Intel technologies, tools and techniques for power and energy efficiency analysis

Andrey Semin
Sr. Staff Engineer
HPC Technology Manager, EMEA

International Conference on Energy-Aware High Performance Computing (EnA-HPC)
September 1st - 2nd 2014, Dresden, Germany

intel

Notices

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document 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.

This document contains information on products in the design phase of development.

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

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as STREAM , NPB, NAMD and Linpack, 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.

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.

Intel, Xeon and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

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

Optimization Notice

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

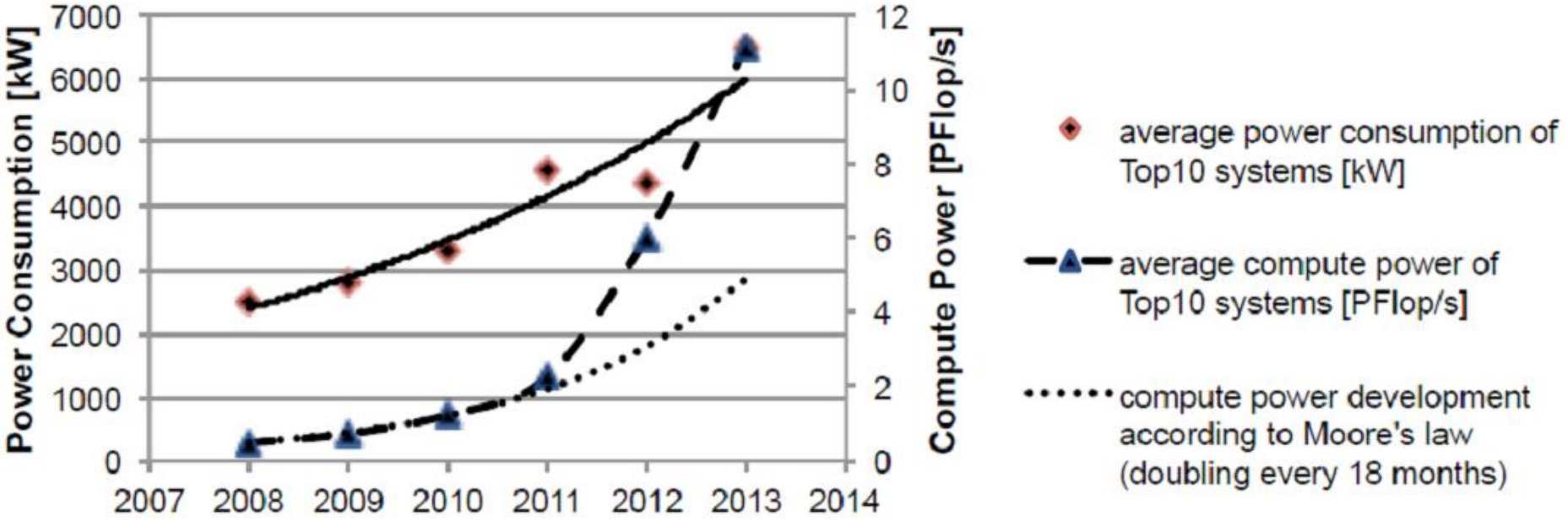
Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Agenda

- Power and energy efficiency metrics for datacentres
- Intel tools and techniques for server power monitoring
- Energy monitoring with Intel tools example

Average power and performance of Top10 supercomputers



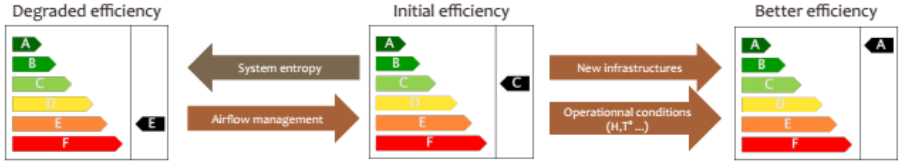
Performance and power of large HPC systems grows faster than Moore's law

Source: Top500 list http://www.top500.org, Courtesy Shailen Sobhee.

Power and energy efficiency

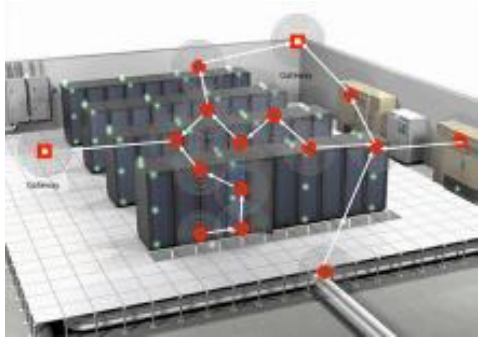
What is power/energy efficiency of application and datacentre?

- PUE, Green500, etc.



Why care?

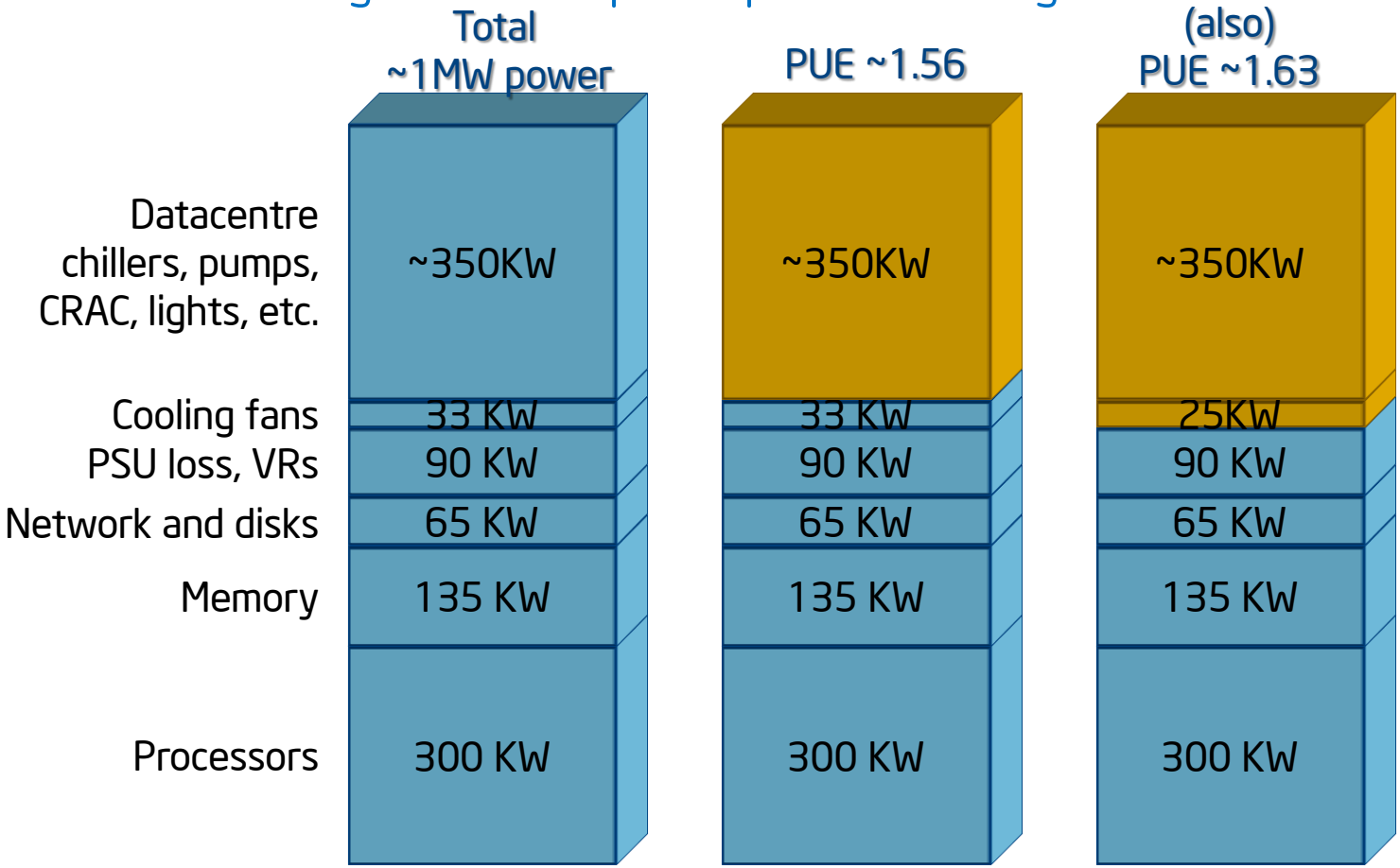
- Can measure => can improve



How to routinely measure?

- Need tools, techniques, etc.

Understanding datacentre power profile and usage effectiveness metrics: PUE



Source: own estimates for 1300 node HPC cluster in 2013. See backup for more details.

Improved approach: introduce ITUE and TUE

Dr. Michael Patterson (Intel) and EEHPCWG Propose a New Metric
ISC13 Gauss Award Winner – Most Outstanding Paper: TUE, a new energy-efficiency metric applied at ORNL's Jaguar

1. **Introduce ITUE:** ITUE is a "PUE-type" metric for the IT equipment
2. **Introduce TUE** - a proposed new metric that uses ITUE and PUE

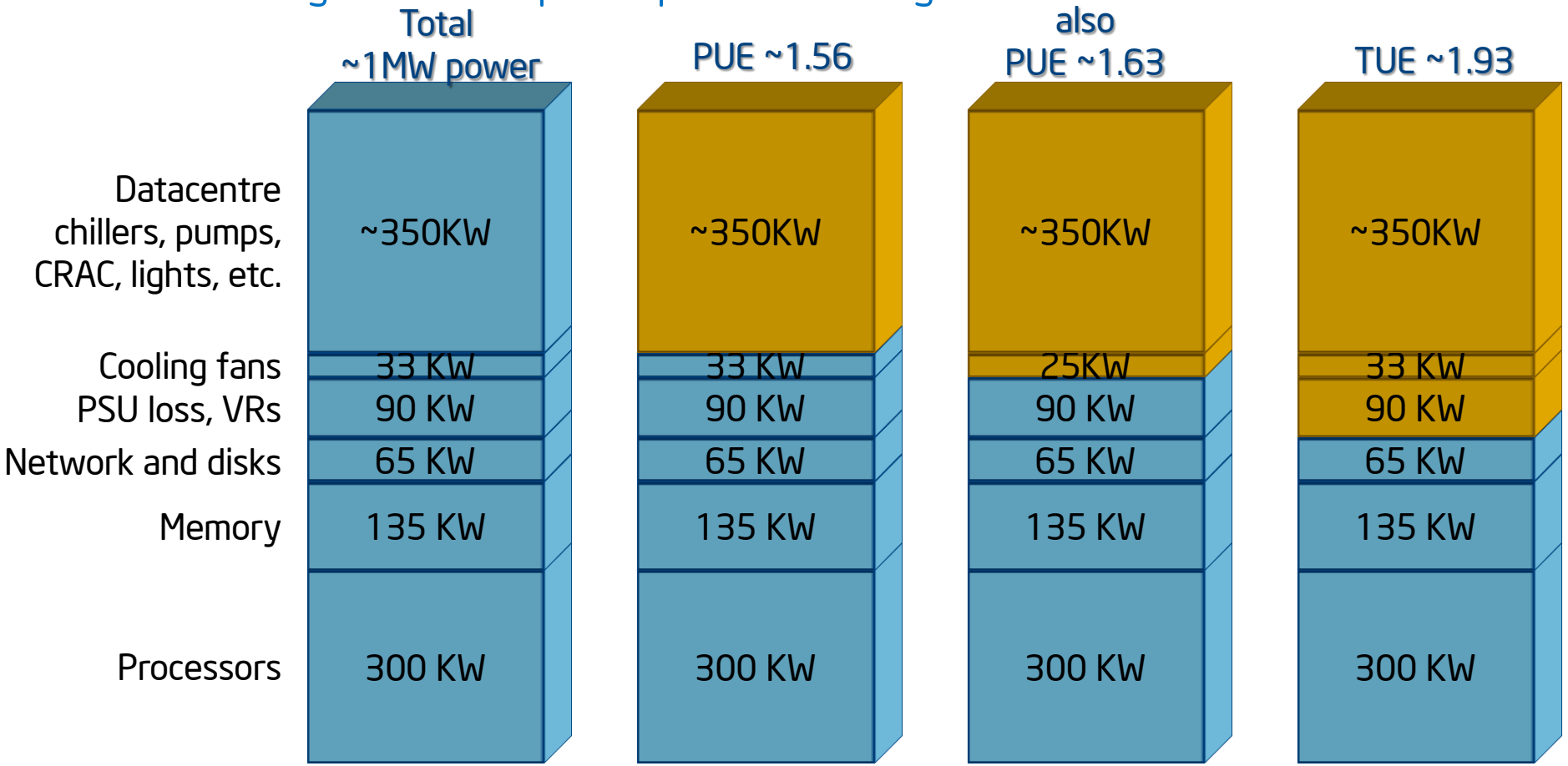
TUE is a calculated value:

$$TUE = PUE \times ITUE$$

$$ITUE = \frac{Total\ Energy\ into\ the\ IT\ equipment\ (productive + infrastructure)}{Total\ Energy\ into\ the\ Compute\ Components\ (productive)}$$

"*Productive*" components = CPUs, Memory, Storage, Networking
"*Infrastructure*" = Power Supplies, Voltage Regulators, Fans

Understanding datacentre power profile and usage effectiveness metrics: TUE



Source: own estimates for 1300 node HPC cluster in 2013. See backup for more details.

Tools and techniques to measure ITUE

$$ITUE = \frac{Total\ Energy\ into\ the\ IT\ equipment\ (productive + infrastructure)}{Total\ Energy\ into\ the\ Compute\ Components\ (productive)}$$

1. Measure Energy into Compute Components:

E.g. power into CPUs, memory, disks, network switches and integrate over time, such as

$$Energy(T) = \int_{t_0}^{t_0+T} Power(t)dt$$

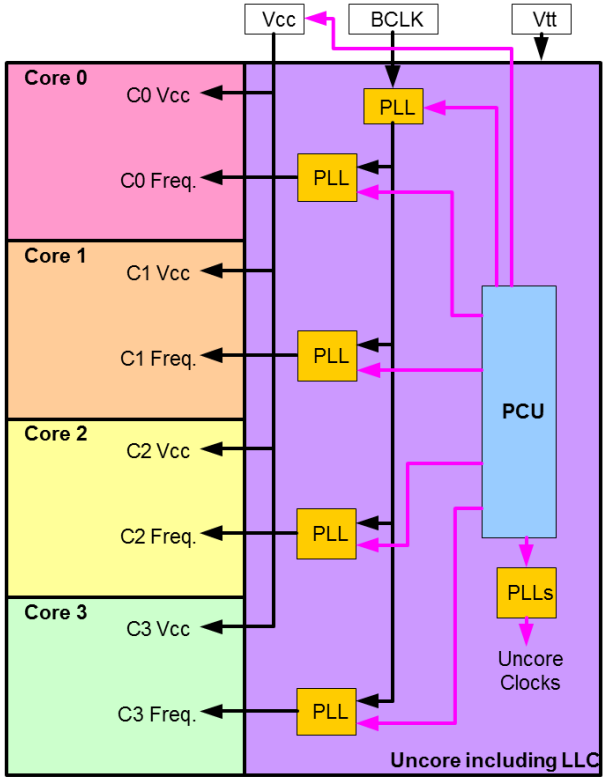
2. Measure Energy into IT Equipment:

E.g. AC power for servers, switch systems, storage arrays and integrate over

Agenda

- Power and energy efficiency metrics
- Intel tools and techniques for server power monitoring
- Energy monitoring with Intel tools example

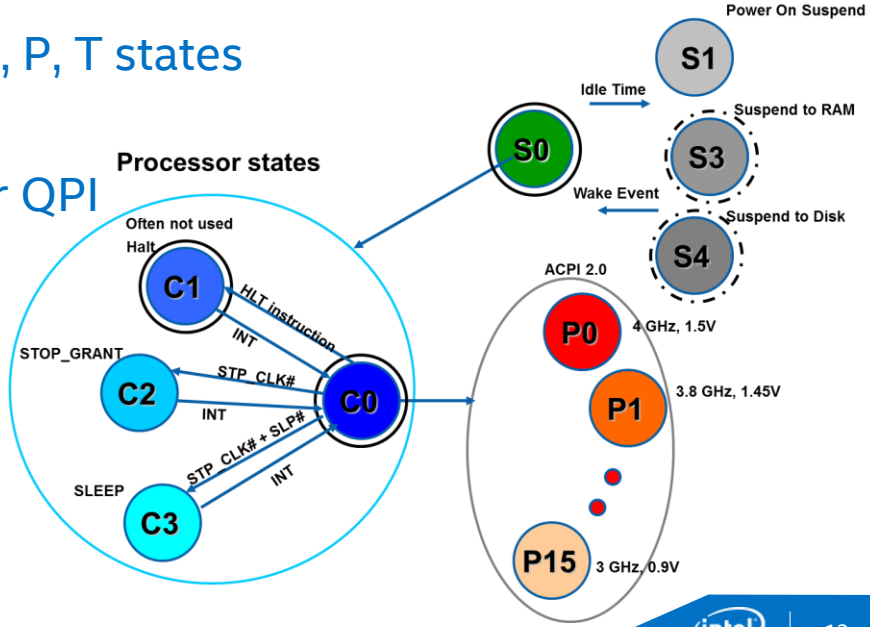
1. Measure Energy into Productive Components with Power Control Unit (PCU) inside Intel processors



Monitors voltage, temperature, power management requests

Controls

- Core/Package C, P, T states
- Memory states
- Power states for QPI



1. Measure Energy into Productive Components: the counters

PACKAGE_POWER_SKU_UNIT (606h): defines units **x** for time, energy ($\mathbf{unit} = \frac{1}{2^x}$ joule and power

CPU related RAPL counters:

- PACKAGE_ENERGY_STATUS (611h): accumulated energy by the CPU socket
- PACKAGE_POWER_SKU (614h): defines min, max and TDP power of the CPU

DRAM related RAPL counters:

- DRAM_ENERGY_STATUS (619h): to monitor DRAM power consumption
- DRAM_POWER_INFO (61Ch): defines min, max and TDP power of DRAM

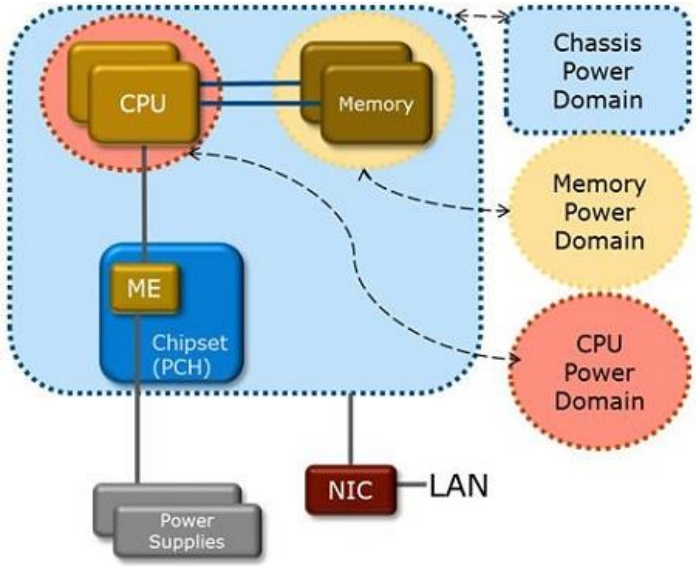
Counters values are updated ~1 ms. Require root privileges to read Model Specific Registers

Tools: Intel PCM, turbostat, perf, Likwid, Intel Vtune, and many others:

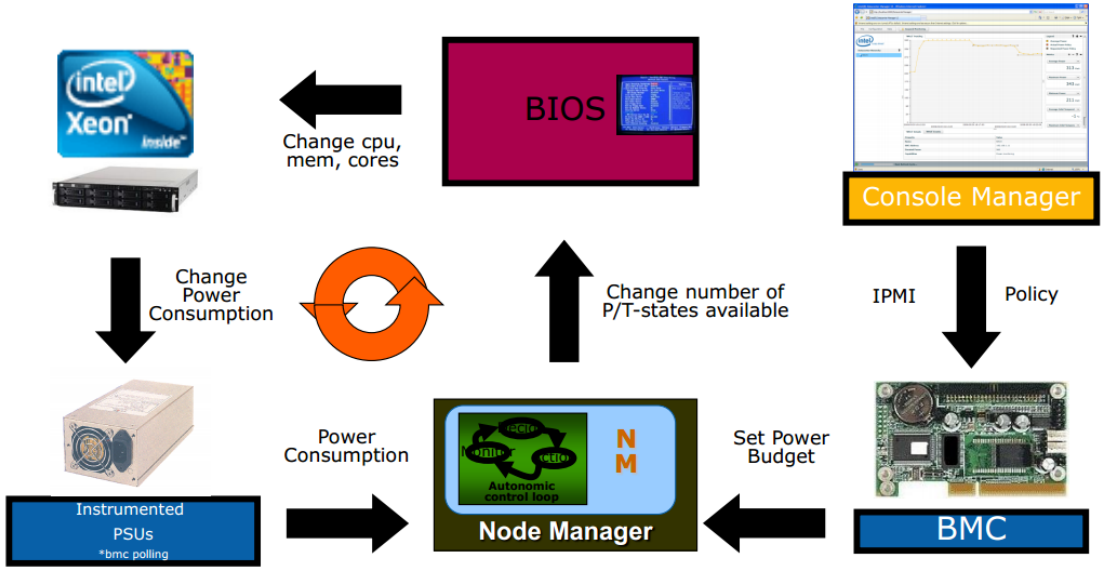
```
# unit=$(echo "ibase=16; 1/2^$(rdmsr -X 0x606 -f 12:8)" |bc -l)     // ~1.52*10^-5
# counts=`a=$(rdmsr -X 0x611); sleep 1; echo "ibase=16; $(rdmsr -X 0x611)-$a"|bc`
# echo "$counts * $unit" |bc -l                                    // 97.30
```

2. Measure Energy into IT Equipment: servers with Intel Intelligent Power Node Manager

Architecture



Operation



The most recent version added support for Intel Xeon Phi Co-processors (PCIe domain)

2. Measure Energy into IT Equipment: Tools

IPMI Command example*:

- Net function: 2Eh-2Fh, Command: C8h "Get Intel NM Statistics" (see also "Reset NM Statistics")

```
#!/bin/sh
RET=$(ipmitool -t 0x2c -b 6 raw 0x2e 0xc8 0x57 0x01 0x00 0x01 0x00 0x00)
# RET =  "57 01 00 5e 00 53 00 45 02 93 00 f2 c8 01 54 eb 47 02 00 50"
if [ "$(echo $RET | cut -c1-8)" = "57 01 00" ]
then
 CUR=`echo $RET |awk '{print "ibase=16;",toupper($4),"+",toupper($5),"*FF"}' | bc`
 MIN=`echo $RET |awk '{print "ibase=16;",toupper($6),"+",toupper($7),"*FF"}' | bc`
 MAX=`echo $RET |awk '{print "ibase=16;",toupper($8),"+",toupper($9),"*FF"}' | bc`
 AVG=`echo $RET |awk '{print "ibase=16;",toupper($10),"+",toupper($11),"*FF"}'| bc`
 echo "CUR:$CUR MIN:$MIN MAX:$MAX AVG:$AVG "       # CUR:94 MIN:83 MAX:579 AVG:147
else
 echo "CUR:NAN MIN:NAN MAX:NAN AVG:NAN"
fi
```

Tools:

- Intel NodeManager Reference Kit (NMRK), Intel DCM, FreeIPMI, may OEM tools (Cisco EnergyWise Suite, Dell OpenManage Power Center, IBM Tivoli Monitoring for Energy Management, etc).

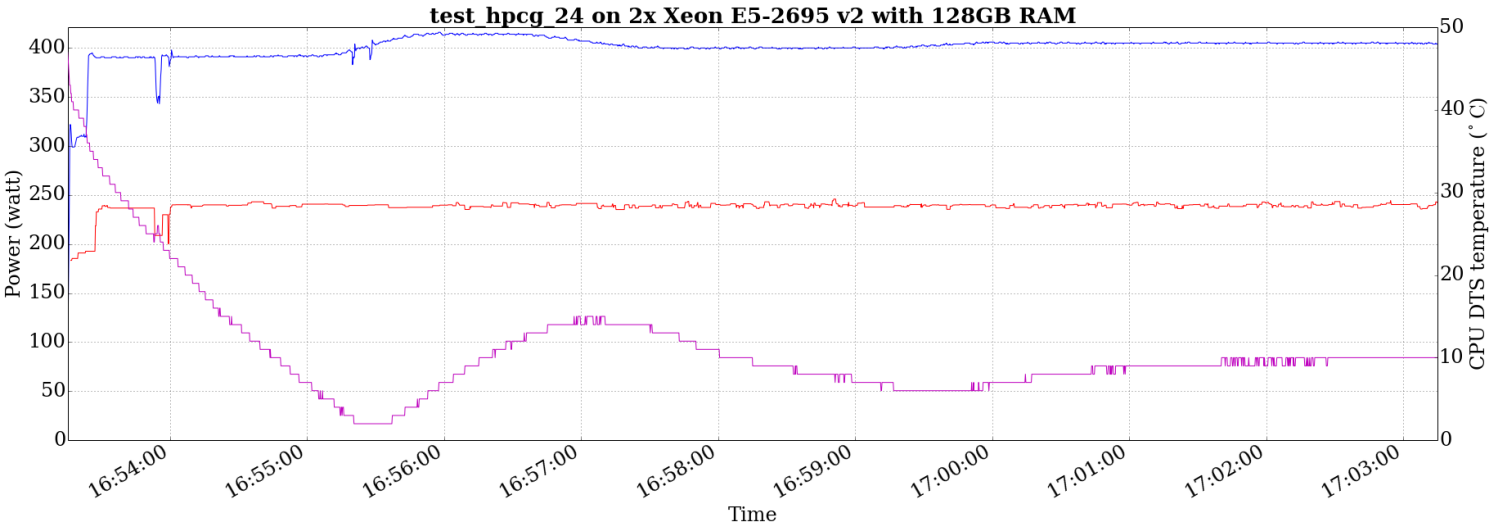
* Source: Intel® Intelligent Power Node Manager 2.0: Specification, Intel DocID: 322999, August 2013

Agenda

- Power and energy efficiency metrics
- Intel tools and techniques for server power monitoring
- Energy monitoring with Intel tools example

HPCG: understanding system power behaviour

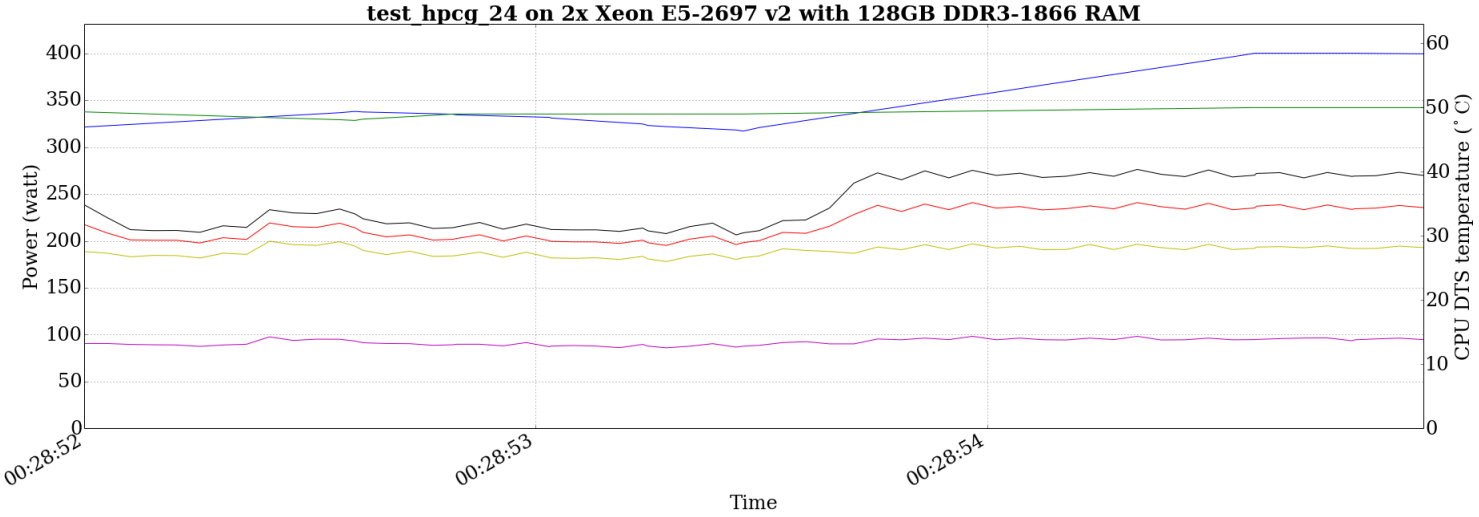
1U chassis, 2c 12 cores, 115W TDP, 128GB RAM DDR3-1600 LV



Quiz: what was happening there?

HPCG: improved system

4U chassis, 2x 12 cores, 130W TDP, 128GB RAM DDR3-1866



Energy and power metering delivered using built-in technologies in Intel platforms

Summary

- Efficiency optimization: new and better metrics should be adopted and used
- Intel delivers technologies for power and energy monitoring and optimization
- Open source and commercial tools are available for community to use RAPL and Intel Power NodeManager

... and one more thing ...

Meet the book of the year… :)

Apress open

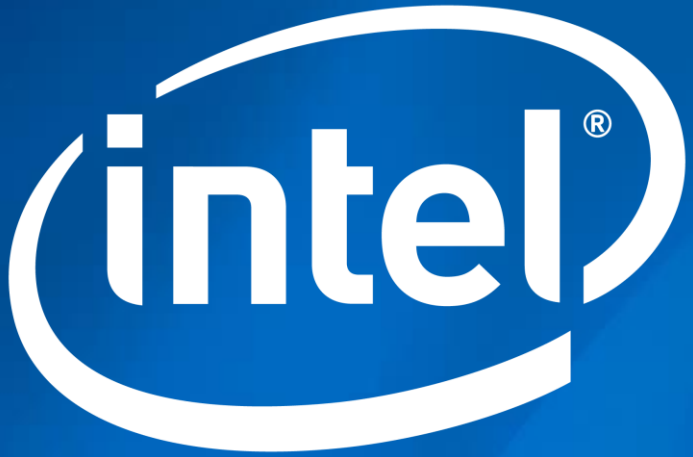
Table of Contents:

Selected For intel Intel's Recommended Reading List

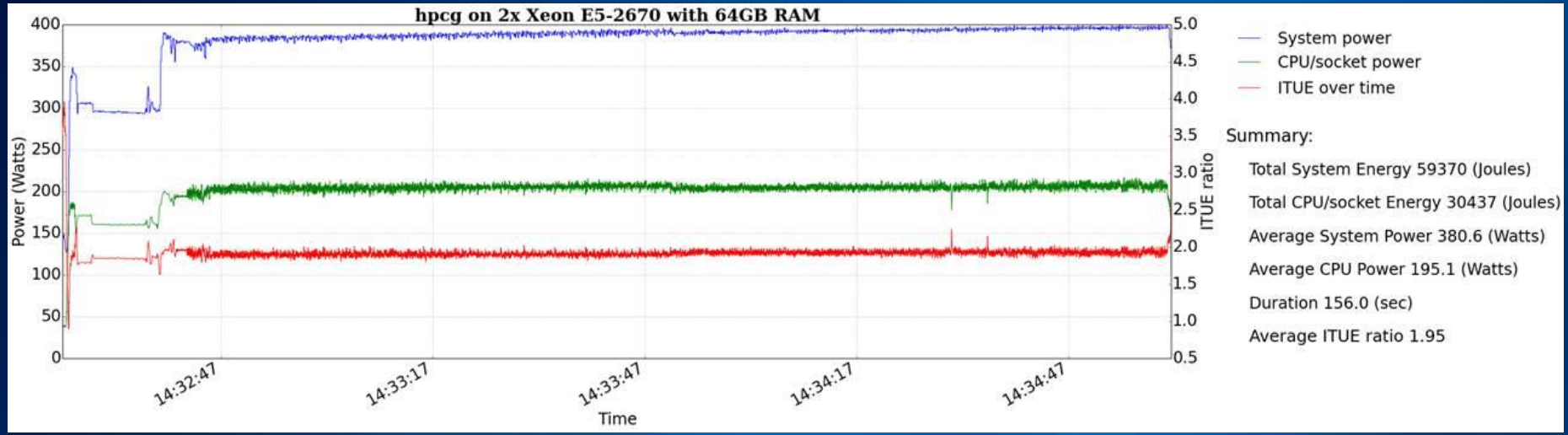
ISBN-13 (pbk): 978-1-4302-6496-5
ISBN-13 (electronic): 978-1-4302-6497-2

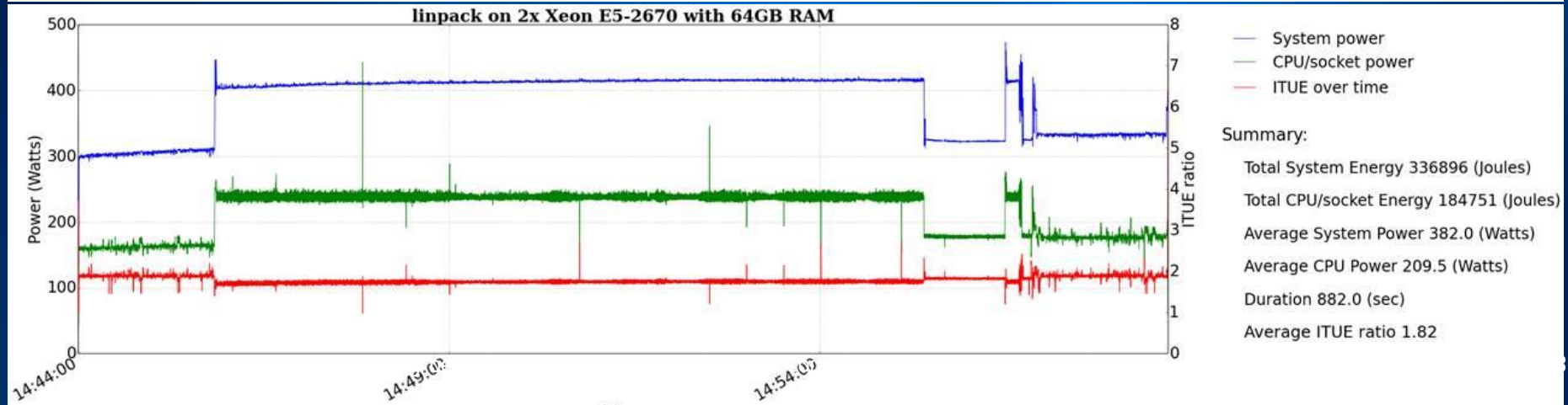
Authors: A.Supalov, A.Semin, M.Klemm,C.Dahnken

Order now at http://www.apress.com/9781430264965


intel

Observing ITUE: using built-in technologies





Legal Disclaimer

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document 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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Intel, Intel Xeon, Intel Xeon Phi™, Intel® Atom™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

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

Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Intel® Advanced Vector Extensions (Intel® AVX)* are designed to achieve higher throughput to certain integer and floating point operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you should consult your system manufacturer for more information.

*Intel® Advanced Vector Extensions refers to Intel® AVX, Intel® AVX2 or Intel® AVX-512. For more information on Intel® Turbo Boost Technology 2.0, visit http://www.intel.com/go/turbo

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

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors.

These optimizations include SSE2®, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804