



SuperMUC & First Experiences using the improved Power Measurement Methodology (Booth #537)

Torsten Wilde, Herbert Huber, Axel Auweter (HPC Group, Leibniz Supercomputing Centre)

Charles Archer, Torsten Bloth, Achim Bömelburg, Ingmar Meijer, Steffen Waitz (IBM)



Leibniz Supercomputing Center



❑ Size

- IT = 3160.50 m²
- Cooling = 4643.00 m²
- Electrical = 1750.00 m²

❑ Power connectivity

- Dedicated redundant 10 MW 20 kV power line

❑ UPS

- 1.6 MVA diesel generator and 1040 kVA static UPS for highly critical LRZ services (300 kW for critical IT equipment)
- 6 x 1.6 MVA (SuperMUC) & 3 x 1.1 MVA fly wheel UPS

❑ Cooling

- 7 chillers with a total Cooling Power (CP) of 7.2 MW
- 11 cooling towers (16 MW total CP) and one 4 MW evaporation tower

SuperMUC rendered on SuperMUC



SuperMUC Phase 1: Technical Highlights



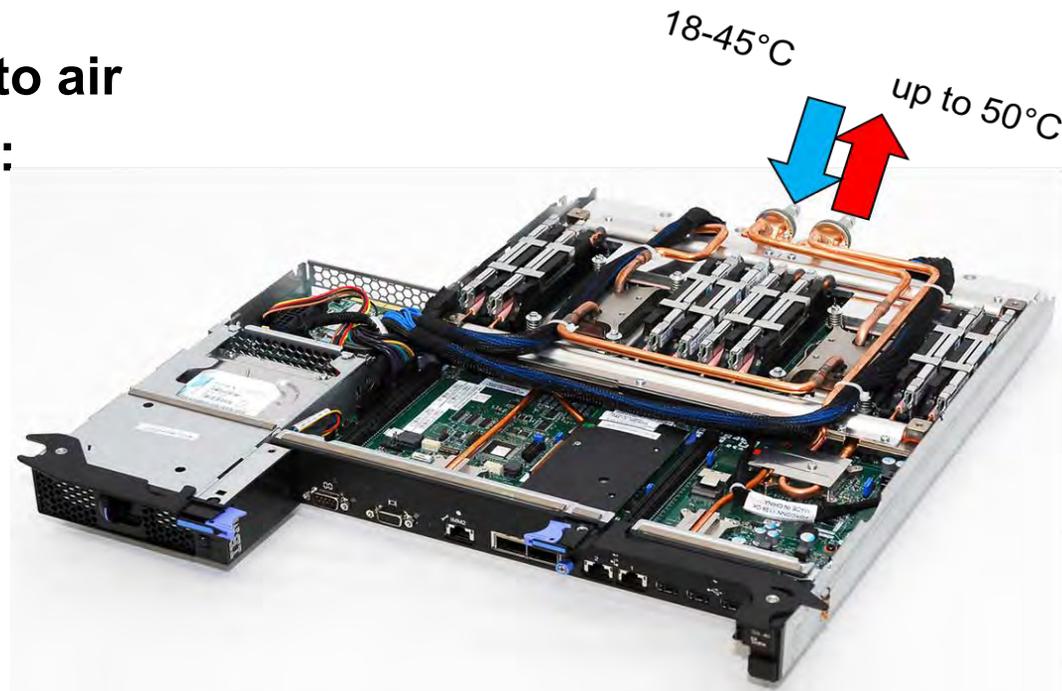
- ❑ **3 PetaFlop/s Peak performance (9216 IBM System x iDataPlex M4 Direct Water Cooled nodes, 147456 E5-2680 cores - Intel Sandy Bridge)**
- ❑ **324 TB of main memory**
- ❑ **Mellanox Infiniband FDR10 Interconnect, Fat Tree Topology**
- ❑ **SLES10 operating system with IBM PE and IBM LoadLeveler**
- ❑ **Large common File Space for multiple purpose**
 - 10 PByte File Space based on IBM GPFS and DDN SFA12000 storage controllers with 200 GByte/s aggregated I/O Bandwidth
 - 2 PByte NAS Storage with 10 GByte/s aggregated I/O Bandwidth
- ❑ **Innovative Technology for Energy Efficient Computing**
 - **Direct Warm Water Cooling**
 - **Energy-aware Scheduling**



Direct water cooled IBM iDATAPLEX dx360 M4 node



- ❑ **Based on Aquasar Prototype**
 - Jointly developed by IBM Labs in Böblingen and Zürich
- ❑ **Compute Nodes (Processors & Memory) Cooled with Warm Water up to 45°C (inlet)**
 - **Outlet up to 50°C**
- ❑ **only 10% of Heat released into air**
- ❑ **Typical operating conditions:**
 - T_{air} 25 – 35°C
 - T_{water} 18 – 45°C

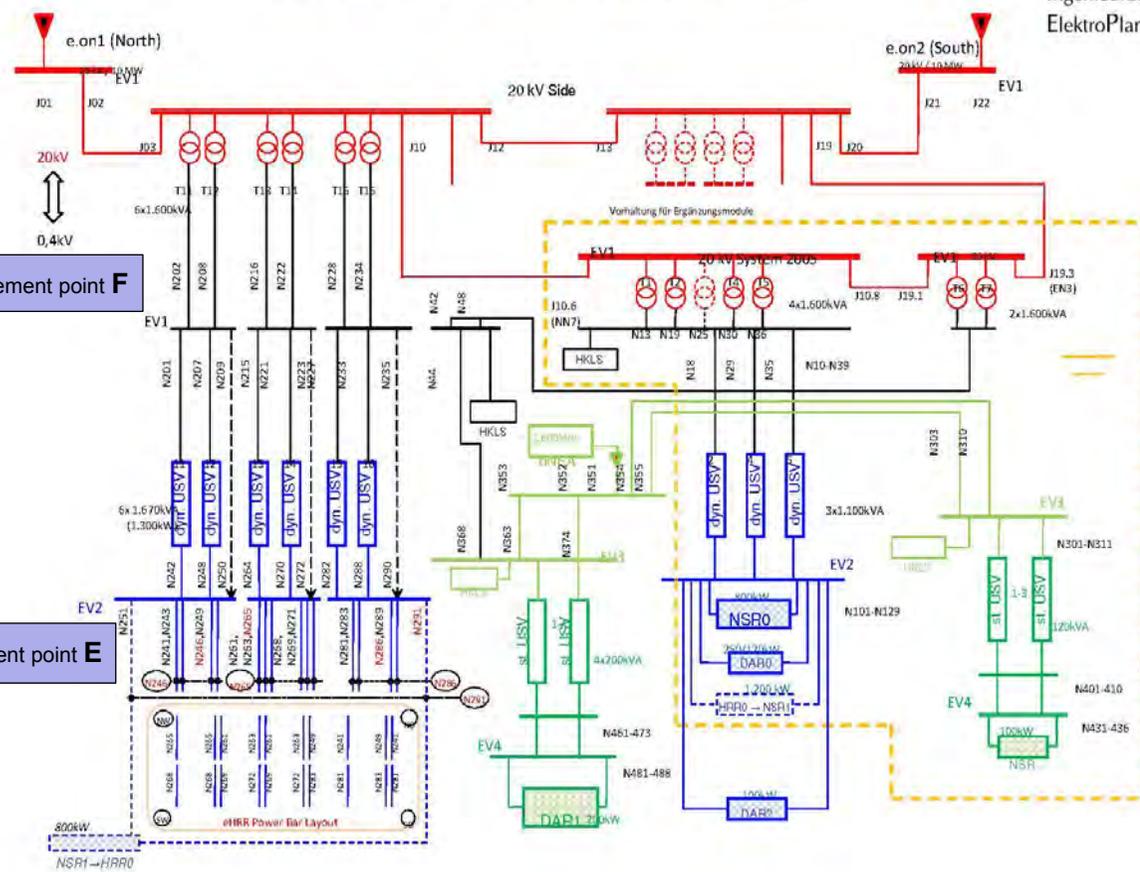


LRZ Infrastructure Power and Energy Measurement Points (1)



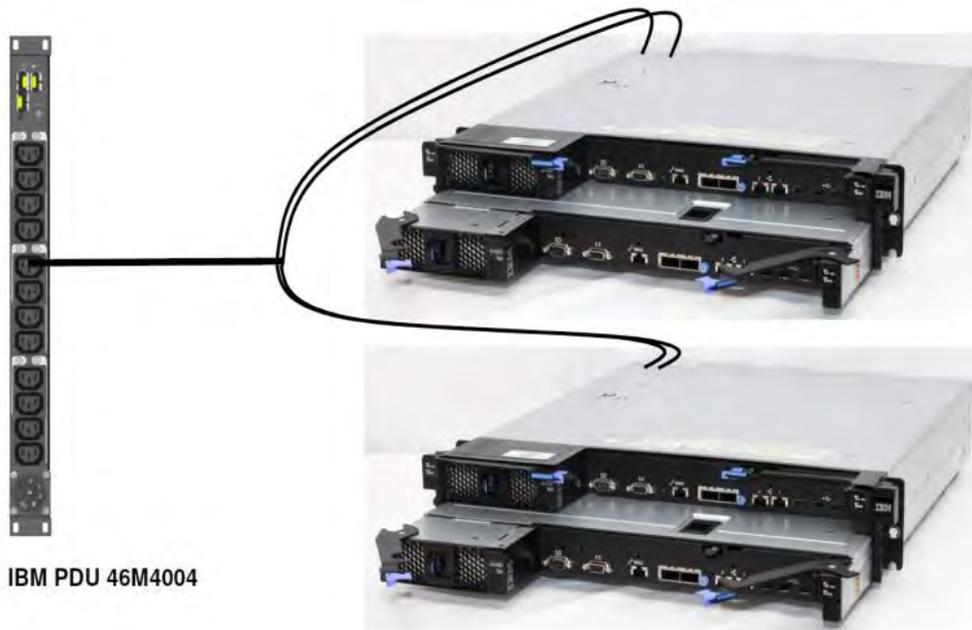
LRZ 20 kV ... 400 V Power Distribution Scheme

IEP
Ingenieurbüro
ElektroPlanung



- Socomec Diris A40/A41 meters at measurement points 1 and 2
- Multi-function digital power & **continuously integrating energy meter** (15 minutes readout interval)
- 1s internal measurement updating period
- Measurements up to the 63th harmonic
- IEC 61557-12 certified
- Energy: IEC 62053-22 Class 0,5S accuracy
- Power: **0.5% accuracy**

SuperMUC Power and Energy Measurement Points (2)



IBM PDU 46M4004

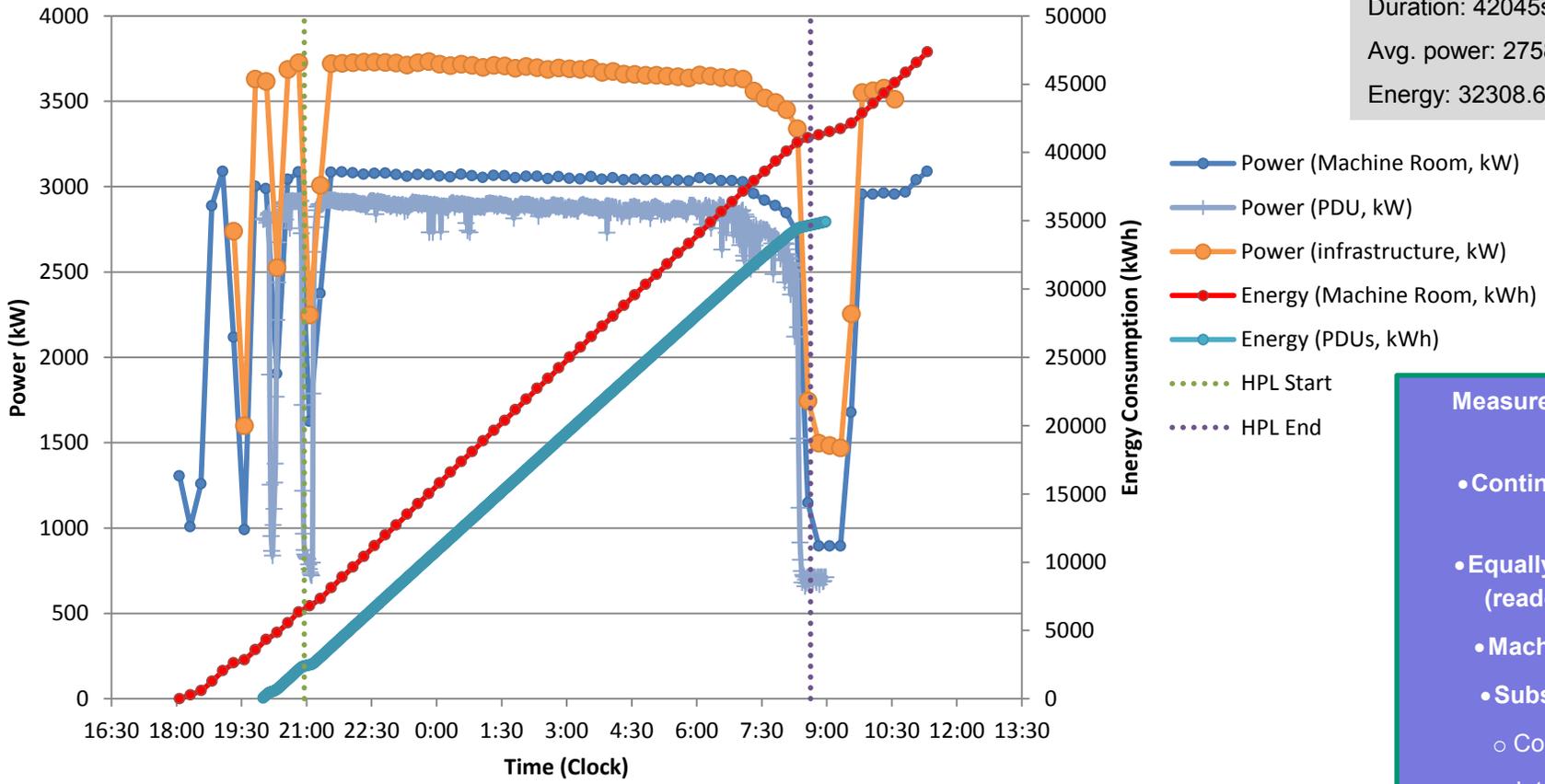
- IBM 46M4004 PDUs are sampling Voltage, Current and Power with a frequency of 120 Hz.
- Power values are averaged over 60 seconds
- One PDU outlet provides power to 4 SuperMUC compute nodes
- One minute readout interval
- RMS Current and Voltage measurements with $\pm 5\%$ accuracy over the entire range
- **Inlet Statistics:**
 - Load Watts (W) - Present Value, Min, Max
 - Cumulative Kilowatt Hours - Present Value, Min, Max
- **Individual Outlet Statistics:**
 - Output Voltage (V) - Present Value, Min, Max
 - Output Current (A) - Present Value, Min, Max
 - Output Power Factor (0.0 - 1.0) - Present Value, Min, Max
 - Load Watts (W) - Present Value, Min, Max
 - Cumulative Kilowatt Hours - Present Value, Min, Max

SuperMUC Green500 Submission Data (Expected Classification Level: L3)



SuperMUC HPL Power Consumption (Infrastructure, Machine Room & PDU Measurements)

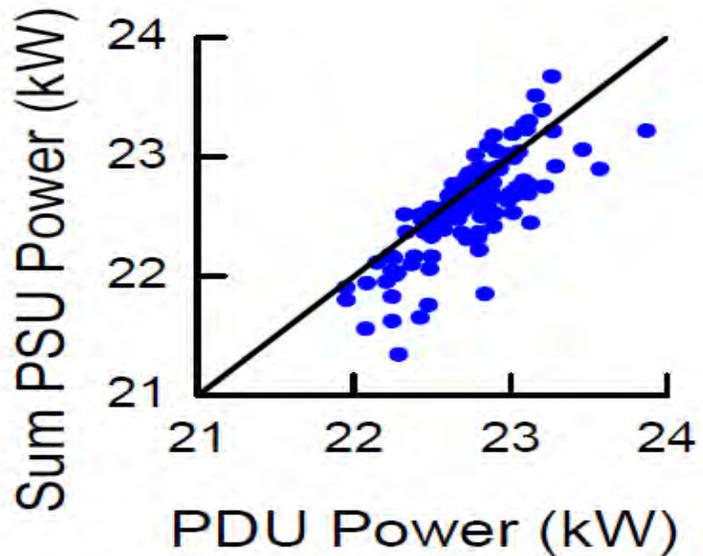
Linpack HPL run May 17, 2012 – 2.582 PF
 Run Start: 17.05.2012 20:56, 965,40 kW
 Run End: 18.05.2012 08:37, 711,02 kW
 Duration: 42045s or 11.68 hours
 Avg. power: 2758.87 kW
 Energy: 32308.68 kWh



Measurement Notes (Energy PDUs):

- Continuous integrated total energy
- Equally spaced time series (readout interval: 1 min)
- Machine fraction: 100%
- Subsystems included:
 - Computational Nodes
 - Interconnect Network

Green500 Measurement Methodology



| |
|---|
| Energy efficiency HPL core phase <i>(single number in GFlops/Watt)</i> |
| 9,380E-01 (PDU, 10 minutes resolution, without cooling) |
| 9,359E-01 (PDU, 1 minutes resolution, without cooling) |
| 9,305E-01 (PDU, 1 minutes resolution, cooling included) |
| 8,871E-01 (machine room measurement) |
| 7,296E-01 (infrastructure measurement) |

Per Node:

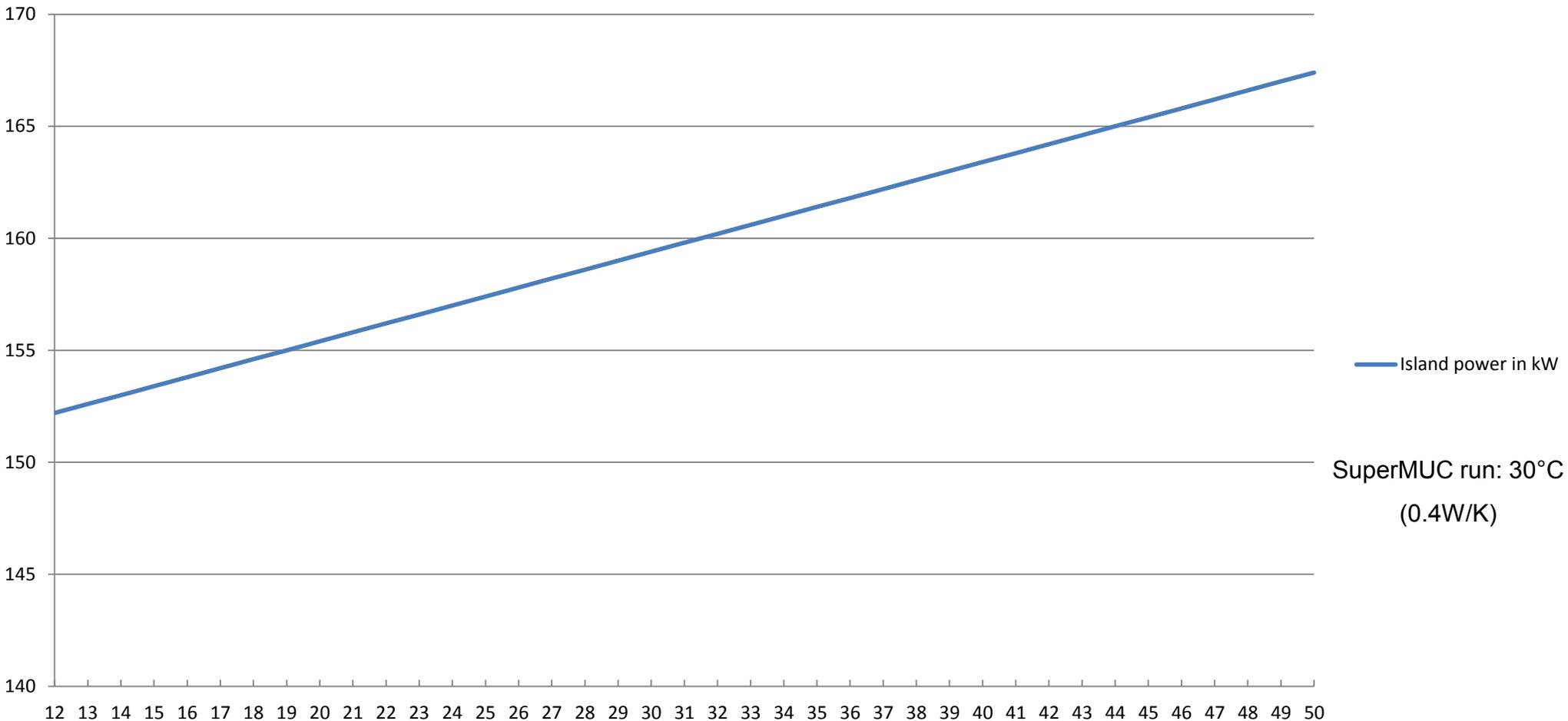
| Freq [GHz] | Power AC [W] | Power DC [W] | Performance [Gflops] | Gflops per W |
|------------|--------------|--------------|----------------------|--------------|
| 2.7 Turbo | 374 | 325 | 348.7 | 0.93 |
| 2.7 | 320 | 283 | 310 | 0.97 |
| 2.5 | 289 | 255 | 288 | 1.00 |
| 2.2 | 249 | 219 | 254 | 1.02 |

Complete system: 0.94 Gflops per W (2.7GHz)

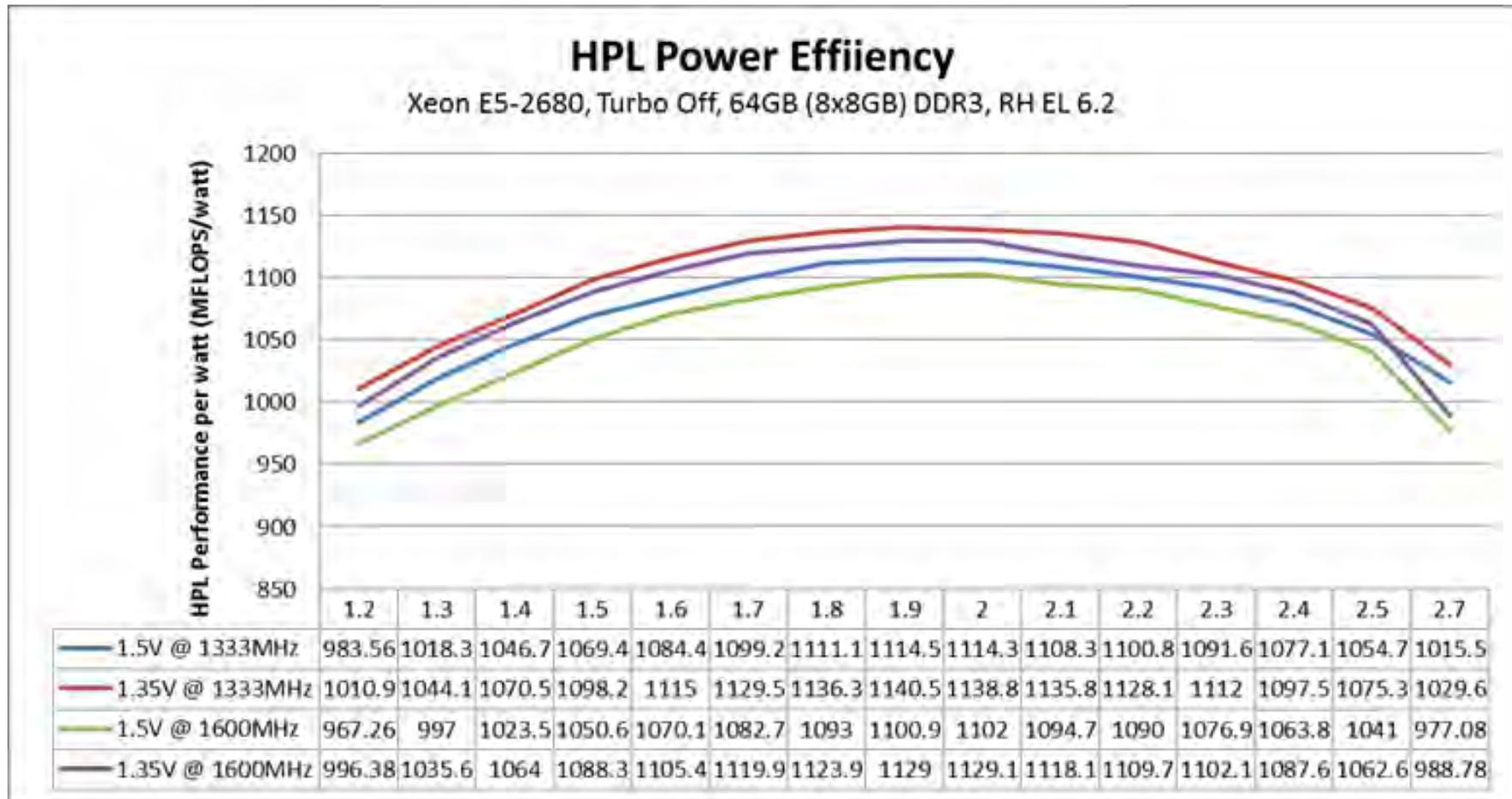
Change: Provide Environmental Data



Island power in kW (2.7GHz, turbo off) vs inlet temperature in C°



Change: Provide CPU frequency and RAM setting



Provided by Andrey Semin, Sr. Engineer HPC, Intel

- ❑ **Well defined measuring procedure**
- ❑ **Raises energy awareness**
- ❑ **Measured Green500 energy efficiencies depend on**
 - System size → smaller is better
 - Processor frequency settings → lower might be better
 - Memory type and speed → slower is better
 - Cooling environment → lower temperature is better but not really more Green
 - Tuned HPL micro kernels and bios settings

❑ **Aggregate same architectures**

- Flops per W should be nearly identical for the same architectures (smaller installations have a slight edge)
- Look at the current list

❑ **Remove all derived systems from the main list**

- I can do simple math, thank you very much

Improvements from a HPC Site perspective



- ❑ **Minimum measurement accuracy for different levels need to be defined**
 - E.g. Level 3 - less than 5%
- ❑ **Different work load mix needed**
- ❑ **Include energy re-use possibilities with system setup**
 - E.g. waste heat re-use
 - Re-use need to be promoted
- ❑ **Move towards more complete system energy consumption**
 - HPC system
 - All infrastructure the HPC system provides
 - Energy required for cooling system(s)

- ❑ **Flops per W just **ONE** indicator out of many**
 - Efficient Linpack machines are not the holly grail for energy efficient HPC
 - Single number hides and muddies complexity associated with efficient HPC architectures
 - Different application domains have different architectural requirements
 - Focus and work with user communities
- ❑ **Don't loose sight of TCO**
 - Cooling requirements, waste heat reuse, infrastructure