Power Management in Intel® Architecture Servers

White Paper

Intel® Architecture Servers

During the last decade, Intel has added several new technologies that enable users to improve the power efficiency of Intel® architecture based platforms. Because these technologies have been introduced in different generations of microprocessors and chipsets over several years, it may not be apparent how they work together and impact the performance of higher-level applications. This paper explains each power-related technology and how they interplay in an Intel architecture-based platform. We present data showing how users can reduce the power consumed by a system when applications don't need full processing power and also how the system can deliver higher performance with a power boost during periods of high loads. Understanding these silicon-level technologies will enable application developers, IT managers, and system users to harness Intel architecture's full potential to deliver energy-efficient performance.

April 2009

Table of Contents

Background . . . 3

Demand Based Switching (DBS) . . . 4

Introduction to DBS . . . 4

Internal Operation of DBS . . . 4

Turbo Boost . . . 5

Introduction to Intel® Turbo Boost Technology . . . 5

How Turbo Mode Works . . . 5

A Test Case for Turbo Mode . . . 6

Intel® Intelligent Power Node Manager . . . 7

Introduction to Node Manager . . . 7

How Node Manager Works . . . 7

Working Together . . . 9

Interactions and Usage Models . . . 9

Experimental Data and Analysis . . . 11

Glossary . . . 12

References . . . 12

Background

A computer system requires electrical power to perform various activities, such as fetching data and application programs, executing instructions delivered by software, displaying output results, communicating with users through various interfaces, and interacting with other devices on the network. A study of power consumption in a data center shows that almost 50 percent of incoming power is consumed by air-conditioning and power-delivery subsystems, even before reaching the servers in a rack. Servers consume the remaining 50 percent, which can be further broken down into the various elements as shown in Figure 1.

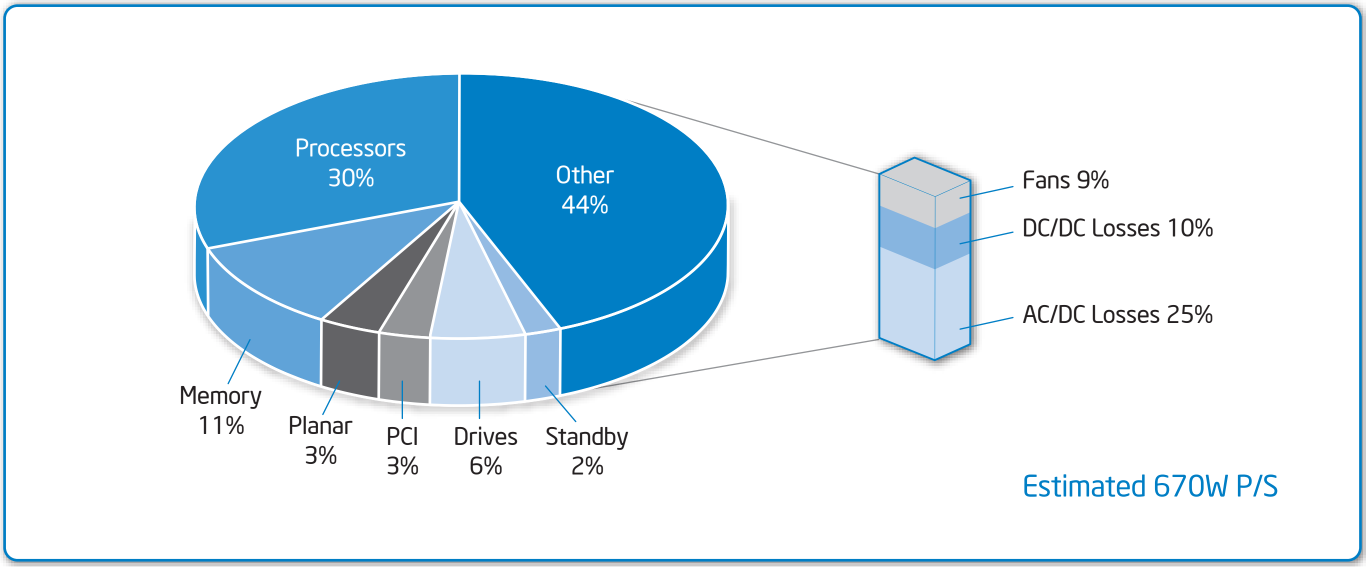


Figure 1. System averages are based on a dual-socket quad-core reference board. Actual values may vary depending on the system configuration and manufacturer.

This paper begins by addressing the ~30 percent power consumed by the processors in a server, and then introduces the concept of managing a system's power with intelligent time-based policies. We will look at three key technologies: Demand Based Switching (DBS), Intel® Turbo Boost Technology and Intel® Intelligent Power Node Manager (NM). The remainder of the paper is organized in different sections introducing each concept independently, their interplay in a platform and managing a rack-level power consumption in the data center.

Demand Based Switching (DBS)

Introduction to DBS

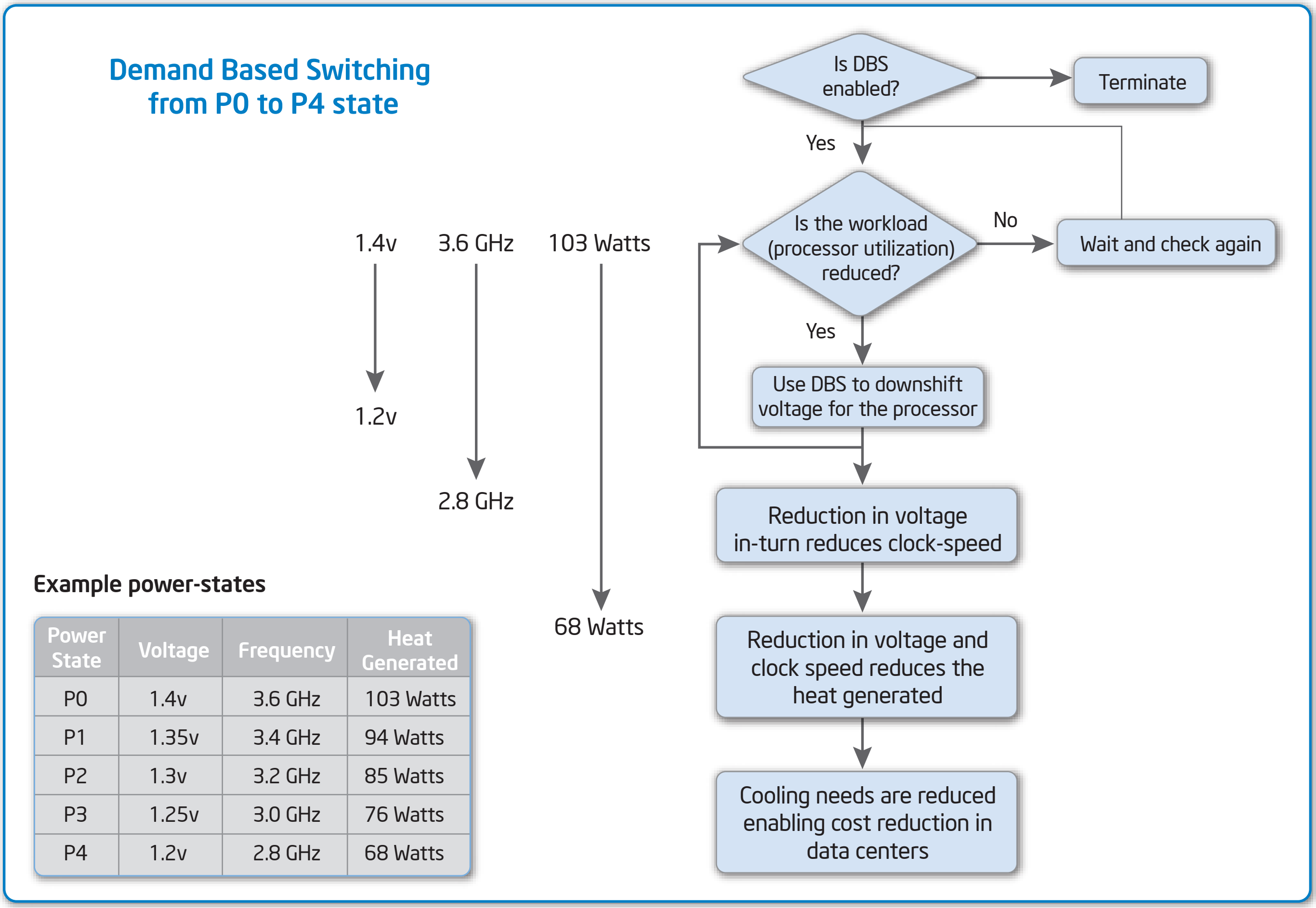
DBS is a power-management technology developed by Intel, in which the applied voltage and clock speed of a microprocessor are kept at the minimum necessary levels for optimal performance of required operations [1]. A microprocessor equipped with DBS operates at a reduced voltage and clock speed until more processing power is required. This is achieved by monitoring the processor's use by application-level workloads, reducing the CPU speed when it is running idle while increasing it as the load increases. This technology was introduced as Intel® SpeedStep® Technology in the server marketplace.

Typically a processor without DBS enabled always runs at the rated speed and consumes corresponding power, independent of the workload, even though the processor is capable of operating at lower operating voltage and frequency combinations. So there is an opportunity to reduce power when the workload levels are lower.

Internal Operation of DBS

Processor performance states (P-states) are a predefined set of frequency and voltage combinations at which a given processor can operate correctly, albeit at different performance levels. With higher frequencies, one will experience faster performance, but to achieve that the voltage also needs to be higher, which makes the processor consume more power (power consumption is proportional to the product of voltage squared and operating frequency).

The table inset in Figure 2 gives examples of a processor's different P-states. In the example, a savings of 35 watts occurs when the processor runs at the lowest performance state. This can be achieved by turning ON the DBS functionality on the BIOS Setup screen. This state should always be selected during lean periods of application usage. After enabling the DBS functionality, one can set policies in the operating system (OS) to become effective at different workloads. When application workloads change, it may change a processor's utilization, initiating a reduction in the processor voltage and clock speed. In turn, the processor's power consumption and corresponding heat generation will drop, leading to cost savings in server power consumption and data center cooling requirements.


Demand Based Switching
from P0 to P4 state
1.4v
3.6 GHz
103 Watts
1.2v
2.8 GHz
68 Watts
Is DBS enabled?
Terminate
Yes
Is the workload (processor utilization) reduced?
No
Wait and check again
Yes
Use DBS to downshift voltage for the processor
Reduction in voltage in-turn reduces clock-speed
Reduction in voltage and clock speed reduces the heat generated
Cooling needs are reduced enabling cost reduction in data centers

Example power-states

Power State	Voltage	Frequency	Heat Generated
P0	1.4v	3.6 GHz	103 Watts
P1	1.35v	3.4 GHz	94 Watts
P2	1.3v	3.2 GHz	85 Watts
P3	1.25v	3.0 GHz	76 Watts
P4	1.2v	2.8 GHz	68 Watts

Figure 2. Demand Based Switching leading to CPU power reduction.

Turbo Boost

Introduction to Intel® Turbo Boost Technology

A processor with turbo mode capabilities can run at frequencies higher than the advertised frequency of the processor if the physical processor package is operating below its rated maximum temperature, current, and power limits.

Turbo mode uses available power headroom to run the active processor cores at higher frequencies. Its availability is independent of the number of cores, but the turbo mode frequency depends on the number of active cores. The amount of time a system spends in turbo mode depends on the system workload and operating environment.

How Turbo Mode Works

CPUs typically operate at a fixed maximum frequency regardless of the workload. However, most applications allow operation below the maximum power rating. Headroom may also be available if some cores are in idle mode as shown in Figure 3, or as long as the package is operating within its thermal limit. Turbo mode speeds up the CPU to utilize available power headroom, as needed, to get an extra performance boost [3].

Turbo mode functionality is enabled and disabled using the BIOS setup, which provides this option only when the processor supports this functionality. When it is enabled, BIOS enables this feature in the processor and publishes a _PST ACPI table with one extra P-state: p0.

Num of P-states in turbo mode = Num of P-states in non-turbo mode + 1

Turbo mode operates under OS control and is engaged only when the OS requests a transition to a P0 state. No OS changes are required to use turbo mode. The Intel® Xeon® processor 5500 series does not support each core running at different ratios when turbo mode engages, thus all active cores will run at the resolved p0 turbo mode frequency. Active cores are in a C0 state (working state, not a sleep state). Figure 4 explains how turbo mode works.

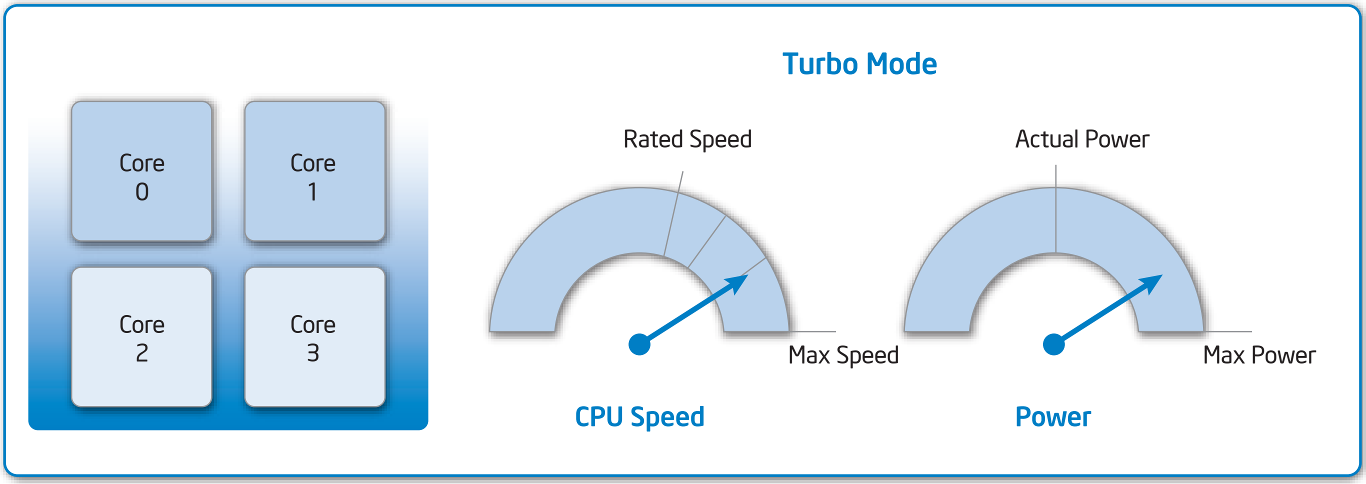


Figure 3. If two cores are off, the remaining active cores can run at a higher frequency while the processor package stays within the overall power and thermal limits.

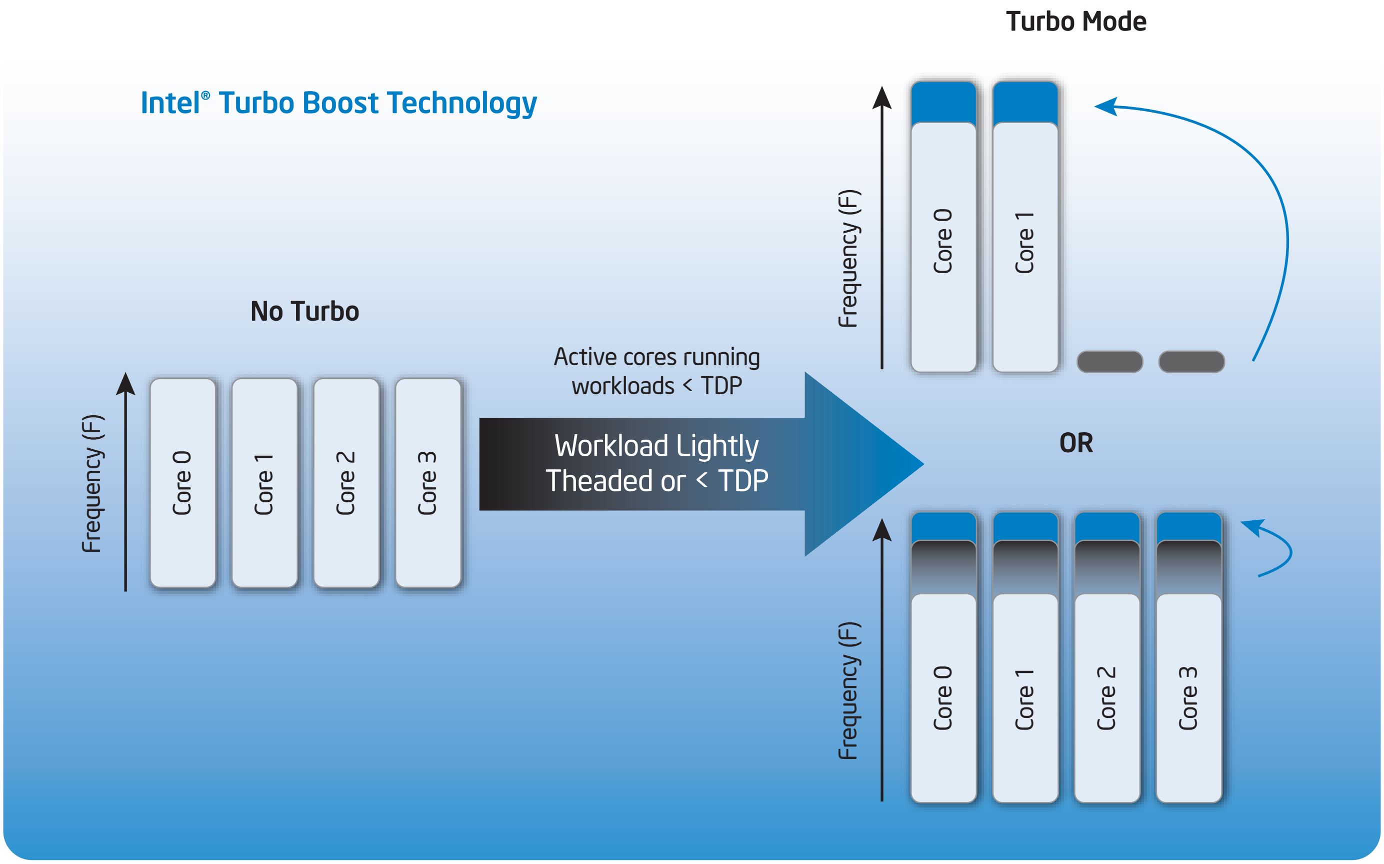


Figure 4. Turbo mode uses the available headroom in processor package power limits.

In the example shown, when all four cores are fully utilized no headroom remains, but if two cores are idle their power consumption will also be lower because the frequency will drop to a lower operating point, assuming DBS was turned ON. The remaining two cores can operate at a higher-than-rated frequency until the overall package-level thermal and power limits are reached. Similarly, if three cores are inactive, the remaining core can run at a still higher frequency.

Note that there is only one turbo state frequency defined, but its value depends on how many cores are active. With more inactive cores, the turbo frequency of the remaining active cores is higher, while maintaining the overall package within its power and thermal limits.

There is also a mode, called the enhanced legacy turbo mode, in which all the cores are working but not at their full capacity, as shown at the bottom right of Figure 4. In this case the core frequency can be increased until the thermal package limit is reached. In all circumstances, the package always stays within its thermal limits.

A Test Case for Turbo Mode

In the first scenario, the user can use the Advanced Configuration and Power Interface (ACPI) dump utility to check the number of P-states published by BIOS when turbo mode is enabled. The system will have one extra P-state when turbo mode is enabled.

For example if a processor is marked with frequency 2 at .6 GHz, the BIOS will publish three P-states (P2–2.4, P1–2.5, P0–2.6) with turbo mode disabled; when turbo mode is enabled BIOS will publish (P3–2.4, P2–2.5, P1–2.6, P0–2.67).

P-states Without Turbo Mode
P0 2.6 GHz
P1 2.5 GHz
P2 2.4 GHz

P-states With Turbo Mode
P0 2.67–2.8 GHz
P1 2.6 GHz
P2 2.5 GHz
P3 2.4 GHz

In the second scenario a user can use a utility which displays the operating frequencies of the processors. The user can see the transition of the CPU to the P0 state, which will be more than the marked frequency of the processor, and can verify that all active cores are in turbo mode. In the above example, the user will see a frequency between 2.67–2.8 GHz when turbo mode is engaged, whereas with turbo mode disabled the maximum operating frequency will be 2.67 GHz.

Thus, turbo mode is not available all the time. The CPU switches to turbo mode based on the temperature, current, and power limits of the processor. Typically this mode benefits scalar applications that are single threaded and thus their performance is directly impacted by a core's operating frequency. Whenever some of the cores will be inactive, the remaining cores that are being used will run at a higher frequency showing higher performance for the scalar applications.

Intel® Intelligent Power Node Manager

Introduction to Node Manager

The Intel®Intelligent Power Node Manager is a system-level technology that reports and manages power consumption of all components in a server. While previously explained technologies optimize power consumption at a component level, NM manages power at the platform level to ensure that system-level power and thermal policies are implemented in a holistic manner.

How Node Manager Works

Figure 5 shows the NM architecture. The NM components can be implemented in many alternative ways. In Intel-developed server boards the components are implemented as part of the Manageability Engine (ME) of the Intel® 5500 series chipset, and communicates with the external management software via the Base-board Management Controller (BMC).

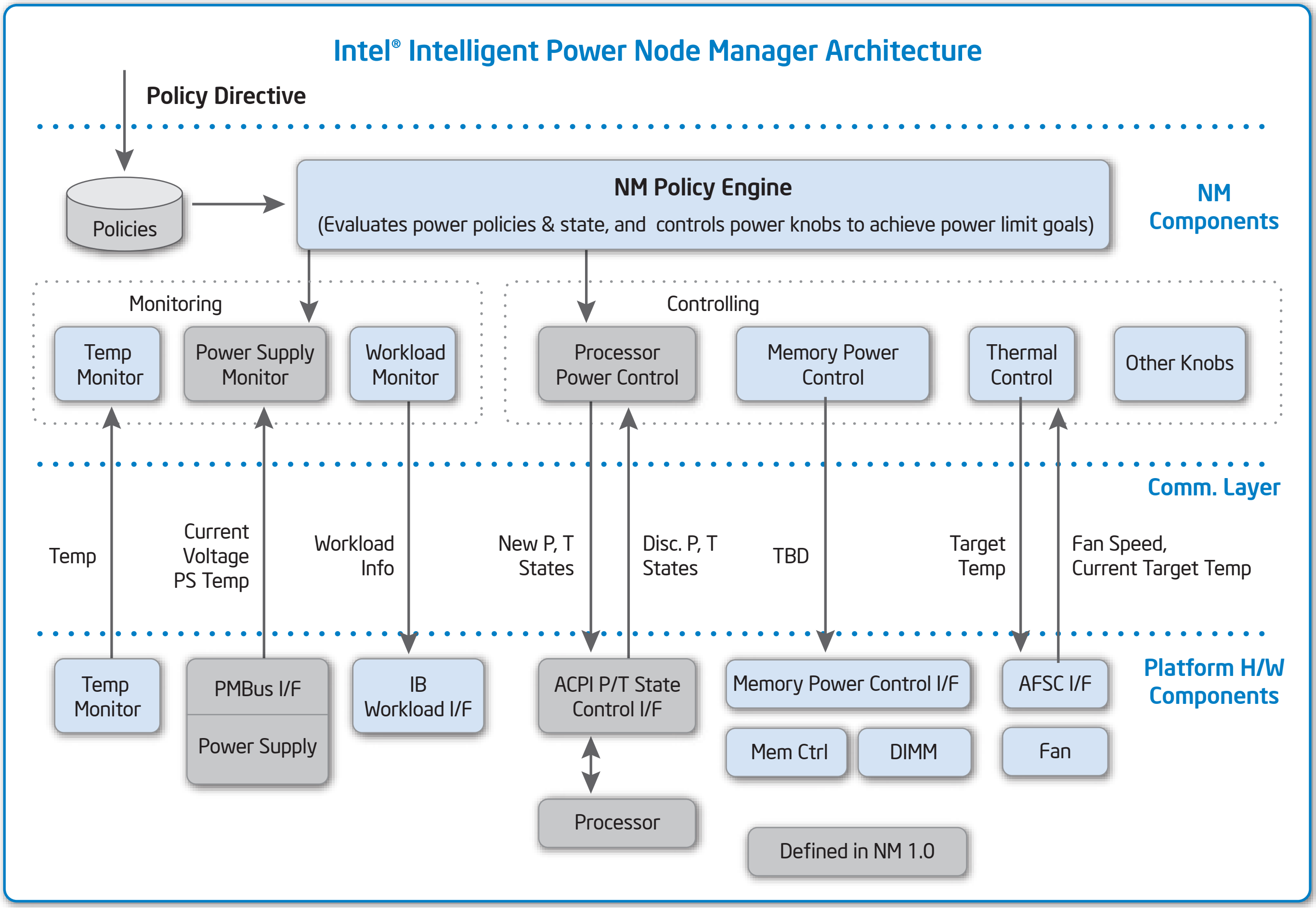


Figure 5. Node Manager architecture and its interactions with system components.

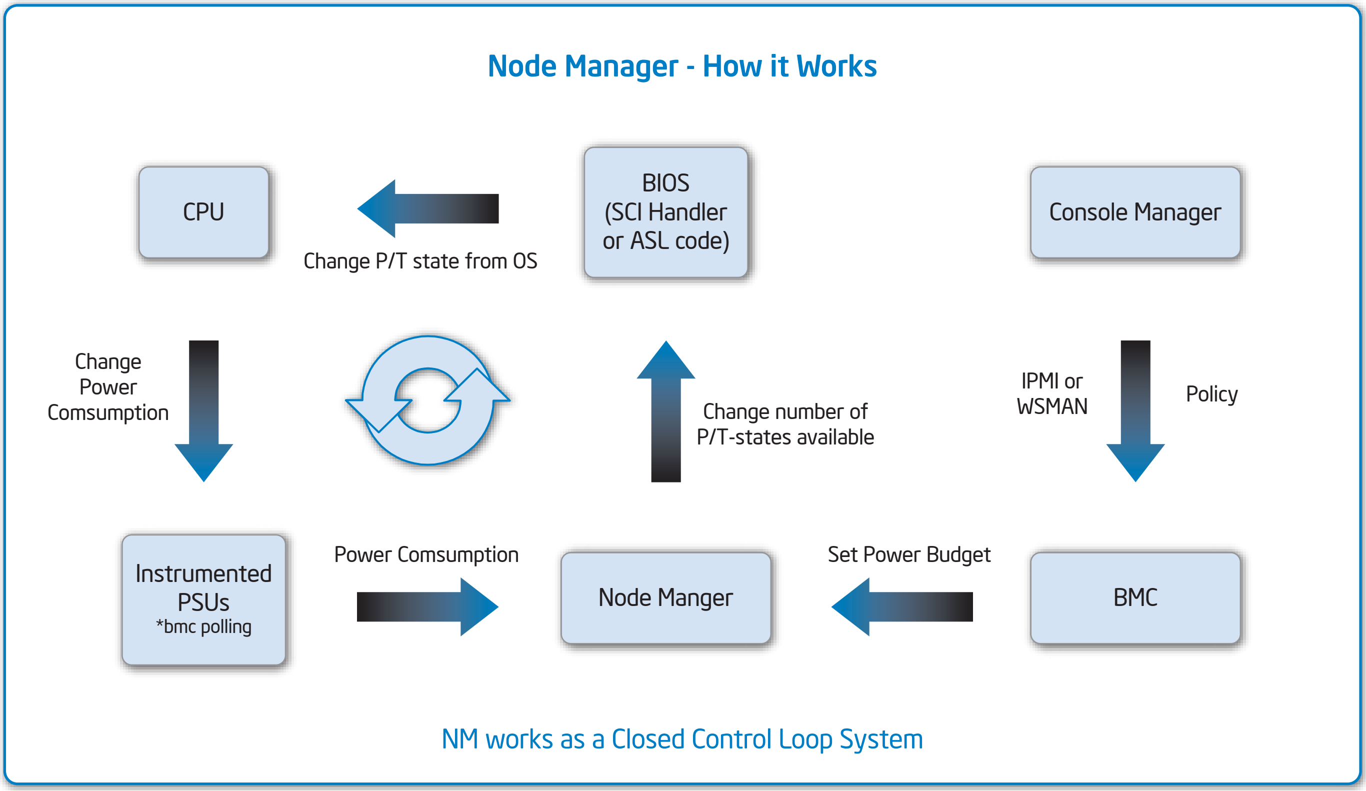

Node Manager - How it Works
CPU
Change P/T state from OS
BIOS (SCI Handler or ASL code)
Console Manager
Change Power Comsumption
Change number of P/T-states available
IPMI or WSMAN
Policy
Instrumented PSUs *bmc polling
Power Comsumption
Node Manger
Set Power Budget
BMC
NM works as a Closed Control Loop System

Figure 6. Node Manager control flow.

NM works as a simple feedback system, as shown in Figure 6, in which the user specifies a maximum power limit during a given time period. The system-level power supply, using the PMBUS standard, measures and reports the power draw into the system. If it exceeds the specified limits during a given operation time, a feedback loop that uses ACPI Source Language code provided by BIOS raises the interrupts to the OS power management, and an ACPI-enabled OS puts the processor in lower P-states, and/or t-states, and impact other components such that the overall system power is reduced.

If system power can't be brought into the specified limits, a user-specified action along with the power policy is implemented. This action may be to raise an alarm (that is, through a SNMP trap to a management agent) or, in an extreme case, the user can even choose to shut down the system. The latter may be needed, for example, to keep a group of systems within the rack-level system limits to avoid a circuit-breaker trip of the entire rack. Another desired usage is to prolong the uptime of a server running on a back-up supply if the main power supply goes down, as often happens in emerging markets.

Working Together

Interactions and Usage Models

Several applications can benefit from the boosted frequency that is higher than the marked frequency of the CPU. Turbo mode increases in performance in multi-threaded and single-threaded workloads by increasing the frequency, while DBS can reduce power when a higher CPU performance is not needed. Node Manager allows a server to stay within pre-determined power limits, similar to a car having an automatic gear-shifting system, as shown in Figure 7.

The combination of all these power management technologies enables a user to get the performance boost when needed, but at other times to save on both the energy needed to run a server and the air conditioning needed to remove the heat generated by the server.

The best way to achieve a server's optimal power performance is to turn ON the DBS and turbo modes in BIOS whenever available, and then use management software to determine the appropriate power policies that can be implemented through NM. One strategy is to run the servers at full power during the day when many server applications are in use, reduce the power during the evening when the workload decrease, and run at the lowest level nights and weekends. Another strategy is to monitor the power of all servers in the rack, and if the overall circuit-breaker power is close to being reached, apply lower power limits for the subset of servers that are not running mission-critical applications at that time. This will enable a higher compute density to be available in the rack above the power limits, allowing different servers to run faster at different times of the day or night.

It has been shown that for a single server, up to a 40W savings can be achieved without a performance impact when an optimal power management policy is applied [2]. To determine the optimum policy, it is advisable to understand the application and workload needs, as well as monitor the server usage patterns, over a period of time before setting different limits in the NM. However, DBS and turbo modes should always be turned ON whenever available.

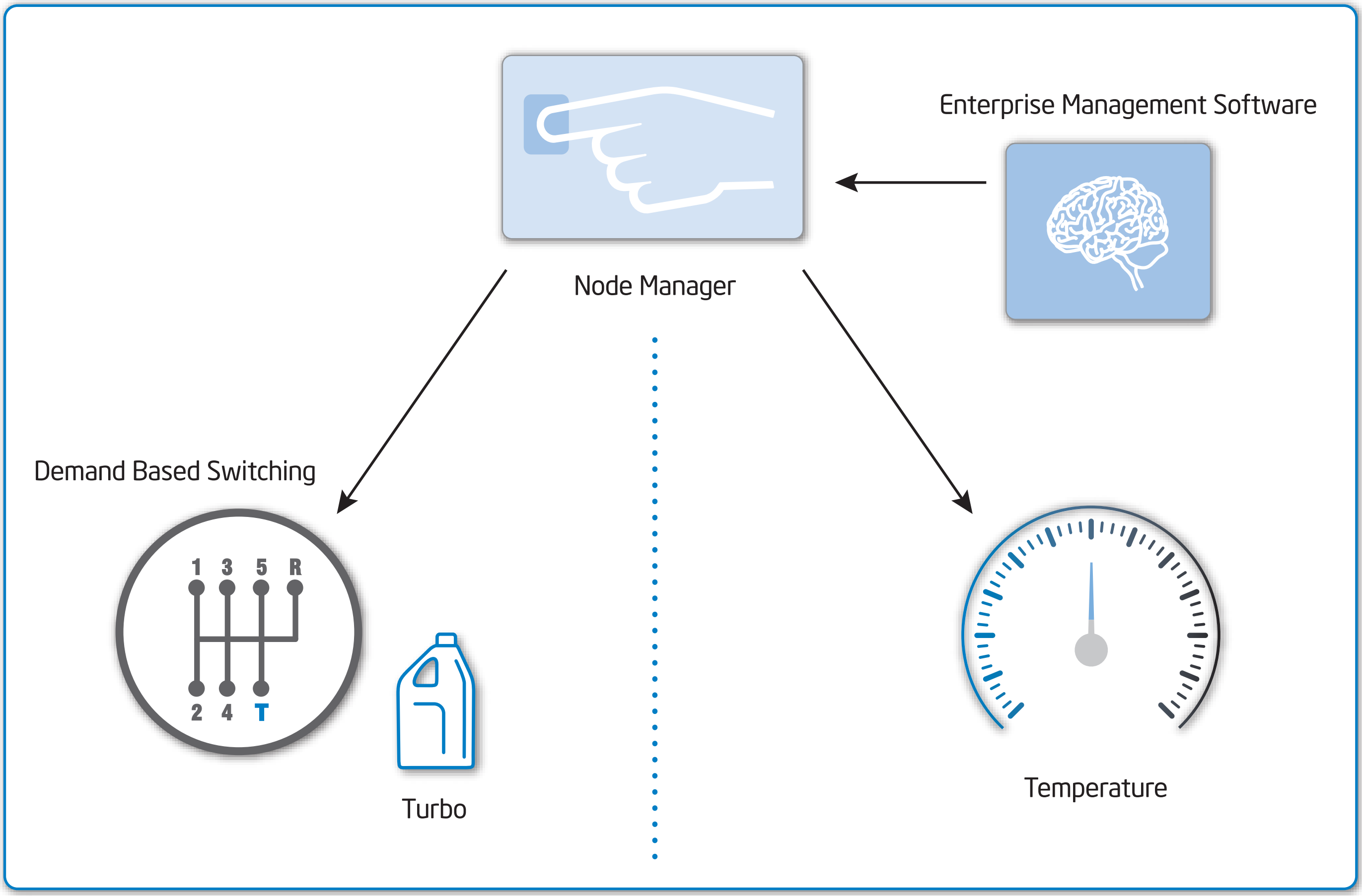


Figure 7. An analogy with a car where extra power may be needed to climb a hill, but reduced power when running idle or low loads.

Experimental Data and Analysis

We put an Intel® Xeon® processor 5500 series platform server to test by stressing the server CPU at different loads while manipulating different NM policies and turning turbo mode on or off.

The data collected are shown in Figures 8 and 9. Server configuration was as follows:

Processor/s	Server Name	Memory	Power Supply	Fans
1 B0 stepping Intel® Xeon® processor 5500 series	SR1625UR	2-512MB DDR3 DIMMs	Redundant power: ▪ Two PMBus enabled power supply modules which are manageable ▪ Power Distribution Board (PDB)	Non-redundant cooling: Ten fixed chassis fans which are not hot swappable

Table 1. Intel® Xeon® processor 5500 series platform server configuration

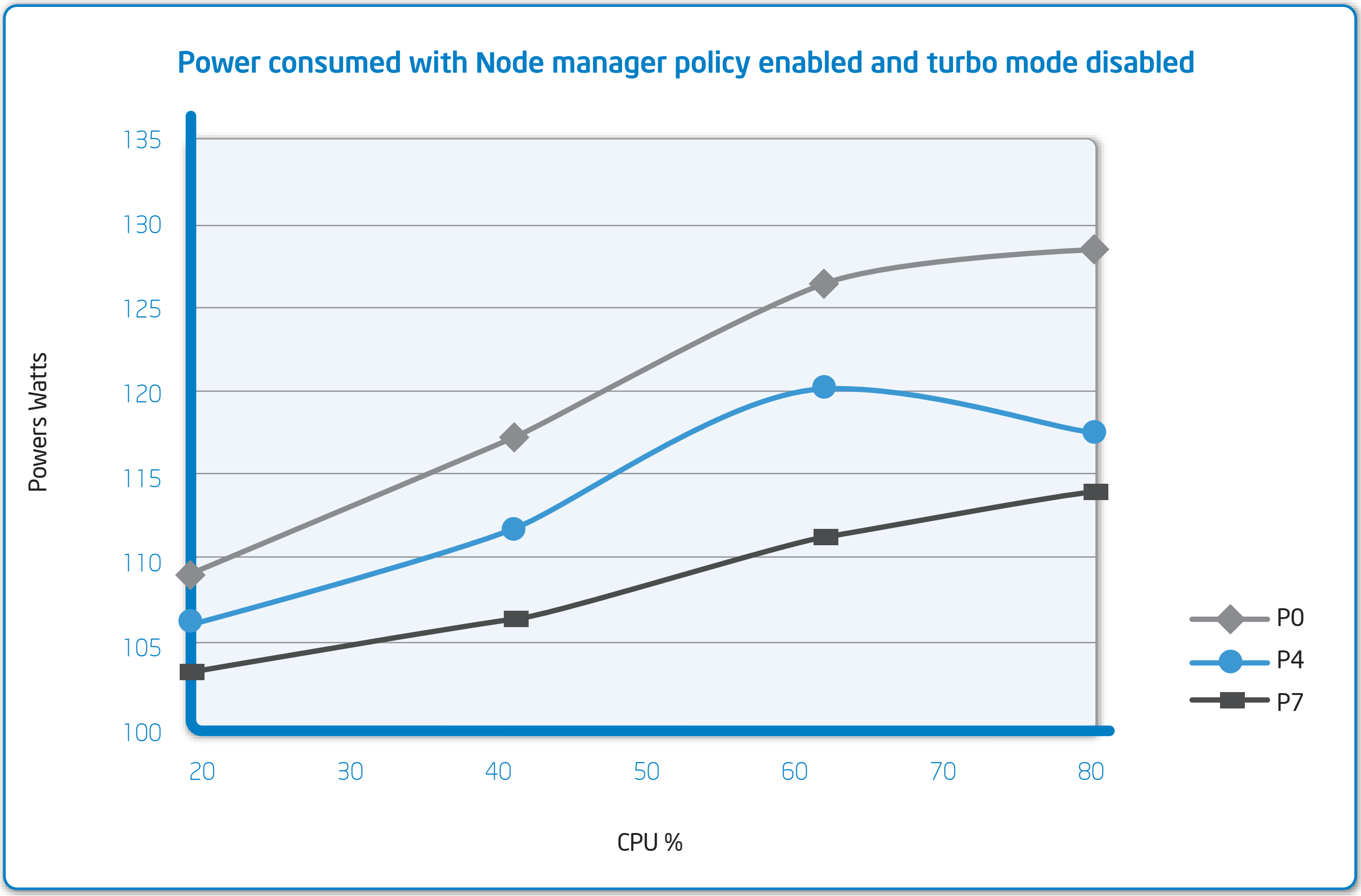


Figure 8. Power variation with P-state change and CPU utilization, without turbo mode.

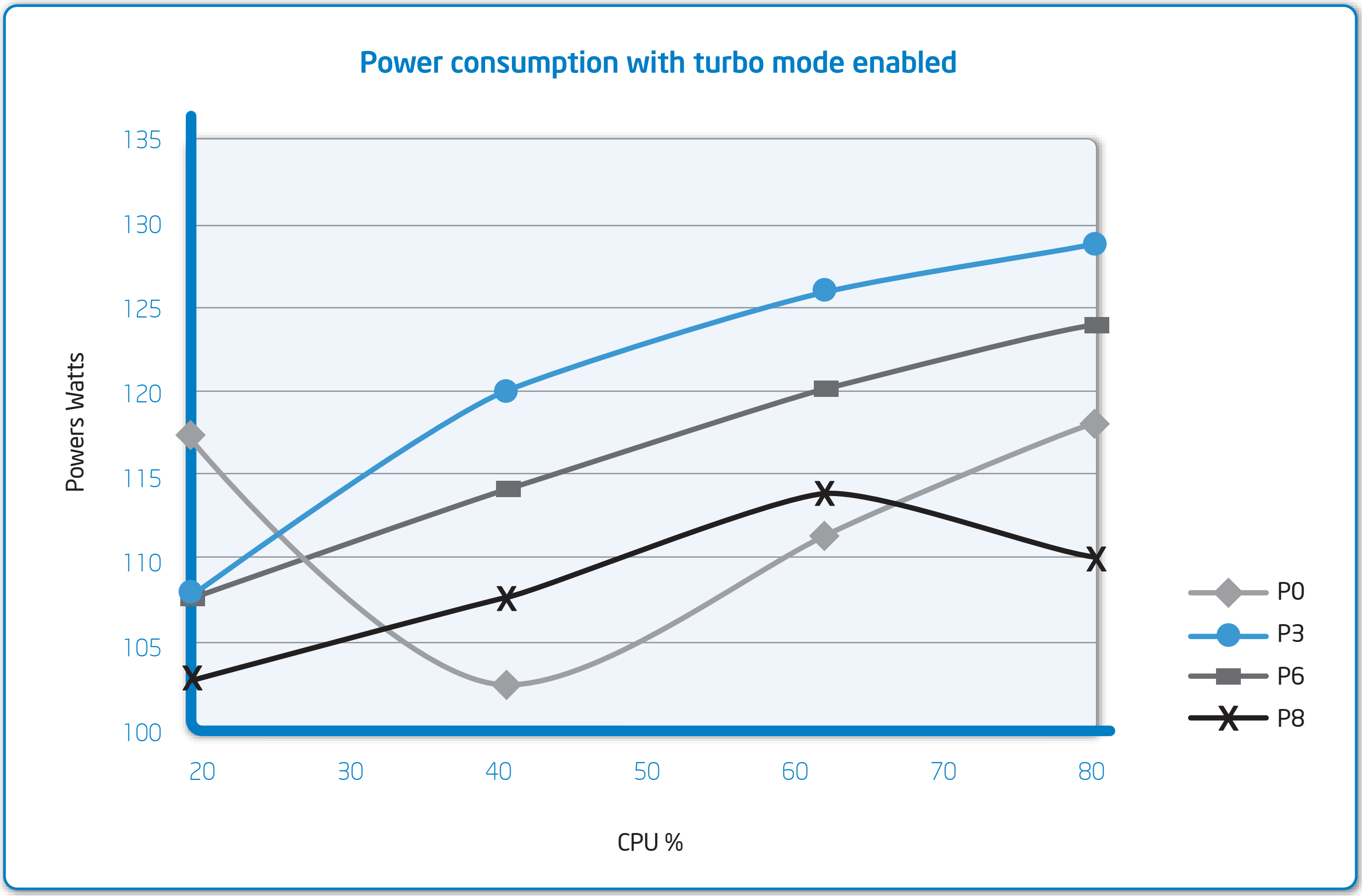


Figure 9. Power variation with P-state change and CPU utilization, with turbo mode.

Several conclusions can be drawn:

- The highest performance state of a turbo mode-enabled system consumes less power overall (an average of 110 watts versus 120 watts without turbo mode enabled).
- To run CPU loads higher than 50 percent, a performance state of P6 can achieve the same power consumption in a turbo-mode enabled server, as a lower performance state of P4 in a server with turbo disabled.
- Turbo mode disabled results in less power consumption with lighter CPU loads—P4 in Figure 8. Also, a server consumes more power with turbo mode enabled—P0 in Figure 9.

To summarize, at high server workloads, the Intel® Xeon® processor 5500 series server should have turbo mode enabled to extract higher performance and lower power consumption from the Intel® Xeon® processor 5500 series. At lower workloads, turbo mode should be disabled to save a server's power consumption.

At the data center level, it has been shown [4] that rack densities can be increased by 20 to 40 percent as long as all the servers are not running at their fully rated power capacity. This allows over-provisioning of compute capacity in a rack, while staying within the rack's power envelope. An obvious benefit is to assign priority to different workloads on different servers at different times of the day. This is akin to a mobile phone operator designing the network capacity to be below the theoretical peak needed because subscribers do not make calls simultaneously. Substantial savings can result for data center operators using Intel's NM technology, while still being able to meet their different customers' peak loads occurring at different times.

Glossary

Advanced Configuration and Power Interface (ACPI): ACPI is an open-standard specification for unified OS-centric device configuration and power management. It brings power management into OS control, as opposed to BIOS central systems. ASL is the ACPI Source Language used for specifying the desired device behavior.

C-state: The processor C-state is the processor's capability to go into various low power idle states (with varying wake-up latencies). Intel architecture-based processors have several C-states representing parts that can be switched off to save power. C0 is the operational state, meaning that the CPU is doing useful work. C1 is the first idle state: The clock running the processor is gated; that is, the clock is prevented from reaching the core, effectively shutting it down in an operational sense. C2 is the second idle state: The external I/O Controller Hub blocks interrupts to the processor. And so on with C3, C4, and others.

P-states: The processor P-state is the capability of running the processor at different voltage and/or frequency levels. Generally, P0 is the highest state resulting in maximum performance, while P1, P2, and so on, will save power but at some penalty to CPU performance.

Server Management Interrupt (SMI): SMI is a special-purpose interrupt that can perform various system management functions including a system's Power controls.

Demand Based Switching (DBS): DBS is a power-management technology in which the applied voltage and clock speed for a processor are kept to the minimum necessary to allow optimum performance of the required operations. A microprocessor equipped with DBS operates at a reduced p-state (voltage and clock speed) until more processing power is actually required. DBS helps to reduce average system power consumption and potentially improves system acoustics.

Intel® Intelligent Power Node Manager (NM): NM is a power-management policy engine that is embedded in Intel® server chipsets. It works with BIOS and OS power management (OSPM) to dynamically adjust platform power to achieve maximum performance and power at a server level, by setting time-based power limit policies, and adjusting P-states. If this power limit can't be reached, an alert or shut-down action can be initiated.

Intel® Turbo Boost Technology: Intel Turbo Boost Technology allows a processor's cores to run faster than the base operating frequency if the package is operating below its power, current, and temperature specification limits. Intel Turbo Boost Technology is activated when the OS requests the highest processor performance state (P0). Maximum frequency depends on the number of active cores. The amount of time the processor spends in the Intel Turbo Boost Technology state depends on the workload and operating environment, providing the extra performance. Intel Turbo Boost Technology increases the performance of both multi-threaded and single-threaded workloads.

References

www.intel.com/support/processors/sb/CS-028855.htm

http://communities.intel.com/servlet/JiveServlet/previewBody/1492-102-1-1723/Node%20Manager%20Baidu%20POC%20WhitePaper%20-%20External.pdf

www.intel.com/pressroom/archive/releases/20080819comp.htm

http://communities.intel.com/openport/blogs/server/2008/04/11/dynamic-power-management-has-significant-values-a-baidu-case-study

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations at www.intel.com/performance/resources/benchmark_limitations.htm.

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION OR SAMPLE. Intel disclaims all liability, including liability for infringement of any proprietary rights, relating to use of information in this specification. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Printed in USA 0409/DJ/MESH/PDF ♻ Please Recycle 321713-001US