TUNING GUIDE AMD EPYC 7003

Workload

Publication	57011
Revision	3.0
Issue Date	Mar, 2022

Date	Version	Changes
Mar, 2021	2.0	Initial public release
Mar, 2022	3.0	Added AMD 3D V-Cache™ information

Audience

This tuning guide describes best practices for optimizing performance using the Data Plane Development Kit (DPDK). It is intended for a technical audience such as application architects, production deployment, and performance engineering teams with:

- A background in configuring servers.
- Administrator-level access to both the server management Interface (BMC) and the OS.
- Familiarity with both the BMC and OS-specific configuration, monitoring, and troubleshooting tools.

Authors

Mark Baker and Jesse Rangel

Note: All of the settings described in this Tuning Guide apply to all AMD EPYC 7003 Series Processors with or without AMD 3D V-Cache™ except where explicitly noted otherwise.

Table of Contents

Chapter 1 Introduction

Default BIOS options generally produce the best overall performance for generic workloads, but these defaults may not be optimal for a specific workload. AMD continually tests various workloads; this tuning guide discusses BIOS options to deliver both maximum performance and performance-per- watt (power efficiency).

- "BIOS Options and Their Benefits" on page 3 lists various BIOS options and the potential benefit of modifying each one.
- "AMD EPYC 7003 BIOS Settings by Workload" on page 13 presents sample workloads and recommended BIOS settings. Keep in mind that these BIOS settings are not "one size fits all" because your specific workload(s) are not identical to synthetic benchmarks.

Note: Not all platforms support all of the BIOS settings described in this tuning guide. Please contact your platform vendor if you cannot see one or more needed settings.

1.1 AMD EPYC 7003 Series Processors

AMD EPYC 7003 Series Processors are built with the leading-edge "Zen 3" core and AMD Infinity Architecture. The AMD EPYC SoC offers a consistent set of features across 8 to 64 cores. Each 3rd Gen EPYC processor consists of up to eight Core Complex Dies (CCD) and an I/O Die (IOD). Each CCD contains one CCX, meaning that each CCD contains up to 8 "Zen 3" cores. The CCDs connect to the I/O Die (IOD) to access memory, I/O, and each other via AMD Infinity Fabric™ technology. 3rd Gen AMD EPYC processors support up to 8 memory channels, 4 TB of high-speed memory per socket, and 128 lanes of PCIe® Gen 4.

3rd Gen AMD EPYC Series processors are built with the following specifications:

3rd Gen AMD EPYC 7003 Series Processors	
Process technology	7nm
Max Processor speed	4.1 GHz
Max number of cores	64
Max memory speed	3200 MT/s
Max memory capacity	4 TB per socket
Peripheral Component Interconnect	128 lanes (max) PCIeGen4

Table 1-1: Table 1 AMD EPYC™ 7003 Series Processors

Some AMD EPYC™ 7003 Series Processors introduce AMD's new 3D V-Cache die stacking technology that enables denser, more efficient chiplet integration. AMD 3D Chiplet architecture stacks L3 cache tiles vertically to provide 768 MB of L3 cache per socket up with to 96MB of L3 cache per CCD, while still providing socket compatibility with existing AMD EPYC 7003 Series Processors. Applications that take advantage of AMD 3D V-cache can see significant performance gain and lower overall TCO..

See *Overview of AMD EPYC™ 7003 Series Processors Microarchitecture* (available from AMD EPYC Tuning Guides) to learn more about the AMD EPYC 7003 Series Processor microarchitecture.

1.2 Operating Systems

AMD recommends using the latest available OS version. Please see AMD EPYC™ 7003 Series Processors Minimum Operating System (OS) Versions for detailed OS version information.

Chapter 2 BIOS Options and Their Benefits

2.1 Infinity Fabric Settings

This section discusses BIOS settings related to AMD Infinity Fabric technology.

2.1.1 Link Speed

Lowering the link speed decreases cross-socket bandwidth and increases cross-socket latency but can also save uncore power (CPU power not consumed by the cores) to either:

- Increase core frequency.
- Reduce overall power consumption.

Setting	Options
xGMI Link Max Speed	• 10.667 Gbps • 13 Gbps • 16 Gbps • 18 Gbps (not supported on all systems)

Table 2-1: Link speed settings

2.1.2 Dynamic Link Width Management (DLWM)

xGMI Dynamic Link Width Management saves power during periods of low socket-to-socket data traffic by reducing the number of active xGMI lanes per link from 16 to 8. Latency may increase in some scenarios involving low-bandwidth, latency-sensitive traffic as the processor transitions from a low-power xGMI state to full-power xGMI state. Setting **xGMI Link Width Control** to **Manual** and specifying a **Force Link Width** eliminates any such latency jitter. Applications that are not sensitive to both socket-to-socket bandwidth and latency can use a forced link width of 8 (or 2 on certain platforms) to save power, which can divert more power to the cores for boost.

Setting	Options
xGMI Link Width Control	• **Auto:** Hide the **Max Link Width** and **Force Link Width** control options. • **Manual:** Show **Max Link Width** and **Force Link Width** control options.
xGMI Max Link Width	• **0:** Max width x8, min width x8 (x2 on certain platforms). • **1:** Max width x16, min width x8 (x2 on certain platforms).
xGMI Max Link Width Control	• **Auto:** Hide the **xGMI Max Link Width** control. • **Manual:** Show the **xGMI Max Link Width** control.

Table 2-2: DLWM settings

xGMI Force Link Width Enable	• **Unforce:** Use automatic xGMI Link Width selection. • **Force:** Use the xGMI Force Link Width link width.
xGMI Force Link Width	• **0:** Use width x2 (not supported on all platforms). • **1:** Use width x8. • **2:** Use width x16.

Table 2-2: DLWM settings (Continued)

2.1.3 Power States

Enable or disable Algorithm Performance Boost (APB). By default, the AMD Infinity Fabric selects between a full- and low-power fabric clock and memory clock based on usage. Latency may increase in some scenarios involving low-bandwidth, latency-sensitive traffic as the processor transitions from low to full power. Setting **APBDIS** to 1 (APB disabled) and specifying a fixed Infinity Fabric P-state of 0 forces the AMD Infinity Fabric and memory controllers into full-power mode and eliminates latency jitter. Setting a fixed AMD Infinity Fabric P-State of 1 on certain CPU OPNs and memory population options reduces both memory latency and memory bandwidth, which may benefit applications that are sensitive to memory latency.

Setting	Options
APB Disable (APBDIS)	• **0:** Dynamically switch Infinity Fabric P-state based on link usage. • **1:** Enable fixed Infinity Fabric P-state control.
Fixed SOC P-State	• **P0:** Highest-performing Infinity Fabric P-state. • **P1:** Second-highest-performing Infinity Fabric P-state. • **P2:** Third-highest-performing Infinity Fabric P-state. • **P3:** Lowest-performing Infinity Fabric power P-state.

Table 2-3: Power state settings

2.1.4 C-States

Much like CPU cores, the AMD Infinity Fabric can enter lower-power states while idle, but a delay occurs when transitioning back to full-power mode that causes some latency jitter. Disabling this feature for workloads requiring low latency and/or bursty I/O will increase both performance and power consumption.

Setting	Options
DF C-states	• **Disabled:** Prevent the AMD Infinity Fabric from entering a low-power state when the processor has entered Cx states. • **Enabled:** Allow the AMD Infinity Fabric to enter a low-power state when the processor has entered Cx states.

Table 2-4: C-state settings

2.2 NUMA and Memory Settings

This section describes NUMA- and memory-related BIOS settings.

2.2.1 ACPI SLIT and SRAT

This setting controls automatic or manual generation of distance information in the ACPI System Locality Information Table (SLIT) and NUMA proximity domains in the System Resource Affinity Table (SRAT). Some hypervisors and operating systems do not perform L3-aware scheduling, and some workloads will benefit from having the L3 declared as a NUMA domain. In dual-socket systems, the remote socket distance can affect memory allocation decisions. Setting this to a value of at least 32 (32 recommended) may improve scheduling of lightly-threaded workloads. Setting this to a value less than 32 (22 recommended) may improve scheduling of heavily-threaded workloads. In general:

- If a workload spans two sockets, then set the distance to < 32.
- If the workload can be confined to a socket, then set the distance to 32.

Setting	Options
ACPI SRAT L3 Cache as NUMA Domain	• **Disable:** Do not report each L3 cache to the OS as a NUMA domain. • **Enable:** Report each L3 cache to the OS as a NUMA domain.
ACPI SLIT Distance Control	• **Auto:** Use default remote- and same-socket distances • **Manual:** Enable remote- and same-socket distance controls
ACPI SLIT Remote Relative Distance	• **Near:** Let BIOS select default values that describe remote cores as relatively close to each local core. • **Far:** Let BIOS select default values that describe remote cores as relatively far away from each local core.
ACPI SLIT <various> Distance	• Enabled when **ACPI SLIT Distance Control** set to **Manual**. • Each item can be set to a value from 10–255 decimal to specify the relative distance of domains.

Table 2-5: ACPI SLIT and SRAT settings

2.2.2 NUMA Nodes per Socket (NPS)

This setting enables a trade-off between minimizing local memory latency for NUMA-aware or highly parallelizable workloads vs. maximizing per-core memory bandwidth for non-NUMA-friendly workloads. NPS2 and/or NPS4 may not be an option on certain OPNs or with certain memory populations.

- **NPS1:** The default configuration (one NUMA domain per socket) is recommended for most workloads. All eight memory channels are interleaved.
- **NPS2:** Every four memory channels are interleaved with each other. This may provide a compromise between memory latency and memory bandwidth when using 200 Gbps network adapters.
- **NPS4:** Every pair of memory channels is interleaved. This is recommended for HPC and other highly-parallel workloads.

This setting is independent of **ACPI SRAT L3 Cache as NUMA Domain** in "ACPI SLIT and SRAT" on page 5. Enabling **ACPI SRAT L3 Cache as NUMA Domain** means this setting will determine the memory interleaving granularity.

Setting	Options
NUMA Node Per Socket	• **NPS0:** Interleave memory accesses across all channels in both sockets (not recommended). • **NPS1:** Interleave memory accesses across all eight channels in each socket and report one NUMA node per socket unless **L3 Cache as NUMA** is enabled. • **NPS2:** Interleave memory accesses across groups of four channels (ABCD and EFGH) in each socket and report two NUMA nodes per socket unless **L3 Cache as NUMA** is enabled. • **NPS4:** Interleave memory accesses across pairs of two channels (AB, CD, EF, and GH) in each socket and report four NUMA nodes per socket unless **L3 Cache as NUMA** is enabled.

Table 2-6: NPS settings

2.2.3 Memory Clock Speed

By default, the 3rd Gen AMD EPYC processor BIOS runs at the maximum clock frequency allowed by the platform and DIMM. This configuration allows maximum memory bandwidth and lowest latency for the processor. Lowering the memory clock speed reduces memory controller power consumption and allows the rest of the SoC to consume more power, thereby potentially boosting performance elsewhere for certain workloads.

Setting	Options
Memory Clock Speed	• **Auto:** Determine maximum memory speed based on SPD information from populated DIMMs and platform memory speed support. • **Values 400 MHz–1600 MHz:** run the DRAM memory clock at the specified speed (The DRAM memory clock is half of the DDR rate.)

Table 2-7: Memory clock settings

2.2.4 Transparent Secure Memory Encryption (TSME)

This feature provides hardware memory encryption of all data stored on system DIMMs that is invisible to the OS and slightly increases memory latency.

Setting	Options
TSME	• **Auto / Disabled:** Disable transparent secure memory encryption. • **Enabled:** Enable transparent secure memory encryption.

Table 2-8: TSME settings

2.3 Power Efficiency Settings

2.3.1 Core Clock Dynamic Power Management (CCLK DPM)

Enabling this feature causes the SoC Efficiency mode to maximize the performance-per-watt by using a dynamic power management algorithm to opportunistically reduce the core clocks. This algorithm is targeted at throughput-based server workloads that exhibit a stable load below the SoC maximum capabilities. The default **Auto** setting maximizes SoC performance.

Setting	Options
EfficiencyModeEn	• **Auto:** Optimize core clock dynamic power management for performance. • **Enabled:** Optimize core clock dynamic power management for power efficiency.

Table 2-9: CLK DPM settings

2.3.2 Power vs. Performance Determinism Settings

The **Determinism** slider selects between:

- **Performance (default for most OPNs):** Uniform performance across identically configured systems in a datacenter. Set cTDP and PPL to the same value, as described in "Processor Cooling and Power Dissipation Limit Settings" on page 8.
- **Power:** Maximum performance of any individual system with varying performance across the datacenter.

Setting	Options
Determinism Control	• **Auto:** Hide the **Determinism Slider** control. • **Manual:** Show the **Determinism Slider** control.
Determinism Slider	• **Auto:** Determined by OPN fusing. • **Power:** Ensure maximum performance levels for each CPU in a large population of identically-configured CPUs by only throttling CPUs when they reach the same cTDP. • **Performance:** Ensure consistent performance levels across a large population of identically-configured CPUs by throttling some CPUs to operate at a lower power level.

Table 2-10: Power/performance settings

2.3.3 Processor Cooling and Power Dissipation Limit Settings

Configurable Thermal Design Power (cTDP) allows modifying the CPU cooling limit and the Package Power Limit (PPL) allows modifying the CPU Power Dissipation Limit. Many platforms configure cTDP to the maximum CPU-supported value. Most platforms also set the PPL to the same value as the cTDP. Both values must be identical when using **Performance** determinism, as described in "Power vs. Performance Determinism Settings" on page 7. If you are using **Power** determinism, then you can reduce system operating power by setting PPL to a value lower than cTDP. The CPU will control CPU boost to keep socket power dissipation at or below the specified PPL.

Setting	Options
cTDP Control	• **Manual:** Set custom configurable TDP. • **Auto:** Use platform- and OPN-default TDP.
cTDP	• **Values 85–280:** Set configurable TDP, in watts.
Package Power Limit Control	• **Manual:** Set customized PPL. • **Auto:** Use platform- and OPN-default-PPL.
Package Power Limit	• **Values 85–280:** Set PPL, in watts.

Table 2-11: cDTP settings

2.3.4 ACPI—Collaborative Processor Performance Control (CPCC)

Enabling CPCC allows the OS to help maintain energy efficiency by controlling when and how much turbo can be applied. ACPI 5.0 introduced this feature. Not all operating systems support CPCC. Microsoft began supporting CPCC with Windows® Server® 2016.

Setting	Options
CPCC	• **Disabled:** Disabled. • **Enabled:** Allow the OS to make performance/power optimization requests using ACPI CPPC.

Table 2-12: CPCC settings

2.4 Processor Core Settings

2.4.1 Cache Prefetchers

Most workloads benefit from the L1 & L2 Stream Hardware prefetchers gathering data and keeping the core pipeline busy, but some workloads are very random in nature. These workloads will perform better when one or both prefetchers are disabled. Both prefetchers are enabled by default.

Setting	Options
L1 Stream HW Prefetcher	• **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.
L2 Stream HW Prefetcher	• **Disable:** Disable prefetcher. • **Enable:** Enable prefetcher.

Table 2-13: Cache prefetecher settings

2.4.2 Symmetric Multithreading (SMT) Settings

Enabling SMT causes neutral to negative performance impacts on some workloads, especially HPC. Also, some application licenses count the number of hardware threads enabled instead of the physical core count. It may therefore be best to disable SMT on your AMD EPYC 7003 Series Processor.

Some operating systems lack the x2APIC support required to support more than 255 threads. Disable SMT if you are running a non-x2APIC OS in a system with dual 64-core processors.

Setting	Options
SMT Control	• **Disable:** Single hardware thread per core. • **Auto:** Two hardware threads per core.

Table 2-14: SMT settings

2.4.3 Core Boost Frequency Settings

This setting limits the maximum boost frequency but does not set a fixed frequency. Some workloads don't need maximum core frequency to achieve acceptable performance. Limiting the maximum core boost frequency can reduce power consumption. The SoC will not exceed the maximum algorithm-allowable frequency if **BoostFmax** is set too high. Actual boost performance depends on many factors, including the other settings discussed in this tuning guide.

Setting	Options
BoostFmaxEn	• **Manual:** Use specified BoostFmax setting. • **Auto:** Use default BoostFmax setting.
BoostFmax	• **Values 0x0–0xFFFFFFFF:** fMax frequency limit to apply to all cores in MHz

Table 2-15: Core boost settings

2.5 I/O Settings

2.5.1 APIC Settings

Interrupt delivery is generally faster when using x2APIC compared to the legacy xAPIC mode, but not all operating systems include AMD x2APIC support. AMD recommends this mode if your OS supports it, including for configurations with fewer than 256 logical processors.

Setting	Options
Local APIC Mode	• **xAPIC:** Use xAPIC, scales to only 255 hardware threads. • **x2APIC:** Use x2APIC, scales beyond 255 hardware threads but not supported by some legacy OS versions. • **Auto:** Use x2APIC only if 256 hardware threads in system, otherwise use **xAPIC**.

Table 2-16: APIC settings

2.5.2 SR-IOV Settings

SR-IOV requires enabling PCIe Alternative Routing-ID interpretation (ARI) on both root complexes and endpoints. ARI devices interpret the PCI address as an 8-bit function number instead of a 3-bit function number, and the device number is implied to be 0.

Setting	Options
PCIe ARI Support [SRIOV]	• **Disable:** Disable Alternative Routing ID interpretation. • **Enable:** Enable Alternative Routing ID interpretation.

Table 2-17: SR-IOV settings

2.5.3 PCIe Ten Bit Tag

A PCIe adapter must support 10-bit extended tags to achieve maximum PCIe Gen 4 bandwidth. This boosts adapter performance by allowing a 3x increase over the previous number of non-posted requests. Not all PCIe Gen 4 devices support 10-bit extended tags, which can cause issues during boot. Disabling this feature allows the server to boot if the adapter is having issues.

Setting	Options
PCIe Ten Bit Tag Support	• **Disable:** Disable PCIe 10-bit tags for all devices. • **Enable:** Enable PCIe 10-bit tags for supported devices. • **Auto:** Enabled

Table 2-18: PCIe 10-bit settings

2.5.4 LCLK and Preferred I/O Settings

The LCLK is an internal clock within the AMD EPYC 7003 Series Processor that controls the communication frequency between the Northbridge I/O (NBIO) logic and the rest of the I/O die (IOD). The PCIe root complex (RC) is part of the NBIO, and each RC has its own independently controlled LCLK to control the upstream and downstream communication speed. Each SoC NBIO has its own independently-controlled LCLK setting, and you should only adjust the specific NBIO that your device is connected to. Preferred I/O and Enhanced Preferred I/O improve DMA write performance for devices on a single PCIe bus. See the latest version of *Infinity Fabric Options* (AMD document #56970) for additional information.

Setting	Options
Root Complex LCLK Frequency	• **Auto:** Dynamically controlled by SoC. • 593 MHz.
Preferred I/O	• **Manual:** Enable Preferred I/O for the bus number specified by **Preferred I/O Bus**. • **Auto:** Disabled.
Preferred I/O Bus	• **Values 00h–FFh:** Specify the bus number for the device(s) for you wish to enable preferred I/O
Enhanced Preferred I/O	• **Auto:** Disabled. • Disabled. • Enabled.

Table 2-19: LCLK settings

2.5.5 Input-Output Memory Management Unit (IOMMU) Settings

Enabling the IOMMU allows devices such as the AMD EPYC processor-integrated SATA controller to present separate IRQs for each attached device instead of one IRQ for the subsystem. The IOMMU also allows operating systems to provide additional protection for DMA capable I/O devices. If you believe the IOMMU is limiting performance, then leave it enabled in BIOS and disable it via OS options (e.g., `iommu=pt` on the Linux® kernel command line). Enabling IOMMU is required when using x2APIC.

Setting	Options
IOMMU	• **Disabled:** Disable IOMMU. • **Enabled:** Enable IOMMU.

Table 2-20: IOMMU settings

This page intentionally left blank.

Chapter 3 AMD EPYC 7003 BIOS Settings by Workload

Use these guidelines for general-purpose workloads. Some cases list the benchmarks used in order to better describe the workloads used to obtain the recommended settings. Default settings are used when labeled default.

3.1 General-Purpose Workloads

3.1.1 Infinity Fabric Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
xGMI Link Max Speed	*default*	18 Gbps	18 Gbps	*default*
xGMI Link Width Control	*default*	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	*default*
xGMI Force Link Width Enable	*default*	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*	*default*
APBDIS	*default*	*default*	*default*	*default*
Fixed SOC P-State [see APBDIS]	*default*	*default*	*default*	*default*
DF C-States	*default*	*default*	*default*	*default*

Table 3-1: Infinity Fabric settings

3.1.2 NUMA and Memory Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
ACPI SRAT L3 Cache as NUMA Domain	Enabled	*default*	*default*	Enabled
ACPI SLIT Distance Control	*default*	*default*	*default*	*default*
ACPI SLIT Remote Relative Distance	*default*	*default*	*default*	*default*
ACPI SLIT <various> Distance	*default*	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	4	4	2 or 1	4
Memory Clock Speed	*default*	*default*	*default*	*default*
TSME	*default*	Disabled	Disabled	*default*

Table 3-2: NUMA and memory settings

3.1.3 Power Efficiency Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
EfficiencyModeEn	*default*	*default*	*default*	Enabled
Determinism Control	Enabled	Enabled	Enabled	*default*
Determinism Slider	Power	Power	Power	*default*
cTDP Control	Manual	Manual	Manual	*default*
cTDP	OPN Max	OPN Max	OPN Max	*default*
Package Power Limit Control	Manual	Manual	Manual	*default*
Package Power Limit	OPN Max	OPN Max	OPN Max	*default*
CPPC	*default*	*default*	*default*	*default*

Table 3-3: Power efficiency settings

3.1.4 Processor Core Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
L1 Stream HW Prefetcher	*default*	*default*	*default*	Disabled
L2 Stream HW Prefetcher	*default*	*default*	*default*	Disabled
SMT Control	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	1
BoostFmax	*default*	*default*	*default*	2500

Table 3-4: Processor core settings

3.1.5 I/O Settings

Setting	CPU Intensive	Java Throughput	Java Latency	Power Efficiency
PCIe ARI Support [SRIOV]	*default*	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*	*default*
Root Complex LCLK Frequency	*default*	*default*	*default*	*default*
Preferred I/O	*default*	*default*	*default*	*default*
Preferred I/O Bus	*default*	*default*	*default*	*default*
IOMMU	*default*	Enabled	Enabled	*default*

Table 3-5: I/O settings

3.2 Memory and I/O Intensive Workloads

3.2.1 Infinity Fabric Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency	Accelerator Throughpout
xGMI Link Max Speed	*default*	*default*	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*	Manual	*default*
xGMI Max Link Width	*default*	*default*	*default*	x16	*default*
xGMI Max Link Width Control	*default*	*default*	*default*	Manual	*default*
xGMI Force Link Width Enable	*default*	*default*	Force (Windows) *default* (Linux)	Force	*default*
xGMI Force Link Width	*default*	*default*	x16	x16	*default*
APBDIS	*default*	1	1	1	*default*
Fixed SOC P-State [see APBDIS]	*default*	*default*	P0	P0	*default*
DF C-States	*default*	*default*	Enabled	Enabled	*default*

Table 3-6: Infinity Fabric settings

3.2.2 NUMA and Memory Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency	Accelerator Throughput
ACPI SRAT L3 Cache as NUMA Domain	Enabled	*default*	Enabled (Win) *default* (Linux)	Enabled	*default*
ACPI SLIT Distance Control	*default*	*default*	*default*	*default*	*default*
ACPI SLIT Remote Relative Distance	*default*	*default*	*default*	*default*	*default*
ACPI SLIT <various> Distance	*default*	*default*	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	*default*	*default*	*default*	4	1
Memory Clock Speed	*default*	*default*	*default*	*default*	*default*
TSME	*default*	*default*	Disabled (Win) *default* (Linux)	*default*	*default*

Table 3-7: NUMA and memory settings

3.2.3 Power Efficiency Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency	Accelerator Throughput
EfficiencyModeEn	*default*	*default*	*default*	*default*	*default*
Determinism Control	*default*	Enabled	Enabled	Enabled	Enabled
Determinism Slider	*default*	Power	Power	Power	Power
cTDP Control	*default*	*default*	*default*	*default*	*default*
cTDP	*default*	*default*	*default*	*default*	*default*
Package Power Limit Control	*default*	*default*	*default*	*default*	*default*
Package Power Limit	*default*	*default*	*default*	*default*	*default*
CPPC	*default*	*default*	*default*	*default*	*default*

Table 3-8: Power efficiency settings

3.2.4 Processor Core Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency	Accelerator Throughput
L1 Stream HW Prefetcher	*default*	*default*	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*	*default*	*default*
SMT Control	*default*	*default*	*default*	*default*	*default*
BoostFmaxEn	*default*	*default*	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*	*default*	*default*

Table 3-9: Processor core settings

3.2.5 I/O Settings

Setting	Memory Throughput	Storage I/O Throughput	NIC Throughput	NIC Latency	Accelerator Throughput
Local APIC Mode	*default*	*default*	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	Enabled (Win) *default* (Linux)	Enabled	Enabled
Root Complex LCLK Frequency	*default*	*default*	593 MHz	593 MHz	*default*
Preferred I/O	*default*	*default*	Enabled (if only one NIC)	Enabled (if only one NIC)	*default*
Preferred I/O Bus	*default*	*default*	NIC bus number	NIC bus number	*default*
IOMMU	*default*	*default*	Disabled (Linux) *default* (Win)	Enabled	Enabled

Table 3-10: I/O settings

3.3 Virtualization and Containers

3.3.1 Infinity Fabric Settings

Setting	VMware vSphere Optimized	Linux KVM Optmized	Containers
xGMI Link Max Speed	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*
xGMI Force Link Width Enable	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*
APBDIS	*default*	*default*	*default*
Fixed SOC P-State [see APBDIS]	*default*	*default*	*default*
DF C-states	*default*	*default*	*default*

Table 3-11: Infinity Fabric settings

3.3.2 NUMA and Memory Settings

Setting	VMware vSphere Optimized	Linux KVM Optmized	Containers
ACPI SRAT L3 Cache as NUMA Domain	*default*	Enabled	*default*
ACPI SLIT Distance Control	*default*	*default*	*default*
ACPI SLIT Remote Relative Distance	*default*	*default*	*default*
ACPI SLIT <various> Distance	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	4	4	4
Memory Clock Speed	*default*	*default*	*default*
TSME	*default*	*default*	*default*

Table 3-12: NUMA and memory settings

3.3.3 Power Efficiency Settings

Setting	VMware vSphere Optimized	Linux KVM Optmized	Containers
EfficiencyModeEn	*default*	*default*	*default*
Determinism Control	*default*	Enabled	*default*
Determinism Slider	*default*	Performance	*default*
cTDP Control	*default*	*default*	*default*
cTDP	*default*	*default*	*default*
Package Power Limit Control	*default*	*default*	*default*
Package Power Limit	*default*	*default*	*default*
CPPC	*default*	*default*	*default*

Table 3-13: Power efficiency settings

3.3.4 Processor Core Settings

Setting	VMware vSphere Optimized	Linux KVM Optmized	Containers
L1 Stream HW Prefetcher	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*
SMT Control	Enabled	Enabled	Enabled
BoostFmaxEn	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*

Table 3-14: Processor core settings

3.3.5 I/O Settings

Setting	VMware vSphere Optimized	Linux KVM Optmized	Containers
Local APIC Mode	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*
Root Complex LCLK Frequency	*default*	*default*	*default*
Preferred I/O	*default*	*default*	*default*
Preferred I/O Bus	*default*	*default*	*default*
IOMMU	*default*	Enabled	*default*

Table 3-15: I/O settings

3.4 Database and Analytics

3.4.1 Infinity Fabric Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway
xGMI Link Max Speed	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*
xGMI Force Link Width Enable	*default*	*default*	*default*
xGMI Force Link Width	*default*	*default*	*default*
APBDIS	1	*default*	*default*
Fixed SOC P-State [see APBDIS]	P0	*default*	*default*
DF C-States	Disabled	*default*	*default*

Table 3-16: Infinity Fabric settings

3.4.2 NUMA and Memory Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway
ACPI SRAT L3 Cache as NUMADomain	*default*	*default*	*default*
ACPI SLIT Distance Control	*default*	*default*	*default*
ACPI SLIT Remote Relative Distance	*default*	*default*	*default*
ACPI SLIT <various> Distance	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	*default*	*default*	*default*
Memory Clock Speed	1	4	4
TSME	*default*	*default*	*default*

Table 3-17: NUMA and memory settings

3.4.3 Power Efficiency Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway
EfficiencyModeEn	*default*	*default*	*default*
Determinism Control	Enabled	*default*	*default*
Determinism Slider	Power	*default*	*default*
cTDP Control	Manual	*default*	*default*
cTDP	OPN Max	*default*	*default*
Package Power Limit Control	Manual	*default*	*default*
Package Power Limit	OPN Max	*default*	*default*
CPPC	*default*	*default*	*default*

Table 3-18: Power efficiency settings

3.4.4 Processor Core Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway
L1 Stream HW Prefetcher	*default*	*default*	*default*
L2 Stream HW Prefetcher	*default*	*default*	*default*
SMT Control	*default*	Enabled (if 32 cores or less per server)	Enabled
BoostFmaxEn	*default*	*default*	*default*
BoostFmax	*default*	*default*	*default*

Table 3-19: Processor core settings

3.4.5 I/O Settings

Setting	RDBMS Optimized	Big Data Analytics Optimized	IoT Gateway
Local APIC Mode	*default*	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	*default*	*default*
PCIe Ten Bit Tag Support	*default*	*default*	*default*
Root Complex LCLK Frequency	*default*	*default*	*default*
Preferred I/O	*default*	*default*	*default*
Preferred I/O Bus	*default*	*default*	*default*
IOMMU	*default*	Enabled	Enabled

Table 3-20: I/O settings

3.5 HPC and Telco Settings

3.5.1 Infinity Fabric Settings

Setting	HPC	OpenStack® NFV	OpenStack® for RealTime Kernel (NFV)
xGMI Link Max Speed	*default*	*default*	*default*
xGMI Link Width Control	*default*	*default*	*default*
xGMI Max Link Width	*default*	*default*	*default*
xGMI Max Link Width Control	*default*	*default*	*default*
xGMI Force Link Width Enable	*default*	*default*	Enabled
xGMI Force Link Width	*default*	*default*	x16
APBDIS	1	*default*	*default*
Fixed SOC P-State [see APBDIS]	P0	*default*	*default*
DF C-States	*default*	*default*	*default*

Table 3-21: Infinity Fabric settings

3.5.2 NUMA and Memory Settings

Setting	HPC	OpenStack® NFV	OpenStack® for RealTime Kernel (NFV)
ACPI SRAT L3 Cache as NUMA Domain	*default*	*default*	*default*
ACPI SLIT Distance Control	*default*	*default*	*default*
ACPI SLIT Remote Relative Distance	*default*	*default*	*default*
ACPI SLIT <various> Distance	*default*	*default*	*default*
NUMA Nodes per Socket (NPS)	4	2	2
Memory Clock Speed	*default*	*default*	*default*
TSME	*default*	*default*	*default*

Table 3-22: NUMA and memory settings

3.5.3 Power Efficiency Settings

Setting	HPC	OpenStack® NFV	OpenStack® for RealTime Kernel (NFV)
EfficiencyModeEn	*default*	*default*	*default*
Determinism Control	Enabled	*default*	Enabled
Determinism Slider	Power	*default*	Performance
cTDP Control	*default*	*default*	*default*
cTDP	*default*	*default*	*default*
Package Power Limit Control	*default*	*default*	*default*
Package Power Limit	*default*	*default*	*default*
CPPC	*default*	*default*	*default*

Table 3-23: Power efficiency settings

3.5.4 Processor Core Settings

Setting	HPC	OpenStack® NFV	OpenStack® for RealTime Kernel (NFV)
L1 Stream HW Prefetcher	*default*	Enabled	Enabled
L2 Stream HW Prefetcher	*default*	Enabled	Enabled
SMT Control	Disabled	*default*	*default*
BoostFmaxEn	*default*	Disabled	Disabled
BoostFmax	*default*	*default*	*default*

Table 3-24: Processor core settings

3.5.5 I/O Settings

Setting	HPC	OpenStack® NFV	OpenStack® for RealTime Kernel (NFV)
Local APIC Mode	X2APIC	*default*	*default*
PCIe ARI Support [SRIOV]	*default*	Enabled	Enabled
PCIe 10-Bit Tag Support	*default*	Enabled	Enabled
Root Complex LCLK Frequency	*default*	*default*	*default*
Preferred I/O	Enabled (if only one NIC)	Enabled (if only one NIC)	Enabled (if only one NIC)
Preferred I/O Bus	IB NIC bus number	NIC bus number	NIC bus number
IOMMU	Enable*	*default*	*default*

Table 3-25: I/O settings

* = For HPC, enable IOMMU in BIOS. Within Linux, add the boot command `iommu=pt` to set the IOMMU to passthrough mode.

This page intentionally left blank.