Trends in Energy-Efficient Supercomputing

BoF Organizers:
Wu Feng, Erich Strohmaier, Natalie Bates, and Tom Scogland
The Ultimate Goal of “The Green500 List”

• Raise awareness (and encourage reporting) of the energy efficiency of supercomputers
  – Drive energy efficiency as a first-order design constraint (on par with performance).

Encourage fair use of the list rankings to promote energy efficiency in high-performance computing systems.
Agenda

• The Green500 & Its Evolution: Past, Present, Future (Wu Feng)
  [ Discussion and Q&A ]
• Status of L1/L2/L3 Measurements (Natalie Bates)
  [ Discussion and Q&A ]
• MEGWARE Report on L2/L3 Measurement (Axel Auweter)
  [ Discussion and Q&A ]
• The 28th Green500 List (Wu Feng)
  – Trends and Evolution
  – Awards
  [ Discussion and Q&A ]
The Green500 and Its Evolution: Past, Present, and Future

Wu Feng
Brief History:
From Green Destiny to The Green500 List


4/2005: Workshop on High-Performance, Power-Aware Computing
  – Keynote address generates initial discussion for Green500 List

  – Workshop on High-Performance, Power-Aware Computing
  – Jack Dongarra’s CCGSC Workshop “The Final Push” (Dan Fay)

9/2006: Founding of Green500: Web Site and RFC (Chung-Hsing Hsu)
  – http://www.green500.org/ Generates feedback from hundreds

11/2007: Launch of the First Green500 List (Kirk Cameron)
  – http://www.green500.org/lists/green200711

Virtual SC BoF, SC’20, Nov. 2020
© 2020, W. Feng
Evolution of Green500

- **11/2010:** Updated Green500 Official Run Rules Released
- **06/2011:** Collaborations Begin on Methodologies for Measuring the Energy Efficiency of Supercomputers (Natalie Bates)
- **06/2013:** Adoption of New Power Measurement Methodology, version 1.0 (EE HPC WG, The Green Grid, Green500, TOP500)
- **01/2016:** Adoption of New Power Measurement Methodology, version 2.0 (EE HPC WG, The Green Grid, Green500, TOP500)
- **05/2016:** Green500 Merges with the TOP500
  - Unified run rules, data collection, and posting of power measurements via the TOP500 ([http://www.green500.org](http://www.green500.org) → [http://www.top500.org/green500](http://www.top500.org/green500))
  - Enable submissions of both performance-optimized (TOP500) and power-optimized (Green500) numbers, *but* with the following constraints...
Evolution of

• Submission of alternate performance and power numbers is *allowed* to the Green500 but with the following constraints:
  – The same *full machine* that was used for the TOP500 run is used for the Green500 run.
  – The same *problem size* that was used for the TOP500 run is used for the Green500 run.
Legacy Assumptions (circa 2007)

- Measuring a small part of a system and scaling it up does *not* introduce too 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

These assumptions were re-visited by EE HPC WG, The Green Grid, Top500, and Green500 (2011-2015)

Evolution of the Power Profile of the HPL Core Phase

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
Agenda

- The Green500 & Its Evolution: Past, Present, Future *(Wu Feng)*
  [ Discussion and Q&A ]
- Status of L1/L2/L3 Measurements *(Natalie Bates)*
  [ Discussion and Q&A ]
- MEGWARE Report on L2/L3 Measurement *(Axel Auweter)*
  [ Discussion and Q&A ]
- The 28th Green500 List *(Wu Feng)*
  - Trends and Evolution
  - Awards
  [ Discussion and Q&A ]
Status of L1/L2/L3 Submissions

N. Bates, W. Feng, E. Strohmaier, and T. Scogland
SC 2020 Green500 BoF
What’s in FLOPS/Watt?

• Submissions
  – Submitted
  – Derived
  – No number

• Measurement methodology
  – Level 1 (L1), Level 2 (L2) and Level 3 (L3)
State of Green500 Submissions

- Green500 June 2020 List Power Source
  - No Number: 316
  - Submitted: 185
  - Derived: 21

- 2020 submissions only
  - No Number: 41
  - Submitted: 12
  - Top 100 No Number entries

<table>
<thead>
<tr>
<th>TOP500 Name</th>
<th>Computer</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 Gadi</td>
<td>PRIMERGY CX2570 M5</td>
<td>National Computational Infrastructure</td>
<td>Fujitsu / Lenovo</td>
<td>Australia</td>
</tr>
<tr>
<td>29 Roxy</td>
<td>Apollo 2000</td>
<td>Government</td>
<td>HPE</td>
<td>USA</td>
</tr>
<tr>
<td>37 Flow</td>
<td>PRIMEHPC FX1000</td>
<td>Nagoya University</td>
<td>Fujitsu</td>
<td>Japan</td>
</tr>
<tr>
<td>56 Betzy</td>
<td>Bull Sequana XH2000</td>
<td>UNINETT Sigma2 AS</td>
<td>Atos</td>
<td>Norway</td>
</tr>
</tbody>
</table>

Virtual SC BoF, SC’20, Nov. 2020
© 2020, W. Feng
What is the Difference Between the Three Levels?

- Increasing accuracy and precision; L1→L3
- Ease of measurement is variable by site & system

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine fraction</td>
<td>Largest of:</td>
<td>Largest of:</td>
<td>Whole system</td>
</tr>
<tr>
<td></td>
<td>• 2 kW</td>
<td>• 10 kW</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• 1/10 of the system</td>
<td>• 1/8 of the system</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• 15 nodes</td>
<td>• 15 nodes</td>
<td></td>
</tr>
<tr>
<td>Subsystems included</td>
<td>Only compute and network</td>
<td>All participating subsystems, estimates allowed</td>
<td>All participating subsystems must be measured</td>
</tr>
<tr>
<td>Meter accuracy</td>
<td>Minimum 5%</td>
<td>Minimum 2%</td>
<td>Revenue grade</td>
</tr>
<tr>
<td>Measurements to report</td>
<td>Average power, core phase</td>
<td>Average power, full run</td>
<td>Energy, full run</td>
</tr>
</tbody>
</table>
Why Make L2/L3 Submission?

- More accurate and precise information
- Increase ability and experience with HPC system-level power & energy measurements
- Some use case examples:
  - Architectural trending, system modeling
  - Procurement & data-center provisioning
  - Operational improvements
  - Validate component-level measurement
Why Make a L3 Submission?
(Feedback from Previous Green500 BoFs)

Los Alamos National Laboratory (LANL)
• Level 3 measurements encouraged diverse organizational teamwork
• Level 3 measurements laid the groundwork for future green monitoring

Swiss Supercomputing Center (CSCS)
• It is always good to have reliable information about your data center
• Doing a reliable Level 3 measurement is not harder than Level 2 or Level 1

RIKEN
• While doing the Shoubu System B Level 3 measurement, the submission team realized an opportunity for optimizing their cooling sub-system
Who Has Made an L2/L3 Submission?

- AIST
- AWE
- Calcul Québec/Compute Canada
- CSC (Center for Scientific Computing)
- CEA/TGCC-GENCI
- Commissariat à l'Énergie Atomique (CEA)
- Facebook
- Forschungszentrum Jülich (FZJ)
- Fujitsu Numazu Plant
- HLRN at ZIB/Konrad Zuse-Zentrum Berlin
- Joint Center for Advanced HPC
- Lawrence Livermore National Laboratory
- MIT/MGHPCC
- National Supercomputing Center in Wuxi
- Sandia National Laboratories
- SENAI CIMATEC

- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Oak Ridge National Laboratory
- Rensselaer Polytechnic Institute (RPI)
- Sandia National Laboratories (SNL)
- Swiss National Supercomputing Centre, CSCS
- Technische Universität Darmstadt
- RIKEN
- Preferred Networks
- NVIDIA Corporation
- Science and Technology Facilities Council
- Universitaet Mainz
- University of Tokyo

(through June 2020 Green500 List)
Gaining Traction:
Level 2 and Level 3 Measurements

![Bar graph showing the number of Level 2 submissions from 2017-06 to 2019-11.]

![Bar graph showing the number of Level 3 submissions from 2017-06 to 2019-11.]

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
What Needs to be Improved in the L2/L3 Methodology and Submission Process?

Swiss National Supercomputing Centre (CSCS)
• [Methodology] Only use only L3 measurement. “No harder than L1 or L2 …”
• [Submission] Need a better way to provide a report and supporting files. The free form box is insufficient.

Fujitsu Numazu Plant
• [Methodology] L3 too difficult from an infrastructure perspective. L2 good.
• [Submission] Make power reporting mandatory for every system.

Los Alamos National Laboratory (LANL)
• [Methodology] Useful document on L2/L3 but intimidating for first-time users
• [Methodology] Need contact info for quick questions or detailed discussions
• [Methodology] Need a list of known metering/measurement equipment
  – LANL needed to contact vendors to ensure that meters met the requirements
What’s Next?

- Should the reporting of power consumption be mandatory for a Top500 Submission?
- Should L2 be the new submission standard?
- What else?

Thank you!

http://eehpcwg.llnl.gov
natalie.jean.bates@gmail.com
Status of L1/L2/L3 Submissions

N. Bates, W. Feng, E. Strohmaier, and T. Scogland

SC 2020 Green500 BoF
What’s in FLOPS/Watt

• Submissions
  – Submitted
  – Derived
  – No number

• Measurement methodology
  – Level 1 (L1), Level 2 (L2) and Level 3 (L3)
State of Green500 Submissions

- Green500 June 2020 List Power Source
  - No Number 316
  - Submitted 185
  - Derived 21

- 2020 submissions only
  - No Number 41
  - Submitted 12
  - Top100 No Number entries

<table>
<thead>
<tr>
<th>TOP500 Name</th>
<th>Computer</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 Gadi</td>
<td>PRIMERGY CX2570 M5</td>
<td>National Computational Infrastructure</td>
<td>Fujitsu / Lenovo</td>
<td>Australia</td>
</tr>
<tr>
<td>29 Roxy</td>
<td>Apollo 2000</td>
<td>Government</td>
<td>HPE</td>
<td>USA</td>
</tr>
<tr>
<td>37 Flow</td>
<td>PRIMEHPC FX1000</td>
<td>Nagoya University</td>
<td>Fujitsu</td>
<td>Japan</td>
</tr>
<tr>
<td>56 Betzy</td>
<td>Bull Sequana XH2000</td>
<td>UNINETT Sigma2 AS</td>
<td>Atos</td>
<td>Norway</td>
</tr>
</tbody>
</table>

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
What is the Difference Between the Three Levels?

- Increasing accuracy and precision; L1 → L3
- Ease of measurement is variable by site & system

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
</table>
| **Machine fraction** | Largest of:  
  • 2 kW  
  • 1/10 of the system  
  • 15 nodes | Largest of:  
  • 10 kW  
  • 1/8 of the system  
  • 15 nodes | Whole system |
| **Subsystems included** | Only compute and network | All participating subsystems, estimates allowed | All participating subsystems must be measured |
| **Meter accuracy** | Minimum 5% | Minimum 2% | Revenue grade |
| **Measurements to report** | Average power, core phase | Average power, full run | Energy, full run |
Why Make L2/L3 Submission?

• More accurate and precise information
• Increase ability and experience with HPC system-level power & energy measurements
• Some use case examples:
  – Architectural trending, system modeling
  – Procurement & data-center provisioning
  – Operational improvements
  – Validate component-level measurement
Why Make a L3 Submission?
(Feedback from Previous Green500 BoFs)

Los Alamos National Laboratory (LANL)
- Level 3 measurements encouraged diverse organizational teamwork
- Level 3 measurements laid the groundwork for future green monitoring

Swiss Supercomputing Center (CSCS)
- It is always good to have reliable information about your data center
- Doing a reliable Level 3 measurement is not harder than Level 2 or Level 1

RIKEN
- While doing the Shoubu System B Level 3 measurement, the submission team realized an opportunity for optimizing their cooling sub-system
Who Has Made an L2/L3 Submission?

- AIST
- AWE
- Calcul Québec/Compute Canada
- CSC (Center for Scientific Computing)
- CEA/TGCC-GENCI
- Commissariat a l'Energie Atomique (CEA)
- Facebook
- Forschungszentrum Juelich (FZJ)
- Fujitsu Numazu Plant
- HLRN at ZIB/Konrad Zuse-Zentrum Berlin
- Joint Center for Advanced HPC
- Lawrence Livermore National Laboratory
- MIT/MGHPCC
- National Supercomputing Center in Wuxi
- Sandia National Laboratories
- SENAI CIMATEC
- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Oak Ridge National Laboratory
- Rensselaer Polytechnic Institute (RPI)
- Sandia National Laboratories (SNL)
- Swiss National Supercomputing Centre, CSCS
- Technische Universität Darmstadt
- RIKEN
- Preferred Networks
- NVIDIA Corporation
- Science and Technology Facilities Council
- Universität Mainz
- University of Tokyo

(through June 2020 Green500 List)
Gaining Traction: Level 2 and Level 3 Measurements

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
What Needs to be Improved in the L2/L3 Methodology and Submission Process?

Swiss National Supercomputing Centre (CSCS)
• [Methodology] Only use only L3 measurement. “No harder than L1 or L2 …”
• [Submission] Need a better way to provide a report and supporting files. The free form box is insufficient.

Fujitsu Numazu Plant
• [Methodology] L3 too difficult from an infrastructure perspective. L2 good.
• [Submission] Make power reporting mandatory for every system.

Los Alamos National Laboratory (LANL)
• [Methodology] Useful document on L2/L3 but intimidating for first-time users
• [Methodology] Need contact info for quick questions or detailed discussions
• [Methodology] Need a list of known metering/measurement equipment
  – LANL needed to contact vendors to ensure that meters met the requirements
What’s Next?

• Should the reporting of power consumption be mandatory for a Top500 Submission?
• Should L2 be the new submission standard?
• What else?

Thank you!

http://eehpcwg.llnl.gov
natalie.jean.bates@gmail.com
Agenda

• The Green500 & Its Evolution: Past, Present, Future (Wu Feng)
  [ Discussion and Q&A ]

• Status of L1/L2/L3 Measurements (Natalie Bates)
  [ Discussion and Q&A ]

• MEGWARE Report on L2/L3 Measurement (Axel Auweter)
  [ Discussion and Q&A ]

• The 28th Green500 List (Wu Feng)
  – Trends and Evolution
  – Awards

  [ Discussion and Q&A ]
MEGWARE Report on L2/L3 Measurement
Agenda

- The Green500 & Its Evolution: Past, Present, Future *(Wu Feng)*
  [ Discussion and Q&A ]
- Status of L1/L2/L3 Measurements *(Natalie Bates)*
  [ Discussion and Q&A ]
- MEGWARE Report on L2/L3 Measurement *(Axel Auweter)*
  [ Discussion and Q&A ]
- The 28th Green500 List *(Wu Feng)*
  - Trends and Evolution
  - Awards
  [ Discussion and Q&A ]
Trends: How Energy Efficient Are We?
Trends: How Energy Efficient Are We?

Green500 Rank
- Top
- Mean
- Median
- Bottom

Projection from ISC’20

Year
- 2007
- 2008
- 2009
- 2010
- 2011
- 2012
- 2013
- 2014
- 2015
- 2016
- 2017
- 2018
- 2019
- 2020
- 2021
- 2022
- 2023

MFLOPS/W
- 0
- 10
- 100
- 1000
- 10000
- 25000
- 33000
- 50000

Virtual SC BoF, SC’20, Nov. 2020 © 2020, W. Feng
Trends: How Energy Efficient Are We?

Green500 Rank

Year

MFLOPS/W

Virtual SC BoF, SC'20, Nov. 2020 © 2020, W. Feng
Trends in Efficiency (2012 – Present):
Homogeneous vs. Heterogeneous Systems

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
Trends in Power: Max, Mean, Median, Min

Legend: 
- Max
- Mean
- Median
- Min

Power (kW)

Year

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
Efficiency vs. Performance

Exascale Goal: Green500

Exascale @ 20MW

Fastest Supercomputer (Fugaku)

Greenest Supercomputer (NVIDIA DGX SuperPOD)

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
Level 2 and Level 3 Submissions over Time
Trends Towards Exascale
Exascale Computing Study: Technology Challenges in Achieving Exascale Systems

• Goal
  – “Because of the difficulty of achieving such physical constraints, the study was permitted to assume some growth, perhaps a factor of 2X, to something with a maximum limit of 500 racks and 20 MW for the computational part of the 2015 system.”

• Realistic Projection?
  – “Assuming that Linpack performance will continue to be of at least passing significance to real Exascale applications, and that technology advances in fact proceed as they did in the last decade (both of which have been shown here to be of dubious validity), then [...] an Exaflop per second system is possible at around 67 MW.”
Trends: Extrapolating to Exaflop (June 2020)

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng
Trends: Extrapolating to Exaflop (June 2020)

Power extrapolated to exaflop (MW)

List release year

38.2 MW
67.6 MW
<table>
<thead>
<tr>
<th>Green500 Rank</th>
<th>GFLOPS/W</th>
<th>Name</th>
<th>Site</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26.20</td>
<td>NVIDIA DGX SuperPOD</td>
<td>NVIDIA Corporation</td>
<td>NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband</td>
</tr>
<tr>
<td>2</td>
<td>26.04</td>
<td>MN-3</td>
<td>Preferred Networks</td>
<td>MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect</td>
</tr>
<tr>
<td>3</td>
<td>25.01</td>
<td>JUWELS Booster Module</td>
<td>Forschungszentrum Juelpich (FZJ)</td>
<td>Bull Sequana XH2000, AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR Infiniband/ParTec ParaStation ClusterSuite</td>
</tr>
<tr>
<td>4</td>
<td>24.26</td>
<td>Spartan2</td>
<td>Atos</td>
<td>Bull Sequana XH2000, AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR Infiniband</td>
</tr>
<tr>
<td>5</td>
<td>23.98</td>
<td>Selene</td>
<td>NVIDIA Corporation</td>
<td>NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband</td>
</tr>
<tr>
<td>6</td>
<td>16.88</td>
<td>A64FX prototype</td>
<td>Fujitsu Numazu Plant</td>
<td>Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D</td>
</tr>
<tr>
<td>7</td>
<td>16.28</td>
<td>AiMOS</td>
<td>RPICenter for Computational Innovations (CCI)</td>
<td>IBM Power System AC922, IBM POWER9 20C 3.45GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband</td>
</tr>
<tr>
<td>8</td>
<td>15.74</td>
<td>HPC5</td>
<td>Eni S.p.A.</td>
<td>PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband</td>
</tr>
<tr>
<td>9</td>
<td>15.57</td>
<td>Satori</td>
<td>MIT/MGHPCC Holyoke, MA</td>
<td>IBM Power System AC922, IBM POWER9 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2</td>
</tr>
<tr>
<td>10</td>
<td>15.42</td>
<td>Supercomputer Fugaku</td>
<td>RIKEN Center for Computational Science</td>
<td>Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D</td>
</tr>
</tbody>
</table>
Brief Analysis of Top 10 Machines on

- 8 out of 10 machines are accelerator-based
  - 1 MN-Core, 4 NVIDIA Ampere GPU, 3 NVIDIA Volta GPUs
- CPU Vendor Distribution in Top10
  - 4 AMD Zen-2 (Rome), 2 Intel Cascade Lake, 2 Fujitsu ARM, 2 IBM power
- From June 2020 Green500,
  - Stayed in top 10 (7)
    - MN-3 (#1→#2), Selene (#2→#5), A64X(#4→#6), AiMOS (#5→#7), HPC5 (#6→#8), Satori (#7→#9), Supercomputer Fugaku (#9→#10)
  - Slid out of top 10 (2)
    - Summit (#8→#11), Marconi-100 (#10→#12)
  - Dropped out due to performance cutoff (1)
    - NA-1
- New in top 10 (3)
  - NVIDIA DGX SuperPOD (#1), JUWELS Booster Module (#3), Spartan2 (#4)
- Country-wise distribution in Top 10
  - 4 from United States, 3 from Japan, 1 each from Germany, France, Italy
NVIDIA DGX SuperPOD, an NVIDIA DGX A100 System at NVIDIA Corporation, CA, USA

is ranked

No. 1 in the Green500

among the World’s TOP500 Supercomputers

with 26.2 Gflops/Watt on the Linpack Benchmark

on the Green500 List published at the SC20 Conference, November 16, 2020

Congratulations from the Green500 Editors
MN-3, a Preferred Networks Systems at Preferred Networks, Japan

is ranked amongst **Level-3 measured systems** as

**No. 1 in the Green500**

among the World’s TOP500 Supercomputers

with 26.0 Gflops/Watt on the Linpack Benchmark

on the Green500 List published at the SC20 Conference, November 16, 2020

Congratulations from the Green500 Editors
Acknowledgements

• Key Contributor
  – Vignesh Adhinarayanan

• Energy-Efficient HPC Working Group (Lead: Natalie Bates) and TOP500 (Erich Strohmaier, Jack Dongarra, Horst Simon)

• YOU!
  – For your contributions in raising awareness in the energy efficiency of supercomputing systems

Virtual SC BoF, SC'20, Nov. 2020
© 2020, W. Feng