White Paper

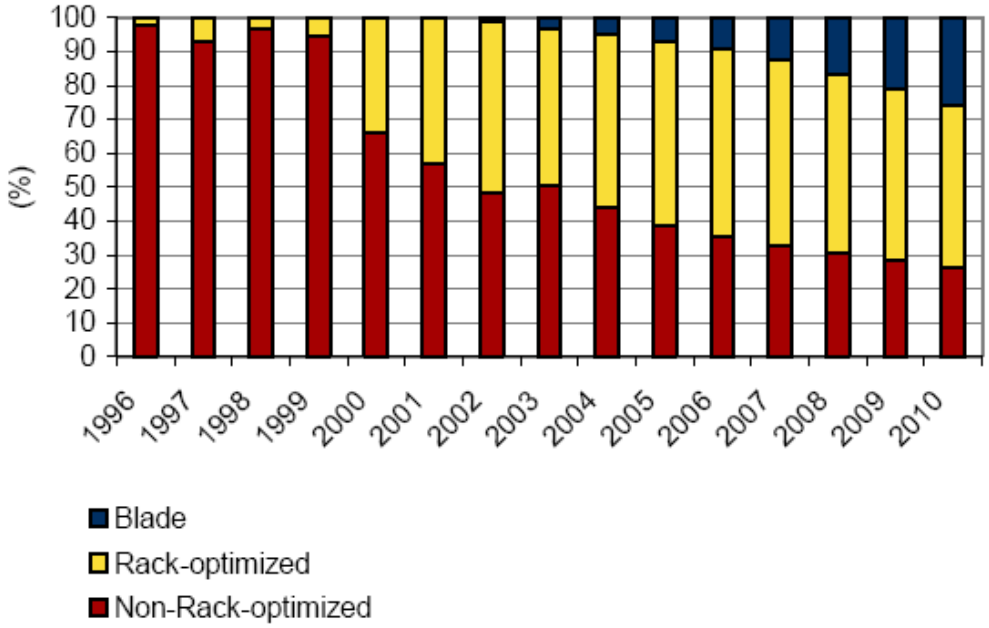
Intel® Intelligent Power Node Manager

Power consumption is the biggest area of concern in the server domain. Between 2000 and 2005, server electricity use more than doubled, growing at a rate of 14 percent each year. The power consumption cost of servers is rapidly growing and management costs have gone higher forcing IDC to identify power management as the critical need of the hour.

By 2007, Gartner had predicted that power and cooling spending will exceed server spending. By 2008, 50% of all data centers will have insufficient power and cooling capacity to meet the demand of high density equipment. Power and cooling costs will increase to more than one-third of the total IT budget

Enhancements in processor performance and density of servers have been the prime contributors to an increase in power consumption. The Power and cooling spend on server is supposed to increase to $30 billion approx in 2010 from $10 billion in 1996 as indicated in the following Table 1.

Worldwide Server Installed Base by Form Factor, 1996–2010



Source: IDC, 2006

While the components of a server system - processor, operating system, and so on - employ different techniques to check power consumption, areas that could potentially help to reduce power consumption are largely untapped.

1. Intel® processors and power management

Intel® processors support many features for power management. Prominent among them are demand based switching using the Intel® SpeedStep® Technology and clock modulation.
Intel also provides other power efficiency techniques such as the following:
- Enhanced HALT state (C1E) or Turbo Mode
- Quad-Rank Fully Buffered DIMM (FB-DIMM)
- Closed Loop Thermal Throttling (CLTT)

The Intel® 5500 series processor (released in early Q1 09) at runtime dynamically manages cores, threads, cache, interfaces, and power to deliver outstanding energy efficiency and performance on demand.
For details on Intel® 5500 series processor power efficiency design, refer to the paper "*First the Tick, Now the Tock: Next Generation Intel® Microarchitecture*".

In addition, Intel is announcing the availability of yet another tool of Intel servers based on Intel® 5500 series chipset using Intel® 5500 series processor. This tool is a management interface at the hands of an IT system administrator, termed as the Intel® Node Manager (NM) Technology.

This whitepaper outlines the NM technology interface, the hardware required to exercise this technology and application scenarios to make maximum use of NM in Intel® 5500 series chipset servers. Key SMB use cases have also been addressed.

2. Overview of Node Manager

Node Manager (NM) is a platform resident technology that enforces power and thermal policies for the platform. These policies are applied by exploiting processor subsystem knobs (such as processor P and T states) that can be used to control power consumption. Node Manager enables power and thermal management by exposing an external interface to management software through which platform policies can be specified. It also implements specific data center power management usage models such as power limiting. With the current implementation NM 1.5 only power control policies are available.

To effectively manage power consumption in servers, NM exposes interfaces to measure input power and platform consumed power at high CPU loads. A range of average power consumed at lowest loads is available by a call to this interface. This value can then be used by the IT administrator set a power policy budget. The policy interface of NM allows the following parameters to be set

1. Power Budget: This specifies the power budget allocated to the node in watts.
2. Time limit – This specifies the time limit within which the server needs to operate at the budgeted power limit. At the end of the time limit the server goes back to previous power consumption limit. This time limit could be a recurring one based days of the week and a 24 hr time cycle
3. Grace period –
4. Power action – action to be taken when power threshold is crossed.
5. power threshold value – user can set a power threshold value which when crossed will be used by NM to take the action specified by the user

3. Intel® 5500 series chipset systems and their hardware support needed for Node Manager

Intel® Server Chassis	Chassis-specific Server Management Features	Chassis-dependent Sensors Managed by Management Controller	Intel Policy-Based Node Manager Support
SR1600UR	• Non- redundant power: One power supply module that is not manageable • Non-redundant cooling: Ten fixed chassis fans which are not hot swappable	• Ten fan tachometer sensors • One front panel temperature sensor	No
SR1625UR	• 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	• Ten fan tachometer sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One power unit redundancy sensor	Yes
SR1630BC	• Non-redundant power: One power supply module which is not manageable • Non-redundant cooling: Two fixed chassis fans which are not hot swappable	• Two fan tachometer sensors • One front panel temperature sensor	No
SR2600URBRP	• Redundant power: Two PMBus enabled power supply modules which are manageable • Power Distribution Board (PDB) • Non-redundant cooling: Three fixed chassis fans which are not hot swappable • Chassis Intrusion	• Three fan tachometer sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One physical security sensor: Chassis intrusion offset	Yes
SR2600URLX	• Redundant power: Two PMBus enabled power supply modules which are manageable • Power Distribution Board (PDB) • Redundant cooling: Six hot swappable chassis fans • Chassis Intrusion	• Six fan tachometer sensors • Six fan presence sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One physical security sensor: Chassis intrusion offset	Yes

Intel® Server Chassis	Chassis-specific Server Management Features	Chassis-dependent Sensors Managed by Management Controller	Intel Policy-Based Node Manager Support
SC5600BASE	• Non- redundant power: One power supply module which is not manageable • Non-redundant cooling: Five fixed chassis fans which are not hot swappable • Chassis Intrusion	• Five fan tachometer sensors • Two fan presence sensors • One front panel temperature sensor • One physical security sensor: Chassis intrusion offset	No
SC5600BRP	• Redundant power: Two PMBus enabled power supply modules which are manageable • Power Distribution Board (PDB) • Non-redundant cooling: Five fixed chassis fans which are not hot swappable • Chassis Intrusion	• Five fan tachometer sensors • Two fan presence sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One physical security sensor: Chassis intrusion offset	Yes
SC5600LX	• Redundant power: Two PMBus enabled power supply modules which are manageable • Power Distribution Board (PDB) • Redundant cooling: Four hot swappable chassis fans • Chassis Intrusion	• Four fan tachometer sensors • Four fan presence sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One physical security sensor: Chassis intrusion offset	Yes
SC5650BRP	• Redundant power: Two PMBus enabled power supply modules which are manageable • Power Distribution Board (PDB) • Non-redundant cooling: 5 fixed chassis fans which are not hot swappable. • Chassis intrusion	• Five fan tachometer sensors • One front panel temperature sensor • PMBus sensors • Two input power sensors • Two output current sensors • Two power supply temperature sensors • Two power supply status sensors • PDB failure monitoring reflected in the power unit sensor 'failure' offset • One power unit redundancy sensor • One physical security sensor: Chassis intrusion offset	Yes

Table 1 List of Chassis with hardware combination

4. PMBus and Power Supply sensors to enable NM support

The Power Management Bus ("PMBus") is an open standard protocol that defines a means of communicating with power conversion and other devices.
For more information, please see the System Management Interface Forum Web site: www.powerSIG.org.

To support NM capabilities the server should have the power supply device and power supply unit compliant with PMBus specification.

5. Node Manager FAQs for finer control of the server system

1. **I have created and enabled the policy. Why is it not effective?**

There could be multiple reasons for having observed this behavior. Once a policy has been set, it should be effective and should limit the power consumption of the system within the set limit. But for the policy to be really effective following things are necessary –

a. Node Manager Policy should be **enabled**
b. PM Bus power supplies with updated PDB firmware (see the tables for finding compatible power supplies for given server in the table above)
c. The system should be sufficiently loaded to observe the effective/optimum behavior (cpu usage should be > 25% approximately)
d. The power limit set should be greater than the idle power consumption of the system.

Even if all these conditions are met, sometimes you may not see the power consumption of the system coming down to the set power limit of policy. This happens because of two things

e. The system is working at a very high load and the power limit set is very low, so even by using the least P State, Node Manager is not able to bring the power value to the set power limit.
f. T states in the system are not enabled due to which NM is not able to make use of T states and bring the power further down.

2. How do you enable T States?

The example here is on Windows 2008 Enterprise edition. T States work after enabling special power policy (quoting from "Processor Power Management in Windows Vista and Windows Server Longhorn" by Microsoft):

a). Open an elevated command prompt.
b). View the "Allow Throttle States processor power policy" by using the following command:

powercfg -qh scheme_current sub_processor
Power Scheme GUID: 381b4222-f694-41f0-9685-ff5bb260df2e (Balanced)
Subgroup GUID: 54533251-82be-4824-96c1-47b60b740d00 (Processor power management)
Power Setting GUID: 3b04d4fd-1cc7-4f23-ab1c-d1337819c4bb (Allow Throttle States)
Possible Setting Index: 000

Possible Setting Friendly Name: Off
Possible Setting Index: 001
Possible Setting Friendly Name: On
Current AC Power Setting Index: 0x00000000
Current DC Power Setting Index: 0x00000000

c) Copy the globally unique identifier (GUID) for the "Allow Throttle States" processor power policy.
d) Set the "Allow Throttle States" policy value to 1 by using the following command:

powercfg -setacvalueindex scheme_current sub_processor 3b04d4fd-1cc7-4f23-ab1c-d1337819c4bb 1

3. Power Supply Issue

NM Policies work only if PMBus compatible power supplies are connected. Check Table 2 for details.

With a single PSU in location PS1, NM works properly.
NM will report wattage and set policies to limit wattage correctly.

With a single PSU in location PS2, NM will report wattage as 0 and NOT set policies.

4. Why does NM not limit the power in when the chassis of the server is open?

There are two reasons why NM won't limit the power consumption in an open chassis condition:
a) You have a cooling efficiency problem in the closed box and CPU was throttled
b) You have a cooling efficiency problem in the open box and fans start running faster (and consume more power). As NM could not limit the power, the power consumption of the CPU went up. You will observe that the fans are running faster.

6. What are the PMBus compliant supplies for different Servers?

The following table lists the power supply device compatible with the chassis:

Chassis	PS or PDB?	Model Number (on decal)	Part Number (on decal)	Revision (on decal)	Product Name (in product area of the FRU)
SR1600UR	PS	TDPS-600EB A	E30381-XXX	All	TDPS-600EB A
SR1625UR	PDB	AC-074 A	E33447-XXX	All	AC-074 A
	PS	DPS-650QB A	E33446-XXX	S0, S1, S2	DPS-650QBA
	PS	DPS-650QB A	E33446-XXX	S3 & Later	DPS-650QB A
SR2600UR	PDB	AC-075 A	E30692-XXX	All	AC-075 A
	PS	DPS-750PB A	E30694-XXX	S0, S1	DPS-750PBA
	PS	DPS-750PB A	E30694-XXX	S2 & Later	DPS-750PB A
SC5650DP	PS	DPS-600MB Y	E35746-XXX	All	DPS-600MB Y
SC5650BRP	PS	DPS-600SB A	E35862-XXX	All	DPS-600SB A
	PDB	RPS-600-5 A	E35772-XXX	All	RPS-600-5 A
SC5650WS	PS	DPS-1000HB A	E37586-XXX	All	DPS-1000HB A
SC5600BASE	PS	DPS-670DB F	E35756-XXX	All	DPS-670DB F
SC5600BRP	PS	DPS-750PB A	E30692-XXX	All	DPS-750PB A
	PDB	AC-078 A	E33995-XXX	All	AC-078 A
SC5600LX	PS	DPS-750PB A	E30692-XXX	All	DPS-750PB A
	PDB	AC-078 A	E33995-XXX	All	AC-078 A
SC5650DP	PS	DPS-600MB Y	E35746-XXX	S1F, 00F	DPS-600MB Y
SC5650BRP	PS	DPS-600SB A	E35862-XXX	All	DPS-600SB A
	PDB	RPS-600-5 A	E35772-XXX	All	RPS-600-5 A
SR1630BC	PS	TDPS-400CB A	E31614-XXX	S1/00F	TDPS-400CB A

Table 2. List of compatible power supplies

6. Bibliography

Robert Frances Group, January 2006