Configuring and Tuning HP ProLiant Servers for Low-Latency Applications White Paper

Abstract

This document is intended to assist HP customers in configuring, tuning, and optimizing HP ProLiant servers for ultra low-latency applications.

Part Number: 581608-005
April 2013
Edition: 5

Contents

Introduction

Overview

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; 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 (1/10/40 Gigabit Ethernet, InfiniBand, Fibre Channel)
- Middleware—Messaging and database services on the network designed for minimum latency and maximum throughput with reliability
- 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 OS tuning to obtain the lowest-latency performance from HP ProLiant BL c-Class server blades and HP ProLiant DL, ML, and SL servers. While this document contains information pertaining to G7 and earlier ProLiant servers, the primary focus is Gen8 servers and later.

The recommendations to disable System Management Interrupts (SMIs) are intended only for extremely 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 HP continues to reduce their impact with each new generation of ProLiant server.

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 HP.

What's new

The current edition of the *Configuring and Tuning HP ProLiant Servers for Low-Latency Applications White Paper*, 581608-005, includes the following additions and updates:

- Recommended hardware configurations (on page 6):
  - Added information on the E5-2687W processor for the HP ProLiant DL360p Gen8 SE Server

- Added information on the Solarflare SFC9020 10 GbE controller

- Added HP ProLiant DL360p Gen8 SE Server (on page 8)
- Tuning recommendations and explanations (on page 11):
  - Updated the following information in the table:
    - Intel Turbo Boost Technology
    - Thermal Configuration
  - Added the following information to the table:
    - Dynamic Power Capping Functionality
    - Memory Patrol Scrubbing
    - Memory Refresh Rate
    - Memory Power Savings Mode
    - ACPI SLIT Preferences
- Updated Turbo mode information and considerations (on page 13) and added the following subsections:
  - Power consumption (on page 13)
  - Thermal considerations (on page 14)
  - Active cores (on page 14)
  - Other considerations for turbo mode (on page 14)
- Added Disabling Dynamic Power Capping Functionality (on page 15)
- Added Disabling Patrol Scrubbing (on page 15)
- Added Setting the Memory Refresh Rate (on page 15)
- Added Setting Memory Power Savings Mode and ACPI SLIT Preferences (on page 15)
- Updated the following tuning procedures:
  - Tuning with the ROM-based Setup Utility (on page 16)
  - Tuning with the HP ROM Configuration Utility (Gen8 and later) (on page 16)
  - Tuning with conrep (on page 16)
- Linux:
  - Updated RHEL and SLES (on page 19)
  - Added Red Hat MRG Realtime (on page 20)
  - Added Recommended Linux boot-time settings (on page 20)
- Added HP-TimeTest (on page 22)

Recommended hardware configurations

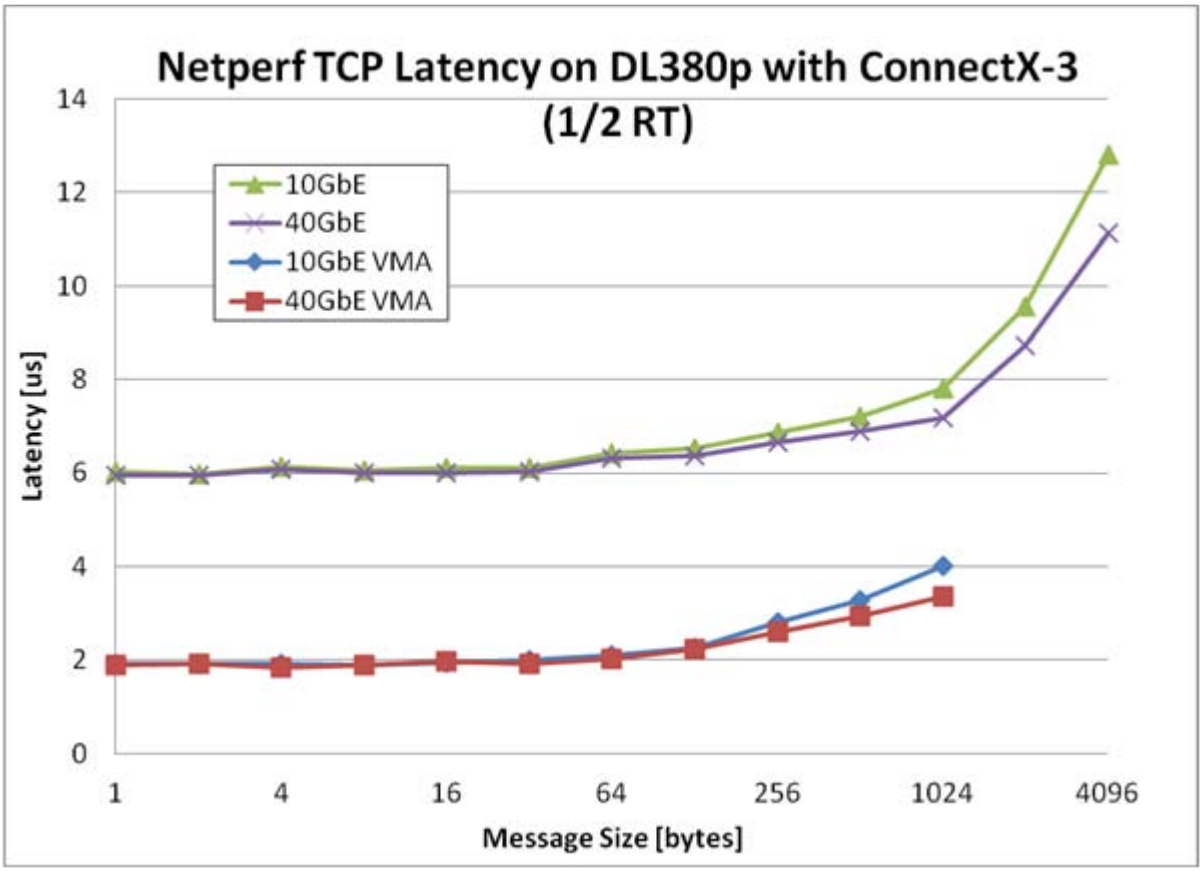
HP recommends the following ProLiant Gen8 hardware configuration when low-latency is required. This information is subject to change and is valid as of the date of publication. For the latest information, see the server QuickSpecs on the HP website (http://www.hp.com/go/support).

- Processor
  - E5-2690 and E5-2643 in BL servers (HP recommends these processors for server blades.)
  - E5-2690 (8c 2.9GHz) and E5-2643 (4c 3.3GHz) in DL servers
  - E5-2687W (3.1 GHz) in the HP ProLiant DL360p Gen8 SE Server. For more information, see "HP ProLiant DL360p Gen8 SE Server (on page 8)."

  For the latest information on supported processors, see the server QuickSpecs on the HP website (http://www.hp.com/go/support).
- Memory
  - 8 or 16 Dual Rank 1600MT/s CAS-11 RDIMMs
  - If installing only one DIMM per channel, consider using 8 Dual Rank 1600MT/s UDIMMs for a 1-clock latency advantage.
  - Each channel should be populated with at least one DIMM.
- PCIe Gen3 architecture
  - The HP ProLiant DL380p Gen8 Server offers three x8 or higher slots that communicate with processor 1 and three x8 or higher slots that communicate with processor 2.
  - The HP ProLiant DL360p Gen8 Server offers three x8 or higher slots that communicate with processor 1.

    Consider a single processor configuration if your application needs approximately 6 cores only. The benefits are as follows:
    - Automatic PCI-to-core affinity (no application rewrite)
    - DDIO performs optimally.
    - Cache snooping is eliminated.
    - No QPI latency
    - Operation at the maximum turbo mode frequency is more likely due to reduced thermal/power load.
    - Even with one processor, there are still two x8 and one x16 PCIe slots for NICs, timing cards, Fusion-io, and so forth.
  - The HP ProLiant BL460c Server Blade has one x8 mezzanine slot that communicates with processor 1 and one x8 mezzanine that communicates with processor 2, plus a FlexibleLOM off processor 1.
- PCIe NIC
  - Mellanox ConnectX-3 based adapters offer ultra low latency and are designed specifically for HP servers in three form factors: PCIe card, FlexibleLOM, and server blade mezzanine. They are sold, integrated, and directly supported by HP. Mellanox ConnectX-3 is the only NIC offering native Gen3 x8 performance (40GbE and FDR InfiniBand).

- The Solarflare SFC9020 10GbE controller is now a supported PCI option for HP ProLiant DL servers. For more information, see the HP 570SFP QuickSpecs on the HP website (http://h18004.www1.hp.com/products/quickspecs/14544_div/14544_div.pdf).
- Additional popular third-party PCIe Ethernet cards for ultra low latency are available from Myricom and can be installed in HP industry-standard ProLiant DL, ML, and SL servers.
- The graph below shows the HP Mellanox FlexibleLOM (Part #649282-B21) in a back-to-back configuration (no switch), running Netperf v2.5.0.



- A half-round trip, with core 2 as an example, uses the following command line:

```
netperf -n 16 -H <peer ip> -c -C -P 0 -t TCP_RR -l 10 -T 2,2 -- -r <message size>
```

- Storage
  - Compared with G7 storage controllers, the HP SmartArray P420 storage controller for Gen8 servers has double the cache size, 6 times the performance with SSD, and double the number of supported drives.
  - HP I/O Accelerator now supports up to 1.2TB MLC in server blade mezzanine cards.

    For more information, see the *HP IO Accelerator for HP BladeSystem c-Class QuickSpecs* on the HP website (http://h18004.www1.hp.com/products/quickspecs/13220_div/13220_div.pdf).
- Tuning

  See "Tuning recommendations and explanations (on page 11)."

HP ProLiant DL360p Gen8 SE Server

HP has made available a new variant of the DL360p Gen8 server designed specifically for low-latency environments, the HP ProLiant DL360p Gen8 SE Server. This special edition server has multiple features in response to the requirements of this market segment:

- Support for the 150W 3.1 GHz Intel Xeon E5-2687W processor. With Turbo Boost, this processor is capable of running at between 3.4 and 3.8 GHz, depending on the number of active cores, with turbo boost stepping of 3/3/3/4/4/5/5/7. Compare this to the Intel Xeon E5-2690 range of 3.3 to 3.8 GHz, with turbo boost stepping of 4/4/4/5/5/7/7/9.
- Enhanced thermal design accommodates the higher wattage processors with standard air cooling.
- A PCIe slot connected directly to each processor. This provides a latency benefit to configurations with two processors and two I/O cards, such as Mellanox ConnectX-3 NICs, for environments that pay careful attention to process placement.
- Updated BIOS to support the new processors and I/O configuration.

As a Special Edition product, the HP ProLiant DL360p Gen8 SE is only available through special order. For more information, or to place an order, contact your HP representative or channel partner. Although it is an Americas product, it can be shipped to other geographies.

Preparing for low-latency configuration

Taking inventories or snapshots

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

- `dmidecode`

  For RHEL before 6.2, obtain v. 2.11 from the nongnu website (http://www.nongnu.org/dmidecode).

- `lspci -vv`
- `conrep` (for ProLiant Gen8 and earlier servers) or `hprcu` (for ProLiant Gen8 servers)

  HP recommends using `hprcu` because of the additional benefits it provides over `conrep`.

- `hpdiscovery`

  To obtain the latest versions of `conrep`, `hprcu`, or `hpdiscovery`, see "Obtaining the Scripting Toolkit (on page 10)."

- `sysctl -a`
- `HP-TimeTest7.2`

  HP-TimeTest is a utility distributed by HP that enables customers to test for jitter in a server. To obtain the HP-TimeTest utility, contact HP (mailto:low.latency@hp.com). Provide your name and your company's name, as well as your mailing address and your HP contact's name.

- Capture kernel boot settings
  - `cat /boot/grub/grub.conf` (for RHEL)
  - `cat /boot/grub/menu.lst` (for SLES)

Upgrading BIOS

Before making BIOS changes for low-latency operation, upgrade the BIOS to the most recent version.

To obtain the most recent BIOS upgrade for HP ProLiant servers:

1. Go to the HP website (http://www.hp.com/go/support).
2. Select **Drivers & Software**.
3. Enter the server model number, and then click **Search**.
4. Select the appropriate product link.
5. Select your operating system.
6. Select the **BIOS - System ROM** category.
7. To obtain 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.

Upgrading 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 network card and iLO 4 firmware to the latest versions.

To obtain the latest network card firmware:

1. Go to the HP website (http://www.hp.com/go/support).
2. Select **Drivers & Software**.
3. Enter the server model number, and then click **Search**.
4. Select the appropriate product link.
5. Select your operating system.
6. Select **Firmware - Network**.
7. Download the appropriate NIC firmware.

To obtain the latest iLO 4 firmware:

1. Go to the HP website (http://www.hp.com/go/support).
2. Select **Drivers & Software**.
3. Enter the server model number, and then click **Search**.
4. Select the appropriate product link.
5. Select your operating system.
6. Select **Firmware - Lights-Out Management**.
7. Click **Obtain software**, and then click the executable file to download it.

Obtaining the Scripting Toolkit

The conrep and `hprcu` utilities can be used to configure the BIOS for minimum latency, and are included in the STK. Use STK 9.40 or later.

For Gen8 servers, SSSTK is now called STK.

`hprcu` is available for Gen8 servers only. `conrep` is available for Gen8 servers and earlier.

`conrep` is the only method available for configuring these options on HP ProLiant G5 servers and HP ProLiant G6 servers that utilize AMD Opteron processors. The utility is one method available for configuring HP ProLiant G6 and G7 servers that utilize Intel Xeon processors.

To install the STK:

1. Go to the HP website (http://www.hp.com/go/ProLiant/STK).
2. Select **STK for Linux**.
3. Select **Download**.
4. Create a new directory.
5. Unpack the archive in the new directory.

Recommended platform tuning

System requirements

The HP BIOS configuration options described in this document include options in HP 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 ("Taking inventories or snapshots" on page 9).
- Upgrade the BIOS ("Upgrading BIOS" on page 9).
- Upgrade the firmware ("Upgrading firmware" on page 10).
- If using a Linux-based server, prepare the server for low-latency tuning.
- Make the recommended changes to the BIOS.

  For tuning recommendations and instructions, see the following sections:

  - Tuning with the ROM-Based Setup Utility (on page 16)
  - Tuning with the HP ROM Configuration Utility (Gen8 and later) (on page 16)
  - Tuning with `conrep` (on page 16)
- Verify the server is configured for low-latency operation ("Verifying the configuration" on page 20).

HP 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 `conrep` and `hprcu` utilities, configuration tools provided by HP.

All HP ProLiant G6 and later Intel-based servers, regardless of the ROM version, support setting Intel Turbo Boost and C-States. For G7 and earlier servers, HP ProLiant 100 Series servers do not support advanced features for iLO Performance Monitoring and Memory Pre-Failure notification.

The following table provides descriptions of the recommended low latency settings for Linux environments. For recommended Windows settings, see "Windows (on page 21)."

Parameter	Value	Description
Intel Virtualization Technology	Disabled	Allows Virtual Machine Managers to utilize virtualization hardware capabilities
Intel Hyperthreading Options	Disabled	Allows Hyperthreading, which adds logical cores but increases computational jitter
Intel Turbo Boost Technology	Enabled	This option allows the processor to make a transition to a frequency that is higher than its rated speed.
Intel VT-d	Disabled	Enables virtualized Directed I/O
Thermal Configuration	First try Optimal Cooling, then repeat with Increased Cooling and then Max Cooling (if available).*	This option enables you to step through the different available cooling settings available in RBSU. Use the one that provides the preferred performance for the lowest power consumption. For more information, see "Thermal Considerations (on page 14)."
HP Power Profile	Maximum Performance	Disables all power management options that may negatively affect performance
HP Power Regulator	HP Static High Performance Mode	Keeps processors in their maximum power/performance state
Intel QPI Link Power Management	Disabled	Precludes placing unutilized QPI links into low power state
Minimum Processor Idle Power Core State	No C-states	Precludes processor transitions into low-power core C-States
Minimum Processor Idle Power Package State	No Package State	Precludes processor transitions into low-power package C-States
Energy/Performance Bias	Maximum Performance	Configures processor subsystems for high-performance/low-latency
Collaborative Power Control	Disabled	Precludes the OS from changing clock frequency
DIMM Voltage Preference	Optimized for Performance	Runs DIMMs at a higher voltage if it increases performance
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.
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.
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 servers.
Processor Power and Utilization Monitoring	Disabled	Disables iLO Processor State Mode Switching and Insight Power Manager Processor Utilization Monitoring, and its associated SMI

Parameter	Value	Description
Memory Pre-Failure Notification	Disabled	Disables Memory Pre-Failure Notification and its associated SMI
Memory Patrol Scrubbing DL580 G7	Disabled	The 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. The default value for this parameter is Enabled.
Memory Refresh Rate	1x Refresh	This option controls the refresh rate of the memory controller. The default value for this parameter is 2x.
Memory Double Refresh DL580 G7	Disabled	This option controls the refresh rate of the DL580 G7 memory controller. The default value for this parameter is Enabled, for a 2x refresh rate.

*If Turbo mode is enabled, then step through the available cooling settings described in "Thermal considerations (on page 14)." Otherwise, the default value is adequate.

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, 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.

Processor	Power	Base frequency	Number of active cores	Turbo-enabled frequency
E5-2690	135 W	2.9 GHz	6-8	3.3 GHz
			4-5	3.4 GHz
			2-3	3.6 GHz
			1	3.8 GHz
E5-2687W	150 W	3.1 GHz	6-8	3.4 GHz
			4-5	3.5 GHz
			2-3	3.6 GHz
			1	3.8 GHz

If the penalty of computational jitter is too severe and you are unable to control temperature and TDP, you should disable Turbo mode. It is possible to maintain a constant number of active cores.

Power consumption

Pushing the processor's TDP limit will result in the processor changing its turbo frequency if the processor consumes too much power. 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, then you can disable a core per processor from within the BIOS, reducing power consumption and providing TDP headroom.

Tests have shown that the E5-2690 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 application.

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 QuickSpec. 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 always operate 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 E5-2687W has a notably higher TDP than the E5-2690, but the T_{case} for the E5-2687W is 5°C (9° F) lower than for the E5-2690, 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 HP 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 (on page 14)," 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 20)."

Disabling Processor Power and Utilization Monitoring and Memory Pre-Failure Notification SMIs

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, G7, and Gen8 servers. Disabling the Memory Pre-Failure Notification SMI has a much smaller effect because it generates an interrupt at a lower frequency: once per hour on G6 and G7 servers and once every five minutes on Gen8 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
- HP 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 uncorrectable 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 HP ProLiant DL580 G7 and HP ProLiant DL980 G7 Servers, 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. 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 affects G6, G7, and Gen8 servers. This option is available in the Service Options menu.

Setting Memory Power Savings Mode and ACPI SLIT Preferences

A new BIOS for many Gen8 platforms dated 20 August 2012, along with previous BIOS releases, provides enhancements that are of interest to Low Latency environments. Later versions of the BIOS are available, but this version is cited as the earliest version to support these settings.

Two new BIOS setting available with this release are Memory Power Savings Mode and ACPI SLIT Preferences. For more information on these settings, see "Tuning recommendations and explanations (on page 11)."

Tuning procedures

Tuning with the ROM-based Setup Utility

To configure BIOS low-latency options using RBSU:

1. Boot the server.
2. When prompted during POST, press **F9** to enter RBSU.
3. When the RBSU menu appears, press **CTRL-A** to display the option for the Service Options menu.
4. Browse through the menus to change the parameters. For more information, see "Tuning recommendations and explanations (on page 11)."

   IMPORTANT: Do not change the other options in the Services Options menu.

5. Verify that the parameters are set as indicated in "Tuning recommendations and explanations (on page 11)."

Tuning with the HP ROM Configuration Utility (Gen8 and later)

To configure BIOS low-latency options using the `hprcu` utility in STK:

1. Change the current directory to the STK/utilities directory:
   ```
   cd STK/utilities
   ```
2. Capture a snapshot of your current settings, specifying "-a" to include the Service Options settings:
   ```
   ./hprcu -a -s -f hprcu_settings.xml
   ```
   `conrep` requires editing the conrep.xml file to access the Service Options settings. With `hprcu`, it is only necessary to specify the undocumented "-a" option.
3. Edit the file `hprcu_settings.xml` to obtain the preferred settings as listed above.
4. Browse through the menus to change the parameters. For more information, see "Tuning recommendations and explanations (on page 11)."
5. Update the BIOS with the modified settings:
   ```
   ./hprcu -a -l -f hprcu_settings.xml
   ```
6. Reboot the server.

Tuning with conrep

`conrep` is a 32-bit executable and requires 32-bit libraries when run on a 64-bit operating system. For example, you may need to install the following list of packages:

- `glibc.i686`
- `nss-softokn-freebl.i686`
- `libxml2.i686`
- `libxml2-devel.i686`
- `zlib-devel.i686`
- `zlib.i686`
- `libstdc++.i686`

- `compat-libstdc++-296.i686`
- `compat-libstdc++-33.i686`

To configure BIOS low-latency options using the `conrep` utility in STK 9.40:

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_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_Double_Refresh_DL580G7">
<helptext><![CDATA[This setting allows the user to change the Memory Double
Refresh setting on the DL580 G7 server.]]></helptext>
<romfamilies>
<romfamily>P65</romfamily>
</romfamilies>
<nvram>0x5F</nvram>
<value id="0x10">Disabled</value>
<value id="0x00">Enabled</value>
<mask>0x10</mask>
</Section>
<Section name="Memory_Patrol_Scrubbing_DL580G7">
```

```
<helptext><![CDATA[This setting allows the user to change the Memory Patrol
Scrubbing setting on the DL580 G7 server.]]></helptext>
<romfamilies>
<romfamily>P65</romfamily>
</romfamilies>
<nvram>0x6F</nvram>
<value id="0x10">Disabled</value>
<value id="0x00">Enabled</value>
<mask>0x10</mask>
</Section>
```

3. Capture a snapshot of your current settings:
   `./conrep -s -x conrep.xml -f conrep_settings.xml`
4. Browse through the menus to change the parameters. For more information, see "Tuning recommendations and explanations (on page 11)."
5. Update the BIOS with the modified settings:
   `./conrep -l -x conrep.xml -f conrep_settings.xml`
6. Reboot the server:
   `reboot`

Recommended operating system tuning

Linux

RHEL and SLES

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

1. Make the following edits:
   - For RHEL systems:

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

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

   `nosoftlockup` prevents RHEL 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. Reboot the server.
3. 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_supplicant  xfs             vpbind          yum-updatesd    \
novell-zmd

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

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

a. Enter the following command:

```
# service irqbalance stop
```

b. Do a one-time run of the irq balancer:

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

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

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

Red Hat MRG Realtime

Red Hat recently 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 realtime performance. The SLUB allocator has pseudo-files named "`cpu_partial`" in the "`/sys/kernel/slab`" file system. To get the best realtime 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 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:

  Edit `/boot/grub/grub.conf` and add "idle=poll" to the kernel line. This is in addition to the "nosoftlockup intel_idle.max_cstate=0 mce=ignore_mce" parameters that should have been added previously.

- For SLES systems:

  Edit `/boot/grub/menu.lst` and add "idle=poll" to the kernel line. This is in addition to the "nosoftlockup intel_idle.max_cstate=0 mce=ignore_mce" parameters that should have been added previously.

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 HP-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-timetest7.2 -v
-f csv -o smi_count
```

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

Consider the following:

- Consider changing the smp_affinity for the IRQs. For example, on a 2p16c server on which you want to leave cores 0 and 8 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(0x00000101,
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 2p16c server on which you want to keep the OS from all cores except for 0 and 8, use the following command:
```
# cset shield --cpu 1-7,9-15 --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

HP 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 ("Obtaining the Scripting Toolkit" on page 10).
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 ("Tuning with conrep" on page 16).

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

- Review the technical information for Windows 2012 on the Microsoft website (http://technet.microsoft.com/en-us/library/hh831415.aspx).
- See the Windows Server 2012 Tuning Guide on the Microsoft website (http://msdn.microsoft.com/library/windows/hardware/jj248719).

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

HP-TimeTest

The original behavior of HP-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 HP-TimeTest to run properly.

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

```
time numactl --physcpubind=3 --localalloc     \ Bind to core 3 and use local memory
nice -n -20                                   \ nice; probably not necessary
/HP-TimeTest/HP-TimeTest7.2                   \ HP-TimeTest7.2 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 1 / p' | dc`   \ threshold is 5 µsec on 2.90 GHz processor
-l `expr 2900000000 \* 60 \* 30 / 44`         \ run for ~30 minutes on 2.90 GHz processor
                                              \ ("44" is # of cycles per loop iteration I get)

-p FIFO,80,-20                                \ Use FIFO scheduling at priority 80; 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.

The HP Low Latency team is working on an updated version of HP-TimeTest, and expects to have it ready for the next update of this document. The following changes are planned:

- Make a change to address the issue of out-of-order processors, which can cause "bleeding" of instructions past the Read Cycle Counter instruction. However, this change will probably not appear in any of the latency plots.
- Check the SMI count when latency spikes are detected, not just at the beginning and end of program execution.
- Add an option to keep the processor core as busy as possible in an attempt to consume maximum power.
- Add an option to specify a run-time instead of a loop count. Such an option might not work when run at a sufficiently high real-time priority.

To provide additional suggestions, contact the HP 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. HP 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 HP 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 11)."
- Run HP-TimeTest to see if you are getting spikes. For more information, contact HP (mailto:low.latency@hp.com). Provide your name and your company's name, as well as your mailing address and your HP contact's name.

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 HP is available on the Mellanox website (http://www.mellanox.com/related-docs/prod_software/Performance_Tuning_Guide_for_Mellanox_Network_Adapters.pdf).

Q. How does HP 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 `hprcu` or `conrep`. To modify the number of enabled cores with `conrep`, use version 3.40 or later, available from STK for Linux 9.20 or later.

Support and other resources

Resources and documentation

The following resources are available:

- *HP ROM-Based Setup Utility User Guide* on the HP website (http://www.hp.com/support/rbsu)
- iLO documentation:
  - *HP iLO 4 User Guide* (for Gen8 servers) on the HP website (http://bizsupport2.austin.hp.com/bc/docs/support/SupportManual/c03334051/c03334051.pdf)
  - *HP iLO 4 Scripting and Command Line Guide* (for Gen8 servers) on the HP website (http://bizsupport2.austin.hp.com/bc/docs/support/SupportManual/c03334058/c03334058.pdf)
  - *HP ProLiant Integrated Lights-Out 3 v1.20 User Guide* (for G7 servers) on the HP website (http://bizsupport2.austin.hp.com/bc/docs/support/SupportManual/c02774507/c02774507.pdf)
  - *HP ProLiant Integrated Lights-Out 3 v1.20 Scripting and Command Line Guide* (for G7 servers) on the HP website (http://bizsupport2.austin.hp.com/bc/docs/support/SupportManual/c02774508/c02774508.pdf)
- *HP Scripting Toolkit for Linux User Guide* on the HP website (http://h20000.www2.hp.com/bc/docs/support/SupportManual/c03297832/c03297832.pdf)
- *HP Scripting Toolkit for Windows User Guide* on the HP website (http://h20000.www2.hp.com/bc/docs/support/SupportManual/c03297836/c03297836.pdf)
- STK on the HP website (http://www.hp.com/go/support)

  The `conrep`, `hcrpu`, and `hpdiscovery` utilities are available in the STK. For more information on downloading STK, see "Obtaining the Scripting Toolkit (on page 10)."
- HP-TimeTest 7.2 utility. To obtain the utility, contact HP (mailto:low.latency@hp.com). Provide your name and your company's name, as well as your mailing address and your HP contact's name.

Before you contact HP

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

- Active Health System log (HP ProLiant Gen8 or later products)

  Download and have available an Active Health System log for 3 days before the failure was detected. For more information, see the *HP iLO 4 User Guide* or *HP Intelligent Provisioning User Guide* on the HP website (http://www.hp.com/go/ilo/docs).
- Onboard Administrator SHOW ALL report (for HP BladeSystem products only)

For more information on obtaining the Onboard Administrator SHOW ALL report, see the HP website (http://www.hp.com/go/OAlog).

- 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

HP contact information

For United States and worldwide contact information, see the Contact HP website (http://www.hp.com/go/assistance).

In the United States:

- To contact HP by phone, call 1-800-334-5144. For continuous quality improvement, calls may be recorded or monitored.
- If you have purchased a Care Pack (service upgrade), see the Support & Drivers website (http://www8.hp.com/us/en/support-drivers.html). If the problem cannot be resolved at the website, call 1-800-633-3600. For more information about Care Packs, see the HP website (http://pro-aq-sama.houston.hp.com/services/cache/10950-0-0-225-121.html).

On a best effort basis only, HP offers technical assistance on low-latency tuning to customers who have followed this guide and still have questions. For more information, contact HP (mailto:low.latency@hp.com). Provide your name and your company's name, as well as your mailing address and your HP contact's name.

Acronyms and abbreviations

ACPI

Advanced Configuration and Power Interface

AMP

Advanced Memory Protection

HPRCU

HP ROM Configuration Utility

iLO

Integrated Lights-Out

LOM

LAN on Motherboard

MRG

Messaging, Realtime, and Grid

POST

Power-On Self Test

RBSU

ROM-Based Setup Utility

SLERT

SUSE Linux Enterprise Real Time Extension

SLIT

System Locality Information Table

SMI

System Management Interrupt

STK

Scripting Toolkit

TDP

Thermal Design Power

Documentation feedback

HP is committed to providing documentation that meets your needs. To help us improve the documentation, send any errors, suggestions, or comments to Documentation Feedback (mailto:docsfeedback@hp.com). Include the document title and part number, version number, or the URL when submitting your feedback.