Hewlett Packard
Enterprise

Configuring and tuning HPE ProLiant Servers for low-latency applications

Technical white paper
Part Number: 581608-009
October 2017
Edition: 10

Contents

Introduction

Low-latency, deterministic system performance is a required system characteristic in the financial services market, where it enables high frequency trading, market data distribution, and exchange data processing. It is also required in other industries such as real-time signal and image processing.

These systems must respond rapidly to external events in a predictable manner. They must do so under heavy workloads, sometimes reaching millions of transactions per second. To achieve this level of performance, system designers must consider the following factors during system design and configuration:

- Hardware—System design, processor type and speed; memory latency, speed, and capacity; network components; and storage subsystem, including SSDs
- OS selection—Operating system kernels specifically designed and tuned for minimum latency and, in some cases, real-time preemption
- BIOS configuration—BIOS support configured for minimum latency and maximum performance
- Networking fabric—Network technology (1/10/25/40 Gigabit Ethernet, InfiniBand, Fibre Channel)
- Middleware—Messaging and database services on the network designed for minimum latency and maximum throughput with reliability. Networking drivers that use transparent kernel bypass libraries, such as VMA for Mellanox ConnectX-4 Lx and OpenOnload forSolarflare Flareon Ultra SFN8522-PLUS.
- End-user applications—Designed to perform multicast messaging accelerated via kernel bypass and RDMA techniques
- Physical distances—Physical separation between the information sources and clients affects overall system performance.

This document presents suggestions and best practice recommendations on BIOS configuration and on basic Linux OS tuning to obtain the lowest-latency performance from HPE ProLiant BL c-Class server blades, HPE Synergy Compute modules, and HPE ProLiant DL, ML, and XL servers. While this document contains information pertaining to Gen9 and earlier ProLiant servers, the primary focus is Gen10 servers.

The recommendations to disable System Management Interrupts (SMIs) are intended only for extreme latency-sensitive use cases. Most customers benefit from the power savings, monitoring, and notifications that the SMIs enable. These SMIs consume less than 0.1% of the server's processing capability, and HPE continues to reduce their impact with each new generation of ProLiant servers; in Gen10 servers, the periodic SMIs have been drastically reduced.

Important
The information in this document is accurate as of the document's release date but is subject to change based on updates made by HPE.

What's new

The current edition of the Configuring and Tuning HPE ProLiant Servers for Low-Latency Applications White Paper, includes the following additions and updates:

- "Recommended hardware configurations" on page 4
  - Updated information to new Intel Xeon Scalable versions of processors available on Gen10 servers
  - Updated recommended memory speed to 2666 MT/s
- Updated the following tuning procedures:
  - Updated BIOS tuning section to discuss Intelligent System Tuning functionality (Workload profiles, Jitter Smoothing, Core Boosting) available on Gen10 servers on page 11
  - "Tuning with HPE iLO RESTful API Interface Tool (Gen9 and Gen10)" on page 16
  - "Tuning with CONREP (Gen8 and older)" on page 17
  - "Recommended operating system tuning" on page 20

Recommended hardware configurations

HPE recommends the following HPE ProLiant Gen10 hardware configuration when low-latency is required. This information is subject to change and is valid as of the date of publication. For the latest information, see the server QuickSpecs on the HPE website (https://www.hpe.com/us/en/support.html).

- The Trade and Match Solution is optimized for applications that perform better at high frequency with lower core count. The latest version is based on Gen10 server technology and provides up to 2.5X the performance of the Gen9 version of the Trade and Match Solution. The base server configuration contains:
  - HPE ProLiant DL380 Gen10 or Apollo 6000 Gen10 server
  - Unique high frequency 8-core processors
  - Enterprise-grade solution including ECC memory and enterprise level components
- Processor
  - Intel Xeon Gold 5122 (4c 3.6 GHz), Gold 6144 (8c 3.5 GHz), Gold 6146 (12c 3.2 GHz), Gold 6142 (16c 2.6 GHz), and Gold 6154 (18c 3.0 GHz) in HPE ProLiant DL, ML, and BL servers, HPE Synergy compute modules, and Apollo System servers. Note: Not all server models support all processor options. See the server's Quickspecs (http://h41370.www4.hpe.com/quickspecs/overview.html) for the list of its supported processors.

  Consider a single processor configuration if your workload does not benefit from a second processor. The benefits of using a single processor are as follows:
  - Yields automatic PCI-to-core affinity (no application rewrite).
  - DDIO performs optimally.
  - Cache snooping is eliminated.
  - No UPI latency is experienced.
  - Simplified CPU core mapping is achieved.
- Memory
  - 8 GB single Rank DDR4-2666MT/s CAS-19 RDIMMs for two DIMMs per channel configurations.
  - If installing only one DIMM per channel, consider using 12 Dual-Rank 2666MT/s 16 GB RDIMMs for improved memory interleaving.
  - Each channel should be populated with at least one DIMM.
- PCIe Gen3 architecture
  - The HPE ProLiant DL380 Gen10 Server offers two x8 slots and one x16 that communicate with each processor. An optional tertiary riser board supports two additional x8 slots. Two additional option slots communicate with processor 1, one x8 FlexibleLOM slot for network options and an x8 Flexible SA slot for storage controller options.
  - The HPE ProLiant DL360 Gen10 Server offers two x8 or higher slots that communicate with processor 1 and one x16 slot that communicates with processor 2. Two additional option slots communicate with processor 1, one x8 FlexibleLOM slot for network options and an x8 Flexible SA slot for storage controller options.
  - The HPE ProLiant BL460c Gen10 Server Blade has one x16 mezzanine slot that communicates with processor 1 and one x16 mezzanine that communicates with processor 2, plus a FlexibleLOM off processor 1.
  - The HPE ProLiant XL230k Gen10 offers up to three optional PCIe expansion slots: one x8 internal, one x16 external, and one x8/x8 internal/external.
- PCIe NIC
  - Mellanox ConnectX-4 Lx-based adapters offer ultra-low latency and are designed specifically for HPE servers in two form factors: PCIe card (640SFP28) and FlexibleLOM (640FLR-SFP28). They are sold, integrated, and directly supported by HPE. The Mellanox ConnectX-4 Lx NIC offers native Gen3 x8 performance (25GbE). Mellanox Connect-IB based adapters offer even greater throughput by offering native Gen3 x16 performance.
  - Additional popular third-party PCIe Ethernet cards for ultra-low latency are available from Solarflare, Myricom, Chelsio, and Exablaze and can be installed in HPE industry-standard ProLiant DL, ML, and SL servers.

- Storage
  - New HPE Smart Array P408x and P816x storage controllers offer 12 Gb/s SAS performance when used in Gen10 servers with 12 Gb/s devices (SSDs or HDDs), which can deliver as much as 60% more IOPS vs. 6Gb/s devices.
  - The E208i-x is the base storage controller on the HPE ProLiant DL360 Gen10 and HPE ProLiant DL380 Gen10. It utilizes a special OS driver to provide RAID functionality to provide a low cost solution. However the driver-based RAID solution may introduce slight OS overhead which could affect latency. For latency-sensitive applications, it is recommended that the E208i be run in AHCI mode and for RAID-based configurations a hardware-based RAID controller such as the HPE Smart Array P408x be used.

Preparing for low-latency configuration

Taking inventories or snapshots

Before you configure servers for low-latency applications, HPE recommends that you take an inventory or snapshot of the following items. This will enable you to track changes during the optimization process.

- `dmidecode`
- `lspci -vv`
- `iLO RESTful Interface tool (for ProLiant Gen9 and later servers)`
- `hpdiscovery`
- `CONREP (for ProLiant Gen9 and earlier servers)`

To obtain the latest versions of CONREP, iLO RESTful Interface tool, or hpdiscovery, see "Obtaining the Scripting Utilities" on page 8,

- `sysctl -a`
- `HP-timetest7.3`
- HPE-TimeTest is a utility distributed by HPE that enables customers to test for jitter in a server. To obtain the HPE-TimeTest utility, contact HPE by emailing to: (low.latency@hpe.com). Please include the name of your HPE contact.
- `cat /proc/cmdline`

Upgrading BIOS and firmware

Before making changes for low-latency operation, be sure that all platform firmware is up-to-date. For low latency, it is especially important to upgrade the BIOS, iLO5 (Gen10), Innovation Engine (Gen10), and network card firmware to the latest versions. HPE offers the Service Pack for ProLiant (SPP) as a comprehensive solution to maintain all of the firmware and software for a server in a single ISO image. Use of the HPE SPP is the recommended method of upgrading the platform firmware. Refer to the "HPE Service Pack for ProLiant Contents Report" for the selected HPE SPP on www.hpe.com to verify latest versions of BIOS, iLO4, and network card firmware are included.

Important

The HPE SPP and BIOS images both require an active warranty or support agreement covering HPE ProLiant servers to be linked to the HPE Support Center profile used to download the components. Please refer to: http://h20564.www2.hpe.com/hpsc/doc/public/display?docId=c04044353 for more information.

To obtain the most recent HPE Service Pack for ProLiant (HPE SPP) upgrade for HPE ProLiant servers:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center.**
3. Enter the server model number, and then click **Go**.
4. Select **Get drivers, software, & firmware**.
5. Select the appropriate product link.
6. Select your operating system.
7. Select the **Application - System Management** category.
8. To obtain the HPE SPP upgrade, do the following:
   - o Download the latest HPE SPP ISO image, and then upgrade the firmware using the instructions included with the ISO. The HPE SPP can be used in online mode from either a Windows- or Linux-hosted operating system, or in offline mode by booting to the ISO.

To obtain the most recent BIOS upgrade for HPE ProLiant servers, if not in HPE SPP:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center**.

3. Enter the server model number, and then click **Go.**
4. Select **Get drivers, software, & firmware**.
5. Select the appropriate product link.
6. Select your operating system.
7. Select the **BIOS - System ROM** category.
8. To update the BIOS upgrade, do one of the following:
   - Download the latest ROMPaq firmware, and then upgrade the firmware using the instructions included with the ROMPaq.
   - Select **Online ROM Flash Component**, click the **Installation Instructions** tab, and then follow the instructions on the Online ROM Flash Component page.

To obtain the latest network card firmware, if not in HPE SPP:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center.**
3. Enter the server model number, and then click **Go.**
4. Select **Get drivers, software, & firmware**.
5. Select the appropriate product link.
6. Select your operating system.
7. Select **Firmware - Network**.
8. Download the appropriate NIC firmware.

Important
Version 2.20 of the iLO 4 firmware has been found to experience periodic SMIs every 15 seconds that cause some latency jitter. It is therefore recommended to use the iLO 4 v2.30 or later release of the firmware to address the problem.

To obtain the latest iLO 4 or iLO 5 firmware, if not in HPE SPP:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center**.
3. Enter the server model number, and then click **Go.**
4. Select **Get drivers, software, & firmware.**
5. Select the appropriate product link.
6. Select your operating system.
7. Select **Firmware - Lights-Out Management**.
8. Click **Obtain software**, and then click the executable file to download it.

Obtaining the Scripting Utilities

For Gen10 and Gen9 servers:

The HPE iLO RESTful Interface Tool can be used to configure BIOS options on Gen9 servers and is the preferred tool for Gen10 systems.

To install the HPE iLO RESTful Interface Tool:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center**.
3. Enter the server model number, and then click **Go.**
4. Select **Get drivers, software, & firmware**.
5. Select the appropriate product link.
6. Select your operating system.
7. Select **Utility.**
8. Click **Download**, next to the appropriate executable file to save it.

For Gen8 and older servers:

The CONREP utility can be used to configure Processor Power and Utilization Monitoring or Memory Pre-Failure Notification for minimum latency, and is included in Smart Start Scripting Toolkit (SSSTK) 9.10 or later. For Gen8 servers, SSSTK is now called STK.

To install the STK:

1. Go to the HPE website (https://www.hpe.com/us/en/support.html).
2. Select **HPE Support Center**.
3. Enter the server model number, and then click **Go.**
4. Select **Get drivers, software, & firmware.**
5. Select the appropriate product link.
6. Select your operating system.
7. Select **Utility - Tools**.
8. Click **Download**, next to the appropriate executable file to save it.

Recommended platform tuning

System requirements

The HPE BIOS configuration options described in this document include options in HPE ProLiant servers to disable the generation of periodic System Management Interrupts (SMIs) used for Power Monitoring and for Memory PreFailure Notification with their attendant latency impact. BIOS options are generally independent of the OS, and a properly tuned low-latency operating system is also required to achieve deterministic performance.

The tuning recommendations described in this document are based on testing and customer interactions. But no single "recipe" can be prescribed. Customers needing a low-latency environment often perform exhaustive testing of the latency impact of various tuning parameters with their application and systems to determine the optimum settings for their environment.

Tuning recommendations and explanations

Consider the following options as part of any deployment in low-latency OS kernel environments:

- Take an inventory or snapshot. See “Taking inventories or snapshots” on page 6.
- `HP-TimeTest7.3`

HPE-TimeTest is a utility distributed by HPE that enables customers to test for jitter in a server. To obtain the HPE-TimeTest utility, contact HPE by emailing to: (low.latency@hpe.com). Please include the name of your HPE contact.

- Capture kernel boot settings
  - For non-UEFI systems (Gen8 and earlier)
    - `cat /boot/grub/grub.conf (for RHEL)`
    - `cat /boot/grub/menu.lst (for SLES)`
  - For UEFI systems (Gen9 and Gen10)
    - `cat /boot/efi/EFI/redhat/grub.conf (for RHEL)`
    - `cat /boot/grub.conf (for SLES 12)`
- Upgrade the BIOS. See "Upgrading BIOS and firmware" on page 6.
- If using a Linux-based server, prepare the server for low-latency tuning. See "Preparing Linux-based servers for low-latency tuning" on page 20.
- Make the recommended changes to the BIOS.
- For tuning recommendations and instructions, see the following sections:
  - “Tuning with the ROM-based Setup Utility (RBSU)” on page 16.
  - “Tuning with HPE iLO RESTful Interface Tool (Gen9 and above)” on page 16.
  - Preparing for low-latency configuration” on page 6.

HPE servers are configured by default to provide the best balance between performance and power consumption. These default settings may not provide the lowest latency. The first step in tuning for low latency is to examine these additional settings that may assist in obtaining optimal low-latency performance. These settings are accessible through RBSU and with the RESTful API utilities, configuration tools that are provided by HPE.

All HPE ProLiant Gen8 and later Intel-based servers, regardless of the ROM version, support setting Intel Turbo Boost and C-States.

The following table provides descriptions of the recommended low-latency settings for Linux environments. Settings that differ for Windows environment are noted in Description column.

Table 1. Recommended low-latency settings for Linux environments.

PARAMETER	VALUE	DESCRIPTION	GENERATION
Workload Profile	Low Latency	Allows setting a group of BIOS options simultaneously to accommodate the targeted class of applications. Default is "General Power Efficiency Compute"; for Apollo systems, the default is High Performance Compute. For more information, see "Intelligent System Tuning" on page 11.	Gen10
Intel Hyperthreading Options	Disabled	Allows Hyperthreading, which adds logical cores but increases computational jitter	Gen8 and above
Intel Turbo Boost Technology	Disabled	Allows processors to make a transition to a frequency that is higher than its rated speed. Note that utilizing Turbo Boost Technology will likely introduce computational jitter due to frequency transitions of the processor cores. For more information, see "Turbo mode information and considerations" on page 13.	Gen8 and above
Intel VT-d	Disabled	Enables virtualized Directed I/O	Gen8 and above
Thermal Configuration	First try Optimal Cooling, then repeat with Increased Cooling and then Max Cooling (if available)*	Steps through the different available cooling settings available in RBSU. Use the one that provides the desired performance for the lowest power consumption. For more information, see "Thermal considerations" on page 14.	Gen8 and above
HPE Power Profile	Maximum Performance	Disables all power management options that may negatively affect performance	Gen8 and Gen9
HPE Power Regulator	HPE Static High Performance Mode	Keeps processors in their maximum power/performance state (automatically set by HPE Power Profile for Gen8 and Gen9 servers, automatically set by Workload Profile for Gen10 servers)	Gen8 and above
Intel QPI Link Power Management	Disabled	Precludes placing unutilized QPI links into low power state	Gen8 and Gen9
Intel UPI Link Power Management	Disabled	Precludes placing unutilized UPI links into low power state (automatically set by Workload Profile for Gen10 servers)	Gen10
Minimum Processor Idle Power Core C-State	No C-States	Precludes processor transitions into low-power core C-States (automatically set by HPE Power Profile for Gen8 and Gen9 servers, automatically set by Workload Profile for Gen10 servers)	Gen8 and above
Minimum Processor Idle Power Package C-State	No Package State	Precludes processor transitions into low-power package C-States (automatically set by HPE Power Profile for Gen8 and Gen9 servers, automatically set by Workload Profile for Gen10 servers)	Gen8 and above
Energy/Performance Bias	Maximum Performance	Configures processor subsystems for high-performance/low-latency (automatically set by HPE Power Profile for Gen8 and Gen9 servers)	Gen8 and above
Collaborative Power Control	Disabled	Precludes the OS from changing clock frequency (automatically set by HPE Power Profile for Gen8 and Gen9 servers, automatically set by Workload Profile for Gen10 servers)	Gen8 and above
DIMM Voltage Preference	Optimized for Performance	Runs DIMMs at a higher voltage if it increases performance. (Gen8 only)	Gen8
Dynamic Power Capping Functionality	Disabled	This option allows for disabling System ROM Power Calibration during the boot process. Doing so accelerates boot times but precludes enabling of a Dynamic Power Cap. (Gen8 and Gen9 only)	Gen8 and Gen9

Table 1. Recommended low-latency settings for Linux environments continued.

PARAMETER	VALUE	DESCRIPTION	GENERATION
Memory Power Savings Mode	Maximum Performance	This option configures several memory parameters to optimize the memory subsystems performance and is configured to Balanced by default.	Gen8
QPI Snoop Configuration	Early Snoop, Home Snoop, or Cluster on Die**	This option allows for the configurations of different snoop modes that impact the QPI interconnect. Changing this option may improve performance in certain workloads. Home Snoop provides high-memory bandwidth in an average NUMA environment (default setting). Cluster on Die may provide increased memory bandwidth in highly optimized NUMA workloads. Early Snoop may decrease memory latency but may also result in lower overall bandwidth as compared to other modes.	Gen9
QPI Home Snoop Optimization	Directory + OSB Enabled	This option allows the disabling of the Directory and Opportunistic Snoop Broadcast (OSB) functionality that is available when the QPI snoop mode is set to Home Snoop. Directory and OSB provides better memory latency and increased bandwidth and is recommended for the vast majority of workloads that benefit from Home Snoop and is enabled by default.	Gen9
ACPI SLIT Preferences	Enabled	This ACPI SLIT describes the relative access times between processors, memory subsystems, and I/O subsystems. Operating systems that support the SLIT can use this information to improve performance by allocating resources and workloads more efficiently. This option is disabled by default on most ProLiant Gen8 and Gen9 servers.	Gen8 and above
Processor Power and Utilization Monitoring	Disabled***	Disables iLO Processor State Mode Switching and Insight Power Manager Processor Utilization Monitoring, and its associated SMI	Gen8 and Gen9
Memory Pre-Failure Notification	Disabled*** Enabled (Gen10)	Disables Memory Pre-Failure Notification and its associated SMI. Note that for Gen10 servers the periodic SMI is only active if ECC errors are detected, significantly reducing the number of SMI events. It is recommended that this be set to Enabled for Gen10 servers.	Gen8 and above
Memory Patrol Scrubbing	Disabled***	Memory Periodic Patrol Scrubber corrects memory soft errors so that, over the length of the system runtime, the risk of producing multi-bit and uncorrectable errors is reduced (automatically set by Workload Profile for Gen10 servers)..	Gen8 and above
Memory Refresh Rate	1x Refresh***	This option controls the refresh rate of the memory controller. The default value for this parameter is 2x for Gen 9 and earlier, and 1x for Gen 10. (automatically set by Workload Profile for Gen10 servers)	Gen8 and above

*If Turbo mode is enabled, then step through the available cooling settings described in “Thermal considerations” on page 14. Otherwise, the default Optimal Cooling setting is adequate.

**QPI Snoop Configuration selection depends on the processor and workload used. See “QPI Snoop Configuration information and considerations (Gen9 only)” on page 12.

***These options are under the Service Options menu (Gen9 and earlier). See “Tuning with ROM Based Setup Utility (RBSU)” on page 16, “Tuning with CONREP” on page 17, or “Tuning with HPE iLO RESTful Interface Tool” on page 166 for details on how to set these options.

Intelligent System Tuning considerations (Gen10 only)

HPE Intelligent System Tuning (IST) is a new set of features that are introduced for HPE ProLiant Gen10 servers to provide enhanced performance tuning. The IST features included in the Gen10 portfolio are Workload Matching, Jitter Smoothing, and Core Boosting. All of these features have an impact in optimizing a server's performance and should be considered for low latency environments. Note that Jitter Smoothing and Core Boosting require an iLO Advanced license to activate and that Core Boosting is available only on select servers and processors. Refer to the "HPE Intelligent System Tuning" whitepaper (https://www.hpe.com/us/en/servers/management/tuning.html#Resources) for more details.

Workload Matching allows a user to select a workload profile in the RBSU that is targeted for applications which fall into specific workload categories. Each of the profiles will set the values for a number of BIOS options that have been determined to benefit the given workload category. For example, the Low Latency workload profile sets BIOS options such as Power Regulator and Turbo Boost to values that are known to benefit latency sensitive environments. The settings that are changed by a workload profile are locked to the intended value and cannot be

changed while the profile is active; these options will be grayed out in the RBSU menu. . To change an option set by a workload profile to a different value, the "Custom" profile must be used.

Processor frequency transitions due to cores entering/leaving different C-states and Turbo Boost states is a source of much computational jitter. While disabling Turbo Boost and C-States eliminate these sources of jitter, it is at the cost of potential performance gains. Jitter Smoothing is meant to address this performance loss. When enabled, Jitter Smoothing will initially run processors at their top-rated Turbo-Boosted frequency (or the frequency set by the Processor Jitter Control Frequency option in RBSU). When a frequency transition is detected, the new lower frequency is now set to be the maximum allowed frequency. This cycle will continue until no more frequency transitions occur. In many cases this threshold frequency will be higher than the processor's base frequency, resulting in improved overall system performance.

When utilizing Jitter Smoothing, a couple of considerations should be kept in mind.

- Enabling Jitter Control will enable Turbo Boost functionality even when the Low Latency workload profile is selected, so both can be used simultaneously.
- If C-States are enabled by the OS, then the Jitter Control set to "Auto" will drop the frequency to the highest enabled C-state when a processor goes into an idle state. This will set the maximum frequency to the C-state frequency, significantly limiting performance. On Linux systems, disable the Intel c-state driver by using the kernel parameter "`intel_idle.max_cstate=0`" to address this problem.
- Intel processors used in Gen10 and in some Gen9 systems have different frequency ranges when applications use AVX or non-AVX instruction sets. In an environment with both AVX and non-AVX applications, Jitter Control set to "Auto" will drop the threshold core frequency to the highest sustainable AVX frequency; this will negatively impact non-AVX performance when AVX workloads are not active. If a small amount of jitter during such transitions between AVX and non-AVX workloads is acceptable, then setting Jitter Control to "Manual" with a corresponding Jitter Control Frequency defined will allow the processor to drop down to the AVX frequencies when needed and jump back up to the Jitter Control Frequency value when AVX instructions are no longer used.

Some ProLiant Gen10 servers offer special processor options that provide HPE's Core Boosting technology, which utilizes more aggressive Turbo Boosting profiles that can provide increased performance over comparable standard processor configurations. Since Core Boosting technology relies on aggressive Turbo Boosting, frequent frequency changes are possible which will introduce computational jitter.

QPI Snoop Configuration information and considerations (Gen9 only)

The QPI Snoop Configuration setting will control how cache snoops are handled. When using the “Early Snoop” option the snoops will be sent by the caching agents; this will provide better cache latency for processors when the snoop traffic is low. The “Home Snoop” option will cause the snoops to be sent from the home agent; this provides optimal memory bandwidth balanced across local and remote memory access. The “Cluster on Die” option will snoop the directory cache first and then the home agent. Using this option will also cause the processor to appear as two NUMA nodes within operating systems, one for each memory controller. This option provides optimal performance for highly NUMA-aware workloads. Note that the “Cluster on Die” option is available only on processors with 10 or more cores.

Xeon E5-2600 v4 series processors add further optimization to the Home Snoop configuration with the "Directory and Opportunistic Snoop Broadcast (OSB)” option under the "QPI Home Snoop Optimization" RBSU setting. Each home agent has a small cache that holds the directory state of migratory cache lines. The home agent will speculatively snoop the remote socket in parallel with the directory read. This enables low cache-to-cache latencies, low memory latencies, and higher memory bandwidths. Home Snoop with Dir+OSB generally will be the best snoop mode for most workloads. See the Table 2 for a summary of the QPI Snoop options.

Table 2. QPI Snoop modes supported in two-socket configurations

	EARLY SNOOP	HOME SNOOP (DEFAULT RBSU OPTION)	HOME SNOOP with DIRECTORY + OSB	CLUSTER ON DIE
Previously available on	E5-2600 (SNB)	E5-2600 v2 (IVB)	E5-2600 v4 (BDW)	E5-2600 v3 (HSW)
Snoop sent by	Caching Agent	Home Agent	Home Agent + speculative remote socket snoop	Directory Cache, then Home Agent
Best used for	Memory latency-sensitive workloads*	NUMA workloads that need maximum local and remote bandwidth	NUMA workloads that need maximum local and remote bandwidth	Highly NUMA-optimized workloads

*For v4 series processors, Home Snoop with Dir+OSB provides better local NUMA latencies with higher memory bandwidth and is recommended over Early Snoop.

Sub-NUMA Cluster (SNC) Configuration and considerations (Gen10 only)

Intel Xeon Scalable processors use a mesh architecture concept to connect processor cores to memory and UPI links, and each core has its own combined Caching and Home agent (CHA). This architecture makes the ring-based QPI snoop modes in older generations of processors obsolete. However, there are still two memory controllers on each processor, so Sub-NUMA clustering (SNC) is introduced to optimize memory access latency and performance. SNC essentially creates two localization domains within the processor where one memory controller and half of the cores are associated to each domain. This feature is best used with workloads that are highly NUMA optimized.

Core frequencies for AVX vs.non-AVX applications information and considerations (Gen9 and Gen10)

With the Intel Xeon v3 series and later processors, Advanced Vector Extensions version 2.0 (AVX2) allow applications to perform 256-bit wide operations for integer and floating-point operations, providing an opportunity for increased performance. However, the power requirements for running AVX instructions are higher than for non-AVX instructions. Therefore the processor's core frequency range will change depending upon whether AVX instructions are executing or not. Beginning with Intel Xeon v4 processors, cores that are executing AVX instructions will be constrained to a lower frequency range (AVX base and AVX Turbo) while running the instructions. The CPU's core frequency will return to the non-AVX frequency range ~1m-sec after the AVX instructions have completed. With the Intel Xeon Scalable series processors, AVX 512 was introduced, allowing 512-bit wide operations to be used with their own set of base and Turbo frequencies.

Table 3 below shows both the AVX (2.0 and 512) and non-AVX frequency ranges for three segment-optimized Xeon Scalable processors. Note that the processor is still also governed by the power/thermal characteristics of the system, so the actual frequency will be determined by both the type of instructions used and the power/thermal conditions.

Turbo mode information and considerations

Intel Turbo Boost can be used to increase the processor's operating clock frequency, but at the risk of computational jitter if the processor changes its Turbo frequency. When that happens, processing stops for a small period of time, introducing uncertainty in application processing time. Turbo operation is a function of power consumption, processor temperature, and the number of active cores. Carefully managing these factors (e.g. utilizing the IST feature Jitter Smoothing, described on p. 11), however, can result in consistent Turbo operation without jitter. The maximum Turbo frequencies for various numbers of active cores for two selected processors are given in the following table.

Table 3. Turbo frequency ranges for certain Xeon Scalable Series processors

PROCESSOR	POWER	BASE FREQUENCY			NUMBER OF ACTIVE CORES	TURBO-ENABLED FREQUENCY		
		AVX 512	AVX 2.0	Non-AVX		AVX 512	AVX 2.0	Non-AVX
Gold 6154	200 W	2.1 GHz	2.6 GHz	3.0 GHz	17-18	2.7 GHz	3.3 GHz	3.7 GHz
					13-16	2.8 GHz	3.3 GHz	3.7 GHz
					9-12	3/1 GHz	3.3 GHz	3.7 GHz
					5-8	3.2 GHz	3.3 GHz	3.7 GHz
					3-4	3.3 GHz	3.4 GHz	3.7 GHz
					1-2	3.5 GHz	3.6 GHz	3.7 GHz
Gold 6144	150 W	2.2 GHz	2.6 GHz	3.5 GHz	5-8	2.8 GHz	3.5 GHz	4.1 GHz
					3-4	3.3 GHz	3.5 GHz	4.1 GHz
					1-2	3.5 GHz	3.6 GHz	4.2 GHz
Gold 5122	105 W	2.7 GHz	3.2 GHz	3.6 GHz	3-4	3.3 GHz	3.6 GHz	3.7 GHz
					1-2	3.5 GHz	3.6 GHz	3.7 GHz

If the penalty of computational jitter is too severe and you are unable to control temperature and keep power consumption less than TDP, you should disable Turbo Mode or reduce the maximum Turbo frequency using Jitter Smoothing.

Power consumption

Pushing the processor's TDP limit will result in the processor changing its Turbo frequency. Because of the risk of processor failure, Intel offers no method to lock a processor into Turbo Mode. Most applications will not consume enough power to exceed the processor's TDP. If you are concerned that yours might but still need to operate at the maximum Turbo Boost frequency, then you can disable a core per processor from within the BIOS or set a core offline from the OS, reducing power consumption and providing TDP headroom.

Tests have shown that the Xeon Gold 6154 processor under heavy computational load is able to stay at the maximum Turbo frequency indefinitely when the system is properly configured, as outlined in this document. However, this is not guaranteed behavior and you should verify this with your workload. Also be aware that due to processor manufacturing variations, some CPUs may be able to stay at full Turbo frequency while other samples of the same processor model may not.

Thermal considerations

The processor's thermal limits are another consideration in maintaining consistent Turbo operation. Ensure that the server's inlet temperature meets the specification in the associated QuickSpecs. Beyond that, there is a BIOS parameter that can be used to regulate the amount of cooling delivered by the fans, but before changing it, note that most configurations will maintain the preferred operating state with the default Optimal Cooling setting. If the system requires more cooling, the server will respond by increasing the fan speed to deliver the necessary cooling.

However, some demanding environments may require a greater base level of cooling. If testing shows that your server's Turbo frequency varies in response to exceeding temperature limits due to varying system load, evaluate the Increased Cooling option, which carries a penalty of increased system power consumption, acoustics, and airflow demand.

The third setting for this parameter is Maximum Cooling, which causes the fans to operate always at their highest speed. Use this setting only if your environment requires it, as it has significantly higher power consumption, acoustic noise, and facility airflow demand.

Keep in mind that different processors have different requirements. The Xeon Platinum 8180 has a notably higher TDP than the Xeon Gold 6144, but the Tcase for the Xeon Gold 6144 is 9°C (16° F) lower than for the Xeon Platinum 8180, making proper cooling especially important.

Active cores

In addition to TDP and thermals, the amount of frequency boost obtained is a function of the number of active cores, which is never more than the number of operational cores as specified by a BIOS setting. Active cores are cores in C0, C1, or C1E State, and HPE recommends disabling C-States in order to keep the number of active cores constant and avoid the attendant latency jitter of changing Turbo frequencies.

Other considerations for Turbo Mode

As noted in "Active cores" page 15, C-States must be disabled in the BIOS. However, some versions of Linux ignore the BIOS setting and must be configured to disable C-States. For more information, see "Recommended Linux boot-time settings" on page 22.

Disabling processor power and utilization monitoring and memory pre-failure notification SMIs (Gen8 and Gen9)

Disabling System Management Interrupts to the processor provides one of the greatest benefits to low-latency environments. Disabling the Processor Power and Utilization Monitoring SMI has the greatest effect because it generates a processor interrupt eight times a second in G6 and later servers. Disabling the Memory Pre-Failure Notification SMI has a much smaller effect because it generates an interrupt at a low frequency: once per hour on G6 and G7 servers, once every five minutes on Gen8 servers, and once every minute on the DL580 Gen8 and all Gen9 servers.

Disabling each option causes some server features to become unavailable. Before reconfiguring BIOS, be sure that none of the features described below are required.

Disabling Processor Power and Utilization Monitoring disables the following features:

- iLO Processor State Monitoring
- Insight Power Manager CPU Utilization Reporting
- HPE Dynamic Power-Savings Mode

Disabling Memory Pre-Failure Notification has the following effects:

- Disables Memory Pre-Failure Warranty Support
- Disables notification when correctable memory errors occur above a pre-defined threshold
- Forces the system to run in Advanced ECC Mode, regardless of the mode configured in RBSU

Important

Online Spare Mode, Mirroring Mode, and Lock-step Mode are not supported when Memory Pre-Failure Notification support is disabled. Supported AMP modes depend on the generation and model of the ProLiant server.

Disabling Memory Pre-Failure Notification does not disable the Advanced ECC mode or correction of errors. Uncorrectable errors are still flagged, logged, and bring the system down. The only difference when this SMI is disabled is that there is no early notification if the correctable error threshold is exceeded.

Disabling Dynamic Power Capping Functionality

Disabling Dynamic Power Capping Functionality prevents the ability to enable a Power Cap via iLO. When this parameter is disabled, the option to enable a Power Cap via iLO is no longer available. Since low-latency installations are unlikely to set power caps, the Dynamic Power Capping Functionality option may be safely disabled in the BIOS. This option accelerates the boot process but does not have any impact on latency when the platform is operating.

Disabling Patrol Scrubbing

Patrol Scrubbing is a feature that scans memory to correct soft memory errors. On the HPE ProLiant Gen9 Server, the Patrol Scrubber re-arms itself through an SMI. The frequency of this event is roughly once per day, but varies based on the amount of installed memory. Low-latency installations can avoid this SMI by disabling Patrol Scrubbing, which is an option in the Service Options menu (Gen8 and Gen9). For Gen10, the option is now under Memory Options menu. On other platforms, Patrol Scrubbing does not require SMI functionality and does not need to be disabled.

Setting the Memory Refresh Rate
An extremely rare potential for memory errors is eliminated by the default memory refresh rate of 2x. Decreasing the rate to 1x will improve memory performance, but with a vanishingly small potential for memory errors. This option is available in the Service Options menu (Gen8 and Gen9). For Gen10, the option defaults to 1x and is now under Memory Options menu.

Tuning with the ROM-based Setup Utility (RBSU)
For new system testing or in environments with a small number of systems where script-based maintenance is not used, the RBSU in the UEFI Systems utilities is the recommended method to configure the BIOS.

To configure BIOS low-latency options using RBSU:

1. Power the server on.
2. When prompted during POST, press `F9` to enter RBSU (Gen8 and earlier) or System Utilities (Gen9 and later).
3. For Gen9 and later servers, select **System Configuration →BIOS/Platform Configuration (RBSU)**.
4. Browse through the menus to change the parameters. For more information, see "Tuning recommendations and explanations" on page 9.

Important
Do not change the other options in the Services Options menu.

---

5. For the parameters marked with "***" in the "Tuning recommendations and explanations" table on page 9, go into the Service Options menu (Gen8 and Gen9):
   a. For Gen8 and earlier, while in the top level of the RBSU menu, press `CTRL-A` to display the option for the Service Options menu. Select **Service Options**.
   b. For Gen9 and DL580 Gen8, press `CTRL-A`., you will be redirected immediately to the Service Options menu.
6. Verify that the parameters are set as indicated in "Tuning recommendations and explanations" on page 9.

Tuning with HPE iLO RESTful Interface Tool (Gen9 and Gen10)

The HPE iLO RESTful Interface tool is useful for scripting the deployment of BIOS options across multiple servers from a common profile file. For complete details on how to utilize the HPE RESTful Interface tool for scripting, please refer to the *HPE RESTful Interface Tool User Guide on the HPE website (See "*Support and other resources" on page 26 for links).

This section will provide details on how to download the service menu options to a json profile file and modify those setting via `hprest`. For Gen10 servers, there are no BIOS options in the service menu that need be changed. The example below is for Gen9 servers. To configure BIOS low-latency options using the HPE iLO RESTful Interface Tool:

1. Edit the `/etc/ilorest/redfish.conf` file and update the following fields:
```
url = https://[iLO IP address]
username = [iLO user account name]
password = [iLO user password]
```
2. Change the current directory to a convenient working directory:
```
cd /home/user/ilorest
```
3. Capture a snapshot of your current Service menu settings (Gen9 only):
```
ilorest rawget /rest/v1/systems/1/bios/service/settings > service.opts.txt
```
4. To disable Processor Power and Utilization Monitoring, disable Memory Pre-Failure Notification, and set Memory Refresh Rate to 1x, and create a patch file, service.patch, with the following format (Gen9 only):
```
{
     "path": "/rest/v1/systems/1/bios/Service/Settings",
     "body": {
         "ProcPwrUtilMonitor": "Disabled",
```

```
            "MemPreFailureNotification": "Disabled",
            "MemRefreshRate": "1xRefresh"
        }
    }
```

5. To disable Patrol Scrubbing, add to the service.patch patch file the following markup in the "body" stanza (Gen9 only):

```
            "MemPatrolScrubbing": "Disabled",
```

Important

The "body" stanza in the `service.patch` file can contain multiple BIOS settings. Each setting should be separated by a comma; the final setting should not have a comma following it. Also, you cannot include settings from the Service RBSU menu and the other menus in the same patch file.

---

6. Update the BIOS with the modified settings:

```
ilorest rawpatch service.patch
```

7. Log out of hprest's iLO session:

```
ilorest logout
```

8. Reboot the server:

```
reboot
```

Tuning with CONREP (Gen8 and older)

CONREP is useful for scripting the deployment of BIOS options across multiple servers from a common profile file. For complete details in how to utilize CONREP for scripting, please refer to the *HP Scripting Toolkit for Linux User Guide* on the HPE website. (See "Support and other resources" on page 26 for links).

Important

Using CONREP to modify BIOS settings may result in different behavior than using RBSU or the HPE RESTful interface tools. In particular, caution must be exercised when changing the HPE Power Profile with CONREP as it will not propagate changes to other BIOS settings in the same manner as the RBSU and the HPE RESTful interface tool will. When changing the HPE Power Profile using CONREP, you will also need to change the settings shown in Table 1 on page 10 that are linked to the HPE Power Profile.

This section will provide details on how to add the service menu options to the default profile file, `conrep.xml` and modify those setting via CONREP. To configure low-latency Service Menu BIOS options using the CONREP utility in STK:

1. Change the current directory to the STK/utilities directory:
   cd STK/utilities

2. Edit the conrep.xml file to include the following stanzas before `</Conrep>` at the end of the file:

```
<Section name="PowerMonitoring">
<helptext>
<![CDATA [This setting determines if Pstate logging and utilization is supported.]]>
</helptext>
<ev>CQHGV3</ev>
<length>1</length>
<value id="0x00">Enabled</value>
<value id="0x10">Disabled</value>
<mask>0x10</mask>
<byte>0</byte>
</Section>
<Section name="DisableMemoryPrefailureNotification">
<helptext>
<![CDATA [This setting allows the user to disable Memory Pre-Failure Notification
support, which will remove the periodic SMI associated with this support. Not
```

```
recommended for anyone except for those who absolutely need every periodic SMI
removed.]]>
</helptext>
<ev>CQHGV3</ev>
<length>1</length>
<value id="0x00">No</value>
<value id="0x20">Yes</value>
<mask>0x20</mask>
<byte>0</byte>
</Section>
<Section name="Memory_Refresh_Rate_Gen9">
<helptext><![CDATA [This setting allows the user to change the Memory Refresh Rate
setting on Gen9 servers.]]></helptext>
<platforms>
<platform>Gen9</platform>
</platforms>
<nvram>0x257</nvram>
<value id="0x00">1x_Refresh</value>
<value id="0x10">2x_Refresh</value>
<value id="0x20">3x_Refresh</value>
<mask>0x30</mask>
</Section>
<Section name="Memory_Refresh_Gen8">
<helptext><![CDATA [This setting allows the user to change the Memory Refresh
setting on Gen8 servers.]]></helptext>
<platforms>
<platform>Gen8</platform>
</platforms>
<nvram>0x261</nvram>
<value id="0x01">1x_Refresh</value>
<value id="0x00">2x_Refresh</value>
<value id="0x02">3x_Refresh</value>
<mask>0x03</mask>
</Section>
<Section name="Memory_Patrol_Scrubbing_Gen9">
<helptext><![CDATA[This setting allows the user to enable or disable the Memory Patrol
Scrubbing setting on Gen9 servers.]]></helptext>
<platforms>
<platform>Gen9</platform>
</platforms>
<nvram>0x257</nvram>
<value id="0x08">Disabled</value>
<value id="0x00">Enabled</value>
<mask>0x08</mask>
</Section>
```

3. Capture a snapshot of your current settings:

```
. /conrep -s -x conrep.xml -f conrep_settings.xml
```

4. To disable Intel Turbo Boost Technology, verify that the `conrep_settings.xml` file contains the following markup (G7 servers):

```
<Section name="Intel_Processor_Turbo_Mode" helptext="Allows Intel processors to
transition to a higher frequency than its rated speed if the processor has available
headroom and is within temperature specification.">Disabled</Section>
```

5. To disable Intel Turbo Boost Technology, verify that the `conrep_settings.xml` file contains the following markup (Gen8 and Gen9 servers):

```
<Section name="Intel_Turbo_Boost_Optimization_Gen8" helptext="Optimize Turbo Boost
heuristics for different situations. For Gen8 or later servers only.">Disabled</Section>
```

6. To disable Processor Power and Utilization Monitoring, verify that the `conrep_settings.xml` file contains the following markup:

```
<Section name="PowerMonitoring" helptext="This setting determines if Pstate logging and
utilization is supported.">Disabled</Section
```

7. To disable Memory Pre-Failure Notification, verify that the `conrep_settings.xml` file contains the following markup:

```
<Section name="DisableMemoryPrefailureNotification" helptext="This setting allows the
user to disable Memory Pre-Failure Notification support, which will remove the periodic
SMI associated with this support.  Not recommended for anyone except for those who
absolutely need every periodic SMI removed.">Yes</Section>
```

8. To disable Memory Patrol Scrubbing (Gen9 servers), verify that the `conrep_settings.xml` file contains the following markup:

```
<Section name="Memory_Patrol_Scrubbing" helptext=" This setting allows the user to
enable or disable the Memory Patrol Scrubbing setting on Gen9
servers.">Disabled</Section>
```

9. Update the BIOS with the modified settings:

```
./conrep -l -x conrep.xml -f conrep_settings.xml
```

10. Reboot the server:

```
reboot
```

Recommended operating system tuning

Linux

Preparing Linux-based servers for low-latency tuning

Before configuring a ProLiant Gen8 or later server for low latency, do the following:

1. Make the following edits:
   - For non-UEFI configurations (Gen8):

     Red Hat (EL 6.x): Edit `/boot/grub/grub.conf` and add `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` to the kernel line

     SLES: Edit `/boot/grub/menu.lst` and add `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` to the kernel line

   - For UEFI configurations (Gen9 and Gen10):

     Red Hat (EL 6.x): Edit `/boot/efi/EFI/redhat/grub.conf` and add `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` to the kernel line

     SLES 12: Edit `/boot/grub2/grub.conf` and add `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` to the kernel line

   - For Red Hat Enterprise Linux Server 7.0 or greater:

     Edit `/etc/default/grub` file and add `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` to the `"GRUB_CMDLINE_LINUX"` line.

Run command:

`# grub2-mkconfig -o /boot/grub2/grub.cfg` (non-UEFI configurations) or

`# grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg` (UEFI configurations)

`nosoftlockup` prevents the kernel from logging an event when a high-priority thread executes continuously on a core for longer than the soft lockup threshold.

`intel_idle.max_cstate=0` prevents the kernel from overriding the BIOS C-State setting.

`mce=ignore_ce` prevents Linux from initiating a poll every five minutes of the Machine Check Banks for correctable errors, which can cause latency spikes. For more information, see the Linux Kernel Archives website (http://www.kernel.org/doc/Documentation/x86/x86_64/boot-options.txt).

2. Disable Error Detection and Correction (EDAC). HPE has developed sophisticated algorithms using the monitored data to better predict a potential memory module failure. This system of monitoring and analysis is known as "HPE Advanced Memory Error Detection Technology". As a result, HPE recommends disabling Linux EDAC. See https://h20564.www2.hpe.com/hpsc/doc/public/display?docId=emr_na-a00016026en_us for more details. To disable EDAC, follow these steps:
   1. Search for EDAC modules and disable EDAC if running
   2. Run:

      ```
      lsmod | grep edac
      ```

   3. For each EDAC module (if any found):

      Add the following to `/etc/modprobe.conf` on OS releases that support `/etc/modprob.conf`

      ```
      alias edac_xxx off
      ```

      Add the following to `/etc/modprobe.d/blacklist.conf` on OS releases that support `/etc/modprobe.d/blacklist.conf`

      ```
      blacklist edac_xxx
      ```

3. Set tuned profile. Tuned is a utility introduced in RHEL 6 and SLES 12 that allows the user to implement a set of OS optimizations as part of a profile. A set of pre-defined profiles is provided that can be used. For RHEL 6, it is recommended that the `"latency-performance"` profile is used for latency-sensitive applications. For RHEL 7 and SLES 12, `"network-latency"` is recommended

for low-latency environments. For details on the performance options being set, please see the `/usr/lib/tuned/`[performance profile]/tuned.conf file for the desired performance profile. To set the desired profile, run the command:

```
# tuned-adm profile latency-performance (RHEL 6) or
# tuned-adm profile network-latency (RHEL 7 and SLES 12)
```

4. Reboot the server.

5. After reboot, run the `stop-services.sh` script to stop extraneous services. The following example stops the services shown and prevents them from starting on subsequent boots:

```
for SERVICE in \

acpid           alsasound     autofs           avahi-daemon     bluetooth     \
conman          cpuspeed      cron             cups             cupsrenice    \
dhcdbd          Dnsmasg       dund             firstboot        hidd          \
ip6tables       Ipmi          irda             kudzu            libvirtd      \
lvm2-monitor    mcstrans      mdmonitor        mdmpd            messagebus    \
multipathd      netconsole    netfs            netplugd         nscd          \
oddjobd         Pand          pcscd            postfix          powersaved    \
psacct          Rdisc         readahead_early  readahead_later  restoresecond \
rhnsd           Rpcgssd       rpcidmapd        rpcsvgssd        saslauthd     \
sendmail        Slpd          smartd           smbfs            suseRegister  \
sysstat         wpa_          xfs              vpbind           yum-updatesd  \
                supplicant

do
 chkconfig --level 2345 $SERVICE off
 service $SERVICE stop
 done
```

Note
For RHEL 7 systems, use the following script to disable services:

```
for SERVICE in \

 avahi-daemon.service        crond.service          dnsmasq.service
 firewalld.service           lvm2-monitor.service   postfix.service
 rpcgssd.service             rpcidmapd.service      rpcsvcgssd.service
 wpa_supplicant.service
do
  systemctl  disable $SERVICE
  systemctl stop $SERVICE
  done
```

6. Use the irqbalancer to preclude some cores from servicing software IRQs:

   a. Enter the following command:

   b. `# service irqbalance stop`

   c. Do a one-time run of the irq balancer, where `"CoreMask"` is a mask of the CPUs on which the OS should not schedule interrupts

   d. `# IRQBALANCE_ONESHOT=1 IRQBALANCE_BANNED_CPUS=${CoreMask} irqbalance`

   e. Wait until the command service `irqbalance status` returns `"irqbalance is stopped."`

f. On SLES 11, the name of the IRQ balancer service is `irq_balancer.`

g. On RHEL 7 and SLES 12, use `systemctl` instead of `service` command to stop irqbalance.

Red Hat MRG Realtime

Red Hat resolved scaling issues for the MRG 2.3 operating system for ProLiant servers with large core counts, such as the DL580 G7 server with four 10-core E7-4870 processors. If you are using MRG 2.3 on servers with a large number of cores, be sure to use a release with a kernel version equal to or greater than the following:

```
kernel-rt-3.6.11-rt30.25.el6rt
```

In addition to having a large number of cores, if your server is running the MRG 2.3 (or later) Realtime kernel, it is using the SLUB memory allocator. The SLUB memory allocator requires additional tuning for real-time performance. The SLUB allocator has pseudo-files named "cpu_partial" in the "/sys/kernel/slab" file system. To get the best real-time performance from the allocator, these files should be set to "0", disabling the cpu_partial logic. This can be done with the following command:

```
# find /sys/kernel/slab -name 'cpu_partial' -exec sh -c 'echo 0 > {}' \;
```

Recommended Linux boot-time settings

The Linux boot parameter "idle=poll" keeps the processing cores in C0 state when used in conjunction with "intel_idle.max_cstate=0." Without it, the processor will enter C1 state.

- For RHEL systems:

  For RHEL 6, edit `/boot/grub/grub.conf (or /boot/efi/EFI/redhat/grub.conf` for UEFI systems) and add `"idle=poll"` to the kernel line. This is in addition to the `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` parameters that should have been added previously.

  For RHEL 7, edit `/etc/default/grub.cfg and add "idle=poll".` This is in addition to the `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` parameters that should have been added previously. After the edit, run the command:

  `# grub2-mkconfig -o /boot/grub2/grub.cfg` (non-UEFI configurations) or

  `# grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg` (UEFI configurations)

- For SLES systems:

  For SLES 11, edit `/boot/grub/menu.lst (or /boot/efi/efi/SuSE/elilo.conf` for UEFI systems) and add `"idle=poll"` to the kernel line. This is in addition to the `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` parameters that should have been added previously.

  For SLES12, edit `/etc/default/grub.cfg and add "idle=poll".` This is in addition to the `"nosoftlockup intel_idle.max_cstate=0 mce=ignore_ce"` parameters that should have been added previously. After the edit, run the command:

  `# grub2-mkconfig -o /boot/grub2/grub.cfg`

Verifying the configuration

To verify your ProLiant server is properly configured for low-latency operation, clear one core (selected at random) of the operating system IRQs, and then run the HPE-TimeTest utility on the randomly selected core:

```
# Core=5
# CoreMask=`echo "16 o 2 $Core ^ p" | dc`
# service irqbalance stop
# until [ "`service irqbalance status`" = "irqbalance is stopped" ] ; do sleep 1 ; done
# IRQBALANCE_ONESHOT=1 IRQBALANCE_BANNED_CPUS=${CoreMask} irqbalance
# sleep 1
# until [ "`service irqbalance status`" = "irqbalance is stopped" ] ; do sleep 1 ; done
# numactl --physcpubind=${Core} --localalloc nice -n -20 ./HP-TimeTest7.3 -v -f csv -o
smi_count -p RR,1
```

On SLES 11, the name of the IRQ balancer service is `irq_balancer.`

On RHEL 7 and SLES 12, use `"systemctl"` to disable and monitor the status of the irqbalance service.

For more information on the HPE-Timetest tool, see "HPE-Timetest" on page 24.

Consider the following:

- Consider changing the smp_affinity for the IRQs. For example, on a server on which you want to leave core 0 for the OS, the following masks off the other processors for all IRQs:

```
for MF in `find /proc/irq -name *smp_affinity` ; do awk -F, \
'{for(i=1;i<NF;i++)printf("00000000,");printf("%8.8x\n",and(0x00000001,
strtonum("0x"$NF)))}' \
$MF > $MF ; done
```

- Consider using cset (http://code.google.com/p/cpuset/) to shield cores from the OS. For example, on a server on which you want to keep the OS from all cores except for 0, use the following command:

```
# cset shield --cpu 1-15 --kthread=on
```

- If running as root, the following command can then be used to move the current PID to the "user" set of cores:

```
# cset proc --move --pid=$$ --threads --toset=user
```

Windows

HPE BIOS low-latency options are supported in Windows Server 2008 and 2012 environments.

To apply the low-latency options in a Microsoft Windows environment:

1. Obtain the STK. See "Obtaining the Scripting Utilities" on page 8.
2. Run the SmartComponent for the most recent version of the STK, note the directory it is in, and then change to it in Windows Explorer or a command window.
3. Run CONREP. See "Tuning with CONREP (Gen8 and older)" on page 17.

For other low-latency tuning recommendations in a Windows environment, do the following:

- See the Windows Server 2012 R2Tuning Guide on the Microsoft website at https://msdn.microsoft.com/en-us/library/windows/hardware/dn529133.

For more information or assistance, contact Microsoft to be put in touch with one of their low-latency experts.

HPE-TimeTest

The original behavior of HPE-TimeTest has been maintained through its many edits, but this behavior is not optimal. For example, it runs at real-time priority 99, but should be run at no higher than 80. On an otherwise idle system, a real-time priority of "1" is adequate for HPE-TimeTest to run properly.

The following provides an example of running HPE-TimeTest with an explanation of each component of the command:

```
time numactl --physcpubind=3   \ Bind to core 3 and use local memory
--localalloc
nice -n -20                          \ nice; probably not necessary
/HP-TimeTest/HP-TimeTest7.3     \ HP-TimeTest7.3 executable
-f csv                           \ output in Comma Separated Variable (csv) format
-o smi                           \ print SMI_count at the beginning and end
-o date                          \ print a timestamp at the beginning and end
-m cycles                        \ latency is determined by cycles (instead of time)
-t `echo '.000005 2900000000 * 0 k \ threshold is 5 µsec on 2.90 GHz processor
1 / p' | dc`
-l `expr 2900000000 \* 60 \* 30 /     \ run for ~30 minutes on 2.90 GHz processor
44`                              \ ("44" is # of cycles per loop iteration I get)
-p RR,1,-20                      \ Use RR scheduling at priority 1; use "nice"
                                 \ of -20 (I suspect irrelevant for RT policies)
```

Generating the output in CSV format allows for easy import into a spreadsheet for plotting.

To provide additional suggestions, contact the HPE low-latency team.

Frequently asked questions

Q. Does disabling Memory Pre-Failure Notification disable memory error correction?

A. Memory errors are still corrected, but notification that the error rate has exceeded a pre-set threshold is disabled. The latency impact of this feature is very small. HPE recommends disabling Memory Pre-Failure Notification only if absolutely necessary.

Q. What memory features are lost if Memory Pre-Failure Notification is disabled?

A. If Memory Pre-Failure Notification is disabled, Online Spare and Mirroring memory modes become unavailable. The system is forced to run in Advanced ECC mode, regardless of the mode set in BIOS. Memory Pre-Failure Warranty Support also becomes unavailable because there is no notification of errors exceeding the programmed threshold.

Q. How does disabling iLO Processor State Monitoring in the HPE ProLiant c-Class enclosure affect power management?

A. Disabling state monitoring does not affect power management.

Q. How can I verify that a server has the low-latency option set?

A. Use one of the following options to verify that the low-latency option is set:

- See the information in "Tuning recommendations and explanations" on page 9.
- Run HPE-TimeTest to see if you are getting spikes. For more information, contact HPE by emailing to: low.latency@hpe.com. Please provide the name of your local HPE representative and region of origin so that we can better serve your request.

Q. Can I interrogate or confirm the memory operating speed?

A. To interrogate or confirm the memory operating speed, ensure your SMBIOS is 2.7 or later and use dmidecode 2.11 or later with the following command:

```
dmidecode -t 17
```

Q. How do I tune a network adapter for optimum low latency?

A. This white paper does not address this topic. Refer to the supplier of the network adapter's controller technology. For example, tuning advice for Mellanox ConnectX-3 adapters integrated and supported by HPE is available on the Mellanox website (https://community.mellanox.com/docs/DOC-2489).

Q. How does HPE recommend I disable cores in ProLiant Gen8 servers?

A. Do the following:

1. From the RBSU menu, navigate to System Options**>**Processor Options**>**Processor Core Disable (Intel Core Select)**.**
2. Enter the number of cores per processor that you want to enable.
   For example, if you have 8-core processors and want to disable 1 core, enter "7" in this field.
3. Boot the server. Verify that the correct information appears during POST; for example, "2 Processor(s) detected, 14 total cores enabled."
   The number of enabled cores can also be modified with hprest (or ilorest for Gen10 and later) or CONREP. To modify the number of enabled cores with CONREP, use version available from STK for Linux 10.0 or later. To modify the number with hprest, use HPE RESTful Interface Tool 1.10 or later.

Q. How do I verify at what Turbo frequency my cores are running?

A. There are a number of utilities that track the real time frequency of each CPU core. For example, for Linux:

1. i7z is an open source utility that provides information on the Intel Core i3, i5, i7, and corresponding Xeon processors. Pre-compiled versions of this utility can be found for most Linux distributions, including Red Hat and SLES.

2. Red Hat Enterprise Linux 6.4 and later provides the utility turbostat as part of its cpupowerutils package.

Both of these utilities will provide real-time information about each core's frequency and percent time in each C-state.

Support and other resources

Resources and documentation

The following resources are available:

- *HP UEFI System Utilities User Guide (for HP ProLiant DL580 Gen8)* on the HP website http://h20566.www2.hpe.com/hpsc/doc/public/display?docId=c03886429
- *HPE UEFI System Utilities User Guide for HP ProLiant Gen10 Servers* on the HP website (https://h20564.www2.hpe.com/hpsc/doc/public/display?docId=emr_na-a00016407ja_jp).
- *HPE RESTful Interface Tool User Guide* on the HPE website (https://hewlettpackard.github.io/python-redfish-utility/#overview)
- iLO documentation:
  - *HPE iLO 5 User Guide* 1.15 (for Gen10 servers) on the HPE website (http://h20564.www2.hpe.com/hpsc/doc/public/display?docId=a00026409en_us)
  - *HPE iLO 4 User Guide* (for Gen8 and Gen9 servers) on the HPE website h20565.www2.hpe.com/hpsc/doc/public/display?docId=c03334051
  - *HPE iLO 4 Scripting and Command Line Guide* (for Gen8 servers) on the HPE website (http://h20566.www2.hpe.com/hpsc/doc/public/display?docId=c03334058)
- *HPE Scripting Toolkit 11.00 for Linux User Guide* on the HPE website http://h20566.www2.hpe.com/hpsc/doc/public/display?docId=a00017067en_us
- *HPE Scripting Toolkit 11.00 for Windows User Guide* on the HPE website http://h20566.www2.hpe.com/hpsc/doc/public/display?docId=a00017070en_us
- *STK* on the HPE website (https://www.hpe.com/us/en/support.html)

  The CONREP, hcrpu, and hpdiscovery utilities are available as part of the STK. For more information on downloading STK, see

  "Obtaining the Scripting Utilities" on page 8.
- HP-TimeTest 7.3 utility. To obtain the utility, contact HPE by emailing: low.latency@hpe.com. Please provide the name of your local HPE representative and region of origin so that we can better serve your requests.

Before you contact HPE

Be sure to have the following information available before you call HPE:

- Active Health System log
- Download and have available an Active Health System log for three days before the failure was detected. For more information, see the *HPE iLO 4 User Guide or HPE Intelligent Provisioning User Guide* on the HPE website at https://www.hpe.com/us/en/support.html
- Onboard Administrator SHOW ALL report (for HPE BladeSystem products only)

  For more information on obtaining the Onboard Administrator SHOW ALL report, see the HPE website h20628.www2.hpe.com/km-ext/kmcsdirect/emr_na-c00705292-47.pdf
- Technical support registration number (if applicable)
- Product serial number
- Product model name and number
- Product identification number
- Applicable error messages
- Add-on boards or hardware
- Third-party hardware or software
- Operating system type and revision level

HPE contact information

For United States and worldwide contact information, see the Contact HPE website at https://www.hpe.com/us/en/contact-hpe.html.

In the United States:

- To contact HPE by phone, call 1-650-687-5817. For continuous quality improvement, calls may be recorded or monitored.
- If you have purchased a Care Pack (service upgrade), see the Support & Drivers website at http://h20565.www2.hpe.com/portal/site/hpsc. If the problem cannot be resolved at the website, call 1-800-633-3600. For more information about Care Packs, see the HPE website at ssc.hpe.com/portal/site/ssc/.

On a best-effort basis only, HPE offers technical assistance on low-latency tuning to customers who have followed this guide and still have questions. For more information, contact HPE by emailing low.latency@hpe.com. Please provide the name of your local HPE representative and region of origin so that we can better serve your request.

Acronyms and abbreviations

ACPI
Advanced Configuration and Power Interface specification

AMP
Advanced Memory Protection

AVX
Intel Advanced Vector Extension

BIOS
Basic Input/Output System

DDIO
Distributed Discrete Input/Output

HPRCU
HP ROM Configuration Utility

iLO
Integrated Lights-Out

LOM
LAN on Motherboard

MRG
Red Hat Enterprise Messaging Realtime Grid platform

POST
Power-On Self Test

QPI
Intel QuickPath Interconnect

RBSU
ROM-Based Setup Utility

SLES
SUSE Linux Enterprise Server

SLIT
System Locality Information Table

SLUB
Unqueued slab memory allocator

SMI
System Management Interrupt

STK
Scripting Toolkit

TDP
Thermal Design Power

UEFI
Unified Extensible Firmware Interface

Sign up for updates

★ Rate this document

c01804533,
October 2017