TECHNOLOGY GUIDE

Intel Corporation

Power Management – Technology Overview

Authors

Kevin Devey

David Hunt

Chris MacNamara

1 Introduction

This document provides a high-level overview of many of the power management features in the 3rd Generation Intel® Xeon® Scalable processors, as well as guidance on how these features can be integrated at a platform level.

Power management features can enable system performance to be maintained while reducing overall energy consumption, or alternatively enable greater system performance at existing power consumption levels. Many of the features can be managed in runtime, allowing fine grain reactions based off system load. These features are exposed to standard orchestration software, thereby enabling feature management to be integrated into existing integration flows.

This technology guide is intended to provide an overview of these technologies. More in-depth enabling details for these technologies are available as part of the Network Transformation Experience Kit.

This document is part of the Network Transformation Experience Kit, which is available at https://networkbuilders.intel.com/network-technologies/network-transformation-exp-kits.

NOTE: The general information contained within this document applies to both the 3rd Generation Intel® Xeon® Scalable processor and the Intel® Xeon® D processor. However, please note that the performance, configurations, and feature set in this document apply specifically to the 3rd Gen Intel® Xeon® Scalable processor and may vary for the Intel® Xeon® D processor.

Table of Contents

Figures

Tables

Document Revision History

REVISION	DATE	DESCRIPTION
001	June 2021	Initial release.
002	February 2022	Added a note regarding Intel® Xeon® D processor in the Introduction section.

1.1 Terminology

Table 1. Terminology

ABBREVIATION	DESCRIPTION
ACPI	Advanced Configuration and Power Interface
CMK	CPU Manager for Kubernetes
DPDK	Data Plane Development Kit (DPDK)
HWP	Hardware-Controlled Performance States
MSR	Model Specific Registers
NFD	Node Feature Discovery
PCU	Power Controller Unit
SST-BF	Intel® Speed Select Technology – Base Frequency (Intel® SST-BF)
SST-CP	Intel® Speed Select Technology – Core Power (Intel® SST-CP)
SST-PP	Intel® Speed Select Technology – Performance Profile (Intel® SST-PP)
SST-TF	Intel® Speed Select Technology – Turbo Frequency (Intel® SST-TF)
TDP	Thermal Design Power
UEFI	Unified Extensible Firmware Interface

1.2 Reference Documentation

Table 2. Reference Documents

REFERENCE	SOURCE
Intel® Speed Select Technology (Intel® SST)	https://www.intel.com/content/www/us/en/architecture-and-technology/speed-select-technology-article.html
Intel® Speed Select Technology - Base Frequency - Enhancing Performance Application Note	https://networkbuilders.intel.com/solutionslibrary/intel-speed-select-technology-base-frequency-enhancing-performance
Intel® Speed Select Technology – Base Frequency (Intel® SST-BF) with Kubernetes* Application Note	https://networkbuilders.intel.com/solutionslibrary/intel-speed-select-technology-base-frequency-with-kubernetes-application-note
Intel® Speed Select Technology - Base Frequency Configuration Automation on OpenStack* Compute Host Application Note	https://networkbuilders.intel.com/solutionslibrary/intel-speed-select-technology-base-frequency-configuration-automation-on-openstack-compute-host
Intel® Speed Select Technology - Base Frequency Priority CPU Management for Open vSwitch* (OVS*) Application Note	https://networkbuilders.intel.com/solutionslibrary/intel-speed-select-technology-base-frequency-priority-cpu-management-for-open-vswitch
Power Management - Enhanced Power Management for Low-Latency Workloads Technology Guide	https://networkbuilders.intel.com/solutionslibrary/power-management-enhanced-power-management-for-low-latency-workloads-technology-guide
Intel® Speed Select Technology – Core Power – Overview Technology Guide	https://networkbuilders.intel.com/solutionslibrary/intel-speed-select-technology-core-power-intel-sst-cp-overview-technology-guide

2 Overview

Two key objectives of many workload deployments are to maximize performance per watt and to reduce its overall carbon footprint through reduced energy consumption over its lifetime. Intel has continued to expand its suite of power management technologies to help achieve both goals. These technologies allow deployments to be tuned for optimal performance per watt at peak and sustained loads, while also enabling significant power savings when the deployment is lightly loaded, reducing overall energy consumption through active power management.

This document provides an overview on many of the power management technologies that Intel has brought to its product portfolio, as well as how they can be used to improve performance per watt and meet corporate sustainability goals.

2.1 P-states

2.1.1 P-states: Review

A Performance State (P-state) is an operating point where a core is operating at a specific frequency and voltage while executing instructions. As the core moves to lower P-states, the frequency and associated voltage drops resulting in reduced power consumption and performance. The Advanced Configuration and Power Interface (ACPI) defines the performance states that are exposed to system software. Software interfaces provided by Enhanced Intel SpeedStep® Technology allow P-states to be controlled per core.

P-states have an associated frequency that varies between SKUs, organized as follows:

- P01 is the Max 1 Core Turbo Frequency, the maximum frequency that can be reached with one core active.
- P0n is the All Core Turbo Boost Frequency range. The level of Turbo Frequency depends on the workload and the operating environment. Turbo is opportunistic as the frequency achieved may fall short of the maximum frequency. See Section 2.3 for further information on Turbo Boost.
- P1 is the Guaranteed Base Frequency of a SKU. All cores can run at this speed while within standard operating conditions. This is sometimes referred to as P1n.
- P2, P3, and all lower P-states are defined as 100 MHz (referred to as a 'bin') below the previous P-state's frequency.
- Pn is the lowest P-state supported by the CPU.

Figure 1 shows an example of the P-state naming for a sample SKU.

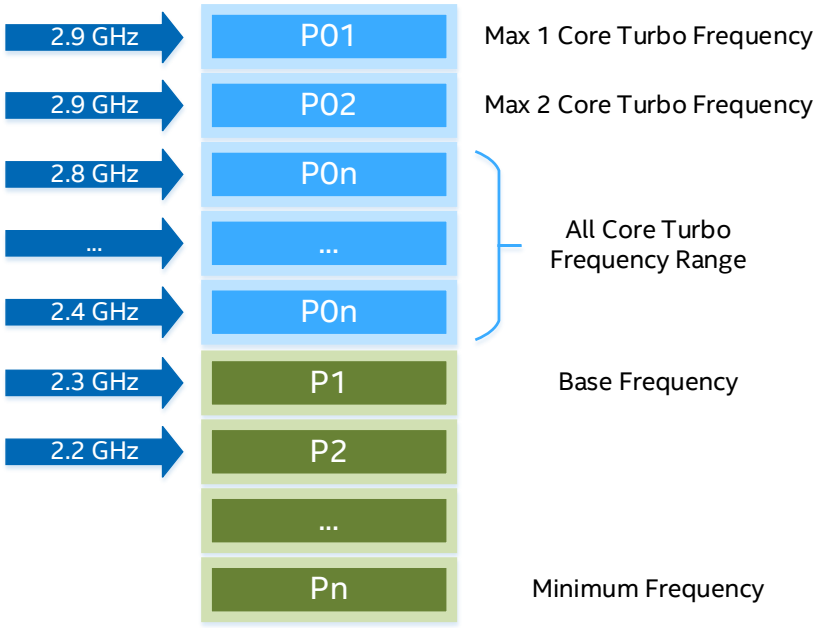

Figure 1. Core P-states Naming

2.1.2 P-states: Management

Hardware: In Intel® Xeon® Scalable processors, Hardware-Controlled Performance States (HWP) give the option to pass P-state control to the hardware, allowing it to make much faster and finer adjustments to the frequency and voltage based on each core's load and OS hints. Without HWP, the OS can directly request P-states through the intel_pstate driver or the acpi_cpufreq driver.

Operating System: P-states can be managed from user space via the Linux kernel system file system "sysfs". By default, the P-states have a management routine known as a governor that decides how to control the frequency in response to the workload.

Orchestration: No orchestration support for P-state control.

DPDK Application: The DPDK Power Management library allows user space applications to save power by dynamically adjusting CPU frequency. The librte_power API is available to request a specific P-state configuration from the kernel. In addition, Poll Mode Drivers are enabled with frequency scaling support.

2.1.3 P-states: Key Takeaway

P-states can be used to dynamically reduce frequency per core, reducing the power consumption of the server.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel Field Applications Engineer (FAE) for more information.

2.2 C-states

2.2.1 C-states: Review

C-states are power states that a CPU can use to reduce power consumption on a per-core level, or on a CPU package level, by powering down portions of the core, package, or both. Disabling portions of the core allows for large power savings but prevents the core from executing instructions.

3rd Generation Intel® Xeon® Scalable processors support several core and package C-states as described below. Note that other processor families may support other C-states, such as C0.1 and C0.2, as defined in Intel® 64 and IA-32 Architectures Software Developer Manuals.

- **Core C-states:**
  - Available States: C0, C1, C1E, C6.
  - C0 is the active state where instructions are executed. No instructions are executed in other Core C-states.
  - Lower C-states (C6 being the lowest) are power optimized, resulting in greater power savings and higher exit latency.
  - If a core reaches C6, the L1 and L2 Caches are flushed to L3 Cache, and data may need to be reloaded into cache after the core exits the power-optimized state.
- **Package C-states[1]:**
  - Available States: PC0, PC1E, PC2, PC6.
  - PC0 is the active state where one or more cores is actively executing instructions. All other PC-states require all cores to be in C6.
  - Lower PC-states (PC6 being the lowest) result in greater power savings and higher exit latency.
  - All PC-states except PC6 still allow snoop, memory, and other traffic to cross the CPU. If a snoop request comes in while the CPU is in PC6, it wakes and moves to PC2 to service the request.

Power savings achieved by these states vary across CPU SKUs and product generations, while the exit latencies vary depending on product generation.

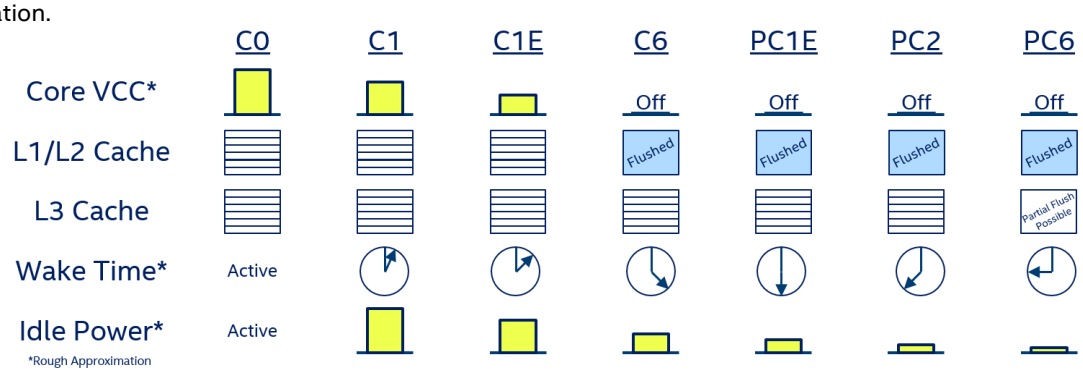

Figure 2. C-state Characteristics

2.2.2 C-states: Management

Hardware: In certain configurations, the Power Control Unit (PCU) in the CPU is responsible for autonomously coordinating core and package C-states while BIOS configuration allows you to limit the C-states available to the platform, ensuring it never goes below your required C-state. Alternatively, intel_idle and the Linux scheduler govern C-states.

Operating System: Core C-states are controlled by the OS as defined by the ACPI Specification. The OS can tell the threads in a core to go to a particular C-state using the MWAIT and MONITOR instruction. In supported CPUs, application software can use the waitpkg instructions to request power-optimized states C0.1 and C0.2.

Orchestration: There is no orchestration level C-state management.

Application: There is no direct application level C-state entry/exit control. Software techniques can be used to influence kernel control of C-states.

2.2.3 C-states: Key Takeaway

C-states can be used to move individual cores or the full CPU Package to an idle state, reducing the power consumption of the server.

[1] Note: The PCx states described here are processor states, and nomenclature varies. For example, ACPI translates processor states to ACPI-defined values.

Further information can be found in the *External Design Specification Volume 1*. Contact your Intel FAE for more information.

2.3 Intel® Turbo Boost Technology

2.3.1 Intel® Turbo Boost Technology: Review

Intel® Turbo Boost Technology automatically provides opportunistic performance improvement by allowing individual cores to operate at a higher frequency. The level of turbo frequency achieved depends on several factors, including processor SKU, number of cores in C0 state, type of workload, estimated power and current, and CPU temperature.

This results in higher frequency in both single and multi-threaded applications when headroom is available.

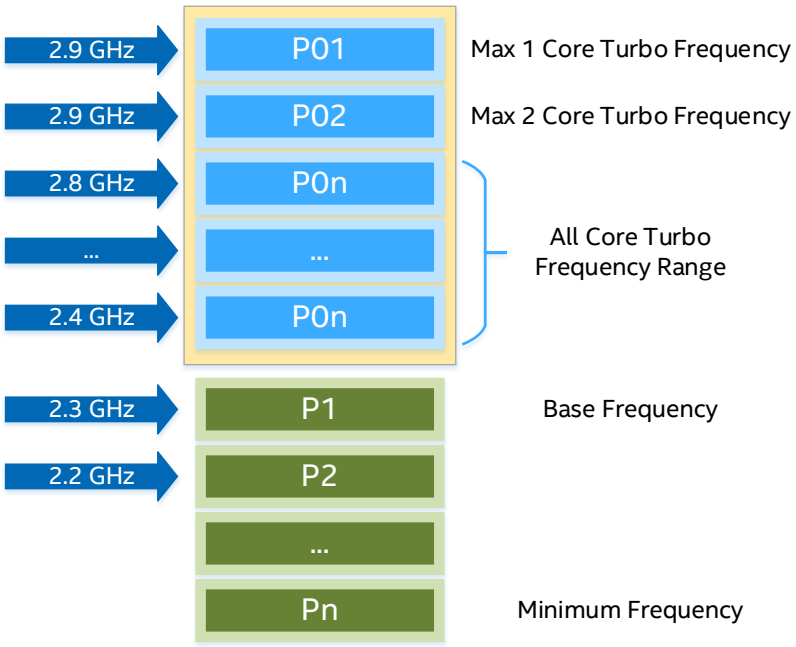

Figure 3. Turbo Boost Frequency Naming

2.3.2 Intel® Turbo Boost Technology: Management

Hardware: The Power Control Unit is responsible for automatically managing the turbo boost algorithm. Intel® Turbo Boost Technology can be enabled/disabled from UEFI.

Operating System: Intel Turbo Boost Technology can be enabled and disabled on a per-core basis via Model Specific Registers (MSRs). MSRs can also restrict the frequency that a core can reach when it enters turbo mode. The OS power management driver can also be used to enable or disable Turbo Boost.

Orchestration: There is no orchestration level management for Intel Turbo Boost Technology.

DPDK Application: The librte_power API allows an application to enable and disable turbo for a given core and allows control of the maximum turbo frequency. The level of control depends on the host OS driver in use. Common OS drivers are intel_pstate and acpi_cpufreq.

2.3.3 Intel® Turbo Boost Technology: Key Takeaway

Intel Turbo Boost Technology provides opportunistic frequency improvements through higher core frequencies of select cores.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for more information.

2.4 Intel® Speed Select Technology – Base Frequency (Intel® SST-BF)

2.4.1 Intel® SST-BF: Review

Intel® Speed Select Technology – Base Frequency (Intel® SST-BF) allows the base frequency of a certain number of cores to be increased (high priority cores) in exchange for lowering the base frequency on the rest of the CPU's cores (low-priority cores), resulting in two frequency tiers across the cores. This can be seen in Figure 4 where, when Intel® SST-BF is active, the first 8 cores run at a higher frequency while the rest of the cores run at a lower frequency. This technology is available in the many 3rd Generation Intel Xeon Scalable processors.

In some workloads, a small number of cores may bottleneck overall system performance. By assigning higher frequencies to these cores with Intel® SST-BF, significant performance improvements can be realized with no change in software.[2]

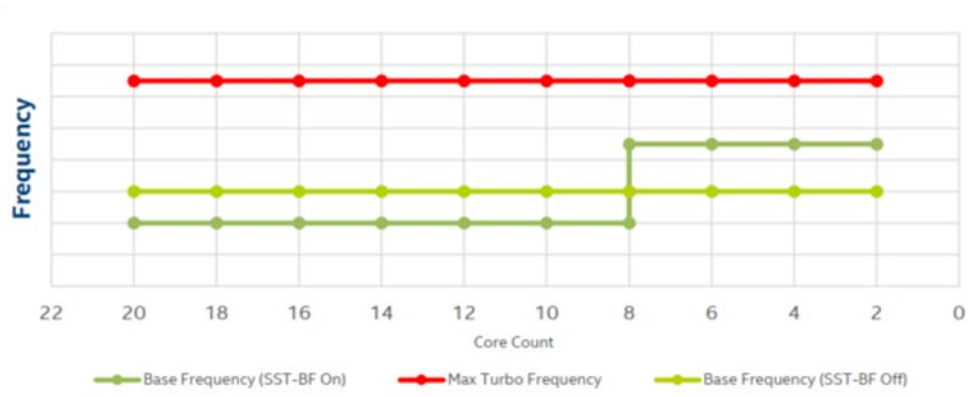

Figure 4. Intel SST-BF Configuration Example

2.4.2 Intel® SST-BF: Management

Hardware: Intel® SST-BF configurations are set by Intel. Intel® SST-BF can be enabled, disabled, and configured via UEFI or via software. This feature requires a CPU that supports Intel SST-BF.

Operating System: The SST Driver is required for Intel® SST-BF to function if not configured by UEFI. An Intel® SST tool is included in the Linux Kernel starting with v5.3. We recommend using Linux Kernel v5.8 or later. The SST Config Tool can be used from the host to configure the platform via a mailbox.

Orchestration: OpenStack Nova (Train release) supports Intel SST-BF when deploying a workload to a single frequency tier. OpenStack Nova (Ussuri release) allows a workload to be deployed across a high and low frequency tier. OpenStack Ironic (Train release) supports IPMI-based configuration, and Redfish-based Ironic support is in progress. Kubernetes CPU Manager (CMK) and Node Feature Discovery (NFD) support Intel SST-BF.

DPDK Application: Librte_power can be used by DPDK applications to leverage Intel SST-BF by identifying higher and lower frequency cores.

2.4.3 Intel® SST-BF: Key Takeaway

Intel Speed Select Technology – Base Frequency allows the user to control and direct base frequency, essentially allowing platforms to power up and prioritize frequency for the most critical workloads at the most critical times.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for information.

See Section 1.2 for reference materials.

[2] For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary.

2.5 Intel® Speed Select Technology – Core Power (Intel® SST-CP)

2.5.1 Intel® SST-CP: Review

Intel® Speed Select Technology – Core Power (Intel® SST-CP) allows the user to assign a priority to each core, which is then used by the Power Control Unit to direct surplus power to the highest priority workloads.

Traditionally, all CPU cores have equal priority when power headroom is available, meaning that the CPU distributes available power to all cores resulting in a uniform increase in frequency regardless of the tasks running on each core. SST-CP offers dynamic prioritization of CPU core power/performance by assigning priority to each CPU core, thereby satisfying the power requirements of each core in a priority order. This allows the user to specify frequency tiers to meet the needs of their workloads. SST-CP can be combined with features such as P-states, Turbo Boost, and other Speed Select Technologies to allow platforms to power up and prioritize frequency for the most critical workloads at the most critical times.

Figure 5 shows this scenario where the Power Control Unit (PCU) distributes surplus power equally to all cores when SST-CP is not being used, compared to distributing surplus power to the high priority core when SST-CP is active. In essence, Intel SST-CP allows the user to set a "Power QoS" for cores in a CPU, whereby CPU cores with highest priority get prioritized access to available power headroom.

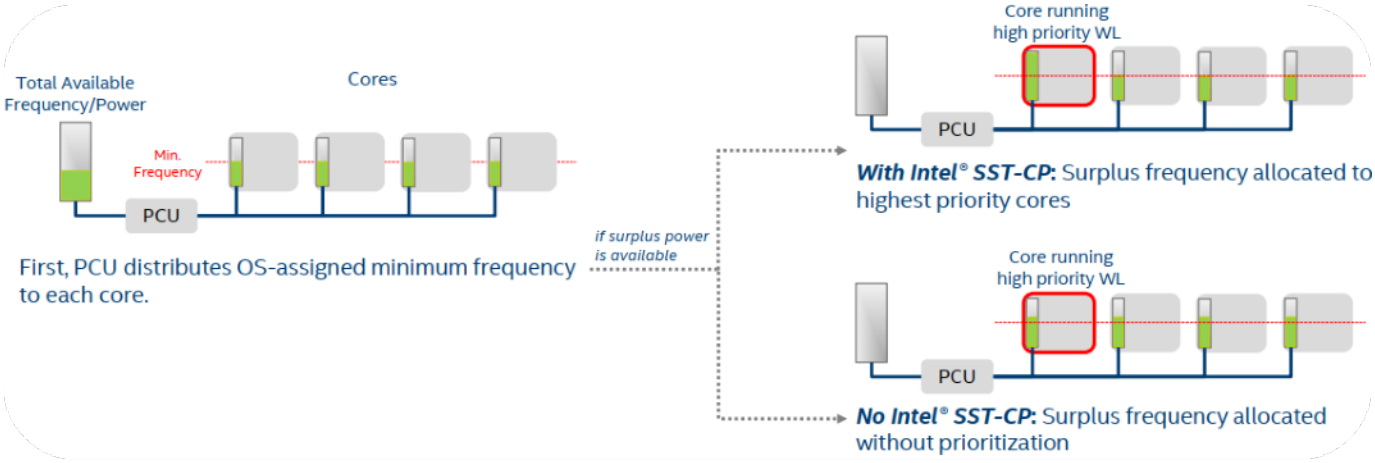

Figure 5. Intel SST-CP Power Allocation Example

2.5.2 Intel® SST-CP: Management

Hardware/Firmware: This feature requires a CPU enabled with Intel SST-CP.

Operating System: The SST Config Tool allows for configuration of Intel® SST-CP via Linux. Intel® SST-CP can be configured for Ordered Throttling where specific cores are prioritized when distributing any remaining power headroom, or Proportional Throttling where power headroom is distributed proportionally across all cores based on user-input weights. The ordered priority type is the recommended way to set up Intel® SST-CP. An Intel® SST tool is included in the Linux Kernel starting with v5.3. We recommend using Linux Kernel v5.8 or later.

Orchestration: OpenStack Nova (Train release) supports Intel® SST-CP. OpenStack Ironic (Train release) supports IPMI-based configuration and Redfish-based Ironic support is in progress. Kubernetes CPU Manager (CMK) and Node Feature Discovery (NFD) support Intel® SST-CP.

DPDK Application: There is currently no librte_power enabling.

2.5.3 Intel® SST-CP: Key Takeaway

Intel® Speed Select Technology – Core Power allows the OS/VMM to direct stranded power to specific cores, allowing prolonged high frequency on certain cores and/or other Speed Select Technologies to enable greater system performance.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for information.

See Section 1.2 for reference materials.

2.6 Intel® Speed Select Technology – Turbo Frequency (Intel® SST-TF)

2.6.1 Intel® SST-TF: Review

Intel® Speed Select Technology – Turbo Frequency (Intel® SST-TF) allows the maximum turbo frequency of a certain number of cores to be increased (high-priority cores) in exchange for lowering the maximum turbo frequency on the rest of the CPU's cores (low-priority cores).

In certain workloads, there are a small number of cores that bottleneck overall system performance. By assigning higher frequencies to these cores with Intel® SST-TF, significant opportunistic frequency improvements can be realized with no change in software.[3] Figure 6 gives an example of the differences between Legacy All-Core Turbo mode and Intel SST-TF mode. With a low number of high-priority turbo cores, a higher turbo frequency can be achieved by these cores. As the number of high-priority cores increases, the turbo frequency reduces until it is at the normal Legacy All-Core Turbo Limit.

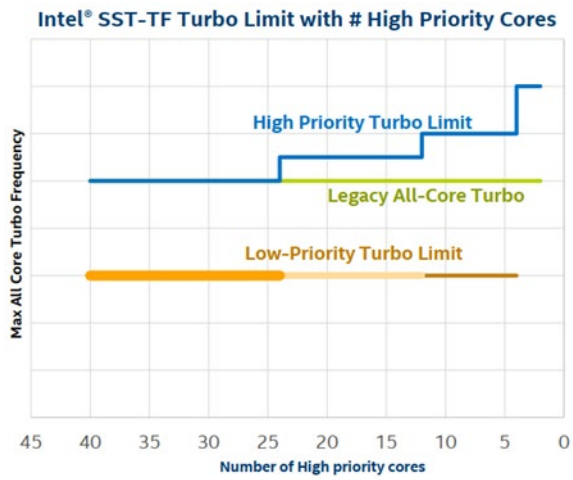

Figure 6. SST-TF Configuration Example

2.6.2 Intel® SST-TF: Management

Hardware/Firmware: This feature requires a CPU enabled with SST-TF.

Operating System: The SST Config Tool allows for configuration of Intel® SST-TF via Linux with the Intel® SST Driver. Linux Kernel v5.5 or later is required for full functionality.

Orchestration: OpenStack Nova (Train) supports Intel® SST-TF when deploying to a single frequency tier. OpenStack Nova (Ussuri release) allows deployment across multiple frequency tiers. OpenStack Ironic is in progress for both IPMI and Redfish.

DPDK Application: No enabling is required.

2.6.3 Intel® SST-TF: Key Takeaway

Intel Speed Select Technology – Turbo Frequency can improve performance of workloads where a small number of cores bottleneck system performance, while staying within the CPU's Thermal Design Power.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for information.

See Section 1.2 for reference materials.

[3] For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary.

2.7 Intel® Speed Select Technology – Performance Profile (Intel® SST-PP)

2.7.1 Intel® SST-PP: Review

Intel® Speed Select Technology – Performance Profile (Intel® SST-PP) allows a single SKU to be used in one of three different configurations of cores, base frequency, and TDP. This allows the platform to be configured based on the workload requirements – a large number of lower frequency cores to handle workloads that can benefit from many threads, a smaller number of higher frequency cores for more frequency sensitive workloads, and a balanced point in between. There is an additional benefit in SKU management – one SKU can be optimized for three different applications with different requirements, resulting in simplified ordering and SKU management.

Figure 7 gives an example of the three configurations enabled by Intel® SST-PP.

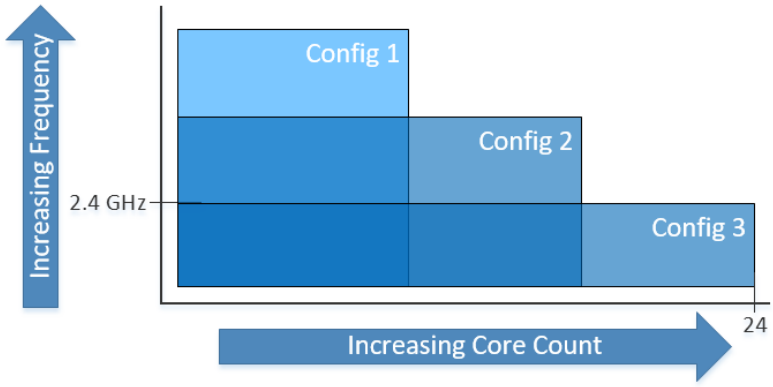

Figure 7. SST-PP Configuration Example

2.7.2 Intel® SST-PP: Management

Hardware/Firmware: Preset configurations are set in hardware by default. UEFI can be used to enable and disable Intel SST-PP, and to allow for static or dynamic runtime configuration. This feature requires a CPU enabled with Intel SST-PP.

Operating System: The SST Config Tool allows for configuration of SST-PP via Linux and requires the Intel® SST Driver. Linux Kernel v5.5 or later is required for full functionality.

Orchestration: Available for IPMI in OpenStack Ironic (Train release), Redfish-based OpenStack Ironic support is in progress. OpenStack Nova-based runtime configuration is also work in progress.

DPDK Application: No DPDK enabling is required.

2.7.3 Intel® SST-PP: Key Takeaway

Intel Speed Select Technology – Performance Profile increases platform flexibility and simplifies SKU management by allowing dynamic changes between preset power, core, and frequency configurations.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for information.

See Section 1.2 for reference materials.

2.8 P-state Transition Latency Improvement

2.8.1 P-state Transition Latency Improvement: Review

P-state transition latency, the time it takes to change the frequency of a core, has been significantly reduced in the 3rd Generation Intel Xeon Scalable processor. Figure 8 provides a comparison between 3rd Generation Intel® Xeon® Scalable processors and the previous generation 2nd Generation Intel® Xeon® Scalable processors while executing a generic workload. When the core frequency is changed on a 2nd Generation Intel Xeon Scalable processor, the core is not processing instructions for ~10-20 microseconds. This block time is eliminated in 3rd Generation Intel Xeon Scalable processors, resulting in seamless core frequency transitions.

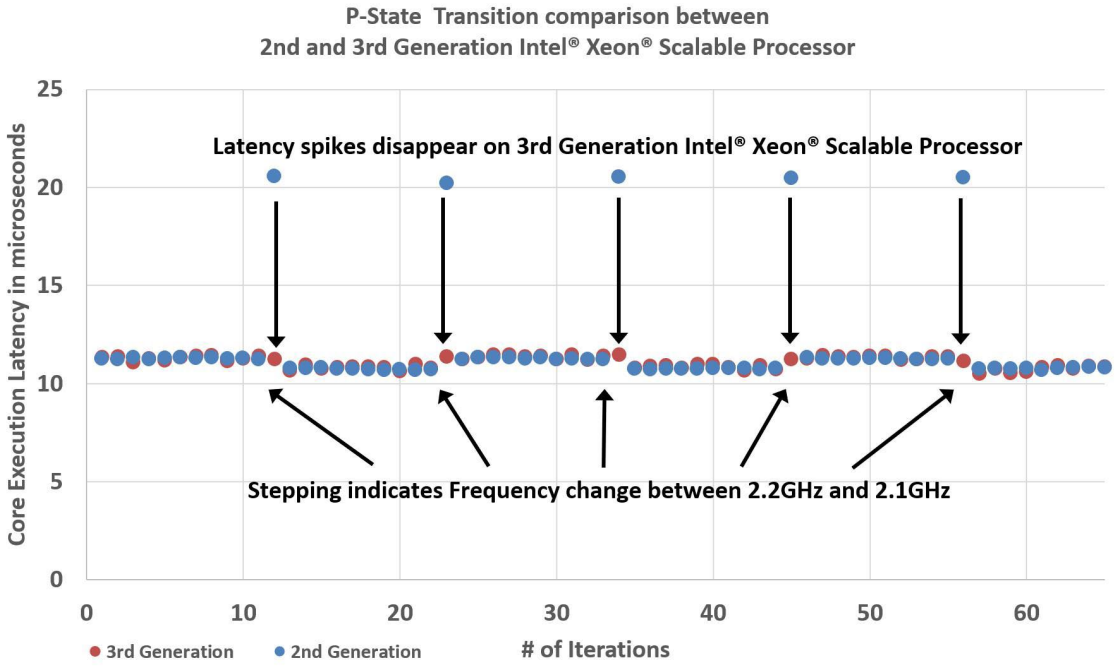

Figure 8. 3rd Generation Intel Xeon Scalable Processor vs. 2nd Generation Intel Xeon Scalable Processor P-state Transition Latency

2.8.2 P-state Transition Latency Improvement: Management

Hardware/Firmware: P-state transition latency improvements do not require any UEFI/hardware enabling. An additional option – GPSS Timer – has been added in UEFI that defines the reaction time to a P-state change request.

The reaction time is an interval timer – essentially a low-pass filter known as the CPU's Global P-state Sequencing (GPSS) timer – within which requests are gathered and queued. After the timer expires (every 500 µs approximately), the last request for a given core is actioned. On the 3rd Generation Intel Xeon Scalable processor, this interval timer is configurable in the BIOS to lower values of 50 µs and 100 µs compared to legacy processors.

Operating System: No changes needed for P-state enabled DPDK Applications.

Orchestration: No Orchestration needed.

DPDK Application: No changes needed for P-state enabled DPDK Applications.

2.8.3 P-state Transition Latency Improvement: Key Takeaway

P-state Transition Latency improvements reduce the potential latency increase and jitter caused by cores switching between different frequencies, allowing P-state control to be used more actively to reduce the power consumption of a server while maintaining deterministic behavior.

You can find further information in the *External Design Specification Volume 1*. Contact your Intel FAE for information.

3 Summary

With each generation of Intel® Xeon® Scalable processors, new features are added to provide greater control over the power and performance balance. This control allows platforms to be optimized for a deployment's key parameters and to react to load conditions to reduce overall power consumption when possible. With the latest set of power management innovations designed into the 3rd Generation Intel® Xeon® Scalable processors, greater performance per watt can be delivered for many applications when compared to previous generation Intel® Xeon® Scalable processors.[4]

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

0222/DN/WIPRO/PDF 637748-002US

[4] For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary.