Planning for Virtualization on Power

Power Technical University
Session VN04
http://www.circle4.com/papers/ptechu2010.pdf

Jaqui Lynch
Solutions Architect
Forsythe Technology Inc.
lynchj@forsythe.com

FORSYTHE

Agenda

- Quick Introduction
- Planning for Shared processors, Shared ethernet, Virtual SCSI
- AMS and AME
- Live Partition Mobility
- NPIV

FORSYTHE

Virtualization Options

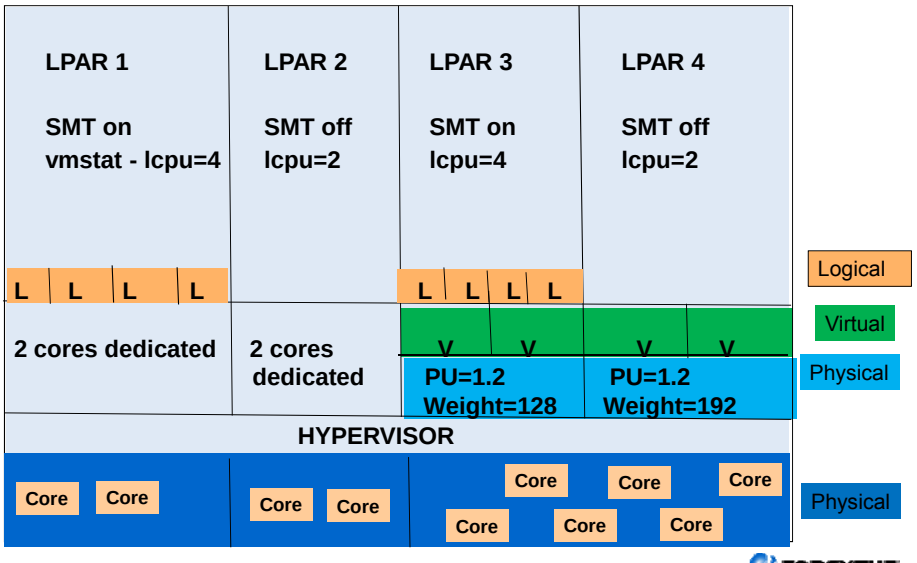
- Real Resources (unvirtualized)
  - Dedicated processors/cores
  - Dedicated fibre or SCSI
  - Dedicated Ethernet
- Virtual or Shared Resources
  - Shared processors/cores
  - Dedicated donating
  - Virtual ethernet - does not require PowerVM
  - Shared ethernet adapter
    - Built on virtual ethernet
  - Shared SCSI (aka Virtual SCSI)
    - Can be SCSI or fibre
  - NPIV
  - Ethernet and SCSI use a custom LPAR called a VIO server
    - Must include processor and memory resources in planning for that LPAR or LPARs
- Hybrids

3

FORSYTHE

Logical Processors

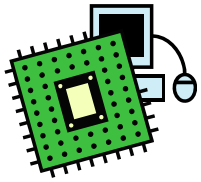
Logical Processors represent SMT threads



4

FORSYTHE

Defining Processors



- Minimum, desired, maximum
- Maximum is used for DLPAR
  - Max can be used for licensing
- Shared or dedicated
- For shared:
  - Desired is also called entitlement
  - Minimum of 1/10 of a core (10 lpars per core)
  - 1/10 of a core is a 1millisecond time slice
  - Partition's guaranteed amount of core is its Entitled Capacity
  - Capped
    - Can't exceed entitlement
  - 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

FORSYTHE

More on the shared processor pool

- Dispatch time for a full core is a 10ms timeslice
- A .4 entitlement means 40% of a core or a 4ms timeslice
  - How this is delivered depends on VPs
  - 2 VPs means 2 x 2ms timeslices on the same or two different cores
- When timeslice ends you will see an ICS (involuntary context switch) if thread still wants to keep running
- A 1.4 entitlement means the LPAR is entitled to 14ms of processing time for each 10ms timeslice (obviously across multiple cores) spread across the VPs
- Hypervisor dispatches excess idle time back to the pool (called a cede)
- Processor affinity tries to take into account hot cache
- LPAR may run on multiple cores depending on entitled capacity, virtual processors and interrupts

FORSYTHE

Uncapped vs Capped

- Capped LPARs can cede unused cycles back but can never exceed entitlement
- Uncapped LPARs can exceed entitlement up to the size of the pool or the total virtual processors, whichever is smaller
- Unused capacity is ceded back
- User defined weighting (0 to 255) is used to resolve competing requests
- Weights are share based
  - 2 LPARs need 3 cores each
  - Only 3 cores available
  - If A is 100 and B is 200 then A gets 1 core and B gets 2 cores
- Use common sense when planning your use of weights and remember the default is 128
  - Prod VIO 192
  - Prod 160
  - Test/Dev 128
  - Have a plan, not necessarily this one – document it well

FORSYTHE

Virtual Processors 1

- Used to tell the operating system how many physical processors it thinks it has
- 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
  - For every 1.00 or part thereof of a processor unit there will be at least 1 VP allocated
  - .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
- VPs Cannot exceed 10x entitlement

FORSYTHE

Virtual Processors 2

- VPs are dispatched to real processors
- Dispatch latency – minimum is 1 millisec and max is 18 millisecs
  - Dispatch latency is the time between a virtual processor becoming runnable and being actually dispatched.
  - http://www.ibm.com/developerworks/wikis/display/virtualization/POWER5+Hypervisor
- VP Folding
- Maximum is used by DLPAR
- Use commonsense when setting max VPs!!!
- In a single LPAR VPs should never exceed Real Processors
- Maximum VPs per partition is 64 or 10 x entitlement (desired PUs)
- Operating system does not see entitlement – it sees configured virtual processors
  - EC=1.4
  - VP = 2
  - Operating system sees 2 processors (proc0 and proc2)
  - Each proc is really 70% of a core in value (7ms)
  - If SMT is on then you will see those broken down into 4 logical cpus (cpu0, cpu1, cpu2, cpu3)

FORSYTHE

Virtual Processors 3

- Both entitled capacity and number of virtual processors can be changed dynamically for tuning
- Virtual processor count does not change entitlement –
  - It is about how the entitlement will be delivered
- Capped
  - Entitlement = 1.4 and 2 x VPs
  - For each 10ms timeslice the LPAR is entitled to 14ms of processing time
  - For 2 VPs that equates to 7ms on each of 2 physical processors
- Uncapped
  - Entitlement = 1.4 and 2 x VPs
    - Start at 7ms each
  - Each VP can grow to 10ms of processor time
  - For 2 VPs that equates to 20ms across 2 physical processors
  - Can't grow beyond 2 VPs even if more are available in the pool
  - VPs becomes a soft cap on the LPAR
  - Be sure to allocate enough VPs to grow for uncapped LPARs

FORSYTHE

Examples

- LPAR 1 - uncapped
  - PU Ent = 2.0
  - PU Max = 6.0
  - Desired VPs = 4.0
  - Can grow to 4 processor units
  - Can't grow beyond 4.0 as VPs cap this
- LPAR2 – Set as Capped
  - PU Ent =2.0
  - PU Max = 6.0
  - Desired VPs = 4.0
  - Can't grow at all beyond 2 processor units (entitlement)
  - A weight of 0 makes the LPAR act as if it is capped
- In both examples above each VP is around .5 of a core or 5ms
  - Ent/VPs

FORSYTHE

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 or 5ms
  - 4 cores with 4 VPs
    - Each dispatch window is 1 processor unit or 10ms
  - Which one matches your workload the best?
- Too few VPs
  - Uncapped LPARs can't grow to use excess cycles as capped by VPs
- Too many VPs
  - May cause excessive processor context switching
  - AIX v5.3 ML3 – virtual processor folding implemented
- ROT
  - While more may be better the maximum is not always the smart choice

FORSYTHE

Multiple Shared Processor Pools

- POWER6 and POWER7 processor-based systems
- Grouping of partitions into subsets called “Pools”
- Default is one pool (pool 0) containing all cores in system
- Can manage processors resources at the subset
  - Can assign caps at the group level
- Can balance cpu resources between partitions assigned to the shared pools and optimize use of processor cycles
- Segment production / development / test / or Oracle or WAS, etc
- Mobility of partitions in a pool is supported
- Maximum: 64 Pools
- Virtual shared pools are created from the HMC interface
- Maximum capacity of a shared processor pool can be adjusted dynamically
- Partitions can be moved between virtual shared processor pools
- Entitled pool capacity dynamically changes as LPARs get added and removed

FORSYTHE

Sharing Ethernet

- Potential Benefits
  - Reduce the number Ethernet adapters, ports
    - Better resource utilization
  - Reduce cabling efforts and cables in frames
  - Reduce number of I/O drawers and/or frames
  - Required for Live Partition Mobility (LPM)
- Issues/Considerations
  - Understand Ethernet adapter/port utilization
  - Understand high availability cluster support requirements
  - Understand implications on backup architecture
  - Understand virtual I/O sizing issues
  - Understand use of link aggregation (also called teaming or etherchannel) and/or VLANS
  - Understand VIO high availability Ethernet options
  - Simplicity!!
- Need to plan as it is not an either/or choice – you can have both dedicated and shared resources

FORSYTHE

Virtual I/O Limits

• Maximum virtual Ethernet adapters per LPAR	256
• Maximum VLANs per virtual adapter	21 VLAN 20 VID, 1 PVID
• Virtual adapters per SEA sharing a physical adapter	16
• Maximum VLAN IDs	4094
• Maximum virtual ethernet frame size	65,408 bytes
• Maximum physical adapters in a link aggregation	8 primary, 1 bkup

- **Just because you can does not mean you should**
  - Do not plan to use anything close to these maximums
  - The documentation alone will cause nightmares

15

FORSYTHE

IVE Notes

- Which adapters do you want? Each CEC requires one.
  - Choices between copper Gb, Fiber 10GB, etc
- Adapter ties directly into GX Bus
  - No Hot Swap and no Swap Out for Different Port Types (10GbE, etc.)
- Not Supported for Partition Mobility, except when assigned to VIOS
- Pay attention to MCS (multi core scaling) – affects scaling and number of logical ports. Default is 4
- Partition performance is at least the same as a real adapter
  - No VIOS Overhead
  - Intra-partition performance may be better than using Virtual Ethernet
- Naming
  - Integrated Virtual Ethernet – Name used by marketing
  - Host Ethernet Adapter (HEA) Name used on user interfaces and documentation

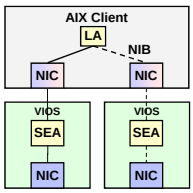
http://www.redbooks.ibm.com/abstracts/redp4340.html
Integrated Virtual Ethernet Adapter – Technical Overview

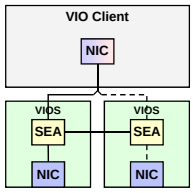
16

FORSYTHE

High Availability VIOS Options

- Network Interface Backup
  - Must be set up in each client.
  - Needs to ping outside host from each client to initiate NIB failover.
  - Load share clients across SEAs but LPAR to LPAR communications will happen through external switches
  - VLAN-tagged traffic is not supported.
  - AIX only.
- Shared Ethernet Adapter Failover
  - Set up in the VIOS's only
  - Optional ping is done in VIOS on behalf of all clients
  - Cannot load-share clients between the primary and backup SEA
  - VLAN-tagged traffic is supported
  - Supported on all AIX, IBM i, Linux



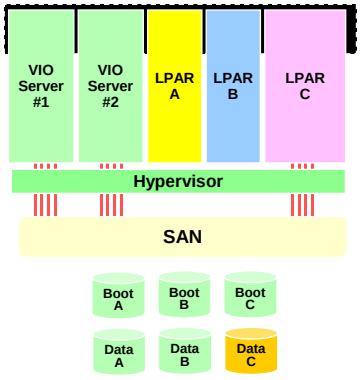


17

FORSYTHE

Sharing Disk Adapters

- Potential Benefits
  - Reduce the number FC adapters and ports and switches
  - Reduce cabling efforts and cables in frames
  - Reduce number of I/O drawers and frames.
  - Required for Live Partition Mobility (LPM)
- Issues/Considerations
  - Understand current SAN adapter / port use
  - Investigate high availability cluster support for virtual I/O
  - Understand implications on backups
  - Understand virtual I/O server sizing
  - Understand availability choices such as dual VIOS, number of HBAs, O/S mirroring, etc
  - This is not an either/or choice – you can have both dedicated and shared resources
  - For LPM need to get away from using LVs – use full hdisks to client LPARs
  - Include LPM potential in planning



18

FORSYTHE

Boot from SAN vs. Boot from Internal Disk

- Performance Advantages
  - 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
- Configuration Advantages
  - 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)
  - Reduce I/O drawers and frames
  - Generally easier to find space for a new image on the SAN
- Boot Advantages
  - Booting through the VIOS could allow pre-cabling and faster deployment of AIX

- Disadvantages
  - SAN must be robust
  - Loss of SAN is disruptive
  - Loss of SAN will prevent a dump if dump device on SAN
- SAN Boot Issues
  - Multi-path driver installs and upgrades more complex
    - They are in use by AIX
    - May have to uninstall and reinstall multi-path software to upgrade it
    - Could require exportvg and importvg or move disk from/to SAN
    - Issue goes away with boot through dual VIOS.

LPM requires LPARs to boot from SAN LUNs (not LVs or internal)

VIOS can boot from internal

FORSYTHE

LPAR Layouts for Planning

LPAR LAYOUTS for P770 Server

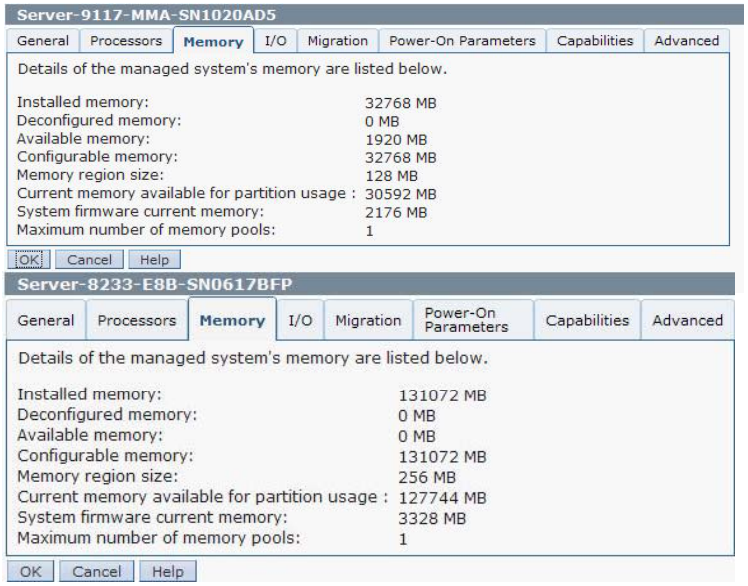
Using Dedicated cores
Server has 32 cores installed with 32 active
Server has 512GB installed with 418GB active

LPAR	Cores Desired	Cores Maximum	Memory Desired GB	Memory Maximum GB	Copper 1GB Eth	Fibre 10GB Eth	Fibre HBA ports
VIO1	0.5	2	3	4	2	1	12
VIO2	0.5	2	3	4	2	1	12
Oracle	16	24	370	384			
WAS	8	16	16	32			
Test Ora	6	8	8	16			
Free	1		10				
Overhead			8				
	32		**418**	**440**	**4**	**2**	**24**

FORSYTHE

Memory Usage

From HMC



Note firmware Use

Also note memory region size

You need to know it for LPM

FORSYTHE

Planning for Memory

This gives a rough estimate
Assumes LMB size is 32mb
each active IVE port adds 102MB
Don't forget memory overhead

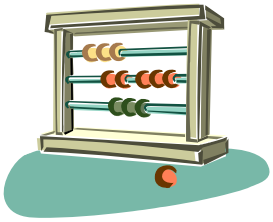
Memory Planning Worksheet
Power7 750
Server 1

Max RAM Capacity	**524288 Ram installed GB**		**131072 Ram Active**		**131072**		
LPAR	**Desired Memory (MB)**	**Maximum Memory (MB)**	**128 Ohead Max Div 64**	**LMB=128 OH/LMB**	**Roundup OH/32**	**Actual Ohead (MB)**	**MB Memory Needed**
VIOS1	2048	4096	64	2.00	2	64	
VIOS2	2048	4096	64	0.50	1	64	
LPAR1	30720	61440	960	7.50	8	960	
LPAR2	23552	47104	736	5.75	6	742	
LPAR3	24576	49152	768	6.00	6	768	
LPAR4	24576	49152	768	6.00	6	768	
LPAR5	22528	45056	704	5.50	6	704	
HYPERVISOR						768	
IVE						106	
Safety Net						512	
MB Total	**130048**	**260096**	**4064**	**33.25**	**33.3**	**5456**	**135504**
GB Total	***127***					***5***	***132***

Hypervisor requires 5.5GB for overhead with these settings
LPARs require 127GB so the total needed is 132GB - this will not fit in 128GB of memory
Solution - reduce maximum setting on larger LPARs to reduce overhead

FORSYTHE

Math 101 and Consolidation



- Old example but a good one
- Consolidation Issues
- Math 101
  - 4 workloads with no growth and peaks of (in rPerf)
    - A 6.03
    - B 2.27
    - C 2.48
    - D 4.87
    - Total = 15.65
    - The proposed 8way is rated at 16.88 rPerf
    - LPARs use dedicated processors
    - Is it big enough to run these workloads in 4 separate dedicated LPARs?
    - NO

FORSYTHE

Why Math 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)

- Same applies in POWER5, 6 and 7 but granularity is 1/10
  - Make sure you round up to 1/10 of a core when sizing
- **Watch for granularity of workload**
- **Convert each workload to cores and then total cores – do not total rPerf**

FORSYTHE

Virtual Ethernet & SEA

- General Best Practices
  - Keep things simple
  - Use hot-pluggable network adapters for the VIOS instead of the built-in integrated network adapters. They are easier to service (hot pluggable)
  - Use dual VIO Servers to allow concurrent online software updates to the VIOS.
  - Configure an IP address on the SEA itself.
    - 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.
  - Be consistent – I always create the network using the same adapter ids in the VIOS and in every client
    - Server = 11-13 for first SEA
    - Client = 2 for first SEA
- Document everything including the commands to create the VEs and SEAs

FORSYTHE

Virtual Ethernet & SEA

- 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
  - Keep the attribute tcp_pmtu_discover set to “active discovery”
  - Set the network tunables (using no) as you would on any other AIX LPAR
  - 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 (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.
- Plan configuration carefully and include CPU and memory in the VIO to support these

FORSYTHE

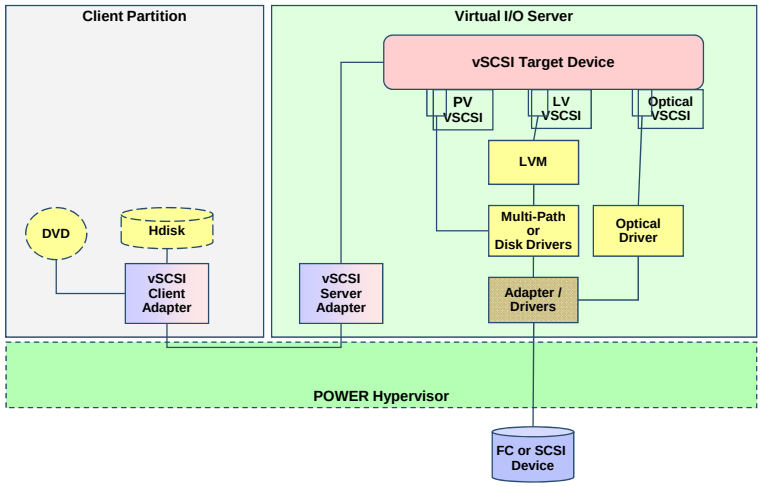
Virtual Ethernet Performance

- Performance - Rules of Thumb
  - Choose the largest MTU size that makes sense for the traffic on the virtual network.
    - Default is 1500 – talk to network people if want to change this
  - CPU utilization for large packet workloads on jumbo frames (9000) 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 to send 64kB of packet data through a Virtual Ethernet connection irrespective of the actual MTU size
    - This results in fewer 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

FORSYTHE

Virtual SCSI Basic Architecture



Source: IBM

FORSYTHE

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
    - If using internal disk for the VIOS then mirror the disk
  - For performance reasons, logical volumes within the VIOS that are exported as virtual SCSI devices should not be striped, mirrored, span multiple physical drives, or have bad block relocation enabled..
    - Remember you can't use LVs if the client is to use LPM
  - 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

FORSYTHE

Virtual SCSI General Notes….

- Notes
  - If you are using FC Multi-Path I/O on the VIOS, set the following fscsi device values (requires switch attach):
    - dyntrk=yes (Dynamic Tracking of FC Devices)
    - fc_err_recov= fast_fail (FC Fabric Event Error Recovery Policy)
      (must be supported by switch)
  - If you are using MPIO on the VIOC, set the following hdisk device values:
    - hcheck_interval=60 (Health Check Interval)
  - If you are using MPIO on the VIOC set the following hdisk device values on the VIOS:
    - reserve_policy=no_reserve (Reserve Policy)

FORSYTHE

Sizing for vSCSI

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/topic/iphb1_p5/iphb1_vios_planning_vscsi_sizing.htm
Virtual SCSI sizing using dedicated processor partitions
Processor entitlement for vSCSI server - based on maximum I/O rates it is expected to serve
VIO Server - 3 client LPARs
LPAR 1 max = 7,000 x 8-KB ops/sec
LPAR 2 max = 10,000 8-KB ops/sec
LPAR 3 max = 5,000 x 128-KB ops/sec
((7,000 × 47,000 + 10,000 × 47,000 + 5,000 × 120,000) / 1,650,000,000) = 0.85 cores
Above is number of 1.65ghz cores needed – for a 4.7ghz p6 it ends up as 0.66 cores

Table 1. Approximate cycles per second on a 1.65 Ghz partition

Disk type	4 KB	8 KB	32 KB	64 KB	128 KB
Physical disk	45,000	47,000	58,000	81,000	120,000
Logical volume	49,000	51,000	59,000	74,000	105,000

For POWER6 and POWER7 these are included in the Workload Estimator
http://www-912.ibm.com/wle/EstimatorServlet

FORSYTHE

VIOS Sizing thoughts 1

- Correct amount of processor power and memory
- Do not undersize memory –
  - I use 3GB each by default for VIOS with a max of 3GB
  - Virtual SCSI server does not cache file data in memory
  - Minimum memory for a VSCSI VIOS is 1GB
  - Minimum memory for a Virtual Ethernet VIOS is 512MB
    - 4MB per physical gbit ethernet adapter (MTU 1500) or 16MB if 9000 for receive buffers
    - 6MB per virtual ethernet adapter
    - The above are per instance
    - System wide mbuf pool per processor – typically around 40MB
- Shared uncapped processors
- Number of virtual processors
  - I normally start with entitlement of 0.5 and VPs of 2 for each VIOS

FORSYTHE

VIOS Sizing thoughts 2

- Higher weight than other LPARs
- Don't forget to add disk for LPAR data as well as boot disks 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
  - An older paper but has some useful pointers
  - 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

FORSYTHE

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 that are used by virtualization
- Shared or dedicated processors
- Capped or uncapped
- If uncapped – number of virtual processors
- Remember boot disks for clients
- Don't forget to add disk for LPAR data for clients
- **Scale by rPerf (or other benchmark data) NOT by ghz when comparing boxes**

FORSYTHE

Best practices

- Plan and plan and 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 (2 VPs) 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
- Determine when to use virtual SCSI (or NPIV) 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
- Tune the VIO server and run performance tools like nmon to gather data
- Be cautious of sizing studies – they tend to undersize memory and sometimes cores and usually do not include the VIO server needs

35

FORSYTHE

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)
- They often round up to whole cores as they don't take into account that you may be consolidating
- I have seen where they just round up to the nearest server size
- They do not test every path when doing sizing studies so if your workload is a little unusual you need to be careful
- Often do not include database buffer areas so memory can be off
- I have seen these be off by 2-3 cores and 40GB of memory so be wary

36

FORSYTHE

Traps for Young Players

- Forgetting Memory and processor Overhead
- Planning for what should and should not be virtualized
- Workload Granularity and Characterization
- Undersizing cores, memory and overhead
  - Hypervisor, I/O drawers, VIOS requirements
  - Setting maximums
- Be cautious of sizing studies – they may undersize memory and cpu needs
- Evaluate each workload to determine when to use resources and when to use dedicated ones
- Consider whether the workload plays well with shared processors
- Based on licensing, use caps wisely when in the shared processing pool
- Chargeback and capacity planning may need to be changed
- Do your management and monitoring tools support virtualized environments?
- **Plan and plan and document! This is critical**

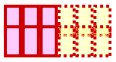
FORSYTHE

Active Memory Expansion

- Innovative POWER7 technology
  - For AIX 6.1 or later
  - For POWER7 servers
  - Must purchase the AME feature for the server
  - Turned on by LPAR
- Uses compression/decompression to effectively expand the true physical memory available for client workloads
- Often a small amount of processor resource provides a significant increase in the effective memory maximum
  - Processor resource part of AIX partition's resource and licensing
- Actual expansion results dependent upon how "compressible" the data being used in the application
  - A SAP ERP sample workload shows up to 100% expansion,
  - BUT your results will vary
  - Estimator tool (amepat) and free trial available

FORSYTHE

Active Memory Expansion - Planning Tool



```
            Active Memory Expansion Modeled Statistics:
                      -----------------------
                Modeled Expanded Memory Size :   8.00 GB

  Expansion      True Memory       Modeled Memory       CPU Usage
   Factor       Modeled Size        Gain                Estimate
  ---------    --------------    -----------------    -----------
     1.21          6.75 GB         1.25 GB [ 19%]         0.00
     1.31          6.25 GB         1.75 GB [ 28%]         0.20
     1.41          5.75 GB         2.25 GB [ 39%]         0.35
   1.51          5.50 GB         2.50 GB[ 45%]          0.58
     1.61          5.00 GB         3.00 GB [ 60%]         1.46

            Active Memory Expansion Recommendation:
                      ---------------------
The recommended AME configuration for this workload is to configure
  the LPAR with a memory size of 5.50 GB and to configure a memory
expansion factor of 1.51.  This will result in a memory expansion of
 45% from the LPAR's current memory size.  With this configuration,
     the estimated CPU usage due to Active Memory Expansion is
  approximately 0.58 physical processors, and the estimated overall
peak CPU resource required for the LPAR is 3.72 physical processors.
```

Sample output

This sample partition has fairly good expansion potential

A nice "sweet" spot for this partition appears to be 45% expansion

- 2.5 GB gained memory
- Using about 0.58 cores additional CPU resource

Example from IBM

Tool included in AIX 6.1 TL4 SP2
Run tool in the partition of interest for memory expansion.
Input desired expanded memory size. Tool outputs different real memory and CPU resource combinations to achieve the desired effective memory.

FORSYTHE

amepat output on a non-busy system

```
# amepat -m 1000 2

Date/Time of invocation                 : Mon Sep  6 11:03:42 CDT 2010
Total Monitored time                    : 2 mins 5 secs
Total Samples Collected                 : 2

System Configuration:
---------------------
Partition Name                  :       p6db53a
Processor Implementation Mode   : POWER7_in_P6_mode
Number Of Logical CPUs                  : 4
Processor Entitled Capacity             : 0.20
Processor Max. Capacity                 : 2.00
True Memory                             : 4.00 GB
SMT Threads                             : 2
Shared Processor Mode                   : Enabled-Uncapped
Active Memory Sharing                   : Disabled
Active Memory Expansion                 : Disabled

System Resource Statistics:     Average         Min             Max
---------------------------     -----------     -----------     -----------
CPU Util (Phys. Processors)      0.02 [  1%]     0.01 [  1%]     0.03 [  2%]
Virtual Memory Size (MB)         945 [ 23%]      945 [ 23%]      946 [ 23%]
True Memory In-Use (MB)          1119 [ 27%]     1119 [ 27%]     1120 [ 27%]
Pinned Memory (MB)               756 [ 18%]      756 [ 18%]      756 [ 18%]
File Cache Size (MB)             157 [  4%]      157 [  4%]      157 [  4%]
Available Memory (MB)            3005 [ 73%]     3005 [ 73%]     3006 [ 73%]
```

This was running on a POWER6 on AIX v7 beta

FORSYTHE

amepat recommendations from amepat report

The recommended AME configuration for this workload is to configure the LPAR with a memory size of 2.50 GB and to configure a memory expansion factor of 1.60. This will result in a memory gain of 60%. With this configuration, the estimated CPU usage due to AME is approximately 0.00 physical processors, and the estimated overall peak CPU resource required for the LPAR is 0.03 physical processors.

NOTE: amepat's recommendations are based on the workload's utilization level during the monitored period. If there is a change in the workload's utilization level or a change in workload itself, amepat should be run again.

The modeled Active Memory Expansion CPU usage reported by amepat is just an estimate. The actual CPU usage used for Active Memory Expansion may be lower or higher depending on the workload.

FORSYTHE

Act Mem Exp – Operations Considerations

- Active Memory Expansion is transparent to applications
- A server using Active Memory Expansion needs an HMC on the server.
- Enabling Active Memory Expansion: does NOT require a server IPL
- Turning Active Memory Expansion on/off for a partition DOES require an IPL of that partition.
  - Changing the expansion factor does NOT require an IPL.
  - If expansion factor set to 1.0 , it is the same as memory expansion turned off
  - But note: AME uses the smaller memory page sizes, not some of the larger page sizes. For a small percentage of clients this page size may be a factor.
- Hardware & software requirements
  - POWER7 server running in POWER7 mode
  - HMC: V7R7.1.0.0 or later
  - Firmware: 7.1
  - AIX 6.1 TL4 SP2 or later

FORSYTHE

Active Memory Expansion – Miscellaneous

- For LPM and AME both servers must have AME
- If server gets really busy AME will continue and workload will slow down as it would anyway when the server gets busy
- WLE (Work Load Estimator) tool does not currently include AME yet
- Do not confuse AME (Act Mem Exp) with AME (AIX Management Edition).
- Enablement is via a “VET” code applied to the VPD information on the anchor card
- You can use Active Memory Expansion in the shared processor pool if the server is enabled.
  - Each partition in the pool turns Act Mem Exp on or off independently.
  - Each partition’s expansion factor is independent.
- 60-day trial period is possible
- It is not an expensive add-on to the server

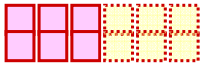
43

FORSYTHE

Active Memory Expansion & Active Memory Sharing

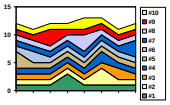
Active Memory Expansion

- **Effectively gives more memory capacity to the partition using compression / decompression of the contents in true memory**
- **AIX partitions only**



Active Memory Sharing

- Moves memory from one partition to another
- Best fit when one partition is not busy when another partition is busy
- AXI, IBM i, and Linux partitions



Active Memory Expansion
Active Memory Sharing

PLUS

- Supported, potentially a very nice option
- Considerations
  - Only AIX partitions using Active Memory Expansion
  - Active Memory Expansion value is dependent upon compressibility of data and available CPU resource

Courtesy IBM

44

FORSYTHE

Live Partition Mobility

45

FORSYTHE

LPM – Live Partition Mobility

- Uses
  - Server Consolidation
  - Workload Balancing
  - Can use to allow power off of unused servers to assist in Green IT efforts
  - Planned maintenance and upgrades
- Inactive partition migration moves a powered-off partition
  - Not a crashed kernel
- Partitions cannot be migrated from failed machines
- Network applications may see a brief (~2 sec) suspension toward the end of the migration, but connectivity will not be lost
- IT IS NOT A REPLACEMENT FOR HACMP OR OTHER HA or DR solutions

46

FORSYTHE

Requirements for LPM

- **PLANNING IS CRITICAL**
- http://www14.software.ibm.com/webapp/set2/sas/f/pm/component.html
- Hardware POWER6 Only
- HMC v7.3.2 with MH01062
- Firmware E*340_039 min
  - http://www14.software.ibm.com/webapp/set2/sas/f/pm/migrate.html
- AIX v5.3 5300-07-01
- AIX v6.1 6100-00-01
- VIO Server 1.5.2.1-FP-11.1 or v2.1
- Two Power6 or Power7 systems managed by HMC or IVM (no mixing)
- PowerVM Enterprise Edition
- Virtualized SAN storage (rootvg and all other vgs)
- Virtualized Ethernet (SEA)
- LPAR being moved cannot be using the HEA/IVE (VIO can though)
- RHEL5 Update 1 and SLES10 Update 1 supported (or later)
- Check the prereq site:
  - https://www-912.ibm.com/e_dir/eserverprereq.nsf
- **No dedicated anything at the time of the move**

47

FORSYTHE

Other LPM Prereqs

- Two servers – POWER6, 6+ or 7 (or mix thereof)
- Managed by a single HMC or IVM on each server
- HMC v7.3.4 introduces remote migration
  - Partitions can migrate between systems managed by different HMCs
- Must be on same subnet as each other
- POWERVM Enterprise Edition
- Check VIOS levels
  - Many new features require v2.1 or higher
- Storage must be virtualized
  - Storage must be zoned to both source and target
  - No LVM based disks
  - hdisks must have reserve_policy=no_reserve
  - See section 3.7 of the LPM red book SG24-7460
- Must use Shared Ethernet Adapter
  - See section 3.8 of the LPM red book SG24-7460
- All resources must be shared or virtualized prior to migration
- **Must have resources available at the target**

48

FORSYTHE

LPM

- Check LPAR on HMC under Capabilities
  - Look for Active and Inactive Partition Mobility Capable=True
- Ensure VIO server is set up as a Mover Service Partition (MSP) under the general tab on the VIO server at each end
  - By default MSP is set to no on a VIO server
- Mover partition must have a VASI (Virtual Asynchronous Services Interface) device defined and configured (done automatically by HMC)
- The pHypervisor will automatically manage migration of CPU and memory
- Dedicated IO adapters must be de-allocated before migration
- cd0 in VIO may not be attached to mobile LPAR as virtual optical device
- Time of Day clocks for VIO servers should be synchronized
- The operating system and applications must be migration-aware or migration-enabled
- Oracle 10G supports LPM
- LMB (memory region) size must be the same on both servers – check on HMC
  - Requires a whole server reboot to change
- **Ensure there is sufficient memory and core at the target for the workload to be moved**

FORSYTHE

Requirements for Remote Migration

- Ability to use LPM between 2 servers on different HMCs

- A local HMC managing the source server
- A remote HMC managing the target server
- Version 7.3.4 or later of the HMC software
- Network access to the remote HMC
- SSH key authentication to the remote HMC
- Plus all the other requirements for single HMC migration

FORSYTHE

NPIV
N_Port ID Virtualization (Virtual FC)

51

FORSYTHE

NPIV Overview

- N_Port ID Virtualization (NPIV) is a fibre channel industry standard method for virtualizing a physical fibre channel port.
- NPIV allows one F_Port to be associated with multiple N_Port IDs, so a physical fibre channel HBA can be shared across multiple guest operating systems in a virtual environment.
- On POWER, NPIV allows logical partitions(LPARs) to have dedicated N_Port IDs, giving the OS a unique identity to the SAN, just as if it had a dedicated physical HBA(s).
- N_Port ID Virtualization
  - Virtualizes FC adapters
  - Virtual WWPNs are attributes of the client virtual FC adapters not physical adapters
  - 64 WWPNs per FC port (128 per dual port HBA)

52

FORSYTHE

NPIV Specifics

- VIOS V2.1 (PowerVM Express, Standard, and Enterprise)
- client OS support: AIX(5.3 and 6.1), Linux(2009), and IBM i (2009)
- POWER6 and POWER7
- 8Gb PCIe HBA
- Unique WWPN generation (allocated in pairs)
- Each virtual FC HBA has a unique and persistent identity
- Compatible with LPM (live partition mobility)
- VIOS can support NPIV and vSCSI simultaneously
- Each physical NPIV capable FC HBA will support 64 virtual ports
- HMC-managed and IVM-managed servers

FORSYTHE

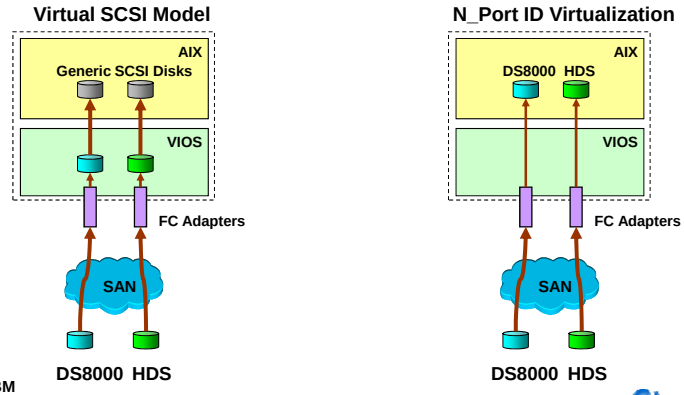
NPIV - Things to consider

- Requires VIOS 2.1 minimum
- First switch must be NPIV enabled
- WWPN pair is generated EACH time you create a VFC. NEVER is it re-created or re-used.
- Just like a real HBA.
- If you create a new VFC, you get a NEW pair of WWPNs.
- Save the partition profile with VFCs in it. Make a copy, don't delete a profile with a VFC in it.
- Make sure the partition profile is backed up for local and disaster recovery! Otherwise you'll have to create new VFCs and map to them during a recovery.
- Target Storage SUBSYSTEM must be zoned and visible from source and destination systems for LPM to work.
- Active/Passive storage controllers must BOTH be in the SAN zone for LPM to work
- Do NOT include the VIOS physical 8G adapter WWPNs in the zone
- You should NOT see any NPIV LUNs in the VIOS
- Load multi-path code in the client LPAR, NOT in the VIOS
- No 'passthru' tunables in VIOS

FORSYTHE

N_Port ID Virtualization Simplifies Disk Management

- N_Port ID Virtualization
  - Multiple Virtual World Wide Port Names per FC port – PCIe 8 Gb adapter
  - LPARs have direct visibility on SAN (Zoning/Masking)
  - I/O Virtualization configuration effort is reduced



Courtesy IBM

55

FORSYTHE

NPIV Planning

Current Allocations to LPARs

LPAR	Desired Core	Des VPs	Max VPs	Memory
p7jaapp1	4	8	16	16GB
p7jaapp2	4	8	16	8GB
p7jaapp3	4	8	16	8GB

For p7jaapp1:

Client adapter	3	FC from VIO1 adapter 30
Client adapter	5	FC from VIO1 adapter 32
Client adapter	4	FC from VIO2 adapter 30
Client adapter	6	FC from VIO2 adapter 32
Client adapter	7	vSCSI from VIO1 adapter 20 (vtopt)
Client adapter	8	vSCSI from VIO2 adapter 20 (vtopt)

I use:	**Server**	**Client**	
	10-19 Ethernet	2	
	20-29 vSCSI	7-8	
	30-39 NPIV	3-6	reserve on LPM target as well

56

FORSYTHE

Sample WWPN Documentation

WWPNs for zoning

All WWPNs start C050 7602 A0F9

VIO Server	LPAR Name	Fibre Card	Top or Bottom	WWPN if NPIV	Server Adapter	Client Adapter	vfchost	App1 1 x 50gb 2 x 100gb B	App2 1 x 50gb 2 x 100gb C
p7javio1	p7jaapp1	fcs0	top-npiv	0008	32	4	2	b	
p7javio1	p7jaapp1	fcs0	top-lpm	0009	32	4	2	b	
p7javio1	p7jaapp1	fcs1	bottom-npiv	000C	33	6	3	b	
p7javio1	p7jaapp1	fcs1	bottom-lpm	000D	33	6	3	b	
p7javio1	p7jaapp2	fcs0	top-npiv	0010	34	4	4		c
p7javio1	p7jaapp2	fcs0	top-lpm	0011	34	4	4		c
p7javio1	p7jaapp2	fcs1	bottom-npiv	0014	35	6	5		c
p7javio1	p7jaapp2	fcs1	bottom-lpm	0015	35	6	5		c

WWPNs for zoning

All WWPNs start C050 7602 A0F9

VIO Server	LPAR Name	Fibre Card	Top or Bottom	WWPN if NPIV	Server Adapter	Client Adapter	vfchost	App1 1 x 50gb 2 x 100gb B	App2 1 x 50gb 2 x 100gb C
p7javio2	p7jaapp1	fcs0	top-npiv	000A	32	5	2	b	
p7javio2	p7jaapp1	fcs0	top-lpm	000B	32	5	2	b	
p7javio2	p7jaapp1	fcs1	bottom-npiv	000E	33	7	3	b	
p7javio2	p7jaapp1	fcs1	bottom-lpm	000F	33	7	3	b	
p7javio2	p7jaapp2	fcs0	top-npiv	0012	34	5	4		c
p7javio2	p7jaapp2	fcs0	top-lpm	0013	34	5	4		c
p7javio2	p7jaapp2	fcs1	bottom-npiv	0016	35	7	5		c
p7javio2	p7jaapp2	fcs1	bottom-lpm	0017	35	7	5		c

FORSYTHE

References

- Active Memory Expansion Forum and Wiki
  - www.ibm.com/developerworks/forums/forum.jspa?forumID=2179&start=0
  - www.ibm.com/developerworks/wikis/display/WikiPtype/IBM+Active+Memory+Expansion
- IBM Infocenter at
  - http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp
  - AIX Commands Reference
  - AIX Performance Management Guide
- White papers
  - "Active Memory Expansion: Overview and Users Guide" by David Hepkin
  - "Active Memory Expansion: Performance Considerations" by Dirk Michel
  - www.ibm.com/systems/power/resources/index.html (then click on white papers) or www.ibm.com/support/techdocs/atsmastr.nsf/Web/TechDocs
- Movie on introduction, technology, use, installation, operation (18 min)
  - http://www.ibm.com/developerworks/wikis/display/WikiPtype/Movies by Nigel Griffiths

FORSYTHE

References

- LPM Prereqs
  - http://www14.software.ibm.com/webapp/set2/sas/f/pm/component.html
- SG24-7460 IBM PowerVM Live Partition Mobility
  - http://www.redbooks.ibm.com/abstracts/sg247460.html?Open
- SG24-7940-03 PowerVM Virtualization : Introduction and Configuration
  - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg247940.html?OpenDocument
- SG24-7590 PowerVM Virtualization Managing and Monitoring
  - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg247590.html?OpenDocument
- REDP-4340-00 IVE Adapter Technical Overview and Introduction
  - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/redp4340.html?OpenDocument
- SG24-7825 PowerVM Migration from Physical to Virtual Storage
  - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg247825.html?OpenDocument

FORSYTHE

Questions???

FORSYTHE