Erich Strohmaier, Lawrence Berkeley National Laboratory and TOP500
Wu Feng, Virginia Tech and Green500

with
Natalie Bates, EE HPC Working Group,
Michael Patterson, Intel and The Green Grid
And others on the Compute System Metrics Team

METHODOLOGIES FOR MEASURING POWER

BoF: Green500 and Its Continuing Evolution
@ SC'14 in New Orleans, LA

UNIFY AND IMPROVE METHODOLOGY

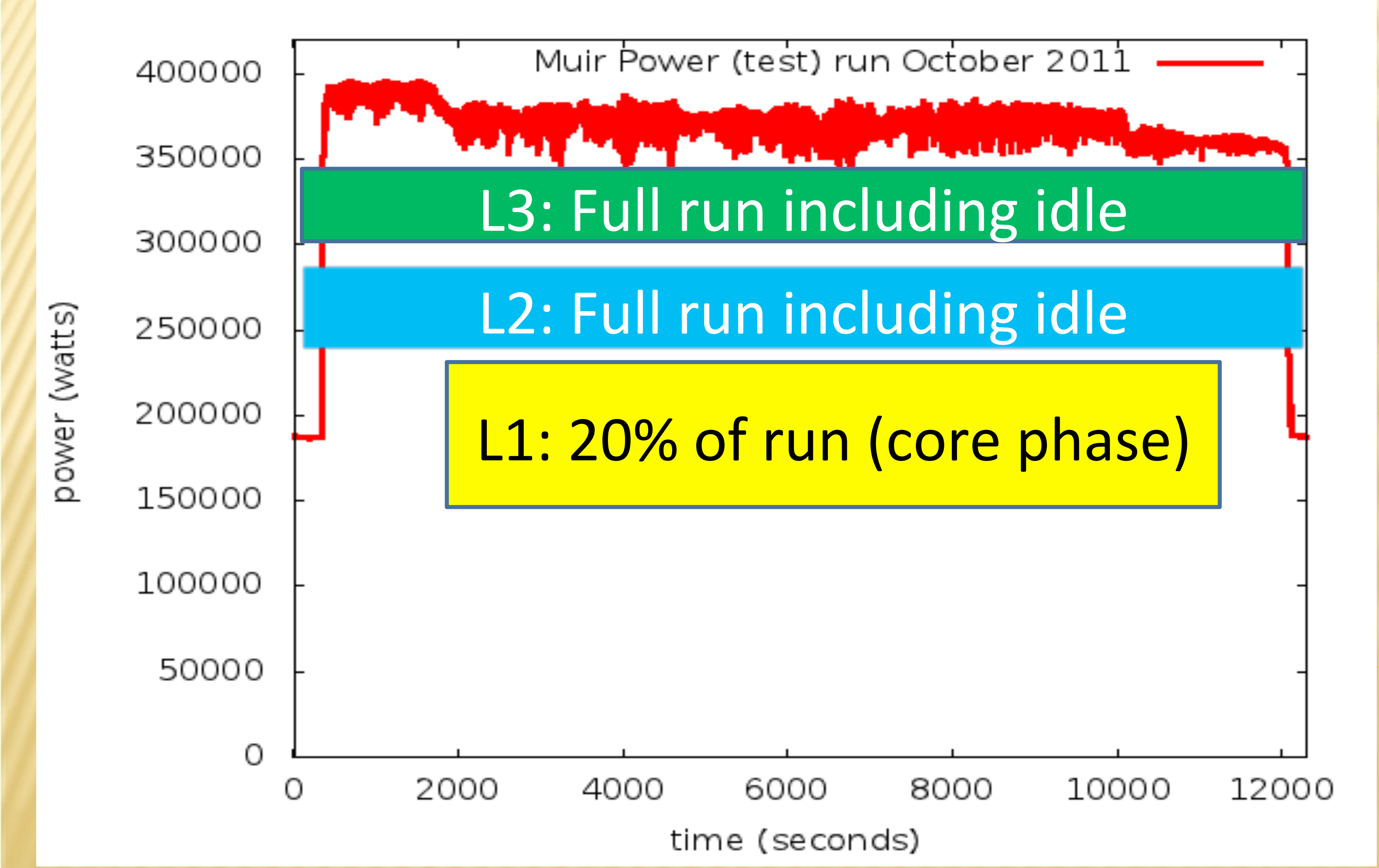
- Issues and concerns with power-measurement methodology
  - Variation in start/stop times as well as sampling rates
  - Node, rack or system level measurements
  - What to include in the measurement (e.g., integrated cooling)
- A collaboration between EE HPC WG, Green500, Green Grid, and Top500 to address these issues and concerns

ADD QUALITY LEVELS AND REFINE ASPECTS

- Three quality levels (currently):
  - Level 1 (L1): basic measurement
  - Level 2 (L2): reasonable effort
  - Level 3 (L3): current best

- Four measurement aspects for each level:
  - Aspect 1: frequency and time extent of measurement
  - Aspect 2: system fraction actually measured
  - Aspect 3: subsystems included
  - Aspect 4: power measurement location

Aspect 1: Time Extent



L3: Full run including idle

L2: Full run including idle

L1: 20% of run (core phase)

Aspect 1: Sampled Data Frequency

Level 3: (L3)

- “Continuously integrated” energy (≥ 120 samples per second)

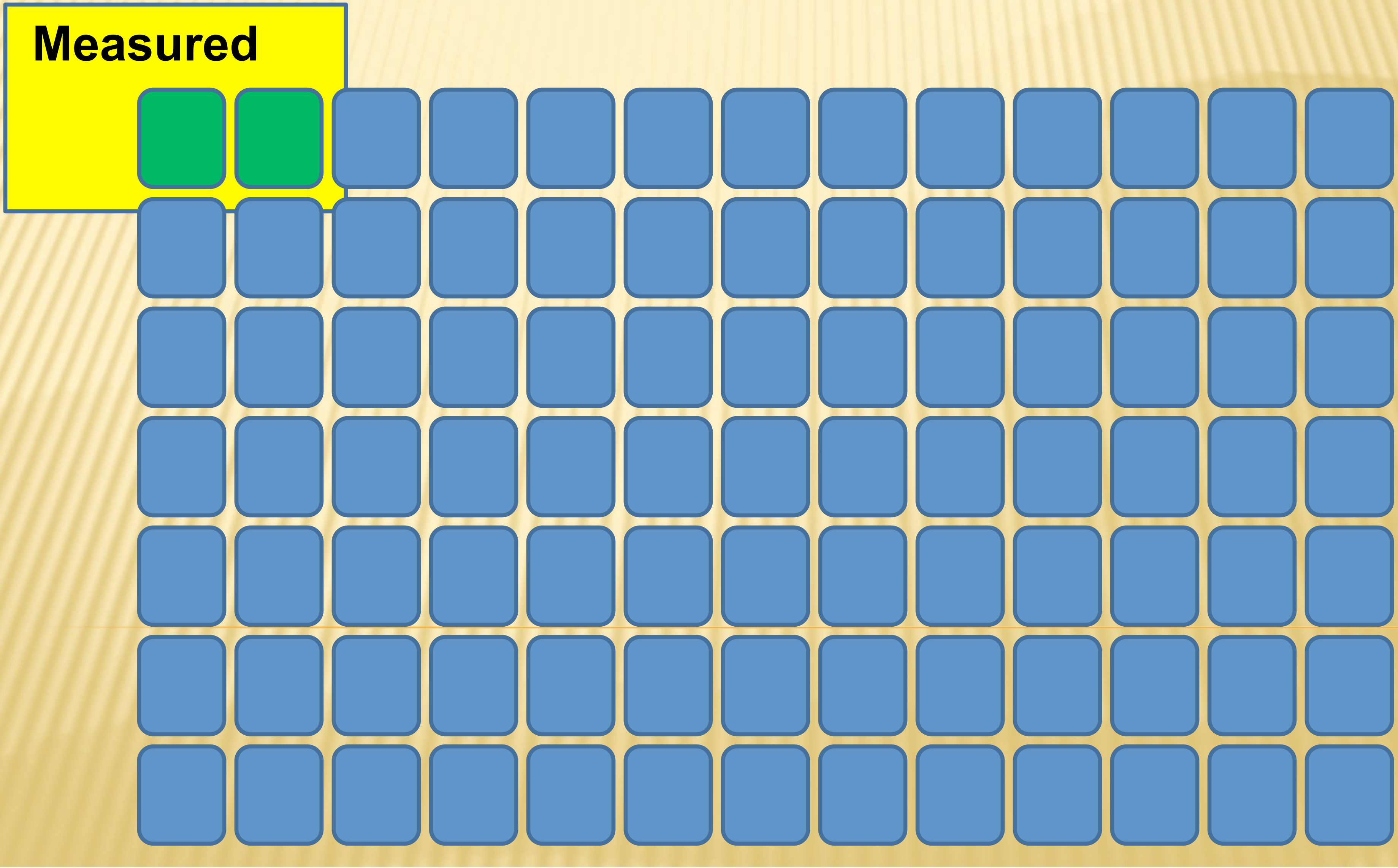
Level 1 and Level 2 (L1 and L2)

- Average power at least once per second

These are *sampling* rates.

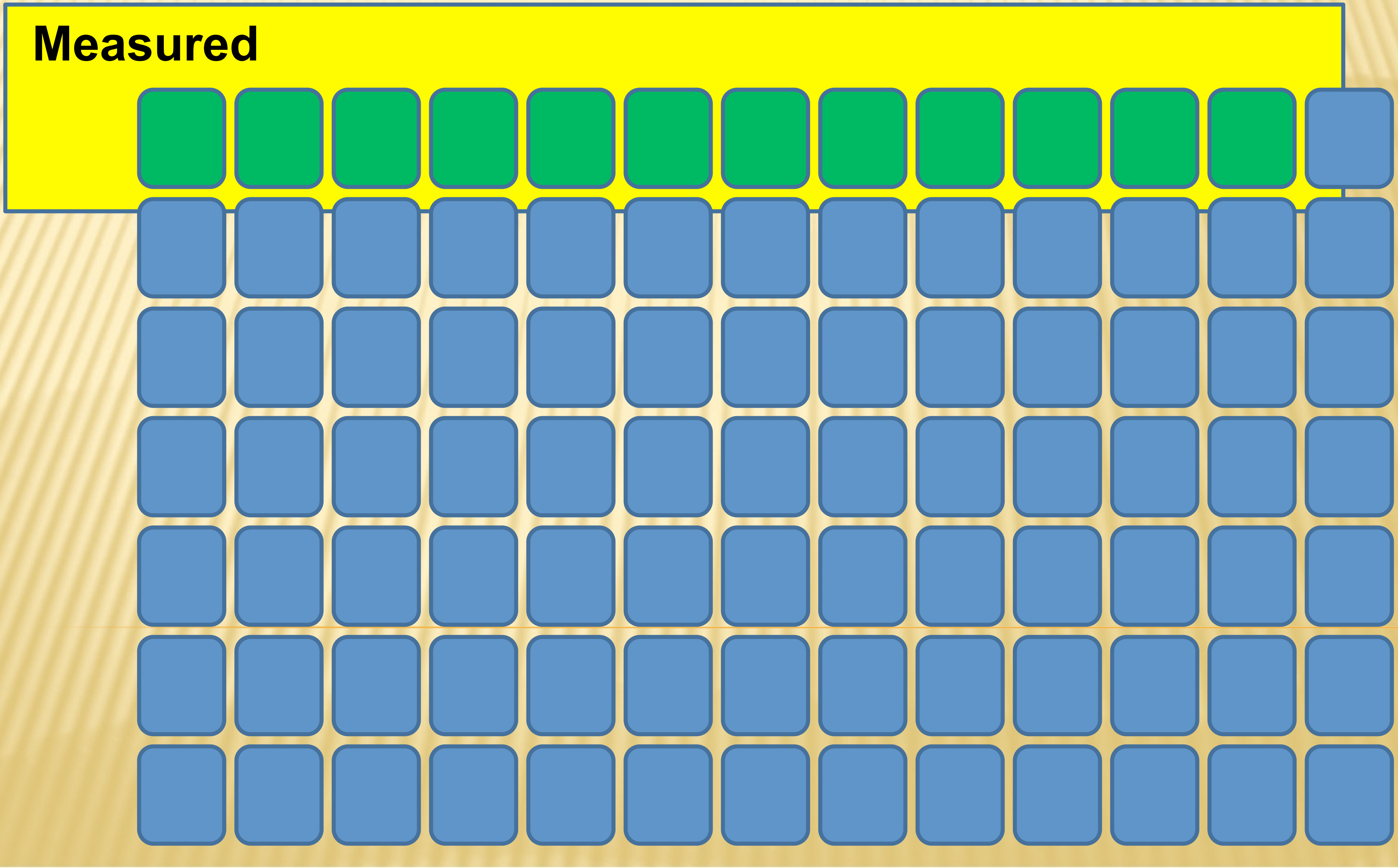
Data at this rate is typically not seen directly, it is internal to the device.

Aspect 2: Machine Fraction
L1: at least 1/64 or 1 kW



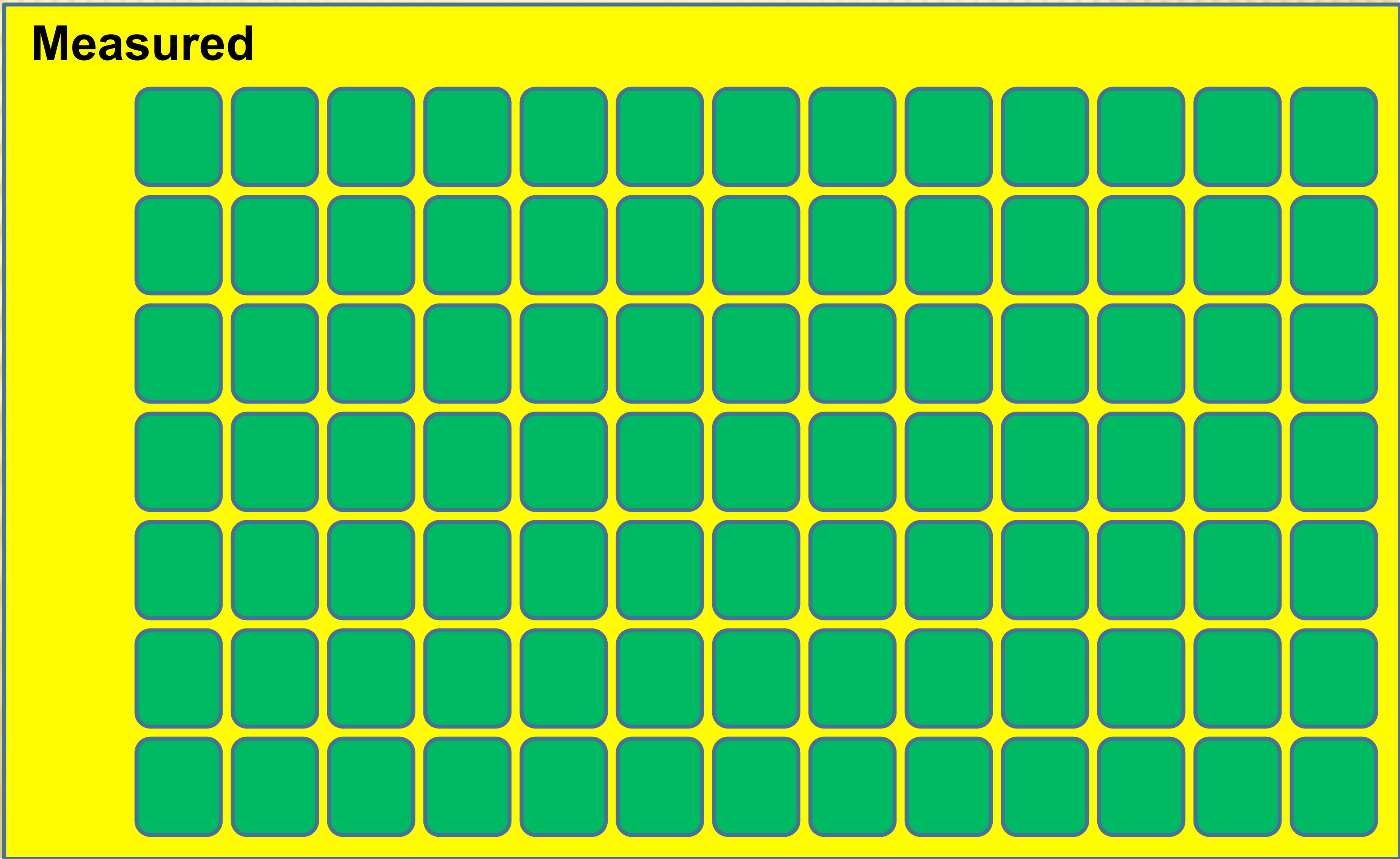
Measured

Aspect 2: Machine Fraction
L2: at least 1/8 or 10 kW



Measured

Aspect 2: Machine Fraction
L3: whole machine



Measured

Aspect 3: Subsystem Inclusion

General Philosophy

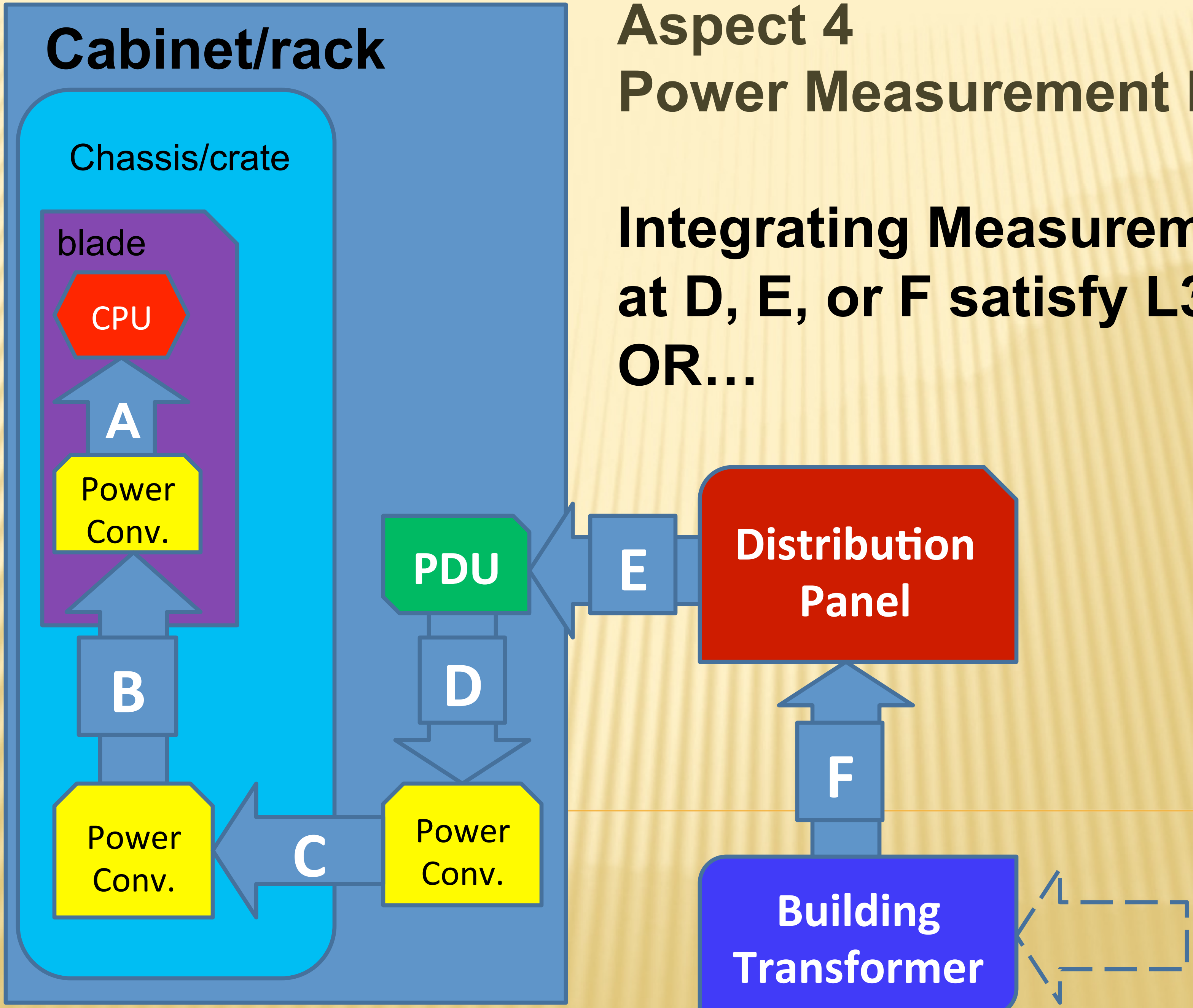
- Include all parts of computational system that participate in the workload

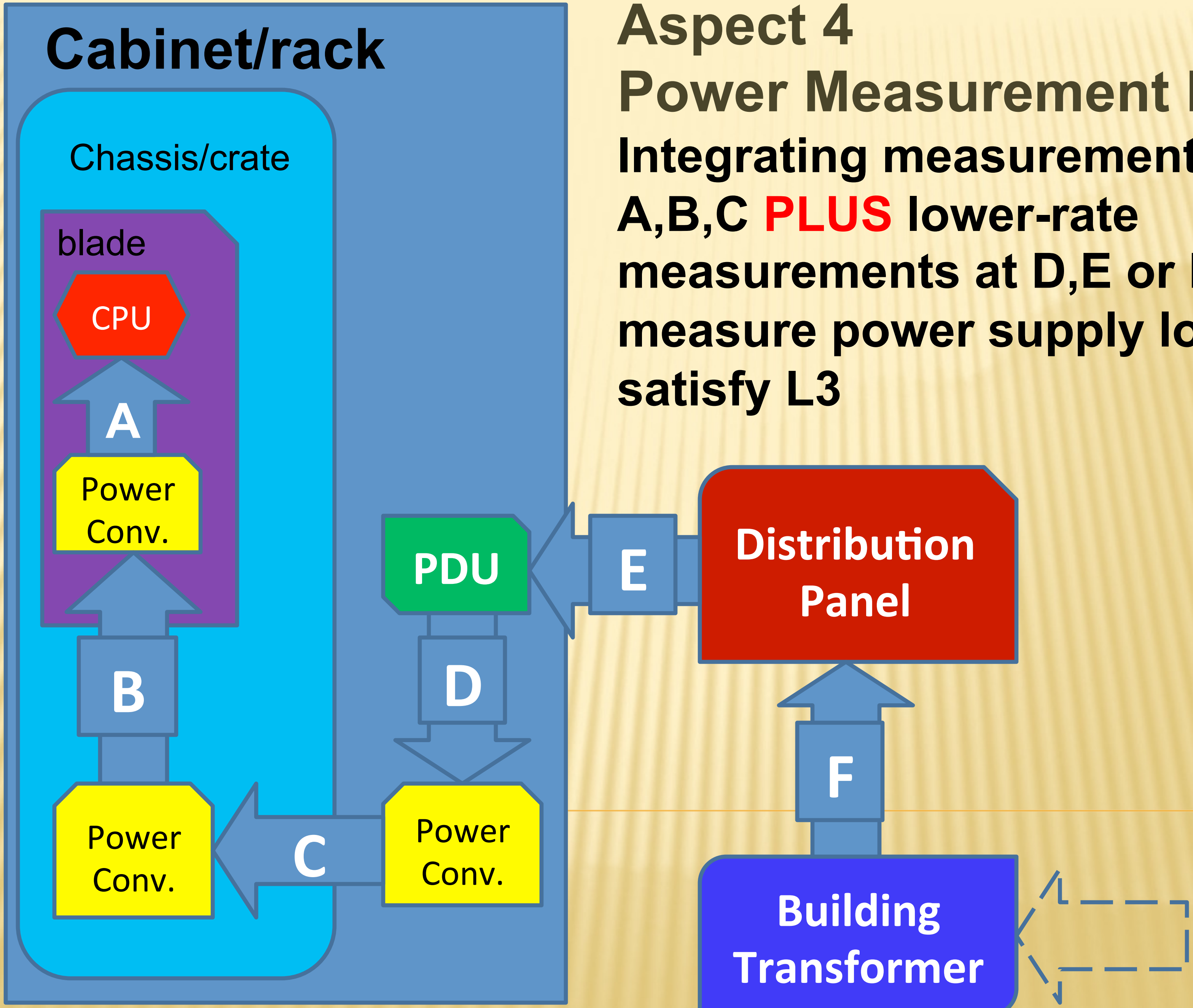
What Must Be Included?

- Processors, memory, cooling power internal to the machine (fans, etc.)
- Internal interconnect network
- Login/compile nodes

Aspect 4
Power Measurement Point

Integrating Measurements at D, E, or F satisfy L3 OR…



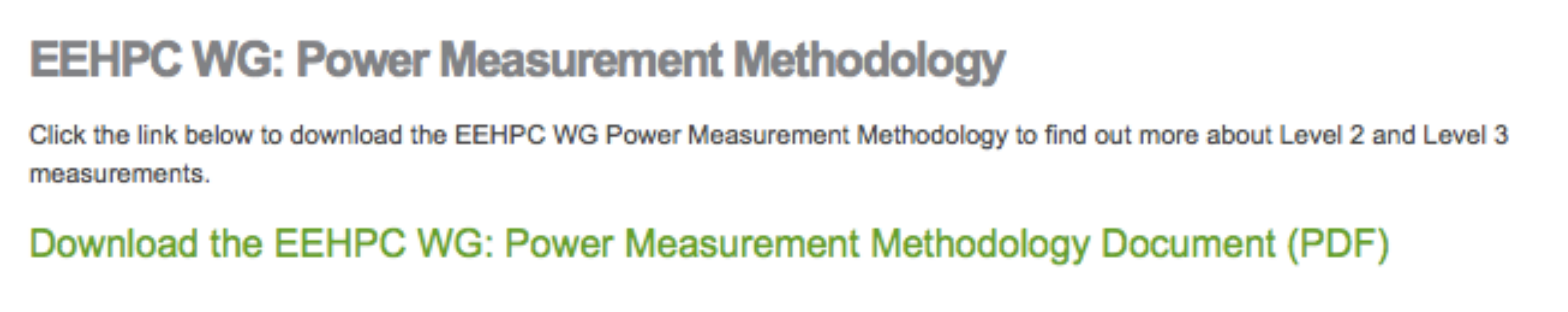


Aspect 4
Power Measurement Point:

Integrating measurements at A,B,C PLUS lower-rate measurements at D,E or F (to measure power supply losses) satisfy L3

WHERE TO FIND THE METHODOLOGY

THE GREEN 500™



EEHPC WG: Power Measurement Methodology

Click the link below to download the EEHPC WG Power Measurement Methodology to find out more about Level 2 and Level 3 measurements.

Download the EEHPC WG: Power Measurement Methodology Document (PDF)

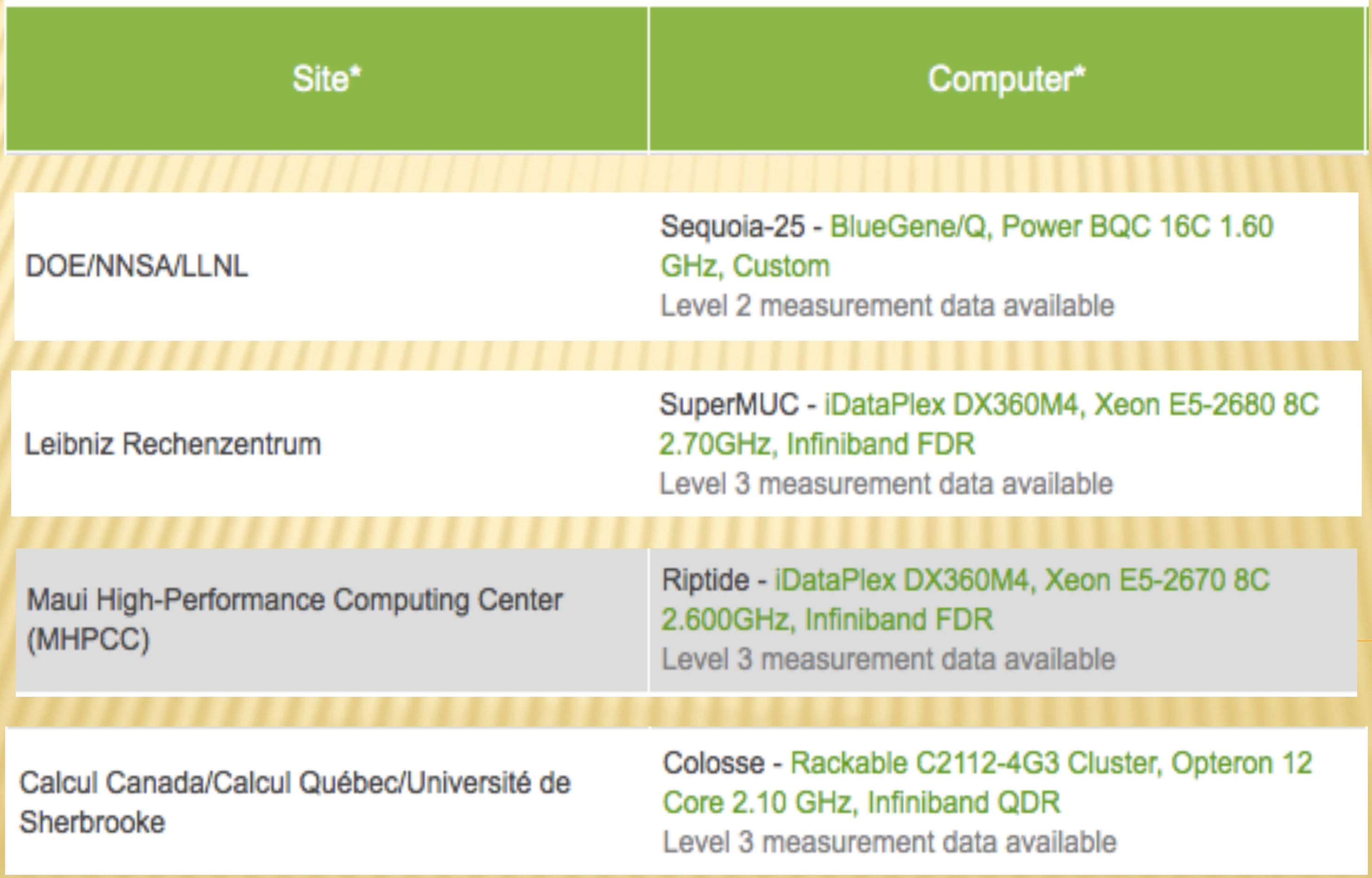
GREEN500 RELEASES NEW METHODOLOGY (2013)

- Green500 accepts higher-precision measurements, denoted as Level 2 and 3
- *“Higher quality measurements… provide much better picture of the real-world costs… as well as a more in-depth picture of how the system handles a Linpack run.”* Green500 Press Release

DEBUT OF NEW METHODOLOGIES

(June 2013 Green500 List)

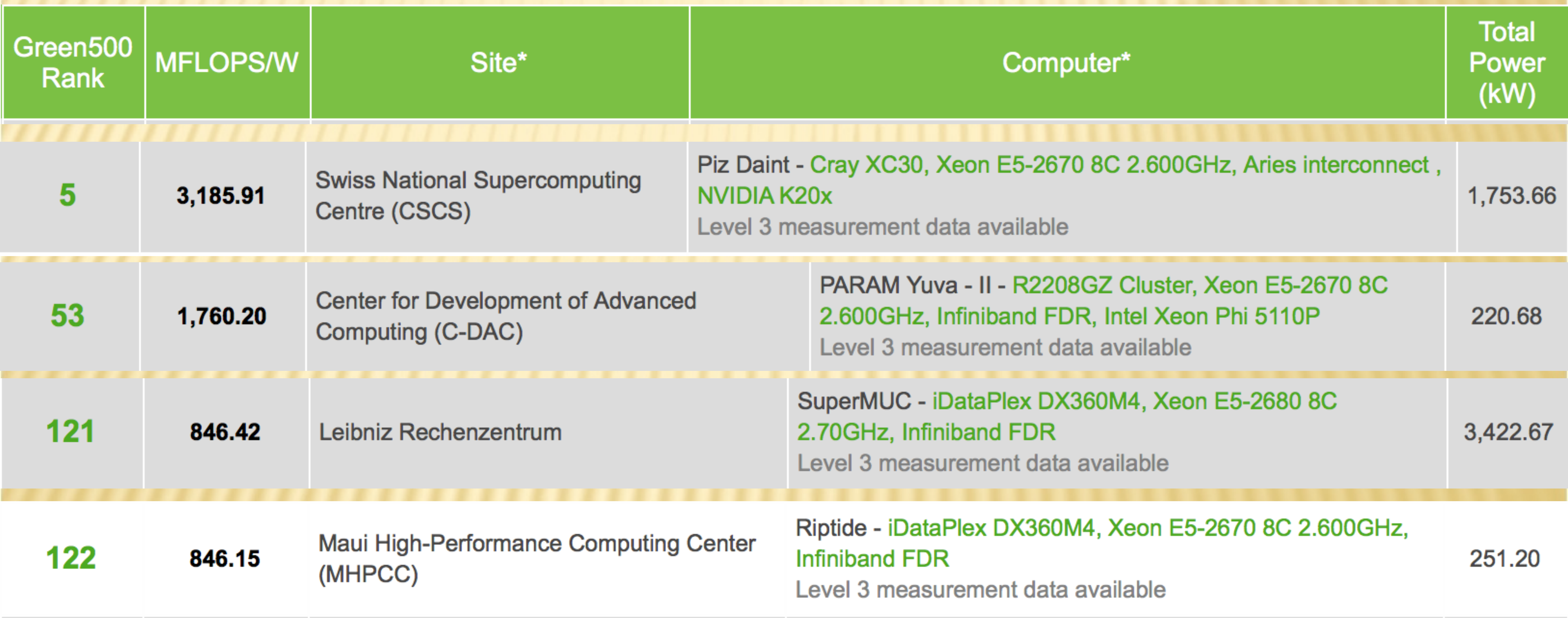
Level 2/3 measurement data available …



Site*	Computer*
DOE/NNSA/LLNL	Sequoia-25 - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom Level 2 measurement data available
Leibniz Rechenzentrum	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR Level 3 measurement data available
Maui High-Performance Computing Center (MHPCC)	Riptide - iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR Level 3 measurement data available
Calcul Canada/Calcul Québec/Université de Sherbrooke	Colosse - Rackable C2112-4G3 Cluster, Opteron 12 Core 2.10 GHz, Infiniband QDR Level 3 measurement data available

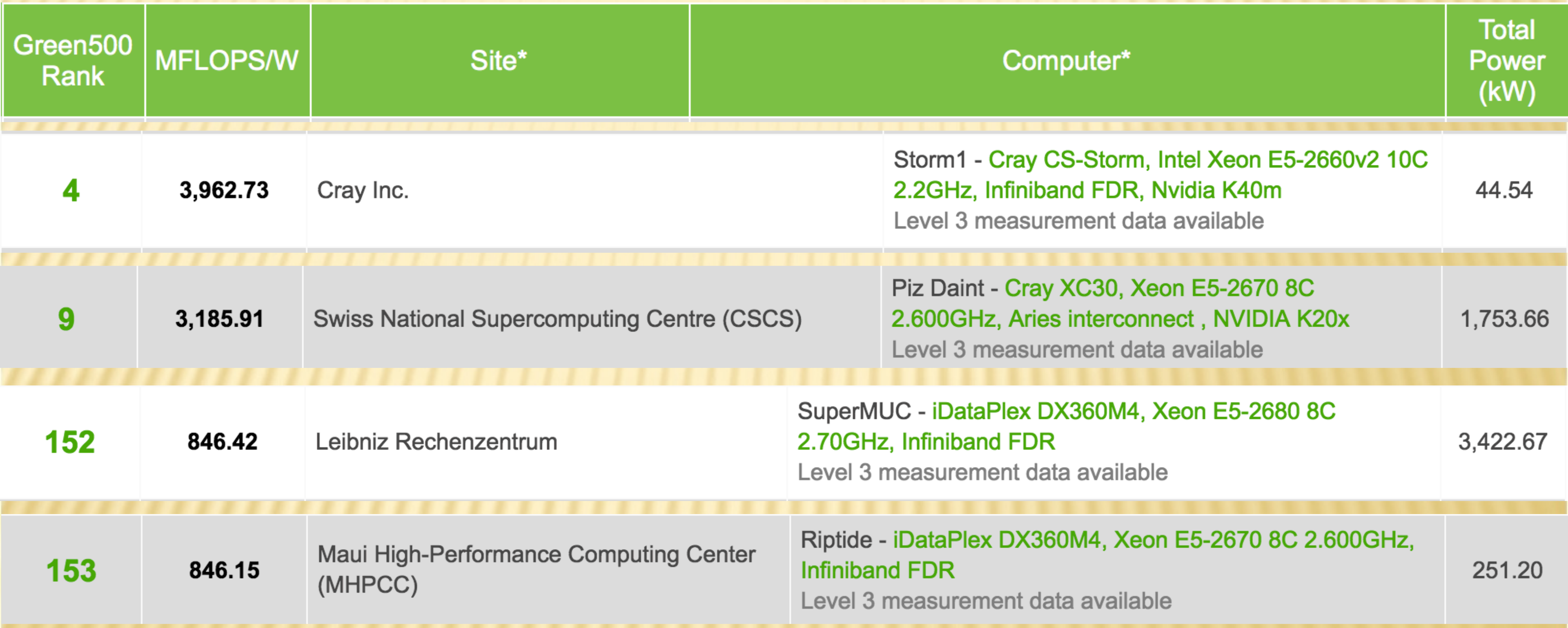
NEW METHODOLOGIES: ONE YEAR LATER

(June 2014 Green500 List)



Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
5	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
53	1,760.20	Center for Development of Advanced Computing (C-DAC)	PARAM Yuva - II - R2208GZ Cluster, Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P Level 3 measurement data available	220.68
121	846.42	Leibniz Rechenzentrum	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR Level 3 measurement data available	3,422.67
122	846.15	Maui High-Performance Computing Center (MHPCC)	Riptide - iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR Level 3 measurement data available	251.20

NEW METHODOLOGIES: NOW



Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
4	3,962.73	Cray Inc.	Storm1 - Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40m Level 3 measurement data available	44.54
9	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
152	846.42	Leibniz Rechenzentrum	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR Level 3 measurement data available	3,422.67
153	846.15	Maui High-Performance Computing Center (MHPCC)	Riptide - iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR Level 3 measurement data available	251.20

Early Adopters and Testers

- Lawrence Livermore National Laboratory
- Leibniz Supercomputing Center
- Oak Ridge National Laboratory
- Argonne National Laboratory
- Universite Laval, Calcul Quebec, Compute Canada
- University of Jaume
- University of Tennessee
- CEA
- Center for Development of Advanced Computing (C-DAC)
- National Center for Atmospheric Research
- Maui High Performance Computing Center
- Swiss National Supercomputing Center (CSCS)

ISSUES TO RESOLVE

- *Refine* ***METHODOLOGY***
  - *System boundary, e.g., file system*
  - *Environmentals, e.g., “either-or -> hybrid” air+liquid cooling*
  - *Measurement instrument specification: accuracy and precision*
- *Identify* ***WORKLOADS*** *for exercising other sub-systems; e.g., memory, storage, I/O*
- *Still need to decide upon* ***METRICS***
  - *Classes of systems (e.g., Top50, Little500, technologies)*
  - *Multiple metrics or a single index*