



## Trends in Energy-Efficient Supercomputing

BoF Organizers:

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



The 28th  
**GREEN**  
500

November 2020

# 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

## Making a Case for a Green500 List\*

W. Feng, IEEE IPDPS HPPAC 2006

- 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 28<sup>th</sup> Green500 List (*Wu Feng*)
  - Trends and Evolution
  - Awards[ Discussion and Q&A ]

COVER FEATURE

## The Green500 List: Encouraging Sustainable Supercomputing

Wu-chun Feng and Kirk W. Cameron  
Virginia Tech

**The performance-at-any-cost design mentality ignores supercomputers' excessive power consumption and need for heat dissipation and will ultimately limit their performance. Without fundamental change in the design of supercomputing systems, the performance advances common over the past two decades won't continue.**





# 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, *SC 2002*, 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

- **11/2010:** 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.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)

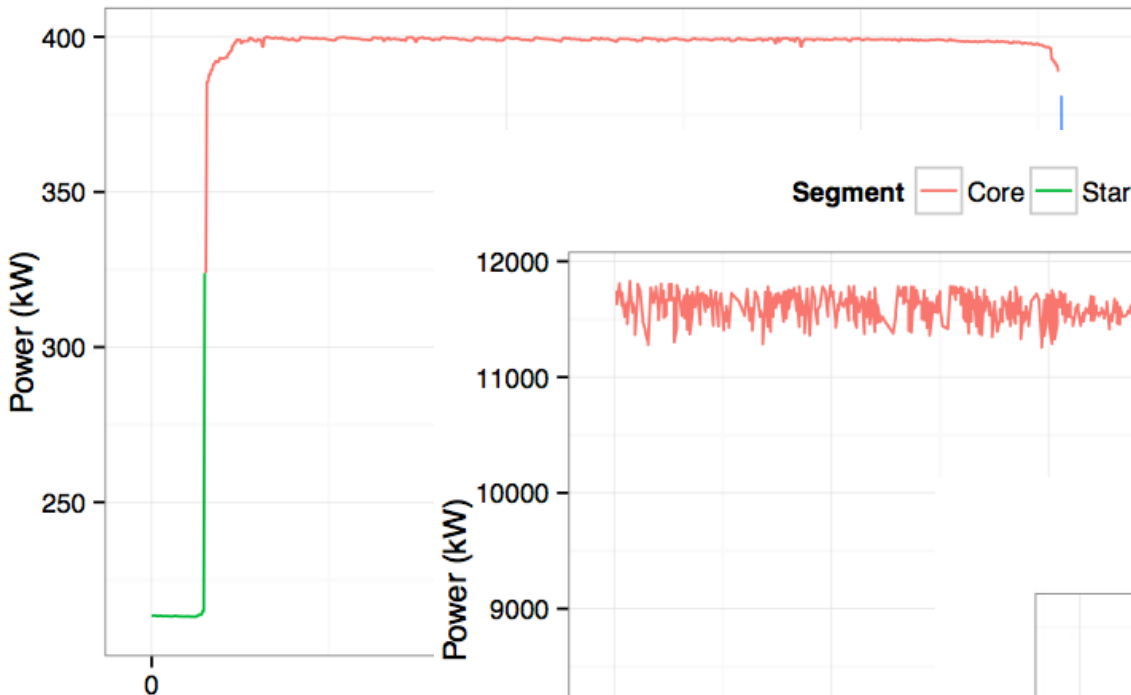
Originally designed to *encourage* reporting of power with accuracy

- 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

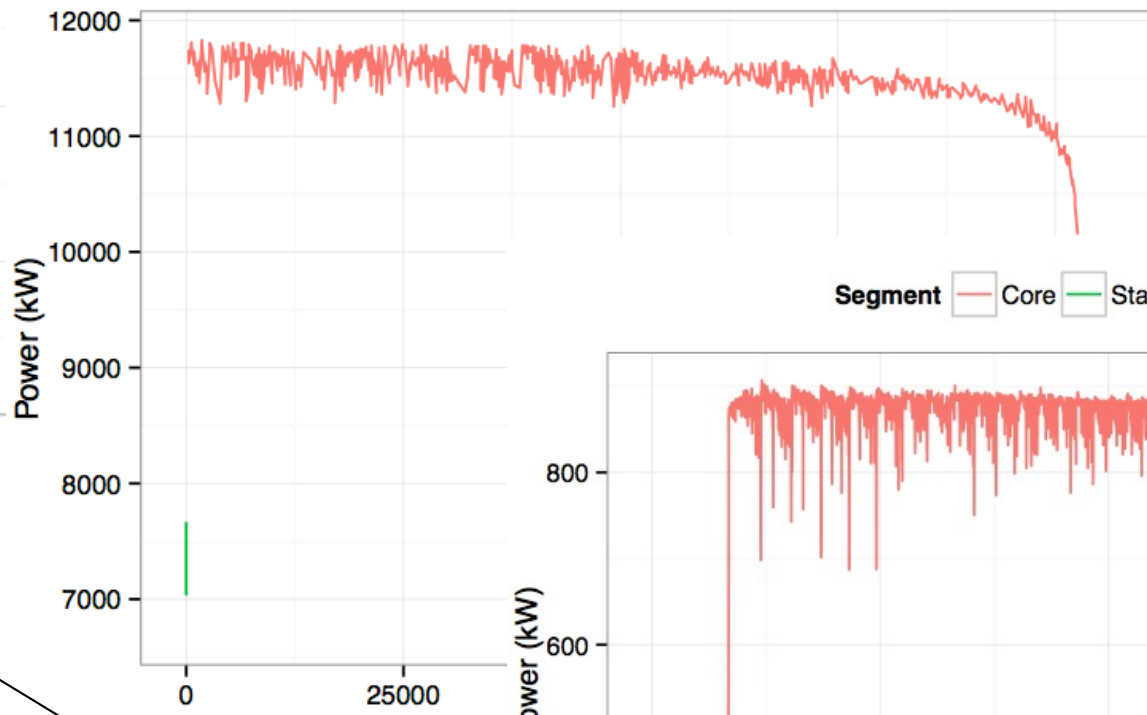
These assumptions were re-visited by *EE HPC WG, The Green Grid, Top500, and Green500* (2011-2015)

D. Rohr et al., “Refining Power Measurement Methodology for Supercomputer-System Benchmarking,” *International Supercomputing Conference*, July 2015.

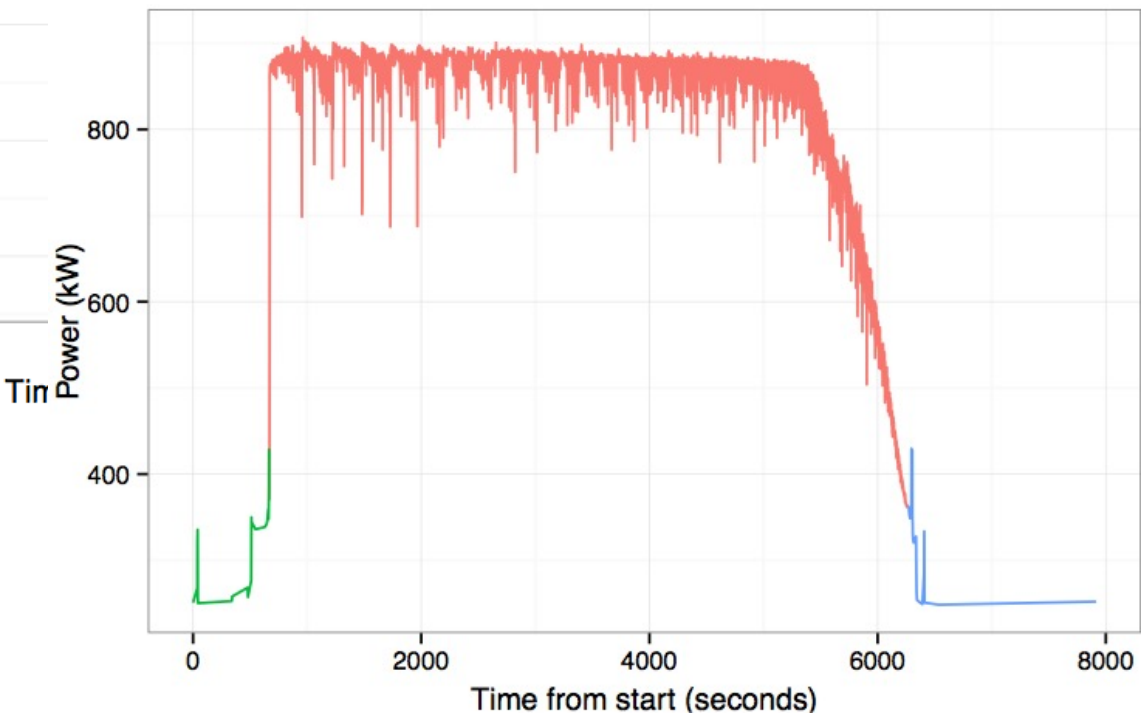
Segment Core Startup Tear-down



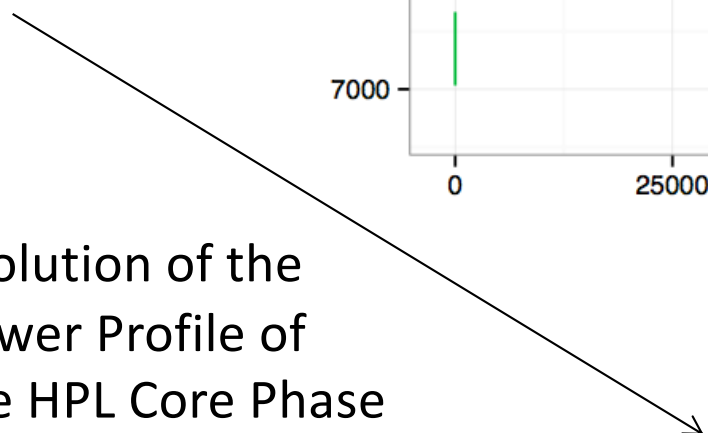
Segment Core Startup Tear-down



Segment Core Startup Tear-down



Evolution of the Power Profile of the HPL Core Phase



# 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 28<sup>th</sup> 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 June2020 List Power Source

- No Number 316
- Submitted 185
- Derived 21

- 2020 submissions only

- No Number 41
- Submitted 12
- Top100 No Number entries

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



# What is the Difference Between the Three Levels?

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

	Level 1	Level 2	Level 3
Machine fraction	Largest of: <ul style="list-style-type: none"><li>• 2 kW</li><li>• 1/10 of the system</li><li>• 15 nodes</li></ul>	Largest of: <ul style="list-style-type: none"><li>• 10 kW</li><li>• 1/8 of the system</li><li>• 15 nodes</li></ul>	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



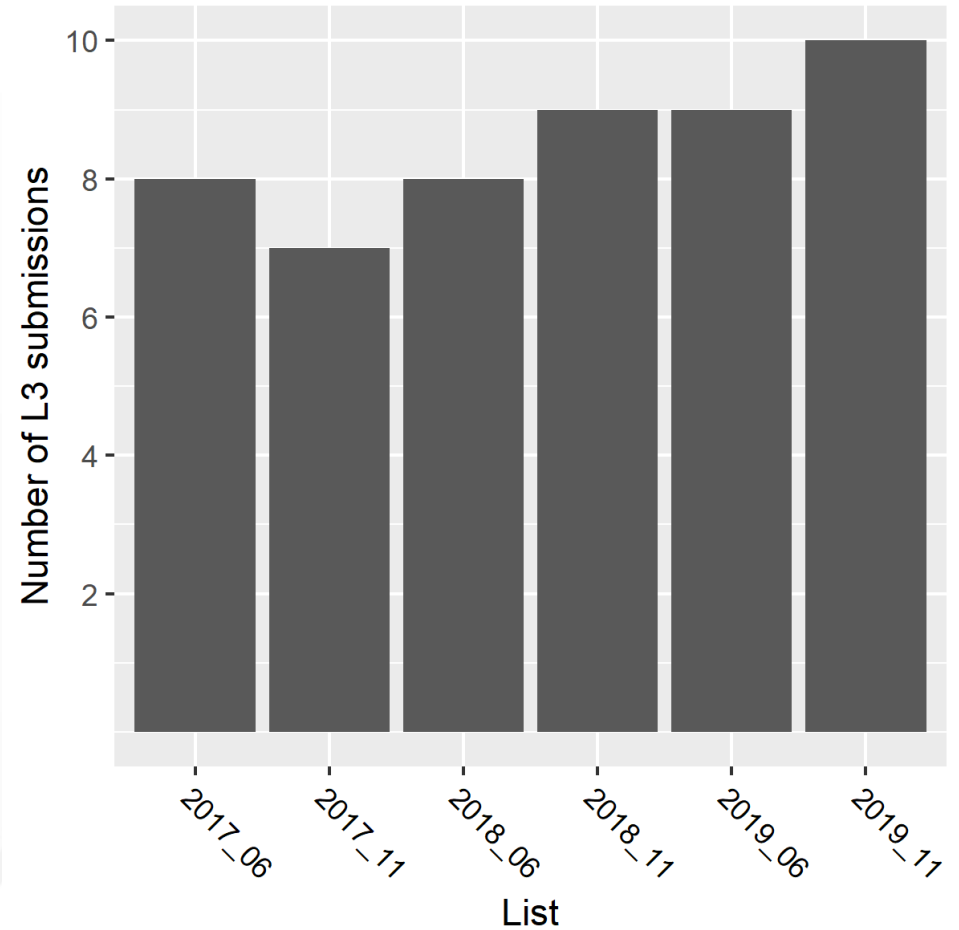
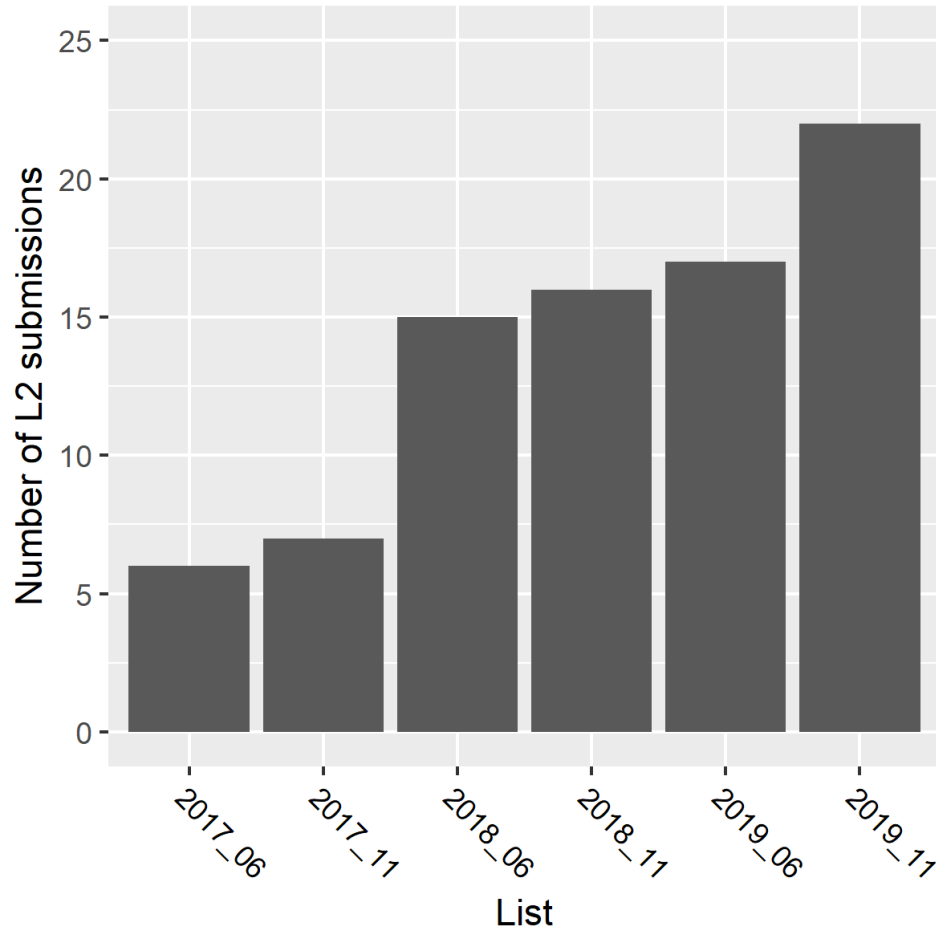
# Who Has Made an L2/L3 Submission?

- AIST
- AWE
- Calcul Québec/Compute Canada
- CSC (Center for Scientific Computing)
- CEA/TGCC-GENCI
- Commissariat à l'Énergie 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





# 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](mailto: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 June2020 List Power Source

- No Number 316
- Submitted 185
- Derived 21

- 2020 submissions only

- No Number 41
- Submitted 12
- Top100 No Number entries

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



# What is the Difference Between the Three Levels?

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

	Level 1	Level 2	Level 3
Machine fraction	Largest of: <ul style="list-style-type: none"><li>• 2 kW</li><li>• 1/10 of the system</li><li>• 15 nodes</li></ul>	Largest of: <ul style="list-style-type: none"><li>• 10 kW</li><li>• 1/8 of the system</li><li>• 15 nodes</li></ul>	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



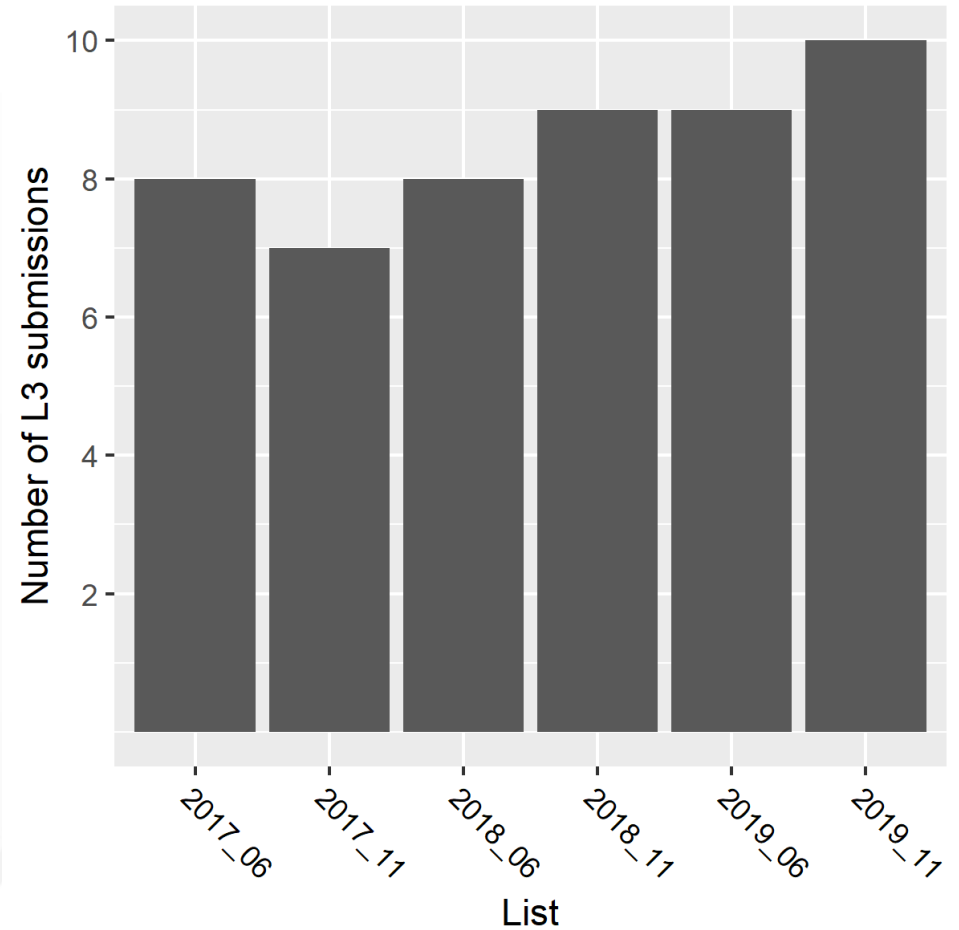
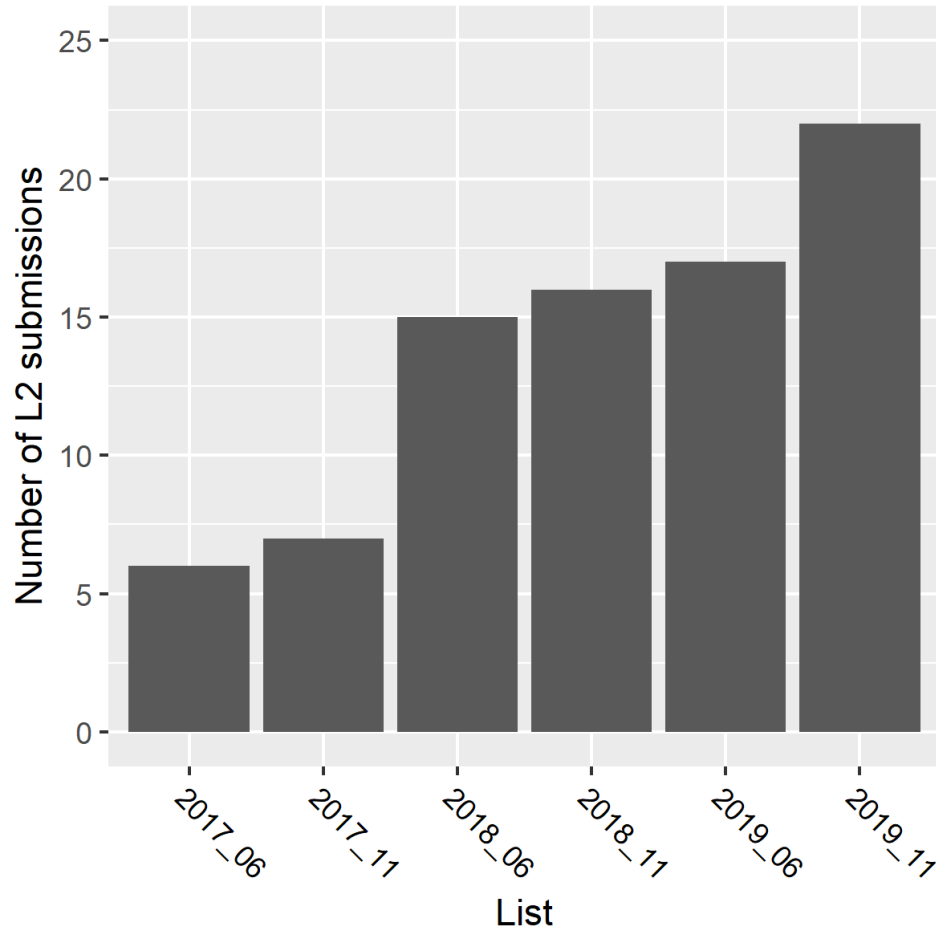
# Who Has Made an L2/L3 Submission?

- 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 Universitaet 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





# 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](mailto:natalie.jean.bates@gmail.com)

# Agenda

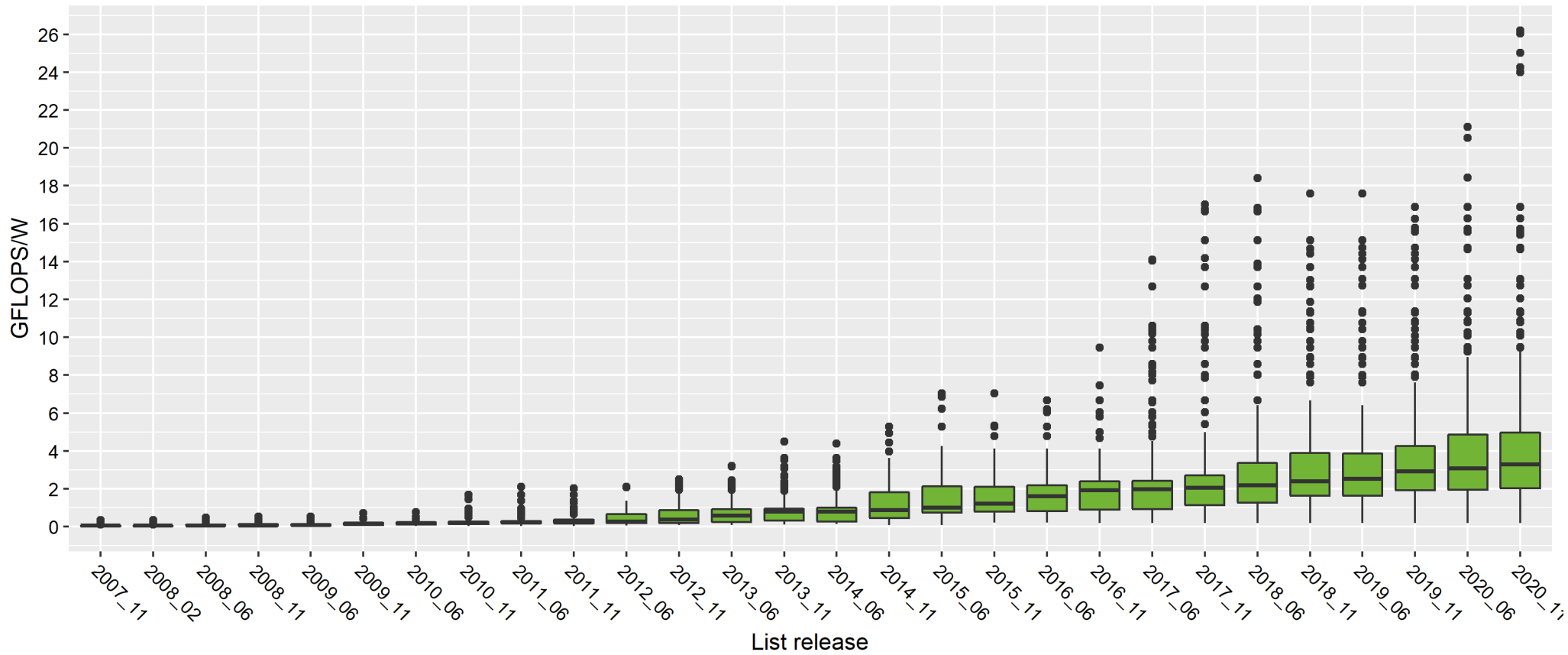
- The Green500 & Its Evolution: Past, Present, Future (*Wu Feng*)  
[ Discussion and Q&A ]
- Status of L1/L2/L3 Measurements (*Natalie Bates*)  
[ Discussion and Q&A ]
- MEGWARE Report on L2/L3 Measurement (*Axel Auweter*)  
[ Discussion and Q&A ]
- The 28<sup>th</sup> Green500 List (*Wu Feng*)
  - Trends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

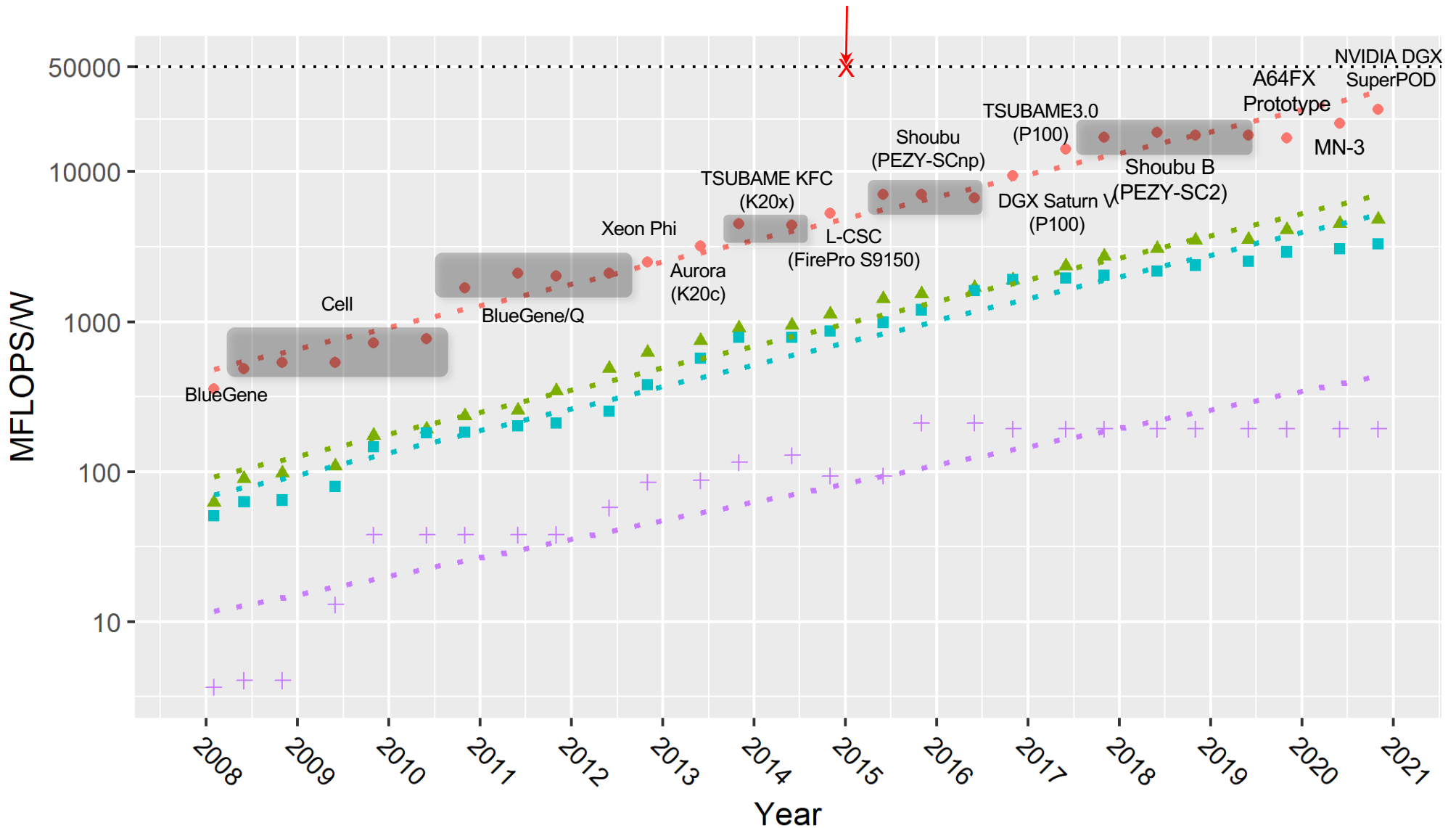
# Trends: How Energy Efficient Are We?





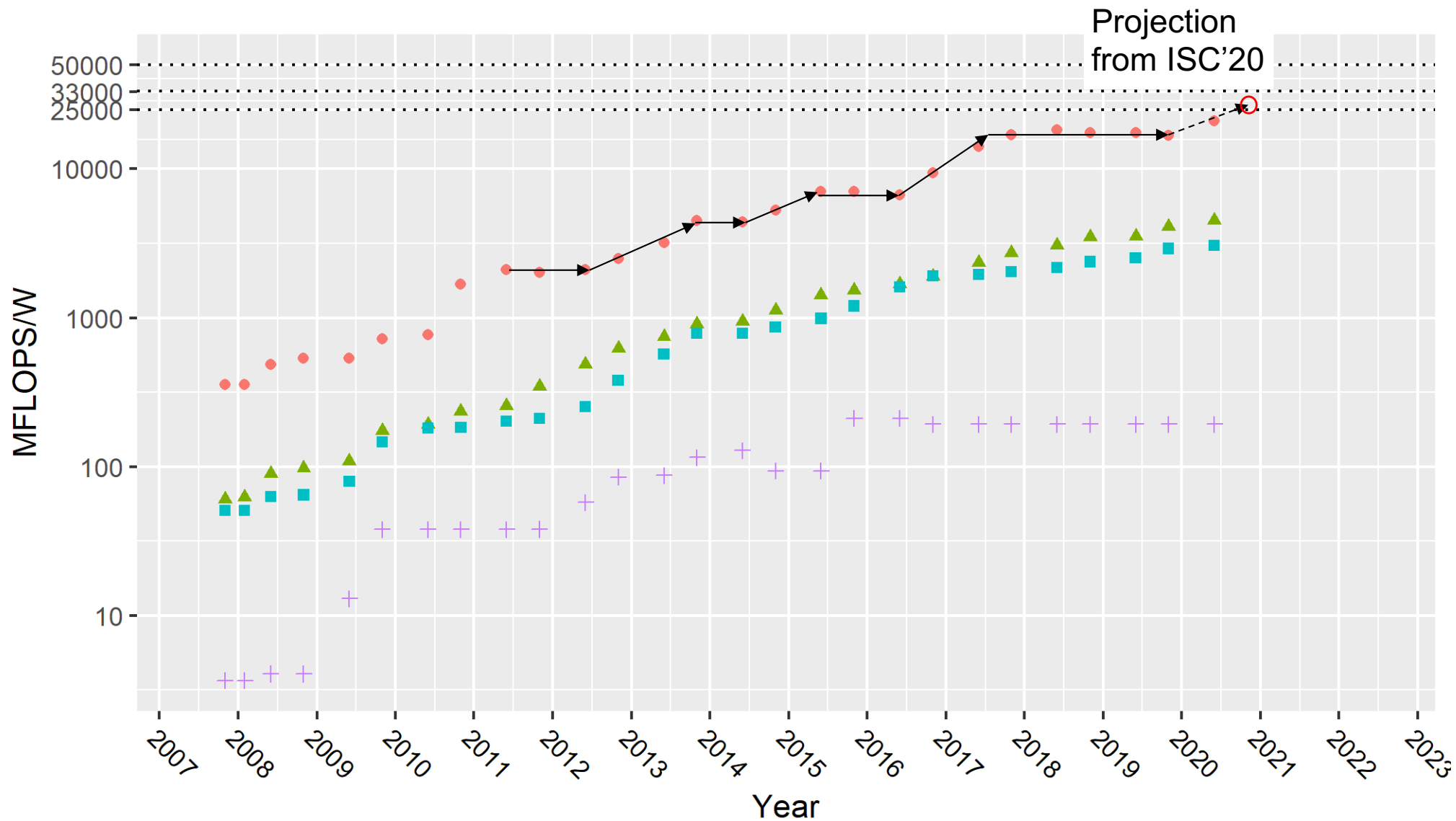
# Trends: How Energy Efficient Are We?

Green500 Rank    Top    Mean    Median    Bottom



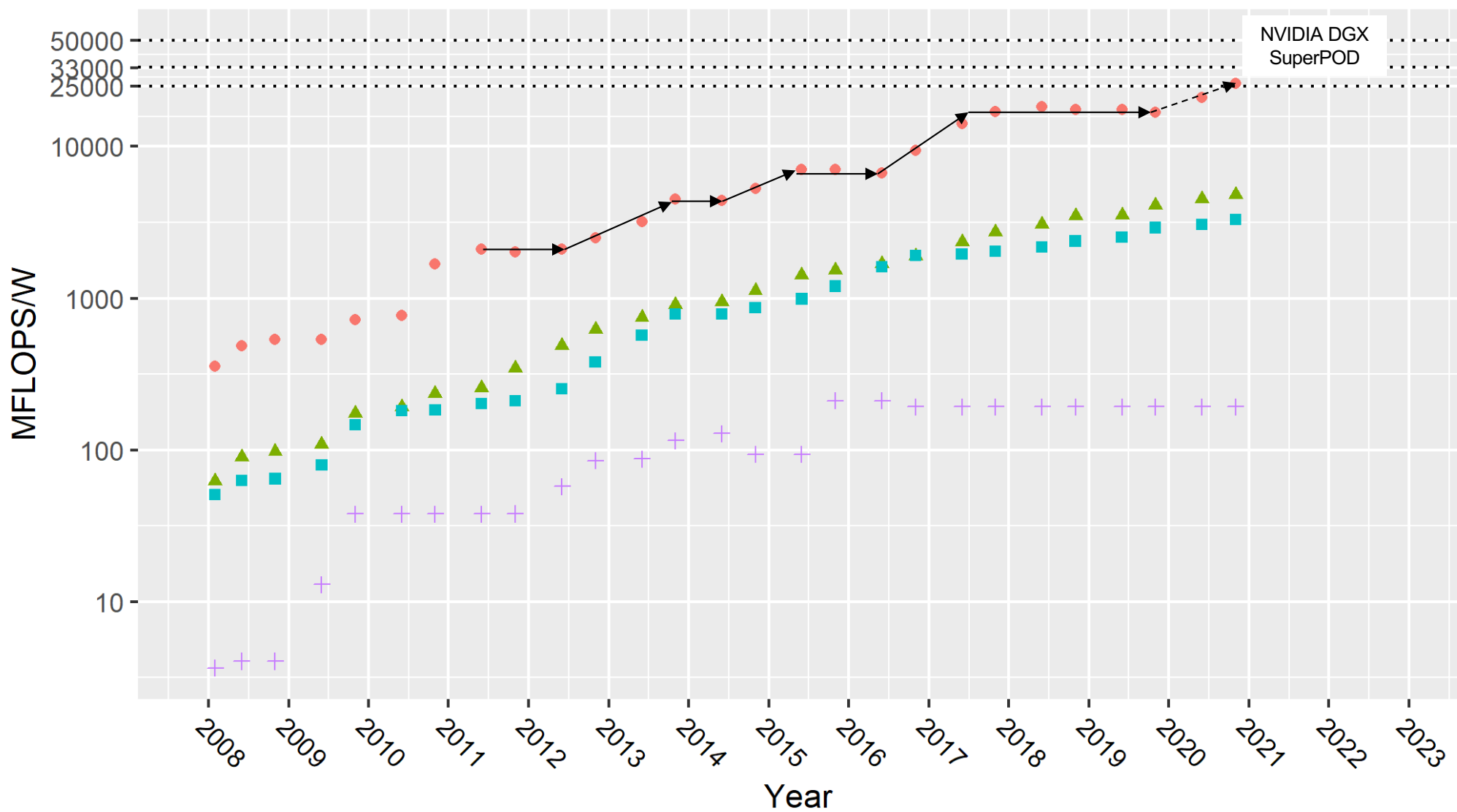
# Trends: How Energy Efficient Are We?

Green500 Rank ● Top ▲ Mean ■ Median + Bottom

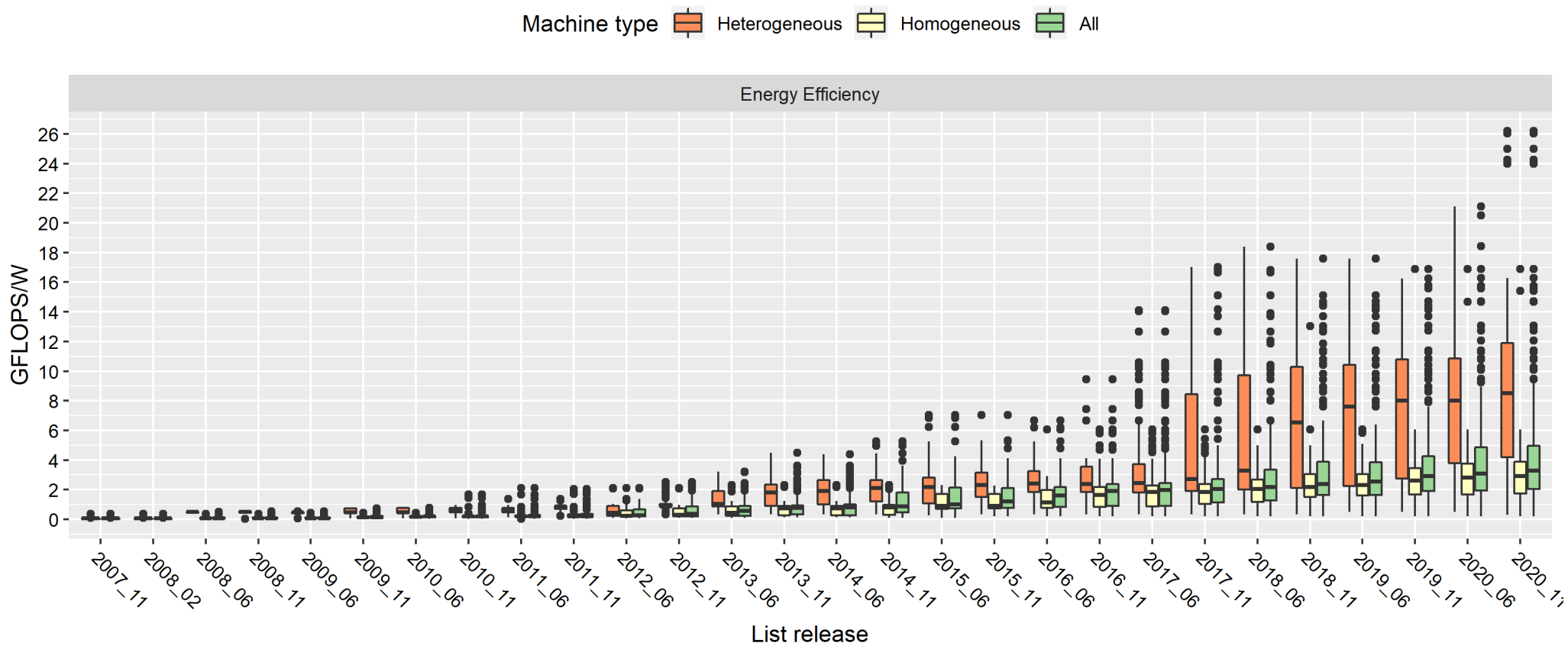


# Trends: How Energy Efficient Are We?

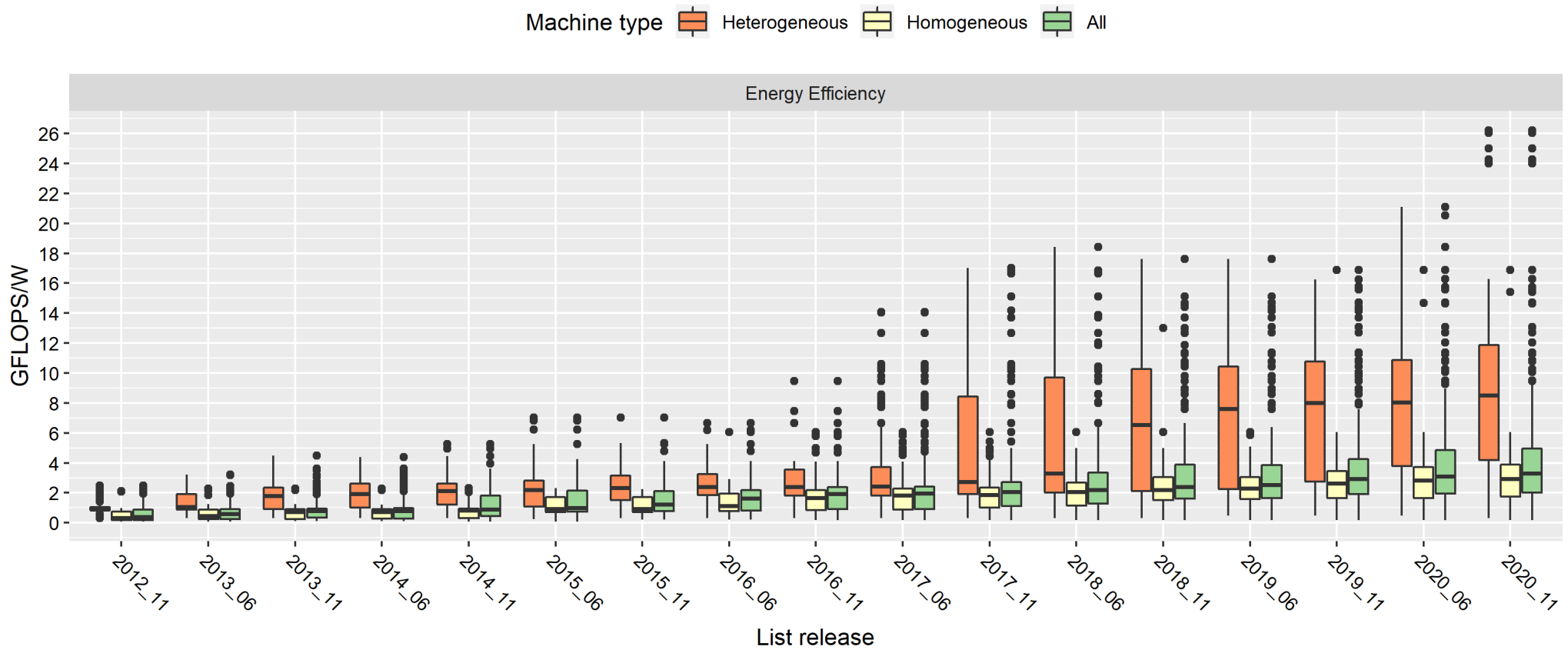
Green500 Rank ● Top ▲ Mean ■ Median + Bottom



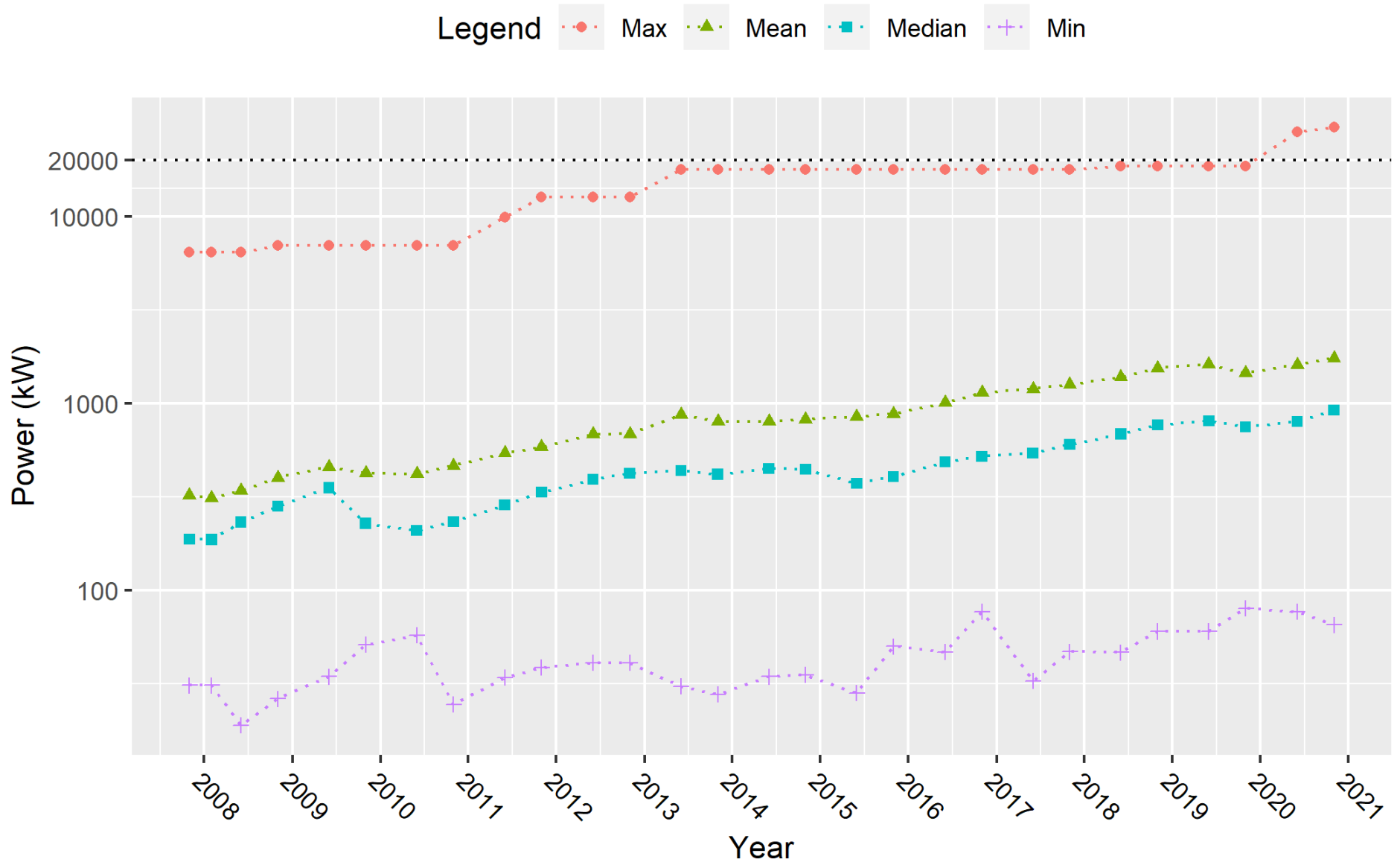
# Trends in Efficiency (2007 – Present): Homogeneous vs. Heterogeneous Systems



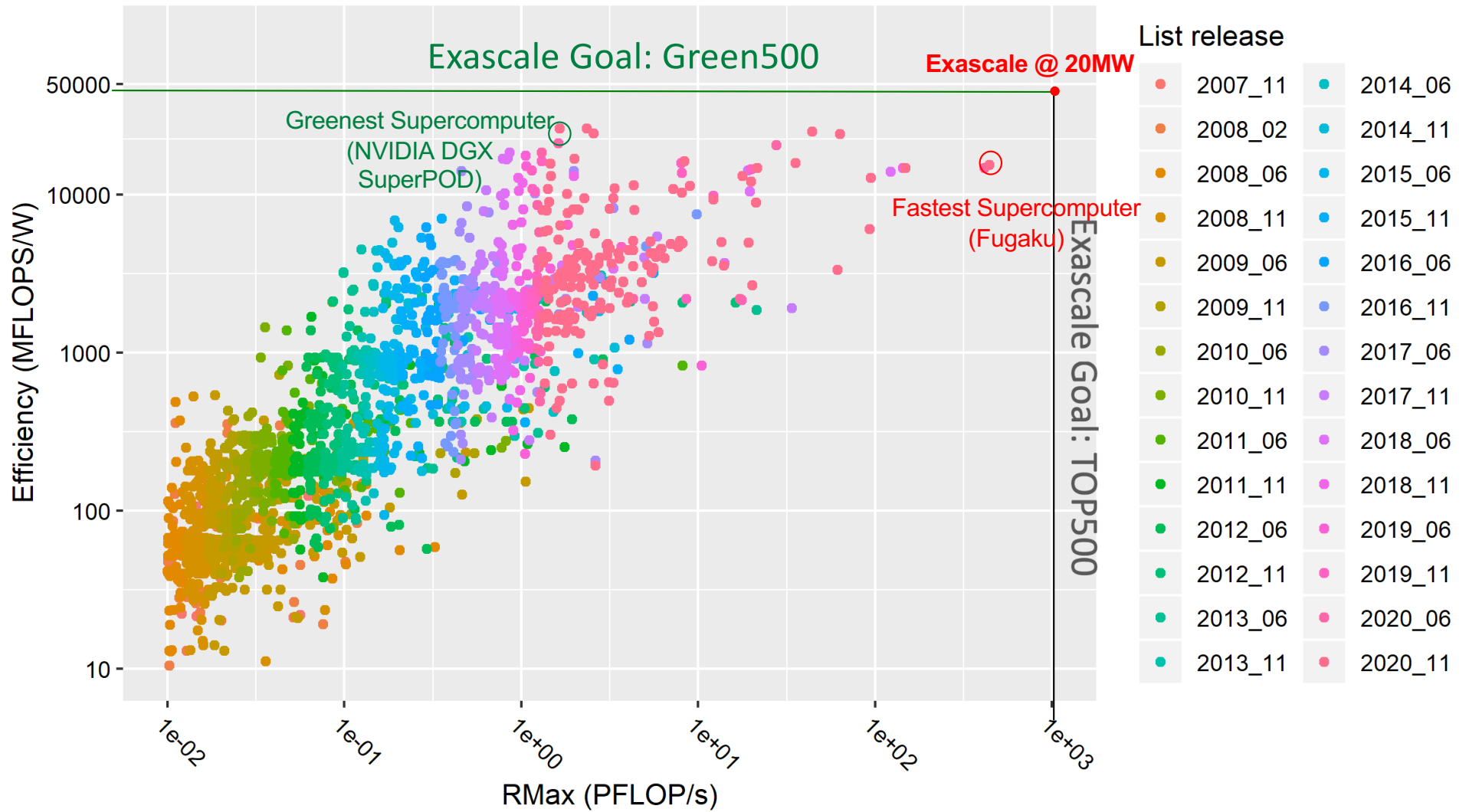
# Trends in Efficiency (2012 – Present): Homogeneous vs. Heterogeneous Systems



# Trends in Power: Max, Mean, Median, Min

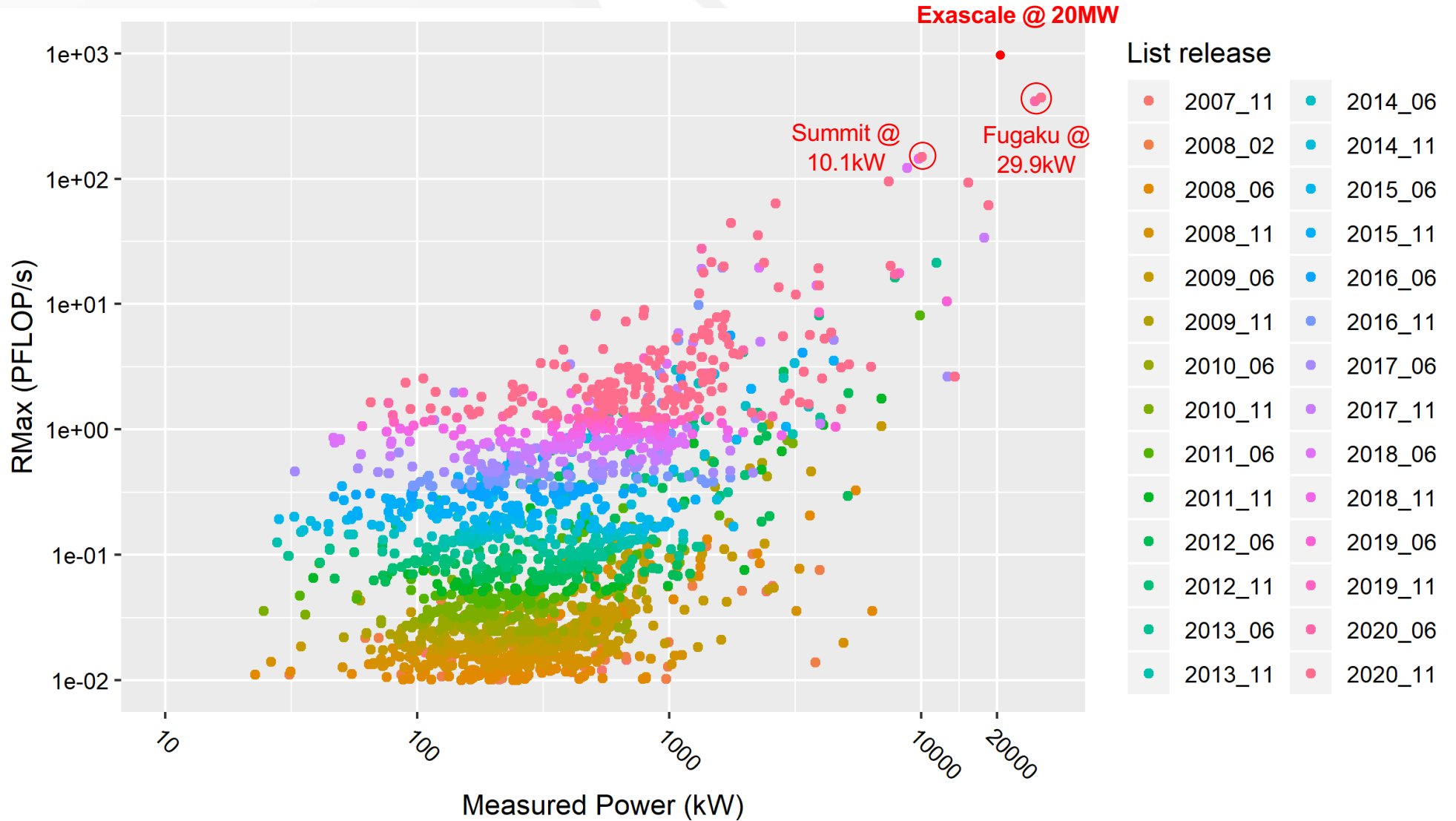


# Efficiency vs. Performance

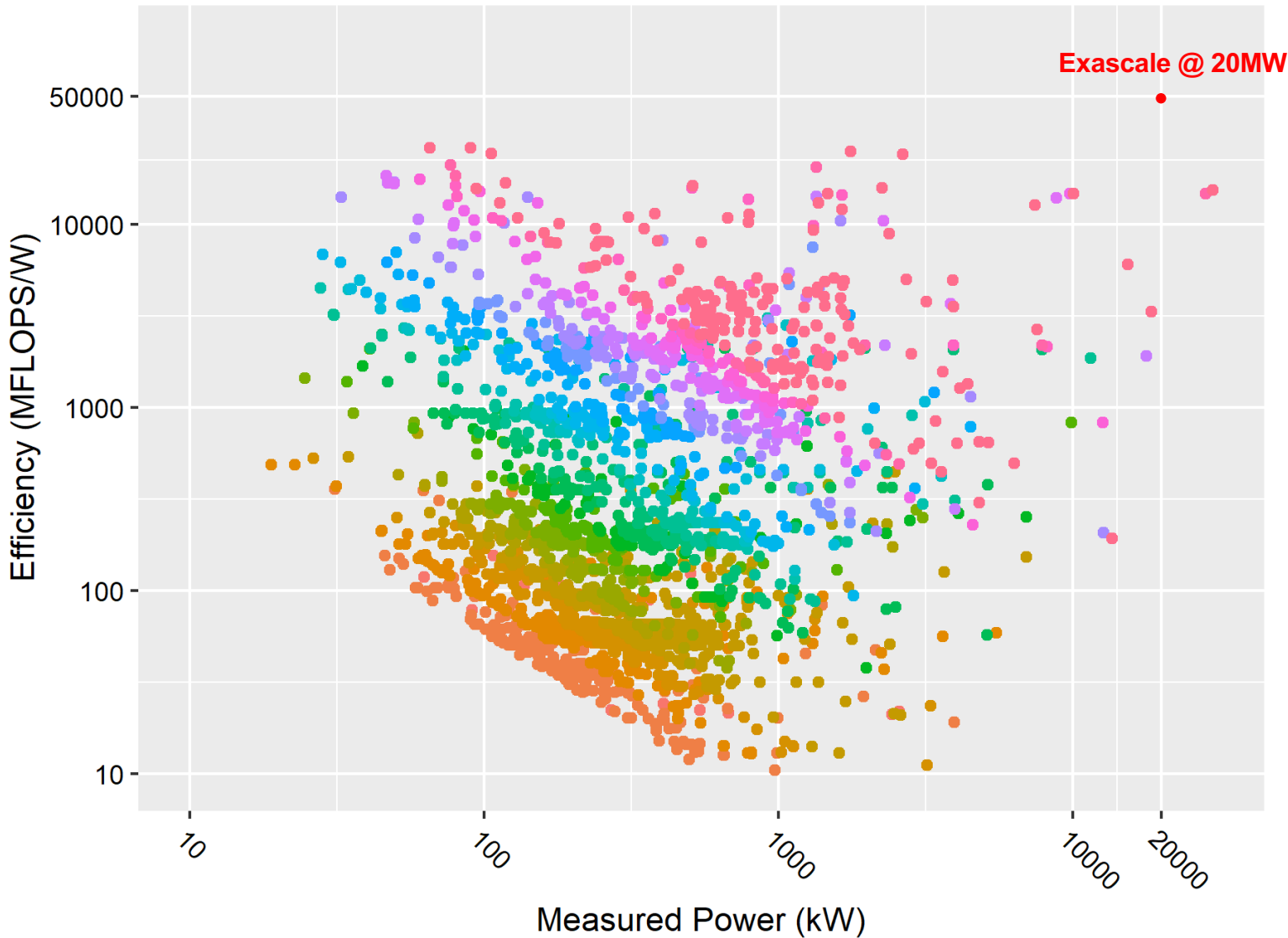




# Performance vs. Power



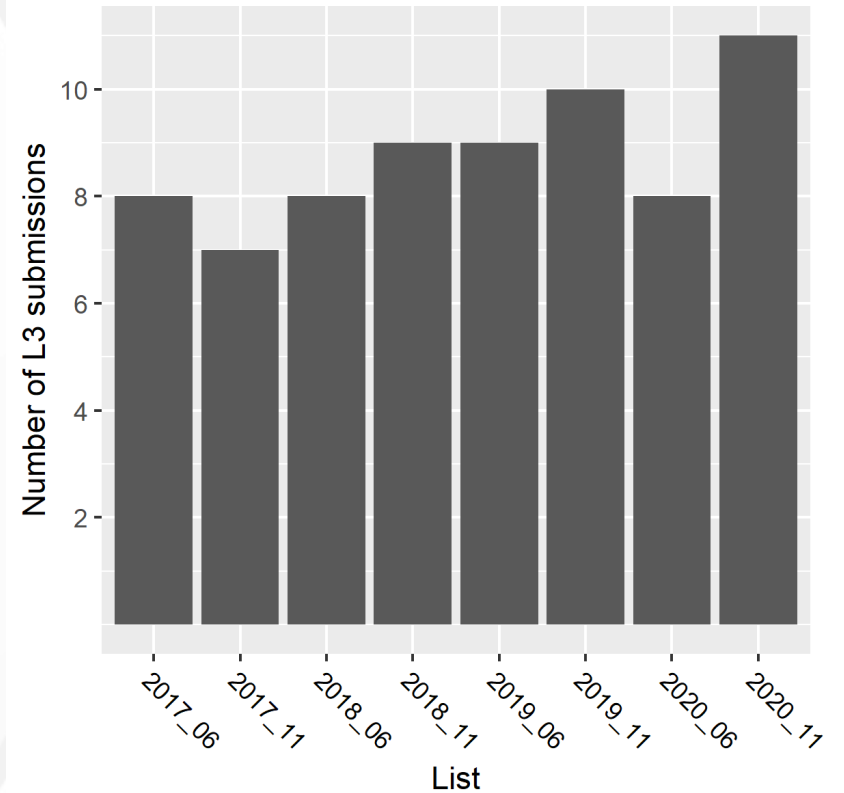
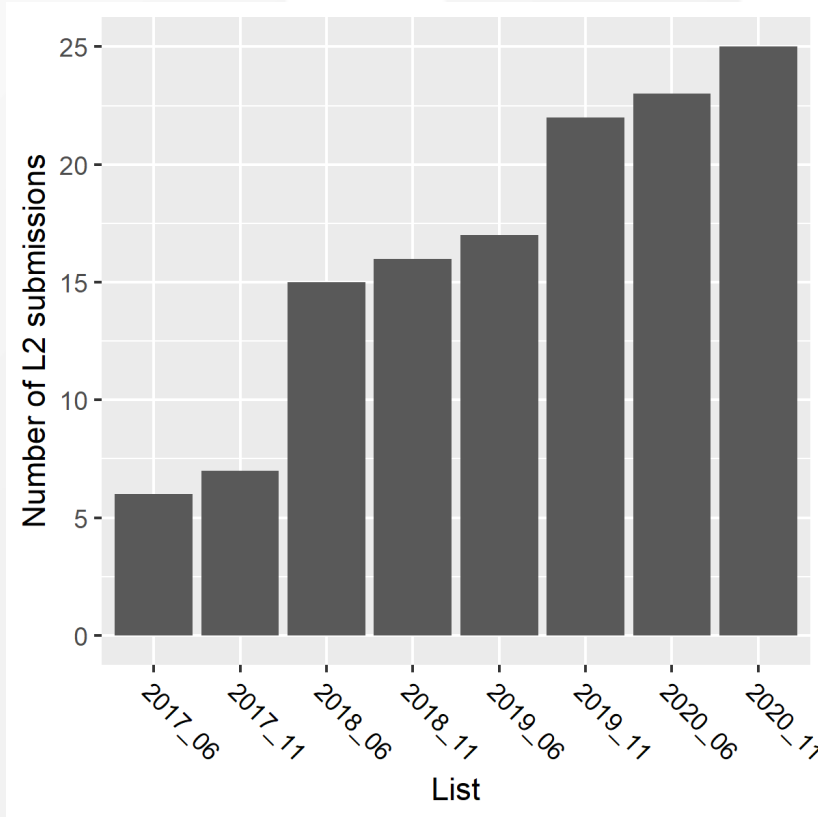
# Efficiency vs. Power



## List release

2007_11	2014_06
2008_02	2014_11
2008_06	2015_06
2008_11	2015_11
2009_06	2016_06
2009_11	2016_11
2010_06	2017_06
2010_11	2017_11
2011_06	2018_06
2011_11	2018_11
2012_06	2019_06
2012_11	2019_11
2013_06	2020_06
2013_11	2020_11

# Level 2 and Level 3 Submissions over Time





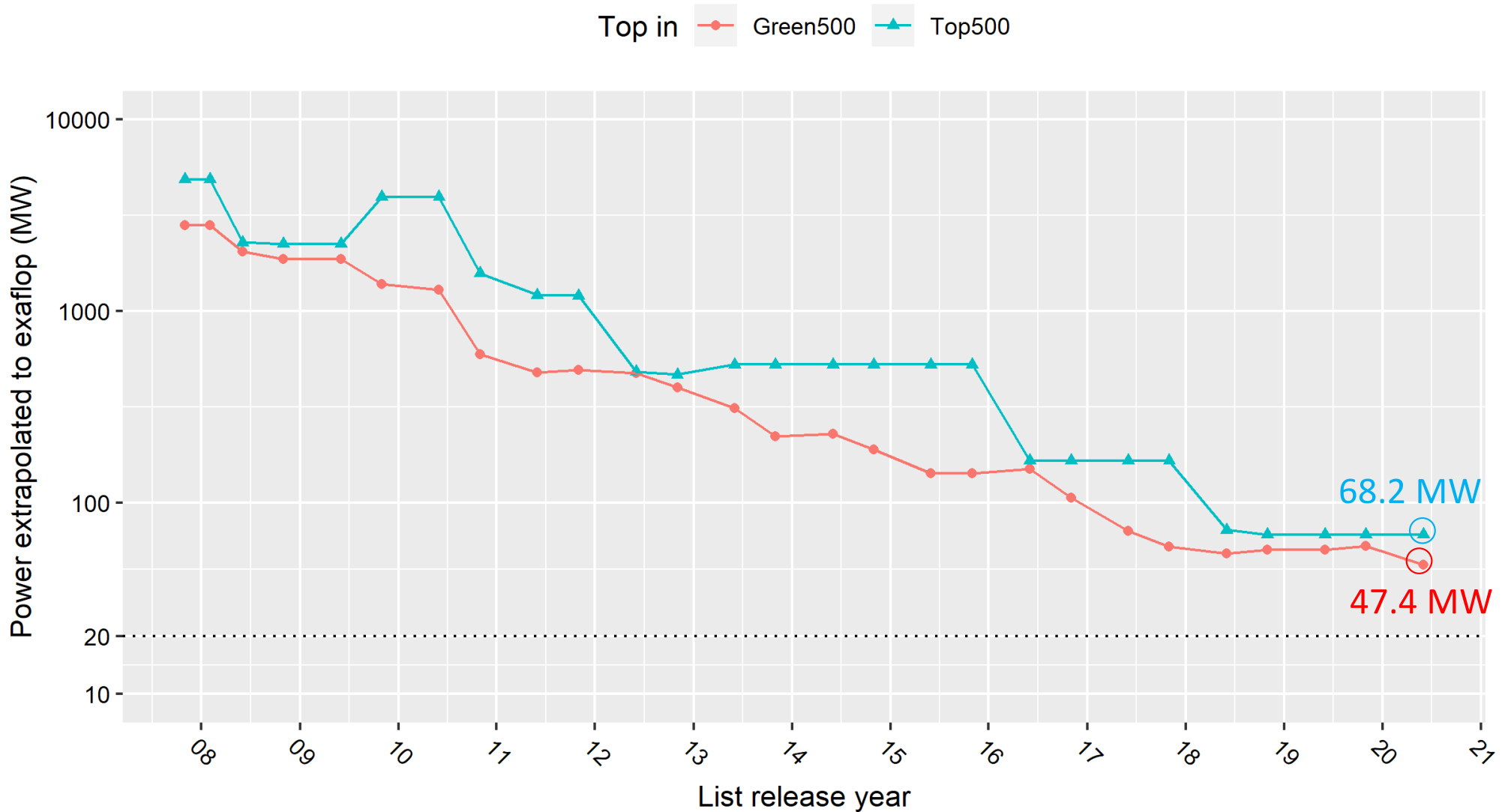
# Trends Towards Exascale

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

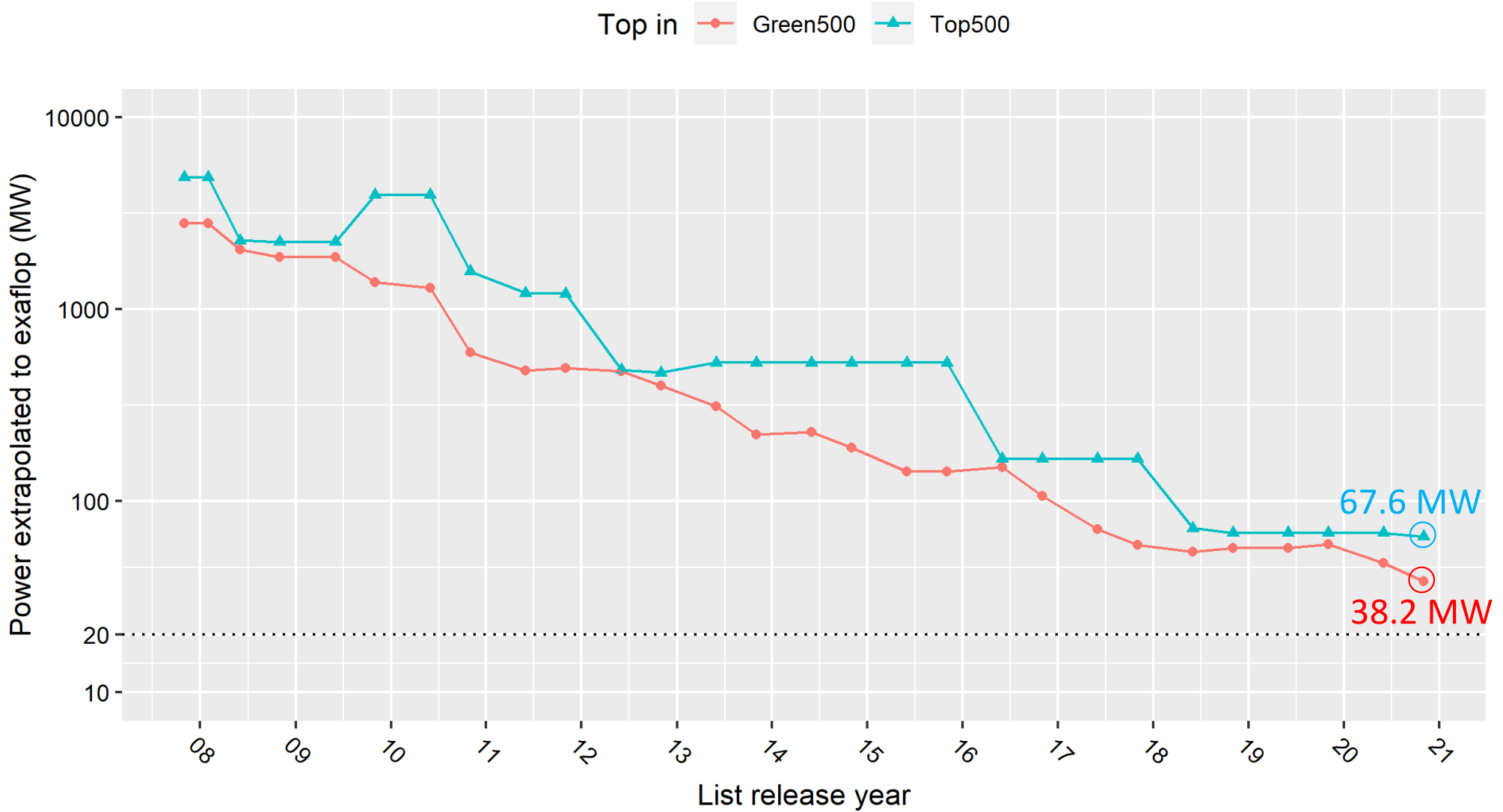
# Exascale Computing Study: Technology Challenges in Achieving Exascale Systems

- 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)



# Trends: Extrapolating to Exaflop (June 2020)





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

# 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



# CERTIFICATE

**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

A handwritten signature in black ink, appearing to read 'Wu-chun Feng', written over a horizontal line.

Wu-chun Feng  
Virginia Tech

A handwritten signature in black ink, appearing to read 'Kirk Cameron', written over a horizontal line.

Kirk Cameron  
Virginia Tech



# CERTIFICATE

## **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

A handwritten signature in black ink, appearing to read 'Wu-chun Feng', written over a horizontal line.

Wu-chun Feng  
Virginia Tech

A handwritten signature in black ink, appearing to read 'Kirk Cameron', written over a horizontal line.

Kirk Cameron  
Virginia Tech



# Acknowledgements

<https://www.top500.org/lists/green500/>

- 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