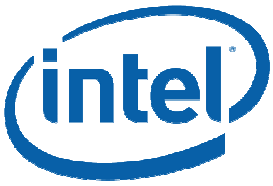

Dynamic Power Optimization for Higher Server Density Racks – A Baidu Case Study with Intel® Dynamic Power Technology

Executive Summary

Intel's Digital Enterprise Group partnered with Baidu.com conducted a proof of concept (POC) project using the Intel® Dynamic Power Node Manager Technology (Node Manager) to dynamically optimize server performance and power consumption to maximize the server density of a rack. The POC engineers used Node Manager to identify optimal control points, which became the basis to set power optimization policies at the node level. A management console – Intel® Datacenter Manager (Datacenter Manager) was used to manage servers at rack-level to coordinate power and performance optimization between servers to ensure maximum server density and perform yield for given power envelope for the rack.

The POC was conducted in Q1' 2008 at Intel-Baidu joint lab under a simulated Baidu production environment. The POC results illustrated that:

- At a single node level, up to 40W savings / system without performance impact when a optimal power management policy is applied
- 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, there could be a rack density increase 20~40% improvement

More detailed findings from this POC are published in this paper.

Contents

Business Overview

Baidu is the biggest 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, over the past years. It also extends to international markets and has started branch offices in Japan, the U.S. and other countries.

Currently Baidu leases racks at datacenter to provide their services on the interent. It pays by rack and each rack is power limited. Increasing server density on a rack is of great interest for Baidu.

Figure 1: Baidu.com Search Portal

Top Business Issues

Datacenter hosting is a primary cost for Baidu business operation, which is proportional to the number of leased racks. Currently, the number of servers on each rack is low due to power constraints. Baidu is interested in increasing server density on a rack within the power constraints set by the datacenter service provider, in order to maximize the performance on each rack.

In current datacenter operation today, there is no dynamic power management technology which allows Baidu to optimize power utilization. As a result, Baidu is facing 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. Available datacenter space is a limiting factor to Baidu's business growth.

- Capacity planning: There is lack of effective means to optimize the 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.

Intel Technology and Solution

Intel® Dynamic Power Node Manager (Node Manager)

Node Manager is an out-of-band (OOB) power management policy engine that is embedded in Intel server chipset. It works with BIOS and OS power management (OSPM) to dynamically adjust platform power to achieve maximum performance/power at node (server) level. 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 supply, provides real-time power consumption data (point in time, or average over an interval), 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

Intel® Datacenter Manager (Datacenter Manager)

Intel® Datacenter Manager is a software add-on to Node Manager to monitor and control power for a group of servers. Intel® Datacenter Manager depends on Intel® Dynamic Power Node Manager. 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® Datacenter Manager features are:

- Group level monitoring of power
- Log and query for trend data
- Group power limiting
- Group level power alerts and notifications
- Support of distributed architectures (across multiple racks)

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. 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.
Set policy to servers on the rack	From the console, 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.

POC Architecture

This POC was set up at Intel-Baidu joint lab on Baidu campus. There were 4 Intel® Bensley servers used in this POC with 2 Intel Quad-Core Xeon® processors with 3 p-states (2.66GHz, 2.33GHz, and 1.99GHz). Each server is configured with 8 GB of memory and PSMI 1.44 instrumented Power Supply. The servers are installed on a rack as a server group, managed by Intel® Datacenter Manager, as shown in Figure 2.

All servers on the rack have the same OS configurations – Optimized version of Linux from Baidu. Each server was configured to run Baidu’s simulated workload stress test.

Datacenter Manager Reference User Interface was used as the group management console. Datacenter Manager monitored the actual power consumptions on each server and aggregated total power consumption at the rack level. It works with Node Manager to set appropriate policies which ensures servers in a rack are delivering the best performance within the rack-level power budget.

For test purpose, we also had a power meter connected to the rack or a server under test to monitor the rack level power consumption, independent of Node Manager and Datacenter Manager as a reference, in order to know Node Manager and Datacenter Manager data is accurate. In most cases, there is a delta between the power meter reading and Node Manager reading, which is why a baseline test is needed before doing the actual tests.

A load runner server was used to generate the workload needed at different level to stress-test servers on the rack. It also gathered result from the workload execution and came up statistics of the workload tests, which would be used for performance analysis.

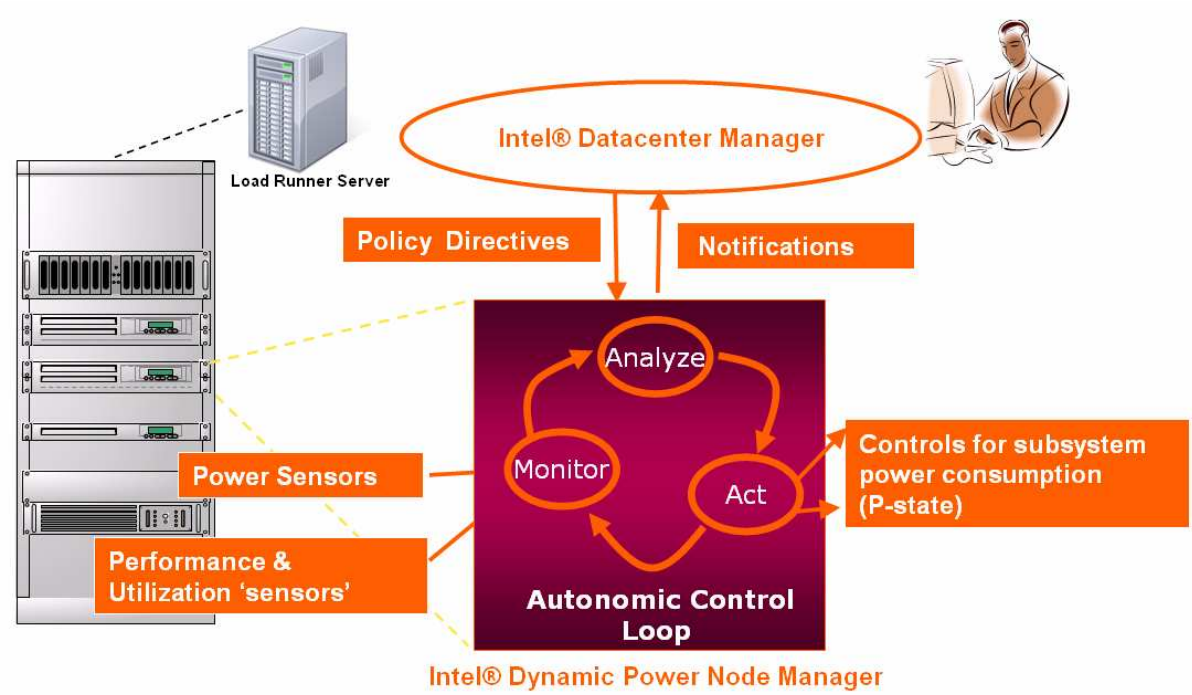

Figure 2: POC Architecture

POC Results

Single Node Testing

The purpose of single node testing is to make sure that Node Manager is well integrated with the server platform, and that the features of Node Manager are functioning properly. More importantly we need to conduct some baseline testing to calibrate that Node Manager power monitoring reading is consistent with the power reading from the wall. In addition, we need to characterize the application and the platform under test to identify the "sweet-spot" for power management; where we get the maximum power reduction with the minimum performance loss. To sum it up, objectives for single node testing are:

- Node Manager feature validation
- Platform power monitoring calibration
- Performance/power optimization

The tests indicate that Node Manager performs well with the Baidu workload configuration. The monitoring and power capping functions work well at different levels of stress test.

CPU use	Node Manager reading (W)	Power meter reading (W)	Delta (W)
100%	371	367	4
87%	357	352	5
74%	330	324	6
62%	307	300	7
50%	275	264	11
37%	261	253	8
25%	253	247	6
13%	220	220	0
0	192	186	8

Table 1: CPU utilization and platform power consumption

The table above shows the result of power monitoring calibration test, which indicate that Node Manager power readings are not far apart from the power meter reading measured directly from the power source plugs. The delta is anywhere between 0~11W depending on the different CPU utilization level. We can use this table to adjust the value read from Node Manager and more accurately estimate the actual power consumption of the entire platform at a given time. We will use this table to calibrate Node Manager power reading for analysis later.

Following the basic power monitoring test and comparison with power meter reading form the power source, we tested the power capping feature of Node Manager. First we observe the range of power consumption that Baidu workload operates without power capping. The optimal workload is when the CPU utilization is about 50~60% with peak power at about 300W (see Figure 4).

With that observation, two levels of power capping were tested: 260W and 200W. The reason was that minimum of 40W power reduction was needed from a 5-server rack in order to add an additional server to achieve our goal of increasing server density. The test was to prove if that would be the case. In addition, it was clear that the server power consumption could not go below idle power which was 200W. By setting the power capping to 200W, the maximum power reduction for the given platform.

Figure 3 is the actual power consumption measurement from Node Manager: 1) no power capping, 2) power capping at 260W, and 3) power capping at 200W. From the diagram, we have the following observation:

- No power capping: the power consumption is around 290W most of the time with a few spikes to 300W.
- Power capping 260W: the power consumption is around 260W most of the time, with quite a few spikes going above 270W, as Node Manager dynamically maintains the platform power against the power capping target.
- Power capping 200W: the power consumption did not go as low as 200W, but rather converged at around 250W. This is because with the workload, the Node Manager cannot force the platform consumption to as low as 200W – 250W is as low as it can go.

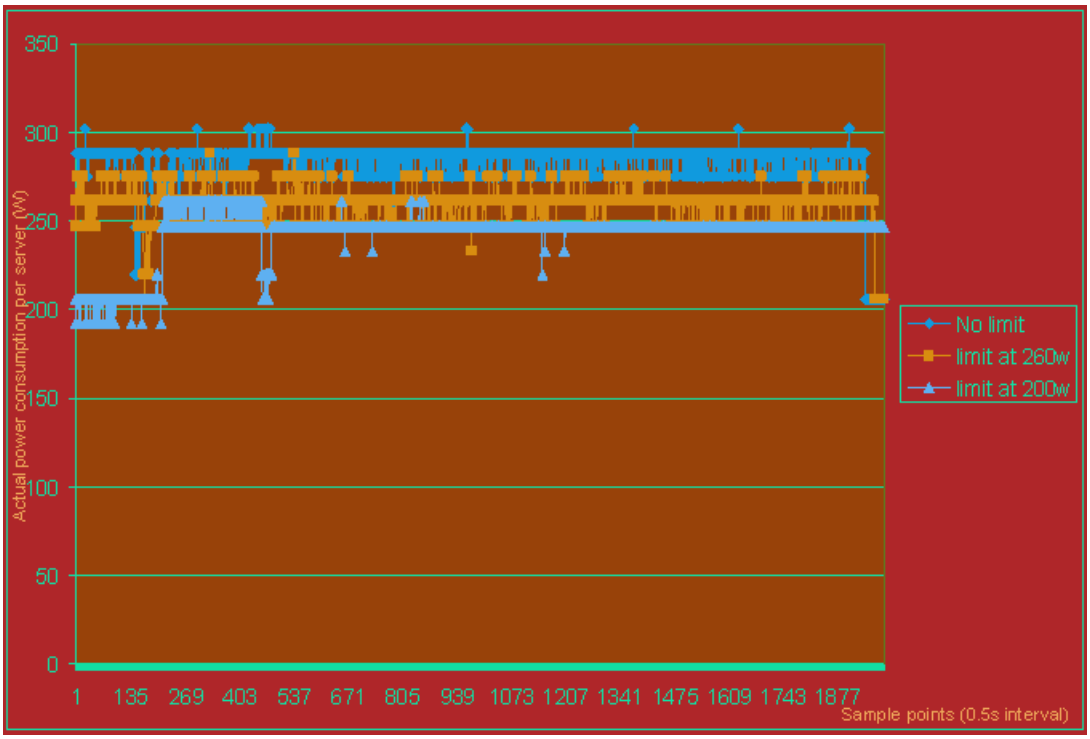

Figure 3: Power capping with different values

Figure 4 shows the performance per queries number under different workload. With this conclusion, it is safe to say that we can reduce platform power consumption to as low as 250W without performance penalty within the optimal range of Baidu workload for the given server under test. This is a significant power reduction comparing to the peak power with no limit, which is ~300W. With the power reduction of 40~50W per server without performance loss, we have "squeezed" enough power at the rack-level among the 5 servers to add another server to the rack. This result gives us enough confidence to start rack-level test by having Intel® Datacenter Manager administrate power policies among servers on the rack and dynamically monitor power consumption at the rack-level.

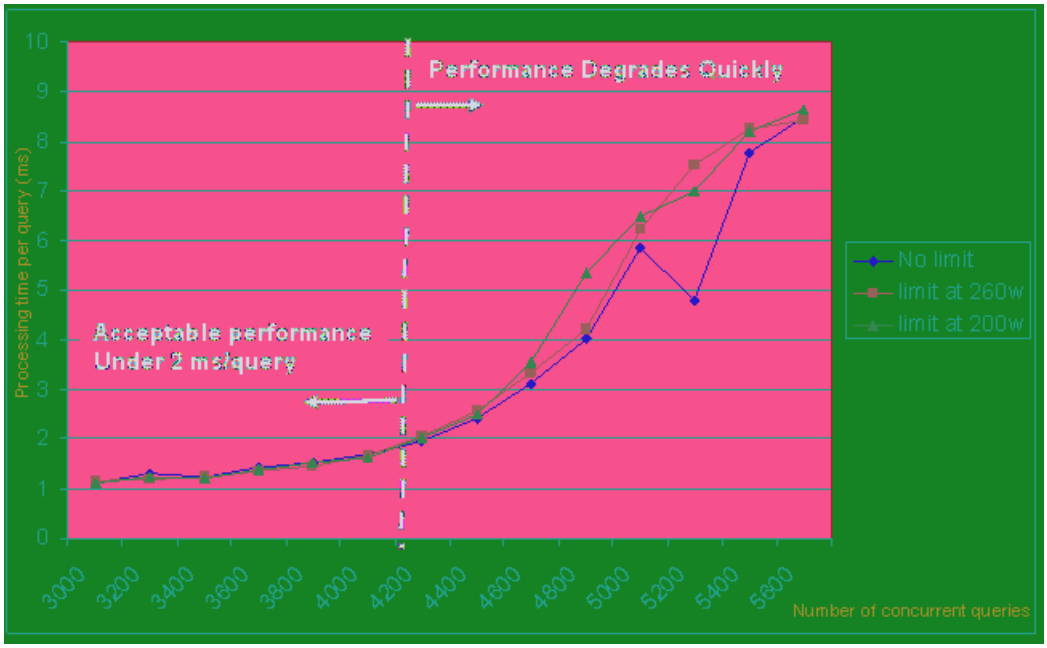

Figure 4: Performance under different power capping

Rack-level testing

Following the single node level test, 3 identical Intel Bensley servers were tested as a group of server on a rack that is managed by Datacenter Manager. We did not use the 5 servers to fill the rack as Baidu datacenter operation today due to test equipment availability. However, as we demonstrated that the power consumption at rack level is linear to the number of servers on the rack since all the servers are running identical workloads, we can extrapolate the scenarios with 5 or more servers with the result of 3 servers on the rack.

Based on the result of single server test, we know that we can cap a single server to 250W with little performance loss. Therefore, during the rack level test with 3 servers, two scenarios were tested: i.e. no power capping policy; and power capping policy at 750W (3x250W). Intel® Datacenter Manager was used to monitor the power consumption and enforce the power capping policies at rack level.

Figure 5 is the aggregated power measured at the rack level with power capping and without power capping policies. As indicated by the chart, when no power capping policy is applied, the rack level power consumption was close to 900W (blue line). When 750W power capping policy was applied, the rack level power consumption was clumped down to close 750W (purple line). The fluctuation in the chart represents dynamic nature of the workload and power management through Intel Node Manager. The spikes of purple-colored line were short and few, which meant Intel Node Manager and Datacenter Manager worked together to enforce and maintain the rack level power policies.

Based on the fact that rack level power consumption is linear to the number of servers on the rack, we can extrapolate that when there are 5 servers on the rack, the power consumption without power capping policy will be close to 1,500W (5x300W). When a power capping policy of 1,250W (5x250W) is applied, Intel Node Manager and Datacenter Manager can work together to enforce and maintain the policy at rack level. Therefore, Intel Node Manager can squeeze ~250W (1,500W-1,250W) of power at the rack level to add another server to the rack within the same power envelope.

Figure 5: Rack-level power capping policy

To extrapolate this result to a 6-server rack with power capping policy of 1,500W (6x250W) enforced by Intel Node Manager and Datacenter Manager, the rack has one more server to do useful work comparing a 5-server rack without power capping policies under the same power envelope (5x300W = 1,500W). This translates into a 20% increase of the rack-level compute capacity within the same power constraint.

Conclusion

The result of the POC showed that Intel Node Manager reduced power consumption at node level by 40~50W without much impact to Baidu workload performance. When the same concept applied at rack level with a combination of Intel Node Manager and Datacenter Manager working together, the rack-level capacity could increase ~20% within the same power envelope without performance impacts.

When the key learning from this POC is applied to the current Baidu datacenter operation of leased racks from China Telecom, there are two strategies to apply Intel Node Manager for power management for improved rack-level capacity within the same power envelope. Let's look more closely of the two strategies and how they compare with today's power allocation strategy.

As indicated in Figure 6, a single rack power limit is at about 2.2KW. Today, there has not been a good way to closely monitor and control the power of each server on the rack, the only way is to run a test, estimate the worst case scenario, and add a safety buffer, so that enough power is

allocated to servers on the rack and they will not run over the power budget at rack-level. This is the case of the left diagram in Figure 6 where each server was allocated ~400W and 5 servers were put on a rack. Obviously, the power budgeted allocated to the rack was not used effectively, which leaves a lot room for improvement.

With Node Manager, the two strategies to improve the rack-level power utilization are the following:

- **Safe guard strategy**: This is a way to set up a safe guard limit for the maximum power consumption value for a given workload. Even though this is not the most optimal power allocation, it is still much better and controllable than nameplate-based estimate as in the left diagram of Figure 6. From the POC, we measured that a server typically consumes 300W or lower. By setting a safe guard policy of 300W in Node Manager, the datacenter managers could easily load 6 servers for total of 1,800W (6x300W) at rack level – a 20% increase of the rack-level capacity. If we really want to push the limit, we may even fit 7 servers on the rack for 2,100W (7x300W) and still within the power envelope (2,200W). As indicated by the center diagram of Figure 6, Node Manager gives datacenter manager the confidence and control to allocate the power budget to increase the rack-level capacity upto ~40% from today's situation (increased 2 servers) assuming we have enough cooling for the additional servers.
- **Fine tuned strategy**: This strategy requires several experiments of power capping at different levels between the maximum and minimum power consumption of a given workload and the impacts on performance at these different capping point. As we learned from this POC, it takes a lot more efforts to find these optimal control points for performance and power trade off, nevertheless, it delivers better rack-level server density and compute capacity. For the case of Baidu, we learned from the POC that 250W was the optimal power capping point with little performance impact to Baidu applications. As indicated by the right diagram of Figure 6, we can apply power capping policy of 250W through Node Manager on each server and easily fit 8 servers on the rack for total of 2,000W (8x250W), still leave ~200W for fluctuations. As a result, we could increase the rack-level capacity by ~60% from today's situation (increased 3 servers) assuming we have enough cooling for the additional servers.

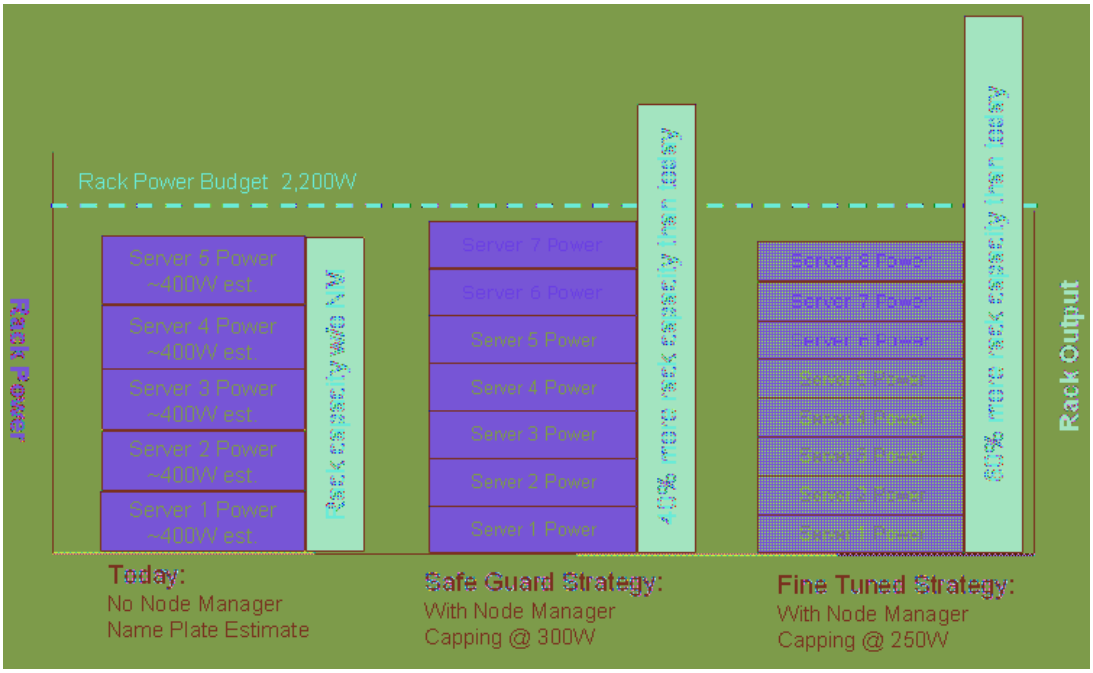

Figure 6: Different power manager strategies and capacity increases

Based on our POC and subsequent analysis, it is clear that Intel Node Manager and Datacenter Manager can safely increase the server density and compute capacity of a rack within the same power envelope through dynamic power management policies. It provides safe guard and monitoring needed for the datacenter managers to ensure that power management policies are enforced properly without major impact to application performance.

Through the POC, we also realized that it takes a lot of additional efforts to integrate Intel Node Manager with a datacenter management environment and derive optimal power management policies. Even though Intel Datacenter Manager makes it easier to integrate at rack-level, there is still a lot more work to do to make Intel Node Manager from a POC to fully integrated functioning systems at Baidu's datacenters.