# IBM Power10 Performance User Guide



**Power10 Performance Quick Start Guides**
**(Power10 QSGs)**
**November 2021**

### Minimum Memory

- For each processor socket, a minimum of 8 of the 16 DIMMs are populated
- In a node, a minimum of 32 of 64 for the DIMMs are populated
- In a 4-Node system, a minimum of 128 of the 256 DIMMs are populated

### DDIMM Plug Rules

- Meet minimum memory allowed (each processor socket a minimum of 8 of the 16 DIMMs are populated)
- All DIMMs under each processor has to be the same capacity
- Feature upgrades will be offered in increments of 4 DDIMM's, all of which have the same capacity.
- The only valid number of DDIMM's plugged into sites connected to a given processor module is 8 or 12 or 16.

### Memory Performance

- System performance improves as the amount of memory is spread across more DDIMM slots. For example, if 1TB is needed in a Node, it is better to have 64 x 32GB DIMMs than to have 32 x 64GB DIMMs.
- Plugging DIMMs that are all the same size will provide the highest performance
- System performance improves as more quads match each other

- System performance improves as more processor DDIMMs match each other
- System performance improves on a multi-drawer system if memory capacity between drawers is balanced.

**Memory Bandwidth**

| DDIMM Capacity | Theoretical MaxBandwidth |
|---|---|
| 32GB, 64 GB (DDR4 @ 3200 Mbps) | 409 GB/s |
| 128GB, 256 GB (DDR4 @ 2933 Mbps) | 375 GB/s |

**Summary**

- For the best possible performance, it is generally recommended that memory be installed evenly across all system node drawers and all processor sockets in the system.  Balancing memory across the installed system planar cards enables memory access in a consistent manner and typically results in better performance for your configuration.
- Though maximum memory bandwidth is achieved by filling all the memory slots, plans for future memory additions should be considered when deciding which memory feature size to use at the time of initial system order.

**P10 Compute & MMA Architecture**

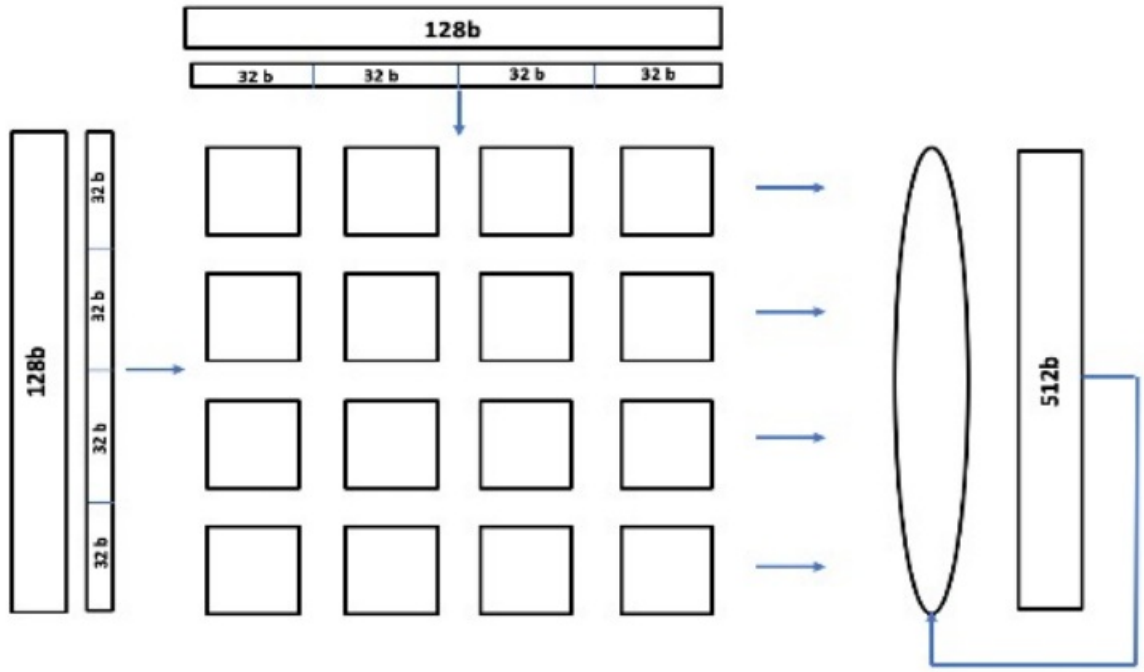
- 2x Bandwidth matched SIMD*
- 8 independent Fixed & Float SIMD engines per Core
- 4 – 32x Matrix Math Acceleration*
- 4 512 bit engine per core = 2048b results / cycles
- Matrix math outer products of Single, Double & Reduced precision.
- MMA Architecture support introduced in POWER ISA v3.1
- Supports SP, DP, BF16, HP, Int-16, Int-8 & Int-4 precision levels.

**P10 MMAApplications & Workload Integration**

- ML & HPC applications with dense linear algebra computations, matrix multiplications, convolutions, FFT can be accelerated with MMA
- GCC version >= 10 & LLVM version >=12 supports MMA through built-ins.
- OpenBLAS, IBM ESSL & Eigen Libraries are already optimized with MMA instructions for P10.
- Easy integration of MMA for enterprise applications, ML frameworks, and Open Community packages via the above BLAS libraries.

**PowerPC Matrix-Multiply Assist Built-in Functions** https://gcc.gnu.org/onlinedocs/gcc/PowerPC-Matrix-Multiply-Assist-Built-in-Functions.html
**Matrix-Multiply Assist Best Practices Guide**
https://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/redp5612.html?Open

| Rank | Operand Type (X,Y) | | | Accumulator | Peak [FL]OPS / cycle | | |
|---|---|---|---|---|---|---|---|
| k | Type | X | $Y^T$ | A | Instruction | Thread | SMT8 Core |
| 1 | Float-64 DP | 4×1 | 1×2 | 4×2 (Fp-64) | 16 | 32 | 64 |
| | Float-32 SP | 4×1 | 1×4 | 4×4 (Fp-32) | 32 | 64 | 128 |
| 2 | Float-16 HP | 4×2 | 2×4 | | 64 | 128 | 256 |
| | Bfloat-16 HP | 4×2 | 2×4 | | | | |
| | Int-16 | 4×2 | 2×4 | | | | |
| 4 | Int-8 | 4×4 | 4×4 | 4×4 (Int-32) | 128 | 256 | 512 |
| 8 | Int 4 | 4×8 | 8×4 | | 256 | 512 | 1024 |




**Virtual Processors**

- The sum of the entitled cores of all shared partitions cannot exceed the number of cores in the shared pool
- Ensure the number of configured virtual processors of any shared partitions on a frame is not more than the number of cores in the shared pool
- Configure the number of virtual processors for a shared partition to sustain peak capacity demand
- Configure the number of entitled cores for a shared partition to average utilization of that partition for better performance
- To ensure better memory and CPU affinity (avoid unnecessary preemptions of the virtual processor), ensure the sum of the entitled cores of all shared partitions close to the number of the cores in the shared pool

**Processor Compatibility Mode**

- There is 2 processor compatibility modes available for AIX: POWER9 and POWER9_base. Default is POWER9_base mode.
- There are 2 processor compatibility modes available for Linux: POWER9 and POWER10 mode. Default is

POWER10 mode.

- After LPM partitions, need to power cycle when changing the processor compatibility mode

**Processor Folding Considerations**

- For share partition running AIX on Power9, the default vpm_throughput_mode = 0, on Power10, the default vpm_throughput_mode = 2. For workloads have long-running jobs, it can potentially help with core usage reduction.
- For dedicated partition running AIX, the default vpm_throughput_mode = 0 on both Power9 and Power10.

**LPAR Page Table Size Considerations**

• Radix page table is supported starting on Power10 running Linux. It can potentially improve workload performance.

**Reference:**
**Hints and tips for Migrating Workload to IBM POWER Systems: https://www.ibm.com/downloads/cas/39XWR7YM**
**IBM POWERVirtualizationBest PracticesGuide: https://www.ibm.com/downloads/cas/JVGZA8RW**

**Ensure OS level is current**
Fix Central provides the latest updates for AIX, IBM i, VIOS, Linux, HMC and F/W. 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. If you cannot move up to the recommended level, then refer to the Known Issue section of the Hints & Tips for migrating Workload to the IBM POWER10 Processor-Based Systems document.
**AIX CPU utilization**
On POWER10, the AIX OS system is optimized for best raw throughput at higher CPU usage when running with dedicated processors. When running with shared processors, the AIX OS system is optimized to reduce CPU usage (pc). If the customer requires to further reduce CPU usage (pc), use the schedule tunable pm_throughput_mode to tune the workload and evaluate the benefits of raw throughput vs. CPU usage.
**NX GZIP**
To take advantage of NX GZIP acceleration on POWER10 systems the LPAR must be in POWER9 compatibility mode (not POWER9_base mode) or POWER10 compatibility mode.
**IBM i**
Ensure the IBM I operating system level is current. Fix Central provides the latest updates for IBM I, VIOS, HMC, and firmware. **https://www.ibm.com/support/fixcentral/**
**Firmware**
Ensure the system firmware level is current. Fix Central provides the latest updates for IBM I, VIOS, HMC, and firmware. **https://www.ibm.com/support/fixcentral/**
**Memory DIMMs**
Follow proper memory plug-in rules. If possible, fully populate memory DIMM slots and utilize similar-sized memory DIMMs.
**Processor SMT level**
To take full advantage of the performance of Power10 CPUs, we recommend clients utilize the IBM i default processor multitasking settings, which will maximize the SMT
level for the LPAR configuration.
**Partition Placement**
Current FW levels ensure optimal placement of the partitions. However, if frequent DLPAR operations are executed on partitions on the CEC, it is recommended the use DPO
to optimize placement.
**Virtual Processors – shared vs dedicated processors**
Utilize dedicated processors for optimal partition level performance.
**EnergyScale**
For the best CPU processor speed, ensure that Maximum Performance is set (default for IBM Power E1080). This

setting is configurable in the ASMI.

**Storage and Networking I/O**

VIOS provides flexible storage and networking functionality. For the best possible performance, utilize native IBM i interfaces for I/O.

**More comprehensive information**

Refer to link: IBM I on Power – Performance FAQ **https://www.ibm.com/downloads/cas/QWXA9XKN**

The enterprise Linux operating system (OS) is a solid foundation for your hybrid cloud infrastructure and for scale-up enterprise software solutions. Recent releases are optimized for best-in-class Power10 Enterprise systems

**Power10**

- SLES15SP3, RHEL8.4 support Power10 native mode
- Compass-mode support to allow clients to migrate from older generation Power systems ( P9 and P8 )
- Default Radix translation support in Power10 mode
- Significant improvement in encryption performance

**Linux + PowerVM**

- Support for PowerVM enterprise features: LPM, Shared CPU Pools, DLPAR
- Innovative solutions: SAP HANA future application growth with 4PB virtual address space
- Reduce time to reload the data: Virtual PMEM support for SAP HANA
- World-class Support & Service

**Supported distros:**

- Starting with Power9 only RedHat and SUSE are supported in PowerVM partitions
- Detailed info on distro support matrix covering older generation HW

**LPM Support:**

- Move Linux logical partitions from older generation Power systems with near-zero application downtime
- Reference: LPM Guide and related information

**Power Specific Packages:**

- PowerPC-utils package: Contains utilities for the maintenance of IBM PowerPC LPARs. Available as part of the distro.
- Advance Toolchain for Linux on Power: Contains latest compilers, runtime libraries.

**Best practices :**

- RHEL provides predefined tunings as part of the tuned service.
- Refer to the latest SAP notes for recommended OS settings for SAP applications. Typically tuned is used in RHEL and capture or sapconf in SLES
- Frequency is managed by the PowerVM. Reference: Energy Management
- Starting Power8 Huge Dynamic DMA Window helps improve I/O performance.

- Starting Power9 24×7-Monitoring is integrated with the perf tool. Allows monitoring the entire system.
- Ensure the system firmware level is current.
- lparnumascore from PowerPC-utils shows the LPAR's current affinity score. DPO can be used to improve the LPAR affinity score.

**More reads :**

- SLES for Power and some compelling features.
- Get started with Linux on Power Systems, Linux on Power Systems servers
- Enterprise Linux community
- IBM Power systems support various network adapters of different speeds and numbers of ports.
- If you are using the same network adapters as your previous system, initially, the same tuning should be used on the new system.
- Most Ethernet adapters support multiple receive and transmit queues whose buffer size can be varied to increase max packet count.
- The default queue settings are different with different adapters and may not be optimal to achieve maximum message rates in a client-server model.
- Using additional queues will increase the CPU usage of the system; so optimal queue setting for a specific workload should be used.

**Higher speed adapter considerations**

- Higher speed networks with 25 GigE and 100 GigE network adapters require multiple parallel threads and tuning of driver attributes.
- If it is a Gen4 adapter, make sure the adapted is seated on a Gen4 slot.
- Additional functions such as compression, encryption, and duplication can add latency

**Changing queue settings in AIX**
To change the number of receive/transmit queues in AIX

- ifconfig enX detach down
- chdev -l entX -a queues_rx=<value> -a queues_tx=<value>
- chdev -l enX -a state=up

**Changing queue settings in Linux**
To change the number of queues in Linux ethtool -L ethX combined <value>

**Changing queue size in AIX**

- ifconfig enX detach down
- chdev -l entX -a rx_max_pkts =<value> -a tx_max_pkts =<value>
- chdev -l enX -a state=up

**Changing queue size in LinuxP:** ethtool -G ethX rx <value> tx <value>

**Virtualization**

- Virtualized networking is supported in the form of SRIOV, vNIC, vETH. 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
- IBM PowerVM Best Practices can be very helpful in VIOS sizing
- If you are using the same storage adapters 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 tuning should be performed.
- If the 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 –
- Changing from Direct Attached Storage (DAS or internal) to Storage Area Network (SAN) or Network Attached Storage (NAS) (or external storage) can increase latency.
- Additional functions such as compression, encryption and deduplication can add latency.
- Reducing the number of Storage LUNs can reduce resources in the server needed to support required throughputs.
- Refer to tuning or setup guides for the new devices to understand these impacts.'
- Virtualization does add latency and can reduce throughput compared to native I/O. Besides the backend hardware, ensure VIOS memory and CPU
- Moving to higher-speed virtualized adapters in VIOS will require adjusting the VIOS configuration in CPUs and memory. IBM PowerVM Best Practices can be very helpful in VIOS sizing.

**Tuning guidelines** – please refer to the IBM Knowledge Center for AIX and Linux guidelines.

**PCIe3 12 GB Cache RAID + SAS Adapter Quad-port 6 Gb x8 Adapter Linux:**

- **https://www.ibm.com/docs/en/power9/9223-42H?topic=availability-ha-asymmetricaccess-optimization**
- **https://www.ibm.com/docs/en/power9/9223-42H?topic=linux-common-sas-raidcontroller-tasks**

**AIX:**

- **https://www.ibm.com/docs/en/power9/9223-42H?topic=aix-multi-initiator-highavailability**
- **https://www.ibm.com/docs/en/power9/9223-42H?topic=aix-common-controller-diskarray-management-tasks**

**IBM**

- **https://www.ibm.com/docs/en/power9/9223-42H?topic=configurations-dual-storageioa-access-optimization**
- **https://www.ibm.com/docs/en/power9/9223-42H?topic=i-common-controller-diskarray-management-tasks**

**PCIe3 x8 2-port Fibre Channel (32 Gb/s) Adapter**

- https://www.ibm.com/docs/en/aix/7.2?topic=iompio-device-attributes
- https://www.ibm.com/docs/en/power9?topic=channel-npiv-multiple-queue-support

**Additional AIX tuning for performance:**

- SCSI over Fiber Channel (MPIO): set multipath algorithm to round_robin for every disk
- NVMe over Fiber Channel: set can attribute to 7 for every NVMe over Fiber Channel Dynamic controller created during the discovery phase

**NVMe Adapter AIX tuning for performance**
Set can attribute to 8 for each NVMe device
**IBM's next-generation C/C++/Fortran compilers that combine IBM's advanced optimizations with the open-source LLVM infrastructure**


| | |
|---|---|
| **LLVM**<br>Greater currency for C/C++ language<br>Faster build speed<br>Community common optimizations<br>Various LLVM-based utilities | **IBM optimizations**<br>Full exploitation of Power architectu<br>Industry-leading advanced optimiz<br>World-class Support & Service |

**Availability**

- 60-day no-charge trial: download from Open XL product page
- Obtain IBM world-class Service & Support through flexible licensing options, from dual-pipe (AAS and PA)
- Perpetual license (per Authorized User or per Concurrent User)
- Monthly license (per Virtual Process Core): target cloud use cases, e.g., on PowerVR instance

**Recommended performance tuning options**

| Optimization Level | Usage rec |
|---|---|
| -O2 and -O3 | Typical sta |
| Link time optimization: -flto (C/C++), -qlto (Fortran) | For worklo |
| Profile guided optimization: -fprofile-generate, -fprofile-use (C/C++)<br>-qprofile-generate, -qprofile-use (Fortran) | For worklo |

For more info please visit: **https://www.ibm.com/docs/en/openxl-c-and-cpp-aix/17.1.0**
**https://www.ibm.com/docs/en/openxl-fortran-aix/17.1.0**

**Full Power10 architecture exploitation with Open XL 17.1.0**

- New compiler option '–mcpu=pwr10' to generate code exploiting Power10 instructions and also automatically tune the optimizations for Power10
- New builtin functions to unlock new Power10 functionalities, e.g., Matrix Multiply Accelerator (MMA)
- New MASS SIMD and vector libraries were added for Power10. All MASS library functions (SIMD, vector, scalar) tuned for Power10 (also Power9).

**Note:** Applications compiled with earlier versions of XL Compilers (e.g., XL 16.1.0) to run on previous Power processors will run compatibly on Power10.
**Binary Compatibility on AIX**
**Note:** XL C/C++ for AIX 16.1.0 already introduced a new invocation xlclang++which leverages the Clang front-end from LLVM project ü C++ objects built with xlC for

- AIX (based on IBM's own front-end) are not binary compatible with C++ objects built with xlclang++ 16.1.0 for AIX
- C++ objects built with xlclang++ 16.1.0 for AIX will be binary compatible with new Open XL C/C++ for AIX 17.1.0
- C compatibility is maintained across all AIX compilers (earlier XL versions for AIX, Open XL C/C++ for AIX 17.1.0)
- Fortran compatibility is maintained between the earlier XLF version for AIX and Open XL Fortran for AIX 17.1.0

**Availability**
The GCC compilers are available on all Enterprise Linux distributions and on
**AIX.**

- The installed GCC version is 8.4 on RHEL 8 and 7.4 on SLES 15. RHEL 9 is expected to ship GCC 11.2.
- There are several ways to obtain a sufficiently recent version of GCC when the default compilers for the distribution are too old to support Power10.
- Red Hat supports the GCC Toolset [1] for this purpose.
- SUSE provides the Development Tools Module. [2]
- IBM provides the latest compilers and libraries via the Advance Toolchain. [3]

**IBM Advance Toolchain**

- The Advance Toolchain provides Power-optimized system libraries along with the compilers, debuggers, and

other tools.
- Building code with the Advance Toolchain can produce the most highly optimized code possible on the latest processors.

**Languages**

- C (GCC), C++ (g++), and Fortran (gfortran), along with others such as Go (GCC), D (GDC), and Ada (gnat).
- Only GCC, g++, and gfortran are usually installed by default.
- The golang compiler [4] is the preferred alternative for building Go programs on Power.

**Compatibility and New Features on Power10**

- Applications compiled with earlier versions of GCC to run on POWER8 or POWER9 processors will run compatibly on Power10 processors.
- GCC 11.2 or later is recommended to exploit all new features available in Power ISA 3.1 and implemented in Power10 processors.
- GCC 11.2 provides access to the Matrix Multiply Assist (MMA) feature provided by Power10 processors. [5]
- MMA programs can be compiled using any of the GCC, LLVM, and Open XL compilers, provided you use sufficiently recent releases.

**IBM Recommended and Supported Compiler Flags [6]**

| | |
|---|---|
| -O3 or -East | Aggressive optimization. -East is essentially equivalent to -O3 -fast-math, which also relaxes restrictions on IEEE floating-point arithmetic. |
| -mcpu=powern | Compile using instructions supported by the Power processor. For example, to use instructions available only on Power10, select -mcpu=power10. |
| -to | Optional. Perform “link-time” optimization. This optimizes code across function calls where the caller and called functions exist in different compilation units, and can often provide a significant performance boost. |
| -unroll-loops | Optional. Perform more aggressive duplication of loop bodies than the compiler normally would. Generally, you should omit this, but on some codes, this can provide better performance. |

**Note:**
Although -mcpu=power10 is supported as early as GCC 10.3, GCC 11.2 is preferred because earlier compilers don’t support every feature implemented in the Power10 processors. Also, objects created using -mcpu=power10 will not run on POWER9 or earlier processors! However, there are ways to create code that is optimized for different processor versions. [7]

[1] Red Hat: Using GCC Toolset.
**https://access.redhat.com/documentation/enus/red_hat_enterprise_linux/8/html/developing_c_and_cpp_applications_in_rhel_8/gcc-toolset_toolsets**.
[2] SUSE: Understanding the Development Tools Module. **https://www.suse.com/c/suse-linux-essentialswhere-are-the-compilers-understanding-the-development-tools-module**/.
[3] Advance Toolchain for Linux on IBM Power Systems.
**https://www.ibm.com/support/pages/advancetoolchain-linux-power**.
[4] Go Language. https://golang.org. [5] Matrix-Multiply Assist Best Practices Guide. **http://www.redbooks.ibm.com/redpapers/pdfs/redp5612.pdf**
[6] Using the GNU Compiler Collection. **https://gcc.gnu.org/onlinedocs/gcc.pdf**

[7] Target-Specific Optimization with the GNUIndirect Function Mechanism.
https://developer.ibm.com/tutorials/optimized-libraries-for-linux-on-power/#target-specific-optimization-
 with-the-gnu-indirect-function-mechanism.
Java applications can seamlessly take advantage of new P10 ISA features on operating systems running in P10 mode by using the Java runtime versions listed below or newer:

**Java 8**

- IBM SDK 8 SR6 FP36
- IBM Semeru Runtime Open Edition 8u302: openj9-0.27.1

**Java 11**

- IBM Semeru Runtime Certified Edition 11.0.12.1: openj9-0.27.1
- IBM Semeru Runtime Open Edition 11.0.12.1: openj9-0.27.1

**Java 17 (drivers may not be available yet)**

- IBM Semeru Runtime Certified Edition 17: openj9-0.28
- IBM Semeru Runtime Open Edition 17: openj9-0.28
- OpenJDK 17

**Performance tuning references:**
IBM WebSphere Application Server Performance Cookbook
**Page Size**
The general recommendation for most Oracle databases on AIX is to utilize 64KB page size and not 16MB page size for the SGA. Typically, 64 KB pages yield nearly the same
performance benefit as 16 MB pages without special management.
**TNS Listener**
Oracle 12.1 database and later releases by default will use 64k pages for text, data, and stack. However, for the TNSLISTENER it still uses 4k pages for text, data, and stack. To
enable 64k pages for the listener uses the export command prior to starting the listener process. Note that running in an ASM based environment that the listener runs out of
GRID_HOME and not ORACLE_HOME.
The documentation for the “strictly setenv” command changed in 12.1 or later releases. The -t or -T was removed in favor of -env or -envs. In the Oracle Listener environment set and export:
– LDR_CNTRL=DATAPSIZE=64K@TEXTPSIZE=64K@STACKPSIZE=64K <tnslistener user >-
VMM_CNTRL=vmm_fork_policy=COR (add the ‘Copy on Read’ command)
**Shared syntax**
The LDR_CNTRL=SHARED_SYMTAB=Y setting does not need to be specifically set in 11.2.0.4 or later releases. The compiler linker options take care of this setting and no longer need to be specifically set. It is not recommended to have LDR_CNTRL=SHARED_SYMTAB=Y specifically set in 12c or later releases.
**Virtual Processor Folding**
This is a critical setting in a RAC environment when using LPARs with processor folding enabled. If this setting is not adjusted, there is a high risk of RAC node evictions under light database workload conditions. Scheda -p -o vpm_xvcpus=2
**VIOS & RAC Interconnect**
A dedicated 10G (i.e, 10G Ethernet Adapter) connection is recommended as a minimum to provide sufficient bandwidth for cluster timing-sensitive traffic. RAC cluster traffic – interconnect traffic should be dedicated and not shared. Sharing of interconnect can cause timing delays leading to node hang/eviction issues.
**Network Performance**
This is a long-standing network-tuning suggestion for Oracle on AIX, although the default remains at 0. TCP Setting of rfc1323=1
**More comprehensive information**
Refer to link: Managing the Stability and Performance of current Oracle Database versions running AIX on Power

Systems including POWER9
https://www.ibm.com/support/pages/node/6355543

**General**

- Use SMT8 mode
- Use dedicated CPU LPARs

**Db2 Warehouse**

- Ensure that a high-speed private network exists between all nodes
- Limit MLN configuration to one node per socket

**CP4D**

- Use PCIe4 for OCP nodes network
- Prior to OCP 4.8, set kernel parameter slub_max_order=0

**Db2 Best Practices**
https://www.ibm.com/docs/en/db2/11.5?topic=overviews-db2-best-practices

**Network**

- For pod network, use private network based on native SRIOV if LPM is not required, otherwise, use VNIC
- For applications that require high bandwidth or low latency, consider using the SR-IOV Network Operator to assign VF directly to a pod
- For services in need of a low timeout, configure the default timeouts for an existing route
- Adjust the desired MTU size of OCP's cluster network

**Operating system**

- Consider increasing the u-limits within the CoreOS Post-install changes
- Refer to the minimum OCP installation requirements for Power platform OCP4.8 installation on Power

**Deployment**

- When deploying applications, note that one vCPU is equivalent to one physical core when simultaneous multithreading (SMT), or hyperthreading, is not enabled. When SMT is enabled, a VCPU is equivalent to a hardware thread.
- Refer to minimum sizing guidelines for workers & master nodes Minimum resource requirements
- Allocate a separate dedicated storage to the built-in container image registry
- Use the following sizing guidelines for OCP's main directories main directories that OpenShift Container Platform components write data to.

Contents



## Documents / Resources

| | |
|---|---|
|  | IBM Power10 Performance [pdf] User Guide<br>Power10, Performance, Power10 Performance |