IBM AIX on Power10 performance topics

The Power10 processor

Table of contents

Preface

This document aims to offer guidance and topics for optimizing performance for IBM® Power10 processor-based servers. It should be noted that this document does not cover all the best practices for PowerVM™, AIX®, or IBM i, and should be used in conjunction with other relevant documentation.

Introduction

To achieve the expected performance on new hardware configurations, you should carefully configure and tune the system. The process begins with good system configuration planning and continues into system setup and deployment. By considering and implementing the provided migration guidelines and establishing sound system tuning and configuration practices, you can better prepare for a successful migration to a Power10 process-based system. It is important to consider best practices when planning any deployment.

This document discusses the design of Power10 hardware, which includes a single chip module (SCM) and a dual chip module (DCM) configuration. Both configurations have specific advantages and are designed to provide capabilities that are applicable to different customer applications.

This document presents topics that provide insights and awareness into commonly encountered performance issues that may be addressed in AIX systems. It should be noted that the topics are not listed in a particular order. There is a common misconception that all tuning changes made on AIX will be applicable to newer hardware and system software. However, for IBM Power servers running on AIX, it is best to use the default tuning and then address any issues presented by a particular workload.

This document also includes AIX specific performance topics that may apply to recent hardware. The focus here is to provide performance topics deemed important and applicable to AIX running on Power10 and includes reference links to more detailed documentation. The topics are presented as an overview and are not intended to be comprehensive. There are new capabilities provided by the Power10 hardware related to performance enhancements that customers would benefit from knowing about.

Power10 family

The Power10 processor chip is a powerful computing solution that offers several key features and capabilities. Here are some of the insights into the Power10 family:

- The Power10 family uses 7nm technology.
- It has up to 24 cores/socket (192 HW threads) with SMT-8 cores in Dual Chip Module (DCM) configuration.
- It has up to 15 cores/socket (240 HW threads) with SMT-8 cores in Single Chip Module (SCM) configuration.
- The processor features a modular building block die and a new core microarchitecture.
- The ISA is optimized for artificial intelligence (AI) and it has a focus on energy efficiency.
- Power10 has hardware enforced security and designed for enterprise use.
- It uses PowerAXON 2.0 and PCIe G5 technologies.
- Power10 supports memory clustering.

Power10 processor chip

The Power10 processor offers a range of innovative features and capabilities, these include:

- Up to 15 SMT8 cores with 2 MB L2 Cache/core for efficient processing.
- Up to 120 simultaneous hardware threads for handling multiple tasks simultaneously.
- Up to 120 MB L3 cache with low latency NUCA management to reduce delays in data retrieval.
- Larger working sets to accommodate complex applications and data.
- L1 instruction cache: 2 x 48 KB 6-way (96 KB total) for faster instruction execution.
- L2 cache: 2 MB 8-way with a 400% improvement in processing speed.
- L2 translation lookaside buffer (TLB): 2 x 4K entries (8K total) for efficient address translation.
- L3 local 8 MB for quick access to frequently used data.
- L3 NUMA (socket) 120 MB for efficient data sharing between cores.
- Data access with reduced latencies to minimize delays.
- L1 data cache access at four cycles nominal with zero penalty for store-forwarding.
- L2 data access at 13.5 cycles nominal for faster data retrieval.
- L3 data access at 27.5 cycles nominal for quick access to larger data sets.
- Translation lookaside buffer (TLB) access at 8.5 cycles nominal for effective-to-real address translation (ERAT) miss, including for nested translation.

Power10 high-end highlights

Power10 processor is a powerful and versatile computing platform designed to deliver exceptional performance, scalability, and reliability. Its advanced architecture and innovative features make it an ideal choice for high-performance computing, enterprise applications, and cloud environments. In this paper, we will dive deeper into the high-end highlights of the Power10, exploring its most impressive features and capabilities.

- Offers a maximum of 240 Power10 Simultaneous Multithreading 8 (SMT8) cores with 10, 12, or 15 core options.
- Features a Modular Scalable Design that allows for up to four 5U Compute Enclosure Drawer (CED) drawers and a 2U Control Unit.
- Provides up to 64TB total memory, with 16TB per drawer, using new OMI (Open Memory Interface) DIMMs that offer increased memory bandwidth of 409 GB/s per socket.
- Has eight Peripheral Component Interconnect Express (PCIe) slots per drawer that support blindswap GEN5 capability.
- Includes new SMP Cables (non-active) with concurrent maintenance capability.
- Offers high-speed 32 Gbs ports for potential future support of Memory Inception/Clustering.
- Provides internal storage with four Non-Volatile Memory Express (NVMe) Flash 7mm U.2 Bays (or 2 15mm U.2 Bays) per drawer.
- Features secure and trusted boot with redundant TPM modules.
- Offers up to 16 I/O expansion drawers, with 4 drawers per Central Electronics Complex (CEC) drawer.
- Equipped with a 2U drawer that contains a Firmware-based System Processor (FSP) based system control unit, similar to the one used in E980 server system.
- Offers terabyte/second sockets and petabyte system memory capacities, with 16 socket SMP to Cluster capabilities.
- Features a New Core Architecture that offers flexibility, larger caches, and reduced latencies.
- Provides hardware-enabled and co-optimized security with the PowerVM hypervisor.
- Offers a 3x improvement over Power9 in terms of energy efficiency.
- Offers 10-20x matrix-math performance/socket compared to Power9.
- Includes 4-32x Matrix Math Acceleration with 5 512b engines per core, resulting in 2048b results/cycle.
- Offers matrix math outer products with double, single, and reduced precision options.
- Provides single-instruction, multiple-data (SIMD) with 8 independent SIMD engines per core for fixed, float, and permute operations.

Chip interconnect – AXON

The Power10 processor utilizes the AXON chip interconnect technology, specifically the PowerAXON 2.0 version. Within a CEC drawer, three PowerAXON1 buses, each 18 bits wide, create a fully connected fabric. This interconnect design allows each Single Core Module (SCM) within a system node to be directly connected to every other SCM in the same drawer at a speed of 32 Gbps. With this on-planar interconnect, there is a total chip-to-chip data bandwidth of 128 GBps, which is a 33% increase compared to the previous implementation on the Power9 processor-based Power E980 systems. The throughput of the interconnect can be calculated as 16 bits per lane multiplied by 32 Gbps, resulting in a rate of 64 GBps per direction, and an aggregated rate of 128 GBps in both directions.

Power10 server configurations

Power10 server features the latest processor design, including Dual Chip Module and Single Chip Module packaging. These configurations range from 4 cores to 24 cores per socket and are available in the following models:

- IBM Power S1014 (9105-41B).
- IBM Power S1022s (9105-22B).
- IBM Power S1022 (9105-22A).
- IBM Power S1024 (9105-42A).
- IBM Power E1050 (9043-MRX) with DCM packaging.
- IBM Power E1080 (9080-HEX) with SCM packaging.

Power10 SCM and DCM configurations

The Power10 processor is available in both single chip module (SCM) and dual chip module (DCM) configurations.

The E1080 SCM configuration allows a maximum of 15 cores on a single socket with a maximum of 16 sockets, providing up to 240 cores. Each node has four sockets, and a maximum of four nodes are allowed in a system. The processor feature #EDP4 provides 60 cores per system node and a total system capacity of 240 cores for a 4-node Power E1080. An SCM can have 10, 12, or 15 Power10 processor cores, and all SCMs in a node must have the same number of cores.

On the other hand, the E1050 DCM configuration allows a maximum of 96 cores when using 24-core DCMs populating 4 sockets. A DCM fills one socket, and the DCMs are available with 12, 18 or 24 cores per socket. The system may be configured with 2, 3, or 4 populated sockets. There is a maximum of one node with four sockets and multi-node configurations are not supported.

The hardware design determines the near, far, and distant NUMA memory and CPU operations, where the CPI required increases as access moves from near to far and distant hardware locations.

AIX default tuning

The AIX operating system on IBM Power systems has been optimized for general performance through extensive testing and best practices tuning defaults set by the IBM Power Systems Performance team. These defaults have been tested against industry standard benchmarks such as SPEC, TPC, HPC and published IBM Power performance report (rPerf) performance data. Therefore, it may not be necessary to make small changes to the AIX configuration, especially if you have many logical partitions (LPARs) to manage.

AIX offers powerful and flexible tuning options, but it is important to test any changes with a specific workload to ensure suitability. Migrating tuning changes between different software releases or hardware platforms should be done with caution as they may cause unexpected performance differences. The best approach is to start with the defaults and evaluate CPU, memory, I/O, and network focus areas.

AIX restricted tunables

The Power Systems Performance team has implemented the use of restricted tunables on AIX to improve the code base. These tunables can only be changed after a detailed analysis of the workload and approval by AIX Performance Development, which may result in a code change or a modification to the default tunable value. This approach aims to minimize the need for customer changes by incorporating improvements into future releases. As a result, most tuning suggestions are unnecessary since the team has already optimized the defaults and resolved bugs in the code.

LPAR placement

LPAR placement significantly impacts performance due to hardware implementation and NUMA architecture, particularly when LPARs become larger than the number of CPUs that can be contained on one node or socket. A node, also known as a central electronic complex (CEC) or CEC drawer, contains sockets. In the Power10 processor, both SCM (Single Core Modules) and DCM (Dual Core Modules) are used, with a limitation on how tightly hardware can be packed based on manufacturing technology, which in this case is 7nm. Balanced resources of memory and CPU across Scheduler Resource Allocation Domains (SRADs) are preferred. The 'lssrad' AIX command can be used to view LPAR placement. For instance, the following output shows a 9080-HEX (P10-E1080) LPAR with SMT-8, 26 VCPU with an entitlement of 26 placed across four SRADs:

```
REF1    SRAD        MEM      CPU
0
          0   252888.00      0-7 24-31 56-63 88-95 120-127 152-159 176-183
200-207
          1   252883.94      8-15 32-39 64-71 96-103 128-135 160-167 184-191
          2   252961.00      16-23 40-47 72-79 104-111 136-143 168-175 192-
199
          3   202915.06      48-55 80-87 112-119 144-151
```

Memory affinity/memory performance

NUMA - Non-Uniform memory access
Near - Same chip or socket
Far - Same node
Distant - Different node

Understanding the hardware architecture and the placement of chips and CPUs is crucial in determining the near, far, and distant CPU and memory configurations, which affects memory affinity and performance. When a single LPAR is node-contained, it has the lowest latency to CPU, memory, and bus activity. However, as CPUs and memory cross sockets and nodes, additional CPU cycles are required for operations, and there can be additional cycles related to locks on kernel resources. Thus, while locking is essential in a multi-CPU architecture, the NUMA effects should also be considered when selecting resources for a single LPAR.

The output of the 'lssrad' command represents the hardware design and architecture, and the best performance is achieved by LPAR configurations with equal resources distributed between SRADs and balanced memory. The DPO command is critical in balancing resources

for optimal CPU and memory affinity, and it is a firmware function with the Power Hypervisor (PHYP) determining resource placement. Memory affinity refers to having memory as close as possible to the CPU accessing it, but the strictest affinity configuration may limit CPU and memory resource choices. Thus, the balance lies in providing efficient resources to an LPAR for running a workload while having enough capacity for enterprise-class workloads. AIX on power is known for its ability to provide capacity for such workloads, and performance is often a balancing act of providing adequate resources for peak workloads while keeping enough capacity available for computational spikes on demand.

Dynamic Platform Optimizer

The Dynamic Platform Optimizer assists in rebalancing CPU and memory resources that have been allocated to LPARs on a system. Depending on the operation performed by the Dynamic Logical Partition (DLPAR), SRADs may not have assigned CPUs or memory, or CPUs may not have memory, etc. The main purpose of this tool is to reallocate resources to maintain the best affinity for efficient LPAR operations. The output of the lssrad AIX command is the best representation of this process. There are no functional issues associated with less optimal placement. However, higher Cycles per Instruction (CPI) may be required to perform CPU and memory operations.

For more information on dynamic platform optimizer documentation refer to the following:

- IBM Power Systems HMC Implementation and Usage Guide
- Dynamic Platform Optimizer

Adapter placement for the 9080-HEX E1080

Adapter Placement for the 9080-HEX E1080 document provides guidelines for the placement of adapters in the EMX0 PCIe Gen3 I/O expansion drawer for the 9080-HEX E1080 system. The document includes information on the slot priorities and placement rules for the supported adapters.

Power10 is SMT-8 capable

The Power10 processor can support up to 8 simultaneous hardware threads within a CPU core, which is known as SMT-8 or Simultaneous Multi Thread. However, the actual number of SMT threads used depends on the processor load and scheduling policies of the AIX scheduler and Power Hypervisor. The rPerf benchmarking numbers indicate the processor modes that the Power10 processor can run on Single Thread (ST) mode and SMT2 mode where work is dispatched to primary and secondary hardware threads. Similarly, in SMT4 mode, work may be dispatched to up to four SMT threads, including primary, secondary, and tertiary threads. This process is automatic and performed by the system.

Model	Processor / cores	Freq. GHz	Cache L1(KB) PerCore	Inst/Data Cache L2/L3/L4 (MB)/ System	LPAR size# cores	rPerf ST	rPerf SMT2	rPerf SMT4	rPerf SMT8
E1080	p10/240	3.55-4.0	96/64	480/1920/-	60	2,250.8	4,492.7	6,314.4	7,998.6

rPerf reference

The rPerf model estimates commercial processing performance relative to other IBM UNIX systems by using an IBM analytical model that considers characteristics from internal workloads, OLTP and SPEC benchmarks. It is important to note that rPerf is not intended to represent any specific public benchmark results and should not be reasonably used in that way.

rPerf estimates are calculated based on the latest levels of AIX and other pertinent software at the time of system announcement. However, actual performance will vary based on application and configuration specifics. The IBM eServer pSeries 640 is the baseline reference system with a value of 1.0.

While rPerf may be used to approximate relative IBM UNIX commercial processing performance, actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration.

It is important to note that the rPerf methodology used for Power10 processor-based systems is identical to that used for Power9 and Power8 processor-based systems. However, variations in incremental system performance may be observed in commercial workloads due to changes in the underlying system architecture.

For mor information on this topic refer to IBM Power Performance Report.

SMT scheduling

Power10 processor-based systems running in shared processor mode have a default value of 2 for the virtual processor management (VPM) throughput mode. This change was made after rPerf testing showed a linear performance increase between SMT and SMT-2. The advantage for shared pool LPARs is a reduction in physical processor consumption with no loss in performance on Power10.

When migrating a system to or from a Power10 processor-based system, the AIX operating system automatically changes the default value of the throughput mode for the VPM. However, the vpm_throughput_mode will not be changed during migration if the system administrator changed the default on the source system before the LPM. The AIX operating system can select the default value of the vpm_throughput_mode tunable parameter of the schedo command during boot operation based on the type of server on which the LPAR is running. The value selected by the AIX operating system is preserved and used on the destination server.

Note: It is recommended to update the operating system level to 7300-00, 7200-05-03-2147, 7200-04-05-2148, 7100-05-09-2148 or later when migrating to or from a Power10. Without the feature that enables changing the vpm_throughput_mode tunable parameter using the schedo command, updating the operating system levels to 7200-05-00-2037, 7200-04-03-2038, and 7100-05-07-2037 might override the value set for the vpm_throughput_mode tunable parameter.

Note: If the value of vpm_throughput_mode has been changed by the administrator, then that value will be retained after an LPM to a Power10 processor-based system. The only condition where the vpm_throughput_mode will be changed to a new default on Power10 is if it was not changed by the user. This means that it was either not changed or restored to default using the schedo -d flag before the LPM. The system assumes that a user change should be conserved.

The AIX scheduler is optimized to provide the best raw throughput and response time on current Power based systems. To achieve this, dispatching preferentially utilizes the primary SMT thread of each core with the net effect of spreading out workload across as many cores as needed/available.

To enable the new mode using the vpm_throughput_mode parameter, the schedo tuning command can be used as shown.

```
# schedo -p -o vpm_throughput_mode=X
# schedo -d vpm_throughput_mode. (restore default value)

# schedo -h vpm_throughput_mode
Help for tunable vpm_throughput_mode:
Purpose: Sets the scaled throughput mode for processor folding.
Values:
```

```
Default: 0
Range: 0 - 8
Type: Dynamic
Unit: level
```

Tuning:

The throughput mode determines the level of SMT exploitation on each virtual processor core before unfolding another core. Increasing the value of the mode will result in fewer cores being unfolded for a given workload, which leads to higher scaled throughput but lower raw throughput. Disabling the option by setting the value to zero will apply the default mode, which is raw throughput. Enabling raw throughput mode using the newer folding algorithm requires setting the value to one.

- The legacy default mode (vpm_throughput_mode=0) operates in raw throughput mode. The workload is dispatched on the primary SMT thread of each unfolded virtual processor as the load increases.
- Note: On Power10, default mode is vpm_throughput_mode=2.
- The enhanced raw throughput mode (vpm_throughput_mode=1) functions similarly to the default mode, but it considers both load and utilization in folding decision. Using this mode may show lower physical CPU consumption than mode 0, depending on the workload.
- In scaled throughput mode SMT2 (vpm_throughput_mode=2), the dispatcher distributes the workload across the primary and secondary SMT threads of a virtual processor instead of spreading them across primary SMT threads only. This mode results in lower physical CPU consumption compared to raw or enhanced raw throughput mode.
- In scaled throughput mode SMT4 (vpm_throughput_mode=4), the dispatcher distributes the workload across the primary, secondary, and tertiary threads first as the load increases.
- In scaled throughput mode SMT8 (vpm_throughput_mode=8), the dispatcher spreads the workload across all SMT threads of a virtual processor (primary, secondary, tertiary, and quarternary threads). This mode provides the lowest physical CPU consumption but also has the highest application response time and lowest per application thread throughput.

Power management

The Dynamic Power Saver feature enables a system to implement algorithms for adjusting the processor core frequency to optimize system performance, while saving power where applicable, or balancing power and performance. The core frequency may exceed 100% at times. This feature can be set via the ASMI or CIM client. The recommended setting is to favor performance over power savings. To reduce the processor power consumption of the managed

system, enable the power saver mode and set it to Power Saving Mode: Maximum Performance.

There are several Power Saver mode options to choose from:

- Static mode reduces power consumption by decreasing the processor clock frequency and voltage to fixed values. This mode delivers predictable performance while reducing power consumption.
- Dynamic Power Saver mode delivers power savings by varying the processor frequency and voltage based on the utilization of the system processors.
- In Dynamic Power Saver Mode, the system firmware balances performance and power consumption.
- Maximum Performance mode causes the processor frequency to be set at a specified fixed value. You can set the maximum limit of the processor frequency and power consumption of the system.

Note that enabling any of the power saver modes causes changes in processor frequencies, changes in processor use, changes in power consumption, and varying performance.

Memory page sizes

Power10 supports all page sizes as shown in the following table.

This feature remains unchanged from previous Power Architectures.

Page size support by AIX and different System p hardware

Page Size	Required Hardware	Requires User Configuration	Restricted
4 KB	ALL	No	No
64 KB	POWER5+ or later	No	No
16 MB	POWER4 or later	Yes	Yes
16 GB	POWER5+ or later	Yes	Yes

Supported segment page sizes

Segment Base Page Size	Supported Page Sizes	Minimum Required Hardware

Page size support by AIX and different System p hardware

Page Size	Required Hardware	Requires User Configuration	Restricted
4 KB	4 KB/64 KB	POWER6®	
64 KB	64 KB	POWER5+	
16 MB	16 MB	POWER4	
16 GB	16 GB	POWER5+	

As with all previous versions of AIX, the default page size in Power10 is 4 KB. A process will continue to use 4 KB pages unless a user explicitly requests another page size to be used.

Increasing the memory page size from 4K to 64K to 16MB results in improved performance. The gain is typically due to reduced page faults and lower cycles per instruction (CPI) when accessing memory larger than the page size. Most enterprise-class Applications and Databases use memory pages larger than 4K to achieve better performance. To check this, you can use the ‘svmon’ or ‘ps’ command. The ‘ps’ command has options to show the various segment page sizes.

Example: ‘ps -Akz’

```
    PID    TTY  TIME     DPGSZ  SPGSZ TPGSZ  SHMPGSZ  CMD
      260      - 1374:12    4K    64K    4K       4K   wait
    65798      -  0:20      4K    64K    4K       4K   sched
   131336      -  0:00      4K    64K    4K       4K   lrud
   196874      -  0:00      4K    64K    4K       4K   vmptacrt
```

DPGSZ – data segment page size
SPGSZ – stack segment
TPGSZ – text segment
SHMPGSZ – Shared Memory segment

The ‘svmon’ command can display a per-process listing that shows segments with mixed page sizes. A segment can have both 4K and 64K pages.

The LDR_CNTRL shell environment variable can be used to specify a larger segment page size than the default of 4K. Following example command can be put into an application startup script to set the shell environment before the processes are started:

```
export LDR_CNTRL=TEXTPSIZE=64K@DATAPSIZE=64K@STACKPSIZE=64K@SHMPSIZE=64K
```

The ‘ldedit’ command can be used to permanently change the page sizes in a binary executable file:

```
ldedit -btextpsize=64K -bdatapsize=64K -bstackpsize=64K /tmp/helloworld
```

Large page usage to back shared memory segments

To back shared memory segments with large pages in an application, you must specify the ‘SHM_LGPAGE’ and ‘SHM_PIN’ flags in the shmget() function. If large pages are not available, the shared memory segment will be backed by 4 KB pages instead.

The physical memory that is used to back large page shared memory, large page data, and heap segments comes from the large page physical memory pool. Therefore, it is important to ensure that the large page physical memory pool has enough large pages to support shared memory, data, and heap large page usage.

For further information on large page usage refer to Large page usage to back shared memory segments.

Power10 OMI memory

The Power10 processor technology introduces the new OMI DIMMs to access main memory. This allows for increased memory bandwidth of 409 GB/s per socket. The 16 available high-speed OMI links are driven by 8 on-chip memory controller units (MCUs), providing a total aggregated bandwidth of up to 409 GBps per SCM. Compared to the Power9 processor-based technology capability, this represents a 78% increase in memory bandwidth.

To optimize system performance, it is recommended to install memory evenly across all system node drawers and processor sockets. This balancing of memory ensures consistent memory access and typically results in better system performance.

It is advisable to consider future memory additions when choosing the memory feature size during the initial system order, even though maximum memory bandwidth is achieved by filling all the memory slots.

Power10 AIX support

AIX 7.1 TL5 SP3
IBM Power10 systems supports AIX 7.1 only in Power8 compatibility mode.

AIX 7.2 TL6 SP2
IBM Power10 systems supports AIX 7.2 only in Power9 compatibility mode.

AIX 7.3
IBM Power systems extends the scalability of AIX 7.3, supporting a maximum of 240 cores (1920 hardware threads) in a single Power10 LPAR.

VIOS 3.1.3
All levels of VIOS 3.1.3 are supported. Lower VIOS levels may be supported based on the update level. See: System to PowerVM Virtual I/O Server maps.

AIX 7.3 Standard Edition

IBM AIX 7.3 Standard Edition provides several new features and improvements that help enhance system availability, scalability, performance, and flexibility while maintaining binary compatibility to ensure a quick and seamless transition to the new release. Additionally, it is designed to work seamlessly with Power10, offering users a frictionless hybrid cloud experience with faster responses, data protection from core to cloud, streamline insights, and automation.

For more information refer to the following:

- IBM delivers enhanced capabilities with IBM AIX 7.3 Standard Edition for additional information on AIX 7.3 Standard Edition.
- IBM AIX Documentation for resources related to IBM AIX, including installation guides, system administration guides, performance tuning guides, and more.
- IBM AIX operating system for IBM Power for performance, reliability, and security.
- Service and Support Best Practices for Power Systems for the list of technical documents on performance and usability of AIX on Power platform.
- Power10 Performance Best Practices for information on AIX, VIOS and network configuration topics.
- Power10 Performance Quick Start Guide for additional information on Power10 features like Memory Performance, MMA Performance, PowerVM Best Practices, AIX Best Practices, and so on.

GZIP NX hardware acceleration

Power10 servers, as well as Power9, have a GZIP based hardware accelerator supported by AIX 7.2 and 7.3. Each processor chip features an on-chip NX accelerator that offers specialized functions for general data compression, GZIP compression, encryption, and random number generation. AIX uses the zlibNX library, which utilizes the accelerator transparently when available. In comparison to the previous Power9 PHYP interface support, the acceleration on Power10 can be used in user mode.

Depending on the AIX level, you can obtain the zlibNX library from the installation media or the Expansion Pack. It is installed by default if you are using AIX 7.3. The AIX 7.3 default installation includes both the pigz (parallel gzip) open-source command and the AIX zlibNX library. Both transparently use the NX GZIP compression accelerator.

To take advantage of NX GZIP acceleration on Power10 processor-based systems, the LPAR must be in Power9 compatibility mode (not Power9 base mode or Power10 compatibility mode). An additional advantage of Power10 and AIX 7.3 is that the code can run in user mode, which provides less overhead and latency, leading to improved performance.

OpenSSL

The OpenSSL software offers a complete and advanced toolkit for secure communications and general-purpose cryptography. In enterprise-class computing, security is a vital concern that usually demands significant computing power. Encryption is a necessary yet resource-intensive operation. Notably, the Power10 features hardware-assisted encryption to allow easy integration of secure protocols for commercial purposes.

For more information on OpenSSL refer to the following:

- Performance improvement in openssh with on-chip data compression accelerator in power9
- AIX Web Download Pack Programs – Latest OpenSSL

Matrix-Multiply Assist (MMA)

The Power10 architecture includes core hardware Matrix-Multiply Assist (MMA) accelerators, which offer a significant performance boost. There are four units dedicated to MMA acceleration, with each capable of producing a 512-bit result per cycle. This results in a 400% increase in Single and Double precision FLOPS, as well as support for reduced precision AI acceleration. Matrix multiplication is a commonly used compute kernel in various applications, particularly in the emerging fields of machine learning and deep learning.

For more information on MMA refer to the following:

- Matrix-Multiply Assist (MMA) exploitation in OpenBLAS on IBM AIX for MMA based computational strength and data bandwidth to handle the demanding AI inferencing and machine learning (ML) workloads.
- Matrix-Multiply Assist Best Practices Guide on redbooks.ibm.com for Matrix-Mulitply Assist (MMA) function.
- AI Acceleration Capabilities Using MMA video on youtube.com for architecture capabilities to accelerate matrix math operations called MMA.
- A matrix math facility for Power ISA ™ processors for numerical linear algebra operations on small matrices, convolution, and discrete Fourier transform.

Power ISA 3.1

The roots of the Power ISA (Instruction Set Architecture) began at IBM Research. In early 1990, the Power (Performance Optimization With Enhanced RISC) Architecture was first introduced with the RISC System/6000 product family. One important addition to ISA 3.1 is the inclusion of a 32-bit instruction prefix that supports PC-relative addressing, up to 34-bit immediate operands, additional operand fields, and additional opcode space.

The Power10 processor core brings in a variety of architecture capabilities. It implements features like prefix instruction support and MMA, which were introduced in Open Power ISA V3.01. MMA is an on-chip AI acceleration capability that speeds up the matrix multiplication compute.

For additional information refer to the following:

- Power ISA on wiki.raptorcs.com for more information on the Power architecture.

- Architecture innovations in Power ISA v3.01 and Power10 video on youtube.com for additional information on the Power architecture.

Summary

This document highlights the importance of system configuration planning, setup, and deployment when migrating to a Power10 process-based system. It provides insights into commonly encountered performance issues and misconceptions about tuning changes made on AIX. The document covers performance topics related to AIX running on Power10 hardware, as well as AIX-specific performance topics that may apply to recent hardware. The focus is to provide important and applicable performance topics and includes reference links to more detailed documentation. New capabilities provided by the Power10 hardware related to performance enhancements are also discussed.

Get more information

For additional information refer to the following:

- IBM Power S1014, S1022s, S1022, and S1024 Technical Overview and Introduction
- IBM Power E1050 Technical Overview and Introduction
- IBM Power E1080 Documentation
- IBM Power E1080 Technical Overview and Introduction
- Hints and Tips for Migrating Workloads to IBM Power9 Processor Based Systems
- IBM Power Virtualization Best Practices Guide

About the author and acknowledgements

Will Quinn - Power Systems Performance, with over 30 years of experience at IBM.
Focal point for customer performance problem escalations into AIX development.
His primary focus is on solving customer performance problems and driving improvements into the AIX operating system. You can reach Will at quinnw@us.ibm.com.

Acknowledgements

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

Special thanks to key contributors from IBM Power Systems Performance:

Dirk Michel - STSM Power Systems Enterprise Performance Architect, responsible for driving system design requirements for performance, covering the full Power portfolio across enterprise and scale-out platforms. His responsibilities include the full stack performance across hardware and software.

Basu Vaidyanathan – STSM Power System Performance, with focus on Enterprise and Cloud customers. Basu has 25+ years of experience in various areas of the system software stack and his specialties include: Hyper Scale Data Center Solutions, System Performance Analysis, Technology mapping to Industry needs, Virtualization Technology Development, AIX Kernel Development, Debugger Development, and inventing!

Arnold Flores - STSM Power Systems Performance with 30 years of experience in AIX development, debugging, and performance.

Matt Accapadi – STSM, Master Inventor, F.A.S.T (Field Assist Support Team) AIX Performance, Primary focus on AIX performance issues escalated by our customers from field production operations.

IBM Corporation New Orchard Road Armonk, NY 10504

Produced in the
United States of America
May 2023