thumb
  - Choose the largest MTU size that makes sense for the traffic on the virtual network.
  - In round numbers, the CPU utilization for large packet workloads on jumbo frames is about half the CPU required for MTU 1500.
  - Simplex, full and half-duplex jobs have different performance characteristics
    - Full duplex will perform better, if the media supports it
    - Full duplex will NOT be 2 times simplex, though, because of the ACK packets that are sent; about 1.5x simplex (Gigabit)
    - Some workloads require simplex or half-duplex
  - Consider the use of TCP Large Send
    - Large send allows a client partition send 64kB of packet data through a Virtual Ethernet connection irrespective of the actual MTU size
    - This results in less trips through the network stacks on both the sending and receiving side and a reduction in CPU usage in both the client and server partitions

Source: IBM

Limits

- Maximum 256 virtual Ethernet adapters per LPAR
- Each virtual adapter can have 21 VLANs (20 VIDs, 1 PVID)
- Maximum of 16 virtual adapters can be associated with a single SEA sharing a single phisical network adapter.
- No limit to the number of LPARs that can attach to a single VLAN.
- Works on OSI-Layer 2 and supports up to 4094 VLAN IDs.
- The POWER Hypervisor can support virtual Ethernet frames of up to 65408 bytes in size.
- The maximum supported number of physical adapters in a link aggregation or EtherChannel is 8 primary and 1 backup.



Source: IBM

IVE Notes (Power6 only)

- Which adapters do you want? Each CEC requires one.
  - Dual 10/100/1000 TX (copper)
  - Quad 10/100/1000 TX (copper)
  - Dual 10/100/1000 SX (fiber)
- Adapter ties directly into GX Bus
  - No Hot Swap
  - No Swap Out for Different Port Types (10GbE, etc.)
- Not Supported for Partition Mobility, except when assigned to VIOS
- Partition performance is at least the same as a real adapter
  - No VIOS Overhead
  - Intra-partition performance may be better than using Virtual Ethernet
- Usage of serial ports on IVE
  - Same restrictions as use of serial ports that were on planar on p5
  - Once an HMC is attached these become unusable
- Naming
  - Integrated Virtual Ethernet – Name used by marketing
  - Host Ethernet Adapter (HEA) Name used on user interfaces and documentation

Virtual SCSI

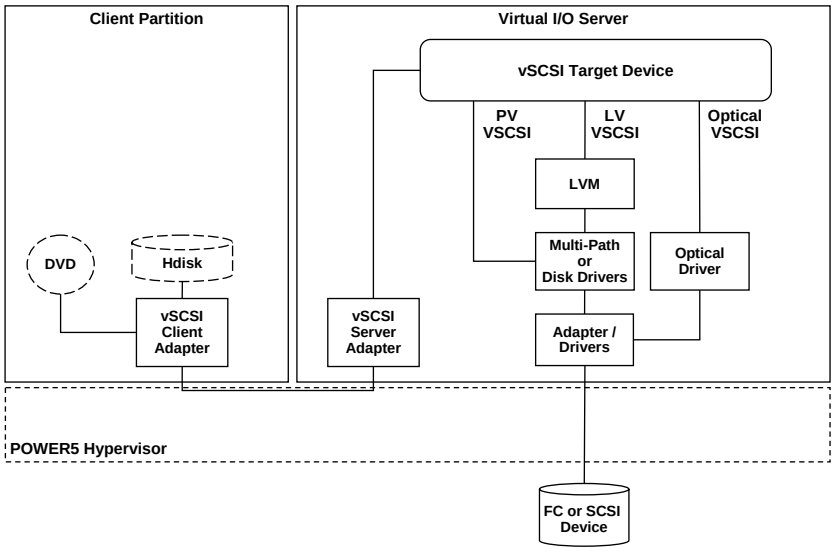
Virtual SCSI General Notes

- Notes
  - Make sure you size the VIOS to handle the capacity for normal production and peak times such as backup.
  - Consider separating VIO servers that contain disk and network as the tuning issues are different
  - LVM mirroring is supported for the VIOS's own boot disk
  - A RAID card can be used by either (or both) the VIOS and VIOC disk
  - Logical volumes within the VIOS that are exported as virtual SCSI devices may not be striped, mirrored, span multiple physical drives, or have bad block relocation enabled
  - SCSI reserves have to be turned off whenever we share disks across 2 VIOS. This is done by running the following command on each VIOS:

    # chdev -l <hdisk#> -a reserve_policy=no_reserve

Source: IBM

Virtual SCSI Basic Architecture

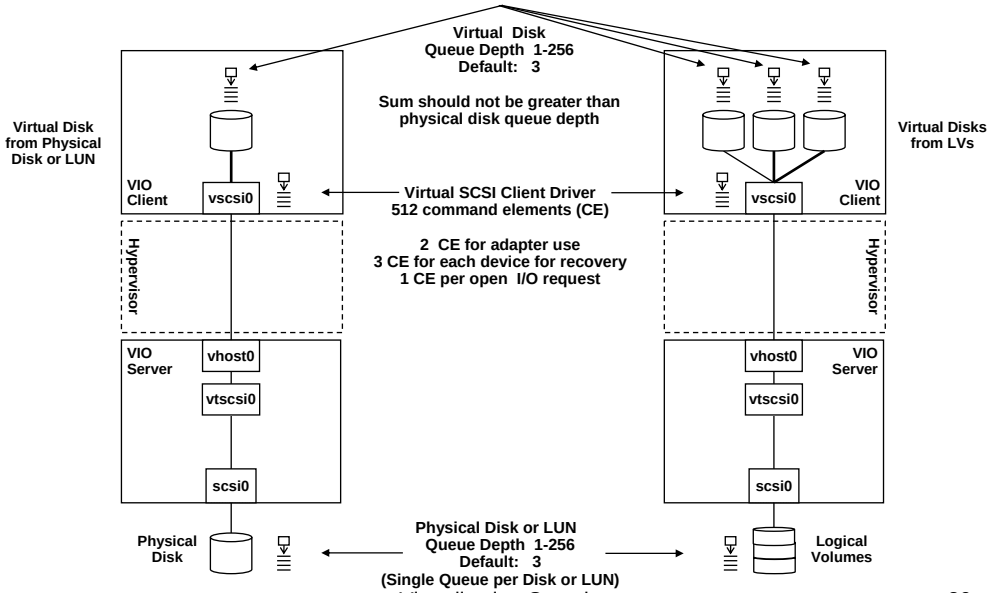


Source: IBM

Virtualization Overview
http://www.mainline.com/

29

SCSI Queue Depth



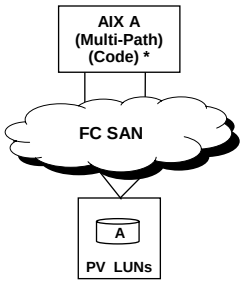
Source: IBM

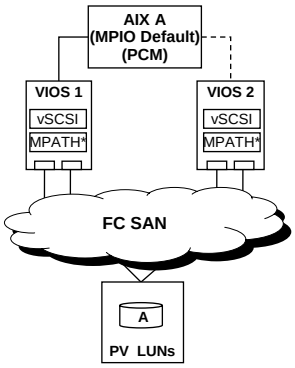
Virtualization Overview
http://www.mainline.com/

30

Boot From SAN

Boot From SAN



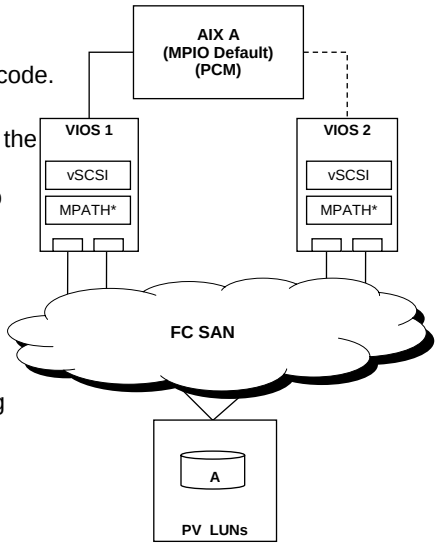


- Boot Directly from SAN
  - Storage is zoned directly to the client
  - HBAs used for boot and/or data access
  - Multi-path code of choice runs in client
- SAN Sourced Boot Disks
  - Affected LUNs are zoned to VIOS(s) and assigned to clients via VIOS definitions
  - HBAs in VIOS are independent of any HBAs in client
  - Multi-path code in the client will be the MPIO default PCM for disks seen through the VIOS.

Source: IBM

Boot from SAN via VIO Server

- Client
  - Uses the MPIO default PCM multi-path code.
  - Active to one VIOS at a time.
  - The client is unaware of the type of disk the VIOS is presenting (SAN or local)
  - The client will see a single LUN with two paths regardless of the number of paths available via the VIOS
- VIOS
  - Multi-path code is installed in the VIOS.
  - A single VIOS can be brought off-line to update VIOS or multi-path code allowing uninterrupted access to storage.



Source: IBM

Boot from SAN vs. Boot from Internal Disk

- Advantages
  - Boot from SAN can provide a significant performance boost due to cache on disk subsystems.
    - Typical SCSI access: 5-20 ms
    - Typical SAN write: 2 ms
    - Typical SAN read: 5-10 ms
    - Typical Single disk : 150 IOPS
  - Can mirror (O/S), use RAID (SAN), and/or provide redundant adapters
  - Easily able to redeploy disk capacity
  - Able to use copy services (e.g. FlashCopy)
  - Fewer I/O drawers for internal boot are required
  - Generally easier to find space for a new image on the SAN
  - Booting through the VIOS could allow pre-cabling and faster deployment of AIX
- Disadvantages
  - Will loose access (and crash) if SAN access is lost.
  - If dump device is on the SAN the loss of the SAN will prevent a dump.
  - It may be difficult to change (or upgrade) multi-path codes as they are in use by AIX for its own need.
    - You may need to move the disks off of SAN, unconfigure and remove the multi-path software, add the new version, and move the disk back to the SAN.
    - **This issue can be eliminated with boot through dual VIOS**.

Source: IBM

Boot from VIOS Additional Notes

- Notes
  - The decision of where to place boot devices (internal, direct FC, VIOS), is independent of where to place data disks (internal, direct FC, or VIOS).
  - Boot VIOS off of internal disk.
    - LVM mirroring or RAID is supported for the VIOS's own boot disk.
    - VIOS may be able to boot from the SAN. Consult your storage vendor for multi-path boot support. This may increase complexity for updating multi-path codes
  - Consider mirroring one NIM SPOT on internal disk to allow booting in DIAG mode without SAN connectivity
    - nim -o diag -a spot=<spotname> clientname
  - PV-VSCSI disks are required with dual VIOS access to the same set of disks

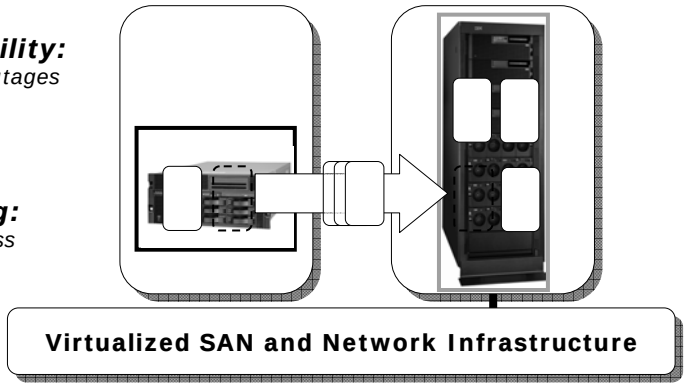
Source: IBM

Other – Sizing, etc

PowerVM Live Partition Mobility

Move running UNIX and Linux operating system workloads from one POWER6 processor-based server to another!

- ✓ ***Continuous Availability:*** *eliminate many planned outages*
- ✓ ***Energy Saving:*** *during non-peak hours*
- ✓ ***Workload Balancing:*** *during peaks and to address spikes in workload*



Source: IBM

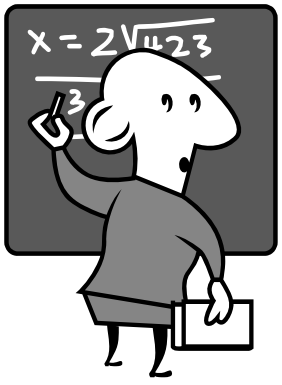
Live Partition Mobility Pre-Reqs

- All Systems in a Migration Set must be managed by the same HMC
  - HMC will have orchestration code to control migration function
- All Systems in a Migration Set must be on the same subnet.
- All Systems in a Migration Set must be SAN connected to shared physical disk – no VIOS LVM-based disks.
- ALL I/O must be shared/virtualized at the time of migration. Any dedicated I/O adapters must be deallocated prior to migration.
- Systems must be firmware compatible (within one release)

Partition Mobility – Other Considerations

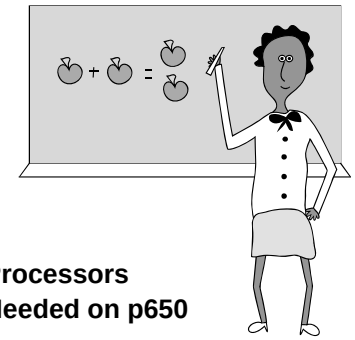
- Intended Use:
  - Workload Consolidation
  - Workload Balancing
  - Workload Migration to Newer Systems
  - Planned CEC outages for maintenance
  - Unplanned CEC outages where error conditions are picked up ahead of time.
- What it is not:
- A Replacement for HACMP or other clustering.
  - Not automatic
  - LPARs cannot be migrated from failed CECs
  - Failed OS's cannot be migrated
- Long Distance Support Not Available in First Release

Math 101 and Consolidation



- Consolidation Issues
- Math 101
  - 4 workloads
    - A 6.03
    - B 2.27
    - C 2.48
    - D 4.87
    - Total = 15.65
    - The proposed 8way is rated at 16.88
    - LPARs use dedicated processors
    - Is it big enough to run these workloads in 4 separate dedicated LPARs?
    - NO

Why micropartitioning is important



- 8w 1.45g p650 is 16.88 rperf
- 2w 1.45g p650 is 4.43 rperf
- So 1w is probably 2.21
- Now back to Math 101

Wkld	Rperf	Processors Needed on p650
A	6.03	3 (6.64)
B	2.27	2 (4.42 - 2.27 is > 2.21)
C	2.48	2 (4.42 – 2.48 is > 2.21)
D	4.87	3 (6.64 – 4.87 is > 4.42)
Total =	15.65	10 (22.12)

- Watch for granularity of workload

On Micropartitioned p5 with no other Virtualization

- 8w 1.45g p650 was 16.88 rperf
- 4w 1.65g p550Q is 20.25 rperf
- So 1w on 550Q is probably 5.06
  - BUT we can use 1/10 of a processor and 1/100 increments
- Now back to Math 101

Wkld	Rperf	Processors 650	Processors 550Q
A	6.03	3	1.2
B	2.27	2	.45
C	2.48	2	.49
D	4.87	3	.97
Total =	**15.65**	**10**	**3.11**

- Watch for granularity of workload
- On the p5 we use fewer processors and we fit!
- p6 is even better

General Server Sizing thoughts

- Correct amount of processor power
- Balanced memory, processor and I/O
- Min, desired and max settings and their effect on system overhead
- Memory overhead for page tables, TCE, etc
- Shared or dedicated processors
- Capped or uncapped
- If uncapped – number of virtual processors
- Expect to safely support 3 LPARs booting from a 146gb disk through a VIO server
- Don't forget to add disk for LPAR data for clients
- Scale by rPerf NOT by ghz when comparing boxes

VIOS Sizing thoughts

- Correct amount of processor power and memory
- Do not undersize memory
- Shared uncapped processors
- Number of virtual processors
- Higher weight than other LPARs
- Expect to safely support 3 LPARs booting from a 146gb disk through a VIO server
- Don't forget to add disk for LPAR data for clients
- Should I run 2 or 4 x VIOS'?
  - 2 for ethernet and 2 for SCSI?
  - Max is somewhere around 10
- Virtual I/O Server Sizing Guidelines Whitepaper
  - http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/perf.html
  - Covers for ethernet:
    - Proper sizing of the Virtual I/O server
    - Threading or non-threading of the Shared Ethernet
    - Separate micro-partitions for the Virtual I/O server

Sysplans and SPT

- System Planning Tool
  - http://www-03.ibm.com/servers/eserver/support/tools/systemplanningtool/
- Sysplans on HMC
  - Can generate a sysplan on the HMC
  - Print it to PDF and you are now documented as to how hardware is assigned to LPARs
- Peer Reviews and Enterprise Reviews
  - They will save you a lot of grief!

Best practices

- Plan plan document!
- Include backup (OS and data) and install methodologies in planning
- Don't forget memory overhead
- Do not starve your VIO servers
  - I start with .5 of a core and run them at a higher weight uncapped
  - I usually give them between 2GB and 3GB of memory
- Understand workload granularity and characteristics and plan accordingly
- Two VIO servers
- Provide boot disks through the VIO servers – you get full path redundancy that way
- Plan use of IVEs – remember they are not hot swap
- Evaluate each workload to determine when to use virtual SCSI and virtual ethernet and when to use dedicated adapters
- Consider whether the workload plays well with shared processors
- Based on licensing, use caps wisely when in the shared processing pool
- Be cautious of sizing studies – they tend to undersize memory and sometimes cores

Sizing Studies

- Sizing studies tend to size only for the application needs based on exactly what the customer tells them
- They usually do not include resources for:
  - Memory overhead for hypervisor
  - Memory and CPU needs for virtual ethernet and virtual SCSI
  - CPU and memory for the VIO servers
  - Hardware specific memory needs (i.e. each active IVE port takes 102MB)
- These need to be included with the results you get
- I have seen these be off by 2-3 cores and 24GB of memory so be wary

Traps for Young Players

- Under-sizing VIOS
- Over-committing boot disks
- Forgetting Memory and processor Overhead
- Planning for what should and should not be virtualized
- Misunderstanding needs
- Workload Granularity
- Undersizing memory and overhead
  - Hypervisor
  - I/O drawers, etc
  - VIOS requirements
  - Setting maximums
- Sizing studies
- Chargeback and capacity planning may need to be changed

Questions?