Considerations and Opportunities for Energy Efficient High-Performance Computing

From Datacenters to Applications

Andrey Semin & Herbert Cornelius

Intel Corporation

ENA-HPC 2013 Conference
Dresden, Germany
2 September 2013

Notices

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

Foreword

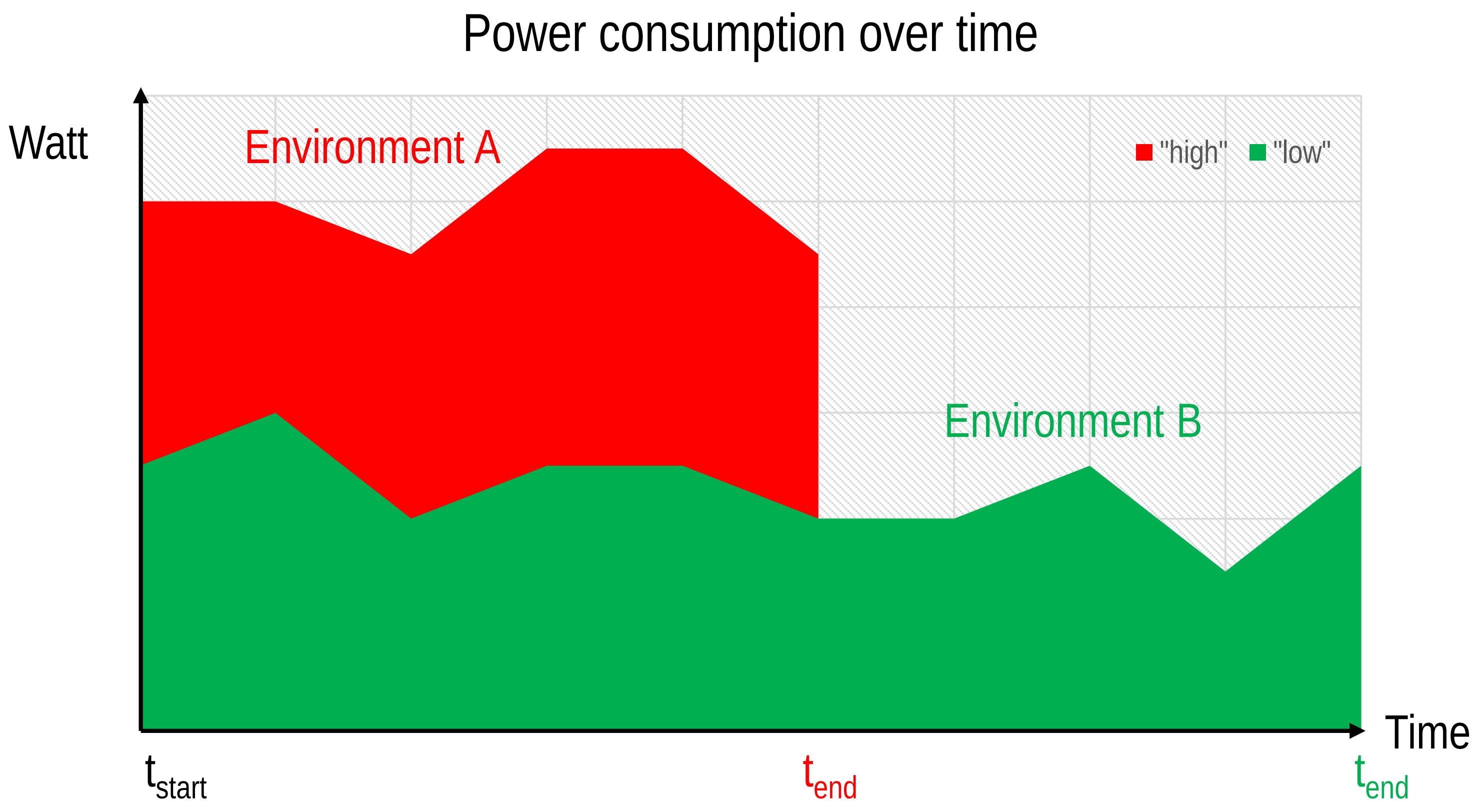
- Any views or opinions presented here are solely those of the authors and do not necessarily represent those of Intel Corporation

- Data presented in this session are results of work completed during 2007-2013 with involvement of many Intel colleagues, customers and fellow travellers

- Acknowledgements:

  - **Intel**: Mike Patterson, Victor Gamayunov, Fan (Frank) Liu, Ram Nagappan, and many others
  - **Samsung**: Peyman Blumstengel, Nicolas Rossetto
  - **Eurotech**: Giampietro Tecchiolli, Paul Arts, Mauro Rossi
  - **RSC**: Egor Druzhinin, Pavel Lavrenko, Nikita Burtsev
  - **JSCC**: Oleg Aladyshev, Pavel Telegin, Boris Shabanov

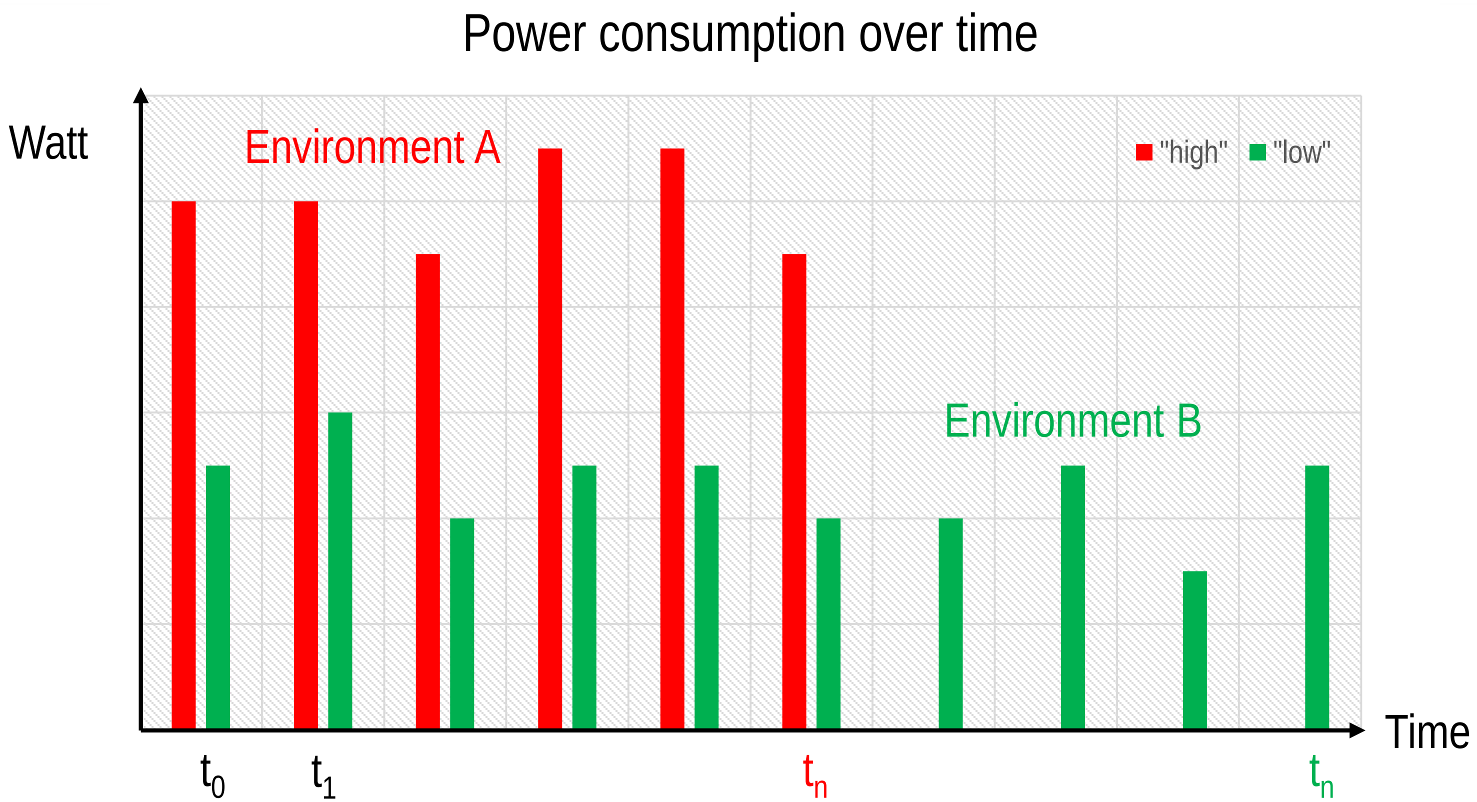
Power consumption and efficiency



$$\int_{t_{start}}^{t_{end}} power\_consumption(t)dt$$

For illustration only.

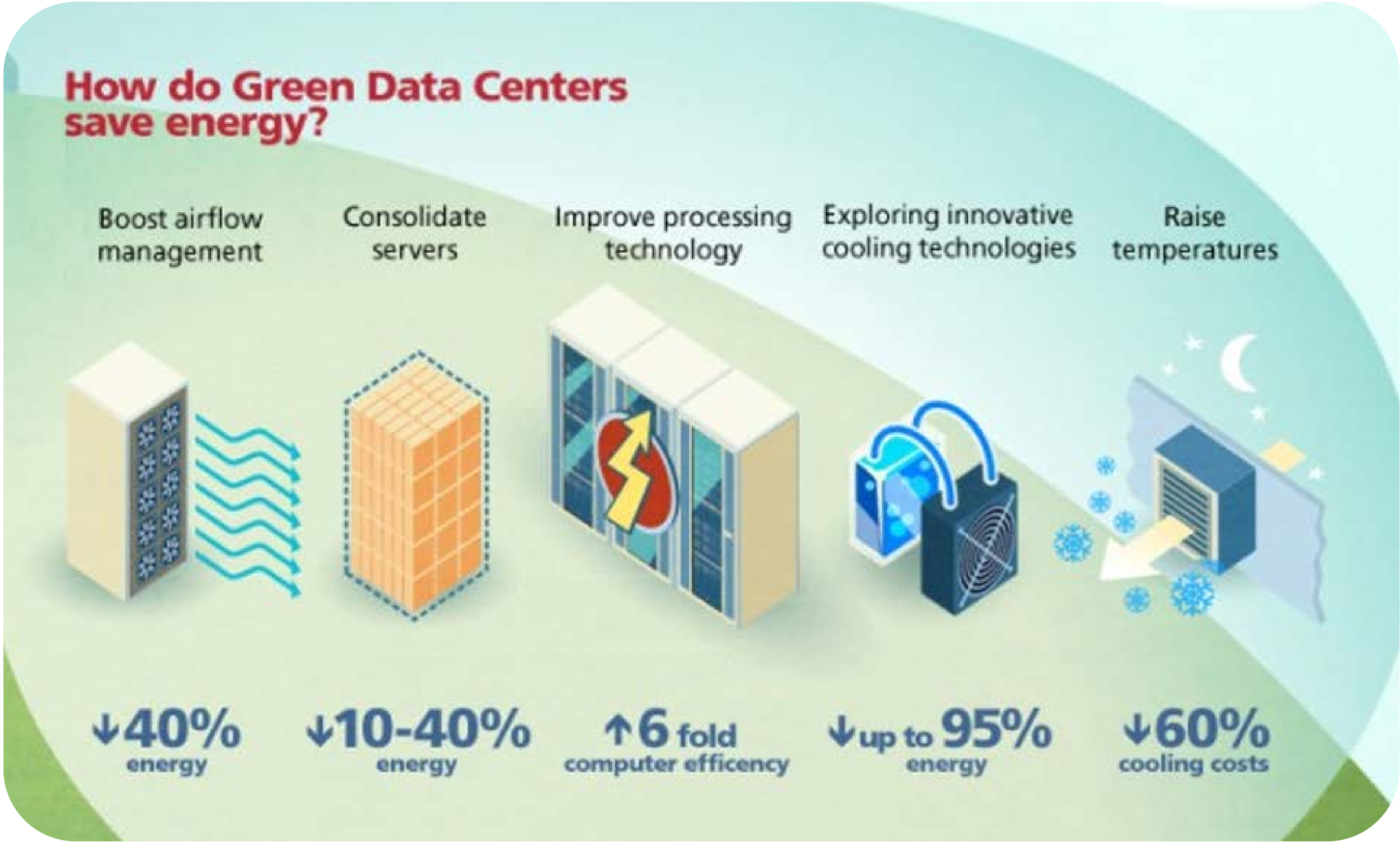
Power consumption and efficiency



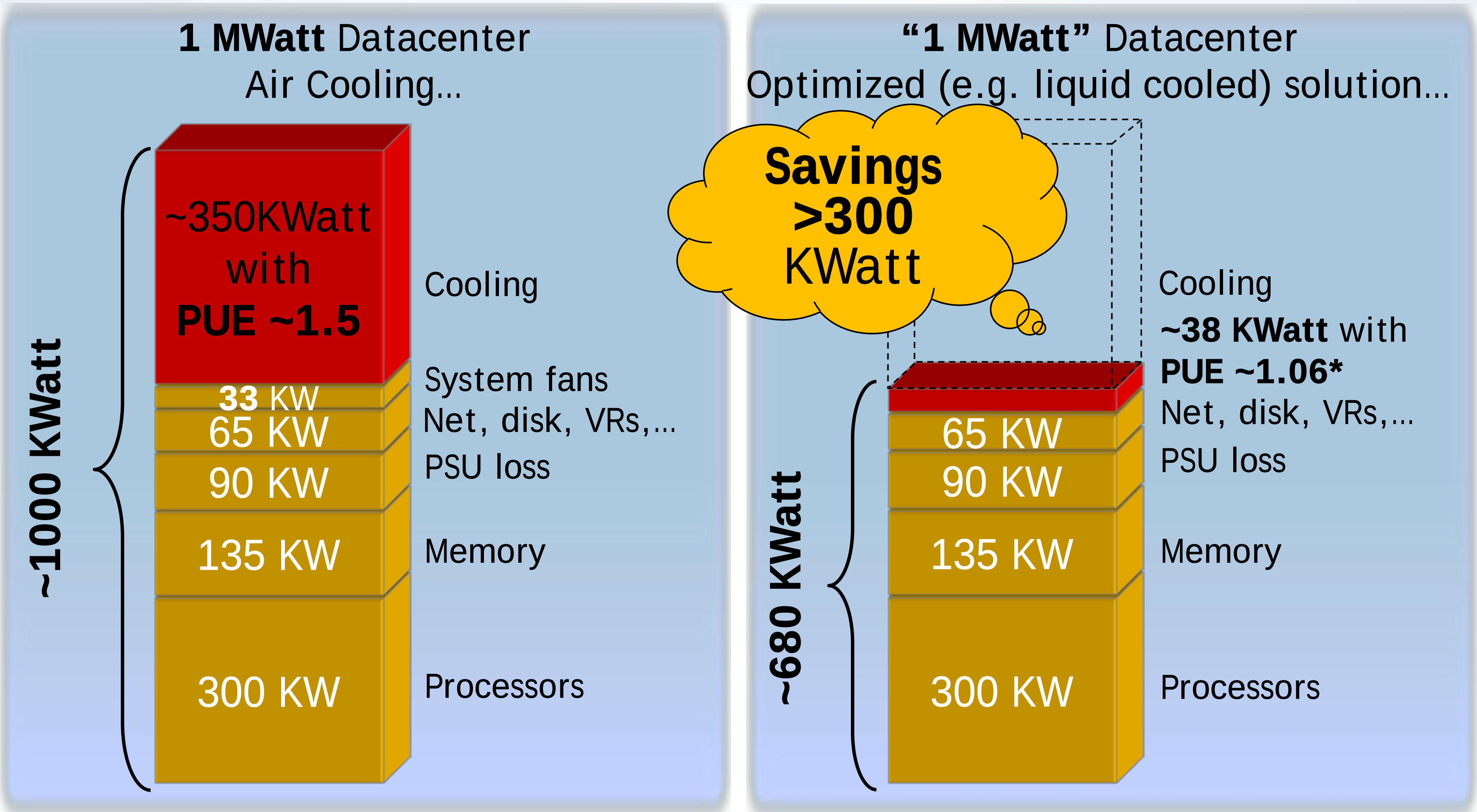
$$\int_{t_{start}}^{t_{end}} power\_consumption(t)dt \cong \sum_{i=0}^{n} measured\_power(t_i)\, t_i$$

For illustration only.

Data Center Energy Savings



Datacenter power consumption breakdown



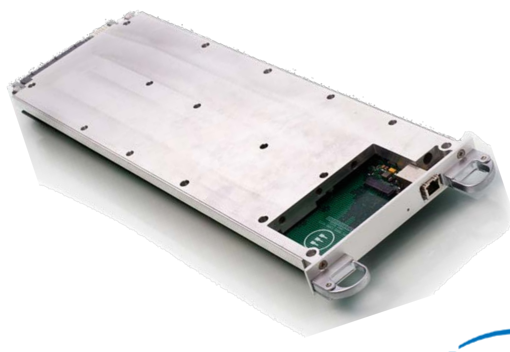
Efficient cooling – the first optimization opportunity

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

* PUE of 1.06 has been achieved on several direct liquid cooling systems running in optimized datacenters with cooling equipment operated in Free Cooling mode.

Available direct liquid cooling options

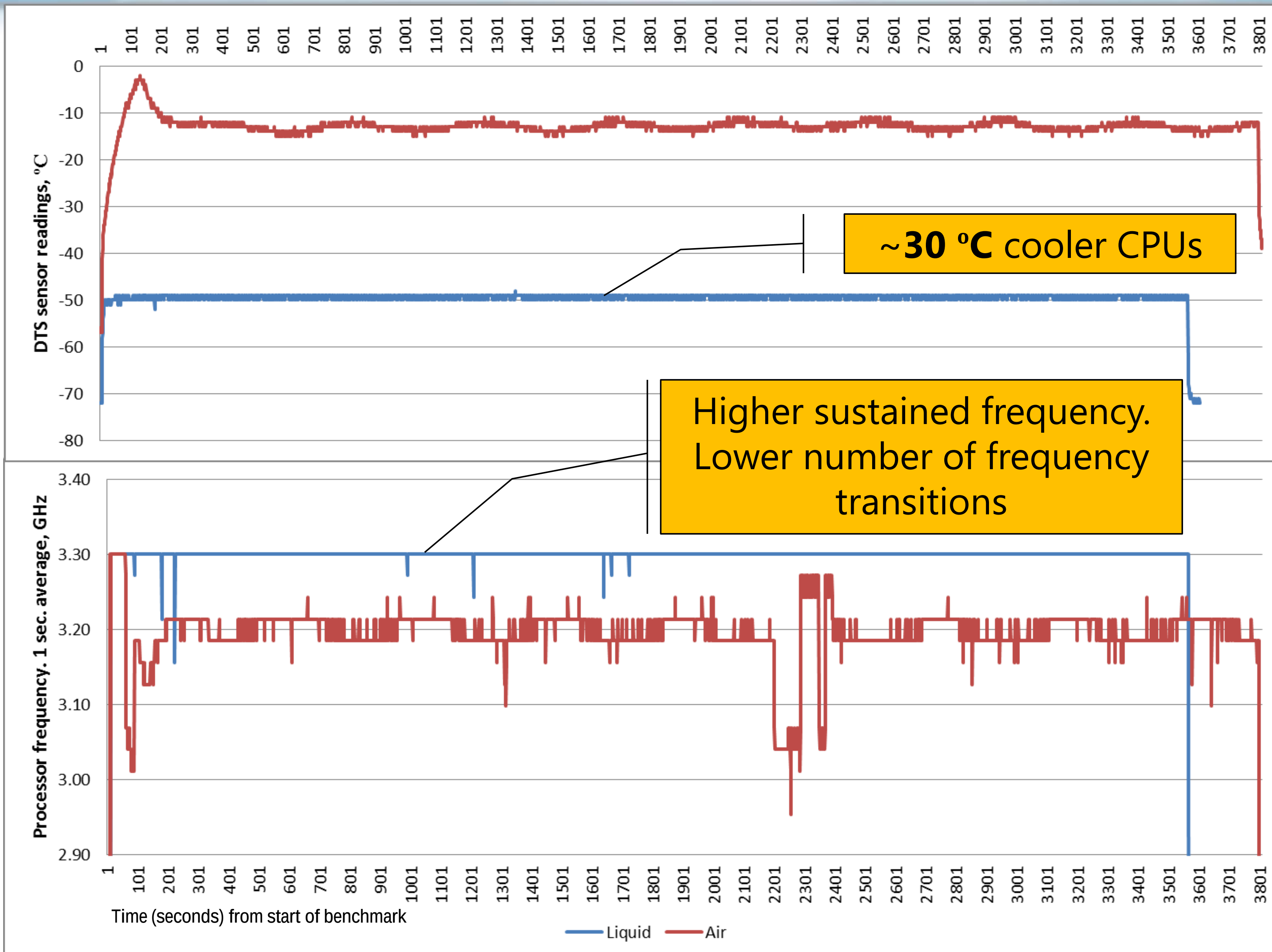
	Submergence of entire server(s)	Partially covered components	Cold-plate covering all components
Pros:	Can use stock servers - still modifications are required to remove fans and disks	May rely on components found in the consumer space Fast to develop new designs due to modular architecture	Highest density Low cost (if the design is right)
Cons:	Is heavy No gains in density if stock servers used Complex handling	Do not remove 100% of heat – need additional air flow Is costly	Can be heavy (but solved) Requires very skilled developers to design the cold-plate



Study #1: liquid vs. air impact on HPC applications

	Air-cooling	Direct liquid cooling
Application	**NAMD** version 2.9 (2012-04-30), x86_64, built: ICC compiler with "–O3 –xAVX" options	
Benchmark input	**ApoA1:** 92224 atoms, 65000 steps (~1h run time), 12A cutoff+PME 4 steps, periodic	
Processor	Intel® Xeon® Processor **E5-2690:** C2 step 2.90GHz, 8 cores, 8GT/s QPI, 135W TDP	
Memory	**64GB** (8*8GB DDR3-1600 Samsung PC3-12800 ECC RDIMM, P/N: M392B1K70DM0-CK0)	
Server board	Intel Server Board **S2600JFF** (AKA Jefferson Pass)	
Power meter	Zimmer **LMG95** Precision Power Meter, measuring at 220V AC	
Cooling	3 dual-rotor fans per board of **Intel Server H2200JF** server chassis	Liquid, Aluminium cold plate «**RSC Tornado**» system

Study #1: observations



Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as NAMD, 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.
Source: Intel Internal Measurements as of March 2013. For more information go to http://www.intel.com/performance

Study #1: summary of the results and key takeaways

	Air cooling	Liquid Cooling	Difference
Application wall time	63 min. 21 sec. (3801 seconds)	59 min. 29 sec. (3569 seconds)	6.5% (1.065x)
Average power (AC220V)	491 Watts	425 Watts	15.5% (1.155x)
Consumed energy	0.518 kWatts*hour (1 864 800 joules)	0.421 kWatts*hour (1 515 600 joules)	23% (1.230x)
Estimated cooling PUE	1.55	1.02-1.1	50% (1.5x)
Estimated total consumed energy (including cooling)	0.80 kWatts*hour	0.44 kWatts*hour	~82% (1.82x)

- Significant (over 1.8x) lower total energy consumption of direct liquid cooled system while running HPC applications
- Application runs faster (over 6%) in liquid cool system due to higher average sustained frequency (+1 bin/100MHz better Turbo upside)
- Average power consumption is lower due to absence of fans (up to 3A*12V each), which offsets higher CPU power draw due to higher clock

Precise control of temperature helps reduce power draw and improve application performance

Study #2: identifying the best memory configuration

Objective: identify the best configuration meeting performance target and consuming lowest amount of electrical energy within 100KW power envelope

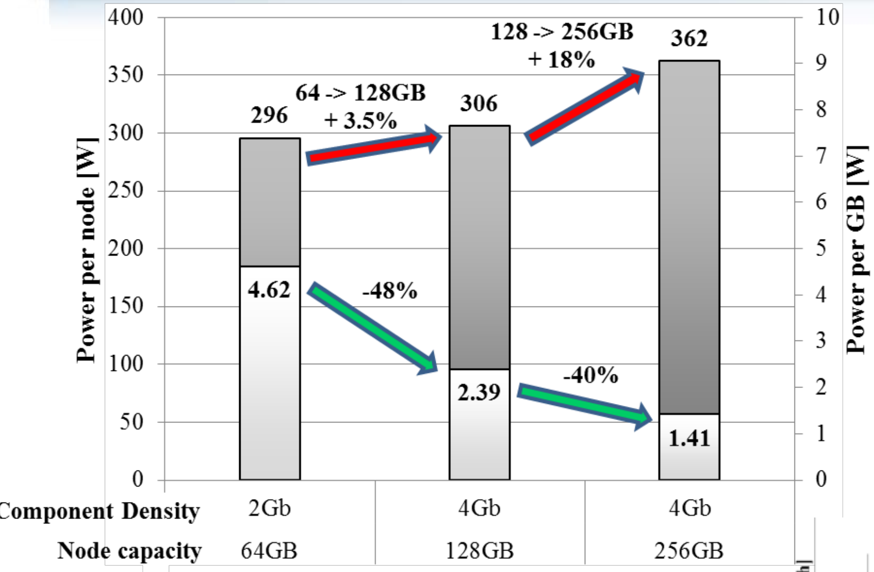
Setup: Intel Server Board S2600CP2J in P4308XXMHGC chassis, 750W PSU. Two Xeon E5-2670 processors and up to 16 DDR3 RDIMMs:

Samsung memory part number	Module density	Speed	Voltage	Component density	Technology node
M393B2G70BH0-YK0	16 GB	1600	1,35 V	4Gb	30nm
M393B2G70BH0-CK0	16 GB	1600	1,5 V	4Gb	30nm
M393B2G70AH0-YK0	16 GB	1600	1,35 V	4Gb	40nm
M393B1K70DH0-YK0	8 GB	1600	1,35 V	2Gb	30nm
M393B1K70DH0-CK0	8 GB	1600	1,5 V	2Gb	30nm
M393B1K70CH0-YH9	8 GB	1333	1,35 V	2Gb	40nm
M393B1K70CH0-CH9	8 GB	1333	1,5 V	2Gb	40nm

Benchmark & workload:

- STREAM 5.9 modified to utilize 85% of installed RAM. TRIAD workload was used
- **Metric**: "energy effectiveness" = amount of data moved per energy unit (in **TB/kWh)**, where the higher value means higher energy effectiveness

Study #2: observations and results



Memory capacity and component density:

- Higher memory density per node consumes more power: +21.5% between 64GB and 256GB
- the power consumption per GB of capacity decreases due to power efficiency of 4Gb component vs. 2Gb.

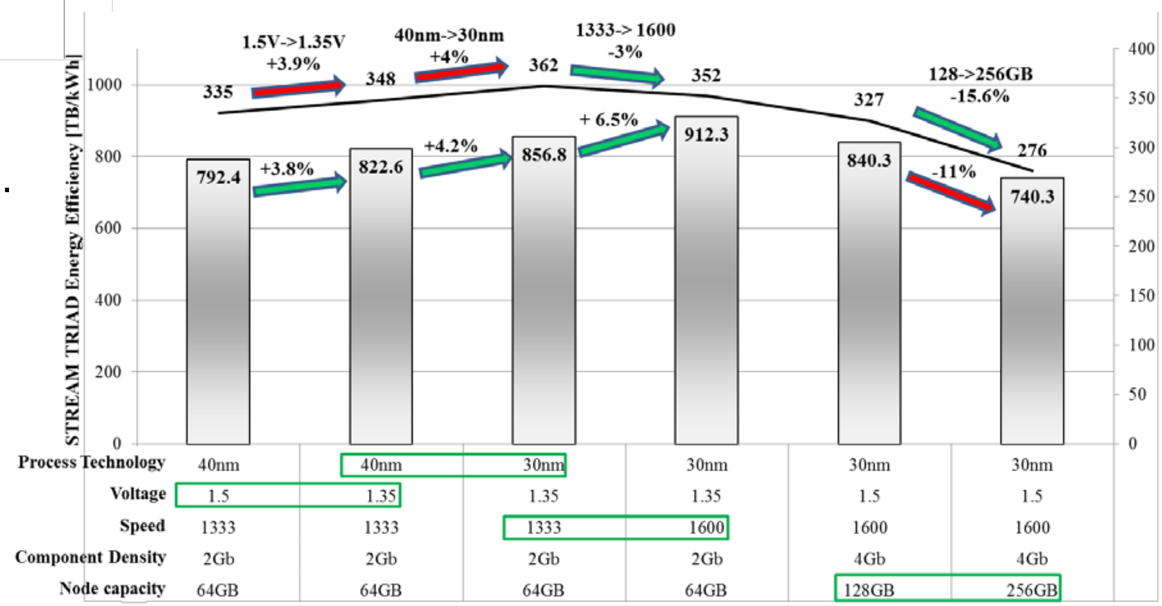
Within 100 KW power limit:

- **276** nodes with 256GB@1600 Mbps, vs.
- **352** nodes with 64GB

i.e. **18% less nodes** with high density modules will provide **3.1x more total** memory in the cluster

The most energy efficient configuration:

- **64GB** capacity per node
- **30nm DRAM process technology**
- running at **low voltage (1.35V)** and **1600Mbps**



Study #3: power limiting impact on energy efficiency

- **Objective**: study impact of power limiting on HPC application performance and [power,energy] efficiency
- **Benchmarks**: NAS Parallel Benchmarks, v.3.3-MPI

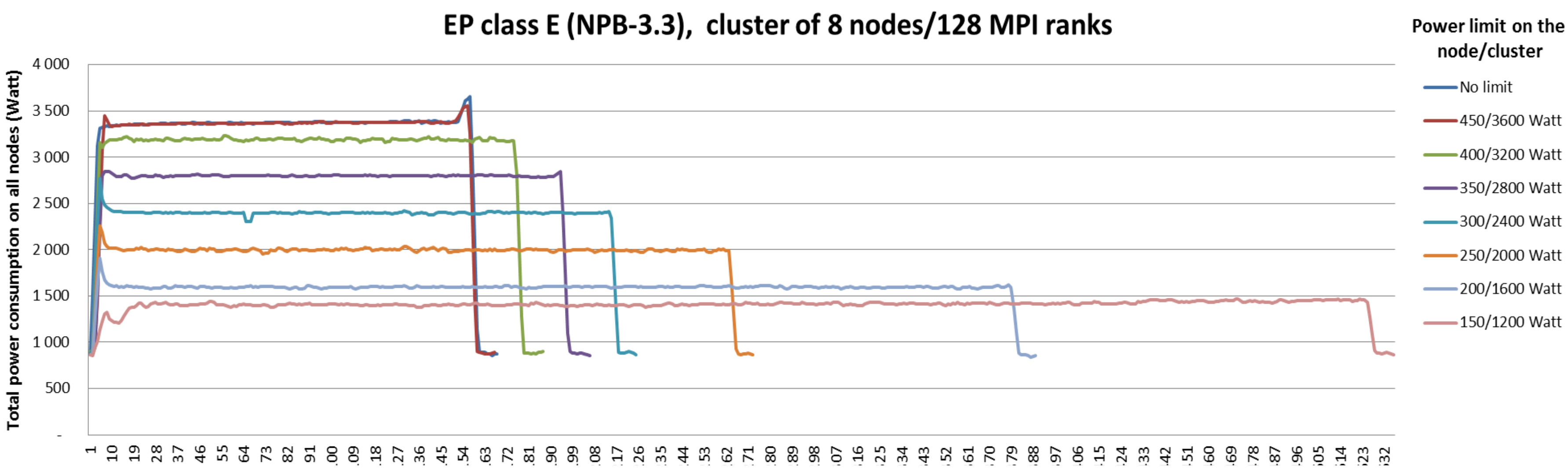
NPB v.3.3	Class	# MPI ranks	PPN	# of nodes	Workload size/# of iterations
CG	E	128	16	8	size: 9000000, iterations: 100
MG	E	128	16	8	size: 2048x2048x2048, iter.: 50
LU	E	128	16	8	size: 1020x1020x1020, iter.: 300
BT	E	144	16	9	size: 1020x1020x1020, iter.: 250
SP	E	144	16	9	size: 1020x1020x1020, iter.: 500
EP	E	256	32	8	size: 2199023255552

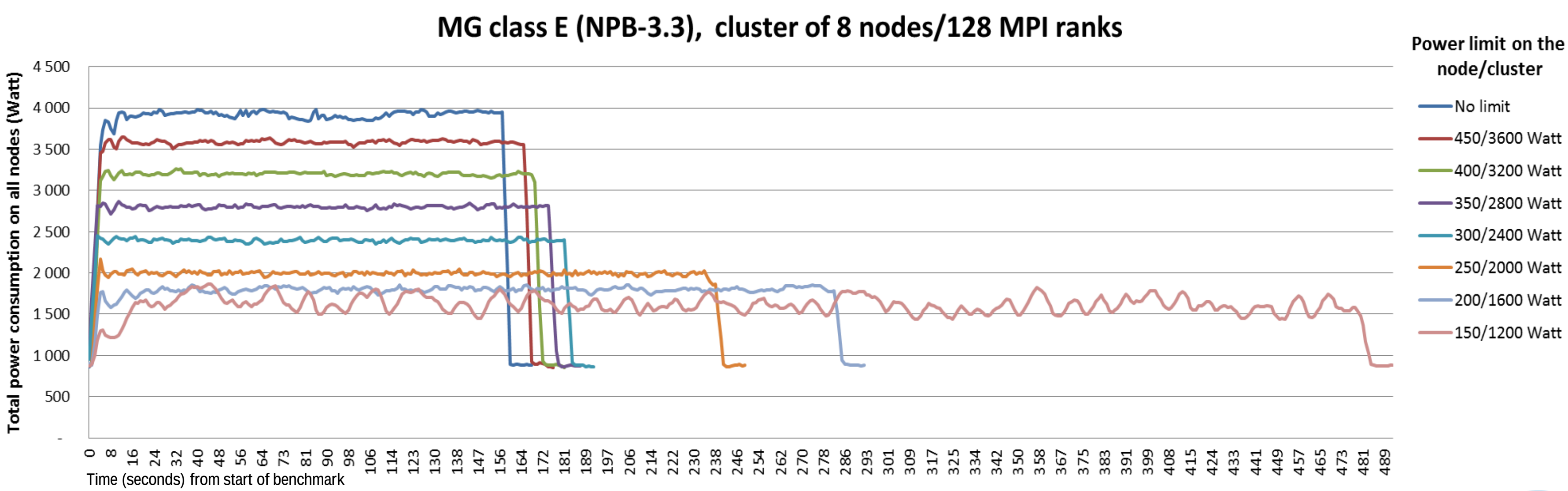
Benchmarks built with Intel Fortran, C/C++ 13.0.1, Intel MPI 4.1.0.024

Systems: cluster 16 nodes, each including

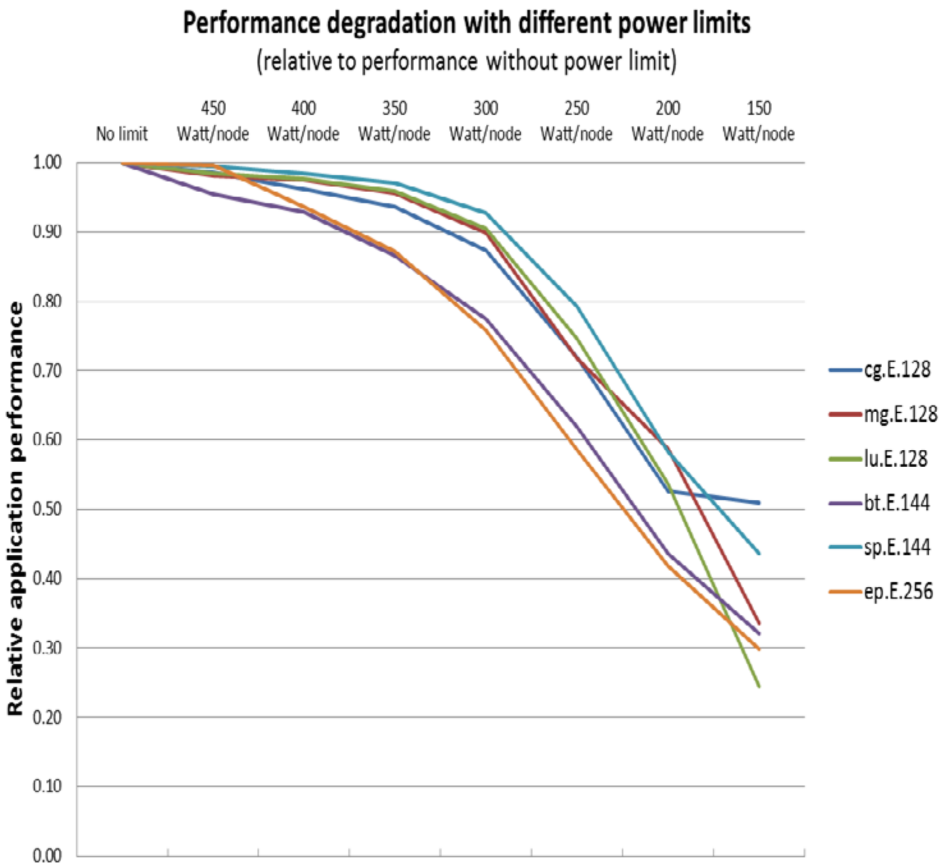
- 2x **Xeon E5-2690**, **64GB** (8x8GB DDR3-1600 RDIMM), FDR **Infiniband**, Intel® **S2600JFF** (Jefferson Pass) with **Intel® Node Manager** enabled
- Power consumption limited using Intel Node Manager to
  no power limit, 450, 400, 350, 300, 250, 200 & 150 Watts per node

Study #3: observations

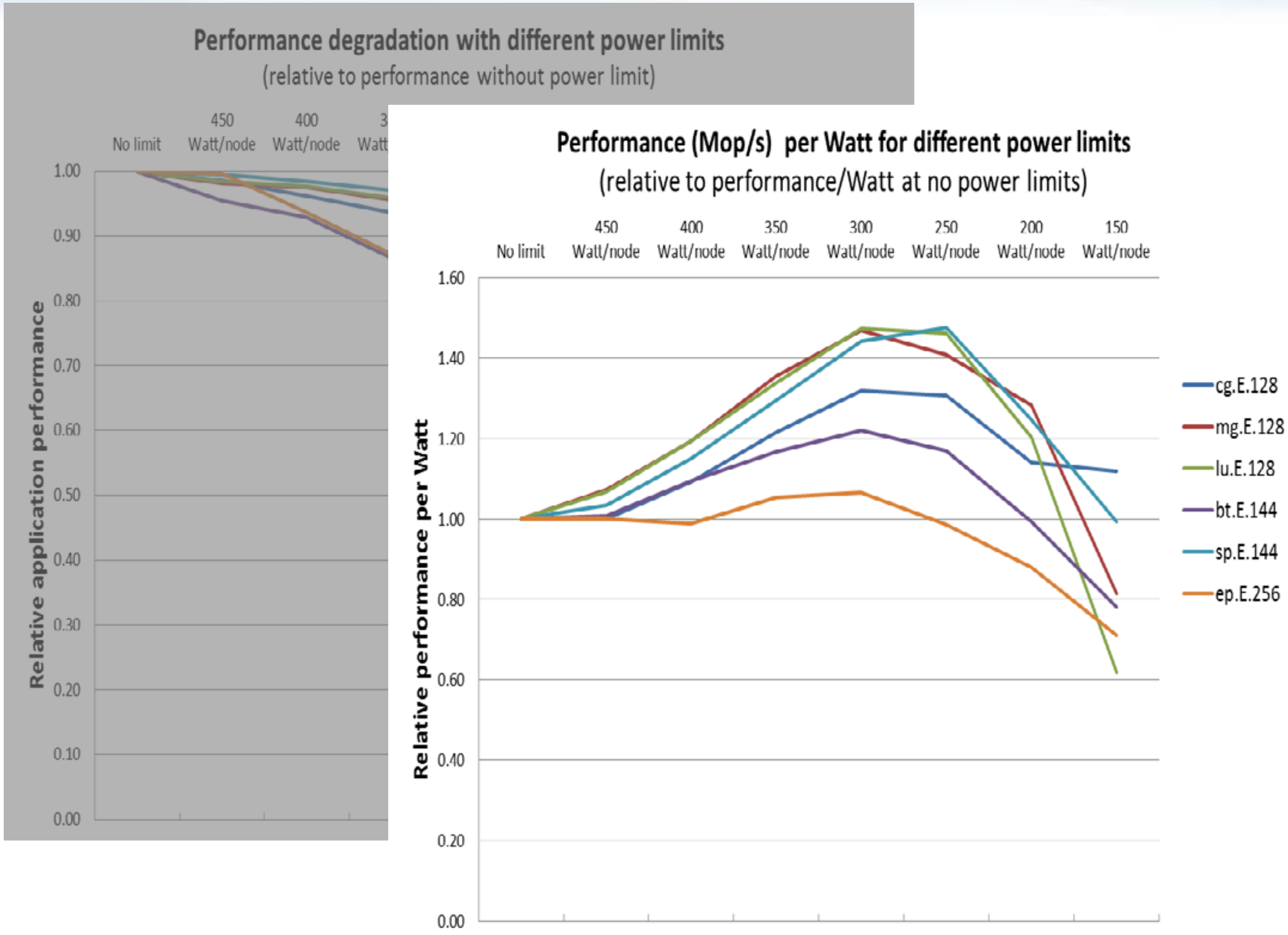




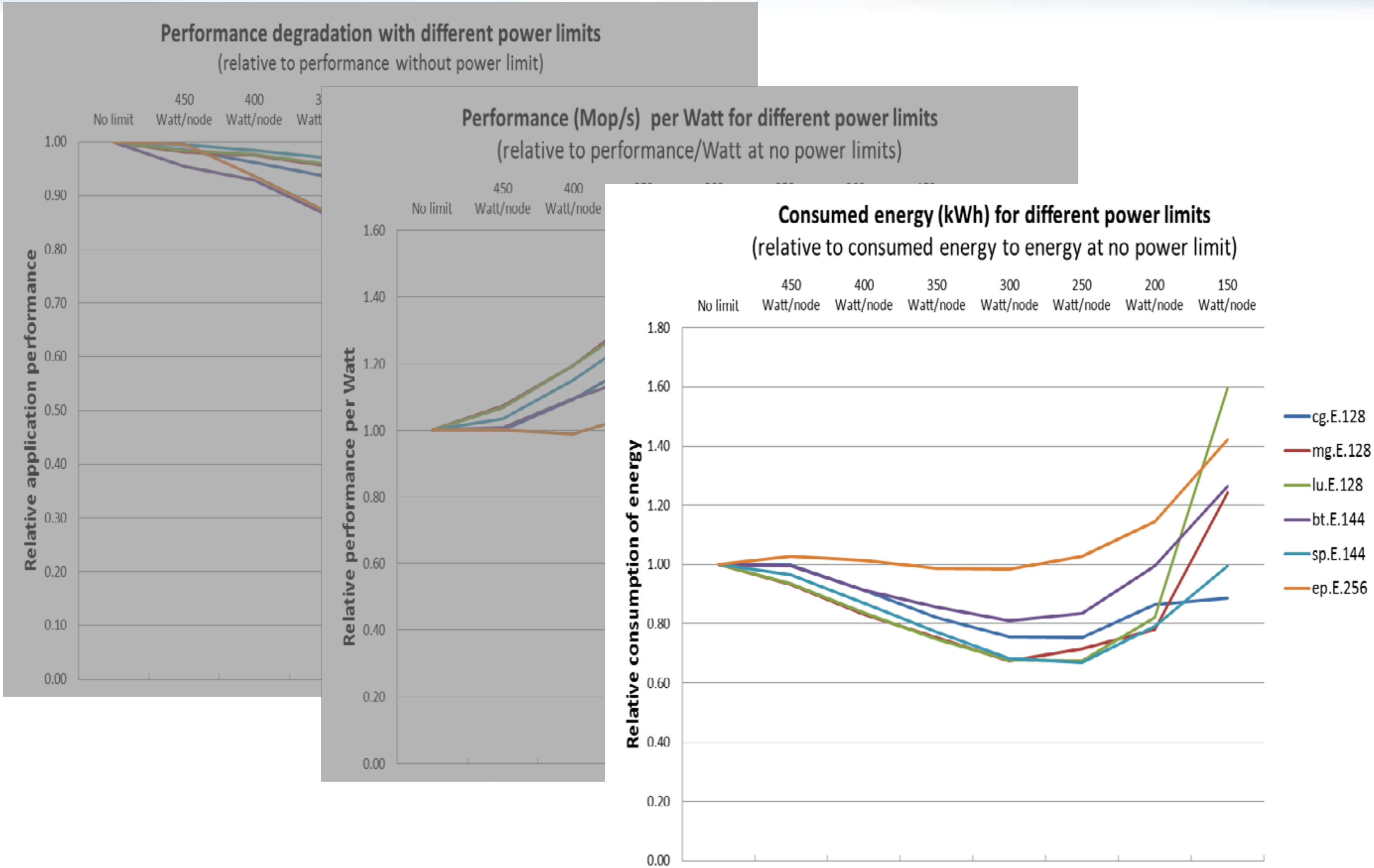
Study #3: observations, cont.



Study #3: observations, cont.



Study #3: observations, cont.



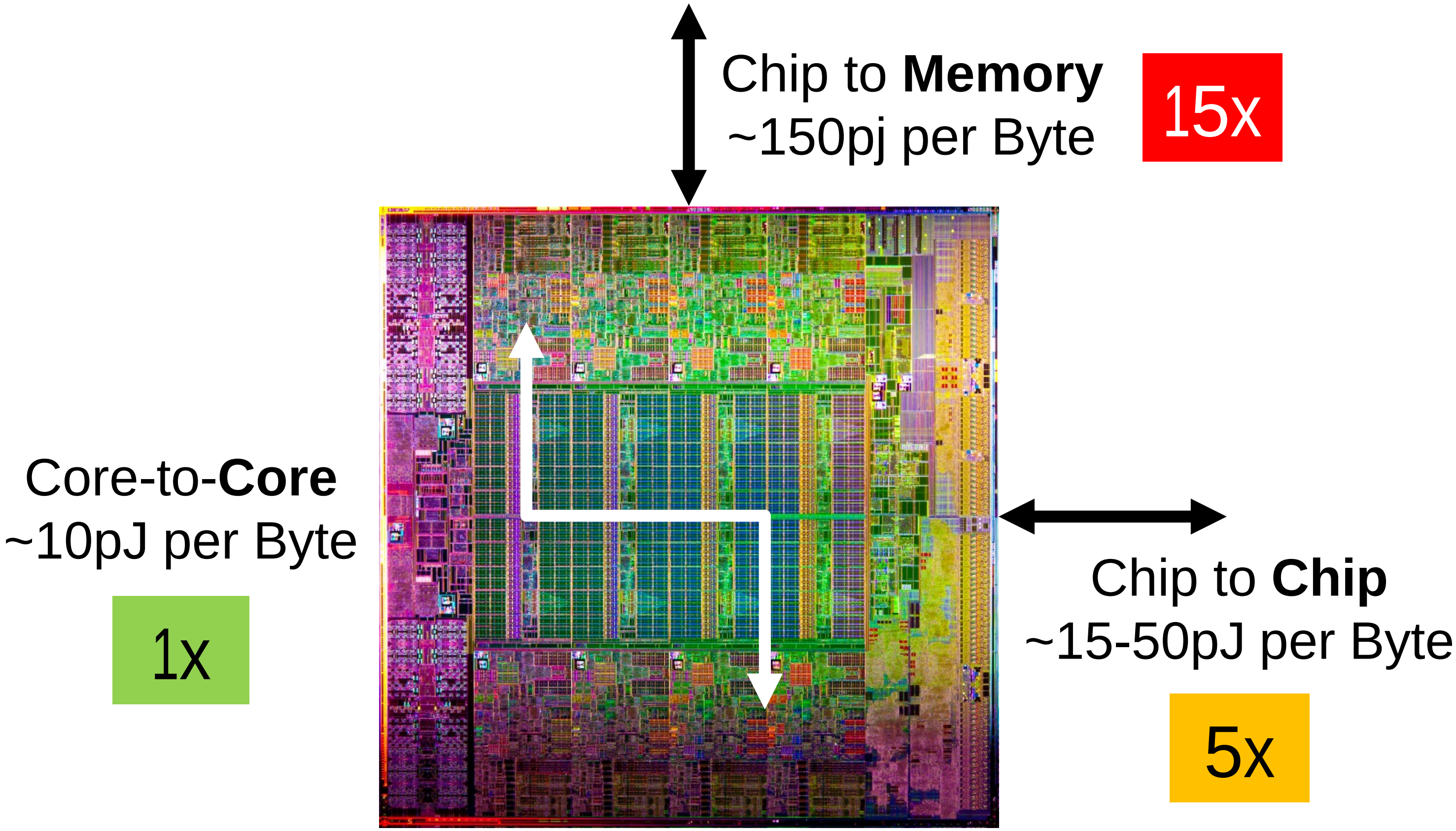
Study #3: Summary of the results and key takeaways

NPB Test	Energy (kWh)		Gain	Most energy efficient power envelope per node (Watt)	Performance/Watt (Mops/Watt)		Gain	Best power envelope per node (Watt) for power/performance
	Total at **no power limit**	**Min** energy			at **no power limit**	Best Perf./Watt		
CG	1.63	1.24	1.31x	**300**	5.83	7.70	1.31x	**300**
MG	0.17	0.12	1.41x	**300**	42.10	61.86	1.47x	**300**
LU	3.87	2.62	1.47x	**300**	46.02	67.87	1.47x	**300**
BT	3.28	2.66	1.23x	**300**	79.16	96.56	1.22x	**300**
SP	4.79	3.2	1.49x	**250**	27.42	40.50	1.44x	**250**
EP	0.145	0.143	1.01x	**350**	4.21	4.49	1.06x	**300**

- Amount of consumed energy varies from application to application and depends on the imposed power limit on the node
- The most **"power efficient"** power limit won't necessarily be the most "**energy efficient"** one!

Right choice of power envelope for application can result in significant energy savings

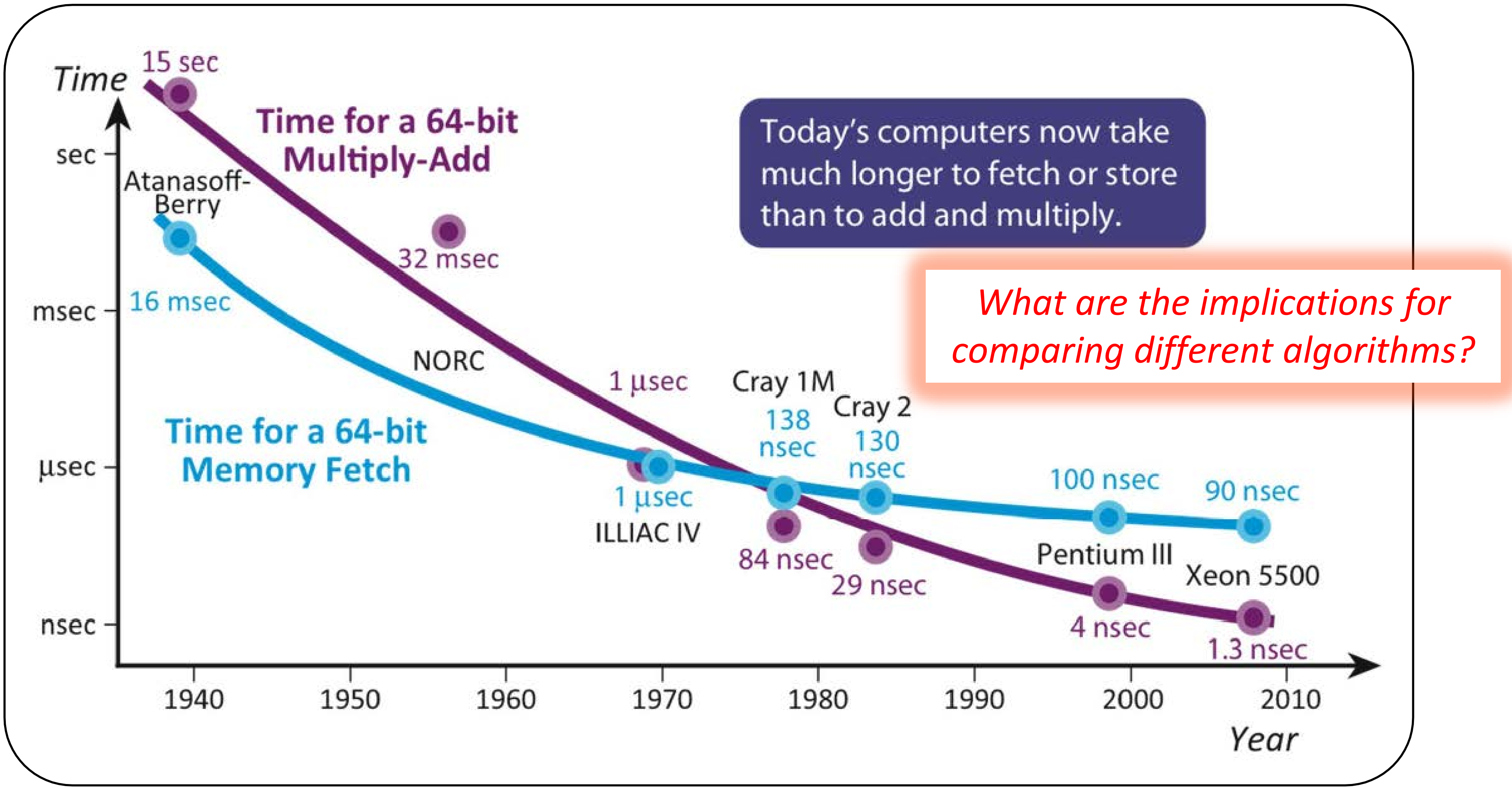
Data movement is expensive



For illustration only.

Integration is key

Food for thoughts – “Memory Wall“

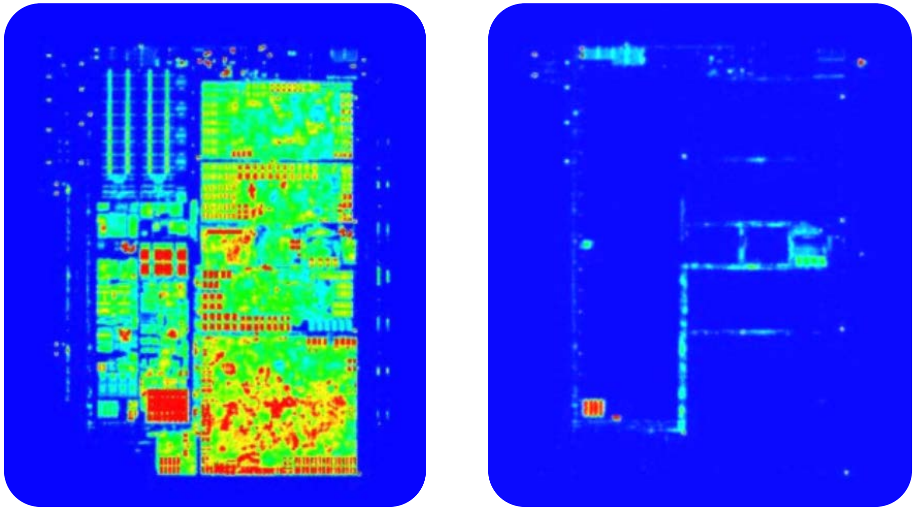


Potential opportunities to reduce power

- Use contiguous memory data access instead of random memory access
- Re-use data as much as possible, in space and time – good cache utilization for lower energy and higher performance
- Use arithmetic to (re-)calculate data instead of loading from memory when data are already available, e.g. (x+y)/2
- Use smart „on-the-fly“ interpolation instead of pure table-lookup from memory
- Use SIMD and multi/manycores for faster computing
- Use asynchronous computing and communication on clusters
- Consider reduced data types if applicable within the memory hierarchy
- Consider reduced arithmetic precision – with caution
- Consider more efficient algorithms to reduce execution time
- Utilize latest generation of processors with advanced power management - for lower energy and higher performance
- Utilize SSDs instead of HDDs for I/O intensive workloads
- Keep TCO in mind

Example of Power Management

- G0 – Working
  - S0 – CPU Fully powered on
    - C0 – Active mode
      - P0
      - ⋮
      - Pn
    - C1 – Auto halt
    - C1E – Auto halt, low freq, low voltage
    - C3 – L1/L2 caches flush, clocks off
    - C6 – save core states before shutdown
- G1 – Sleeping
  - S3 cold – Sleep – Suspend To Ram (STR)
  - S4 – Hibernate – Suspend To Disk (STD), Wakeup on PCH
  - S5 – Soft Off – no power, Wakeup on PCH
- G3 – Mechanical Off



Thermal camera images

C: Core Power States
P: Performance States

One last thing ...

know what you do

A very simple arithmetic example

using IEEE 64-bit DP-F.P.

X_1	X_2	X_3	X_4	X_5	SUM(X_1:X_5)	
1.00E+21	17	-10	130	-1.00E+21	0.00	✗✗
1.00E+21	-10	130	-1.00E+21	17	17.00	✗✗
1.00E+21	17	-1.00E+21	-10	130	120.00	✗
1.00E+21	-10	-1.00E+21	130	17	147.00	✗
1.00E+21	-1.00E+21	17	-10	130	137.00	✓
1.00E+21	17	130	-1.00E+21	-10	-10.00	✗✗✗

Source: Ulrich Kulisch, *Computer Arithmetic and Validity,* de Gruyter Studies in Mathematics 33 (2008), p. 250

"Results can be satisfactory, inaccurate or completely wrong. Neither the computation itself nor the computed result indicate which one of the three cases has occurred."

intel®

BACKUP: Power breakdown

- "1 Mwatt" is approx. for 1300 HPC servers, each including:
  - 2x CPUs at 115W each (running well at TDP, e.g. with Linpack);
  - 16x 8GB DDR3-1600 RDIMM. Memory power estimated to 6.5W loaded power draw per module (with VRs). Internal measurements, and also cross-checked with other publically available sources, e.g. here http://h20000.www2.hp.com/bc/docs/support/SupportManual/c03293145/c03293145.pdf
  - 3x high performance fans inside server totalling 25 Watts… e.g. as in Intel Bobcat peak chassis. Quote them separately as the fans are **absent in the liquid-cooled system configuration;**
  - Others: disk, network adapters (such as IB), on-board VRs and other small components are estimated as 50 Watt per server;
  - Total power conversion efficiency AC to 12V DC taken to 83%;
- Total power consumption of each server is ~410 Watts on DC rails and with est. PSU efficiency of 83% is 493 Watts on AC (internally measured 490-495 Watts on 220VAC under Linpack on Canoe Pass)
- Total power consumption for 1300 servers is then 640 KWats

PUE options:

- If PUE is estimated at 1.5-1.55 (good for air cooled datacenter with free cooling) the total power consumption will be 960-992KW for the datacenter.
- If PUE is estimated at 1.05-1.06 (measured in several liquid cooled datacenter installations) the total power consumption will be 670-680KWatts for the datacenter.

…but **your mileage may vary**, of course

Source: Intel Internal Estimates as of June 2013.

For more information go to http://www.intel.com/performance