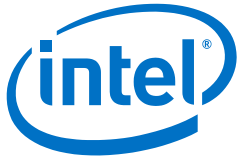
Intel® Node Manager

Intel® DCM



Energy/Power Management capabilities

Ram Nagappan

October 2013

Legal Disclaimers

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo

Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/

Legal Disclaimers: Performance

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, Go to: http://www.intel.com/performance/resources/benchmark_limitations.htm.

Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.

Relative performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.

SPEC, SPECint, SPECfp, SPECrate. SPECpower, SPECjAppServer, SPECjEnterprise, SPECjbb, SPECompM, SPECompL, and SPEC MPI are trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org for more information.

TPC Benchmark is a trademark of the Transaction Processing Council. See http://www.tpc.org for more information.

SAP and SAP NetWeaver are the registered trademarks of SAP AG in Germany and in several other countries. See http://www.sap.com/benchmark for more information.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/software/products.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the third quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "believes," "seeks," "estimates," "may," "will," "should" and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be the important factors that could cause actual results to differ materially from the company's expectations. Demand could be different from Intel's expectations due to factors including changes in business and economic conditions; customer acceptance of Intel's and competitors' products; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Uncertainty in global economic and financial conditions poses a risk that consumers and businesses may defer purchases in response to negative financial events, which could negatively affect product demand and other related matters. Intel operates in intensely competitive industries that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross margin percentage are affected by the timing of Intel product introductions and the demand for and market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductions, marketing programs and pricing pressures and Intel's response to such actions; and Intel's ability to respond quickly to technological developments and to incorporate new features into its products. The gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; start-up costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; product manufacturing quality/yields; and impairments of long-lived assets, including manufacturing, assembly/test and intangible assets. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Expenses, particularly certain marketing and compensation expenses, as well as restructuring and asset impairment charges, vary depending on the level of demand for Intel's products and the level of revenue and profits. Intel's results could be affected by the timing of closing of acquisitions and divestitures. Intel's results could be affected by adverse effects associated with product defects and errata (deviations from published specifications), and by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues, such as the litigation and regulatory matters described in Intel's SEC reports. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.

Agenda

Energy/Power Management capabilities in:

Intel® Xeon Phi Coprocessor

Intel® Xeon Processor E5 V2 Family

Intel® Node Manager

Intel® DCM

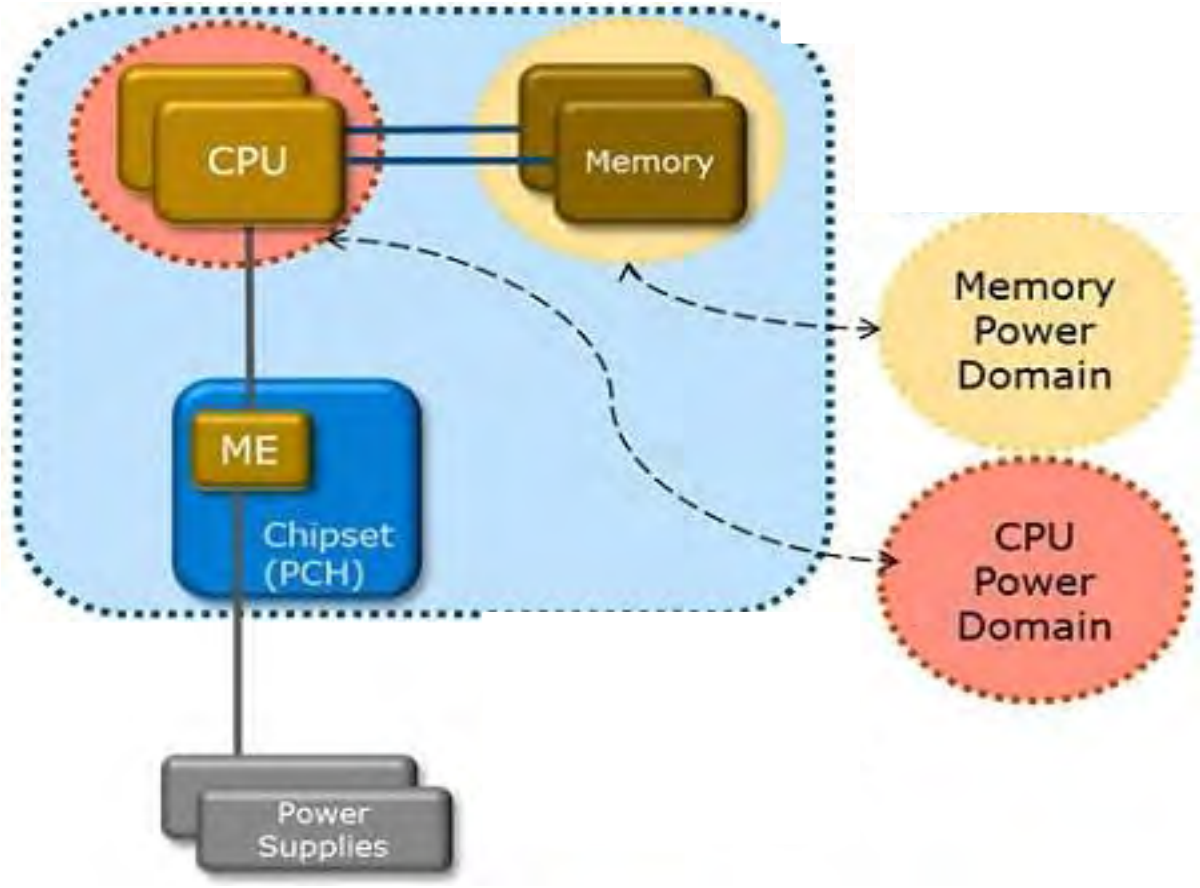
Summary

Intel® Energy/Power Management Capabilities from Component to System

Energy/Power Management capabilities from Component to System

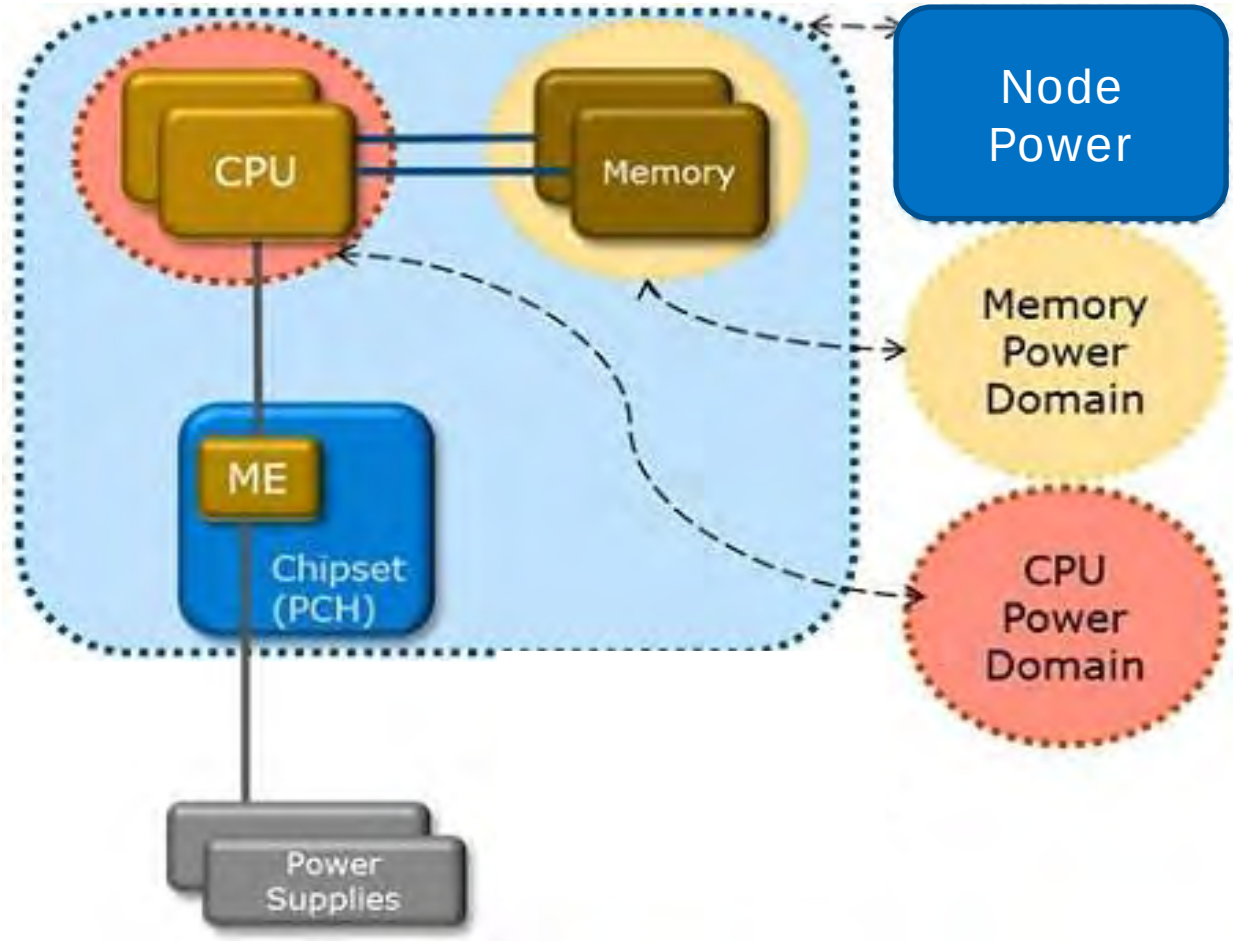
- Component
  - RAPL (Running Average Power Limit) to Monitor and Limit – CPU and Memory Power
- Node
  - Intel Node Manager to Monitor & Limit Node Power
- System/Cabinet
  - Intel Data Center Manager (DCM) to monitor and limit power at System/Cabinet

Intel Xeon Component Level Energy Measurement



RAPL (Running Average Power Limit) to Monitor and Limit – CPU and Memory Power

Intel Node Manager



Node Level Power Monitoring & Limiting Capability using Intel Node Manager

Intel Data Center Manager (DCM)

SDK with Web Service APIs for Data Center Power and Thermal Power Management

Management Console

Easy integration in the Management Console

Intel® DCM SDK (Web Service API)

- Monitor
- Control
- Trend
- Scalability
- Standards

Servers

System/Cabinet power monitoring and limiting using Intel DCM

Intel® Xeon Phi Coprocessor

Intel® Xeon Phi™ Coprocessor Product Lineup

3 Family
Outstanding Parallel Computing Solution
Performance/$ leadership

6GB GDDR5
240GB/s
>1TF DP

3120P

3120A

5 Family
Optimized for High Density Environments
Performance/watt leadership

8GB GDDR5
>300GB/s
>1TF DP
225-245W

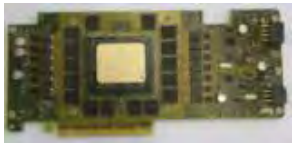
5110P

5120D

7 Family
Highest Performance, Most Memory
Performance leadership

16GB GDDR5
352GB/s
>1.2TF DP

7120P



7120X

Intel Xeon Phi Coprocessor Power States

Coprocessor Power State	7120P/3120P Power (Watts)	5110P Power (Watts)
C0	300	225
C1	<115	<115
PC3	<50	<45
PC6	<30	

P States & Turbo Mode

- All cores run at the same frequency
- Intel® Xeon Phi™ coprocessor 7120 supports Turbo Mode
  - When the card is operating below its specified power and temperature limits, it will enter turbo while still remaining within the power and thermal specifications.

Manageability

- Intel® Xeon Phi™ coprocessor manageability relies on a System Management Controller (SMC) on the PCI Express* card.
- The system provides sensor telemetry information for management by in-band (host) software and out-of-band software via the PCI Express* SMBus.
- SMC monitors power and temperatures within the Intel® Xeon Phi™ coprocessor and through sensors located on the PCI Express* card

Measurements: Components

(Info) Components are the physically discrete units that comprise the node. This level of measurement is important to analyze application energy performance trade-offs. This level is analogous to performance counters and carries many of the same motivations. Components may not only be silicon devices. For example, it would be useful to know how much fan energy is being used by the Muffin fans at the back of the rack or by some active rear door cooling methodology. Also, some systems may have a CDU. How much energy is being used by the CDU for motors, fans.

(enhancing) The ability to measure the current and voltage of each individual component must be provided.

The measurement sampling frequency should be:
(mandatory) 10 samples per second
(important) 100 samples per second
(enhancing) 1000 samples per second

Slide from
Vendor Forum
12 September 2013

(mandatory) The current and voltage data shall be both real electrical measurements and based on heuristic models.

Measurement Capabilities

- Power/energy data is sampled every 50ms
- Two user programmable power threshold levels
  - PL0
    - Defaults to 125% of TDP
  - PL1
    - Defaults to 105% of TDP

Intel® Xeon Processor E5 V2 family

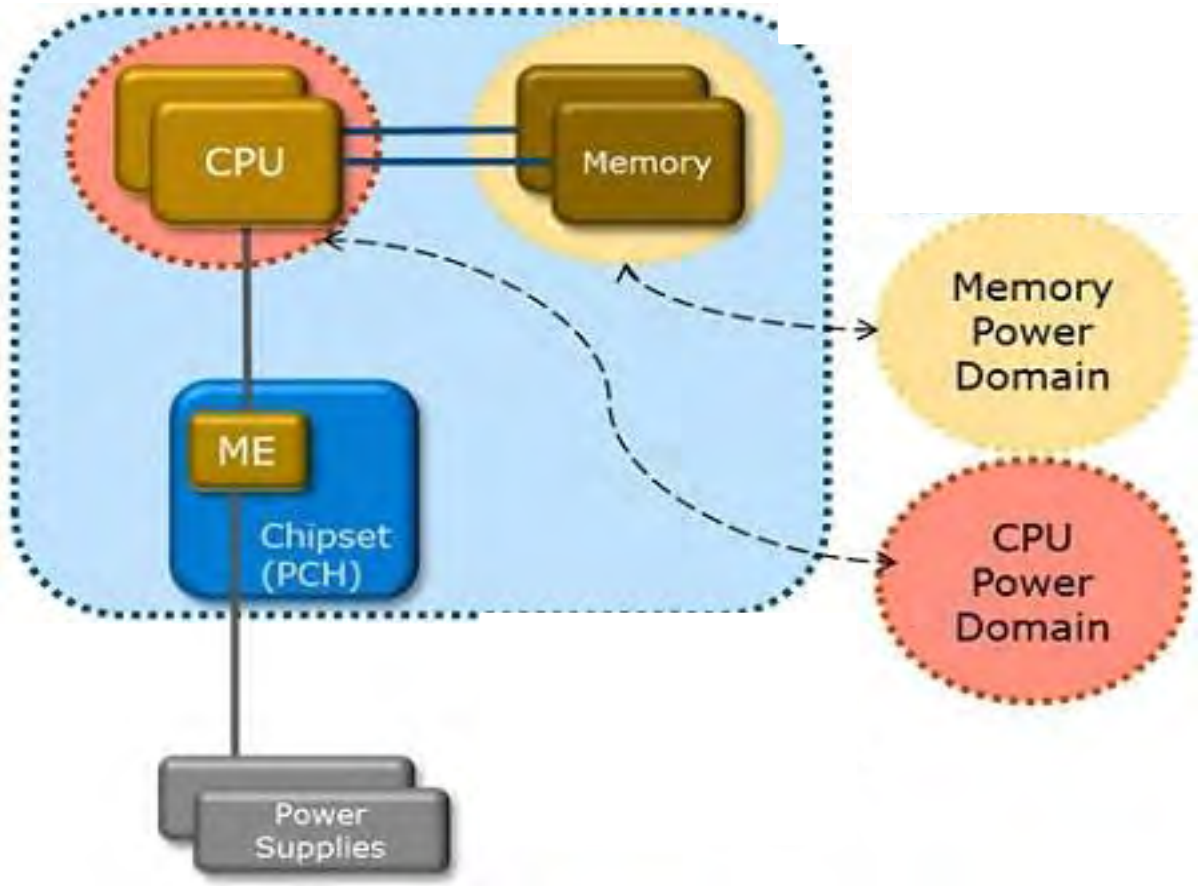
Intel® Xeon® Processor E5-2600 v2 Product Family

Today's focus

Feature	Intel® Xeon® E5-2600 (Sandy Bridge-EP)	Intel® Xeon® E5-2600 v2 (Ivy Bridge-EP)
QPI Speed (GT/s)	8.0, 7.2 and 6.4	
Addressability	46 bits physical, 48 bits virtual	
Cores	Up to 8	**Up to 12**
Threads Per Socket	Up to 16 threads	**Up to 24 threads**
Last-level Cache (LLC)	Up to 20 MB	**Up to 30 MB**
Intel® Turbo Boost Technology[1]	Yes	
Memory Population	4 channels of up to 3 RDIMMs, 3 LRDIMMs or 2 UDIMMs	
Max Memory Speed	Up to 1600	**Up to 1866**
Memory RAS	ECC, Patrol Scrubbing, Demand Scrubbing, Sparing, Mirroring, Lockstep Mode, x4/x8 SDDC	
PCIe* Lanes / Controllers/Speed (GT/s)	40 / 10 (PCIe* 3.0 at 8 GT/s)	
TDP (W)	150 (Workstation only), 130, 115, 95, 80, 70, 60, 50	
Idle Power Targets (W)	15W or higher, 12W for LV SKUs	**10.5W or higher, 7.5W for LV SKUs**

Running Average Power Limit (RAPL)

- RAPL
  - Ability to monitor and Limit CPU and DRAM power



Measurements: Components

(Info) Components are the physically discrete units that comprise the node. This level of measurement is important to analyze application energy performance trade-offs. This level is analogous to performance counters and carries many of the same motivations. Components may not only be silicon devices. For example, it would be useful to know how much fan energy is being used by the Muffin fans at the back of the rack or by some active rear door cooling methodology. Also, some systems may have a CDU. How much energy is being used by the CDU for motors, fans.

(enhancing) The ability to measure the current and voltage of each individual component must be provided.

The measurement sampling frequency should be:
(mandatory) 10 samples per second
(important) 100 samples per second
(enhancing) 1000 samples per second

Slide from
Vendor Forum
12 September 2013

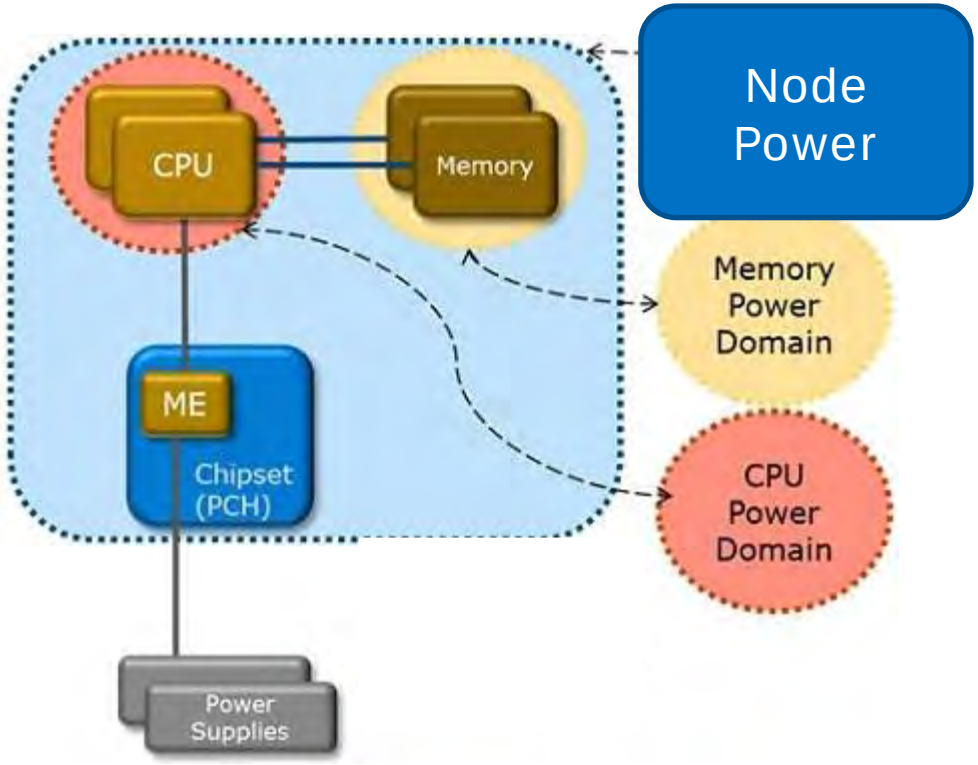
(mandatory) The current and voltage data shall be both real electrical measurements and based on heuristic models.

Measurement Capabilities

- RAPL counters are updated approximately every millisecond
- In Intel's experience the accuracy of the data is good around TDP (Thermal Design Power) levels, and may be 10% off at lower power load levels.
- Intel® Power Governor (power_gov) or Intel® Performance Counter Monitor utility can be used to (a) monitor power and (b) limit power

Intel® Node Manager

Intel Node Manager



- Intel Node Manager periodically queries the Power Supply Unit (PSU), CPU and memory about power consumption.
- It can also query about inlet temperature.
- Based on this data, Intel Node Manager calculates various statistics.
- Intel Node Manager can simultaneously monitor total system power, CPU power and memory power.

Node Manager Features

Platform Level Power Monitoring - monitoring and statistics (min, max, avg) of the total board power. Intel Node Manager reads this either directly from PSU or HSC using the PMBus protocol (PMBus v1.2 preferred) or from an external BMC using IPMI.

Platform Level Power limiting - ability enforce a policy to limit the total board power consumption.

Processor subsystem power monitoring – monitoring and statistics (min, max, avg) of the processor subsystem power

Processor power limiting - ability to enforce a policy to limit the power consumption of the processor package (socket) subsystem

Memory subsystem power monitoring – monitoring and statistics (min, max, avg) on the memory subsystem power

Memory power limiting - ability to enforce a policy to limit the power consumption of the memory subsystem

Inlet air temperature monitoring - monitoring and statistics (min, max, avg) of the inlet air temperature

Measurements: Nodes

(Info) A node level measurement shall consist of a combined measurement of all components that make up a node in the architecture. For example, these components may include the CPU, memory and the network interface. If the node contains other components such as spinning or solid state disks they shall also be included in this combined measurement. The utility of the node level measurement is to facilitate measurement of the power or energy profile of a single application. The *node* may be part of the network or storage equipment, such as network switches, disk shelves and disk controllers.

(important) The ability to measure the current and voltage of any and all nodes shall be provided.

The current and voltage measurements shall provide a readout capability of:

(mandatory) ≥ 1 per second
(important) ≥50 per second
(enhancing) ≥ 250 per second

Slide from
Vendor Forum
12 September 2013

(mandatory) The current and voltage data must be real electrical measurements, not based on heuristic models.

Node Manager Sampling

- How often power monitoring value is updated?
  - Sampled every 100ms
- What is the accuracy?
  - It is based on PMBUS power supply specification
  - Around +/- 5%

Intel® DCM

Intel® Data Center Manager (DCM)

A middleware with web service APIs for data center power and thermal management – easy to integrate in the Management Console

ISV Management Console

DCM Middleware (Web Service API)

MONITOR | CONTROL | TREND

SCALABILITY | STANDARDS

Hardware Protocols

Node Manager IPMI | iDRAC IPMI | iLO/DCMI IPMI | IMM IPMI | CMC HTTPS/WS-MAN | OA SSH/CLI | IMM SSH/CLI | SNMP

Rack Servers

Blade Servers

PDU and UPS

IPMI = Intelligent Platform Management Interface
IMM = Integrated Management Module
SNMP = Simple Network Management Protocol
WS-MAN = Web Services-Management
iDRAC = Integrated Dell Remote Access Controller
CMC = chassis management controller
CLI = command line interface
DCMI = Data Center Manageability Interface
iLO = Integrated Lights-out
OA = Onboard Administrator
SSH = Secure Shell

Intel DCM Features

- Monitor Power and Thermals – Aggregated actual and historical trend data and alerts for racks and groups of servers
- Policy-based Management – Intelligent heuristics engine maintains group power cap on demand
- Scalability – Manage 10000s nodes using agentless technology.
- Retrieve real-time PDU and UPS power data

Measurements: System, Platform and Cabinet

(mandatory) Shall be able to measure the current and voltage of the system, platform(s) and cabinet(s).

The current and voltage measurements shall provide a readout capability of
(mandatory) ≥ 1 per second
(important) ≥ 50 per second
(enhancing) ≥ 250 per second

Slide from
Vendor Forum
12 September 2013

(mandatory) The current and voltage data shall be real electrical measurements, not based on heuristic models

(important) The vendor shall assist in the effort to collect these data in whatever other subsystems are provided (e.g., another vendor's back-end storage system).

(important) Those elements of the system, platform and cabinet that perform infrastructure-type functions (e.g., cooling and power distribution), shall be measured separately with the ability to isolate their contribution to the energy and power measurements.

DCM Sampling

- CPU RAPL updates data every ~1ms
- NM updates data every ~100ms
- DCM updates aggregated data every few seconds
  - Can go faster if needed

Summary

Summary

- Intel delivers complete power management solution for HPC systems
  - Component
    - RAPL - CPU and Memory Power
  - Node
    - Intel Node Manager
  - System/Cabinet
    - Intel Data Center Manager