

A DYNAMIC Approach to Power Budgeting

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 the next-generation Intel® Core™ microarchitecture through innovative software engineering. The ultimate solution exploits the ability to dynamically select optimal P- and T-states of the Intel Core microarchitecture processor so that software components—from Intel® Dynamic Power Technology Node Manager (Node Manager) to Microsoft 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.

“For the first time—by default—every time you run Microsoft Server* 2008, the OS will manage the processor power to the needs of the workload. And it’s all transparent to the user.”

Sean McGrane
Principal Lead Program Manager, Microsoft

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- and T-states on the processor, the power cap (limit) on the platform can be enforced.

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 focused 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 had 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 OS policy



SPOTLIGHT ON IDF

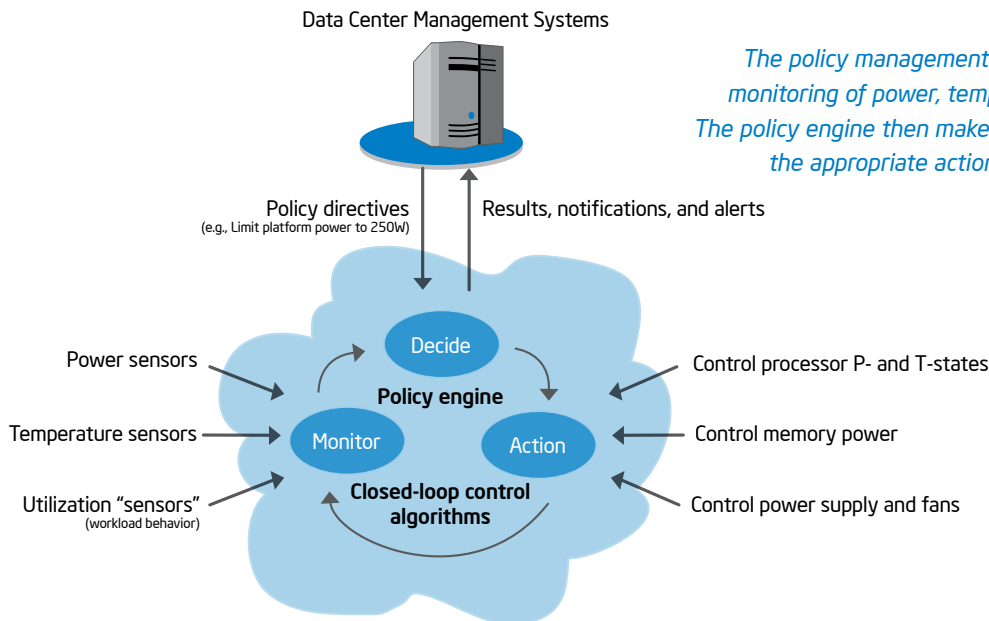
IDF Lecture Session and Live Demonstration: Dynamic Power Management

The data center management strategies pioneered by Intel and Microsoft, based on the capabilities of the Intel® Core™ microarchitecture and the features of the Microsoft Windows Server* 2008 operating system, will be presented during an IDF Lecture Session jointly organized by both companies.

The topics to be included in this session and live demonstration are:

- Power balancing and budgeting techniques in data centers, based on deployments featuring Windows Server 2008 and Hyper-V* virtualization running on a platform powered by Intel Core microarchitecture using Intel® Dynamic Power Technology.
- Dynamic power management approaches to balance workloads using Windows Server 2008 and Hyper-V power policy functionality.
- Energy efficiency optimizations available through Windows Server 2008 and Hyper-V.

These sessions are designed to equip IT managers with the knowledge to better control power allocation and use within data centers. The information should also be helpful to developers, system architects, technologists, and others involved in energy efficiency issues at the corporate level. For more information on upcoming sessions or highlights from previous sessions, visit: www.intel.com/IDF



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.

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, principal lead program manager for 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 establish 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.”

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 P- and T-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 and 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 Intelligent Platform Management Interface (IPMI) 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.

The space and power savings offered through virtualization are substantial, as shown in the comparison that was posted on a Microsoft blog.² These values were determined for Microsoft's internal IT center.

Item	Physical System Cost	Virtual Server Build Cost	Savings
Number of servers required	477 systems at a cost of USD 5,000 each	16 physical host systems at USD 20,000 each	Just under USD 2 million
Hard drive space	19 Tb	8 Tb	11 Tb
Rack Space	30 racks	2 racks	28 racks
Power	525 amps	8 amps	517 amps

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

THE ART OF VIRTUALIZATION AND POWER SAVINGS

Hyper-V is the virtualization component of the Microsoft 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 Microsoft Server 2008 as another operating system role. You boot the standard Server 2008 image and then you can configure it for Hyper-V. When you do that, it enables the hypervisor and Microsoft's virtualization solution that comes with Server 2008. One of the good things is that Hyper-V offers the same power management benefits that every other 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 states (P- and T-states) 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 trade-offs for hardware and software. All of the statistics that we have show that most data centers' 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 Microsoft's 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 ratio is very dependent on the types of workloads being consolidated. 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 Microsoft 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



Microsoft's Hyper-V* dynamically manages processor performance states to the needs of the workload. IT Managers... take a break!

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.

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, principal lead program manager, Microsoft
- Ward Ralston, group product manager, Microsoft
- Lee Purcell, contributing writer, *Intel® Software Insight* magazine

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

There's More



New Intel Magazine Can Help Ignite Your Visual Adrenaline

If computer graphics is your passion, our exciting new magazine, *Intel® Visual Adrenaline*, may be right for you.

Published quarterly, the *Intel Visual Adrenaline* magazine covers all things computer graphics, including using multi-threading and code optimization to render your games and apps, cool tools and development products, and news and articles from industry innovators.

It's free, it's radical, and it's yours today with your subscription to Intel® Software Dispatch for Visual Computing.

Subscribe at www.intelsoftwaregraphics.com