

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

SETTING TRENDS FOR ENERGY-EFFICIENT SUPERCOMPUTING

METHODOLOGY: Erich Strohmaier

UNIFY AND IMPROVE METHODOLOGY

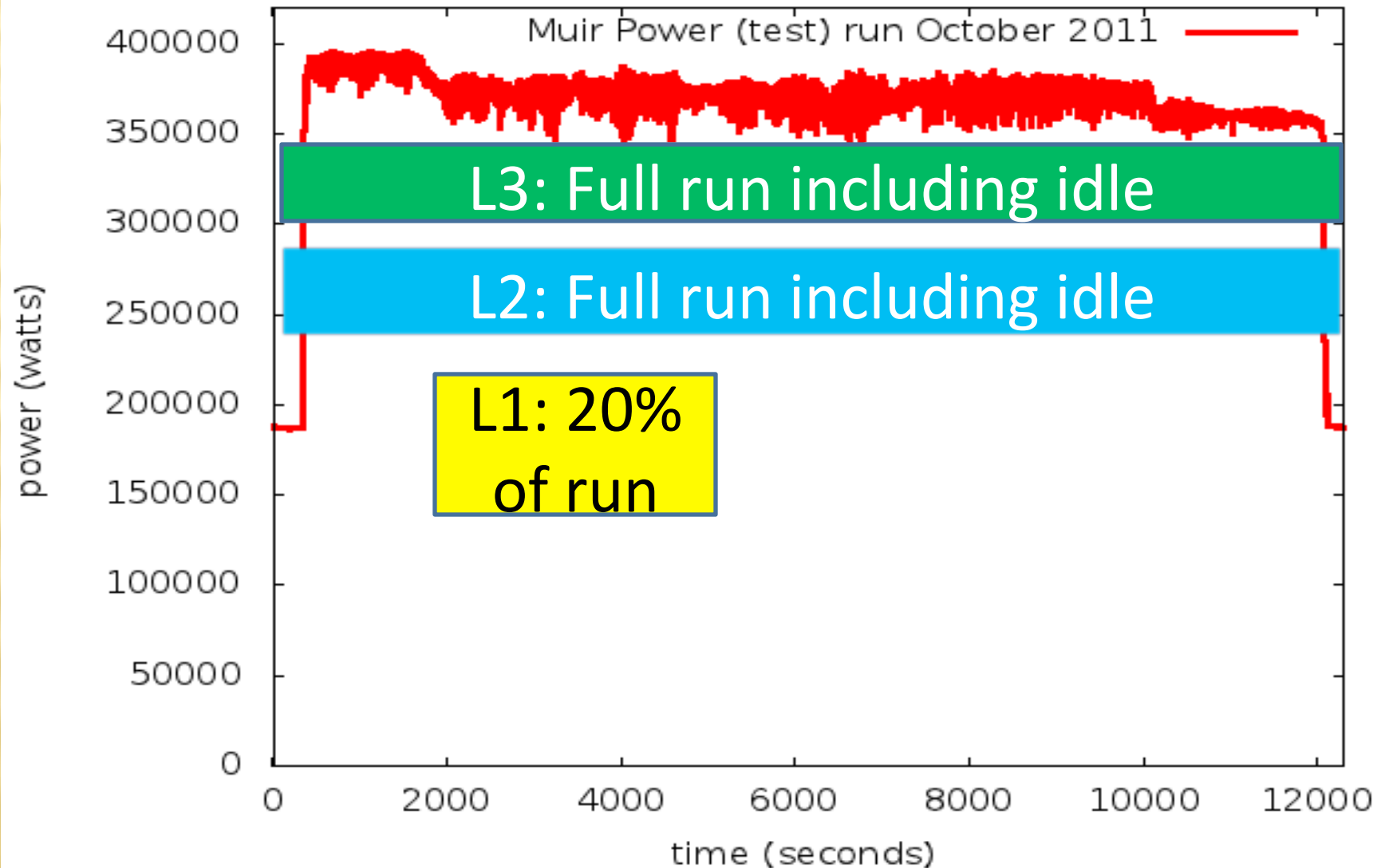
- + Green500 & TOP500 power-measurement methodology issues and concerns
 - + Variation in start/stop times as well as sampling rates
 - + Node, rack or system level measurements
 - + What to include in the measurement (e.g., integrated cooling)
 - + Flexible, but compromises consistency between submissions
-

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



Aspect 1: Sampled Data Frequency

Level 3: (L3)

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

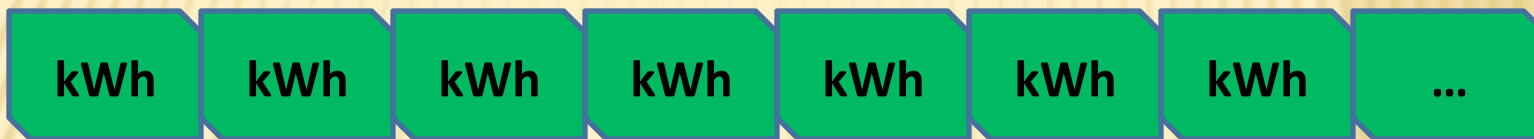
Level 1 and Level 2 (L1 and L2)

- Average power at least once per second

These are *sampling* rates. Data at this rate is typically not seen directly, it's internal to the device.

Aspect 1: Reported Data Requirements

L3: at least 10 reported integrated energy values



L2: at least 10 averaged power values



L1: at least one averaged power value



L1: at least $1/64$ or 1 kW

A 7x13 grid of blue squares. The top-left corner, consisting of the first two squares of the first row, is highlighted in yellow. The rest of the grid is composed of blue squares.

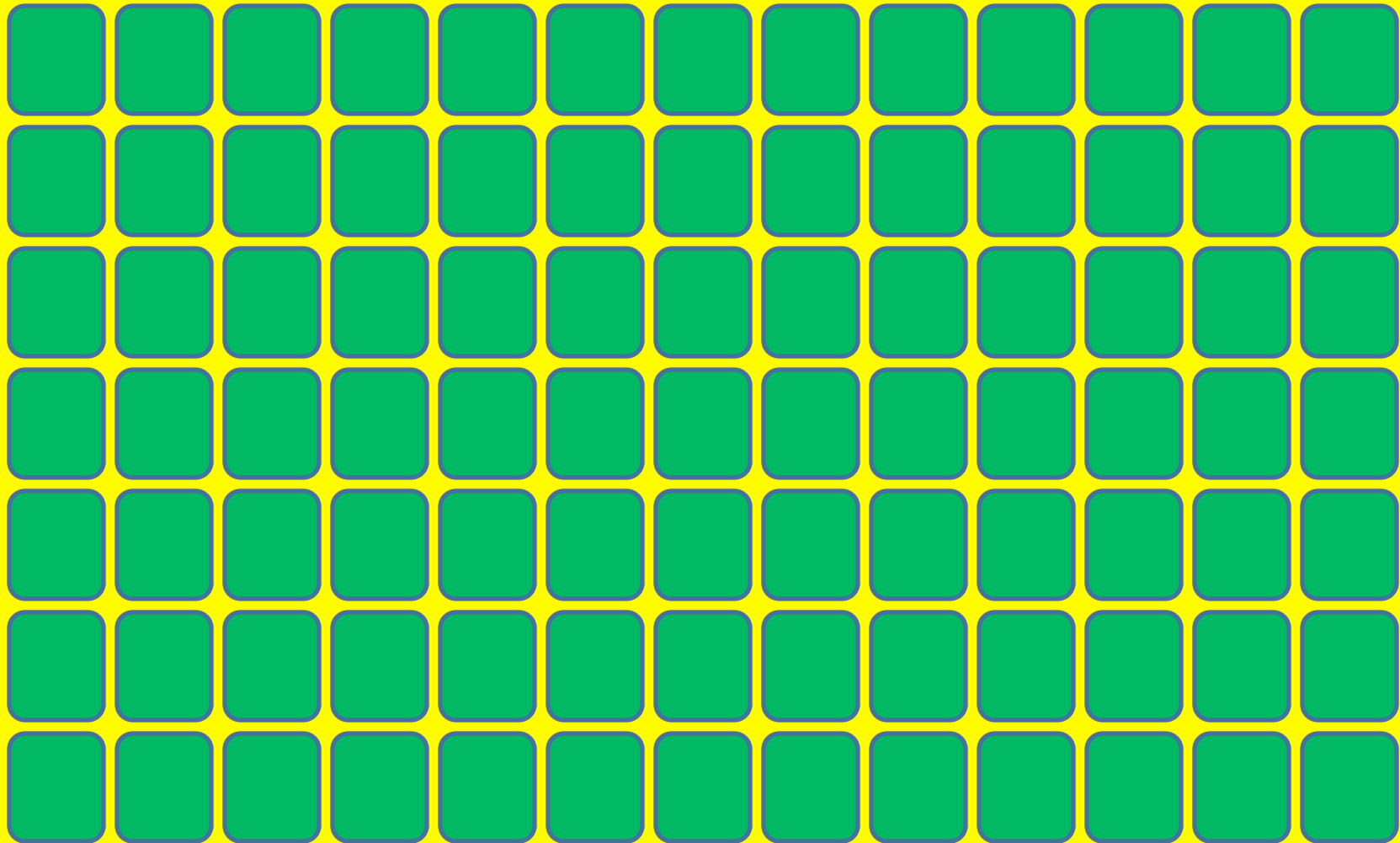
L2: at least 1/8 or 10 kW

A 7x14 grid of colored squares. The top row contains 12 green squares followed by 2 blue squares. The remaining 6 rows consist of 14 blue squares each.

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

Must include:

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

Cabinet/rack

Chassis/crate

blade

CPU

A

Power
Conv.

B

Power
Conv.

C

PDU

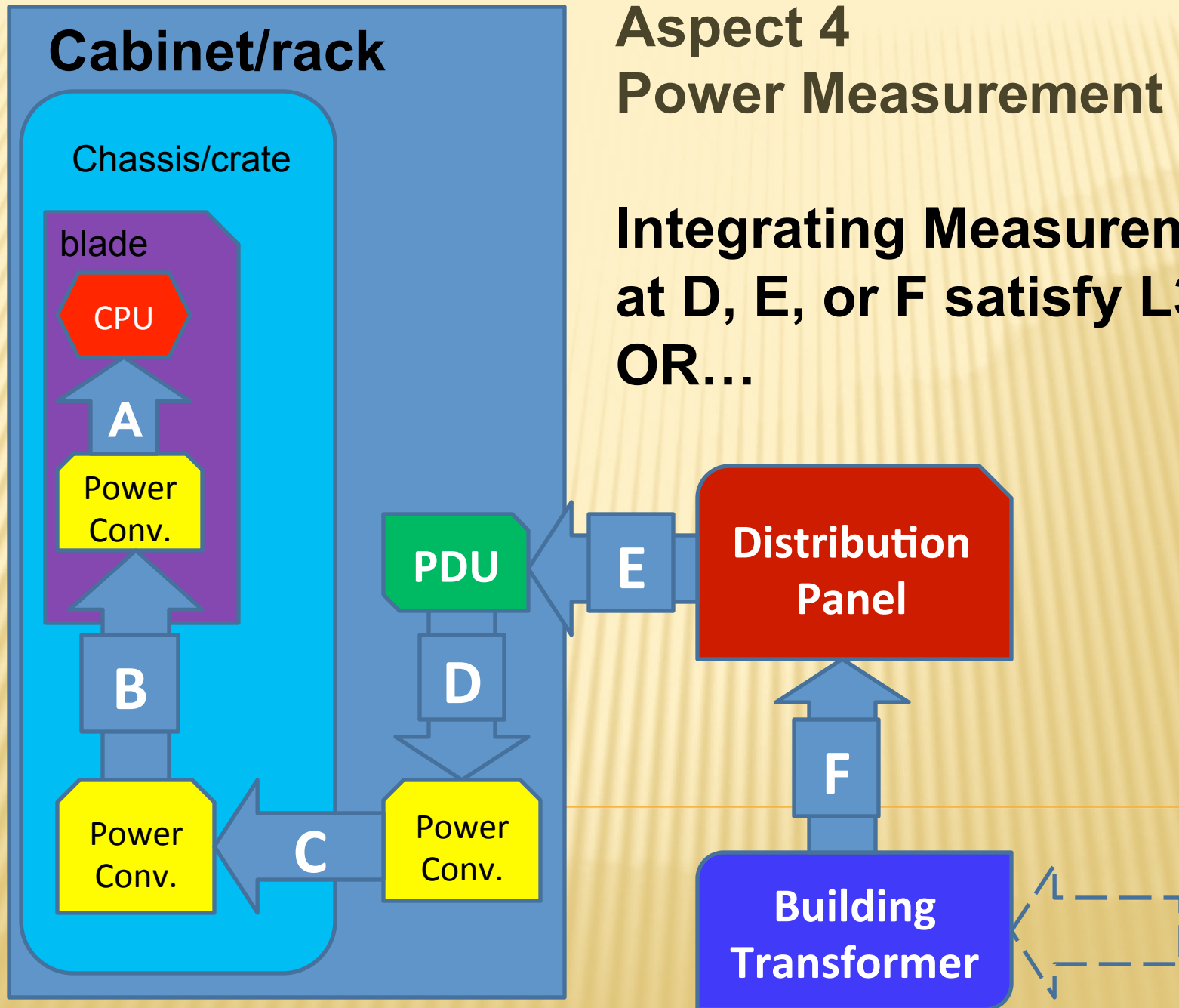
D

Power
Conv.

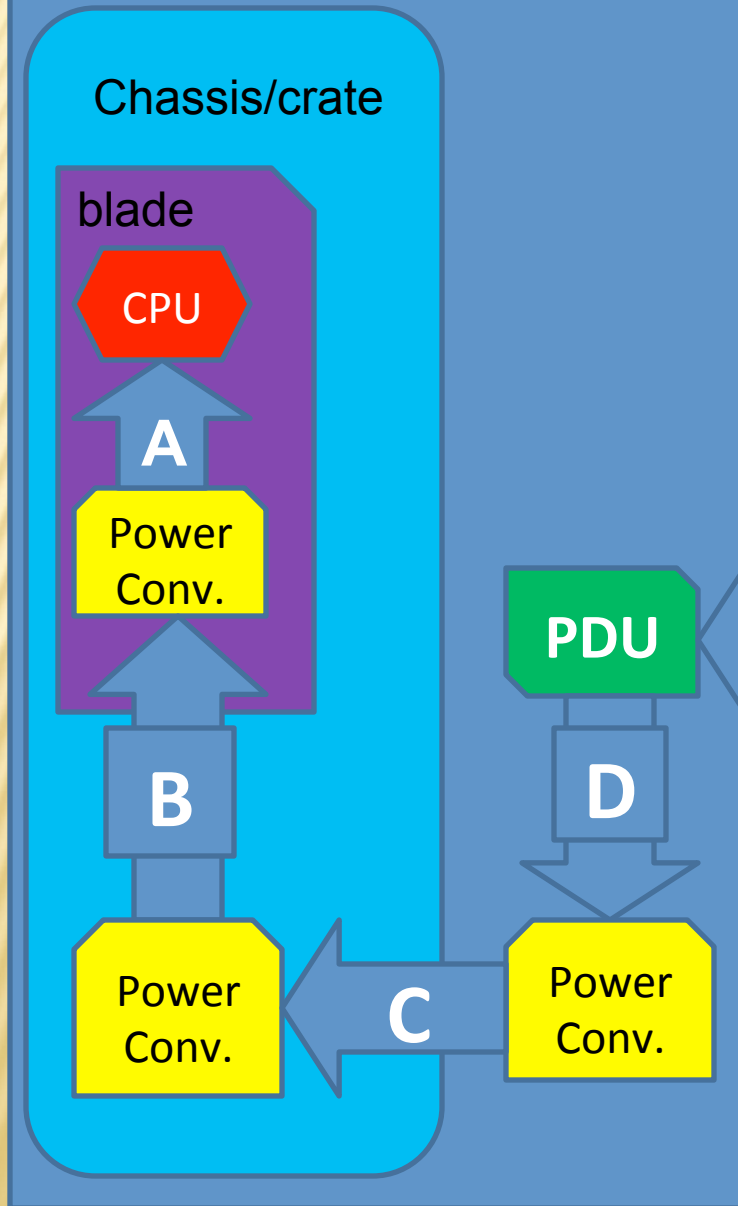
Aspect 4

Power Measurement Point

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



Cabinet/rack



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

ADOPTION: Wu Feng

WHERE TO FIND THE METHODOLOGY



<http://www.green500.org/resources>

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 \(PDF\)](#)

JUNE 2013:

GREEN500 RELEASES NEW METHODOLOGY

- ✗ Green500 accepts higher-precision measurements, denoted as Level 2 and 3
 - ✗ Supplement to
 - ✗ Derived Level 0
 - ✗ Measured Level 1
- ✗ *“Higher quality measurements... provide a 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

Early Adopters and Testers

- ✗ Lawrence Livermore National Laboratory
- ✗ Leibniz Supercomputing Center (LRZ)
- ✗ Oak Ridge National Laboratory
- ✗ Argonne National Laboratory
- ✗ Universite Laval, Calcul Quebec, Compute Canada
- ✗ University of Jaume
- ✗ University of Tennessee
- ✗ Commissariat a l'Energie Atomique (CEA)
- ✗ National Center for Atmospheric Research (NCAR)
- ✗ Maui High Performance Computing Center
- ✗ Universite Sherbrooke, Calcul Quebec, Compute Canada
- ✗ Center for Development of Advanced Computing (C-DAC)
- ✗ Swiss National Supercomputing Center (CSCS)

JUNE 2013 GREEN500 LIST:

Level 2 and 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

Progress Since June 2013 Green500 List

- ✖ June 2013 Green500 List
 - ✖ Four (4) L2/L3 submissions
- ✖ November 2013 Green500 List
 - ✖ Two (2) additional L2/L3 submissions
 - ✖ L3: Center for Development of Advanced Computing (C-DAC)
 - ✖ L3: Swiss National Supercomputing Center (CSCS)
- ✖ Paper describing the methodology
 - ✖ A power-measurement methodology for large-scale, high-performance computing (under review)

ISSUES TO RESOLVE

- ✗ Refine *methodology*
 - ✗ System boundary, e.g., file system
 - ✗ Environmentals, e.g., hybrid air+liquid cooling vs. either-or
 - ✗ Measurement instrument spec: accuracy and precision
- ✗ Identify *workloads* for exercising other sub-systems; e.g., memory, storage, I/O
- ✗ Decide upon *metrics*
 - ✗ Classes of systems (e.g., Top50, Little500, technologies)
 - ✗ Multiple metrics or a single index