Lenovo

Using Intel Speed Select on ThinkSystem Servers Running Linux

Introduces the performance and power management technology from Intel

Compares Intel SST with other Intel power management tools

Describes the four major features of Intel SST

Lists key commands available using the Intel SST Tool

Peng Liu

Abstract

Intel Speed Select Technology is a processor-based power management technology that provides multiple CPU performance configurations for users to choose and set different frequency ranges to different cores according to their computing needs.

This paper introduces this new power management technology. The paper first reviews the legacy technologies and the various frequency range each works on, and points out two of the shortcomings of the legacy technologies. Then the paper introduces Intel Speed Select Technology and its main features features. Finally, the support status of the new technology and its usage on Linux is described.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction

To improve CPU performance as much as possible, many power management technologies has been developed over the years. These technologies work by dynamically adjusting CPU frequency within fixed frequency ranges uniformly on all CPU cores. These fixed and uniform frequency ranges make the power management simple to implement and adequate to systems with single or several cores.

However, with increasing CPU cores being applied to enhance computer performance, legacy power management technologies face the challenges from systems with multiple cores and a variety of current computation scenarios. More flexibility is required to manage power distribution among CPU cores.

Intel® Speed Select Technology is such a power management technology that provides multiple CPU performance configurations for users to choose and set different frequency ranges to different cores according to their computing needs.

Overview of Legacy Intel Power Management Technologies

With increased transistor density and core numbers, the CPU has become more and more capable. However, the power constraints have restricted this progress, leading to Intel implementing a series of mechanisms in the x86 architectures to extract additional performance within these constraints.

Such mechanisms include:

- Enhanced Intel SpeedStep Technology (EIST)

  Introduced with Pentium M in 2005, EIST enables the management of processor power consumption via performance state (P-state) transitions. The states are defined as discrete operating points associated with different voltages and frequencies.

- Turbo Boost Technology (TBT)

  Introduced with Nehalem in 2008, it uses the same principle of leveraging thermal headroom to dynamically and opportunistically increase processor performance for single-threaded and multi-threaded/multitasking environment.

- Hardware-Controlled Performance States (HWP)

  Introduced with Skylake in 2015, it autonomously selects performance states as deemed appropriate for the applied workload and with consideration of constraining hints that are programmed by the OS. The states are a continuous, abstract unit-less, performance value scale that is not tied to a specific performance state/frequency by definition.

Frequency Ranges

In Figure 1 on page 4, the frequencies that the processor can run at are split into several ranges according to the ways they are managed.

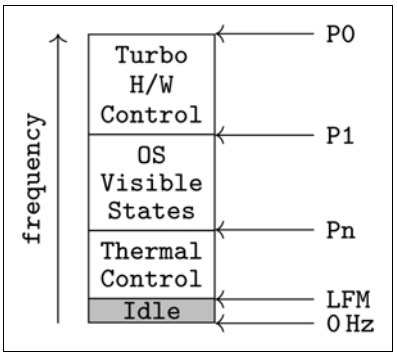

Figure 1 P-states

At the bottom is the idle state which is entered when the processor executes HLT instruction. From Pn to LFM (Low Frequency Mode) is the thermal control range which is entered when the processor is overheated.

P1 to Pn are the guaranteed frequencies at which the processor can chose to run at any time. P1 is the max guaranteed frequency (also Base Frequency) that the processor can chose even with heavy load at worst case conditions, that is, with processors in all cores running heavy operations at P1. Pn is the energy efficient state. EIST and HWP work at this range with the performance change visible to OS. The main difference is that with EIST OS can control the P-states directly, whereas with HWP the processor autonomously selects P-states based on performance guidance hints from OS.

At the top is the turbo frequency range from P0 to P1 which TBT works at. P0 is the max possible frequency (also Turbo Limit). Unlike the guaranteed frequency range, whether the processor can run at a turbo frequency and for how long it can run depend on the power and thermal budgets status.

Figure 2 shows the effect of frequency on the number of active cores. The graph (a) shows that when all four cores of a processor are busy, each of them can maintain the guaranteed frequency P1. But when some of the cores are idle, then the others can boost to a turbo frequency higher than P1 as shown in (b). And this turbo frequency boosts to the highest frequency P0 as all cores but core 0 become idle as shown in (c).

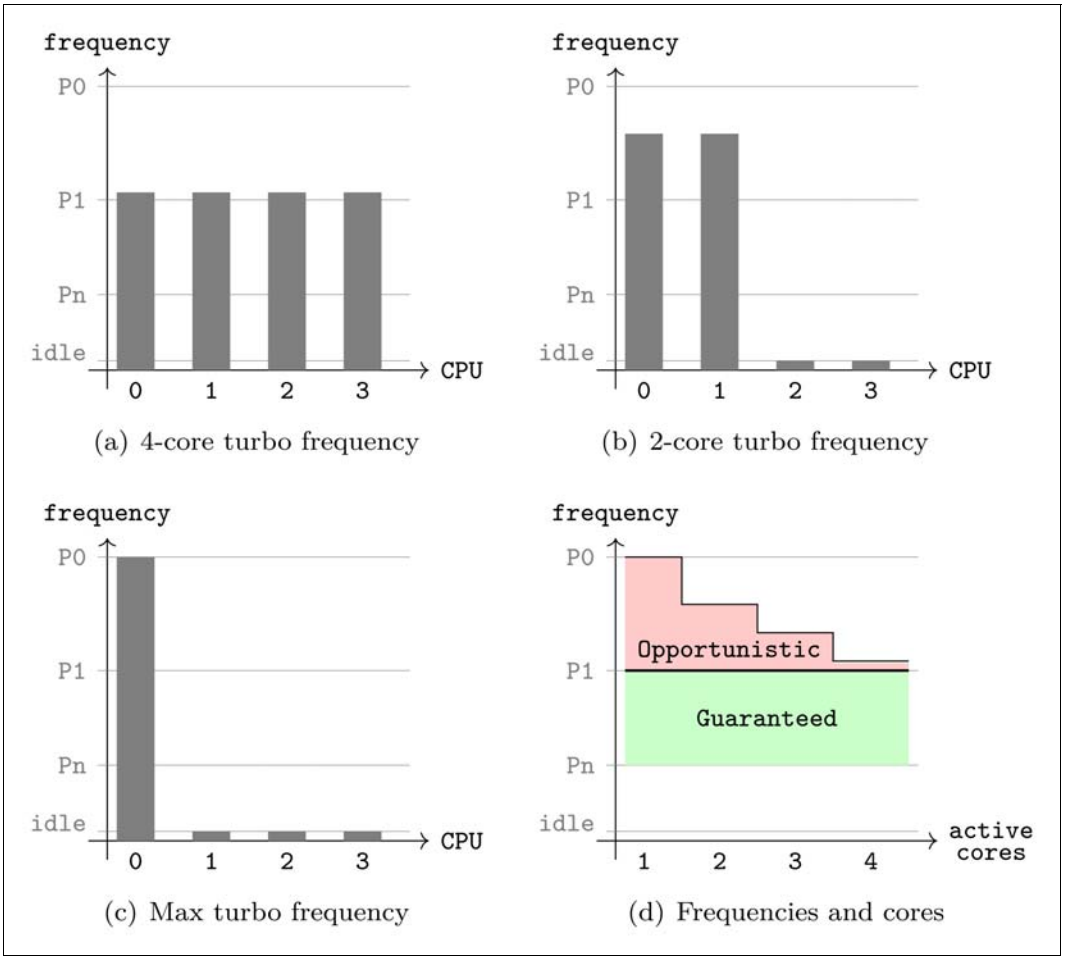

Figure 2 Frequency and cores

When a processor is not busy, it is put to the idle state by OS. The extra power and thermal budgets left thus can be used by busy processors to boost their frequencies. The more inactive processors there are, the higher turbo frequency can be acquired by the active processors as Figure 2 (d) shows. However the opportunistic performance obtained is not always desirable. For example, real-time workloads need sustainable performance.

Shortcomings of Legacy Intel Power Management Technologies

The main shortcoming of above mentioned legacy power management technologies is its lack of configurability on platforms with many cores:

- The base and max turbo frequencies are fixed for a CPU.

  An enterprise may need to accommodate various types of workloads, which include both the thread intensive and non-thread-intensive tasks. To best satisfy the workloads, the enterprise has to acquire servers with different core number and frequency configurations. It may be desirable for the enterprise to want adjust CPU configuration dynamically based on the workloads being run.

- The previous techniques treat all cores equally with the same base and max turbo frequencies.

  With so many cores running various programs we might need to distribute the performance according to program's priorities. For example, with cloud environments it is preferable that some cores running virtual machines of higher privilege have greater base or turbo frequencies than others.

Intel Speed Select Technology (SST) provides a powerful new collection of features that give more granular control over CPU performance. With Intel SST, one server can be configured for power and performance for diverse workload requirements.

Intel SST

Intel SST is a collection of features that improves performance and optimizes the Total Cost of Ownership (TCO) by providing more control over the CPU performance, which handles diverse workloads, varying usages, and unpredictable demands. With Intel SST, one server can do more.

The four features under the Intel SST umbrella are:

- Performance Profile (Intel SST-PP)
- Base Frequency (Intel SST-BF)
- Core Power (Intel SST-CP)
- Turbo Frequency (Intel SST-TF)

These four features are introduced into Intel platforms sequentially as shown in Figure 3.

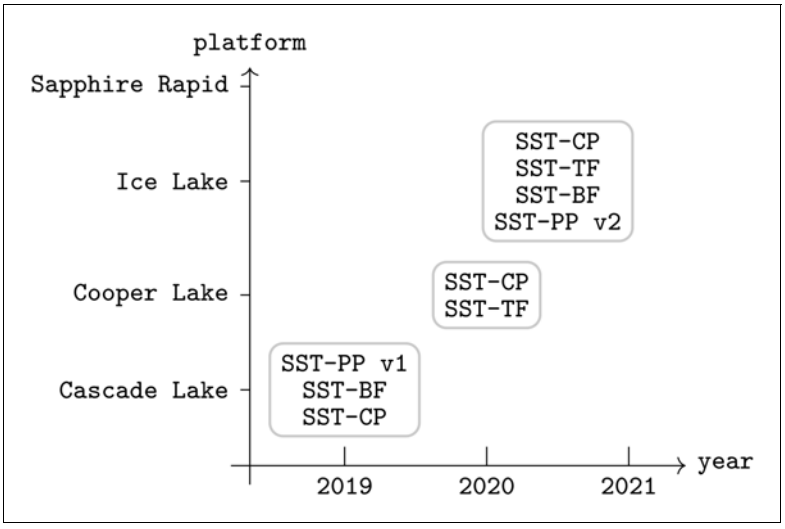

Figure 3 SST feature history

The 2nd Gen Intel Xeon Scalable processor "Cascade Lake" is the first Intel server processor introducing the Intel SST technology. Intel SST-TF is a new addition to the 3rd Gen Intel Xeon Scalable processors "Cooper Lake" and "Ice Lake". Ice Lake server processors also upgraded Intel SST-PP to version 2 that supports both static and dynamic configurations, however Cooper Lake processors do not support Intel SST-PP.

Intel SST features are implemented in the firmware and executed in the Power Control Unit (PCU). The mechanism to control these features are specific to firmware implementation for

Intel Xeon CPUs and are not architectural features. The interface mechanism and semantics of the messages can change in future Xeon CPUs.

Intel SST-PP

Intel SST-PP feature introduces a mechanism that allows the CPU to be configured to run at multiple optimized performance profiles. It is conceived to address requests for flexibility in the base frequency scenario (P1) that the processor figures out and operates at. Intel SST-PP allows higher base frequency at reduced core counts as shown in Figure 4.

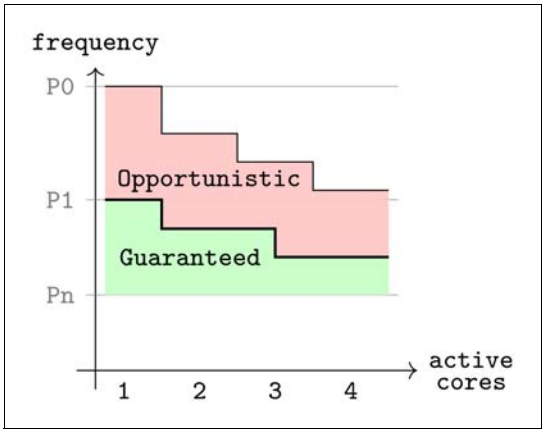

Figure 4 Intel SST-PP frequency and core count relation

Intel SST-PP v1

Each performance profile defines a processor configuration by various performance parameters including active core count, Thermal Design Power (TDP), Streaming SIMD Extensions (SSE) base frequency, and Tjmax. When a profile is chosen, the processor is configured according to the parameter settings in the profile. A configuration is also called a configuration level. The configuration with all core active is called the base configuration or configuration level 0. Intel SST-PP introduces two more configuration levels 3 and 4, which allow the CPU to be configured for lesser cores but higher frequencies.

Configuration level setting can be done statically during boot time by configuring BIOS setup options. For example, most BIOS have Active Core Count setup option that allows users to configure a lesser number of cores to be enabled.

Intel SST-PP v2

Intel SST-PP v2 improves on v1 by allowing run time configuration. Users may need to enable dynamic configuration support by BIOS. When enabled, system boots in the base configuration by default. Configurations can be switched using Intel SST tool, which is introduced in “Intel SST Tool on Linux” on page 11.

During configuration switch, cores may have to be put offline so that the online core number does not exceed the core count allowed in the target configuration. Furthermore, the cores that may be online in a configuration level is not arbitrary. Instead it is predetermined under thermal constrains by BIOS and presented as a configuration parameter active core mask introduced in Intel SST-PP v2.

Besides active core mask, SST-PP v2 also adds new configuration parameters AVX2/AVX3 base frequency and turbo ratio limits for SSE and AVX2/AVX3, CLM P0/P1, and memory frequency.

Usage Scenario

Intel SST-PP enables improved server utilization and reduced qualification costs by allowing one server to be configured for different workload requirements instead of deploying different servers based on the workload requirement. With a single server deployed, the same server can be reconfigured dynamically to one of the supported profiles to suit the specific workload requirements.

Intel SST-CP

Intel SST-CP feature is a mechanism that allows user to define as per core priority and then distributes power among cores according to cores' priorities when surplus frequency is available. Traditionally, the Power Control Unit (PCU) distributes surplus power equitably without priority amongst cores. With Intel SST-CP, we can direct frequency to cores with highest priority or bottlenecks to improve overall performance.

Class of Service

The prioritization between cores is done by a new interface named Class of Service (CLOS). A CLOS is defined by parameters including min and max frequency. Up to 4 CLOS, namely CLOS0, CLOS1, CLOS2, and CLOS3, can be configured on Cooper Lake and Ice Lake.

Each CPU core can be tied to a CLOS and hence an associated priority. CLOS and cores' binding to CLOS can be configured by Intel SST tool at run time.

With Intel SST-CP enabled, PCU first distributes OS-assigned CLOS mini-mum frequency to each core. Then if surplus power is available, PCU distributes it in the way determined by the so-called priority type.

There are two priority types defined.

- Ordered: Priority order is CLOS0 > CLOS1 > CLOS2 > CLOS3, where CLOS0 gets the highest priority. Higher priority cores get the excess budget first.
- Proportional: There is an additional parameter called weight which can be specified for a CLOS. The excess budget is distributed to CLOSes in proportion to their weights.

Intel SST-CP frequency response is not guaranteed because many factors affect the response like workload power consumption, thermal limits, priority group number, core number in each group, etc.

CLOS Configuration Guidance

Unlike Intel SST-PP, the high and low priority cores are not predetermined. With Intel SST-CP, any cores can be configured to any CLOS. But to achieve best performance, the following two principles are to be followed.

- Symmetric high priority cores: Ensure high priority cores are equally split between the sub-NUMA nodes within a CPU.
- Spread out high priority cores: To minimize thermal issues, ensure as much on-die spread as possible for the high priority cores.

Intel SST-BF

The Intel SST-BF feature lets users control and direct base frequency. It is similar to Intel SST-CP in that cores are divided to priority groups among cores whose base frequency is traded off. If some critical workload threads demand constant high guaranteed performance, then this feature can be used to execute the thread at higher base frequency on high priority cores as shown in Figure 5.

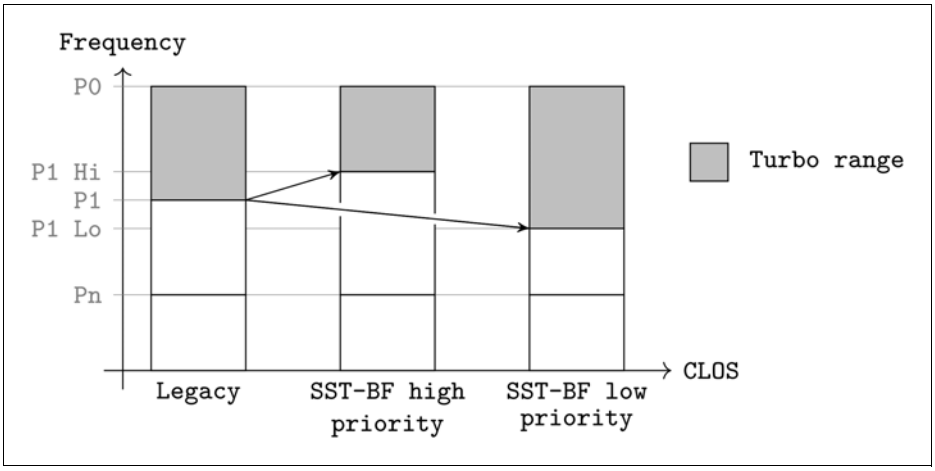

Figure 5 Intel SST-BF base frequency transition

Unlike Intel SST-PP, no low priority cores are required to be offline. Instead SST-BF is only available on Intel SST-PP base configuration. Thus, to enable Intel SST-BF, the base configuration is to be selected first. On Ice Lake, Intel SST-BF implementation requires other power management technologies to be enabled which include EIST, TBT, HWP, and Intel SST-CP in ordered priority type. Intel SST-BF can be enabled and disabled via Intel SST tool.

Intel SST-BF depends on but is distinct from Intel SST-CP. Intel SST-BF supports only two priority levels instead of four. And the performance is guaranteed by Intel SST-BF by adjusting base frequencies, while it is not so with Intel SST-CP. Consequently, the binding of cores to priority levels and other required information of Intel SST-BF is not adjustable but fused into the platform to ensure that the frequency is sustainable. Software is expected to set the min frequency accordingly via CLOS interface.

Intel SST-BF improves system performance for asymmetric workloads, and it does so with lesser reduced performance variability than Intel SST-CP for high priority cores.

Enabling Intel SST-BF

The steps to enable Intel SST-BF are as follows:

1. Enable HWP native mode in BIOS setup menu.
2. Select/enable Intel SST-PP base configuration in BIOS setup menu or at run time if the configuration is not base.
3. Determine if Intel SST-BF is supported in the CPU.
4. Discovers the core mask of high priority capable cores.
5. Enable Intel SST-BF.
6. Configure parameters of each CLOS group.
7. Subscribe or assign cores to CLOS groups based on their intended priority.

Except the first step, the whole Intel SST-BF enabling work can be performed using Intel SST tool.

Intel SST-TF

Intel SST-TF feature is for the scenario that the system is kept busy utilizing all CPUs, but the user wants some configurable option to get high performance on some CPUs. The legacy TBT gives all cores the same turbo frequency. But per-core P-states allow a heterogeneous set of frequencies, and hence power can be distributed differently among core leading to improved performance on cores running high priority threads at the cost of lower or no turbo frequency on other cores.

High and Low Priority Cores

Intel SST-TF feature enables the ability by having software to choose cores, referred to as High Priority (HP) cores, to become capable of running turbo frequencies beyond traditional max frequency limit. Intel SST-TF then clips the max frequency of all other cores, referred to as Low Priority (LP) cores, to the predefined limit. By trading off HP and LP cores' frequency limits, it keeps CPU within defined TDP, reliability, and current budgets.

Since LP cores are uniformly clipped to a fixed frequency no matter how many HP cores are chosen, the headroom achieved by such clipping diminishes when HP core numbers increase, so is the turbo frequency limit of HP cores as shown in Figure 6.

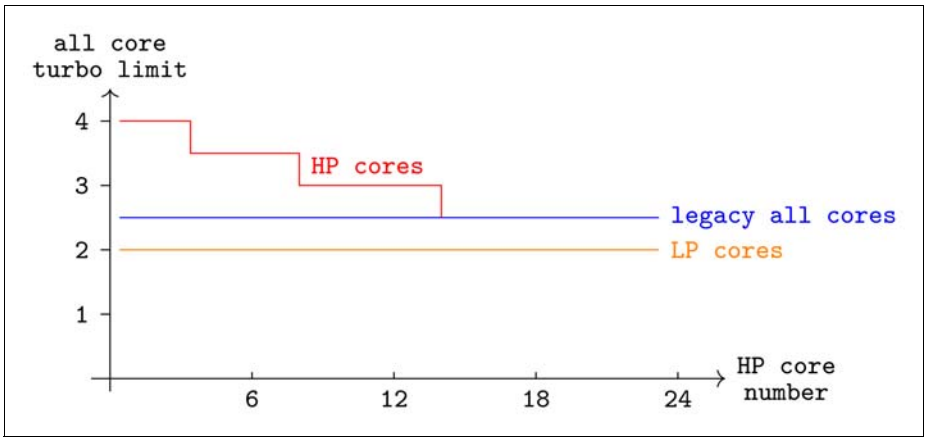

Figure 6 Intel SST-TF turbo frequency prioritization

Like Intel SST-CP, there is no parameter core mask for Intel SST-TF, and any core can be granted high priority at any time using CLOS interfaces. But the LP core turbo limit and HP core turbo limits are characterized in a fused platform-specific Intel SST-TF table that is determined by reliability and power considerations and tied to a specific Intel SST-PP configuration.

Enabling Intel SST-TF

There are four steps to enable Intel SST-TF, and Intel SST tool can be used to perform all of them.

1. Select Intel SST-PP configuration corresponding to Intel SST-TF configuration needed. It is possible that only a certain performance level supports Intel SST-TF.
2. Enable Intel SST-TF if Intel SST-TF fuse is enabled for the SKU.
3. Core prioritization must be configured through CLOS interface.
4. Turbo ratio limits must be configured to realize the Intel SST-TF frequencies for high priority cores.

Usage Scenario

As many cloud, communications, and HPC users perform tasks of varying priorities, the ability to increase performance opportunistically on higher priority cores while reducing performance on lower priority cores can be leveraged to enhance performance on those tasks that matter the most. Increased variability in these cases is offset by the improved performance.

Intel SST Tool on Linux

The support for Intel SST features is introduced to Linux kernel 5.3, and is fully supported in kernel 5.5 and later. The software stack for Intel SST support on Linux is composed of kernel drivers and user space tools as shown in Figure 7.

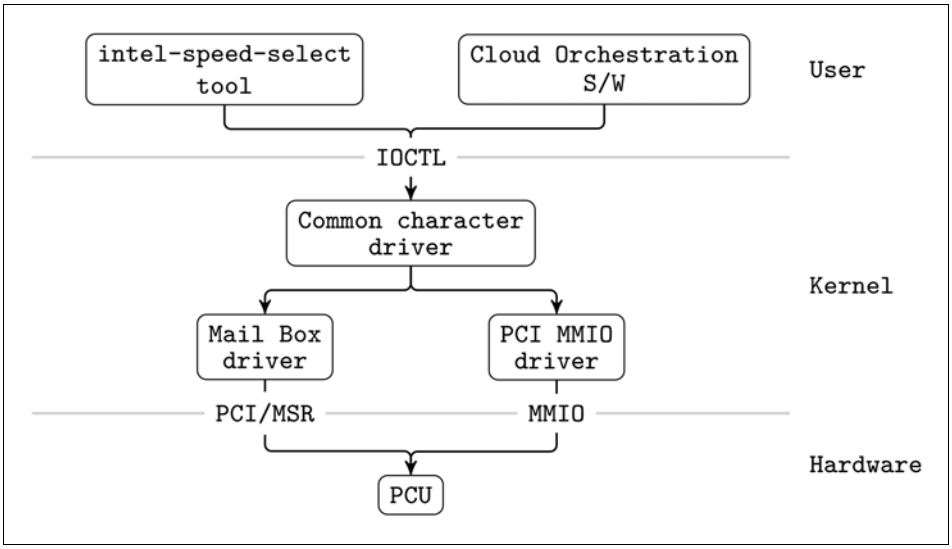

Figure 7 Intel SST software stack on Linux

Intel SST Tool command summary

intel-speed-select is the tool to enumerate and control Intel SST features. The tool is included in the Linux kernel source code. The latest user guide is available from:

https://www.kernel.org/doc/html/latest/admin-guide/pm/intel-speed-select.html

To get help with the tool, execute

```
intel-speed-select [feature [command]] --help
```

There is a multi-level help structure in the tool.

List the top-level help:

```
intel-speed-select --help
```

List a feature help:

```
intel-speed-select feature --help
```

List the help on a feature command:

```
intel-speed-select feature command --help
```

To check the current platform and driver capabilities, execute

```
intel-speed-select --info
```

Intel SST-PP commands

There can be multiple performance profiles on a system.

To get the number of profiles, execute

```
intel-speed-select perf-profile get-config-levels
```

It is possible that they are locked. To check if the system is locked, execute:

```
intel-speed-select perf-profile get-lock-status
```

To get properties of a specific performance level, execute:

```
intel-speed-select perf-profile info -l level
```

Here `-l` option is used to specify a performance level. If the option `-l` is omitted, then this command will print information about all the performance levels.

To get the current performance level, execute:

```
intel-speed-select perf-profile get-config-current-level
```

To the change the performance level to level, execute:

```
intel-speed-select perf-profile set-config-level -l level [-o]
```

If `-o` is specified, CPUs which are not present in the enable_cpu_mask for this performance level will be made offline as well.

Intel SST-CP commands

To enable with the platform default priority type, execute:

```
intel-speed-select core-power enable
```

To disable CLOS based prioritization:

```
intel-speed-select core-power disable
```

To check core-power config options, execute:

```
intel-speed-select core-power config --help
```

To get the current CLOS configuration:

```
intel-speed-select core-power get-config -c CLOS
```

To associate a CPU with a CLOS group:

```
intel-speed-select -c CPU core-power assoc -c CLOS
```

To check the existing association for a CPU:

```
intel-speed-select -c CPU core-power get -assoc
```

Intel SST-BF commands

To get capabilities of Intel SST-BF for the current performance level, execute

```
intel-speed-select base-freq info -l level
```

To enable Intel SST-BF feature

```
intel-speed-select base-freq enable [-a]
```

If `-a` is specified, then it not only enables Intel SST-BF, but also adjusts the priority of cores using Intel SST-CP features.

To disable Intel SST-BF

```
intel-speed-select base-freq disable [-a]
```

Intel SST-TF commands

To get the base turbo capability of performance level, execute

```
intel-speed-select perf-profile info -l level
```

To get the capability

```
intel-speed-select turbo-freq info -l level
```

To enable Intel SST-TF

```
intel-speed-select -c CPU-list turbo-freq enable [-a]
```

The CPU numbers passed with `-c` arguments are marked as high priority, including its siblings. If `-a` is specified, then it enables Intel SST-TF feature and also sets the CPUs to high and low priority using Intel SST-CP.

References

For more information, see these web references:

- Intel speed select technology

  https://www.intel.com/content/www/us/en/architecture-and-technology/speed-select-technology-article.html
- Intel Speed Select Technology in the Linux kernel:

  https://lkml.org/lkml/2019/6/26/1037
- Intel Speed Select Technology User Guide:

  https://www.kernel.org/doc/html/latest/admin-guide/pm/intel-speed-select.html

Author

Peng Liu is an experienced Linux Engineer at the Lenovo Infrastructure Solutions Group, based in Beijing, China. He focuses on storage device drivers and but his interest areas include topics such as I/O frameworks and memory management.

Special thanks to the following people for reviewing the paper during development and providing many valuable suggestions.

- Adrian Huang, Lenovo Taipei
- Rick Hsu, Lenovo Taipei
- Huaisheng Ye, Lenovo Beijing

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on April 14, 2021.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1465

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available from https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®	Lenovo(logo)®	ThinkSystem™
Lenovo XClarity™	ThinkAgile™	

The following terms are trademarks of other companies:

Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.