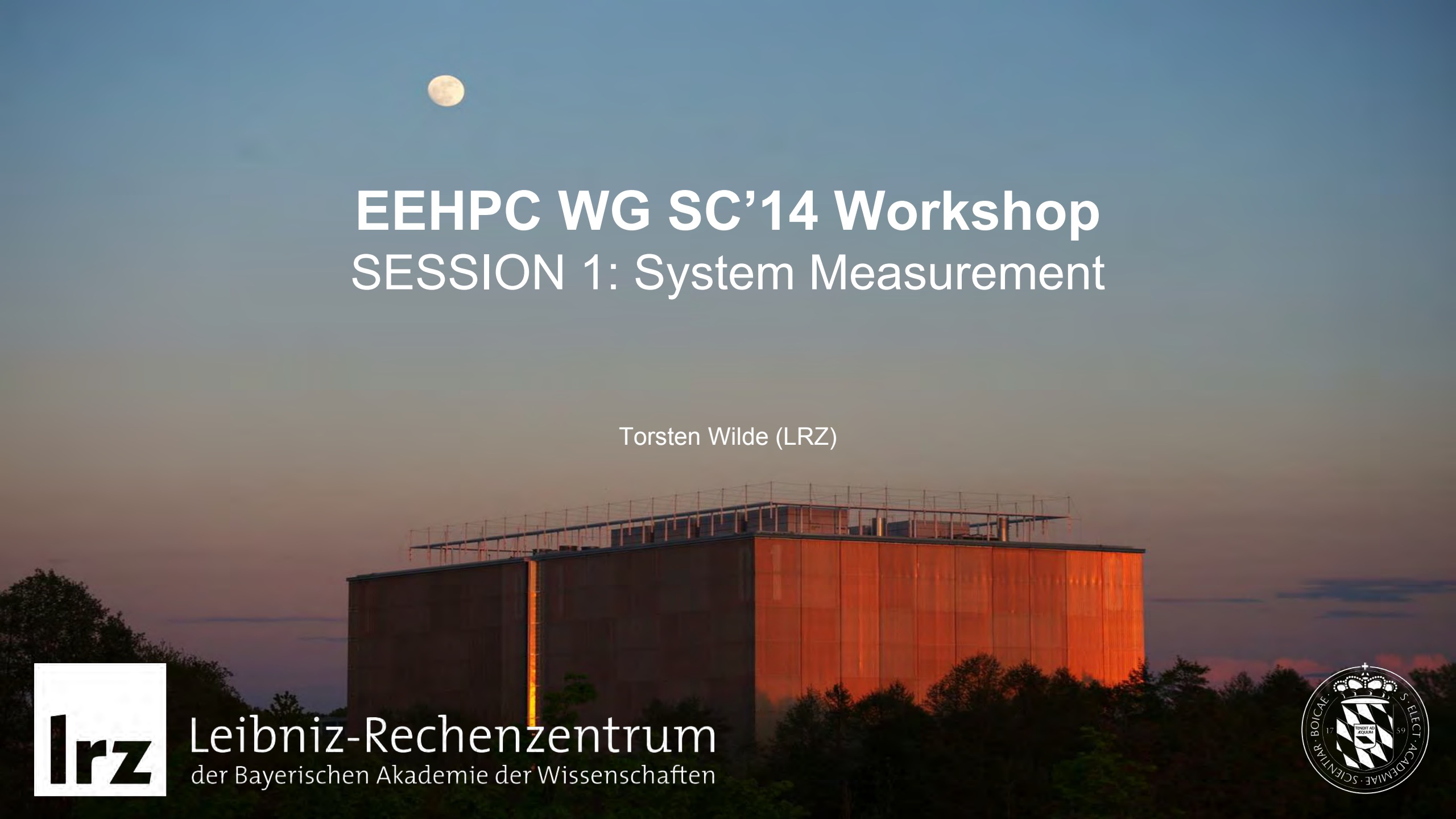

EEHPC WG SC'14 Workshop
SESSION 1: System Measurement

Torsten Wilde (LRZ)

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

Who we are

- EEHPC WG Sub Group:
  - Computing Systems
    - System Workload Power Measure Methodology
- Define a standard and accurate (high quality) measurement methodology to measure power and energy consumption of HPC systems
  - Get comparable results
  - Define what is included and how it is measured
- Support the Green500 list
- Support energy efficient HPC

Outline

- Motivation
- What to expect from the session
  - Node variability
  - Interconnect
  - Workload phases
- Q&A

Why are we here

- Power consumption and facility costs of HPC are increasing.
  - “Can only improve what you can measure”
- What is needed?
  - Converge on a common basis for:
    - METHODOLOGIES
    - WORKLOADS
    - METRICS
  - for energy-efficient supercomputing, so we can make progress towards solutions.

Power/Energy Measurement Methodology

SIMOPEK

	Aspect 1: Time Fraction & Granularity	Aspect 2: Machine Fraction	Aspect 3: Subsystems Measured
Level 1	20% of run: 1 average power measurement	(larger of) 1/64 of machine or 1kW	[Y] Compute nodes [ ] Interconnect net [ ]Storage [ ]Storage Network [ ]Login/Head nodes
Level 2	100% of run: at least 100 average power measurements	(Larger of) 1/8 of machine or 10kW	
Level 3	100% of run: at least 100 running total energy measurements	Whole machine	

Measurement Challenges Example (SuperMUC Lev1-3)

Efficiency Level	Mflops/Watt full run	Mflops/Watt core phase
L1 (+- 5%)	1055	1055
L2 (>10kW) (+- 5%)	1011	917
L2 (>1/8) (+- 5%)	994	900
L3 (+- 0.5%)	887	855

"A power-measurement methodology for large scale, high performance computing", Runner-up Best Paper Award, Proceedings of the 5th ACM/SPEC international conference on Performance engineering, Dublin, Ireland 2014

Level 3 includes:

- Compute nodes
- Interconnect network
- GPFS mass storage systems
- Storage network
- Head/login and management nodes
- Internal warm water cooling system (machine room internal cooling such as water pumps, heat exchangers, etc)
- PDU power distribution losses

The need to improve the measurement methodology

SIMOPEK

- Level 1, 2, and 3 are not comparable
  - How can we change Level 1 and 2 requirements to reach a relative “high quality” result
- Need to question legacy assumptions:
  - Measuring a small part of a system and scaling it up doesn‘t introduce to much of an error
  - The power draw of the interconnect fabric is not significant when compared to the compute system
  - The workload phase of HPL will look similar on all HPC systems

What to expect

SIMOPEK

3 Challenges

- Node variability
  - Daniel Hackenberg, TU Dresden, Germany
- Interconnect
  - Robin Goldstone, Lawrence Livermore National Laboratory
- Workload phase
  - Tom Scogland, Lawrence Livermore National Laboratory and Green500

Summary

- Investigate statistical requirement for number of nodes for level 1 and 2
  - Cherry picking still possible
- Network power is significant
  - Topology advantage not seen for HPL (Green500)
  - Identify metric and benchmark
- „Workload Phase“ needs to be refined and MPI rank variability is a concern

- Easiest might be to just measure everything

- Contact us if you are interested in helping us

Thank You! Questions?