

1. Stay current (don't forget NIM)



3

3

Patch order

1. Read the readmes
2. Usually do HMC first
3. Then server firmware
4. NIM server (it should be standalone and needs to be at highest level) plus I/O firmware
5. Then VIO servers and any I/O firmware
6. LPARs – AIX, IBM i, Linux
7. Note I document everything
8. I write up every step of the upgrade before I do it and then tweak as I go along

4

4

Use FLRT and check Prereqs

FLRT Home Page:

<http://www14.software.ibm.com/webapp/set2/flrt/home>
<https://www-304.ibm.com/support/customercare/flrt/>

FLRT Lite

<http://www14.software.ibm.com/webapp/set2/flrt/liteHome>

VIOS to NIM Master Mapping:

<http://www14.software.ibm.com/webapp/set2/flrt/sas?page=viostable>

System Software Maps for VIOS:

<http://www-01.ibm.com/support/docview.wss?uid=ssm1platformvios>

AIX/VIOS Security Tables:

http://www14.software.ibm.com/webapp/set2/sas/f/flrt3/Sec_APARs.html

VIOS Hiper Tables:

http://www14.software.ibm.com/webapp/set2/flrt/doc?page=hiper#vios_hiper

Also check MPIO driver versions as there are specific requirements for each VIO release

AIX Support Lifecycle

<https://www-01.ibm.com/support/docview.wss?uid=isg3T1012517>



5

5

Stay current with AIX

- Don't forget HMC, server & I/O firmware, VIO updates, etc
- <https://www14.software.ibm.com/webapp/set2/flrt/liteTable?prodKey=aix>
- AIX 7.1 prior to AIX 7100-05-00 went eosps 12/31/2019
 - Latest is 7100-05-08
- AIX 7.2 prior to AIX 7200-04-00 went eosps 9/30/2021
 - Latest is 7200-05-02 (7200-05-03 coming 4Q)
- Use FLRTLITE to check supported levels of AIX, VIO, PowerHA, HMC, etc
 - <https://www14.software.ibm.com/webapp/set2/flrt/liteHome>

6

6

2. Install Hiper and other fixes



7

7

Updating Java, SSH and SSL

- SSH and SSL are obtained from the Web Download Pack which has moved to:
 - <https://www-01.ibm.com/marketing/iwm/iwm/web/pickUrxNew.do?source=aixbp>
- Java is downloaded from Fix Central
- Untar the files and put all ssh, ssl and java files (Java7 and java8) into a directory. I used /usr/local/soft/javasshssl
- Take a backup and use smitty to install or nimadm or NIM
- #lspp -l | grep Java8
- Make sure Java8.sdk and Java8_64.sdk are on
- As of 8/15/2021
 - Latest SSL is openssl-1.0.2.2102.tar.Z
 - Latest SSH is OpenSSH_8.1.102.2103.tar.Z
 - Latest Java8 is 8.0.6.35 – shows as 8.0.0.635
 - Latest Java7 is 7.0.10.90 – shows as 7.0.0.690
 - Latest Java6 is 6.0.0.655 – shows as 6.0.0.655



8

8

Efixes and ifixes

Many security patches are put on using efixes or ifixes

AIX and VIO servers need these to be applied – **use FLRTVC to determine what fixes are needed**

Run flrtvc and download and install the ifixes that are needed

<https://www-304.ibm.com/webapp/set2/sas/f/flrt/flrtvc.html>

You will need to read the .asc file to determine which efixes you need for the level you are running

Efixes & ifixes are downloaded using FTP or http. Links are provided by FLRTVC

`/usr/sbin/emgr -l` lists them

`emgr -P` lists the patches and the packages they affect

To apply a fix change into the directory it is in and then run it in preview mode:

`cd /usr/local/soft/fixes/bind_fix17`

`emgr -p -e IJ25927s2a.200708.epkg.Z`

Remove the `-p` and run again for real if the preview was successful:

`emgr -e IJ25927s2a.200708.epkg.Z`

If you run `emgr -l` and there are no fixes listed then you most likely have security holes that need patching, specifically Java, openssh and openssl.

9

9

Know how to get help

- If your upgrade goes south
- Know how to get a perfpmr
 - <https://www.ibm.com/support/pages/readme-collecting-perfpmr>
 - Keep an up to date copy downloaded so you don't have to download it in a panic
- Know how to run snap
 - <https://www.ibm.com/support/pages/how-collect-snap-powervm-virtual-io-server-vios>
 - <https://www.ibm.com/support/pages/node/670105>
 - Uploading snap to Testcase
 - <https://www.ibm.com/support/pages/node/681391>
- Opening a support ticket
 - <https://www.ibm.com/mysupport/s/my-cases>
- Have your VIO server and AIX LPARs predocumented as you will have to provide lots of information

10

10

3. Keep it simple and consistent



11

11

General

- Keep it simple
- Ensure LMB is the same on all servers if you want to use LPM
- Use hot pluggable adapters rather than built in ones
 - Easier maintenance
- ***All adapters should be desired, not required***
- Don't mix multipath drivers on HBAs
- Run HMC Scanner and/or Sysplan before and after all changes
- Plan for at least one update per year (IBM normally puts out 2)

12

12

General

- Mirror rootvg and page space if on internal disk
- Have a spare disk in your LPAR to use for cloning prior to updates
- NOTE – AIX requires at LEAST 30GB in rootvg – give it 100GB to 150GB
- Keep your rootvg very clean
 - If you need user or application filesystems put them outside of rootvg
- Check page spaces
 - Make sure all page spaces are the same size and on different LUNs
- Add logging and set up dump devices properly
- Check errpt regularly
 - NEVER run at 100% entitlement – ensure it is high enough and there are plenty of VPs and memory
- Backup regularly – use NIM or scripts
 - Backups should include mksysb images that can be used for bare metal restores, not just application data
 - Mksysbs should be taken for all AIX LPARs and VIO servers

13

13

4. Document well



14

14

Documentation is Critical

- Use HMCScanner and Sysplan
- Put together spreadsheet of documentation
- All equipment and serial numbers
- UAK expiration dates on servers (and AIX in the future)
- Customer numbers
- Server and I/O firmware levels, VIO and O/S levels
- HMC information including version, BMC and PNOR, networking etc
- IP addresses
- Resource profiles
- Adapter allocations
- Standards used for network, vSCSI, NPIV mappings
- Actual vSCSI assignments
- Actual NPIV vfcmaps
- Vfchosts and their associated WWPNs
- SEA and virtual ethernet VLAN assignments
- Switchports for SAN and network
- Power needs and PDU mapping
- Anything else you can think of

15

15

5. Backup regularly



16

16

Use alt_disk_copy prior to changes

- Have two disks so you can take a clone
- If rootvg is mirrored you will need to unmirror for maintenance
- No need to mirror if on SAN, always mirror rootvg if on internal disk

```
# lspv | grep root
hdisk0      00ce48c008314b9f      rootvg      active
hdisk1      00ce48c03c8f2115      altinst_rootvg
```

```
# bootinfo -b
hdisk0
```

```
exportvg altinst_rootvg
alt_disk_copy -V -B -d hdisk1
```

I always do a bosboot and rewrite the bootlist before any reboot
Recovery if issues with the upgrade is to point the bootlist to the new disk and reboot

If you want to use FBO (file backed optical) add a 3rd disk in its own VG so rootvg does not get huge

Keep rootvg small and clean

Don't forget mksysb images

17

17

6. Have a Performance Plan



18

18

What are you measuring?

- *Response time* is the elapsed time between when a request is submitted and when the response from that request is returned.
 - Amount of time for a database query
 - Amount of time it takes to echo characters to the terminal
 - Amount of time it takes to access a Web page
 - How much time does my user wait?
- *Throughput* is a measure of the amount of work that can be accomplished over some unit of time.
 - Database transactions per minute
 - File transfer speed in KBs per second
 - File Read or Write KBs per second
 - Web server hits per minute
- Make sure you know the difference and what is important to you

19

19

What makes it go slow? - 1/2

- Obvious:-
 - Not enough CPU
 - Not enough memory
 - Not enough disk bandwidth
 - Number of adapters
 - Queue depth and adapter queues
 - Number of disks
 - Not enough network bandwidth
 - Hardware errors - errpt
 - Software errors - errpt
 - Insufficient CPU entitlement
 - Affects CPU performance
 - Affects network performance for SEA
 - Too many or too few VPs
 - Overallocating CPU pools

20

20

What makes it go slow? – 2/2

- Not so obvious:-

- AIX tuning
- Oracle/DB2 parameters log place, SGA, Buffers
- Read vs write characteristics
- Adapter placement, overloading bus speeds
- Throttling effects – e.g., single-thread dependency
- Application errors
- Background processes (backups, batch processing) running during peak online times?
- Concurrent access to the same files
- Changes in shared resources
- Network Buffers
- Backlevel server or I/O firmware
- No cores left to grow beyond entitlement
- Missing AIX or VIO patches

21

21

Before you start - take a deep breath!



Image courtesy of teefury.com

22

22

Have a Plan

What do you hope to accomplish?

1. Describe the problem.
2. Measure where you're at (baseline).
 1. Perfpmr, your own scripts, nmon, hmcscanner, etc
3. Recreate the problem while getting diagnostic data (perfpmr, your own scripts, etc.).
4. Analyze the data.
5. Document potential changes and their expected impact, then group and prioritize them.
 1. Remember that one small change that only you know about can cause significant problems so document ALL changes
6. Make the changes.
 1. Group changes that go together if it makes sense to do so but don't go crazy
7. Measure the results and analyze if they had the expected impact; if not, then why not?
8. Is the problem still the same? If not, return to step 1.
9. If it's the same, return to step 3.

This may look like common sense but in an emergency that is the first thing to go out the window

Also, find a quiet place to work so you can focus. If you are trying to work on a critical problem have someone in the team who is responsible to report back on the status so you can concentrate on the issue.

Tuning is iterative – you may not get it right the first time

23

23

Basic Checklist for Performance

- Patching
- Ensure your firmware is current.
- Follow the memory plug-in rules.
- Lots of DIMMs
- Ensure OS level is current.
- Evaluate the use of SMT8.
- Right-size your shared LPARs. Use rPerfs for sizing
- VP:E ratio
- Java JDKs
- New compilers
- Recompile code
- DO NOT USE AIX RESTRICTED TUNABLES
- Throughput mode
- Size your VIO servers correctly
- Tune your virtual ethernets

24

24

Avoiding Problems

- Stay current
- Known memory issues with 6.1 tl9 sp1 and 7.1 tl3 sp1
- Java 7.1 SR1 or higher is the preferred Java for POWER7 and POWER8
- Java 6 SR7 is minimal on POWER7 but you should go to at least Java 7
- WAS 8.5.2.2 or higher
- Refer to Section 8.3 of the Performance Optimization and Tuning Techniques Redbook SG24-8171
- HMC v9.1M940SP1 is required for POWER9 S922
 - For FW950 you need 9.1M950
 - – does not support servers prior to POWER7 (specific models and FW on P7)
 - Need to go to v9 anyway as v8 no longer supported
- Remember not all workloads run well in the shared processor pool – some are better dedicated
 - Apps with polling behavior, CPU intensive apps (SAS, HPC), latency sensitive apps (think trading systems)

25

25

7. Provide plenty of Resources



26

26

Resources

- Entitlement
- VPs
- Memory
- HBA queues
 - max_transfer, num_cmd_elems
- HDISK settings
 - max_transfer, queue_depth, reserve, algorithm
- Network
 - Virtual Buffers
 - TCP and UDP send and receive space



27

27

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
- Do not starve your VIO servers or client LPARs!
- Set entitlement and VPs correctly
- Be cautious of sizing studies – they tend to undersize memory and sometimes cores and usually do not include the VIO server needs
- Consider whether the workload will play well with shared processors
- Never underestimate the power of common sense
- **Scale by rPerf (or other benchmark data) NOT by ghz when comparing boxes**
- **Do not size the box by the sum of the entitlements. Leave room for LPARs to shrink and grow**

28

28

CPU Sizing

When sizing the server:

Do not sum the entitlements and size the server based on entitlement

The shared processor pool is supposed to be used to share processors and you need room to grow and shrink

Sum the peak concurrent VPs needed and add some for growth and then size the server

i.e.

LPAR	Entitlement	Max at noon	Max at 6pm
A	.5	1.5	2
B	.8	1	.8
C	6	10	12
D	3	8	2
TOTAL	10.3	20.5	16.8

This system needs 21 cores not 11

29

29

Sample /etc/tunables/rc-tune.sh for AIX and VIO

```
#!/bin/ksh
# First we set the network tuneables
#
/usr/sbin/no -p -o rfc1323=1
/usr/sbin/no -p -o tcp_sendspace=262144
/usr/sbin/no -p -o tcp_recvspace=262144
/usr/sbin/no -p -o udp_sendspace=65536
/usr/sbin/no -p -o udp_recvspace=655360
#
vmo -p -o minfree=1024          we calculate this
vmo -p -o maxfree=2048         we calculate this

#Run ifconfig -a and check the en values - (assuming IP is on en5):
#chdev -l en5 -a tcp_recvspace=262144 -P
#chdev -l en5 -a tcp_sendspace=262144 -P
#chdev -l en5 -a rfc1323=1 -P
#
chdev -l fcs0 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
chdev -l fcs1 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
chdev -l fcs2 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
chdev -l fcs3 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P

num_cmd_elems can be 1024, 2048, 4096 if supported by disk system and VIO is set at least that high
NOTE above tunables cannot be larger than settings in the VIO servers
```

30

30

Sample /etc/tunables/rc-tunebufs.sh

This tunes buffer settings for the two virtual adapters – assumes ent4, ent5 are virtuals

lsdev -C | grep ent will show the adapters so you can pick the right ones

```
#!/bin/ksh
#
chdev -l ent4 -a buf_mode=min -P
chdev -l ent5 -a buf_mode=min -P
chdev -l ent4 -a max_buf_tiny=4096 -P
chdev -l ent4 -a max_buf_small=4096 -P
chdev -l ent4 -a max_buf_medium=512 -P
chdev -l ent5 -a max_buf_tiny=4096 -P
chdev -l ent5 -a max_buf_small=4096 -P
chdev -l ent5 -a max_buf_medium=512 -P
```

31

31

HBA Tuning

- Make the same tuning changes you would make on AIX, but VIO must be set at least as high as clients
- Set num_cmd_elems and max_xfer_size on the fiber adapters on VIO
 - chdev -l fcs0 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
 - chdev -l fcs1 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
- Check these numbers are supported by your disk vendor
- If NPIV also set on clients
- **Client setting cannot be higher than the VIOs**
- **VIO must be rebooted to at least the client value prior to client change.**
- Pay attention to adapter layout and priorities
- NOTE – as of AIX v7.1 tl2 (or 6.1 tl8) num_cmd_elems is limited to 256 on the VFCs so set num_cmd_elems to the high number on the VIO but to no more than 256 on the NPIV clients
- See: <http://www-01.ibm.com/support/docview.wss?uid=isg1IV63282>
- **Increased again to 2048 in July 2016**
- <http://www-01.ibm.com/support/docview.wss?uid=isg1IV76270>
- This upper limit is set in the client LPAR not the VIO server
- BUT the client setting MUST NOT be larger than what is set in the VIO server

32

32

HBA max_xfer_size

The default is
0x100000 /* Default io_dma of 16MB */

After that, 0x200000,0x400000,0x800000 gets you 128MB

After that 0x1000000 checks for bus type, and you may get 256MB, or 128MB

There are also some adapters that support very large max_xfer sizes which can possibly allocate 512MB

VFC adapters inherit this from the physical adapter (generally)

Unless you are driving really large IO's, then max_xfer_size on the HBA is rarely changed beyond 0x200000 which provides a 128MB DMA

Client setting cannot be higher than the VIOs were booted with

33

33

7a. CPU



34

34

Entitlement and VPs

- Utilization calculation for CPU is different between POWER5, 6, 7, 8 and POWER9
- VPs are also unfolded sooner (at lower utilization levels than on P6 and P5)
- May also see high VCSW in lparstat
- This means that starting with POWER7 you need to pay more attention to VPs
 - You may see more cores activated at lower utilization levels
 - But you will see higher idle
 - If only primary SMT threads in use then you probably have excess VPs
- Try to avoid this issue by:
 - Reducing VP counts
 - Use realistic entitlement to VP ratios
 - 10x or 20x is not a good idea
 - Try setting entitlement to .6 or .7 of VPs
 - Per Nigel**
 - For E=0.05 to 0.90 use VPs=1**
 - For E=1 to 4 round up to next whole integer**
 - For E=4 to 8 round up 1 or 2 VPs**
 - Ensure workloads never run consistently above 100% entitlement
 - Too little entitlement means too many VPs will be contending for the cores
 - NOTE – VIO server entitlement is critical – SEAs scale by entitlement not VPs**
 - This applies to all LPARs using the SEA as well**
- All VPs have to be dispatched before one can be redispached
- Performance may (in most cases, will) degrade when the number of Virtual Processors in an LPAR exceeds the number of physical processors**
- The same applies with VPs in a shared pool LPAR – these should not exceed the cores in the pool**

35

35

Scaling VPs after Upgrades

- VPs and entitlements should be reviewed when upgrading servers
- rPerf per core increases as does threading depending on architecture
- Example
- | Server | Cores | smt4 rperf | rperf/core | Increase over p740 |
|----------|-------------|------------|------------|--------------------|
| p740 p7+ | 16 (3.6g) | 197.70 | 12.36 | |
| S822 p8 | 16 (4.1g) | 302.40 | 18.90 | 1.53 |
| S922 p9 | 16 (3.9max) | 313.10 | 19.57 | 1.58 |
- | Server | Cores | smt8 rperf | rperf/core | Increase over S822 SMT4 |
|---------|-------------|------------|------------|-------------------------|
| S822 p8 | 16 (4.1g) | 323.60 | 20.23 | 1.07 |
| S922 p9 | 16 (3.9max) | 394.50 | 24.66 | 1.31 |

The above are only approximations using published rPerf

For P8 and P9 SMT8 gives an additional boost for threaded workloads

The S922 above in SMT8 is 394.5 rPerf (about 1.26 of the S922 SMT4 number)

The S822 is 323.60 (about 1.27 of the S822 SMT4 number)

rPerf is not a guarantee – it is used to give an idea of relative performance between machines

The point to note:

If you are using 16 cores (or VPs) on the p740 or the S822 then you do not need 16 cores (or VPs) on your new S922 unless you have enormous growth or latent demand.

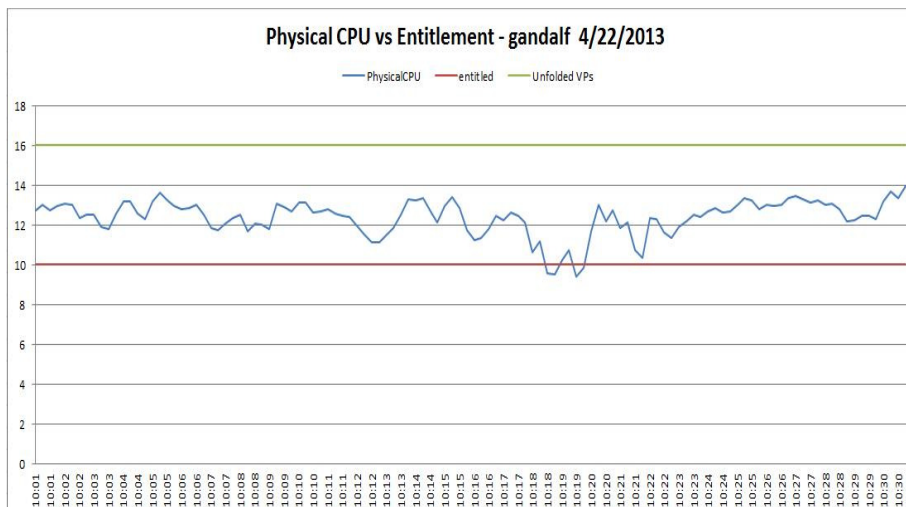
Resize VP assignments LPAR by LPAR, based on rperf needed per LPAR

Too many cores or VPs wastes dispatcher and hypervisor cycles

36

36

Entitlement and vps from lpar tab in nmon



LPAR always above entitlement – increase entitlement
Affects its CPU performance but also its network

37

37

Uncapped vs Capped LPARs

- 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 desired 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 – I often use 255
 - Prod 160
 - Test/Dev 128
- Have a plan, not necessarily this one – document it well
- You can cap an LPAR by:
 - Setting LPAR as capped
 - Set LPAR weight to 0
 - Setting desired CPU (entitlement) to the same as desired VPs
 - Max VPs and max CPU have nothing to do with capping
- Capping and pools are used to ensure licenses are adhered to

38

38

Applications and SPLPARs

- Applications do not need to be aware of Micro-Partitioning
- Not all applications benefit from SPLPARs
- Applications that may not benefit from Micro-Partitioning:
 - Applications with a strong response time requirements for transactions may find Micro-Partitioning detrimental:
 - Because virtual processors can be dispatched at various times during a timeslice
 - May result in longer response time with too many virtual processors:
 - Each virtual processor with a small entitled capacity is in effect a slower CPU
 - Compensate with more entitled capacity (2-5% PUs over plan)
 - Applications with polling behavior
 - CPU intensive application examples: DSS, HPC, SAS
- Applications that are good candidates for Micro-Partitioning:
 - Ones with low average CPU utilization, with high peaks:
 - Examples: OLTP, web applications, mail server, directory servers
- In general Oracle databases are fine in the shared processor pool
- For licensing reasons you may want to use a separate pool for databases

39

39

Throughput mode in POWER7 and POWER8

- P7 and higher with AIX v6.1 TL08 and AIX v7.1 TL02
- Dispatches more SMT threads to a VP core before unfolding additional VPs
- Tries to make it behave a bit more like P6
- For POWER7 & POWER8, the default dispatch algorithm is known as Raw Throughput Mode. AIX will dispatch Virtual Processors once a utilization threshold of 49% has been exceeded. Once all available first SMT threads of all VPs are executing, it then wraps around and uses the second SMT thread for each VP.
- **Raw** provides the highest per-thread throughput and best response times at the expense of activating more physical core
- **Scaled** provides the highest core throughput at the expense of per-thread response times and throughput. It also provides the highest system-wide throughput per VP because tertiary thread capacity is "not left on the table."
- **schedo -p -o vpm_throughput_mode=**
 - 0 Legacy Raw mode (default)
 - 1 "Enhanced Raw" mode with a higher threshold than legacy
 - 2 Scaled mode, use primary and secondary SMT threads
 - 4 Scaled mode, use all four SMT threads
 - 8 Scaled mode, use eight SMT threads (POWER8, AIX v7.1 required)
- Dynamic Tunable
- SMT unfriendly workloads could see an enormous per thread performance degradation

40

40

40

Throughput mode in POWER9

From Steve Nasypany Session at pTechu October 2018

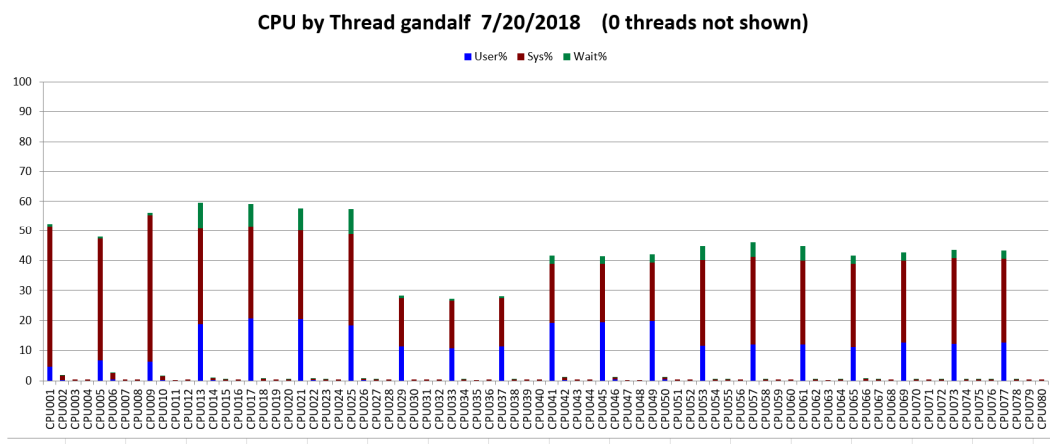
The VP code is aware of the core architecture and will place/collapse smaller workloads slightly more aggressively when workloads are present

- Optimization has the additional impact of reducing physical consumption
- Single thread utilization is calibrated to ~32% in SMT8 (~44% in SMT4), so below the default VP dispatch threshold of ~50% per core
- Because single-threads are calibrated lower and equivalent workloads will overall generate a lower utilization, they are more likely to fall below dispatch threshold, thus lowering physical consumed
- Linux not running on PowerVM dispatches to SMT4 cores and does not use hardware calibration

41

41

nmon SMT4 Threads



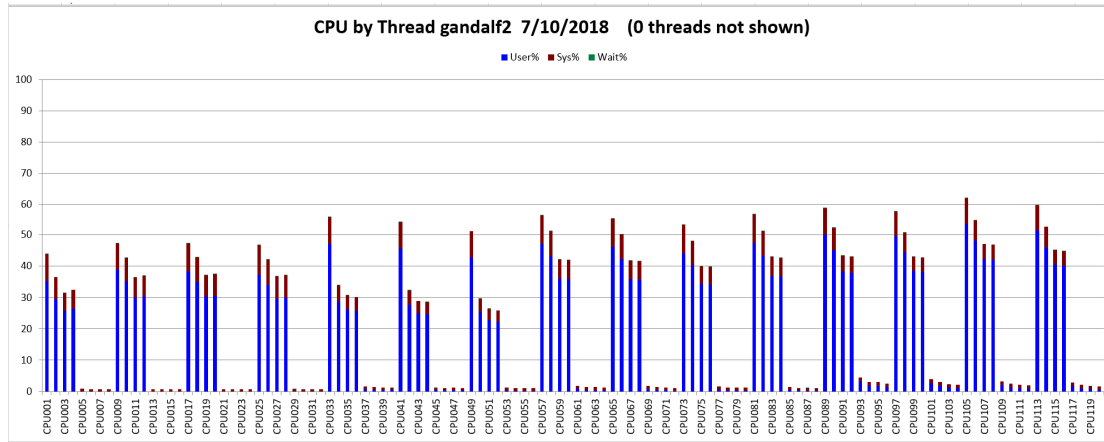
Note mostly primary thread used and some secondary – we could possibly reduce cores/VPs

Considering reducing cores to 16 or 18 from 20 (SMT4)

42

42

nmon SMT8 Threads Requires POWER8 Mode



Planning to keep cores at this level for now

43

43

lparstat dedicated – bbbl tab in nmon

CPU in sys	24
Virtual CPU	20
Logical CPU	80
smt threads	4
capped	1
min Virtual	1
max Virtual	20
min Logical	1
max Logical	160
min Capacity	1
max Capacity	20
Entitled Capacity	20
min Memory MB	262144
max Memory MB	786432
online Memory	524288

NOTE – even though the cores are dedicated it still refers to them as VPs.

So we have 20 physical cores at SMT4 which means 80 LCPUs

44

44

lparstat – bbbl tab in nmon

lparno	3
lparname	gandalf
CPU in sys	24
Virtual CPU	16
Logical CPU	64
smt threads	4
capped	0
min Virtual	8
max Virtual	20
min Logical	8
max Logical	80
min Capacity	8
max Capacity	16
Entitled Capacity	10
min Memory MB	131072
max Memory MB	327680
online Memory	303104
Pool CPU	16
Weight	150
pool id	2

Compare VPs to poolsize
LPAR should not have more VPs than the poolsize

LPAE is in SPP so Virtual CPU is maximum size LPAR can grow to from entitlement
Ent=10, VCPU=16, 16 cores in pool

45

45

lparstat 30 2 SPP

lparstat 30 2 output

System configuration: type=Shared mode=Uncapped smt=4 lcpu=72 mem=319488MB psize=17 ent=12.00

%user	%sys	%wait	%idle	physc	%entc	lbusy	app	vcs	phint
46.8	11.6	0.5	41.1	11.01	91.8	16.3	4.80	28646	738
48.8	10.8	0.4	40.0	11.08	92.3	16.9	4.88	26484	763

lcpu=72 and smt=4 means I have 72/4=18 VPs but pool is only 17 cores – BAD

Ent=12 means I am entitled to 12 cores – don't know VPs so don't know how big this LPAR can grow

psize = processors in shared pool

lbusy = %occupation of the LCPUs at the system and user level

app = Available physical processors in the pool

vcs = Virtual context switches (virtual processor preemptions)

phint = phantom interrupts received by the LPAR

interrupts targeted to another partition that shares the same physical processor
i.e. LPAR does an I/O so cedes the core, when I/O completes the interrupt is sent to the core but different LPAR running so it says "not for me"

NOTE – Must set "Allow performance information collection" on the LPARs to see good values for app, etc
Required for shared pool monitoring

46

46

lparstat 30 2 Dedicated

lparstat 30 2 output

System configuration: type=Dedicated mode=Capped smt=4 lcpu=80 mem=524288MB

%user	%sys	%wait	%idle
16.8	28.7	6.4	48.1
17.0	29.3	5.8	48.0

lcpu=80 and smt=4 means I have 80/4=20 real cores

lbusy = %occupation of the LCPUs at the system and user level

lparstat -h 30 2 output

System configuration: type=Dedicated mode=Capped smt=4 lcpu=80 mem=524288MB

%user	%sys	%wait	%idle	%hypv	hcalls
16.8	29.8	5.4	48.0	61.3	2222545
16.8	30.1	5.1	48.0	61.1	2258600

%hypv

Indicates the percentage of physical processor consumption spent making hypervisor calls.

hcalls

Indicates the average number of hypervisor calls that were started.

47

47

lparstat & mpstat -s POWER8 Mode Example

b814aix1: lparstat 30 2

System configuration: type=Shared mode=Uncapped smt=8 lcpu=48 mem=32768MB psize=2 ent=0.50

%user	%sys	%wait	%idle	physc	%entc	lbusy	app	vcswh	phint
0.0	0.1	0.0	99.9	0.00	0.8	2.3	1.96	244	0
0.0	0.2	0.0	99.8	0.00	1.0	2.3	1.96	257	0

b814aix1: mpstat -s

System configuration: lcpu=48 ent=0.5 mode=Uncapped

Proc0 0.00%								Proc8 0.00%							
cpu0	cpu1	cpu2	cpu3	cpu4	cpu5	cpu6	cpu7	cpu8	cpu9	cpu10	cpu11	cpu12	cpu13	cpu14	cpu15
0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Proc16 0.00%								Proc24 0.00%							
cpu16	cpu17	cpu18	cpu19	cpu20	cpu21	cpu22	cpu23	cpu24	cpu25	cpu26	cpu27	cpu28	cpu29	cpu30	cpu31
0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Proc32 0.00%								Proc40 0.00%							
cpu32	cpu33	cpu34	cpu35	cpu36	cpu37	cpu38	cpu39	cpu40	cpu41	cpu42	cpu43	cpu44	cpu45	cpu46	cpu47
0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

48

48

topas -C

```

Topas CEC Monitor          Interval: 10          Thu Feb 27 08:53:05 2014
Partitions Memory (GB)      Processors
Shr: 5   Mon:86.0 InUse:23.0 Shr: 8   PSz: 16   Don: 0.0 Shr_PhysB 0.02
Ded: 0   Avl:  -       Ded: 0   APP: 16.0 Stl: 0.0 Ded_PhysB 0.00

Host      OS  Mod Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI  pmem
-----shared-----
b740nl1   A71 Ued 32 5.3 32   0 0 0 99  0.03 210 4.00 0.7 0  -
b740vio2  A61 U-d 3.0 2.8 8    0 0 0 99  0.00 256 0.50 0.8 0  -
b740ft2   A71 Ued 32 5.3 4    0 0 0 99  0.00 191 1.00 0.4 0  -
          A61 U-d 3.0 2.8 4    0 0 0 99  0.00 171 0.50 0.6 0  -
b740l1    A71 U-d 16 7.1 16   0 0 0 99  0.00 212 2.00 0.1 0  -

Host      OS  Mod Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw %istl %bstl
-----dedicated-----

```

Shows pool size of 16 with all 16 available
Monitor VCSW as potential sign of insufficient entitlement

51

51

nmon -p

```

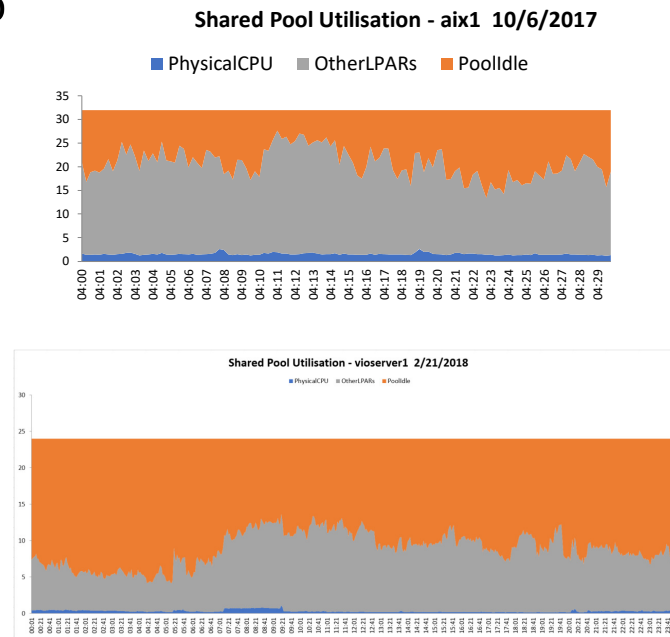
1:topas nmonqc=many-CPUsgggggggHost=b740ft2ggggggRefresh=2 secsgg09:09.31gggggggggggggggggggggggg
x Shared-CPU-Logical-Partition ggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggg
xPartition:Number=14 "b740ft2"
xFlags:LPAred DRable SMT Shared UnCapped PoolAuth Migratable Not-Donating AMSable.
xSummary: Entitled=1.00 Used 0.00 ( 0.5%) 0.0% of CPUs in System
x          PoolCPUs=16 Unused 15.98 0.0% of CPUs in Pool
xCPU-Stats-----Capacity-----ID-Memory-----
xmax Phys in sys 16 Cap. Processor Min 0.10 SPLPAR Group:Pool 32782:0
xPhys CPU in sys 16 Cap. Processor Max 4.00 Memory(MB) Min:Max 1024:65536
xVirtual Online 1 Cap. Increment 0.01 Memory(MB) Online 32768
xLogical Online 4 Cap. Unallocated 0.00 Memory Region LMB 256MB min
xPhysical pool 16 Cap. Entitled 1.00 Time-----Seconds
xSMT threads/CPU 4 -MinReqVirtualCPU 0.10 Time Dispatch Wheel 0.0100
xCPU-----Min-Max Weight-----MaxDispatch Latency 0.0000
xVirtual 1 4 Weight Variable 128 Time Pool Idle 15.9761
xLogical 1 16 Weight Unallocated 0 Time Total Dispatch 0.0046
x-----
xEvent= 0 --- --- SerialNo Old=--- Current=F6934C When=---
x-----
xShared Pools MaxPoolCapacity=16.00 MyPoolMax =16.00 SharedCPU-Total=16.00
xSharedCPU=16 EntPoolCapacity=15.20 MyPoolBusy= 0.02 SharedCPU-Busy = 0.02
xgggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggggg
x
v

```

52

52

nmon Analyser LPAR Tab



53

53

emstat

```
emstat -a 30 2
```

Alignment	Alignment	Emulation	Emulation
SinceBoot	Delta	SinceBoot	Delta
0	0	0	0
0	0	0	0

Check for emulations and alignments

Caused when old code run on new architectures or operating systems

Code should be recompiled whenever possible when upgrades occur

Binary compatibility statement means it will work but the best performance is obtained when you run a version compiled for the architecture

54

54

7b. MEMORY



55

55

Memory Types

- Kernel
 - This is AIX itself and the minimum is around 2GB.
- Persistent
 - Backed by filesystems (called file cache)
 - Raw devices, GPFS, Oracle ASM does not make use of file cache.
- Working storage
 - Dynamic
 - Includes executables and their work areas
 - Backed by page space
 - Shows as avm in a vmstat -l (multiply by 4096 to get bytes instead of pages) or as %comp in nmon analyser or as a percentage of memory used for computational pages in vmstat -v
 - ALSO NOTE – if %comp is near or >97% then you will be paging and need more memory
- Prefer to steal from persistent as it is cheap
- minperm, maxperm, maxclient, lru_file_repage and page_steal_method all impact these decisions

56

56

Plan for Firmware Overhead for adapters

You can use the System Planning Tool or your own tool – I use the spreadsheet below

<http://www.circle4.com/ptechu/memoryplan.xlsx>

Note div - use 64 for all pre p7+ and IBM I, – 128 for p7+ and p8

Each active port on an 8Gb card or higher requires 512MB in the VIO or LPAR

Each VFC (NPIV vfchost) requires 140MB in the VIO

Example:

20 ports per VIO without NPIV would be $20 * 512 = 10GB$ plus VIOS base for each VIOS

if NPIV then we allocate per client so if there are 20 clients on each VIO then each VIO needs:

$20 * 140 = 2800MB$ or 2.8GB above the base

Don't underestimate memory overhead

57

57

Server Memory

- Reserved Memory is based on max memory for an LPAR, not on desired
- This is because memory gets reserved for HPTs (hypervisor page tables)

1	Name	Mode	Min GB	Curr GB	Max GB	ExpFact	AMS->	Weight	Prim VIOS	Sec VIOS	Curr VIOS
2	lpar1	ded	2.25	5.00	100.00						
3	lpar2	ded	2.00	15.75	100.00						
4	lpar3	ded	2.00	15.75	100.00						
5	lpar4	ded	10.00	64.00	100.00						
6	vios2	ded	2.00	6.00	100.00						
7	vios1	ded	2.00	6.00	100.00						
8											

Ent_Sys_Pools OnOff CoD CoD Events LPAR_Summary LPAR_Profiles LPAR_CPU **LPAR_MEM** Physical_Slo

58

58

Server Memory

Status	Type Model	Serial	GHz	CPU Type	Tot Cores	Act Cores	Deco nf Cores	Curr Avail Cores	Pend Avail Cores	Ded Cores	Pool Size	Virt Procs	#LPAR	Tot GB	Act GB	Deconf GB	Firm GB	Curr Avail GB	Pend Avail GB
Operating	8286-41A	xxxxxxx	3.72	PowerPC_POWER8	8	5	3	0.00	0.00	0	2	4	6	128.00	128.00	0.00	13.75	1.75	1.75

System Summary Ent_Sys_Pools OnOff CoD CoD Events LPAR_Summary LPAR_Profiles LPAR_CPU ...

CPU Type	Tot Cores	Act Cores	Deco nf Cores	Curr Avail Cores	Pend Avail Cores	Ded Cores	Pool Size	Virt Procs	#LPAR	Tot GB	Act GB	Deconf GB	Firm GB	Curr Avail GB	Pend Avail GB
PowerPC_POWER8	80	26	0	21.64	21.64	0	26	22	28	3,072.00	1,536.00	0.00	23.00	1,447.00	1,447.00
PowerPC_POWER8	80	26	0	0.31	0.31	0	26	52	39	3,072.00	1,536.00	0.00	38.50	967.50	967.50

Look at Firm GB in HMCScanner under System Summary Tab

Latest is 0.11.42

<https://www.ibm.com/support/pages/hmc-scanner-power-server-config-and-performance-stats>

[https://www.ibm.com/support/pages/sites/default/files/inline-files/\\$FILE/hmcScanner-0.11.42.zip](https://www.ibm.com/support/pages/sites/default/files/inline-files/$FILE/hmcScanner-0.11.42.zip)

59

59

Correcting Paging

From vmstat -v

11173706 paging space I/Os blocked with no psbuf

Isps output on above system that was paging before changes were made to tunables

Isps -a

Page Space	Physical Volume	Volume Group	Size	%Used	Active	Auto	Type
paging01	hdisk3	pagingvg	16384MB	25	yes	yes	lv
paging00	hdisk2	pagingvg	16384MB	25	yes	yes	lv
hd6	hdisk0	rootvg	16384MB	25	yes	yes	lv

Isps -s

Total Paging Space	Percent Used
49152MB	1%

Can also use vmstat -l and vmstat -s

Should be balanced – NOTE VIO Server comes with 2 different sized page datasets on one hdisk.

Best Practice

More than one page volume

All the same size including hd6

Page spaces must be on different disks to each other

Do not put on hot disks

When you mirror rootvg do not unmirror hd6

Mirror all page spaces that are on internal or non-raided disk

If you can't make hd6 as big as the others then swap it off after boot

All real paging is bad

60

60

svmon Memory Breakdown

```
# svmon -G -O unit=auto -i 2 2
```

Unit: auto

	size	inuse	free	pin	virtual	available	mmode
memory	16.0G	8.26G	7.74G	5.50G	10.3G	7.74G	Ded
pg space	12.0G	2.43G					
	work	pers	clnt	other			
pin	5.01G	0K	4.11M	497.44M			
in use	8.06G	0K	202.29M				

Unit: auto

	size	inuse	free	pin	virtual	available	mmode
memory	16.0G	8.26G	7.74G	5.50G	10.3G	7.74G	Ded
pg space	12.0G	2.43G					
	work	pers	clnt	other			
pin	5.01G	0K	4.11M	497.44M			
in use	8.06G	0K	202.29M				

Keep an eye on memory breakdown especially pinned memory. High values can mean someone has pinned something

61

61

Looking for Problems

- lssrad -av
- mpstat -d
- topas -M
- svmon
 - Try -G -O unit=auto,timestamp=on,pgsz=on,affinity=detail options
 - Look at Domain affinity section of the report
- Etc etc

62

62

Memory Problems

- Look at computational memory use
 - Shows as avm in a vmstat -l (multiply by 4096 to get bytes instead of pages)
- System configuration: lcpu=48 mem=32768MB ent=0.50
- | r | b | p | w | avm | fre | fi | fo | pi | po | fr | sr | in | sy | cs | us | sy | id | wa | pc | ec |
|---|---|---|---|--------|---------|----|----|----|----|----|----|-----|-----|----|----|----|----|------|-----|----|
| 0 | 0 | 0 | 0 | 807668 | 7546118 | 0 | 0 | 0 | 0 | 0 | 1 | 159 | 161 | 0 | 0 | 99 | 0 | 0.01 | 1.3 | |
- AVM above is about 3.08GB which is about 9% of the 32GB in the LPAR
 - Shows as %comp in nmon analyser
 - Shows as a percentage of memory used for computational pages in vmstat -v
- **NOTE – if %comp is near or >97% then you will be paging and need more memory**
- Try svmon -P -Osortseg=pgsp -Ounit=MB | more
 - This shows processes using the most pagespace in MB
 - You can also try the following:
 - svmon -P -Ofiltercat=exclusive -Ofiltertype=working -Ounit=MB | more

63

63

Affinity

- LOCAL SRAD, within the same chip, shows as s3
- NEAR SRAD, within the same node – intra-node, shows as s4
- FAR SRAD, on another node – inter-node, shows as s5
- Command is lssrad -av or can look at mpstat -d
- Topas M option shows them as Localdisp%, Neardisp%, Fardisp%
- The further the distance the longer the latency
- Problems you may see
 - SRAD has CPUs but no memory or vice-versa
 - CPU or memory unbalanced
- Note – on single node systems far dispatches are not as concerning
- To correct look at new firmware, entitlements and LPAR memory sizing
- Can also look at Dynamic Platform Optimizer (DPO)

64

64

Memory Tips

Avoid having chips without DIMMs.

Attempt to fill every chip's DIMM slots, activating as needed.

Hypervisor tends to avoid activating cores without "local" memory.

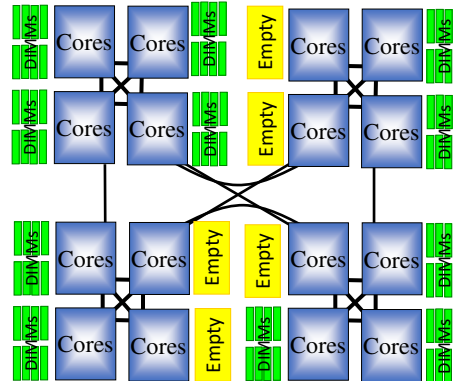


Diagram courtesy of IBM

65

65

E850 mpstat -d

mpstat -dw 15 2 output

System configuration: lcpu=80 mode=Capped

cpu	cs	ics	bound	rq	push	S3pull	S3grd	S0rd	S1rd	S2rd	S3rd	S4rd	<i>local near far</i>			
													S5rd	S3hrd	S4hrd	S5hrd
0	92	31	1	1	0	0	44	95.1	0.6	0.0	1.2	0.0	3.1	100.0	0.0	0.0
1	3	0	0	0	0	0	0	82.1	12.0	0.0	6.0	0.0	0.0	100.0	0.0	0.0
2	2	0	0	0	0	0	0	95.3	4.7	0.0	0.0	0.0	0.0	100.0	0.0	0.0
3	2	0	0	0	0	0	0	95.1	4.9	0.0	0.0	0.0	0.0	100.0	0.0	0.0
4	314	21	1	1	0	0	95	97.4	0.1	0.0	0.6	0.0	2.0	100.0	0.0	0.0
5	14	2	0	0	0	0	0	94.5	5.5	0.0	0.0	0.0	0.0	100.0	0.0	0.0
6	3	0	0	0	0	0	0	96.4	3.6	0.0	0.0	0.0	0.0	100.0	0.0	0.0
7	2	0	0	0	0	0	0	95.3	4.7	0.0	0.0	0.0	0.0	100.0	0.0	0.0
..... lots more lines here																
72	280	38	1	1	0	224	32	91.0	0.0	0.0	3.0	0.0	6.0	66.5	8.6	24.8
73	2	0	0	0	0	0	0	95.6	4.4	0.0	0.0	0.0	0.0	95.6	0.0	4.4
74	2	0	0	0	0	0	0	95.1	4.9	0.0	0.0	0.0	0.0	95.1	0.0	4.9
75	2	0	0	0	0	0	0	94.9	5.1	0.0	0.0	0.0	0.0	94.9	0.0	5.1
76	161	25	0	0	0	87	14	93.0	0.0	0.0	2.9	0.0	4.1	77.1	7.4	15.4
77	2	0	0	0	0	0	0	94.3	5.7	0.0	0.0	0.0	0.0	94.3	0.0	5.7
78	1	0	0	0	0	0	0	93.8	6.2	0.0	0.0	0.0	0.0	93.8	0.0	6.2
79	2	0	0	0	0	0	0	92.3	7.7	0.0	0.0	0.0	0.0	92.3	2.6	5.1
ALL	6771	996	13	13	0	1888	2170	93.8	0.3	0.0	2.0	0.0	3.9	89.5	3.5	6.9

The above is for a multi socket system (E850) so pay attention to the numbers under near and far
On a single socket system (S814) I would expect to see everything local (s3hrd)

66

66

E880 mpstat -d SMT8

System configuration: lcpu=120 mode=Capped

local near far

cpu	cs	ics	bound	rq	push	S3pull	S3grd	S0rd	S1rd	S2rd	S3rd	S4rd	S5rd	S3hrd	S4hrd	S5hrd	%nsp
0	6727	338	1	1	0	2100	1834	68.5	9.9	0.0	17.7	3.9	0.0	83.0	17.0	0.0	101
1	5766	265	0	0	0	1874	1634	67.7	10.7	0.0	17.6	4.0	0.0	81.8	18.2	0.0	101
2	5441	268	0	0	0	0	0	73.1	10.6	0.0	16.4	0.0	0.0	95.2	4.8	0.0	101
3	5667	292	0	0	0	0	0	73.6	10.1	0.0	16.3	0.0	0.0	95.1	4.9	0.0	101
4	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	88.9	11.1	0.0	101
5	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	101
6	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	101
7	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	71.4	28.6	0.0	101
8	5895	298	1	11	0	1974	1654	67.5	11.0	0.0	17.5	4.1	0.0	81.5	18.5	0.0	101
9	5758	291	0	0	0	1948	1600	67.7	10.7	0.0	17.5	4.1	0.0	81.9	18.1	0.0	101
10	5242	235	0	0	0	0	0	71.2	11.9	0.0	17.0	0.0	0.0	94.6	5.4	0.0	101
11	6319	332	1	9	0	0	0	74.1	10.0	0.0	15.9	0.0	0.0	94.9	5.1	0.0	101
12	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	101
13	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	71.4	28.6	0.0	101
14	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	101
15	0	0	0	0	0	0	0	46.2	53.8	0.0	0.0	0.0	0.0	76.9	23.1	0.0	101
118	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	66.7	33.3	0.0	101
119	0	0	0	0	0	0	0	0.0	100.0	0.0	0.0	0.0	0.0	50.0	50.0	0.0	101
ALL	265322	12290	43	112	0	60475	99151	72.1	10.0	0.0	13.9	4.0	0.0	86.7	13.3	0.0	0

67

67

vmstat -v Output

3.0 minperm percentage
90.0 maxperm percentage
45.1 numperm percentage
45.1 numclient percentage
90.0 maxclient percentage

1468217 pending disk I/Os blocked with no pbuf
11173706 paging space I/Os blocked with no psbuf
2048 file system I/Os blocked with no fsbuf
238 client file system I/Os blocked with no fsbuf
39943187 external pager file system I/Os blocked with no fsbuf

pbufs
pagespace
JFS
NFS/VxFS
JFS2

numclient=numperm so most likely the I/O being done is JFS2 or NFS or VxFS
Based on the blocked I/Os it is clearly a system using JFS2
It is also having paging problems (vmstat and lsps commands)
pbufs also need reviewing (lvmo command)

68

68

vmstat -v Output

```
memory pools      4
numperm           10.1%
numclient         10.1%
uptime            up 16 days,  4:46
```

17308 pending disk I/Os blocked with no pbuf
4538737 paging space I/Os blocked with no psbuf

1972 file system I/Os blocked with no fsbuf
 8724 client file system I/Os blocked with no fsbuf
 7077 external pager file system I/Os blocked with no fsbuf
 89.9 percentage of memory used for computational pages

pbufs
pagespace
 JFS
 NFS/VxFS
 JFS2

numclient=numperm so most likely the I/O being done is JFS2 or NFS or VxFS
 Based on the blocked I/Os it is clearly a system using JFS2
 It is also having paging problems (vmstat and lsps commands)
 pbufs also need reviewing (lvmo command)

Also note that computational was 89.9%

69

69

Memory Pools and fre column

- fre column in vmstat is a count of all the free pages across all the memory pools
- When you look at fre you need to divide by memory pools
- Then compare it to maxfree and minfree
- This will help you determine if you are happy, page stealing or thrashing
- You can see high values in fre but still be paging
- You have to divide the fre column by mempools
- In below if maxfree=2000 and we have 10 memory pools then we only have 990 pages free in each pool on average. With minfree=960 we are page stealing and close to thrashing.

kthr				memory		page				faults				cpu			
r	b	p	avm	fre	fi	fo	pi	po	fr	sr	in	sy	cs	us	sy	id	wa
70	309	0	8552080	9902	75497	9615	9	3	84455	239632	18455	280135	91317	42	37	0	20

Assuming 10 memory pools (you get this from vmstat -v)

9902/10 = 990.2 so we have 990 pages free per memory pool

If maxfree is 2000 and minfree is 960 then we are page stealing and very close to thrashing

70

70

Calculating minfree and maxfree

```
vmstat -v | grep memory
3 memory pools
```

```
vmo -a | grep free
maxfree = 1088
minfree = 960
```

Calculation is:

$\text{minfree} = (\max(960, (120 * \text{lcpu}) / \text{memory pools}))$

$\text{maxfree} = \text{minfree} + (\text{Max}(\text{maxpagehead}, \text{j2_maxPageReadahead}) * \text{lcpu}) / \text{memory pools}$

So if I have the following:

Memory pools = 3 (from vmo -a or kdb)

J2_maxPageReadahead = 128

CPUS = 6 and SMT on so lcpu = 12

So minfree = $(\max(960, (120 * 12) / 3)) = 1440 / 3 = 480$ or 960 whichever is larger

And maxfree = $\text{minfree} + (128 * 12) / 3 = 960 + 512 = 1472$

I would probably bump this to 1536 rather than using 1472 (nice power of 2)

The difference between minfree and maxfree should be no more than 1K per IBM

If you over allocate these values it is possible that you will see high values in the "fre" column of a vmstat and yet you will be paging.

71

71

nmon Monitoring

• nmon -ft -AOPV^dML -s 15 -c 120

- Grabs a 30 minute nmon snapshot
- A is async IO
- M is mempages
- t is top processes
- L is large pages
- **O is SEA on the VIO**
- P is paging space
- V is disk volume group
- d is disk service times
- ^ is fibre adapter stats
- W is workload manager statistics if you have WLM enabled

If you want a 24 hour nmon use:

```
nmon -ft -AOPV^dML -s 150 -c 576
```

May need to enable accounting on the SEA first – this is done on the VIO

```
chdev -dev ent* -attr accounting=enabled
```

Can use entstat/seastat or topas/nmon to monitor – this is done on the vios

```
topas -E
nmon -O
```

VIOS performance advisor also reports on the SEAs

72

72

POWER9 Performance Best Practices

A brief checklist

This document is intended as a short summary for customers on key items that should be looked at when planning a migration. For a more in-depth and more complete set of recommendations, please refer to the document links provided on the second page

Description	Instructions	Description	Instructions
Ensure firmware is current	Fix Central provides latest updates. Latest FW levels as of this writing : FW940 for ALL POWER9 systems. Use the FLRT tool to obtain the recommended levels for a given platform. NOTE: Ensure required HMC level is installed when updating FW.	LPM	• VIOS 3.1 runs on native POWER9 mode • VIOS 2.2.5 and above provide performance improvements over prior versions. • Use of 10Gb dedicated network for LPM is preferred. • Use of dual MSPs can improve performance (minimum requirements: VIOS 2.2.5 and FW860)
Memory DIMMs	All POWER9: Follow proper memory plug-in rules	AIX Tunables / VIOS Tunables	• Tuning a VIOS is not recommended unless directed by VIOS/AIX support. • Restricted tunables should not be modified (unless directed by AIX/VIOS development) • Tunables should not be migrated across HW or AIX levels.
Ensure OS level is current	Fix Central provides the latest updates for AIX, IBM I, VIOS, Linux, HMC and FW. In addition to that, the FLRT tool provides the recommended levels for each H/W model. Use these tools to maintain your system up to date.	AIX CPU utilization	The AIX OS system is optimized for best raw throughput at higher CPU usage. If the customer requires to reduce CPU usage (pc), use the schedo tunable vpm_throughput_mode to tune the workload and evaluate the benefits of raw throughput vs. CPU usage.
SMT8	To take full advantage of the performance of POWER9 CPUs, we recommend clients evaluate the use of SMT8. Proper sizing is also recommended to maximize the POWER9 improvements. We recommend when moving to SMT8 to reboot the partition to get the best performance of this change.	VIOS configuration	• Shared Ethernet adapters using a 10Gb, 40Gb or 100Gb Mellanox adapter as a backing device should enable the "flp_n_run" attribute, via chdev, on the network adapter port : chdev ent#-a flp_n_run=yes • If configured with shared processors: ➢ Assign total entitlement of all VIOS partitions to be 10-15% of cores in shared pool and assign CPU ratio of 2:1 (vCPUs:ec). Refer to the PowerVM Best Practices for additional recommendations ➢ Assign uncapped mode and set variable weight capacity of VIOS partition higher than all client LPARs serviced by VIOS • For performance and flexibility, it is recommended to use IBM I to virtualize internal storage to IBM I. If you must use VIOS, follow the wiki at the following link. • For vFC, ensure no more than 64 client connections total per physical fcs adapter on the VIOS. Also, ensure no more than 64 storage ports configured per vFC adapter on the client. These are physical limits; practical limits may differ based on workload. • For vSCSI disks, ensure the queue_depth for virtual disks is less than or equal the queue_depth of the physical disk in the VIOS. • For vSCSI adapters, ensure you configure VTDs based on the following formula: Max VTDs = (512 - 2) / (virtual_q_depth + 3) • Only enable the largesend attribute on the SEA (physical adapter backing the SEA) if all LPARs serviced by the VIOS are AIX partitions.
40GbE adapter	RHEL7: For network bandwidth sensitive workloads, we recommend increase the receive queue size from 1024 to 8192.	Virtual Ethernet adapters on AIX	• Increase the virtual Ethernet (vETH) device driver buffers if the partition is dropping packets on the virtual interface even when running with entitled CPU capacity, e.g., chdev -l ent# -a max_buf_xxxx=NNNN - NOTE: For desired buffer size adjustments, refer to 'AIX on Power - Performance FAQ' link below • Set largesend on vETH adapter to improve performance (AIX): chdev -l ent# -a mtu_bypass=on (or) ifconfig ent# largesend
Sizing a system	• When migrating to POWER9, we recommend considering using SMT8, and size the LPARs based on the SMT8 rPerf values; in many instances, this will likely reduce the number of V/Ps required. • Use Workload Estimator (WLE) for sizing LPARs for CPU consumption as it provides better sizing results.		
Right-size your Shared LPARs	• Assign entitled capacity (EC) to sustained peak utilization for LPARs with critical SLA requirements • Assign EC to average utilization and number of virtual CPUs to peak utilization (physical core consumption) for LPARs with non-critical SLA		
Partition Placement	Current FW levels ensure optimal placement of the partitions. However, if constant DLPAR operations are executed on partitions on the CEC, it is recommended the use DPO to optimize placement (requires current level of FW).		
Compilers	• IBM XL C/C++ for AIX V16.1 and XL Fortran V16.1 added support for POWER9. Also adds support for C++11 and C++14. • IBM XL C/C++ for Linux V16.1.1 & XL Fortran V16.1.1 support for P9 ISA • Advanced Toolchain : 11.0-3 and later • gcc: Version 7 of gcc is recommended for P9 ISA support. Also includes support for "-mtune=power9"		
vNIC	• AIX - Change the following vNIC interface settings as follows: chdev -l ent# -a rx_queue_num=8 -a tx_queue_num=8 -a rx_queue_elem=1024 -a use_rec_q_val=no • Linux distros - Update to latest kernel. Also, set rx/tx queues to maximum ethtool -L eth# rx 8 tx 8		
Java	• IBM JDK8 SR5 is the minimum level to exploit POWER9 • Open JDK 1.8 provides partial support for P9 ISA • Use of 64k size pages normally increases application performance		
IBMi	Ensure Technology Updates are current (see link below)		

Above at: http://www14.software.ibm.com/webapp/set2/sas/f/best/power9_performance_best_practices.pdf

73

73

8. Use NTP



74

74

Setup NTP

```
#vi /etc/ntp.conf
```

Comment out broadcast and add:

```
server 0.pool.ntp.org
```

```
server 1.pool.ntp.org
```

```
#ntpdate 0.pool.ntp.org
```

Update rc.tcpip to start ntp at boot

Now start NTP

```
#startsrc -a "-c /home/padmin/config/ntp.conf" -s xntpd
```

You can substitute your own NTP servers for the ones above if you have them

75

75

9. Update I/O firmware



76

76

Update I/O Firmware

May also need to update HMC, server firmware, VIO first – use FLRTVC to check and read the readmes

As root run: `lsmcode -A`

Check on Fix Central under Power I/O Firmware

You will need to know what kind of adapters you have (feature codes)

If you are updating the primary you can let it failover or you can force a failover

`chdev -l ent14 -a ha_mode=standby`

When done `chdev -l ent14 -a ha_mode=auto`

Example updating a 5899 network adapter with code uploaded to server

`cd /software/adapters/5899`

`rpm -ivh --ignoreos e414571614102004.10240310.aix.rpm`

`diag -T download -d ent0`

Updated all 4 ent0-ent3

You may have to unconfigure the SEA to do this (see next slide)

77

77

Update I/O Firmware – unconfigure SEA

You may have to unconfigure the SEA to do this (SEA here is on ent14)

`ifconfig en14 down`

`ifconfig en14 detach`

`rmdev -l ent14`

If IP is on the real then:

`ifconfig en4 down`

`ifconfig en4 detach`

`rmdev -l ent4`

You may also have to do this for the aggregate as well

Then:

`cfgmgr`

Check microcode went on and sea came back

`smitty tcpip` and check ip is now on

`ifconfig -a`

78

78

Useful Links

- Jaqui Lynch Articles
 - <http://www.circle4.com/jaqui/eserver.html>
- Jaqui's Movie Replays
 - <http://www.circle4.com/movies>
- Jaqui's Youtube Channel
 - <https://www.youtube.com/user/adespota4/>
- Nigel Griffiths AIXpert Blog
 - <https://www.ibm.com/support/pages/aixpert-blog-nigel-griffiths-mrnmon>
- Nigel Griffiths Twitter – mr_nmon
 - https://twitter.com/mr_nmon
- Nigel Griffiths YouTube
 - <https://www.youtube.com/nigelargriffiths>
- Gareth Coates – Tricks of the POWER Masters
 - <https://www.ibm.com/support/pages/node/1116939>
- Gareth Coates Twitter – power_gaz
 - https://twitter.com/power_gaz

81

81

Useful Links

- ESS Website to download base software
 - <https://www.ibm.com/servers/eserver/ess/index.wss?lnk=msdDO-enss-usen>
- HMC Scanner
 - <https://www.ibm.com/support/pages/node/1117515>
 - [https://www.ibm.com/support/pages/sites/default/files/inline-files/\\$FILE/hmcScanner-0.11.42.zip](https://www.ibm.com/support/pages/sites/default/files/inline-files/$FILE/hmcScanner-0.11.42.zip)
- AIX 7.2 Performance Guide
 - https://www.ibm.com/support/knowledgecenter/ssw_aix_72/performance/performance_pdf.pdf
 - https://www.ibm.com/support/knowledgecenter/en/ssw_aix_72/navigation/performance.html
- SG24-8171 – Power Systems Performance Optimization including POWER8
 - <http://www.redbooks.ibm.com/redbooks/pdfs/sg248171.pdf>
- SG24-8453 - AIX Modernization and Enhancements
 - <http://www.redbooks.ibm.com/redbooks/pdfs/sg248453.pdf>
- SG24-8171 – Power Systems Performance Optimization
 - <http://www.redbooks.ibm.com/redbooks/pdfs/sg248171.pdf>
- IBM POWER Systems Performance Report April 13, 2021
 - POWER7 to POWER9
 - <https://www.ibm.com/downloads/cas/K90RQOW8>

82

82

Backup Slides



83

83

Terminology

Run Queue

- Run queue length is another well known metric of CPU usage
 - It refers to the number of software threads that are ready to run, but have to wait because the CPU(s) is/are busy or waiting on interrupts
 - The length is sometimes used as a measure of health, and long run queues usually mean worse performance, but many workloads can vary dramatically
 - It is quite possible for a pair of single-threaded workloads to contend for a single physical resource (batch, low run queue, bad performance) while dozens of multi-threaded workloads share it (OLTP, high run queue, good performance)

84

84

Terminology

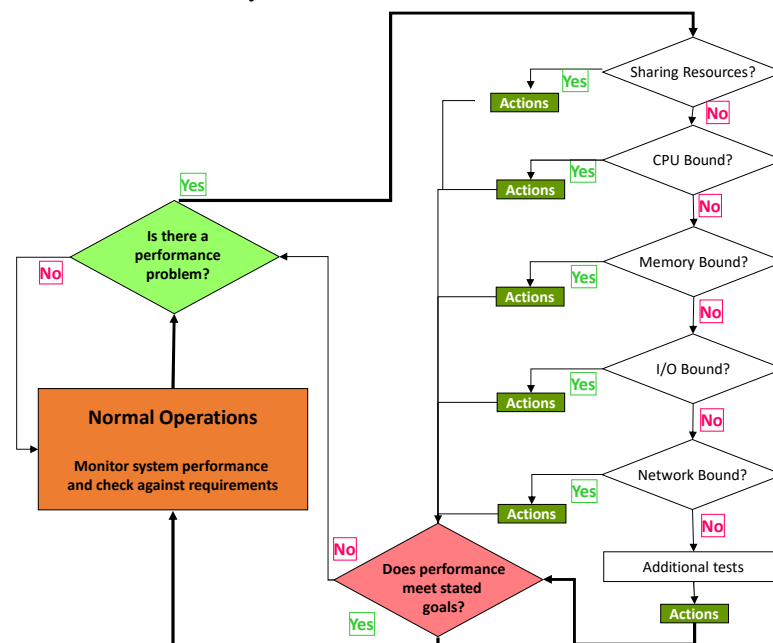
Context Switches

- Context switches
 - The number of times a running entity was stopped and replaced by another
 - Collected for Threads (operating system) and Virtual Processors (hypervisor)
 - There are voluntary and involuntary context switches
- How Many “context switches” are Too Many?
 - No rules of thumb exist
 - Voluntary: Not an issue because it means no work for the CPU
 - Involuntary: Could be an issue, but generally the bottleneck will materialize in a easier to diagnosis metric; such as, CPU utilization, physical consumption, entitlement consumed, run queue
 - Establish a baseline and compare when system encounter performance problems
- Tool outputs
 - `vmstat` reports total context switches as “cs”
 - `sar -w` as “cswch/s”
 - AIX `lparstat` reports virtual processor context switches as “vcsw”
 - AIX `mpstat` reports per logical processor context switches (involuntary: *ilcs* and voluntary: *vlcs*). The sum of *vlcs* should roughly equal VP *vcsw*.

85

85

Performance Analysis Flowchart



86

86