HPE ProLiant Server Power Management

Red Hat Enterprise Linux 6.x and 7.x and
SUSE Linux Enterprise Server 11 and 12

Contents

Technical white paper

Abstract

Power management is crucial to data center power provisioning. This document provides a brief overview of the processor-based power-saving features supported on HPE ProLiant servers, and the power management features that are embedded in HPE ProLiant servers. This document also discusses how these features are used and their relationship to the Red Hat® Enterprise Linux® 6.x (RHEL 6.x), RHEL 7.x, SUSE Linux Enterprise Server 11 (SLES 11), and SLES 12 operating systems. It includes new features available with HPE ProLiant Gen10 and later Intel®-based servers.

Introduction

The RHEL 6.x, RHEL 7.x, SLES 11, and SLES 12 operating systems running on HPE ProLiant servers use processor-based features to achieve better power efficiency. These processor-based features include the following:

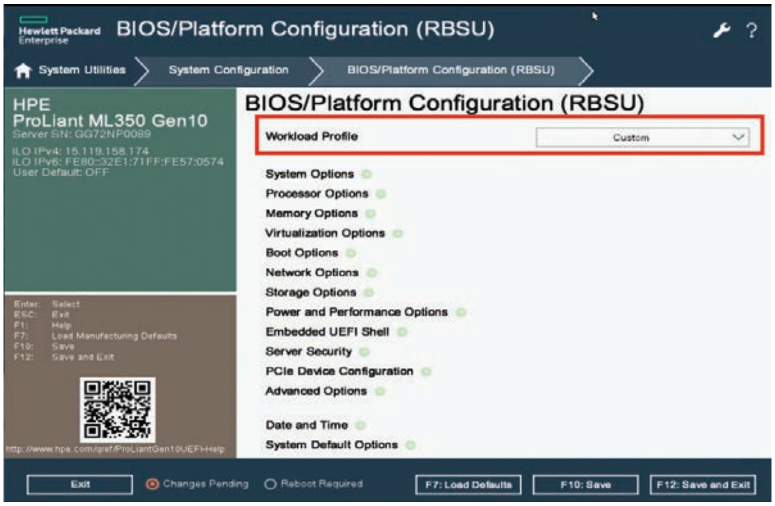
- Performance states (P-states) define a set of fixed operating frequencies and voltages, where P0 represents the highest operating frequency and voltage. You can save power by entering P-states with lower frequency and voltage levels. Either the platform firmware or the operating system controls the P-states.
- Power states (C-states), excluding the C0 state, represent idle states and determine the power consumed when a processor is idle. C0 is a non-idle state with higher C-states representing idle conditions with increasing power savings. The operating system controls the C-states.
- Throttle states (T-states) define a set of fixed frequency percentages that can be used to regulate the power consumption and the thermal properties of the processor. HPE ProLiant servers can reserve the use of T-states for the system firmware.

In addition, HPE ProLiant servers are also capable of using the various processor states to support innovative power management features that are operating system independent and are implemented in the hardware and firmware:

- HPE Power Regulator provides a facility to efficiently control processor power usage and performance, either statically or dynamically, depending on the mode selected.

HPE Workload Profile

HPE Workload Profiles as shown in Figure 1 are a configurable option to deploy BIOS settings based on the workload customer intends to run on the server. HPE ProLiant Gen10 server provided **Workload Profiles** contain settings for the most common workload categories in ROM Based Setup Utility (RBSU). Each profile is designed to obtain specific performance results. Not all profiles require that the options be set to specific settings and only the dependent settings for that profile are changed in **Power and Performance Options**. For example, when Workload Profile is in **General Power Efficient Compute**, the HPE Power Regulator is configured to **Dynamic Power Saving Mode** for power efficient computing management. For more information of Workload Profile settings, see the UEFI System Utilities User Guide for HPE ProLiant Gen10 Servers and HPE Synergy.



BIOS/Platform Configuration (RBSU)

HPE
ProLiant ML350 Gen10
Server SN: GG72NP0089
ILO IPv4: 15.119.158.174
ILO IPv6: FE80::32E1:71FF:FE57:0574
User Default: OFF

Workload Profile: Custom

Figure 1. Configuring Workload Profile settings via HPE RBSU

HPE Power Regulator

HPE Power Regulator is a configurable processor power-usage feature that allows you to choose from several options for (1) enabling the server to manage P-states or (2) delegating control of regulating P-states to the operating system.

HPE Power Regulator is implemented within the firmware on Intel-based HPE ProLiant server.

HPE ProLiant Gen10 servers provide the following HPE Power Regulator modes, which you can select from the ROM Based Setup Utility (RBSU) or through HPE Integrated Lights Out 5 (iLO 5).

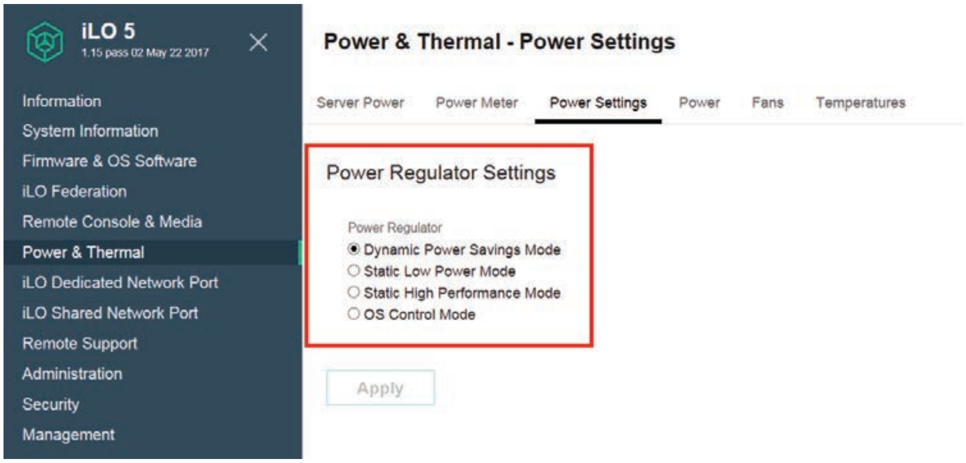
Table 1. HPE Power Regulator modes for HPE ProLiant servers

HPE Power Regulator Mode	Description
Dynamic Power Savings Mode	The firmware is capable of managing the P-states. However, when the Collaborative Power Control (CPC) setting is enabled in RBSU, the OS and the firmware collaborate to attain the desired frequency for a processor. When CPC is disabled, this mode allows the firmware to exclusively control the P-states of a processor to match the server load. On HPE ProLiant Gen10 server, Dynamic Power Saving Mode is the default mode with the CPC setting enabled.
Static Low Power Mode	The firmware controls the P-states. The P-state of the processor is static, and it is set to the P-state that corresponds to the lowest operating frequency supported by the processor.
Static High Performance Mode	The firmware controls the P-states. The P-state of the processor is static, and it is set to P0, which corresponds to the highest operating frequency supported by the processor.
OS Control Mode	The Linux operating systems control P-states and manage the P-states according to the policy set by the administrator via the OS.

For the Static Low Power Mode and Static High Performance Mode described previously, Hewlett Packard Enterprise recommends that you disable CPC to ensure that the firmware has exclusive control of the P-states. CPC is located within the **Power and Performance Options** in RBSU.

The OS Control Mode allows HPE ProLiant server firmware to delegate management of P-states to the operating systems. For general-purpose workloads where a balance of high performance and power efficiency is desired, OS Control Mode is recommended.

You can adjust the HPE Power Regulator settings through the HPE iLO 5 interface as shown in Figure 2 or the RBSU as shown in Figure 3. You must reboot the system to change the transitions to and from the OS Control Mode, but you can change the system between the other three modes dynamically through HPE iLO 5 interface.

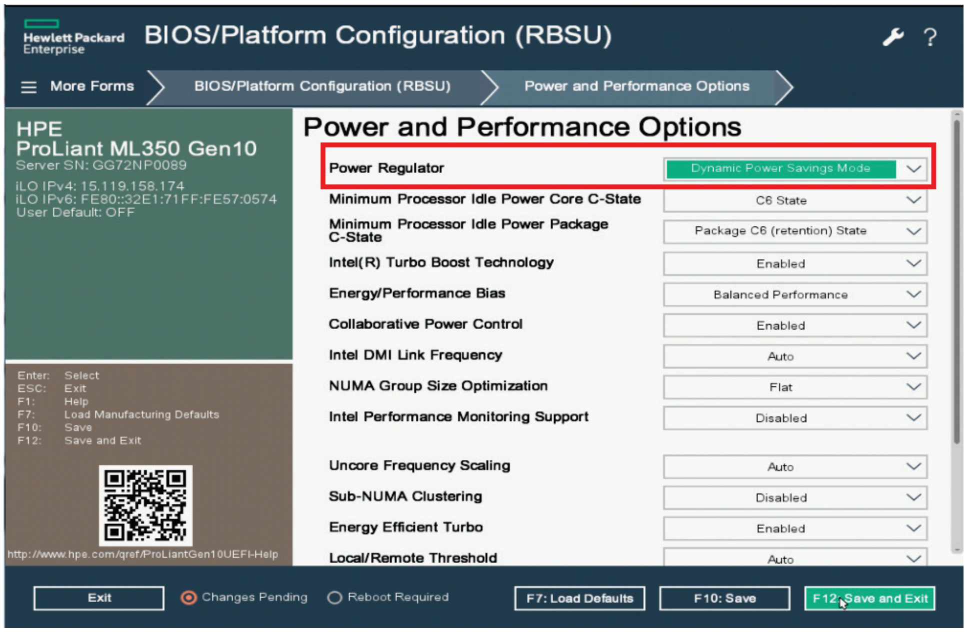


Power & Thermal - Power Settings

Power Regulator Settings

Power Regulator
● Dynamic Power Savings Mode
○ Static Low Power Mode
○ Static High Performance Mode
○ OS Control Mode

Figure 2. Configuring Power Regulator Settings via HPE iLO 5


Hewlett Packard Enterprise
BIOS/Platform Configuration (RBSU)
More Forms
BIOS/Platform Configuration (RBSU)
Power and Performance Options
Enter: Select
ESC: Exit
F1: Help
F7: Load Manufacturing Defaults
F10: Save
F12: Save and Exit
http://www.hpe.com/qref/ProLiantGen10UEFI-Help
Exit
Changes Pending
Reboot Required
F7: Load Defaults
F10: Save
F12: Save and Exit

HPE
ProLiant ML350 Gen10
Server SN: GG72NP0089
iLO IPv4: 15.119.158.174
iLO IPv6: FE80::32E1:71FF:FE57:0574
User Default: OFF

Power and Performance Options

Power Regulator	Dynamic Power Savings Mode
Minimum Processor Idle Power Core C-State	C6 State
Minimum Processor Idle Power Package C-State	Package C6 (retention) State
Intel(R) Turbo Boost Technology	Enabled
Energy/Performance Bias	Balanced Performance
Collaborative Power Control	Enabled
Intel DMI Link Frequency	Auto
NUMA Group Size Optimization	Flat
Intel Performance Monitoring Support	Disabled
Uncore Frequency Scaling	Auto
Sub-NUMA Clustering	Disabled
Energy Efficient Turbo	Enabled
Local/Remote Threshold	Auto

Figure 3. Configuring Power Regulator settings via HPE RBSU

HPE ProLiant Power Management

Linux operating system manages the power usage of HPE ProLiant servers by adjusting the processor P-states when the HPE Power Regulator setting in RBSU is configured in OS Control Mode. Typically, within the Linux operating system, a governor[1] dictates the policy, while the actual P-state transition is accomplished by a suitable P-state driver. Linux operating system offers a choice of governors, each implementing a different policy.

For **Red Hat Enterprise Linux 6**, the default P-state driver is “acpi-cpufreq” P-state driver and the default governor for “acpi-cpufreq” P-state driver is the on-demand governor, which dynamically adjusts the processor P-states to match the load on the server. The other governors of “acpi-cpufreq” P-state driver in RHEL 6 are

- Userspace, which enables the user space program (cpuspeed)
- Performance, which selects the P-state corresponding to the highest supported frequency

For **SUSE Linux Enterprise Server 11**, the default P-state driver is “acpi-cpufreq” P-state driver and the default governor for “acpi-cpufreq” P-state driver is the on-demand governor. The other governors of acpi-cpufreq P-state driver in SLES 11 are

- Userspace, which enables the user space program (cpuspeed).
- Performance, which selects the P-state corresponding to the highest supported frequency.
- Conservative, which much like the “ondemand” governor, sets the CPU frequency depending on the current usage. It differs in behavior in that it gracefully increases and decreases the CPU speed rather than jumping to max. speed the moment there is any load on the CPU. This behavior is more suitable in a battery-powered environment.
- Powersave, which forces the CPU to run at its minimum allowed frequency constantly.

[1] Governor defines the power characteristic of the system CPU, which affects CPU performance, and it is used to decide what frequency should be set within the CPUfreq policy.

For **Red Hat Enterprise Linux 7 and SUSE Linux Enterprise Server 12**, the default P-state driver is “intel_pstate” P-state driver. For “intel_pstate” P-state driver, there are two governors, one is “performance”, which selects the P-state corresponding to the highest processor frequency and the other is “powersave”, which sets the CPU to run the minimum allowed frequency constantly. The default governor for RHEL 7.x is “performance” and the default governor for SLES 12 is “powersave”.

The following tables list the P-state drivers on Intel-based platforms under OS Control Mode in different Linux operating system.

Table 2. P-state drivers under OS Control Mode for RHEL 6.x and RHEL 7.x

Processor family	P-state driver (RHEL 6.x)	P-state driver (RHEL 7.x)
Intel® Xeon®	acpi-cpufreq	intel_pstate

Table 3. P-state drivers under OS Control Mode for SLES 11 and SLES 12

Processor family	P-state driver (SLES 11)	P-state driver (SLES 12)
Intel Xeon	acpi-cpufreq	intel_pstate

On Intel-based HPE ProLiant platforms, Linux operating system natively supports the Intel Demand-Based Switching with Enhanced Intel SpeedStep® Technology.

To manage the processor’s power consumption, the firmware must communicate information about the processor P-states and their associated frequencies to the OS. These information are available in the files and directories under “/sys/devices/system/cpu”.

For Linux operating system, the “cpupower” tool can be used to provide information about P-states of the processors in the system in a user-friendly format. When invoked with “cpupower –c all frequency-info”, it will display information about all processor cores. This includes the P-state driver, the frequency range defined by the processor, the available frequency steps (that is, the P-states), the available and current governors, and the current frequency. It also provides options to display specific information or information for specific CPUs as illustrated in the sample output. Example 1 shows how cpupower supports options to display information specific to CPU 0 in RHEL 7.3. Example 2 shows how cpupower supports options to display information specific to CPU 0 in SLES 11 SP4 environment.

Example 1. Output for CPU 0 in OS Control Mode (RHEL 7.3)

```
# cpupower -c 0 frequency-info
analyzing CPU 0:
   driver: intel_pstate
   CPUs which run at the same hardware frequency: 0
   CPUs which need to have their frequency coordinated by software: 0
   maximum transition latency: 0.97 ms.
   hardware limits: 1.20 GHz - 2.50 GHz
   available cpufreq governors: performance, powersave
   current policy: frequency should be within 1.20 GHz and 2.50 GHz.
      The governor “performance” may decide which speed to use
      within this range.
current CPU frequency is 1.95 GHz (asserted by call to hardware).
boost state support:
   Supported: yes
Active: yes
```

Example 2. Output for CPU 0 in OS Control Mode (SLES 11 SP4)

```
# cpupower -c 0 frequency-info
analyzing CPU 0:
  driver: acpi-cpufreq
  CPUs which run at the same hardware frequency: 0
  CPUs which need to have their frequency coordinated by software: 0
  maximum transition latency: 10.0 us.
  hardware limits: 1000 MHz - 1.80 GHz
  available frequency steps: 1.80 GHz, 1.80 GHz, 1.70 GHz, 1.60 GHz, 1.50 GHz, 1.40 GHz, 1.30 GHz,
1.20 GHz, 1.10 GHz, 1000 MHz
  available cpufreq governors: conservative, userspace, powersave, ondemand, performance
  current policy: frequency should be within 1000 MHz and 1.80 GHz.
                  The governor “ondemand” may decide which speed to use
                  within this range.
  current CPU frequency is 1000 MHz (asserted by call to hardware).
  boost state support:
    Supported: yes
    Active: yes
```

The governor used under the OS Control Mode can be changed dynamically by modifying the value in the “/sys/devices/system/cpu/cpu*/cpufreq/scaling_governor” for each CPU. Also, using “cpupower frequency-set” with the “–g” option can switch the governor at runtime. For more information about cpupower commands, see the man pages.

Idle power states (C-states)

Processor power use at idle is a crucial factor in determining power consumption of a server when there is no workload to execute. Typically, when a processor has no work to perform, the operating system places the processor in a halt state signified as C1. Newer-generation processors support deep C-states (C2 to C6, where C6 is the deepest state), allowing Linux operating system to take advantage of these states. The deeper the C-state, the more power the processor can save. Although C-states can significantly reduce power consumption, the drawback of going to a deeper C-state is the latency associated with the time it takes for the processor to wake up and resume executing instructions. Information about the C-states for system processors is available in “/sys/devices/system/cpu/cpu*/cpuidle/state*”.

Note
You can configure the server to not utilize the idle C-states by choosing the **No C-states** setting in RBSU.

Additional power management features

RHEL 6.x, RHEL 7.x, SLES 11, and SLES 12 provide a comprehensive set of features for managing the power usage of HPE ProLiant servers.

Tickless idle is a kernel feature that eliminates periodic timer ticks when system CPUs are idle. The Linux kernel uses periodic timer events, called timer ticks, for each CPU. These ticks facilitate process accounting, scheduler load balancing, and maintaining per-CPU timer events. However, one drawback is that when systems are idle, these periodic ticks wake up the system and bring it out of its power saving “sleep” state—roughly every millisecond for newer kernels operating at 1000 Hz. This causes unnecessary power consumption. With tickless idle, the periodic timer tick has been eliminated when the CPU is idle. This allows the CPU to remain in power saving states for longer periods. The result is lower overall system power consumption and reduced costs.

Additional tools are available for monitoring system power consumption. For example, using the PowerTOP[2] tool, you can identify the processes responsible for waking up a processor from its idle state, and thereby drive up power consumption. You can see the PowerTOP documentation for further information on what the output of PowerTOP represents, and learn tips and tricks on how to best tune your server for maximum power savings.

The cpupower tools can not only be used to give an overview of all CPU power-related parameters that are supported on a given machine but also the integrated monitoring framework. It can access both kernel-related parameters as well as hardware statistics and thus ideally be suited for performance benchmarks. Example 3 shows how cpupower with subcommand monitor to report processor topology, and monitor frequency and idle power state statistics. For detailed information, see the cpupower monitor man page.

Example 3. Output for cpupower monitor in non-idle state

```
# cpupower monitor
    |Nehalem                    || Mperf              || Idle_Stats
CPU | C3   | C6   | PC3  | PC6  || C0   | Cx   | Freq || POLL | C1-S | C1E- | C6-S
   0|  0.00|  0.00|  0.00|  0.00|| 27.35| 72.65|  2507||  0.00| 16.65| 56.92|  0.05
  20|  0.00|  0.00|  0.00|  0.00|| 25.24| 74.76|  2512||  0.03|  8.74| 66.99|  0.02
   1|  0.00|  0.00|  0.00|  0.00|| 25.46| 74.54|  2599||  0.00| 15.50| 59.86|  0.14
  21|  0.00|  0.00|  0.00|  0.00|| 24.45| 75.55|  2596||  0.00|  1.95| 74.53|  0.02
   2|  0.00|  0.00|  0.00|  0.00|| 25.67| 74.33|  2627||  0.00|  9.98| 65.16|  0.11
  22|  0.00|  0.00|  0.00|  0.00|| 25.22| 74.78|  2632||  0.00|  9.11| 66.52|  0.08
   3|  0.00|  0.00|  0.00|  0.00|| 27.44| 72.56|  2651||  0.01| 14.78| 58.40|  0.28
  23|  0.00|  0.00|  0.00|  0.00|| 29.09| 70.91|  2662||  0.00|  8.18| 63.44|  0.11
   4|  0.00|  0.00|  0.00|  0.00|| 32.28| 67.72|  2690||  0.01| 10.86| 57.67|  0.08
  24|  0.00|  0.00|  0.00|  0.00|| 25.02| 74.98|  2692||  0.00|  8.77| 67.11|  0.09
   5|  0.00|  0.00|  0.00|  0.00|| 28.29| 71.71|  2691||  0.00| 15.53| 56.84|  0.09
  25|  0.00|  0.00|  0.00|  0.00|| 99.72|  0.28|  2692||  0.00|  0.08|  0.27|  0.00
   6|  0.00|  0.00|  0.00|  0.00|| 24.64| 75.36|  2687||  0.03| 12.66| 63.61|  0.07
  26|  0.00|  0.00|  0.00|  0.00|| 28.46| 71.54|  2688||  0.00|  8.39| 63.93|  0.11
   7|  0.00|  0.00|  0.00|  0.00|| 26.07| 73.93|  2516||  0.00|  9.48| 65.29|  0.08
  27|  0.00|  0.00|  0.00|  0.00|| 24.84| 75.16|  2515||  0.00| 10.05| 65.93|  0.13
   8|  0.00|  0.00|  0.00|  0.00|| 68.26| 31.74|  2692||  0.00|  2.05| 29.96|  0.05
  28|  0.00|  0.00|  0.00|  0.00|| 30.39| 69.61|  2691||  0.00| 11.89| 58.51|  0.08
```

[2] An introduction to PowerTOP: 01.org/powertop.

```
   9|  0.00|  0.00|  0.00|  0.00|| 76.06| 23.94|  2443||  0.00|  2.76| 21.42|  0.07
  29|  0.00|  0.00|  0.00|  0.00|| 28.17| 71.83|  2491||  0.00| 13.11| 59.58|  0.04
  10|  0.00|  0.00|  0.00|  0.00|| 25.83| 74.17|  2627||  0.00| 12.54| 62.47|  0.09
  30|  0.00|  0.00|  0.00|  0.00|| 25.99| 74.01|  2619||  0.00| 10.00| 64.88|  0.06
  11|  0.00|  0.00|  0.00|  0.00|| 25.35| 74.65|  2603||  0.00| 11.38| 64.18|  0.07
  31|  0.00|  0.00|  0.00|  0.00|| 25.33| 74.67|  2614||  0.00|  9.23| 66.37|  0.03
  12|  0.00|  0.00|  0.00|  0.00|| 27.47| 72.53|  2687||  0.00|  5.57| 67.66|  0.21
  32|  0.00|  0.00|  0.00|  0.00|| 24.55| 75.45|  2689||  0.00|  9.12| 67.33|  0.04
  13|  0.00|  0.00|  0.00|  0.00|| 25.62| 74.38|  2662||  0.00| 14.88| 60.20|  0.28
  33|  0.00|  0.00|  0.00|  0.00|| 24.07| 75.93|  2665||  0.00|  9.99| 66.93|  0.06
  14|  0.00|  0.00|  0.00|  0.00|| 79.56| 20.44|  2692||  0.00|  2.98| 17.78|  0.00
  34|  0.00|  0.00|  0.00|  0.00|| 26.79| 73.21|  2692||  0.00| 10.65| 63.53|  0.10
  15|  0.00|  0.00|  0.00|  0.00|| 25.67| 74.33|  2687||  0.00| 11.08| 63.91|  0.30
  35|  0.00|  0.00|  0.00|  0.00|| 24.01| 75.99|  2686||  0.00| 10.06| 66.78|  0.11
  16|  0.00|  0.00|  0.00|  0.00|| 25.52| 74.48|  2606||  0.00| 10.66| 64.71|  0.11
  36|  0.00|  0.00|  0.00|  0.00|| 25.30| 74.70|  2603||  0.00| 10.70| 64.88|  0.09
  17|  0.00|  0.00|  0.00|  0.00|| 25.98| 74.02|  2638||  0.00|  9.37| 65.55|  0.09
  37|  0.00|  0.00|  0.00|  0.00|| 24.79| 75.21|  2643||  0.00| 12.26| 63.78|  0.12
  18|  0.00|  0.00|  0.00|  0.00|| 25.67| 74.33|  2620||  0.00|  9.84| 65.36|  0.10
  38|  0.00|  0.00|  0.00|  0.00|| 25.69| 74.31|  2595||  0.00| 10.32| 64.95|  0.07
  19|  0.00|  0.00|  0.00|  0.00|| 37.64| 62.36|  2486||  0.00|  9.91| 53.11|  0.13
  39|  0.00|  0.00|  0.00|  0.00|| 25.83| 74.17|  2596||  0.00| 10.63| 64.45|  0.09
```

On Intel-based HPE ProLiant servers, Intel processors may contain support for Hardware-Controlled Performance States (HWP), which autonomously selects performance states while utilizing OS supplies performance guidance hints. When HWP is enabled, the processor autonomously selects performance states as deemed appropriate for the applied workload and with consideration of constraining hints that are programmed by the OS. These OS-provided hints include minimum and maximum performance limits, preference toward energy efficiency or performance, and the specification of a relevant workload history observation time window. The means for the OS to override HWP's autonomous selection of performance state with a specific desired performance target is also provided; however, the effective frequency delivered is subject to the result of energy efficiency and performance optimizations.

Summary

HPE ProLiant servers is designed to save power when under load and when idle. The processor-based power management features supported in the hardware are enabled by the firmware automatically. They are also used in close coordination between the server's firmware and the Linux operating systems. Typically, you do not have to activate these features; they are enabled by default.

Appendix A: Table for default P-state driver on HPE ProLiant Gen10 servers

RHEL 7.x and SLES 12

Table 4. Default P-state driver on HPE ProLiant Gen10 server (RHEL 7.x and SLES 12)

Collaborative Power Control (CPC)	HPE Power Regulator	
	OS Control Mode	**Other modes**[3]
CPC enabled	Intel P-state driver	Intel P-state driver
CPC disabled	Intel P-state driver	No driver[4]

RHEL 6.x and SLES 11

Table 5. Default P-state driver on HPE ProLiant Gen10 server (RHEL 6.x and SLES 11)

Collaborative Power Control (CPC)	HPE Power Regulator	
	OS Control Mode	**Other modes**[5]
CPC enabled	acpi-cpufreq driver	No driver[6, 7]
CPC disabled	acpi-cpufreq driver	No driver[8]

Appendix B: References

For additional information, see the resources listed in the following table.

Resource description	Web address
HPE ProLiant Gen10 servers	hpe.com/us/en/servers/gen10-servers.html
Enhanced Intel SpeedStep Technology and Demand-Based Switching on Linux	software.intel.com/en-us/articles/enhanced-intel-speedstepr-technology-and-demand-based-switching-on-linux
Linux cpufreq kernel documentation	kernel.org/doc/Documentation/cpu-freq/
Linux cpuidle kernel documentation	kernel.org/doc/Documentation/cpuidle/
UEFI System Utilities User Guide for HPE ProLiant Gen10 Servers and HPE Synergy	hpe.com/support/UEFIGen10-UG-en
HPE Integrated Lights Out (iLO) 5 User Guide	hpe.com/support/ilo5-ug-en
An introduction to PowerTOP	01.org/powertop
RHEL 6 Power Management Guide	access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html-single/Power_Management_Guide/index.html
RHEL 7 Power Management Guide	access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/7/html/Power_Management_Guide/index.html
SLES 11 System Analysis and Tuning Guide	suse.com/documentation/sles11/pdfdoc/book_sle_tuning/book_sle_tuning.pdf
SLES 12 System Analysis and Tuning Guide	suse.com/documentation/sles-12/pdfdoc/book_sle_tuning/book_sle_tuning.pdf

[3, 5] Other modes are "Dynamic Power Savings Mode," "Static High Performance Mode," and "Static Low Power Mode."

[4, 6, 8] No driver means there is no CPU scaling driver in the operating system. Since there is no CPU scaling driver loaded, there is no CPUfreq folder listed in "/sys/devices/system/cpu/cpu*/".

[7] pcc_cpufreq driver is no longer to be used for HPE Gen10 platform; therefore, there is no CPUfreq driver when CPC is enabled in other modes with RHEL 6.x and SLES 11 SPx.

Appendix C: Known issues

Issue description	Symptom	Comment
scaling_cur_freq file is missing in Red Hat Enterprise Linux 7.3	scaling_cur_freq file can't be found in the cpufreq folder "/sys/devices/system/cpu/[cpuid]/cpufreq/" when intel_pstate driver is loaded in RHEL 7.3.	Generic issue for RHEL 7.3, not HPE specific. Fixed in RHEL 7.4. Red Hat Knowledgebase: access.redhat.com/solutions/3116471

Next steps

Hewlett Packard Enterprise welcomes your feedback. To make comments and suggestions about product documentation, send a message to techdocs_feedback@hpe.com and include the document title and part number if available in your message. All submissions become the property of Hewlett Packard Enterprise.

Learn more at

hpe.com/us/en/servers/gen10-servers.html

a00020432ENW, July 2017

Hewlett Packard Enterprise