SuperMUC & First Experiences using the improved Power Measurement Methodology

Herbert Huber, Axel Auweter, Torsten Wilde, High Performance Computing Group, Leibniz Supercomputing Centre

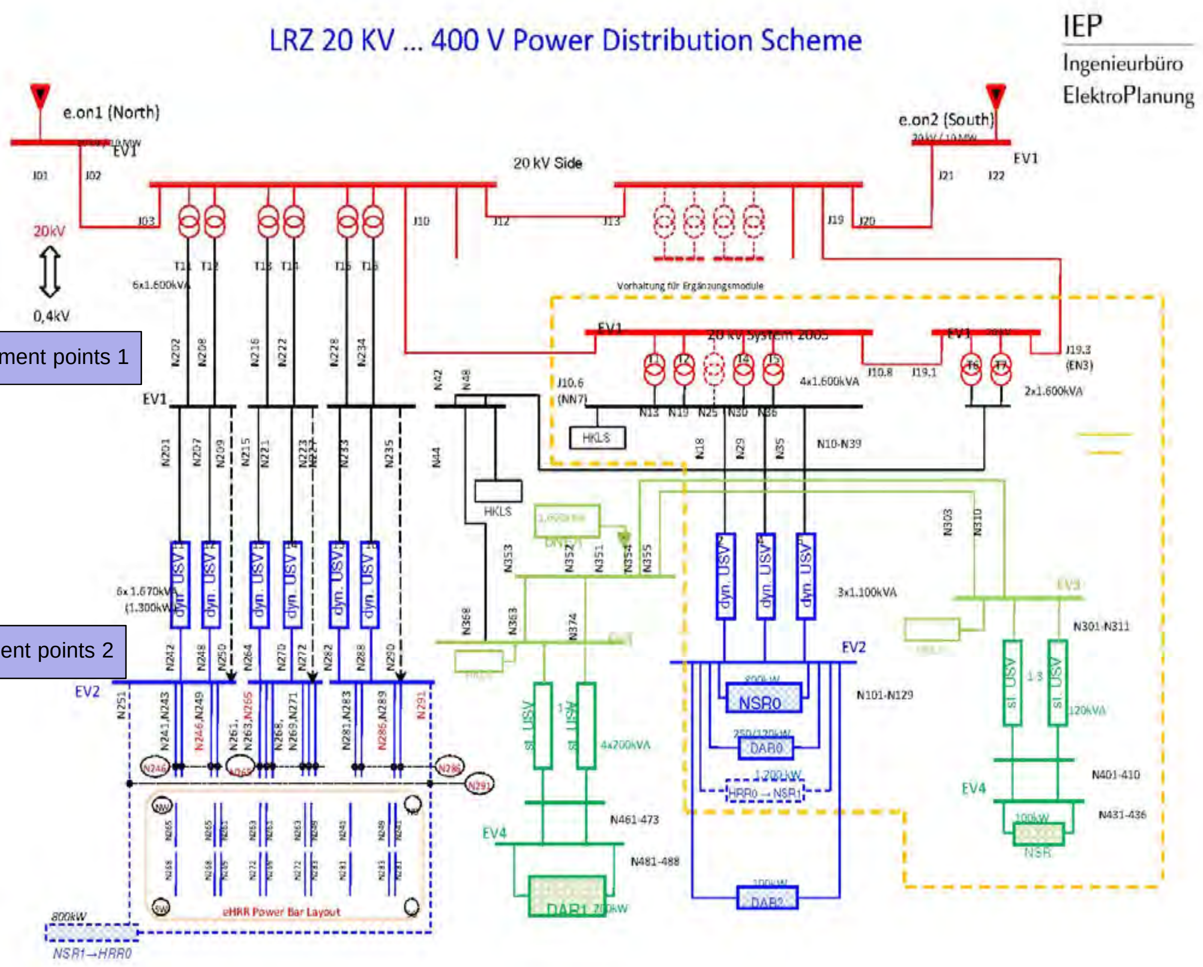
Charles Archer, Torsten Bloth, Achim Bömelburg, Ingmar Meijer, Steffen Waitz, IBM

SuperMUC Phase 1: Technical Highlights

- ❑ **3 PetaFlop/s Peak performance (9216 IBM System x iDataPlex M4 Direct Water Cooled nodes, 147456 Intel E5-2680 cores)**
- ❑ **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**

LRZ Infrastructure Power and Energy Measurement Points (1)

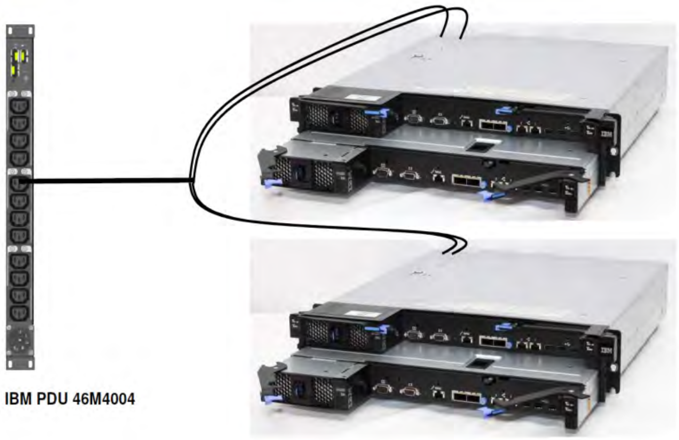


Measurement points 1

Measurement points 2

- 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)



Report and Analyze	Level 1	Level 2	Level 3
Aspect 1: requirements of measured values for ac measurement	1 instantaneous power measurement per second	1 instantaneous power measurement per second	continuously integrated total energy
Aspect 1: requirements of reported values for submission	one average power covering at least 20% of the run	time series of equally-spaced averaged power values	time series of equally spaced total energy values
Aspect 2: machine fraction	at least 1/64 of the machine or 1 kW	at least ? of the machine or 10 kW	whole machine
Aspect 3:	subsystems included	subsystems included	subsystems included
Aspect 3:	Point in power distribution where measurement is taken	Point in power distribution where measurement is taken	Point in power distribution where measurement is taken
required analyzed values for submission	core phase average power	core phase average power and whole application average power	core phase average power and whole application average power

- 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 ±5% accuracy over the entire range
- **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)

Linpack HPL run May 17,2012 – 2.582 PF

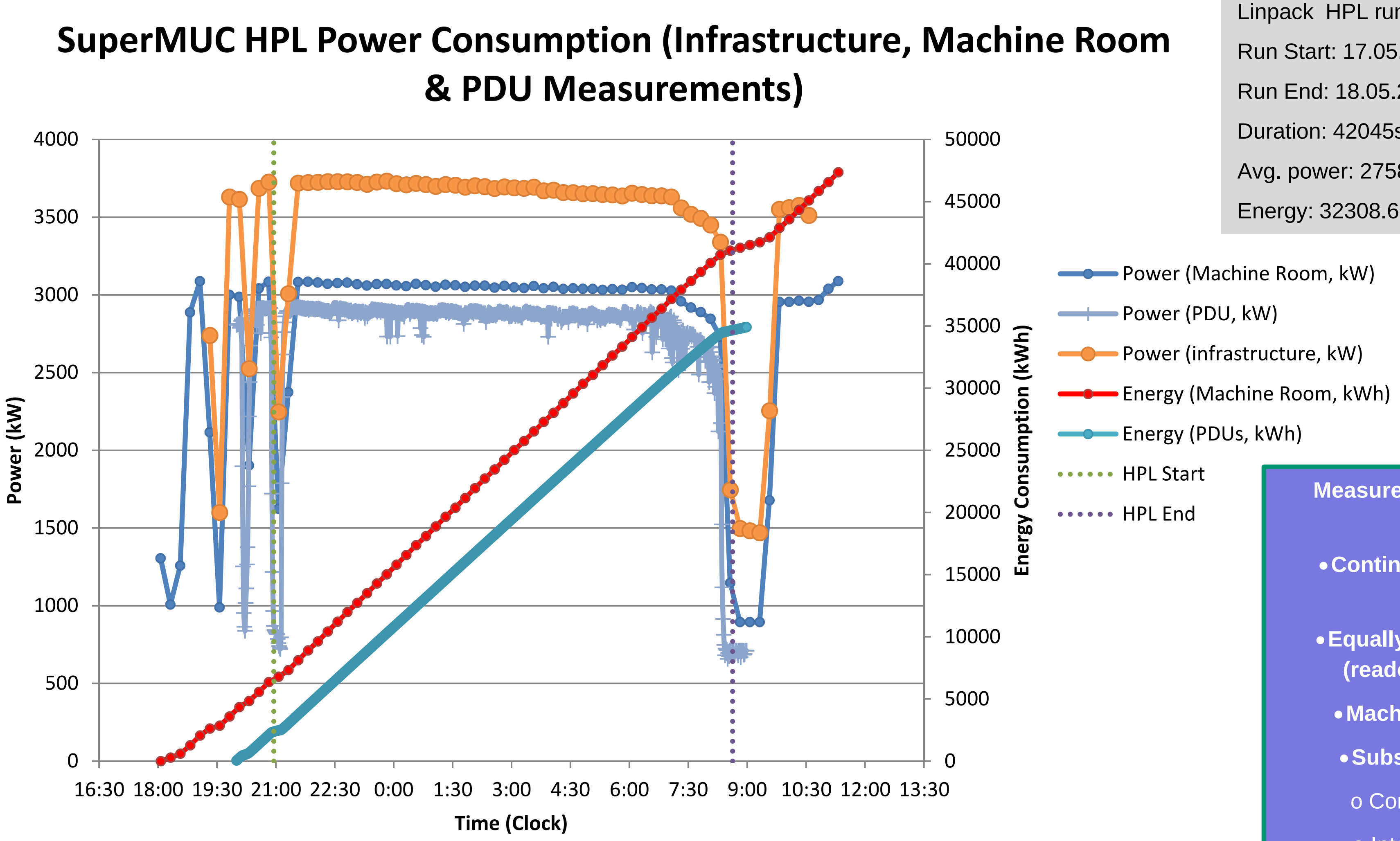
Run Start: 17.05.2012 20:56, 965,40 kW

Run End: 18.05.2012 08:37, 711,02 kW

Duration: 42045s or 11.68 hours

Avg. power: 2758.87 kW

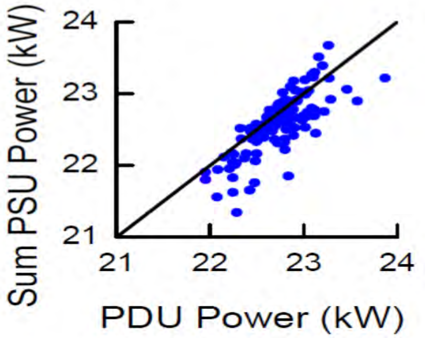
Energy: 32308.68 kWh



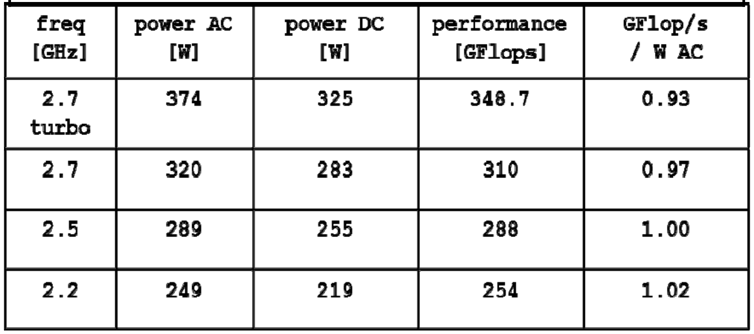
Measurement Notes (Energy PDUs):

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

Green500 Measurement Methodology: Roadblocks, Issues, Concerns, …



Energy efficiency *(single number in GFlops/Watt)*
9,380E-01 (PDU, 10 minutes resolution, whole run, without cooling)
9,359E-01 (PDU, 1 minutes resolution, whole run, without cooling)
9,305E-01 (PDU, 1 minutes resolution, whole run, cooling included)
8,871E-01 (machine room measurement, whole run)
7,296E-01 (infrastructure measurement, whole run)



freq [GHz]	power AC [W]	power DC [W]	performance [GFlops]	GFlop/s / W AC
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

- ❑ **Minimum measurement accuracy for L3 need to be defined (less than 5% ?)**
- ❑ **What needs to be measured for a L3 Green500 submission?**
  - ● Compute nodes only?

  or

  - ● all **system components** needed to run Linpack (e.g., communication network, fans, …)?
  - ● AC power consumption of system including all AC/DC conversion losses?

  or

  - ● DC power consumption of system (where AC/DC power conversion losses are added using a mathematical model)?
- ❑ **Measured energy efficiencies depend on**
  - ● System size → smaller is better
  - ● Processor frequency settings → lower is better
  - ● Memory type and speed, …. → slower is better
  - ● Machine room temperature → lower is better but not really more Green !
  - ● Node Bios settings, number of installed I/O adapters, …
  - ● Accuracy and time resolution of measurement equipment