

# Intelligent Power Optimization for Higher Server Density Racks

## A Baidu\* Case Study with Intel® Intelligent Power Technology

### White Paper

Digital Enterprise  
Group

Q1 2008

Intel®  
Corporation

### Executive Summary

Intel® partnered with Baidu.com\* and conducted a proof of concept (PoC) project using the Intel® Intelligent Power Node Manager (Node Manager) and Intel® Data Center Manager to intelligently optimize server performance and power consumption to optimize server density. The engineers implementing the PoC used Node Manager to identify optimal control points, which became the basis to define node level power optimization policies. Intel® Data Center Manager (Data Center Manager) Reference Graphical User Interface was used to manage servers as a group to carry out rack power policies while minimizing performance impact of all the managed nodes in the rack. The policies enable increasing rack density and workload yield even when the whole rack is subject to power limitations.

The PoC was conducted initially on site at Baidu in Q1 2008 at Intel-Baidu joint lab with a configuration close to production configurations running Baidu search workloads. The runs in the first round were carried out with Bensley generation platforms. An updated set of runs was carried out in Q1'2009 using pre-production units with the more advanced Nehalem platform.

The test servers used were provisioned with dual X5560 processors. These processors feature eleven ACPI p-states, allowing the frequency to vary from 2.8 GHz to 1.6 GHz instead of the three states available in the Bensley generation. The additional p-states allow finer grained power control and an expanded power dynamic range.

Here are some global results:

- For the Bensley generation it was possible to trim power consumption by 40 W running the Baidu search workload while staying within the acceptable performance boundaries determined by Baidu. For the Nehalem platform the range has been expanded to 70 watts, allowing Baidu increased operational flexibility.
- At rack level, up to 20% additional capacity increase could be achieved within the same rack-level power envelope when aggregated optimal power management policy is applied
- Comparing with today's datacenter operation at Baidu, by using Intel® Node Manager and Intel® Data Center Manager, there could be a rack density increase from 5 to 7/8 servers – a 40%+ improvement

# Table of Contents

<b>1. Business Overview</b> .....	<b>3</b>
1.1 Top Business Issues .....	4
<b>2. Intel® Technology and Solution</b> .....	<b>4</b>
2.1 Intel® Intelligent Power Node Manager (Node Manager) .....	4
2.2 Intel® Data Center Manager (Data Center Manager) .....	5
<b>3. POC Use Cases</b> .....	<b>5</b>
<b>4. POC Architecture</b> .....	<b>6</b>
<b>5. POC Results</b> .....	<b>7</b>
<b>6. Conclusion</b> .....	<b>8</b>

# 1. Business Overview

Baidu is the largest search company in China, accounting for over 60% of search market share in China. Its market share in the Chinese domestic market has grown steadily in the past few years. Baidu's reach extends to international markets, with established branch offices in Japan, the U.S. and other countries.

Currently Baidu uses China Telecom\* as its data center hosting provider. Baidu is billed by the rack. Each rack comes with strict power limits of no more than 10 A per rack or 2.2 KW at 220 V.

Because racks are currently power limited, a significant amount of rack space can't be used without hitting power envelope limits. Since Baidu is also billed by the rack, Baidu is highly motivated to maximize the number of servers per rack while staying within the 10 A current limit.

As Baidu grows, the company will eventually operate company-owned data centers. Given their previous experience, power consumption remains one of the top platform management concerns, and is not expected to abate even with company operated data centers.

Figure 1: Baidu.com Search Portal



## 1.1 Top Business Issues

As mentioned above, data center hosting represents a major operational cost for Baidu, costs being proportional to the number of racks leased from China Telecom. Power constraints limit the number of servers that can be placed in a rack, and a significant amount of space goes unused and wasted. An easy way to increase the compute yield is to tightly manage the number of servers within the power envelope. Unfortunately currently there is no accurate means to measure power consumption against the power limit. Hooking a power meter would only give a snapshot in time in what is a very dynamic environment. The closest data would be using a derated nameplate figure. This figure is still overly conservative because it needs to account for the worst case in power consumption. In other words, there is no dynamic power management technology which allows Baidu to optimize power utilization. To summarize, Baidu is facing the following power management challenges:

- **Over-allocation of power:** Power allocation to servers does not match actual server power consumption. Power is typically allocated for worst case scenario based on server nameplate. Static allocation of power budget based on worst case scenario leads to inefficiencies and does not maximize use of available power capacity and rack space.
- **Under-population of rack space:** As a direct result of the over-allocation problem, there is a lot of empty space on racks. When Baidu needs more compute capacity, they have to pay more for additional racks. China Telecom on the other hand, is operating at capacity and Baidu has leased most of the available racks in Beijing. Available datacenter space is limiting factor to Baidu's business growth.
- **Capacity planning:** Baidu does not have means to forecast and optimize power and performance dynamically at rack level. To improve power utilization, datacenters needs to track actual power and cooling consumption and dynamically adjust workload and power distribution for optimal performance at rack and datacenter levels.

## 2. Intel® Technology and Solution

### 2.1 Intel® Intelligent Power Node Manager (Node Manager)

Node Manager is an out-of-band (OOB) power management policy engine embedded in Intel® server chipsets. Processors carry the capability to regulate their power consumption through the manipulation of the P- and T-states. Node Manager works with the BIOS and OS power management (OSPM) to perform this manipulation and dynamically adjust platform power to achieve maximum performance and power for a single node. Node Manager has the following features:

- **Dynamic Power Monitoring:** Measures actual power consumption of a server platform within acceptable error margin of +/- 10%. Node Manager gathers information from PSMI instrumented power supplies, provides real-time power consumption data singly or as a time series, and reports through IPMI interface.
- **Platform Power Capping:** Sets platform power to a targeted power budget while maintaining maximum performance for the given power level. Node Manager receives power policy from an external management console through IPMI interface and maintains power at targeted level by dynamically adjusting CPU p-states.
- **Power Threshold Alerting:** Node Manager monitors platform power against targeted power budget. When the target power budget cannot be maintained, Node Manager sends out alerts to the management console.

## 2.2 Intel® Data Center Manager (Data Center Manager)

Intel® Data Center Manager is software technology that provides power and thermal monitoring and management for servers, racks and groups of servers in datacenters. It builds on Intel® Intelligent Power Node Manager and customers existing management consoles to bring platform power efficiency to End Users. Data Center Manager implements group level policies that aggregate node data across the entire rack or data center to track metrics, historical data and provide alerts to IT managers. This allows IT managers to establish group level power policies to limit consumption while dynamically adapting to changing server loads. The wealth of data and control that Data Center Manager provides allows data centers to increase rack density, manage power peaks, and right size the power and cooling infrastructure. It is a software development kit (SDK) designed to plug-in to software management console products. It also has a reference user interface which was used in this POC as proxy for a management software product. Key Intel® Data Center Manager features are:

- Group (server, rack, row, PDU and logical group) level monitoring and aggregation of power and thermals
- Log and query for trend data for up to one year
- Policy driven intelligent group power capping
- User defined group level power alerts and notifications
- Support of distributed architectures (across multiple racks)

## 3. POC Use Cases

In this POC we focused on use cases to test Node Manager features at node level first. A baseline test is needed to identify the optimal control points at the node level for Baidu search workload. We then used these optimal control points as the base for rack level policy definition. A summary of use cases is listed below:

Use Case Title	Description
Get power consumption on each server	Using the Intel® Node Manager features to dynamically gather point in time power consumption from each server on the rack
Estimate total power consumption of a rack	Estimate rack level power consumption by summing up node level power consumption; display on, and notify console as appropriate
Optimize rack level policy within a given power envelop and server workload	At rack level, analyze the power consumption of each server, overall power consumption, rack level power envelope, and targeted performance goals (utilization, response time, query queue length, etc.) as well as other factors important to Baidu to determine the optimal power distribution policy. Baidu will set the policy and optimization strategy based on their work load and priority.
Apply a common policy to the servers in a rack	Using Intel® Data Center Manager, set policy to each rack in terms of particular power budget target that the server has to observe
Node-level monitoring and tracking against policy	Leveraging Node Manager features to adjust server power consumption to the target set by the policy within 60 seconds and maintain at the target until further notice
Node-level alert and notification	Use Node Manager feature to detect and send alert when a server fail to reach policy target in 60 seconds or maintain the target during operation.
Alert handling and mitigation	Once an alert is received, the console needs to automatically decide upon a course of action to mitigate the risk – ignore, set a new policy, or shutdown the troubled server, etc.

Note: The last two use cases "Node-level alert and notification" and "Alert handling and mitigation" are not covered in this POC.

#### 4. POC Architecture

This POC was set up at Intel-Baidu joint lab on Baidu campus. The rig consisted of four Nehalem-EP based servers used in this POC with 2 Intel® Quad-Core Xeon® processors with eleven p-states (2.8 down to 1.6 GHz). Each server was configured with 18 GB of DDR3 memory and PSMI 1.44 instrumented power supplies. The servers are installed on a rack as a server group, managed by Intel® Data Center Manager, as shown in Figure 2.

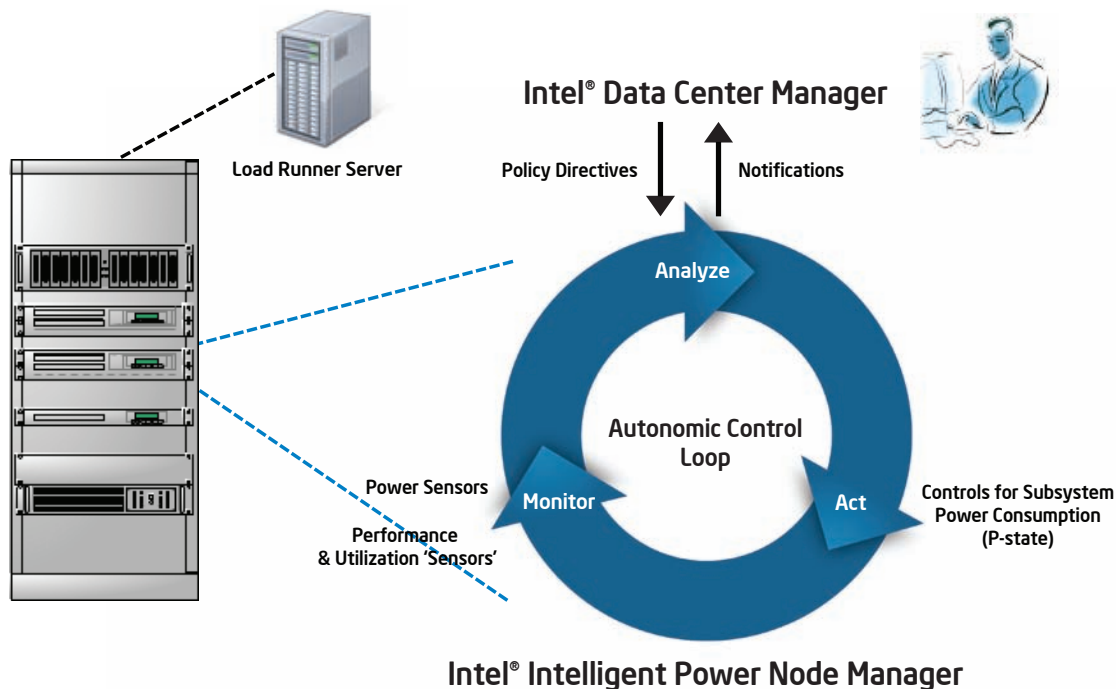
The servers in the rack were loaded with the same operating system configuration, Red Hat® Linux Enterprise Version 4.0 Update 3 with customized kernel patches from Baidu. Each server was configured to run Baidu's workload stress test which measured the number concurrent queries a server could process every second. Each server under test also had Node Manager configured.

Intel® Data Center Manager was used to see server and rack level actual power consumption data. The Reference User Interface was used as the group management console. Data Center Manager monitored the actual power consumption on each server and aggregated total power consumption for the group (rack level.) Data Center Manager combines instances Node Manager across multiple servers to set appropriate policies thereby providing a mechanism to optimize power consumption for each node, yet stay within the constraint of a the rack-level power budget.

For test purposes, we also connected watt meters to the rack and servers under test to monitor the rack level power consumption as an independent confirmation to the power numbers reported by the instrumented PSMI power supplies and the aggregated numbers provided by Data Center Manager. In most cases, there is a small discrepancy between watt meter and Node Manager figures. Baseline numbers were noted before running the actual tests.

A load runner server was used to generate graduated loads for the servers in the rack. The tool also had a capability to generate execution statistics from the various workload tests carried out.

Figure 2: POC Architecture



### 5. POC Results

System architects wonder about the performance tradeoffs involved when voltage and frequency scaling are applied to a CPU running a workload. The figure below depicts parametric sensitivity curves of response times versus power capping.

An initial set of calibration runs to establish asymptotic values for workloads in terms of number of threads (one query per thread) and response times at full power. The asymptotic query time is about 0.6 ms. At this point the system is considered fully loaded. Increasing the number of queries per second beyond this point starts degrading the response times appreciably.

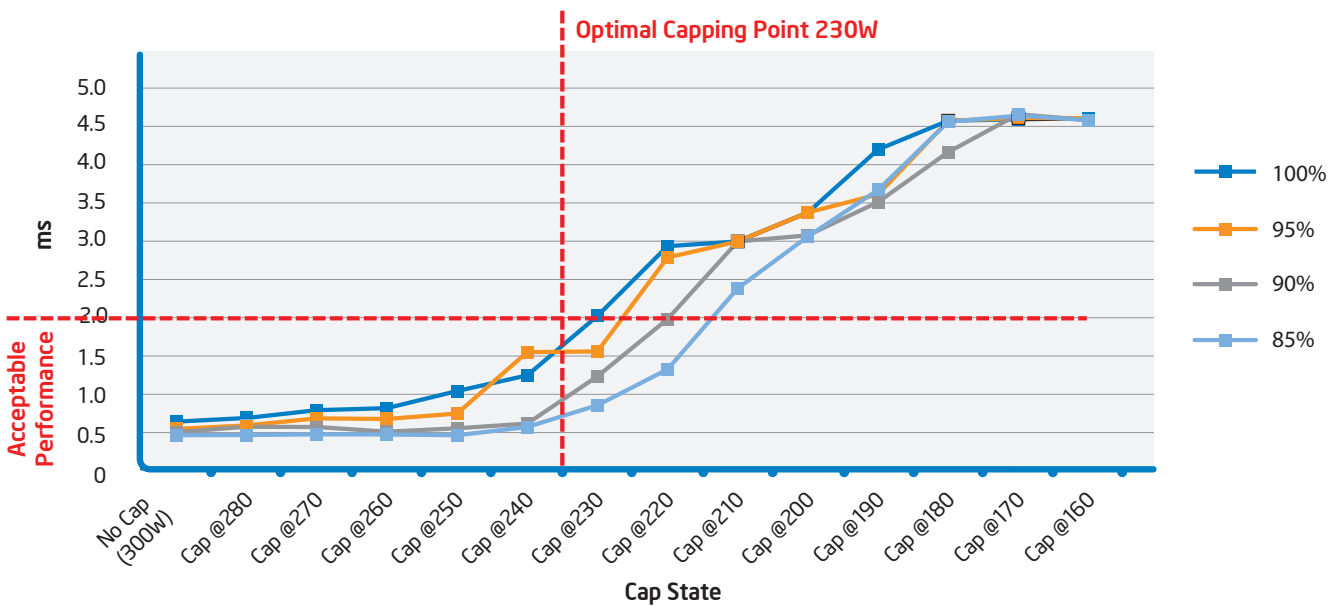
Additional plots were made with reduced number of threads to depict performance sensitivity to the number of threads. Plots were made at baseline, and 95, 90 and 85 percent of the number of original queries per second.

Each curve represents the system response time as a function of power capping, from no capping at all to capping level set at 160 watts while the workload is held constant. Note a gentle degradation in response time until the power capping level reaches 230 watts.

Baidu set the threshold for acceptable response time at 2 milliseconds. This threshold sets a practical limit for a power rollback of about 70 watts before the system crosses the threshold. Counter intuitively, for moderate power capping response time is slightly better for the lower thread counts plotted. This may indicate some thread processing overhead.

The compression after the cap is set to 180 watts is due to the lower range of power control authority being reached.

Figure 3: Effect of Power Capping on Search Response Time



## 6. Conclusion

The 70 watt number is a precise actual number for power consumption derived from the instrumented power supplies in the Nehalem-based servers. This knowledge allows increasing the number of servers in a rack without exceeding the power limits imposed by the hosting provider. There are two approaches that we can apply to benefit from the results of this measurement.

As indicated in Figure 4, the hosting provider imposes a 10A current limit per rack or about  $10\text{ A} \times 220\text{ V} = 2.2\text{ KW}$  ceiling per rack. The circuit breakers are actually built with 100 percent headroom. However, servers are provisioned to not exceed the nominal limit of 10 A, and will trip instantly if the current draw is doubled.

Without the monitoring facilitated by the instrumented power supplies, the rack provisioning planners need to assume the more conservative figure of a derated nameplate figure, which amounts to 400 watts per server.

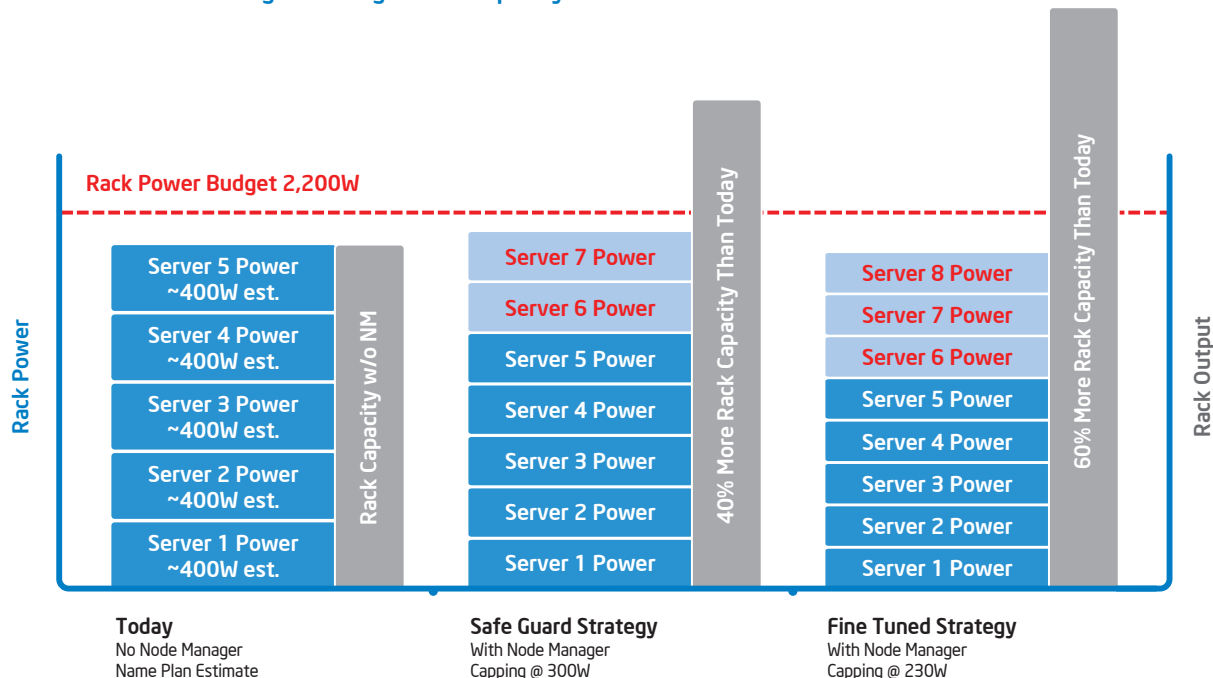
Alternatively, for legacy platforms without power monitoring capability, this number is calculated through a pre-test measurement, adding extra power draw to account for the worst case from configuration changes during the server lifetime plus an additional safety margin.

In this baseline case, if the best known figure is 400 watts, this means that a rack cannot be loaded beyond five servers without exceeding the power quota. For servers with the common 2U form factor a rack can usually take 20 servers. This means the rack must be left 75 percent empty with a significant waste of available rack space.

Having the more precise power consumption figures from Node manager, the two strategies to improve the rack-level power utilization are the following:

- **Safeguard strategy:** This is a way to set up a safeguard limit for the maximum power consumption value for a given workload. In other words, we look at the historical record of actual power consumption readings provided by Node Manager and set a cap at the maximum power level to minimize power excursions beyond this set limit. This number is still significantly lower than the derated nameplate number. From the POC, we measured that a server typically consumes 300W or less. Hence if each server is operated with a Node Manager imposed ceiling of 300W, now we can load up the rack with seven servers for a total rack consumption of  $300\text{ watts} \times 7 = 2.1\text{ KW}$ , or 40 percent increase over the original, more conservative assumption of 400 watts per rack because we know that the servers won't consume more than 300 watts, and even if they do, the 300 watt policy limit will make excursions very unlikely. We assume that the hosting data center has sufficient thermal capacity to sustain the extra servers. Because the power limit has been set to the expected peak power usage, power capping will kick on only very rarely, and no performance impact is expected.
- **Fine tuned strategy:** This strategy requires several experiments of power capping at different levels yielding results similar to the ones described under the PoC results. The goal of these experiments is to determine the feasible capping range that still yields acceptable performance. These experiments take work but will deliver a finely tuned system that not only stays within the power envelope but also delivers acceptable performance according to preset criteria. This strategy requires active power capping at all times. For the experiments we performed, we learned that we could roll back power consumption down to 230W without undue deterioration in search response times. With power capped at 230 watts, we can now install eight servers in the rack for a total consumption of 1.9 KW ( $8 \times 230\text{W}$ ). If we set rack-level policy at 1,900W, we can now install three extra servers in the rack up from the original five. In other words, other things being equal, rack capacity had been increased by 60 percent.

Figure 4: Different Power Manager Strategies and Capacity Increases



Looking into the PoC results and subsequent analysis, it is clear that Intel® Node Manager and Data Center Manager allow a significant increase in server density and compute capacity of a rack through dynamic power management policies. The rack consumption still stays within the present power envelope with acceptable performance impact. Node Manager Technology provides a monitoring capability and a defined power consumption ceiling allowing data center managers to safely increase rack loading between 40 and 60 percent.

The power management approaches described in this document are relatively simple, derived from an initial study. The results reported in this study are by no means definitive. For workloads of different characteristics, additional, more aggressive power management approaches are possible that promise to increase the rack yield even further.

For more information on Intel® Intelligent Power Node Manager  
and Intel® Data Center Manager visit:  
[datacentermanager.intel.com](http://datacentermanager.intel.com)

Copyright © 2008 Intel Corporation. All rights reserved.

Intel, the Intel logo, Core 2 Duo, Celeron, and vPro are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. 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.

\*Other names and brands may be claimed as the property of others.

