



Trends in Energy-Efficient Supercomputing

BoF Organizers:

Wu Feng, Erich Strohmaier, Natalie Bates, and Tom Scogland



The 26th
GREEN
500

November 2019

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 and its Evolution: Past, Present, Future (*Wu Feng*)
- The 26th Green500 List (*Wu Feng*)
 - Trends and Evolution
- Status of L1/L2/L3 Measurements (*Erich Strohmaier & Natalie Bates*)
 - What's Next?
- Green500 Presentations from L2/L3 Reporting Sites
 - A64FX Prototype, #1 on Green500 (*Toshiyuki Shimizu, Fujitsu*)
 - AiMOS, #3 on Green500 (*Chris Carothers, RPI*)
 - Astra, #3 on Green500 (*David Martinez, SNL*)



The Green500 and its Evolution: Past, Present, and Future

Wu Feng

The
GREEN
500

Brief History:

From Green Destiny to The *Green500* List

2/2002: Green Destiny (<http://sss.lanl.gov/> → <http://sss.cs.vt.edu/>)

- “Honey, I Shrunk the Beowulf!” 31st ICPP, August 2002.
- “High-Density Computing: A 240-Processor Beowulf in One Cubic Meter, SC02, November 2002.

4/2005: Workshop on High-Performance, Power-Aware Computing

- Keynote address generates initial discussion for *Green500* List

4/2006 and 9/2006: Making a Case for a *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>

Evolution of

- 11/2009: Experimental Lists Created
 - *Little Green500*: More focus on LINPACK energy efficiency than on LINPACK performance in order to foster innovation
 - ~~*HPCC Green500*: Alternative workload to evaluate energy efficiency~~
 - ~~*Open Green500*: Enabling alternative innovative approaches for LINPACK to improve performance and energy efficiency, e.g., mixed precision~~
- 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)

Evolution of

- **05/2016:** Green500 Merges with TOP500
 - Unified run rules, data collection, and posting of power measurements via the TOP500 (<http://www.green500.org> → <http://www.top500.org/green500>)
 - Enable submissions of both performance-optimized (TOP500) and power-optimized (Green500) numbers



Evolution of

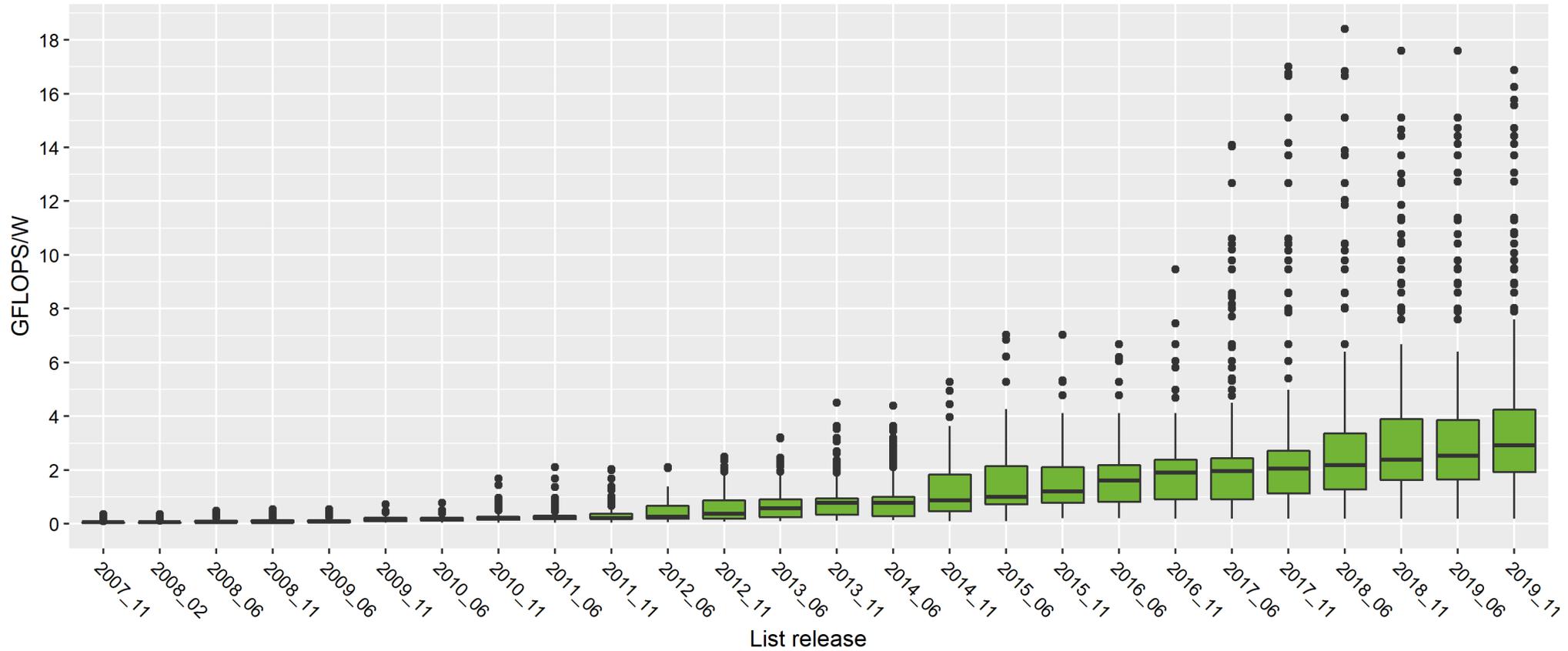
- Submission of alternate performance and power numbers is *allowed* to the Green500 but with the following constraints:
 - The same **full machine** that was used for the TOP500 run is used for the Green500 run.
 - The same **problem size** that was used for the TOP500 run is used for the Green500 run.



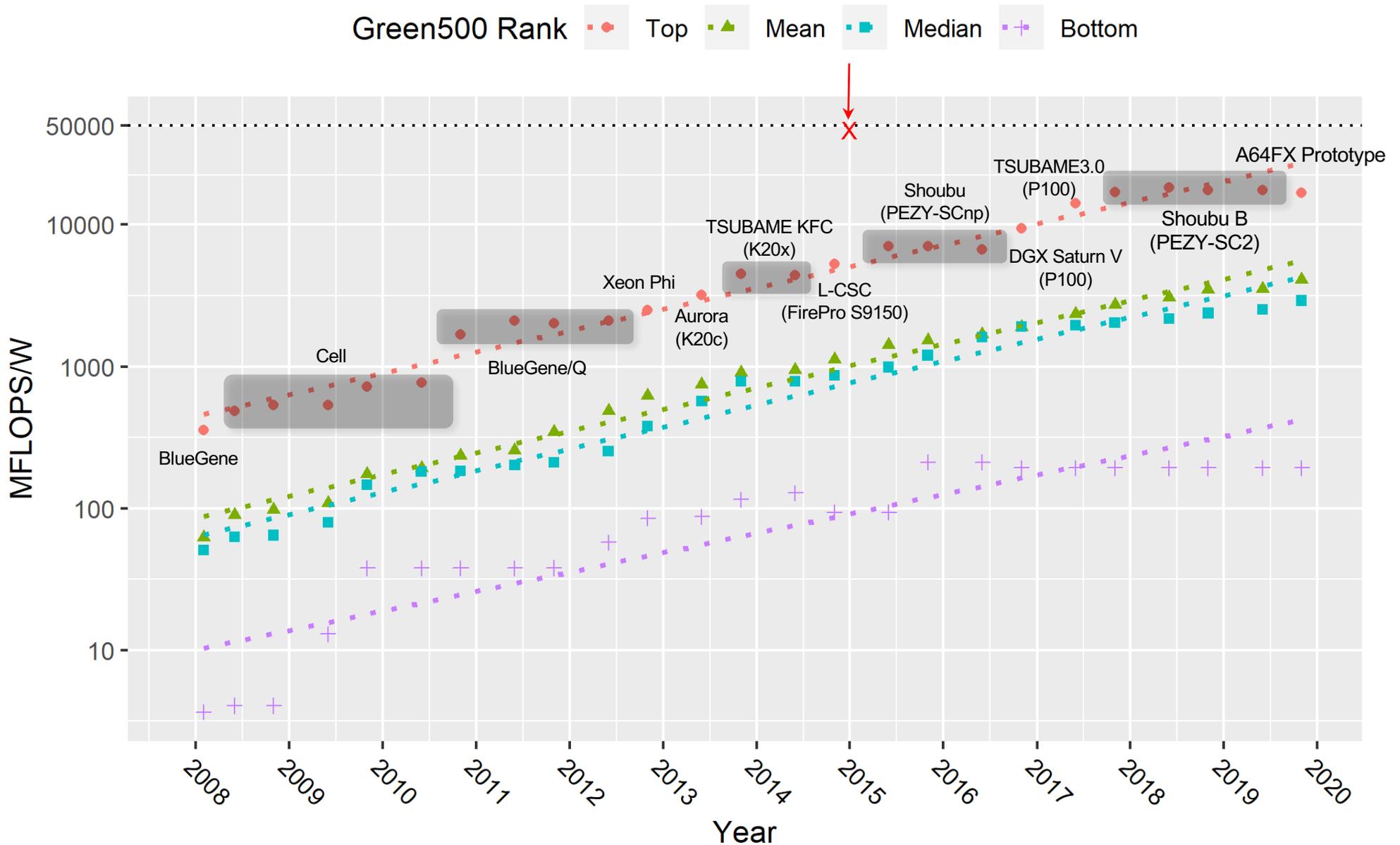
Agenda

- The Green500 and its Evolution: Past, Present, Future (*Wu Feng*)
- The 26th Green500 List (*Wu Feng*)
 - Trends and Evolution
- Status of L1/L2/L3 Measurements (*Erich Strohmaier & Natalie Bates*)
 - What's Next?
- Green500 Presentations from L2/L3 Reporting Sites
 - A64FX Prototype, #1 on Green500 (*Toshiyuki SHIMIZU, Fujitsu*)
 - AiMOS, #3 on Green500 (*Chris CAROTHERS, RPI*)
 - Astra, #3 on Green500 (*David Martinez, SNL*)

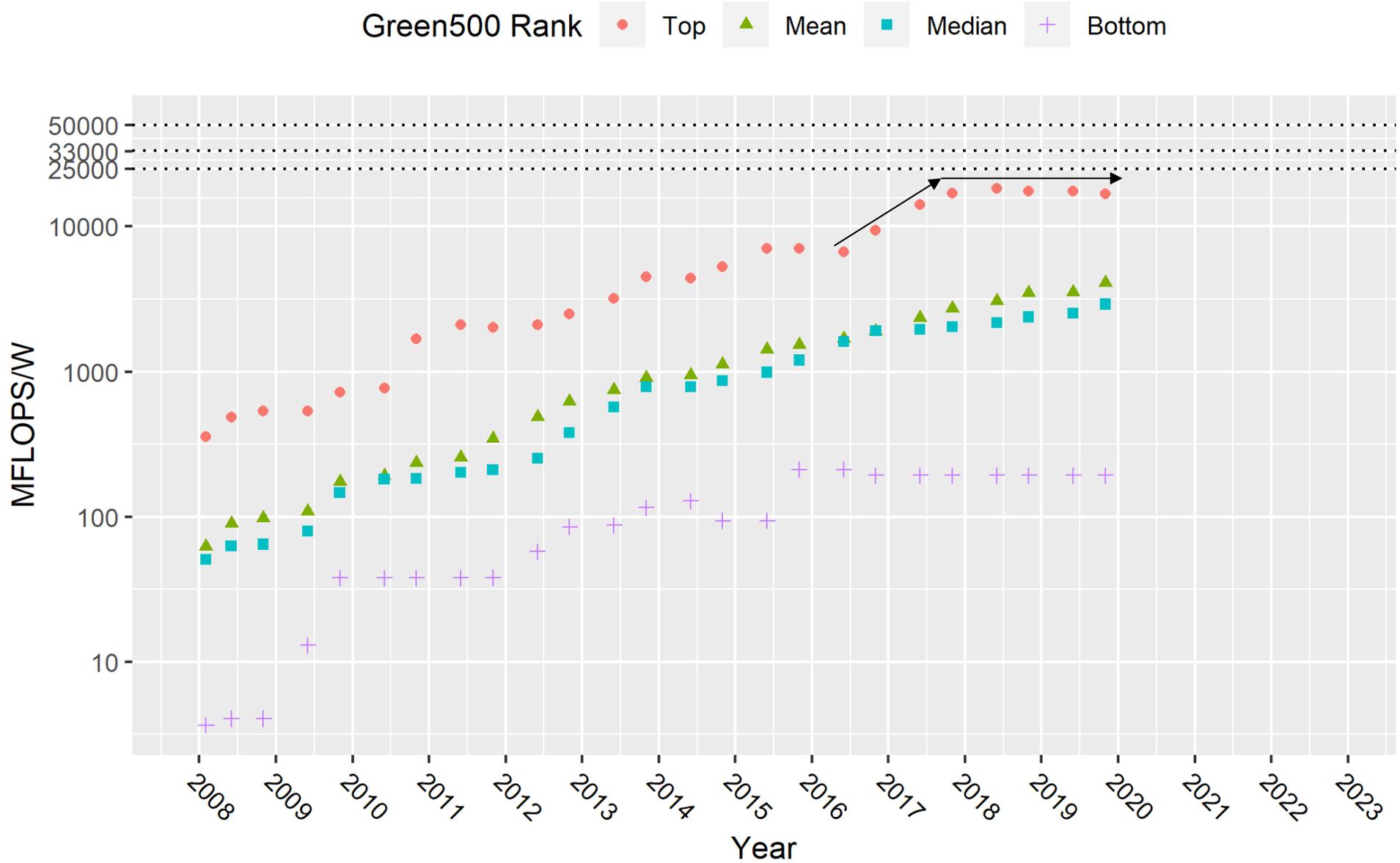
Trends: How Energy Efficient Are We?



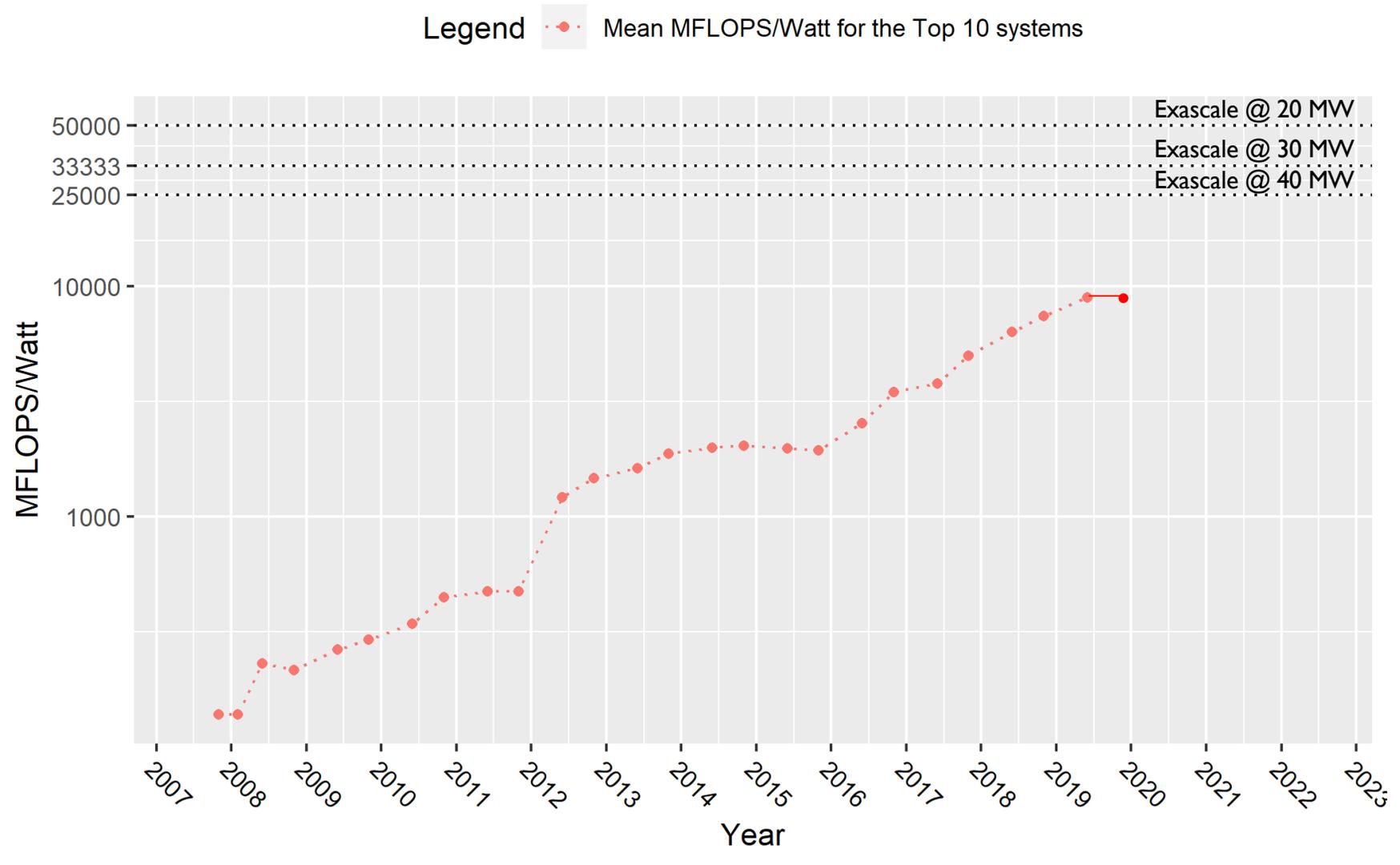
Trends: How Energy Efficient Are We?



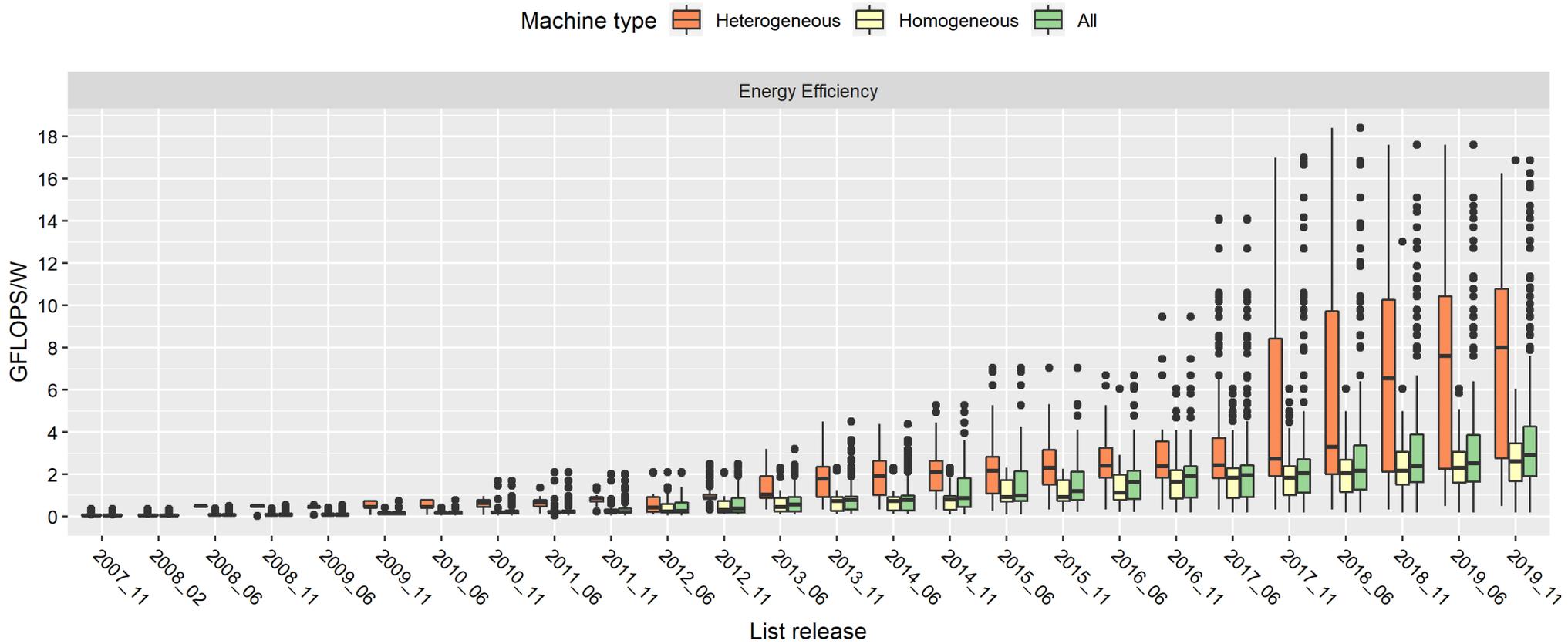
Trends: How Energy Efficient Are We?



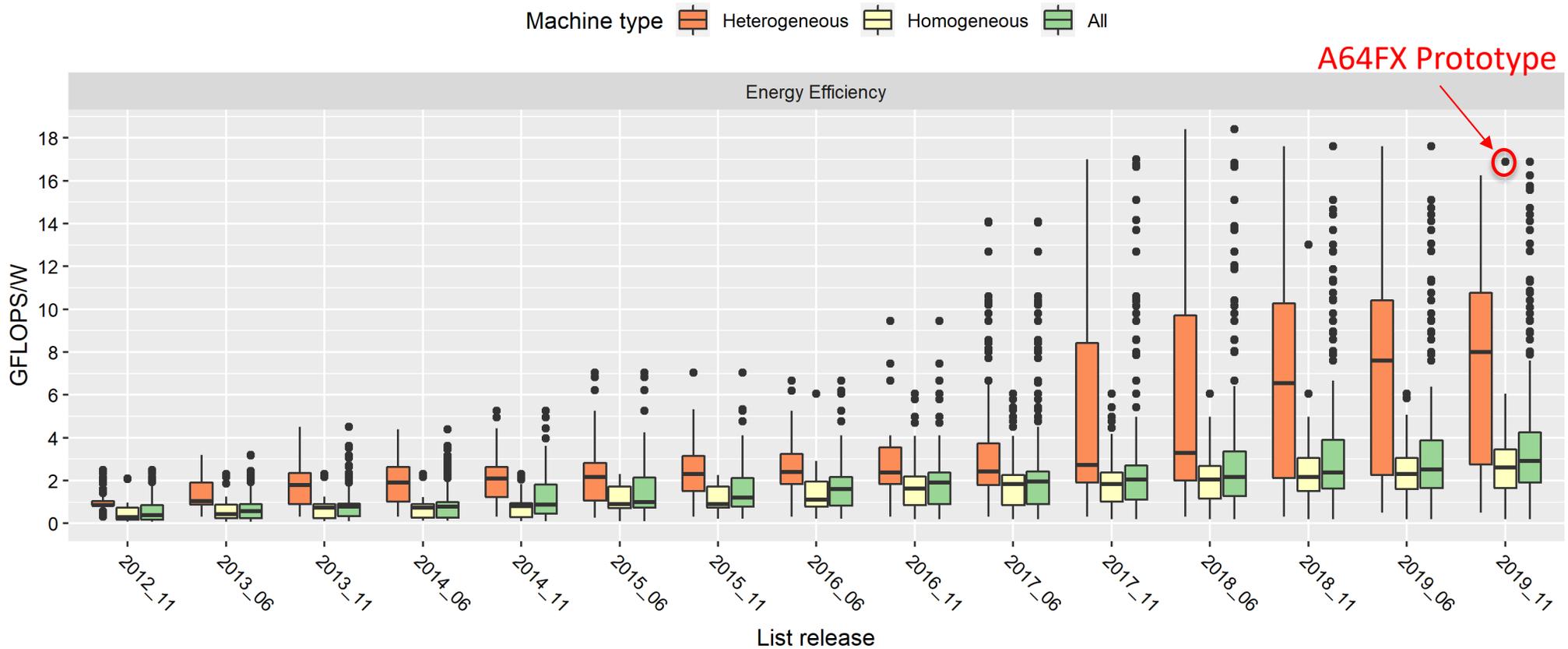
Trends: How Energy Efficient Are The Fastest Supercomputers?



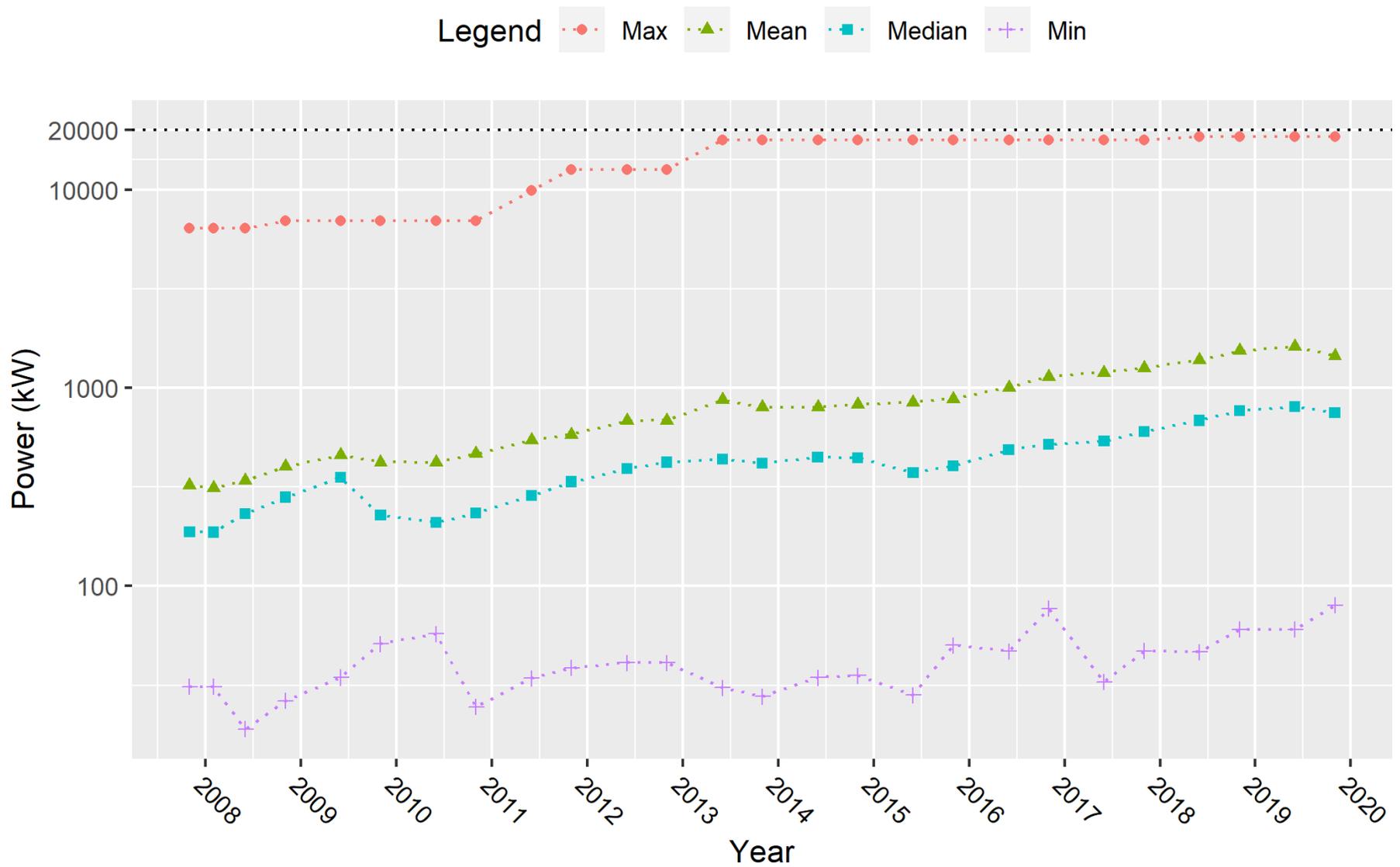
Trends in Efficiency (2007 – Present): Homogeneous vs. Heterogeneous Systems



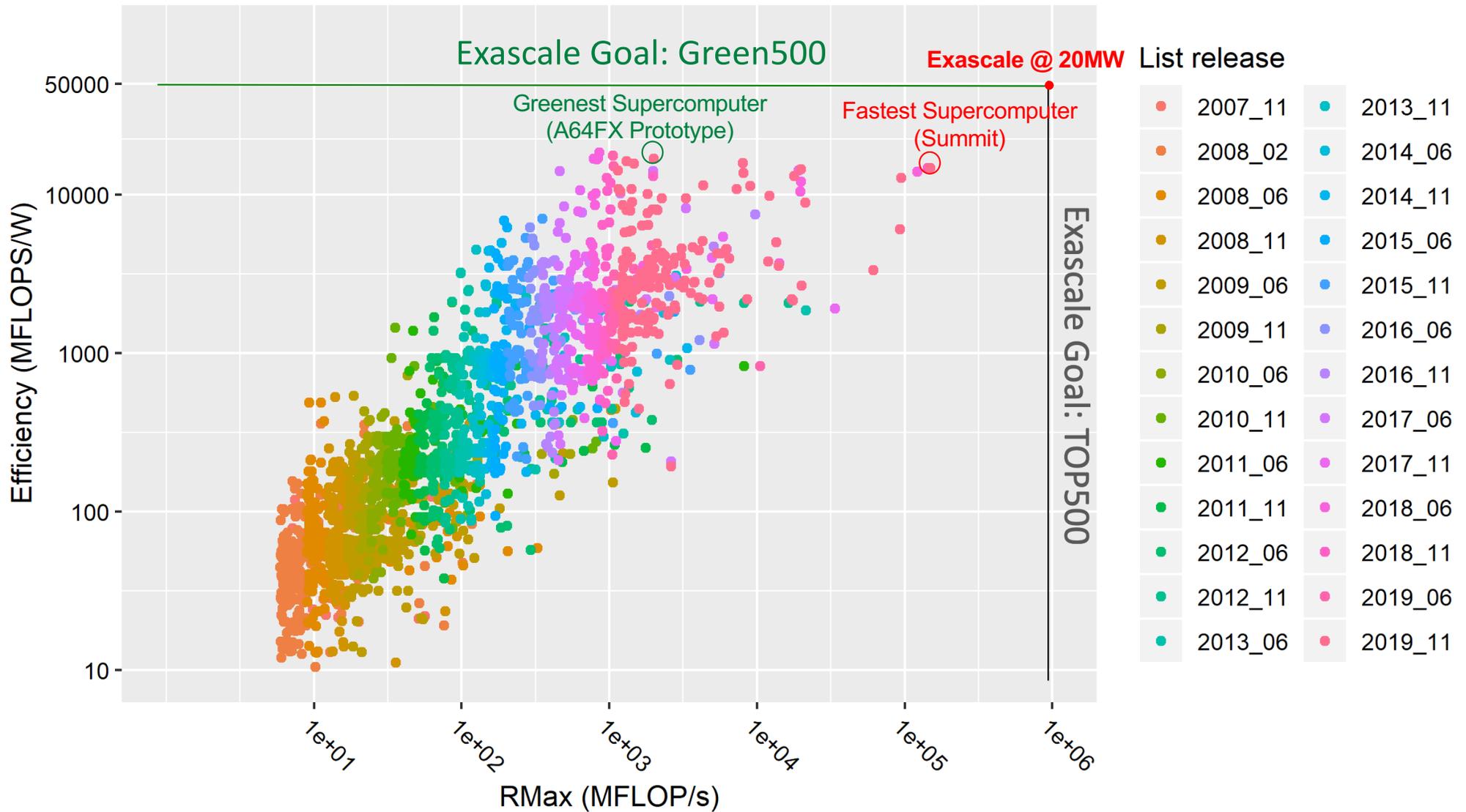
Trends in Efficiency (2012 – Present): Homogeneous vs. Heterogeneous Systems



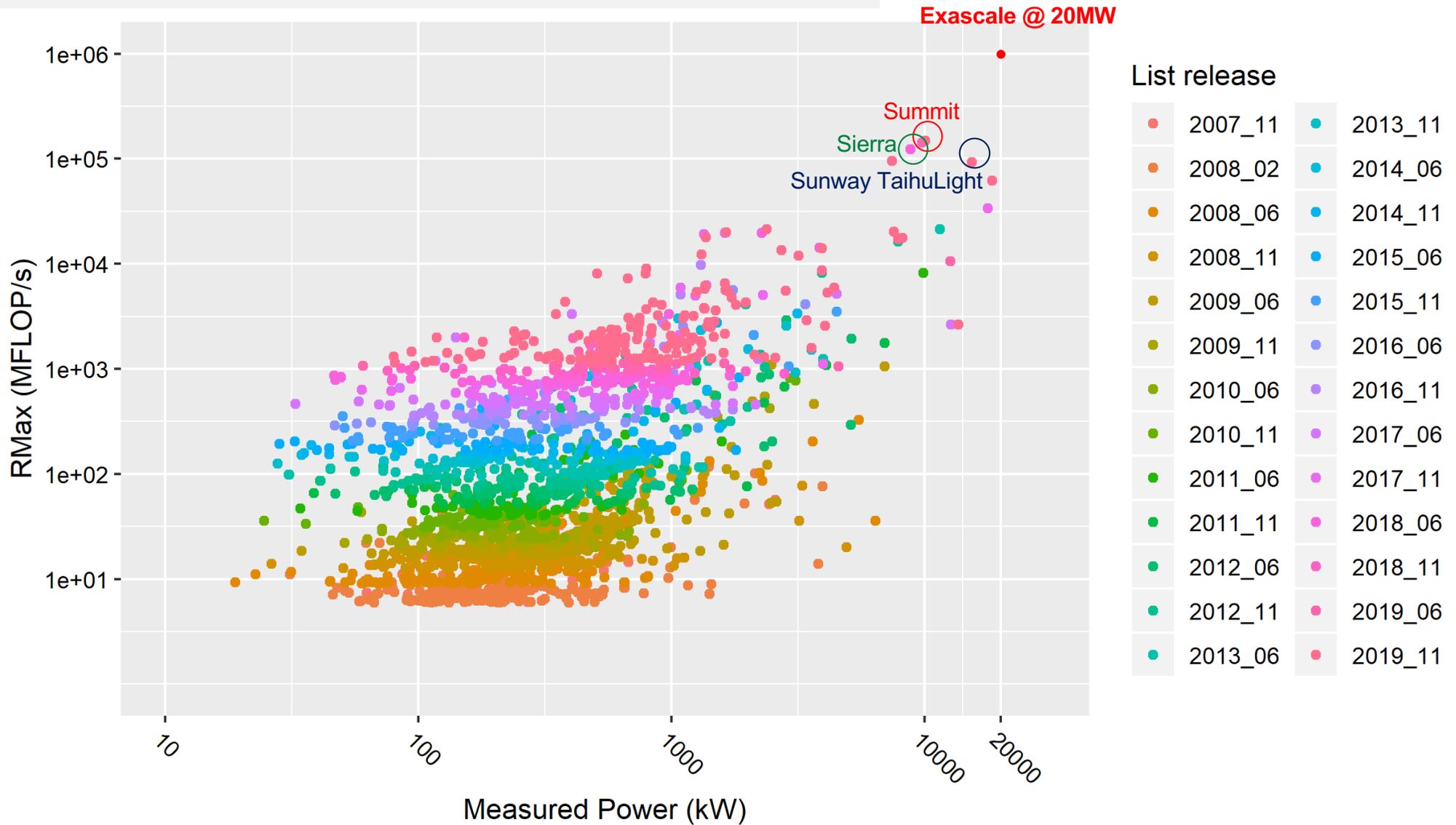
Trends in Power: Max, Mean, Median, Min



Efficiency vs. Performance



Performance vs. Power



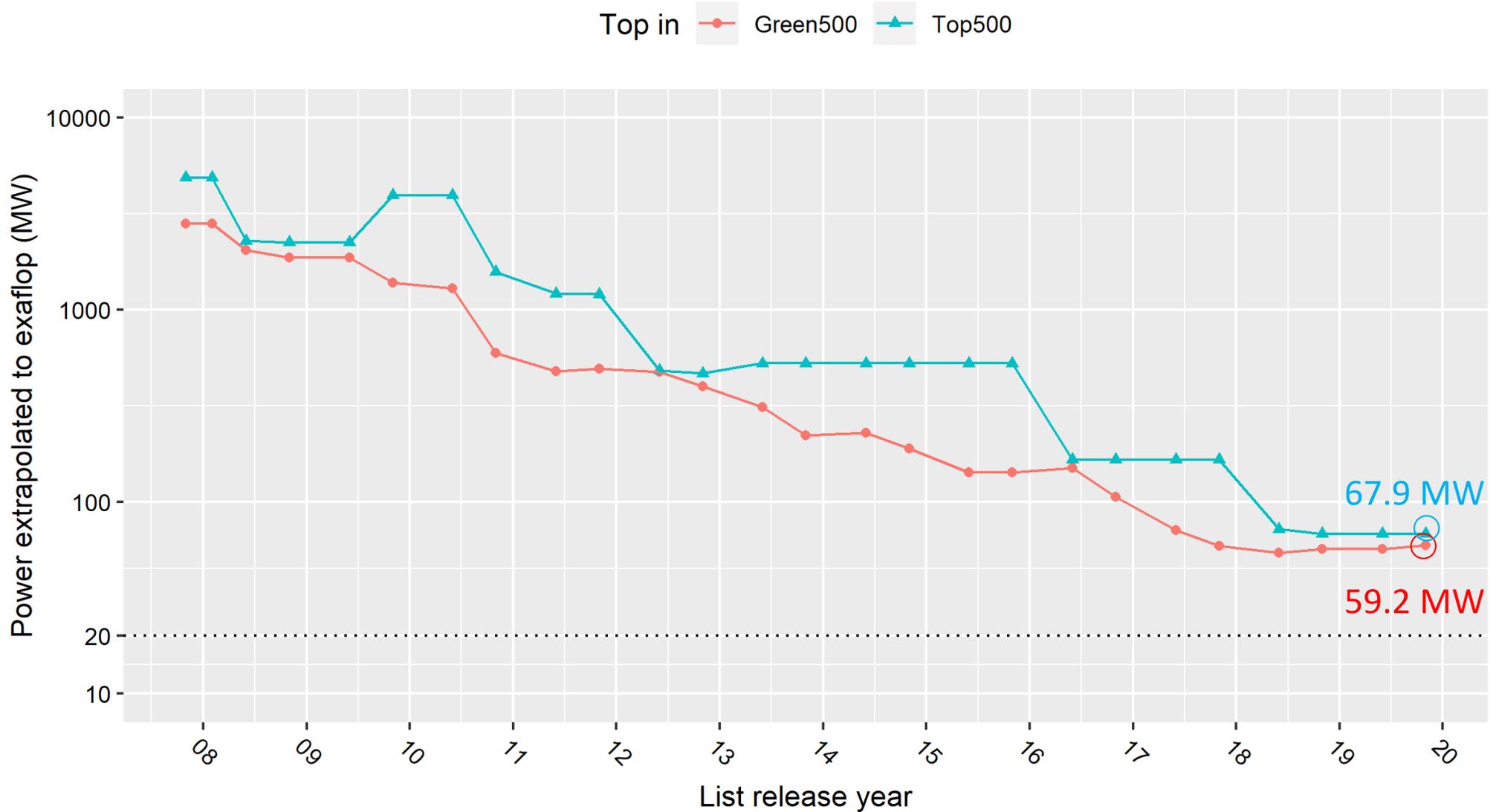
Trends Towards Exascale

The Green500 BoF, SC'19, November 2019

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 (Nov 2019)



Green500 Rank	GFLOPS/W	Name	Site	Computer
1	16.876	A64FX prototype	Fujitsu Numazu Plant	Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D
2	16.256	NA-1	PEZY Computing K.K.	ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz
3	15.771	AiMOS	Rensselaer Polytechnic Institute Center for Computational Innovations (CCI)	IBM Power System AC922, IBM POWER9 20C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100
4	15.574	Satori	MIT/MGHPCC Holyoke, MA	IBM Power System AC922, IBM POWER9 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2
5	14.719	Summit	DOE/SC/Oak Ridge National Laboratory	IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband
6	14.423	AI Bridging Cloud Infrastructure (ABCI)	National Institute of Advanced Industrial Science and Technology (AIST)	PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR
7	14.131	MareNostrum P9 CTE	Barcelona Supercomputing Center	IBM Power System AC922, IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Tesla V100
8	13.704	TSUBAME3.0	GSIC Center, Tokyo Institute of Technology	SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2
9	13.065	PANGEA III	Total Exploration Production	IBM Power System AC922, IBM POWER9 18C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100
10	12.723	Sierra	DOE/NNSA/LLNL	IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband

Brief Analysis of Top 10 Machines on



- 9 out of 10 machines are accelerator-based
 - 1 PEZY-SC2, 7 NVIDIA Volta GPUs, 1 NVIDIA Pascal GPU
- From June 2019 Green500 to November 2019 Green500,
 - Stayed in top 10 (6)
 - Summit (#3 -> #5), AI Bridging Cloud Infrastructure (#4 -> #6), MareNostrum P9 CTE (#5 -> #7), TSUBAME3.0 (#6 -> #8), PANGEA III (#7 -> #9), Sierra (#8 -> #10)
 - Slid out of top 10 (2)
 - Advanced Computing System (#9 -> #11), Taiwania 2 (#10 -> #12)
 - Dropped out due to performance cutoff (2)
 - Shoubu system B (#1), DGX SaturnV Volta (#2)
- Changes in Top 10
 - New in top 10 (4)
 - A64FX prototype (#1), NA-1 (#2), AiMOS (#3), Satori (#4)
 - Peak efficiency drops from 17.6 GFLOPS/Watt → 16.9 GFLOPS/Watt
- Country-wise distribution in Top 10
 - 4 from Japan, 4 from United States, 1 from Spain, 1 from France

Agenda

- The Green500 and its Evolution: Past, Present, Future (*Wu Feng*)
- The 26th Green500 List (*Wu Feng*)
 - Trends and Evolution
- Status of L1/L2/L3 Measurements (*Erich Strohmaier & Natalie Bates*)
 - What's Next?
- Green500 Presentations from L2/L3 Reporting Sites
 - A64FX Prototype, #1 on Green500 (*Toshiyuki Shimizu, Fujitsu*)
 - AiMOS, #3 on Green500 (*Chris Carothers, RPI*)
 - Astra, #3 on Green500 (*David Martinez, SNL*)



Updated Status of L1/L2/L3 Submissions

N. Bates, W. Feng,
E. Strohmaier, and T. Scogland
SC 2019 Green500 BoF



Measuring Power: Level 1 (L1), Level 2 (L2), Level 3 (L3)

- State of Green500 submissions
- What is the difference between the three levels?
- Why make a L2/L3 submission?
- Who has made a L2/L3 submission?
- What needs to be improved in the L2/L3 methodology and submission process?
- Should L2 be the new submission standard?

Should the reporting of power consumption be mandatory for a TOP500 submission? (Suggestion by Fujitsu “A64FX Prototype” representatives.)

State of Green500 Submissions

- Only 130 submissions. Down from 188 submissions in 6/2019. Down from a high of 320+ submissions.
 - 98 L1; 22 L2; 10 L3
- How to encourage *more* submissions?
 - Award #1 only to submitters of a L2/L3 measurement?
 - Propose a phase-in period for L2/L3 measurements and abolish L1 measurements?
 - Mandate power submission for every supercomputer?
 - Others?





What is the Difference Between the Three Levels?

	Level 1	Level 2	Level 3
Granularity	One per second, evenly spaced across the core phase	One per second, evenly spaced across the <i>full run from idle to idle</i>	Continuously integrated energy across the full run from idle to idle
Measurements to report	<ul style="list-style-type: none"> Core-phase average 	<ul style="list-style-type: none"> Core-phase average <i>Full-run average</i> <i>10 measurement's in core</i> <i>Idle power</i> 	<ul style="list-style-type: none"> Core-phase energy Full-run energy 10 measurement's in core Idle power
Machine fraction	Largest of: <ul style="list-style-type: none"> 2 kW 1/10 of the system 15 nodes 	Largest of: <ul style="list-style-type: none"> <i>10 kW</i> <i>1/8 of the system</i> 15 nodes 	Whole system
Subsystems included	Compute and network	All participating subsystems: compute, network, storage if used, etc.	All participating subsystems: compute, network, storage if used, etc. (all measured)
Meter accuracy	Minimum 5%	<i>Minimum 2%</i>	Revenue grade (accepted by SPEC power) or documented 1% or better



Why Make a L2/L3 Submission?

- Difficult to obtain a system-level measurement, but it has value *beyond Green500 and Top500*
 - Do we need to go beyond lists? Of course ...
- Examples of use cases for system-level measurements :
 - Architectural trending, system modeling (design, selection, upgrade, tuning, analysis),
 - Procurement & data-center provisioning (see power constraints for proposed exascale supercomputers)
 - Operational improvements
 - Day-to-day workloads vs HPL
 - TCO combining HPC system and room Infrastructure
 - Validate component-level measurement (by summing them up!)



Why Make a L3 Submission?

(Feedback from Previous Green500 BoFs)

LANL Description of Benefits from L3 Submission

- Team Interaction
- L3 measurements laid the groundwork for future green monitoring
- Power monitoring, one of LANL's top issues now

Level 3 Reporting Systems at LANL

(Note: All Penguin Computing)

- LANL CTS-1 Grizzly (#119), LANL CTS-1 Fire (#126), LANL CTS-1 Ice (#127)



Who Has Made an L2/L3 Submission?

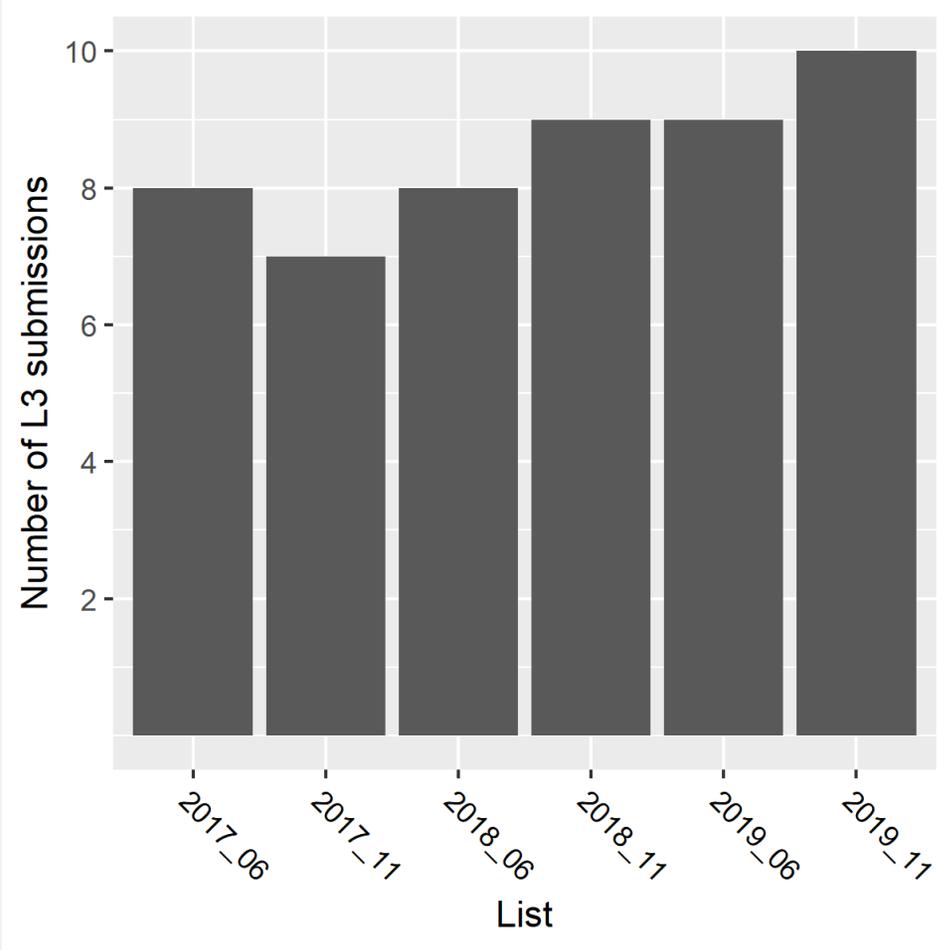
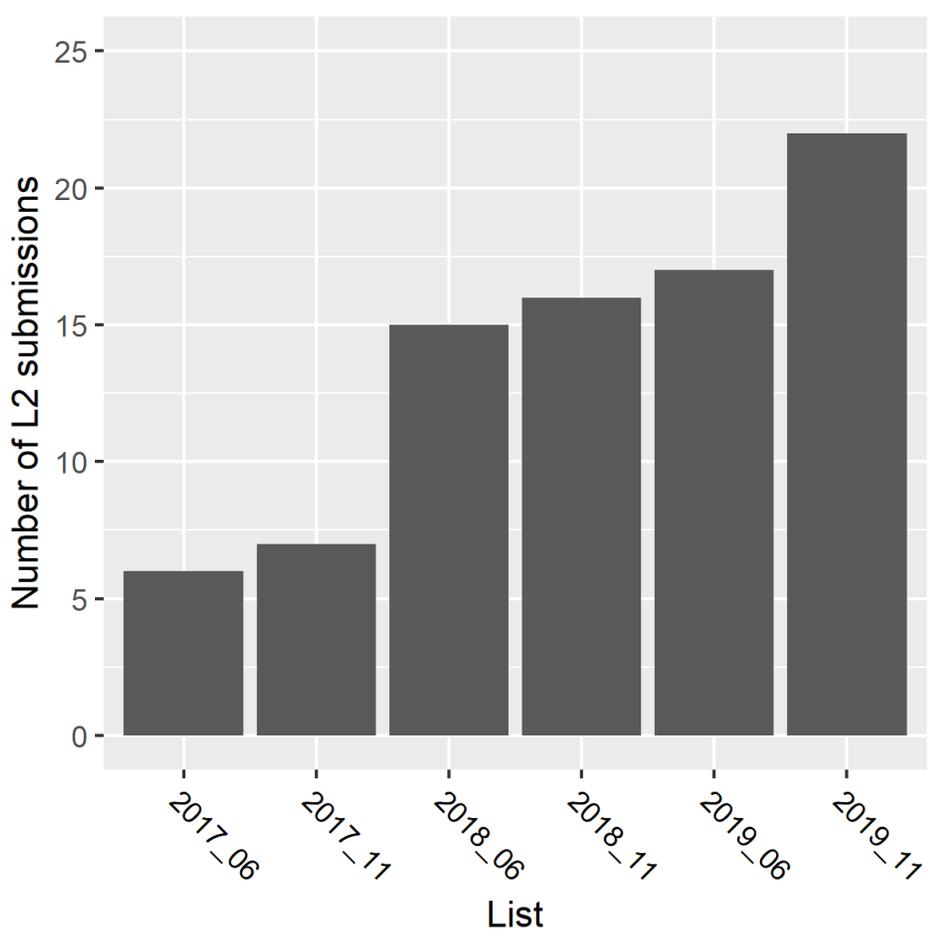
L2

- AIST
- AWE
- Calcul Québec/Compute Canada
- CSC (Center for Scientific Computing)
- CEA/TGCC-GENCI (2)
- 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 (3)
- SENAI CIMATEC
- Science and Technology Facilities Council
- Universitaet Mainz
- University of Tokyo

L3

- Lawrence Livermore National Laboratory (2)
- Los Alamos National Laboratory (3)
- Oak Ridge National Laboratory
- **Rensselaer Polytechnic Institute (RPI)**
- **Sandia National Laboratories (SNL)**
- Swiss National Supercomputing Centre, CSCS (2)

Gaining Traction: Level 2 and Level 3 Measurements





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



Thank you!

<http://eehpcwg.llnl.gov>

natalie.jean.bates@gmail.com

Agenda

- The Green500 and its Evolution: Past, Present, and Future (*Wu Feng*)
- Status of L1/L2/L3 Measurements (*Erich Strohmaier*)
 - What's Next?
- Green500 Presentations from L2/L3 Reporting Sites
 - AI Bridging Cloud Infrastructure, #4 on Green500 (*Hiroataka Ogawa, AIST*)
 - Shoubu System B, #1 on Green500 (*Sunao Torii, Exascaler Inc.*)
 - Summit, #3 on Green500 (*James Rogers, ORNL*)
- The 24th Green500 List (*Wu Feng*)
 - Trends and Evolution
- Discussion and Q&A

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