Emerging Trends on the Evolving Green500: Year Three

Tom Scogland, Balaji Subramaniam, and Wu-chun Feng
{tom.scogland,balaji,wfeng}@vt.edu
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060

Abstract

It has been traditionally viewed that as the scale of a supercomputer increases, its energy efficiency decreases due to performance that scales sub-linearly and power consumption that scales at least linearly with size. However, based on the first three years of the Green500, this view does not hold true for the fastest supercomputers in the world. Many reasons for this counterintuitive trend have been proposed — with improvements in feature size, more efficient networks, and larger numbers of slower cores being amongst the most prevalent.

Consequently, this paper provides an analysis of emerging trends in the Green500 and delves more deeply into how larger-scale supercomputers compete with smaller-scale supercomputers with respect to energy efficiency. In addition, our analysis provides a compelling early indicator of the future of exascale computing. We then close with a discussion on the evolution of the Green500 based on community feedback.

I. Introduction

Like the unanticipated SPUTNIK I launch in 1957, the arrival of the Japanese Earth Simulator in 2002 became known as *Computenik* due to its unanticipated computing prowess, which obliterated U.S. domination in supercomputing. Unlike the space race begun by SPUTNIK I, the "arms race in supercomputing" has yet to end. The emphasis on speed has resulted in the construction of supercomputers that consume exorbitant amounts of energy and require elaborate cooling facilities to function [1], [2], [10].

To address this issue, we created the Green500 [4] with the goal of bringing greater visibility to the energy efficiency of supercomputers and to provide a complementary view to the TOP500 [11]. As the race towards building an exascale supercomputer continues, the high-performance computing (HPC) community has come to realize the importance of "being green" in supercomputing. Thus, in this paper, we track the progress of green HPC, analyze trends that have emerged over time, and deliver insights and implications for the future of supercomputing.

With the DARPA IPTO exascale study clearly indicating that power will be *the* obstacle to exascale computing [3], the tracking and analysis of power and energy efficiency by the Green500 will provide early indicators for exascale computing and its projected 68-megawatt (MW) power envelope. Two trends in the Green500 suggest it may well be possible to meet, and perhaps exceed, current expectations of the energy efficiency of the first exascale supercomputer. First, while it is well known that FLOPS/watt scales sub-linearly as the size of the machine its run on increases [8], the upper ranks of the Green500 are not consistently populated with smaller machines. Quite the contrary in fact, many of the greenest machines are larger machines. We analyze a few common explanations for this phenomenon in this paper. In addition to the discussion of scaling, we present implications of scaling and other trends in efficiency to the development of exaflop machines, including what it would take to power one today.

The rest of the paper is organized as follows. Section II provides background on the motivation behind the Green500. Section III presents an analysis of high-level trends in the Green500 along with their implications. The scaling of energy efficiency at the very high end is discussed in Section IV. A discussion of the insights that the Green500 provides into the future of exascale computing follows in Section V. Section VI presents the evolution of the Green500 list and its practices. Finally, Section VII presents concluding remarks and a discussion of future work.

This work was supported in part by NSF grant CCF-0848670, a DoD National Defense Science & Engineering Graduate Fellowship (NDSEG), and Supermicro.

II. Background

In the past, energy efficiency was the last item on the minds of supercomputer designers. Consequently, in 2008, the annualized energy cost for a single 1U server surpassed its purchase cost, as shown in Figure 1 from Belady [2]. In addition, the focus on performance, as defined by speed, translates into a a fast-rising total cost of ownership (TCO), as implicitly evidenced by the fast-rising annual infrastructure and energy (I&E) cost in Figure 1.

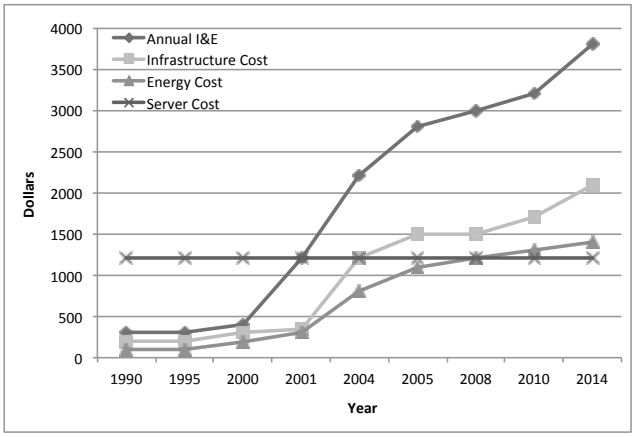

Fig. 1. Annual amortized costs in the data center

By 2014, the infrastructure and energy (I&E) cost has been projected to account for 75% of the TCO of a 1U server in [2]. Take that in comparison to the early 1990s, when that same cost accounted for merely 20% of the TCO. Simply continuing to "feed the beast," which supercomputing has become, is no longer an environmentally or financially sustainable option. As such, the energy efficiency of supercomputers must be improved. Beyond the single computer, larger groups of machines, such as datacenters and HPC centers, are already feeling constrained by the physical limits of incoming and cooling wattage, like those of the National Security Agency (NSA) [7], [13], Google [10] and Yahoo! [6].

As a result of the pressing need for improvements in energy efficiency, the mission of the Green500 is to *raise awareness* of energy efficiency in the supercomputing community and bring energy efficiency to the same level of importance as performance, as defined by speed. Currently, the Green500 list uses the LINPACK benchmark, provided by the TOP500, to rank the most energy efficient supercomputers in the world. The workload is an algorithm to solve a dense linear system of equations of the form $Ax = b$ of the order N. It performs LU factorization on the coefficient matrix and computes the solution by backward substitution. The data is distributed on a two-dimensional PxQ grid using a cyclic scheme for better load balancing and scalability. The benchmark reports performance in terms of floating-point operations per second (FLOPS).

In line with the goals of the Green500 and feedback from the HPC community, three additional exploratory lists were created in 2009 and described in [5]. These lists were named the Little Green500, the HPCC Green500, and the Open Green500. The Little Green500 came in response to requests for a broader interpretation of what constitutes a supercomputer, particularly given that the focus of the Green500 is *not* performance (as defined by speed) but energy efficiency. It also took a cue from the belief that performance does not scale as quickly as power, thus allowing smaller machines to achieve higher efficiency than their larger counterparts.

The HPCC Green500 is, as its name implies, a list that is intended to adopt the "High Performance Computing Challenge" (HPCC) benchmarks [9] in place of LINPACK. The HPCC benchmark suite was created to overcome the limitations of the LINPACK benchmark. The LINPACK benchmark only stresses the raw computing power of the processor in the supercomputer. In contrast, the HPCC benchmark suite stresses multiple components of a supercomputer, including the processor, memory, and network interconnect, and examines the system with greater variety of memory access patterns than the LINPACK benchmark. The benchmark suite consists of seven benchmarks and 28 tests, which have memory access patterns ranging from low to high spatio-temporal data locality. Lastly, the Open list came in response to the demand for less stringent run rules than those usually demanded by the Green500 and the TOP500 — specifically, a lifting of the restriction on using mixed-precision floating point in the solver of LINPACK.

III. Analysis

There are certain trends that we track across the years, showing us how the world of supercomputing is becoming greener and where the innovation is coming from. In this section, we present an analysis of the Green500 and the progress made in green supercomputing over the past three years.

We have observed a persistent and steady climb in the energy efficiency of the machines on the list, mostly in the top half of each list. The average efficiency, as shown in Figure 2, has nearly tripled in the last three years. More impressively though, the maximum energy efficiency more than doubled over the last *six months*. In June 2010, the three most efficient supercomputers on the Green500 were the QPACE clusters in Germany at 773.4 MFLOPS/watt. On the Little Green500 in June, the most efficient supercomputer was GRAPE-DR at 815.4 MFLOPS/watt.

By November 2010, the BlueGene/Q supercomputer, an in-house prototype of what will become the Sequoia supercomputer at Lawrence Livermore National Laboratory, more than doubled GRAPE-DR's energy efficiency with 1684.2 MFLOPS/watt. Never before in the short history of the list has the maximum jumped so far so fast. Even the #2 machine, the GRAPE-DR special-purpose cluster, doubles the efficiency of the previous leader, i.e., QPACE.

It has been a topic of interest in our previous analyses as to whether the energy efficiency of supercomputing would track Moore's Law over time. As of last year, the curve was on track, with the average energy efficiency doubling over the first 24 months of the existence of the Green500. Now with 36 months behind us and after the jump in maximum energy efficiency provided by BlueGene/Q and GRAPE-DR, the maximum energy efficiency has increased *five-fold* in 36 months. Average efficiency, on the other hand, remains on track, having increased three-fold in the same time frame.

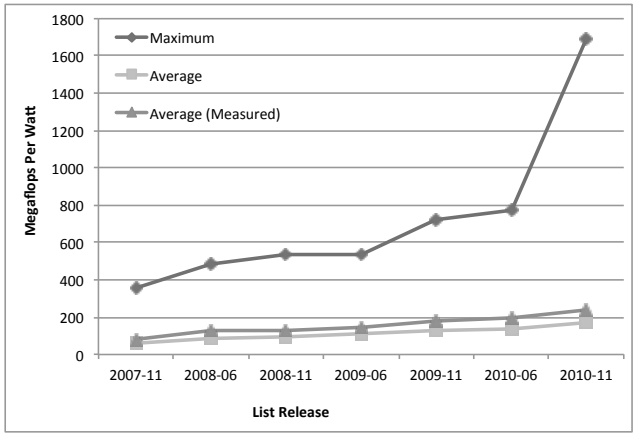

Fig. 2. Efficiency statistics across Green500 releases.

Overall efficiency, as in Figure 2, does not really show the whole picture though. While Figure 2 shows the efficiency in terms of flops per watt, it ignores how much of the machine is being wasted through performance inefficiency. Figure 3 shows the performance efficiency (calculated as $(R_{max}/R_{peak}) * 100$) versus the Green500 ranking of the systems in the June 2010 and November 2010 lists. Clearly, machines which had less than 60% performance efficiency made it to the top of both the lists. For example, the Mole cluster with performance efficiency as low as 18% occupied the 8th and 19th positions in June 2010 and November 2010, respectively.

This trend has high correlation with the emergence of accelerator-based systems. Such supercomputers currently come in two flavors: (1) Cell-based systems, such as

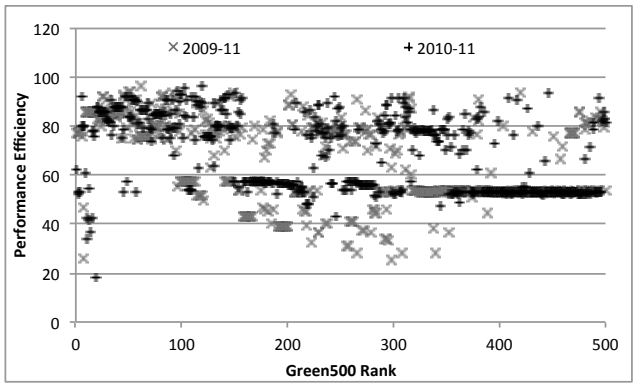

Fig. 3. Performance efficiency of systems over time

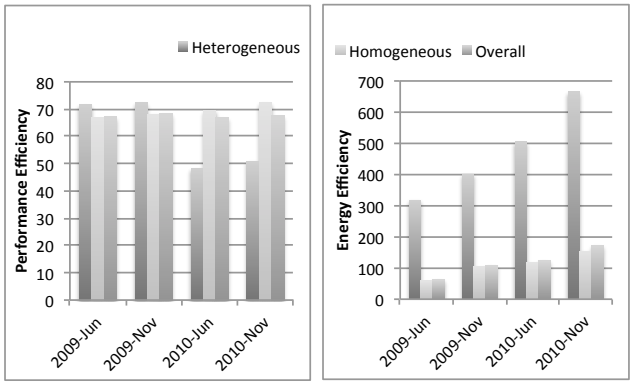

Fig. 4. Performance efficiency of heterogeneous and homogeneous systems over time

QPACE and (2) GPU-based systems, such as TSUBAME-2.0. To further investigate this trend, we analyzed the performance and energy efficiency of heterogeneous and homogeneous systems over time. Figure 4 shows the performance and energy efficiency of heterogeneous and homogeneous systems over time. The performance efficiency of heterogeneous systems during last year was well over 70%. However, we see a decline in the performance efficiency of heterogeneous systems this year. The cause is that Cell-based supercomputers dominated the heterogeneous part of supercomputing last year, while this year saw the dominance of GPU-based supercomputers on the Green500 list.

Why do GPU-based supercomputers have such low performance efficiencies? Traditionally, GPUs are meant to perform embarrassingly parallel, single-precision calculations. However the benchmark used for ranking the systems in Green500, LINPACK, is neither embarrassingly parallel nor are its computations single-precision. Therefore, performance efficiency is reduced. In spite of these

	2007	2008	2009	2010
Average Rank	76	123	162	106
Lowest Rank	176	496	445	404
Highest Rank	1	1	2	1

TABLE I. Statistics on the TOP500 ranks of the 30 greenest supercomputers over time

issues, the energy efficiency of heterogeneous systems has increased with the emergence of more GPU-based supercomputers. These systems are highly energy efficient, as the presence of only 15 such system increases the average efficiency of the entire list from 156 MFLOPS/watt (average energy efficiency of homogeneous systems) to 173 MFLOPS/watt in November 2010. We expect the energy efficiency of the list to reach new highs with the emergence of GPUs that perform double-precision operations more efficiently.

IV. Efficiency of Scale

The FLOPS/watt metric used by the Green500 has long been a source of heated debate. One of the many concerns is that it is biased towards smaller supercomputers, as noted in [5], [8]. The performance of scientific benchmarks, such as LINPACK, do not scale linearly with respect to processor or node count, whereas power consumption scales at least linearly. As a result, smaller supercomputers should have better energy efficiency according to the FLOPS/watt metric. However, as mentioned earlier, the data in the Green500 list does not universally support this relationship. As illustrated in Table I, while smaller machines that are ranked lower in the TOP500 do appear high on the Green500, the average rank of machines from the greenest 30 are in the top 10-15% of the TOP500. Furthermore, in all the Green500 lists thus far, the fastest supercomputer in the world (as ranked by the TOP500) was also amongst the 30 greenest in the world. The only exception came in 2009 when Jaguar was the fastest supercomputer in the world (i.e., #1 on the TOP500) but was not amongst the 30 greenest supercomputers. Discussions with the community have brought up several potential causes for the phenomenon, three of which we discuss below.

A. Processor Minimum Feature Size

The idea here is that the smaller the transistors on a chip, the more efficient the chip may be. Because the fastest machines are at the bleeding edge, we expect that these machines will possess the newer processor fabrication technology first. If this were the case, then

Interconnect	2007	2008	2009	2010
Custom/Proprietary	26	12	18	16
InfiniBand	4	18	12	13
Gig-E	0	0	0	1
Green 30 % custom	87%	40%	60%	53%
Overall % custom	13%	12%	9%	10%

TABLE II. Interconnect statistics for the greenest 30 machines

it would stand to reason that the added efficiency from a more efficient use of die space and lower energy per transistor would account for newer and faster machines to rise to the top of the Green500. That said, it does not hold out. Figure 5 shows that the greenest machines have had larger feature size (on average) since 2007. Only with the latest Green500 have the greenest machines had the smaller feature size (on average). In fact, up until this year, the greenest machines averaged a less dense manufacturing process than both the rest of the list and the top performing supercomputers.

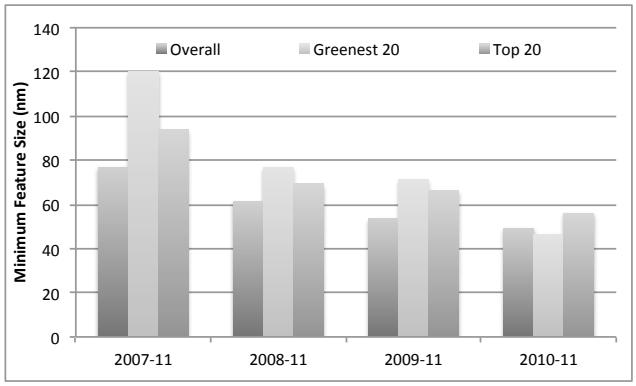

Fig. 5. Average minimum feature size

B. Custom Interconnects

Another indicator of machines that fare well on the Green500 list, despite their size, is their interconnect. As shown in Table II, the 30 greenest machines on the list are traditionally either custom or InfiniBand-based interconnects. The custom networks in machines like those of the BlueGene architecture, QPACE, and K-Computer by Fujitsu provide an efficiency edge. While we expected to find that the custom interconnects improved performance efficiency, and thus energy efficiency, some machines, such as QPACE, possess custom interconnects and deliver worse performance efficiency but better energy efficiency. This mattered most for the lists from 2007-2009. In the 2010 list, however, the faster energy-efficient machines use

InfiniBand, as in the case of TSUBAME-2.0 and GRAPE-DR, or Gigabit Ethernet, in the case of EcoG. While this has been an important consideration in the past, the fact that InfiniBand and even Gigabit Ethernet can bring large machines to the top suggests that it is not the whole story.

C. More Processing Elements

The last and best correlating cause is the number of processor cores in each node. In the first list in 2007, most nodes had one to four CPUs of one or two cores each. The top 20 however were composed entirely of BlueGene systems, which contained 64 cores per node in BlueGene/L and 128 cores per node in BlueGene/P. In 2008, the 19 greenest machines were either BlueGene/P or Cell Broadband Engine (Cell BE) nodes. While Cell BE only offered 18 cores per node, it provided a higher bandwidth interconnect between the cores. In 2009, this trend was further reinforced not only by BlueGene/P and Cell BE machines but also by the GRAPE-DR supercomputer, which contains 16,000 cores in every node, and the NUDT TH-1 GPU-accelerated supercomputer, comprised of more than 1,600 cores per node (including GPU cores). Finally, in 2010, the average number of cores per node in the greenest 10 supercomputers rose to more than 2,300. The average for the rest of the list was in the range of 10-15 cores per node. Invariably, the architectures with more tightly coupled and smaller compute units rose towards the top of the Green500. Even so, there was a new entry in the greenest 10 this past year, the Fujitsu K computer used conventional SPARC64 processors with only tens of cores to achieve the rank of the fourth most efficient supercomputer on the Green500.

D. Summary

While each of the three causes discussed here have shown some correlation with the rise of faster machines on the Green500 list, none is solely responsible. It seems that it is a combination of these factors and more. This suggests that moving toward efficient interconnects and more tightly coupled processing elements will help as we face the challenge of exascale.

V. Implications for Energy-Efficient Exascale Supercomputing

The RoadRunner supercomputer broke the petaflop barrier in June 2008 and since then DARPA's Exascale Computing Study [3] has shown that the HPC community has shifted its focus to achieving the next major milestone in performance, an exascale system. However, as mentioned in the report, exascale systems are predicted to consume 68 megawatts of power even when making highly optimistic assumptions. Consequently, the HPC community is beginning to realize that power consumption is one of the biggest impediments to designing such large scale systems. In this section, we seek to glean insights into the future of multi-petaflop and exaflop computing.

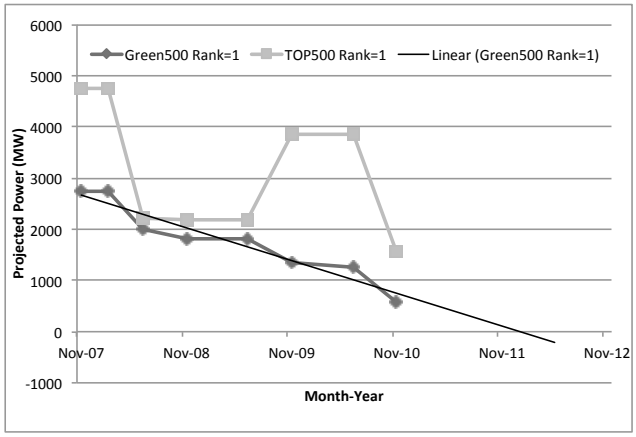

Fig. 6. Power projections for exascale systems

Figure 6 shows the predicted power of an exascale system when projected from the highest ranked system on the Green500 and TOP500 lists over time. While this is a naïve projection, based on assuming that both power and performance scale linearly, it gives a reasonable approximation of the cost to power an exascale machine made out of today's hardware. While the exascale projection as of November 2010 is still more than 500 megawatts (MW), the downward trend in projected power consumption of an exascale system is encouraging.

Most recently, the energy efficiency of the highest ranking machine on the Green500 list has seen more than two-fold increase since June 2010. However, projections of power consumption, when extrapolated from today's #1 machine on the TOP500, show that an exascale supercomputer would consume over 1.5 *gigawatts*. The increase in projected power consumption for an exascale supercomputer from November 2009 to June 2010 is due to the fact that the highest-ranked machine on the TOP500 was the Jaguar supercomputer, a large-scale traditional cluster. This emphasizes the importance of innovative designs for achieving energy-efficient exascale supercomputing.

Something we look forward to seeing is how this trend progresses over the coming two to three years. As shown by the trend line in Figure 6 if things continue as they are now, our supercomputers will be *generating* energy in approximately 18 months. Given that computers generating power is a most unlikely outcome, it will be interesting to

see how the progression of efficiency changes over the coming years.

VI. Evolution of The Green500

In this section, we present the evolution of *The Green500* with respect to its run rules and exploratory lists.

A. Official Run Rules

The official run rules for the Green500 continue to evolve. With the latest November 2010 release, the Green500 now accepts submissions of a machine subset. Specifically, the run rules allow for a partial machine to be submitted, so long as the total performance of the partial machine remains faster than the 500th-ranked machine on the TOP500 list.

The intent of the aforementioned change was to combat the projected trend of performance scaling sub-linearly while power consumption scaled at least linearly. While Section IV shows that some machines did not find this to be an issue, this change "levels the playing field" for machines of all sizes, as long as they remain faster than the 500th-ranked machine on the TOP500 list.

For the coming year, there has been discussion of further clarifications and modifications to the run rules. For example, NCSA scientists at the University of Illinois propose that a larger portion of the LINPACK run be used to measure power consumption. The current run rules only require power measurement over an inner 20% of the run, specifically the LU factorization phase. NCSA's experience [12], however, showed that the power consumed in the middle of the run showed enough variation and decreased enough over time that they suggest measuring from 10% into the run until 90% into the run, covering 80% and capturing the bulk of the LU factorization phase.

In addition, there have been requests from the community for a more formal specification of our methodologies for *estimating* power consumption of machines for which we do not have measured power available.

B. Exploratory Lists

Last year, the Green500 launched three exploratory lists: Little, Open, and HPCC. Each was designed to address a specific comment or desire of the community: different benchmarks, lower "barrier to entry," and allowance of mixed-precision, floating-point arithmetic, respectively. Now that the lists have existed for a year, the Little list has clearly become the most successful. It has received multiple unique submissions, including the GRAPE-DR in June and the Jazz cluster in November. The Little Green500 now contains 402 machines with measured numbers whereas the main Green500 "only" contains 272 machines with measured numbers, leaving 228 to be derived. On a related note, the Open list has been incorporated into the Little list in order to allow for mixed-precision submissions.

As for the HPCC list, it is active but remains undisplayed on the Green500 web site after a year, due in large part to the limited number of entries. This limited participation was expected for two reasons. First, the HPCC benchmark suite does not report a single number that allows for easy comparison between machines. In fact, it reports many numbers, which must be compared separately, making machines difficult to rank overall. Second, running the HPCC benchmarks and tuning them to run on a machine is a laborious exercise, one that Green500 submitters have voiced as an undue burden. Furthermore, getting these benchmarks to run on heterogeneous systems is a further challenge.

VII. Conclusion

With the third year of the Green500 drawn to a close, this paper presented an analysis of the data collected since its launch. We found that the trends in energy efficiency favor reaching the power goals set by DARPA in [3]. Our analysis also identified certain aspects of design, especially larger numbers of less intelligent cores, that likely contribute towards a higher FLOPS/watt rating.

Overall, this past year was an exciting year for the Green500, with the greatest increase in maximum energy efficiency that we have ever seen as well as the rise of GPUs as both high-performance and highly energy-efficient accelerators.

Based on community feedback, we are considering updating the run rules with particular focus on the following:

- Re-defining the time interval for which power measurement is performed during a LINPACK run.
- Accommodating different types of system design, e.g., provide rules to extrapolate power for machines in which the compute and network racks are separate.

Finally, as highlighted by the exascale computing study [3], the design of an exascale system will require innovations in all the components of the system, not just floating-point units. To encourage such innovations, we plan to study the relation between system parameters such as memory bandwidth and energy efficiency in the future.

Acknowledgements

The authors wish to thank Yang Jiao, Mark Gardner, Heshan Lin, and Kirk Cameron for their support in sustaining the Green500 List and the Green500 community

for their continued support and feedback, particularly participants at the Green500 Birds-of-a-Feather sessions over the past three years.

References

[1] D. Atwood and J. G. Miner. Reducing Data Center Cost with an Air Economizer. White Paper: Intel Corporation, August 2008.

[2] C. Belady. In the Data Center, Power and Cooling Cost More Than the IT Equipment It Supports. *Electronics Cooling Magazine*, 13(1), May 2007.

[3] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, K. Yelick, and P. Kogge. Exascale Computing Study: Technology Challenges in Acheiving Exascale Systems.

[4] W. Feng and K. Cameron. The Green500 List: Encouraging Sustainable Computing. *IEEE Computer*, December 2007.

[5] W. Feng and H. Lin. The Green500 List: Year two. In *Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, pages 1–8. IEEE, 2010.

[6] D. Filo. Serving up greener data centers. http://ycorpblog.com/2009/06/30/serving-up-greener-data-centers/.

[7] S. Gorman. NSA Risking Electrical Overload. In *The Baltimore Sun*, August 2006.

[8] C. Hsu, W. Feng, and J. Archuleta. Towards Efficient Supercomputing: A Quest for the Right Metric. In *IPDPS '05: Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Workshop 11*, page 230.1, Washington, DC, USA, 2005. IEEE Computer Society.

[9] P. Luszczek, D. Bailey, J. Dongarra, J. Kepner, R. Lucas, R. Rabenseifner, and D. Takahashi. The HPC Challenge (HPCC) benchmark suite. In *SC06 Conference Tutorial*. Citeseer, 2006.

[10] J. Markoff and S. Hansell. Hiding in Plain Sight, Google Seeks More Power. In *The New York Times*, June 2006. http://www.nytimes.com/2006/06/14/technology/14search.html.

[11] H. Meuer. The TOP500 Project: Looking Back over 15 Years of Supercomputing Experience. www.top500.org, 2008.

[12] C. Steffen. Personal communication, 2010.

[13] NSA Electrical Power Upgrade, January 2008. http://cryptome.org/nsa010208.htm.