IBM Power Systems Performance

Hints and Tips for Migrating Workload to IBM Power9 and Power10 Processor-Based Systems

Version 2.0
2022

Disclaimer – Hints & Tips for Migrating Workload to IBM Power9 and Power10 Processor-Based Systems

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs, or services available in all countries in which IBM operates or does business.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS"

WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY

DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A

PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.

Statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Acknowledgements

We would like to thank the many people who made invaluable contributions to this document. Contributions included authoring, insights, ideas, reviews, critiques, and reference documents.

Our special thanks to key contributors from IBM Power Systems Performance:

Abraham Alvarez - Power Systems Performance

Clark Anderson - Power Systems Performance

Qunying Gao - Power Systems Performance

Aravinda Herle - Power Systems Performance

Dirk Michel - Power Systems Performance

Rick Peterson - Power Systems Performance

Joel Peterson - Power Systems Performance

Puvichakravarthy Ramachandran - Power Systems Performance

Theresa Xu - Power Systems Performance

Our special thanks to key contributors from IBM Development:

Juan M Casas - AIX development

Arnold Flores - AIX development

Bob G Kovacs - AIX development

Brian Veale - AIX development

Our special thanks to key contributors from IBM Global Support:

Janine A Jackson - AIX/VIOS L3 support

Evan Zoss - AIX/VIOS L3 support

1 Preface

This document is intended to provide guidance, best practice recommendations, and a checklist for migrating workloads from Power7 and Power8 systems to Power9 and Power10 systems. This document by no means covers all the PowerVM, AIX or IBM i best practices and should be used in conjunction with other PowerVM, AIX or IBM i documents.

2 Introduction

When workloads are deployed on new hardware configurations, care must be given to the configuration and tuning of the new system to achieve the expected performance. This starts with good system configuration planning and continues into system setup and deployment. By spending some time to consider and implement these migration guidelines and to establish sound system tuning and configuration practices, you will be better prepared for a successful migration to a Power9 and Power10 process-based systems.

3 Planning Workload Migration

Live Partition Mobility (LPM) has made it easy to move the running partitions from one Power system to another without downtime. The work of planning the migration should include all different areas which pertain to your applications, as well as some of the key areas that we discuss below, so that your migration runs smoothly, and you obtain the best performance from your hardware. Some of these include:

3.1 Install the latest firmware on the Power9 and Power10 system, update HMC software to latest version.

3.2 The support website FixCentral provides updates for the latest OS updates for IBM i, VIOS, AIX. Where possible, update the OS levels of the logical partitions to be migrated to the latest levels, or at the very least the minimum levels recommended for the platform.
NOTE: Be aware that installing only the minimum levels can potentially leave your partitions or workloads vulnerable to issues that have been resolved in some of the latest updates of the operating systems. See section 4 and 4.1 in this document for detail link.

3.3 The Supported Linux Distribution website lists the supported linux releases for Power10

3.4 Power9 and Power10 makes more efficient use of the 8 hardware SMT threads available per CPU (when running in SMT8 mode). When migrating from a Power7 or Power8 platform, consider the use of SMT8, in addition to considering reducing the allocation of CPUs (in dedicated CPU LPARs), or reducing VPs and CPU entitlement (on shared CPU LPARs). Refer to Virtual Processor and Entitlement Considerations (section 6) in this document for additional information.

3.5 Placement of the partition is important in order to obtain the best performance in your Power9 and Power10 systems. In order to optimize the initial placement of your partitions, it is best to create the most important and/or largest partitions first, followed by less important and/or smaller partitions. If necessary, placement can be optimized after the partitions have been created by using DPO on the HMC against either the whole managed node or the partition you are most interested in. It is also suggested to run DPO if you are constantly running DLPAR operations across partitions on your managed node; constantly running DLPAR (either add or remove) can lead to resources being placed in non-optimal locations. To check the affinity score of your system or LPARs, you can use the lsmemopt (see the HMC documentation for further information).

After a live migration, the cpu and memory affinity of the LPAR could be different. Use the following commands to check the topology:

AIX: lssrad -av

Linux: numactl (may require reboot of the logical partition). lparnumascore -d 4 can be used to understand the affinity and affinity score of the LPAR before a reboot.

IBM i: Collection Services includes metrics that can be useful for monitoring authority activity. The primary metric to monitor is AFSCORE (affinity score) in the file QAPMSYSAFN. You can use the iDoctor tool to examine this Collection Services information in the System tab.

3.6 Capacity planning is important when considering processor migration. Consider the application behavior (e.g., highly multi-threaded workloads vs single-threaded workload) when setting performance improvement goals and expectations. See the following documents for more information:

IBM i on Power -Performance FAQ
IBM Power Systems Performance Report

3.7 EnergyScale performance can allow higher processor frequency under the right conditions. Consider the workload, the system environment, and EnergyScale configuration to better understand how performance can be affected.

See the following documents for more information on Power9 and Power10 EnergyScale and frequency changes:
IBM EnergyScale for POWER9 Processor-Based Systems
POWER9 EnergyScale - Configuration & Management
IBM EnergyScale for Power10 Processor-Based Systems
IBM Power10 based systems Redbooks

4 Software Requirements

The following lists latest software for migration.

AIX Requirements:
https://www-01.ibm.com/support/docview.wss?uid=ssm1platformaix

VIOS Requirements:
https://www-01.ibm.com/support/docview.wss?uid=ssm1platformvios

Linux Requirements / Support:
https://www.ibm.com/docs/en/linux-on-systems?topic=lpo-supported-linux-distributions-virtualization-options-power10-linux-power-servers

HMC / Firmware Supported Combinations
https://www-945.ibm.com/support/fixcentral/main/transform?xml=https://download.boulder.ibm.com/ibmdl/pub/software/server/firmware/sfw_fixSupportedCombos.xml&title=HMC+/+Firmware+Supported+Combinations

HMC and HMC Virtual Appliance on Fix Level Recommendation Tool (FLRT) Lite
https://www14.software.ibm.com/support/customercare/flrt/liteTable?prodKey=hmc

Power Code Matrix - Supported HMC Hardware
https://www.ibm.com/support/pages/node/6554904

IBM i Technology Updates:
www.ibm.com/ibmi/techupdates

IBM i Support: Recommended fixes:
https://www.ibm.com/support/pages/ibm-i-support-recommended-fixes

4.1 Known issues

AIX

There are some issues that have been found to make an impact in performance on LPARs which are migrated to Power10 based systems. We recommend that the LPARs which are planned to be migrated, are updated to the following AIX levels to prevent being exposed to performance issues after the migration to Power10:

Description	VRMF Levels containing tunable change (out-of-box)	Assoc. Release Date	APARs
Shed mode	7300-00 (730) 7200-05-03-2135 (72X) 7200-05-02-2113 (72V) 7200-04-04-2114 (72Q) 7100-05-08-2113 (71c)	12/10/21 9/10/21 4/16/21 6/25/21 4/16/21	NO_APAR IJ30874 IJ30849 IJ31336 IJ30840
VPM Fold threshold	7300-00 (730) 7200-05-03-2135 (72X) VIOS_SP_3.1.2.30 (VIOS) 7200-04-05-2148 (72Q) 7100-05-09-2135 (71c)	12/10/21 9/10/21 2/11/22 2/11/22 9/10/21	NO_APAR IJ31874 IJ33110 IJ32201 IJ32060
Scheduler tuning	7300-00 (730) 7200-05-03-2135 (72X) VIOS_SP_3.1.2.30 (VIOS) 7200-04-05-2148 (72Q) 7100-05-09-2135 (71c)	12/10/21 9/10/21 2/11/22 2/11/22 9/10/21	NO_APAR IJ31873 IJ34066 IJ34487 IJ32391

Java related:

Description	Link	APARs
Java Error on AIX7.3: negative offset caused illegal instr by assembler	https://www.ibm.com/support/pages/apar/IJ34951	IJ34951
Containier crash: need to consider portable-aot cases during code patching	https://www.ibm.com/support/pages/apar/IJ34950	IJ34950

5 Processor Compatibility Mode

The processor compatibility is configured in the logical partition profile on the HMC.

If you want to leverage the full features of the system, you should configure the processor compatibility mode to default.

If you want to be able to migrate to a system with an older processor generation, the processor compatibility mode of your partition needs to be set to a processor compatibility mode supported by the destination system. For example, if you have Power10, Power9, and Power8 system you may want to set the processor compatibility mode to Power8 so that you can migrate the partition to any of the systems.

The compatibility mode of logical partitions will be preserved across migration unless you change it. The change of the compatibility mode of the logical partition is not dynamic, it requires a shutdown and restart of the logical partition. When you restart the logical partition, hypervisor checks the configured processor compatibility mode and determines whether the operating environment supports that mode. If the operating environment supports the configured processor compatibility mode, the hypervisor assigns the logical partition the configured processor compatibility mode. If the operating environment does not support the configured processor compatibility mode, the hypervisor assigns the logical partition the most fully featured processor compatibility mode that is supported by the operating environment.

For more detail on processor compatibility modes, refer to the following:
Power9: https://www.ibm.com/docs/en/power9/9009-42A?topic=mobility-processor-compatibility-modes
Power10: https://www.ibm.com/docs/en/power10/9080-HEX?topic=mobility-processor-compatibility-modes

To check the compatibility mode that your partition is running from the OS refer to the following commands:
AIX: lsconf
Linux: LD_SHOW_AUXV=1 /bin/true | grep _PLATFORM
IBM i: need to check partition profile on HMC

For AIX the default SMT level is SMT4 on both Power7 and Power8 systems, and the default is SMT8 on Power9 and Power10 systems with compatibility mode of Power8 and above. After migrating from Power7 or Power8 to Power9 and Power10 system, if you don’t reboot AIX, it will continue running in SMT4 mode. After rebooting AIX, if logical partition operates in Power7 mode, AIX will remain in SMT4 mode. If the logical partition operates in Power8 mode or above, AIX will change from SMT4 to SMT8 mode. If you want to preserve SMT4 mode across logical partition reboots, you need to run smtctl and bosboot command.

For Linux, the default SMT level is SMT8. For IBM i: the default SMT level is SMT8.

The following tables describe the features supported with various compatibility modes on Power9 and Power10 systems.

Platform	Compatibility mode	FW Level	AIX Level	P9 PMU	P10 PMU	SMT8	XIVE	NX GZIP	P10 MMA	P10 optimized memcpy
Power9	Power7	FW910	7.1 TL 4							
Power9	Power8	FW910	7.1 TL 4			X				
Power9	Power9_Base	FW910	7.2 TL 4	X		X				
Power9	Power9	FW940	7.2 TL 4	X		X	X	X		
Power10	Power8	>=FW1010	7.1 TL 5 SP2			X				
Power10	Power9_Base	>=FW1010	7.2 TL 4 SP2	X		X				
Power10	Power9	>=FW1010	7.2 TL 4 SP 2	X		X	X	X		
Power10	Power10	>=FW1010	7.3 TL 0		X	X	X	X	X	X

Platform	Compatibility mode	FW Level	Linux Level (min)	P9 PMU	P10 PMU	SMT8	XIVE	NX GZIP	P10 MMA	P10 optimized memcpy
Power9	Power8	FW910	RHEL 8.0			X				
Power9^	Power9_Base	FW910	SLES12 SP3 SLES15 RHEL8 RHEL9	X		X				
Power9^	Power9	FW940	SLES12 SP3 SLES15 RHEL8 RHEL9	X		X	X			
Power10	Power9_Base	>=FW1010	SLES12 SP5 RHEL8.2	X		X				
Power10	Power9	>=FW1010	SLES12 SP5 RHEL8.2	X*		X	X			
Power10	Power10	>=FW1010	SLES15 SP3 RHEL8.4		X	X	X		X	
Power10	Power10	>=FW1010	RHEL8.5		X	X	X		X	X
Power10	Power10	>=FW1010	SLES15 SP4 RHEL8.6 RHEL9.0		X	X	X	X	X	X

^ The minimum supported Linux level varies depending on the different Power9 offerings. Please visit this link for the right support statement : Linux Support on Power

*Power9 compatibility mode has limited PMU capabilities. Basic PMU events are supported.

Platform	Compatibility mode	FW Level	IBM i Level(min/max)	P9 PMU	P10 PMU	SMT8	XIVE	NX GZIP	P10 MMA	P10 optimized memcpy
Power9	Power7	FW910	7.1/7.3	X						
Power9	Power8	FW910	7.1/7.4	X		X				
Power9	Power9_Base	FW910	7.1	X		X				
Power9	Power9	FW940	7.1	X		X				
Power10	Power8	>=FW1010	7.3/7.4		X	X				
Power10	Power9_Base	>=FW1010	7.3		X	X				
Power10	Power9	>=FW1010	7.3		X	X				
Power10	Power9	>=FW1010	7.4		X	X	X	X		
Power10	Power10	>=FW1010	7.3		X	X			X	
Power10	Power10	>=FW1010	7.4		X	X	X	X	X	
Power10	Power10	>=FW1010	7.5		X	X	X	X	X	

XIVE: eXternal Interrupt Virtualization Engine. Power9 and Power10 system supports a larger number of interrupts sources and delivers interrupts directly to virtual processors without going through hypervisor.
NX GZip: Power9 and Power10 processor-based servers support on-chip accelerators that perform various functions such as compression, decompression of data. In Power10 mode, the NX GZip allows direct user level access. For linux specific support, refer the libnxgz github space.
P9 PMU: Performance Monitor Unit. PMU is a programmable component of microprocessor core on the chip. The PMU provides a programmable interface for monitoring and collecting various hardware performance event counters. Partitions on Power9 platform need to run in Power9 compatibility mode to have full access to the Power9 PMU.
P10 PMU: Performance Monitor Unit. PMU is a programmable component of microprocessor core on the chip. The PMU provides a programmable interface for monitoring and collecting various hardware performance event counters. Partitions on Power10 platform need to run in Power9 compatibility or Power10 native mode to have full access to the Power10 PMU.
P10 MMA: Power10 processor-based servers support the Matrix Math Accelerator (MMA) that can be used to accelerate enterprise AI inferencing.

Power10 Radix support for Linux:

Starting from Power10, phyp enables lpar support for radix page table. For partitions running Linux on Power10, the default mode is Power10/Radix mode. A Linux partition will run with HPT mode after live partition migration from Power9 or earlier systems, partition deactivation/activation is required to switch to Radix mode.

6 Virtual Processor and Entitlement Considerations

As mentioned earlier in this document, the Power9 and Power10 processors have improvements for all SMT modes (e.g. SMT2, SMT4, SMT8). When using SPLPARs (shared processor LPARs), it is therefore recommended that you follow these guidelines in order to obtain the best results in your migration to Power9 and Power10.

With these new improvements, in general, workloads will show better performance when run in SMT8 mode, but at the same time, workloads running in SMT2 mode, will also see a significant improvement.

When planning the shared processor configuration, you need to understand well the goal that you have in mind when migrating to a Power9 and Power10 based systems.
Your VP and entitlement configurations previously determined for Power7 and Power8 based systems should be revisited and most likely reduced when migrating to a Power9 and Power10 based system using these guidelines to achieve your intended goal.

As a rule of thumb, it is recommended that you assign the VPs and entitlement to the LPAR as follows:

- Use the peak utilization of the LPAR as the baseline to assign VPs.
- Use the average utilization of the LPAR as the baseline to assign entitlement.

For further guidance on virtualization, see the following document:
IBM Power Virtualization Best Practices.

Given the higher thread strength on Power9 and Power10 processors, to maximize the processing capacity of the processor, we recommend using all the hardware threads available by using the SMT8 mode (default setting) on the partition. As noted in section 5 above, AIX will default to SMT8 on Power9 and Power10 at boot time. If migrating from Power7 or Power8, a reboot will be required to switch to the SMT8 default.

In addition to using SMT8, if the goal of the migration is to reduce the processor capacity utilized by the LPAR, you must reduce the number of CPUs (dedicated LPAR) or VPs (shared LPAR) allocated to the LPAR on the new Power9 and Power10 based system. As a baseline, evaluate the capacity required to run your existing workload on a Power9 and Power10 system based the AIX capacity rating (rPerf value) which can be found on the IBM Power Systems Performance Report or, for IBM i capacity planning, the CPW rating which is listed in Section 3.6 in this document.

> (AIX only)
> In addition to the reduction of VPs, you can consider setting the schedo tunable "vpm_throughput_mode" to a value of 2 on Power9 processor-based systems. By default, running in raw throughput mode (i.e. vpm_throughput_mode=0) AIX spreads all available

work across as many VPs as available, dispatching work **first** to the primary SMT thread of each VP, then the secondary SMT threads, and so on, thus providing the best performance in most workloads. This configuration, however, can lead to a higher PC (processor capacity) utilized by the LPAR. Depending on the workload, you can reduce the PC by setting the vpm_throughput_mode tunable to 2, thus having AIX to schedule work to the primary and secondary threads equally. A more detailed discussion of the different modes can be found in the IBM Power Virtualization Best Practices. On Power10 processor-based systems, the default for "vpm_throughput_mode" is set to a value of 2 for shared processor partitions.

One way you can review the CPU utilization of an AIX partition is by looking at the output of the mpstat command. In the following example you can see that mostly the primary threads of each of the VPs assigned to this LPAR are busy doing work.

```
cpu  min  maj  mpc  int   cs  ics   rq  mig lpa sysc  us sy wa  id   pc  %ec  lcs   Time
    0    0    0    0  265   94    1    1    1 100  288  12 65  0  22 0.00   0.1  225 18:55:39
  1    0    0    0   14    0    0    0    1 100    0   0  0  0 100 0.00   0.1   15 18:55:39
  2    0    0    0    9    0    0    0    1 100    0   0  1  0  99 0.00   0.0   10 18:55:39
  3    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.00   0.1   10 18:55:39
  4    0    0    0    9    0    0    0    1 100    0   0  1  0  99 0.00   0.0   10 18:55:39
  5    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.00   0.1   10 18:55:39
  6    0    0    0    9    0    0    0    1 100    0   0  1  0  99 0.00   0.0   10 18:55:39
  7    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.00   0.1   10 18:55:39
  8    0    0    0  120   32   11    1    2 100    0 100  0  0   0 0.32  10.7  100 18:55:39
  9    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   11 18:55:39
 10    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   11 18:55:39
 11    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   11 18:55:39
 12    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   10 18:55:39
 13    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 14    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   10 18:55:39
 15    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 16    0    0    0  115    1    1    1    1 100    0 100  0  0   0 0.32  10.8  100 18:55:39
 17    0    0    0   17    0    0    0    1 100    0   0  0  0 100 0.12   4.1   19 18:55:39
 18    0    0    0   14    0    0    0    1 100    0   0  0  0 100 0.06   1.9   15 18:55:39
 19    0    0    0   12    0    0    0    1 100    0   0  0  0 100 0.12   4.1   13 18:55:39
 20    0    0    0   11    0    0    0    1 100    0   0  0  0 100 0.06   1.9   12 18:55:39
 21    0    0    0   10    0    0    0    1 100    0   0  0  0 100 0.12   4.1   11 18:55:39
 22    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   11 18:55:39
 23    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   11 18:55:39
 24    0    0    0   99    1    1    1    1 100    0 100  0  0   0 0.32  10.7   99 18:55:39
 25    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 26    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   10 18:55:39
 27    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 28    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   10 18:55:39
 29    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 30    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.06   1.9   10 18:55:39
 31    0    0    0    9    0    0    0    1 100    0   0  0  0 100 0.12   4.1   10 18:55:39
 32    0    0    0    0    0    0    0    1 100    0   0 54  0  46 0.00   0.0    0 18:55:39
 33    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 34    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 35    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 36    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 37    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 38    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 39    0    0    0    0    0    0    0    1 100    0   0  0  0 100 0.00   0.1    0 18:55:39
 40    0    0    0    0    0    0    0    1 100    0   0  6  0  94 0.00   0.0    0 18:55:39
 41    0    0    0    0    0    0    0    1 100    0   0  5  0  95 0.00   0.0    0 18:55:39
 42    0    0    0    0    0    0    0    1 100    0   0  4  0  96 0.00   0.0    0 18:55:39
 43    0    0    0    0    0    0    0    1 100    0   0  5  0  95 0.00   0.0    0 18:55:39
 44    0    0    0    0    0    0    0    1 100    0   0  4  0  96 0.00   0.0    0 18:55:39
 45    0    0    0    0    0    0    0    1 100    0   0  4  0  96 0.00   0.0    0 18:55:39
 46    0    0    0    0    0    0    0    1 100    0   0  4  0  96 0.00   0.0    0 18:55:39
 47    0    0    0   17    0    0    0    1 100    0   0 41  0  59 0.00   0.0   18 18:55:39
ALL    0    0    0  892  128   14    4   49 100  288  32  0  0  68 3.02 151.1  852 18:55:39
```

This workload could be compacted to run mostly on the first two VPs, if the system were to run with vpm_throughput_mode=2, thus reducing the PC of the system.

To compare Power9 and Power10 processor-based systems to previous models, use the following documents:

IBM Power Systems Performance Capabilities Reference
or
IBM Power Systems Performance Report

For best performance, we recommend that the system be fully populated with DIMMs, rather than only partial population of DIMMs. This will allow the hypervisor a better chance to place the LPARs in an optimal location on the system.

Improving the LPAR placement will also improve the latency of most workloads, as it will make optimal use of the memory and CPU resources on the system.

For small partitions, it is best to contain the partition on a single SRAD (Scheduler Resource Allocation Domain) or numa node when possible. The hypervisor attempts to place the LPAR in an optimal location, based on:

- CPU entitlement assigned to the partition
- Amount of memory allocation to the partition
- I/O devices allocated to the partition
- Available (free) memory / CPU on the available SRADs

More detailed information regarding the above can be found in the IBM Power Virtualization Best Practices.

An easy way to check at a high level the placement of the partition you can run the following commands:

<table>
<tr><th></th><th>Command to run</th><th>Sample output</th></tr>
<tr><td>AIX</td><td>lssrad -av</td><td>REF1   SRAD   MEM   CPU
0
          0   63722.19   0-63
          1   63495.00   64-127
          2   63495.00   128-191
          3   63495.00   192-255
1
          4   63495.00   256-319
          5   63472.00   320-383
          6   63744.00   384-447
          7   63738.44   448-511</td></tr>
<tr><td>IBM i</td><td>rmnodeinfo macro</td><td>Node statistics for 2 nodes across 2 node groups
Node # | 0 | 1
|===
Hardware node ID | 1 | 2 |
Node group ID | 0 | 1 |
Hardware group ID | 0 | 2 |
of Logical procs | 16 | 16 |
of procs folded | 8 | 16 |
main store pages | 000748DB | 0007B8A8 |</td></tr>
<tr><td>Linux</td><td>numactl -H</td><td>available: 2 nodes (0,8)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
node 0 size: 261646 MB
node 0 free: 38220 MB
node 8 cpus: 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159
node 8 size: 261719 MB
node 8 free: 237764 MB
node distances:</td></tr>
</table>

Placement of the LPAR can be improved, when possible, by running the optmem command from the HMC which manages this system.

7 I/O Considerations

If you are using the same network and storage adapters and the same storage subsystem configuration as your previous system, initially, the same tuning should be used on the new system. If additional performance is desired from the existing system, then normal network and storage tuning should be performed.

If the I/O subsystems have not changed, or the storage and network devices that connect to those adapters are capable of higher throughput or lower latency than the previous subsystem components, then the influence the I/O subsystems exert on the perceived speed of the applications will either be imperceptible or possibly show improved speed.

If the network or storage subsystems are appreciably different on the newer system than the prior system, the following list of considerations could negatively impact the perceived speed of applications.

- Devices that connect to adapters:
    - Changing from Direct Attached Storage (DAS or internal) to Storage Area Network (SAN) or Network Attached Storage (NAS) (or external storage) can increase latency. Less dedicated write cache space resulting in lower write cache efficiency, longer I/O data path lengths and the potential of other activity sharing the network and storage servers may cause increased latency.
    - Higher speed networks. Many users are moving up to 25, 50 or 100 GigE networks which require revisiting tuning.
    - Additional functions such as compression, encryption and deduplication can add latency.
    - Refer to tuning or setup guides for the new devices to understand these impacts.
- Configurations:
    - Network and storage fabric topology changes can result in slower or fewer number of paths.
    - Different storage protection levels result in differing performance impacts on certain workloads. For example, RAID 10/1 (or mirroring) has less negative impact on storage write performance than RAID5, which has less impact than RAID6.
    - Many users are implementing higher level of security which may require IPsec or other security protection. These can negatively affect network bandwidth and latency.
- Sizing:
    - Reducing the number of Storage LUNs can reduce resources in the server needed to support required throughputs. For example, do not size storage subsystems on capacity alone. Do take into account performance of both logical and physical devices, thus the number of them.

        - For example, NVMe devices have a virtualization layer built into their controllers that allows the physical storage space to be partitioned into multiple logical devices called 'namespaces'. Partitioning into more logical devices can allow the host OS or VIOS to utilize more CPU and memory resources for the I/O and allow more parallelism, which can, depending upon a myriad of variables, allow the host to take better advantage of the high throughput SSDs can provide. Namespaces can be created and deleted from the NVMe manager screens.
    - Storage server sizing tools may make estimates without taking into account effects of server impacts like internal bus bandwidth limitations and/or latencies. Derating their output by some percentage, 5-10%, is advisable if the server impacts are unknown.
- Virtualization:
    - Virtualization does add latency and can reduce throughput compared to native I/O. Besides the backend hardware, ensure VIOS memory and CPU amounts are enough to provide the required throughput and response times.
    - Moving to higher speed virtualized network adapters in VIOS will require adjusting the VIOS configuration in CPUs and memory.
    - IBM PowerVM Best Practices Redbook at http://www.redbooks.ibm.com/abstracts/sg248062.html can be helpful sizing VIOS.
- AIX Specific Tuning Effects:
    - Maximizing the number of processors used to handle I/O completion interrupts promotes higher throughput capabilities. It is advised to set the 'Desired processors' and 'Maximum processors' value in the partition's profile as defined in an HMC to Power of 2 values to better enable AIX or VIOS to use more processors for I/O interrupts.
    - Storage: Ensure the per device and per adapter queue depths and number of channels per adapter/port are sufficiently high enough to overcome the I/O path latency to be able to reach desired throughputs.
        - SSDs that attach with the NVMe protocol have large command and response queues called 'channels'. It is recommended to increase the number of channels for workloads that require high command throughput. 'High' is defined as at or approaching the device's io/s limit for the storage I/O workload that the application generates. Increasing the number of queues will not typically lower response times, nor will it increase throughput for storage I/O workloads that stress high data throughputs, where units are GB/s, and the I/O lengths are typically 16KB/io or larger.

          The number of NVMe channels, or 'nchans', can be altered via smitty menus or the command line. For smitty menus use: "Devices", then "NVMe Manager", then "Change / Show Characteristics of a NVMe Controller", then choose your drive, then the "Number of Channels" field can be selected and altered. The following command line will also work.
            - chdev -l nvmeX -a nchan=8 -P (where X = 0, 1, etc)

- ○
  - Network: For higher speed >=25 GigE adapters the default settings for number of transmit and receive queues and entries in the queues are a starting point. For 100 GigE adapter please consult the 100 GigE adapter tuning guide.
  - IPsec: If you are not setting any IPsec rules then IPsec should be turned off. Due to the architecture of IPsec, increasing the number of transmit and receive queues on an adapter interface may not improve single client IPsec performance.
- IBM i Specific Effects:
  - Be aware that storage devices that support IBM i's native block length, 4160 or 520 bytes, can result in more efficient I/O (less CPU usage per I/O) than storage devices that only support 4096 or 512 byte block lengths.

8 Cloud consideration

- OCP 4.9 or higher provides general support for Power10
- Build your Java application container image using OS UBI + JDK or JRE packages using minimum level of IBM SDK V8.0.6.36 / V11.0.12.1 or later:
  - Semeru Runtimes: https://developer.ibm.com/languages/java/semeru-runtimes/downloads
  - IBM SDK: https://www.ibm.com/support/pages/java-sdk-downloads-version-80
  - Java UBI from RedHat catalog use the image on 2021/09/23 or later

9 Consider Lab Services engagement and/or benchmark

The IBM Systems Lab Services organization https://www.ibm.com/it-infrastructure/services/lab-services is available to assist you with resolving system, application, and database performance problems. Formal and informal training opportunities are also available where you learn how to use the performance tools and resolve performance problems on your own.

If you need additional help in assessing the potential impact of a system migration, benchmarking a system environment, or identifying ways to improve the performance of your environment, please contact IBM Lab Services at ibmsls@us.ibm.com.

10 Migration checklist

- Plan the migration.
- Install latest required software, apply the available fixes.
- Set appropriate processor compatibility mode for logical partitions before and after migration.
- Plan the virtual processor and entitlement for logical partition to best fit your operation and performance requirement.
- Follow I/O consideration guide.
- Consider engagement with IBM Systems Lab Services as described in Section 9.