A Dynamic Approach to Power Budgeting

Microsoft Corporation
Intel Corporation

July 28, 2008

Abstract

In an effective object lesson that speaks to the future of hardware and software development, Intel and Microsoft found ways to unlock the dynamic power capabilities of next-generation Intel® Core™ microarchitecture through innovative software engineering. The ultimate solution exploits the ability to dynamically select optimal P-states and T-states of the Intel Core microarchitecture processor so that software components—from Intel Dynamic Power Technology Node Manager (Node Manager) to Windows Server® 2008 and Microsoft® Hyper-V™—can use these features and interoperate seamlessly.

This information applies to the Windows Server 2008 operating system.



Author's Disclaimer and Copyright: Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of third-party vendors and their devices. All products, dates, and plans are based on current expectations and subject to change without notice.

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

Copyright © 2008. Intel Corporation. All rights reserved. \*Other names and brands may be claimed as the property of others.

Windows Hardware Engineering Conference - WinHEC Sponsors’ Disclaimer: The contents of this document have not been authored or confirmed by Microsoft or the WinHEC conference co-sponsors (hereinafter “WinHEC Sponsors”). Accordingly, the information contained in this document does not necessarily represent the views of the WinHEC Sponsors and the WinHEC Sponsors cannot make any representation concerning its accuracy. THE WinHEC SPONSORS MAKE NO WARRANTIES, EXPRESS OR IMPLIED, WITH RESPECT TO THIS INFORMATION.

Microsoft, Hyper-V, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

Contents

[Introduction 3](#_Toc204765620)

[Transforming Data Center Power Use 3](#_Toc204765621)

[Worldwide IT Spending on Servers, Power, and Cooling and Management/Administration 4](#_Toc204765622)

[Collaborative Work Yields Impressive Results 4](#_Toc204765623)

[The Art of Virtualization and Power Savings 6](#_Toc204765624)

[Contributors 7](#_Toc204765625)

[**About the writer, Lee Purcell:** 8](#_Toc204765626)

[Related Information 8](#_Toc204765627)

[Future Development Possibilities 8](#_Toc204765628)

[Next-Generation Intel Capabilities and Intel® Dynamic Power Technology Node Manager 8](#_Toc204765629)

# Introduction

In an effective object lesson that speaks to the future of hardware and software development, Intel and Microsoft found ways to unlock the dynamic power capabilities of next-generation Intel® Core™ microarchitecture through innovative software engineering. The ultimate solution exploits the ability to dynamically select optimal P-states and T-states of the Intel Core microarchitecture processor so that software components—from Intel Dynamic Power Technology Node Manager (Node Manager) to Windows Server® 2008 and Microsoft® Hyper-V™—can use these features and interoperate seamlessly. The result is an instructive exercise in the advantages of coordinating design projects between hardware and software teams. Deep engineering of this sort can achieve notable success when communication and collaboration toward a common goal guide the design decisions.

Encouraged by the success of this project, Microsoft and Intel plan to deliver a lecture series in association with the Intel Developer Forum to explain practical power budgeting and balancing techniques on Intel Core microarchitecture-based platforms to IT administrators and personnel. Interested developers should also gain insight into techniques by which the innate capabilities of a processor, such as Intel Core microarchitecture, can be accessed and utilized by operating systems and software applications.

The comprehensive solution created by Intel and Microsoft responds effectively to the challenge of rising data center power costs and the related issue of cooling increasingly dense server equipment configuration. Early proof-of-concept deployments have dramatically showcased the benefits of the technology and benchmark results reported from several test sites have confirmed substantial power savings. It’s unlikely that this kind of success could have been achieved without the hardware and software features working together in unison.

# Transforming Data Center Power Use

Power and cooling costs in data centers have risen sharply over the past several years—as much as 400 percent1 in the span of a decade. As a portion of the total operating budget of the data center, these costs can represent as much as 40 to 50 percent of the total. These cost factors place a substantial burden on IT administrators and directors responsible for server management and deployment, particularly as they come at a time when many data centers are trying to increase the range and scope of their business services.

As discussed in the Intel Technology Journal, *Dynamic Data Center Power Management: Trends, Issues, and Solutions*, faced with these issues, businesses have three fundamental choices to transform data center power use:

* Expand power and cooling capacity.
* Build new data centers.
* Employ a power management solution that maximizes the usage of existing capacity.

## Worldwide IT Spending on Servers, Power, and Cooling and Management/Administration

A power management solution presents a less costly, less disruptive approach to addressing the challenge, and this was the solution that Intel and Microsoft engineering teams decided to adopt. By combining the strengths of a server operating system virtualization solution and dynamic node manager with the capabilities of Intel Core microarchitecture, the server can operate at varying power levels and frequencies, depending on conditions that can be determined through monitoring the data center environment.



The technologies involved dovetail very neatly and complement each other well, characteristics that were enhanced further through extensive 2-1/2 year collaboration between Intel and Microsoft. The resulting solution maintains data center performance levels while achieving signiﬁcant power savings. At a time when energy costs are skyrocketing, data center administrators now have a clear solution that allows them to increase server rack and dynamically load balance power in the data center, deploy additional business services, and apply strict power budgeting controls without adversely affecting the responsiveness of their services or the overall efficiency of the data center in meeting business requirements.

# Collaborative Work Yields Impressive Results

Policy management is the key to power budgeting in the data center. So, the focus for the design team—including engineers from Intel and Microsoft—was on being able to implement policies through the Windows Server 2008 operating system that would cap off thermal levels or cap off the power that a particular system uses. The capabilities of Node Manager were important to this objective, as were the capabilities that the processor provides, which includes being able to vary P-states and T-states. By changing the frequency and the voltage at which the processors are operating, you can essentially control the power usage of the processor. These are capabilities that have been available on earlier Intel processor-based platforms, but the Intel Core microarchitecture platform improves and extends the features available for this purpose.

Integral to the approach is Node Manager, which consists of a combination of a baseboard management controller (BMC) and ﬁrmware, including a manageability engine that is built into the Intel Core microarchitecture platform. Part of the design collaboration involved ensuring that the operating system and Intel Core microarchitecture processor communicate effectively, using the Advanced Conﬁguration and Power Management Interface (ACPI) speciﬁcation (<www.acpi.info>) protocol, through code embedded in the ﬁrmware. The Node Manager uses these capabilities to monitor and enforce the thermal and power management policies. By modulating the P-states and T-states on the processor, the power cap (limit) on the platform can be enforced.

The policy management mechanisms rely on the monitoring of power, temperature, and workloads. The policy engine then makes decisions to determine the appropriate action to minimize power use.



From the Windows Server 2008 perspective, similar power-management capabilities are included. The operating system uses the same P- and T‑state knobs to regulate and control power use when using its out-of-box default balanced power policy. This power policy is in effect during normal operation as well as power-budgeting and capping scenarios. Much of the development effort was on coordinating the operation of the components so that they work together effectively. The policy manager, which has a higher visibility, had to be able to work using ACPI and the BIOS communication with the operating system to function smoothly. The IT manager sets a budget or power policy through a management console, which sends power policy to Node Manager on each server. Node Manager uses ACPI and the BIOS to communicate power policy information to the operating system.

The ultimate solution, from the point of view of the end user, carries out the operating system policy management functions seamlessly. Third-party management consoles can be integrated into the solution to provide necessary data center monitoring and oversight features. Realistic measurement of server power levels under varying conditions was an essential part of the development process. “One of the things we had to validate,” Shiv Kaushik, Intel fellow and director of system software, said, “was what happens when power is being restricted by Node Manager. It is important that Windows® work cooperatively with Node Manager in such a way that we get the best power efﬁciency we can, while still staying within the budget. This is one of our mutual accomplishments that we’re going to talk about and show off at IDF.”

Sean McGrane of Microsoft worked closely with Intel engineers throughout the engagement. “Another task that we wanted to accomplish with Intel,” McGrane continued, “was come up with a metric—a way to measure the power efﬁciency of a particular combination of hardware and software. We both decided the best way to do that was to use industry standard benchmarks, such as TPC or SPECpower. Test or performance runs started at zero-percent utilization for that particular benchmark and ran all the way through to 100-percent utilization.” At each point on the curve, the engineering teams measured the throughput at that level of the benchmark and measured how much power was used. Those values allowed them to determine the throughput per watt or the relative power efficiency for a given combination of software and hardware (for each particular benchmark). This helped established a baseline measurement on existing hardware and with existing software so that whenever changes were made to either the hardware or the software, the team could rerun the exact same set of tests and compare the resulting power efficiency results against the baseline measurement.

This technique proved invaluable to the development. “We can look at what we call the load line,” McGrane said, “which is the power efficiency line for that particular benchmark. And, we can see where the changes have improved power efficiency (or sometimes where it has degraded power efficiency). This gives us a really good way to actually measure the value as we add new features—either into the software or into the hardware. It required a lot of work with our Intel colleagues to come up with the correct set of hardware configurations and the correct set of processes to actually do that. I think that this work has been really valuable to us.”

# The Art of Virtualization and Power Savings

Hyper-V is the virtualization component of the Windows Server 2008 operating system. It offers additional opportunities for data center power savings through virtualization. About the uses of Hyper-V, McGrane said, “We treat Hyper-V in Windows Server 2008 as another operating system role. You boot the standard Windows Server 2008 image and then you can configure it for Hyper-V. When you do that, it enables the hypervisor and the Microsoft virtualization solution that comes with Windows Server 2008. One of the good things is it that Hyper-V offers the same power management benefits that every other Windows Server 2008 role has, so when Hyper-V is operating and the hypervisor is enabled, the performance state management works exactly the same way. Hyper-V will dynamically manage the processor performance state (P-state and T-state) based on the utilization level of the processors.”



Server consolidation is one obvious benefit of deploying Hyper-V. “From a power perspective,” McGrane continued, “Hyper-V allows customers to reduce the number of servers that they actually use in the data centers. It’s part of the power metrics that we’ve been doing to try and figure out the power efficiency tradeoffs for hardware and software. All of the statistics that we have show that most data center’s servers or most servers that are operating in a data center are running at very low utilization, typically somewhere below 20-percent utilization. This is very inefficient from a power perspective. With the Microsoft Hyper-V virtualization solution, you can consolidate many of those workloads to a single server without having to rewrite any of the software running within the virtual machines (VMs).

Microsoft has determined that some fairly high consolidation ratios are possible with Hyper-V. This represents a high potential for energy savings. If, for example, a data center can consolidate five workloads from five under-utilized servers to one server running Hyper-V, that is four less servers being powered in the data center. Consolidation numbers even higher than this have been possible within some environments. “Consolidation ratios depend a lot on the workloads being deployed, but I think we typically see,” McGrane said, “around three virtual machines have been consolidated to each processor core—as a rule of thumb for the consolidation rate that’s possible with virtualization and with Hyper-V. That’s a substantial opportunity to save power.”

 In summing up the Intel and Microsoft engagement, McGrane said, “With Windows Server 2008, for the first time—by default—the operating system will manage the power of the processors to the needs of the workload. The operating system basically monitors the utilization level of the processors and based on that utilization level it will automatically raise or lower the performance states on the processors. It will only use the processor power that’s required to drive the workloads at any point in time. And, this all happens transparently to the user.”

# Contributors

This article includes contributions from the following individuals:

* Sean McGrane, Microsoft
* Ward Ralston, Microsoft
* J.K. Lialias, Intel
* David Filani, Intel
* Shiv Kaushik, Intel
* Anil Aggarwal, Intel
* Irwinder “Richie” Singh, Intel

Many thanks to all for the information and insights provided and the generous investment of time in helping produce this article.

### **About the writer, Lee Purcell:**

Based in Arlington, VT, Lee Purcell is the executive director of Lightspeed Publishing LLC, offering businesses new ways to communicate through digital media. A long-term member of Vermont Businesses for Social Responsibility ([www.vbsr.org](http://www.vbsr.org)) and Renewable Energy Vermont ([www.revermont](http://www.revermont)), Lee is also the author of ten trade computer books on digital media and Web technologies, which have been translated into seven languages. He blogs on alternative energy topics at lightspeedpub.blogspot.com.

# Related Information

## Future Development Possibilities

Increasingly, hardware and software developers must work collaboratively to create far-reaching solutions that span architectures, as Intel and Microsoft discovered in this long-term engagement that proactively addressed the challenge of data center power budgeting. While the aforementioned strategies exercise energy-efficiency features of Intel® Xeon® processors, future development will target other server subsystems including system memory (DRAM).

By dynamically balancing power in the data center, IT administrators can deliver what the business needs without breaking the service level contract they have with their utility company. They can stay within a fixed budget and manage their energy bills in a predictable, consistent manner. The entire solution relies on essentially moving power from one portion of the data center to another dynamically throughout the day—depending on where the need for power is the most urgent.

## Next-Generation Intel Capabilities and Intel® Dynamic Power Technology Node Manager

The unique architectural characteristics of Intel Core microarchitecture make possible the power budgeting solution that is described in this article, particularly the flexibility in dynamically modifying the T-states and P-states to run the processor at reduced power levels and varying frequencies. IT managers are now able to define a power budget for the data center. This policy is enforced by Node Manager, which serves as an out-of-band power management policy engine. Embedded in the silicon of the Intel server chipset, Node Manager works in combination with the operating system power management functions and the BIOS. Optimal performance/power ratios are maintained for each server by dynamically raising and lowering platform power in response to changing environmental and system conditions.

With node power management capabilities, customers can actually see how much power is being used. They might evaluate the situation for a few weeks and notice one server has never gone above 200 watts. In such circumstance, they could quite safely lower the budget from 500 watts to 300 watts, without restricting the performance capability of the workload. By doing so, they have the freed up 200 watts they could then use to deploy more servers into a rack. Being able to deploy more servers into a rack and fully utilize the power and cooling infrastructure helps resolve the biggest problem the customer has today: ensuring there is enough power and cooling capacity in the data center infrastructure.

Among the features provided by Node Manager:

* **Power monitoring.** Measures power consumption dynamically for each server (within a margin of plus or minus 10 percent) and generates report data to the remote management interface.
* **Power capping at the platform level.** Enforces the current power policy, as received from the IPMI interface linked to the server management console, by dynamically altering processor P-states, staying within the allocated power budget.
* **Management console alerts.** Monitors the platform power usage and when thresholds cannot be maintained, Node Manager generates alerts to the server management console.

For more information, visit these sites:

* **To learn more about Hyper-V, visit:** **http://www.microsoft.com/windowsserver2008/en/us/virtualization-consolidation.aspxfi an.com/**
* **To learn more about Intel Core microarchitecture, visit:** [**http://www.intel.com/pressroom/archive/releases/20070328fact.htm**](http://www.intel.com/pressroom/archive/releases/20070328fact.htm)

1 Dynamic Data Center Power Management: Trends, Issues, and Solutions. Filani, David; Jackson He, Sam Gao, et al. Intel Technology Journal, February 21, 2008, pg 59.

2 <http://blog.windowsvirtualization.com/wss/microsoft-it-going-green-with-virtualization-today>