Computer Science and Information Systems 13(1):151–171 DOI: 10.2298/CSIS150228063F

Current Prospects Towards Energy-Efficient Top HPC Systems

Sonja Filiposka[1,2], Anastas Mishev[1], and Carlos Juiz[2]

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, R. Macedonia
{sonja.filiposka, anastas.mishev}@finki.ukim.mk
[2] Architecture and Performance of Computer and Communication Systems Group, University of the Balearic Islands, 07122Palma de Mallorca, Spain
cjuiz@uib.es

Abstract. Ever since the start of the green HPC initiative, a new design constriction has appeared on the horizon for the top supercomputer designers. Today's top HPCs must not only boast with their exascale performances, but must take into account reaching the new exaflops frontiers with as minimum power consumption as possible. The goals of this paper are to present the current status of the top supercomputers from both performance and power consumption points of view. Using the current and available historical information from the Top and Green HPC lists, we identify the most promising design options and how they perform when combined together. The presented results reveal the main challenges that should become the focus of future research.

Keywords: energy efficiency, green computing, HPC, power consumption, performances.

1. Introduction

HPC increasingly becomes a mainstream computing paradigm, not only for scientific, but also for many other applications. Getting the most of its performance, demonstrated as the number of floating-point operations per second (flops) is an imperative, both for the manufacturers and users. Two times a year, at the International Supercomputing Conference, the Top500 [1] lists are compiled, containing the most powerful HPC systems as measured with the High Performance Linpack benchmark (HPL) [2]. The first list was produced back in 1993 and since then, the world has seen a 5 orders of magnitude increase in the peak performance capabilities of the most powerful computing systems. But, adding more performance comes at a high cost: the need for more electricity to power these systems. Currently, the most power consuming system on the Top500 list requires almost 20MW of energy that equals the production of a small thermal power plant.

The increased interest in putting power constraints to these supercomputers led to the establishment of another competitive list, the Green500 [3]. The list uses the same benchmarking tools, but instead of looking only at the flops, the ranking value is the ratio of performance and power used, expressed as MFlops per Watt.

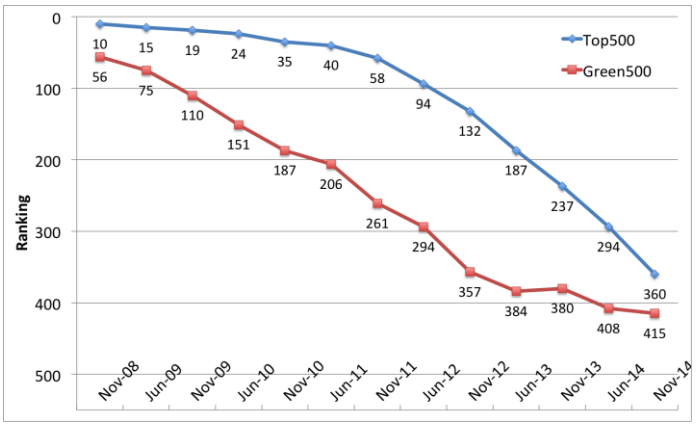


Fig. 1. Top500 and Green500 ranking of an example HPC system (Dawning 5000A, QC Opteron 1.9 Ghz, Infiniband, Windows HPC 2008 at the Shanghai Supercomputer Center, data labels are the exact ranking values in the corresponding lists)

By reviewing the historical changes in the lists and analyzing the transformations in different system characteristics we can track the history of technological development and design modifications that have led to today's petaflops top supercomputers [4]. The two lists unfold a different story about the strengths and weaknesses in the HPC systems development. Thus, after the first glance on the top stats, we find ourselves wondering on the way the top supercomputers standings change over time as more powerful newcomers enter the lists. As an example of this slow exit from the scene, the dynamics of both lists are presented in Fig. 1, tracking a single example system through a period of 6 years. From our observed example we can conclude that during the first few years the top high performing supercomputers are slowly falling down on the Top500 list, and as they are getting older their ranking drops are becoming more pronounced. Compared to the green performances of the same system we observe an almost steady drop along the Green500 list, which leads towards the conclusion that the new arrivals on the list are more frequently exhibiting higher green performances thus supporting the statement of increased awareness for a green HPC design of future systems.

By comparing both lists, past and present, many important milestones in the HPC development can be noted. Additionally, many of the technology improvements, performance or power wise, leave their significant mark on these list. Deeper understanding of these lists, their trends and correlations, can help guide research and industry towards the exascale supercomputers in the years to come. Detailed inspection and cross comparison of these systems can unearth lots of information about the power efficient architectures that should be utilized and developed in the future in order to further improve the green HPC initiative. The main goal of this paper is to infer the future of efficiency and power consumption of HPC systems by analyzing the technologies developed today showcased by the leadership-class computer systems. The objective is to provide an insight of the performance versus power trends for the current best architectures and to determine the direction in terms of processors, interconnections, system family and alike that shows a steady state improvement and paves the way for the future power aware efficient HPC systems.

2. Related work

Since its establishment in 1993, the Top500 list has become the arena where major players in the HPC industry, research and applications have measured their progress. The only metrics used was maximum performance (Rmax), measured in flops. This competitive development race led to the construction of systems that had huge power requirements, needed in order to sustain or increase their performance.

Recognizing the unsustainability in designing power hungry HPC systems, in the recent years an effort towards an energy efficient HPC design is on the rise. This green HPC initiative led to the proposal for a new list, the Green500 [5]. The main idea of this proposal was to establish additional, power-aware metrics to rank the most powerful HPC systems. Since 2007, parallel to the Top500 list, the Green500 [3] is published.

Over time both lists are becoming a rich source of historical data for the development of HPC systems, that can be used to better explain today's technology and predict future trends [6]. Thus, the Green500 list has been used as a basis for making first order projections for the power consumption of future supercomputers [7]. However, the presented conclusions are mostly based on the analysis of the interdependence of the power or performance vs. green efficiency, which means that the projections are based on functions that separately analyze the already intrinsic relationship between the green metrics and one of the two variables used to calculate it. Thus the results are mainly due to the existing natural correlation between the metrics, as we present further on in this paper. Similar approach can also be seen in [8], where the authors build a slightly different composite metrics, which is again based on the performance and efficiency, while in [9] experimental validations of the power measurement methodologies used in the Green500 list are presented which puts realistic weight on the analysis that can be made using the information from the list.

When considering the building blocks of most HPC systems, trying to analyze the energy efficiency of the system components one can also take into account another different benchmark. The SPECpower_ssj2008 [10], based on the standard SPEC benchmark, established itself as the standard server power-aware benchmark. It uses special methodology to measure the power consumed during benchmarking. The benchmark result is the power consumption at different load levels when processing a business transaction with a typical server side Java application related to the achieved performance scores. The results from all different load levels are used to compute an overall power-performance metric. An in-depth analysis of the SPECpower_ssj2008 list from 2010 is given in [11]. Unfortunately, the SPECpower results database is not regularly updated, which makes this data set incomplete, out-of-date and hard to work with. Also, it seems that some of the most significant vendors have recently stopped publishing their SPECpower results. Due to these reasons it seems that the top lists remain the only sources of information that can provide a global overview of the current status of the top supercomputers architecture performances and characteristics.

The overall efficiency of the HPC systems depends on many of its elements. There have been efforts to refine the flops/watt metrics in order to enable a deeper reflection of the contribution of different components of a HPC system to the overall efficiency. One such proposal is The Green Index introduced in [12]. The idea behind this approach is to have a single number value that will capture the system-wide energy efficiency.

Starting from 2013, the Green500 was extended with additional data such as: system family, interconnect, etc. Even though this data was already available in the Top500 list, the matching of both lists was difficult due to incomplete information and ambiguities between different systems, making a holistic performance/power analysis tedious and prone to errors. Based on these recent additions, the influence of various design options on the efficiency, along with some initial historical trends, most promising designs and main challenges that should direct future research efforts are given in [13]. In this paper we update and extend the work started in [13] throwing light on the most recent advancements as well as on the interdependencies of the different parameters that describe the top HPC systems.

3. Global Analysis of Current and Previous Top HPC Systems

Since high power consumption of the HPC system inevitably leads to increased construction costs as well as increased energy spent on cooling equipment, thus creating problems with the system reliability and availability, the green HPC initiative should be a major concern for researchers and vendors. Hence, today, power effective architectures and power management have become essential for HPC systems. The overall energy efficiency of supercomputers has modestly improved during the first years of the green HPC initiative (2007-2010), see Fig. 2. Improvements in this period can be observed over the full range of machines, and are due to the "plucking" of low hanging fruit in energy efficiency, e.g. using existing low-power microprocessors [5]. Following this initial stage, further improvements turned towards energy-centric architectural designs. After November 2010, the rise in the green metric (MFlops/W) performances coincides with the rise of the tide of the heterogeneous systems that introduce a large number of accelerator cores in order to boost the performances of the traditional multicore processors. As an example, the first three in the last Green500 list (Nov 2014) are hybrid systems wherein more than 80% of the total cores are accelerator cores, represented with AMD FirePro cores – Green1, ARM based PEZY-SC many-cores – Green2 and NVIDIA K20x cores – Green3. The initial boost of these power friendly architectures enters a small stagnation period during 2011-2012, just to make another big leap in 2013 that continues throughout 2014. Thus, we are currently facing a staggering improvement of almost 5300 MFlops/W compared to the first Green1 system in 2007 that exhibited modest ~360 MFlops/W, which is an astonishing improvement of around 1475%. This result has been the cornerstone of one of the goals of this paper: to analyze what are the critical decisions in HPC architectural design that have made this progress possible, as well as how to continue with this positive trend in the future.

However, analyzing the history and events that influence the green performance metric is only half of the story. The other half is to connect it with the trends that are evident in the unscrupulous Top500 list where the raw Rmax performances is the key metric since this is the list that defines the contents of the Green500. If we combine the Green500 improvements with their corresponding standings in the Top500 listing, as presented with the annotated data labels in Fig. 2, it is evident that the greenest HPCs are far from the best of the best performing HPC systems, entering the first Top100 only once in November 2011. This result points towards the need to close the gap between

the usage of low power accelerators and the organizational and structural design of the system that will squeeze out the maximum of their computing potential, thus placing the best green systems higher on the Top500 list.

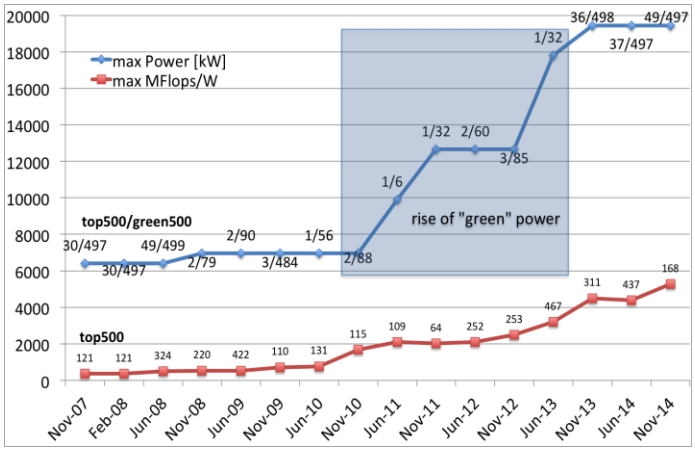


Fig. 2. Timeline of maximum power consumption and green performances presented as MFlops/W of Top500 and Green500 supercomputers (*annotated with top500/green500 index for maximum power and top500 index for maximum green performances which always represent the top 1 on the green500 list*)

Another concerning fact that arises from our timeline analysis is the rise of the maximum power consumption of the top HPC systems. It is interesting to note that the rise of the maximum power consumption follows right after the initial leap of green awareness in Nov. 2010. As it can be concluded from the annotated labels for maximum power consumption on Fig. 2 (representing the standings on the Top500 and Green500 lists), the large increase in power consumption from this period onward, although alarming, does not result with power hungry systems that are not green friendly. On the contrary, the top power HPC systems during the period of Nov. 2010 up to Jun. 2013 were obviously being constructed in order to reach the top 3 on the Top500 list, but are simultaneously placed well within the first 100 on the Green500 list thus being very good energy efficient examples.

Although efforts are being made in making the HPC systems more power efficient, it seems that this is actually not a case since the power consumption in average and maximum values is increasing even more steeply in the last years compared to the overall green performance. Also, there is an alarming trend from Nov, 2013 continuing throughout 2014, where the top power hungry HPC is one of the most non-green systems today. It seems that this machine (QUARTETTO) will continue to hold the "prominent" first position in maximum power consumption in the following time period since at the same time it belongs to the first Top50 systems, meaning that it will take a while until it is finally out of the Top500 list due to low performances. What is even more concerning is the fact that this particular system is also a hybrid based on a combination of NVIDIA K20x and Xeon Phi accelerator cores that represent more than 70% of the total number of cores. This evidently points out a very significant issue: the

path towards an energy-efficient high-performing HPC system is not as simple as building a hybrid system with a great number of accelerator cores, it must be designed in a way that this system will be well balanced.

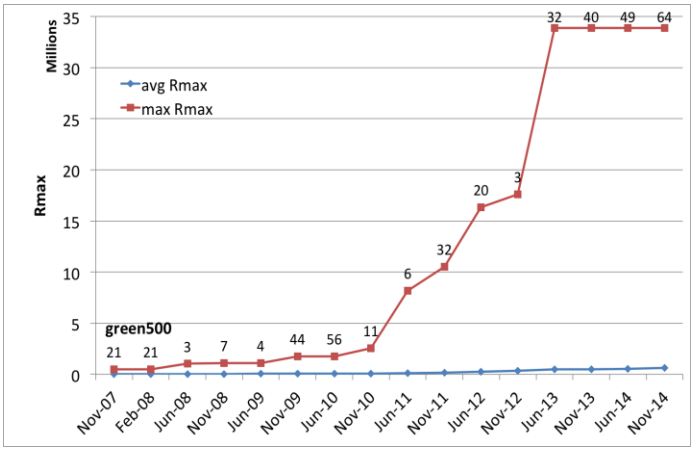


Fig. 3. Timeline of the maximum and average performances of top supercomputers represented by Rmax in GFlops (*annotated with the Green500 index of the corresponding Top1 supercomputer according to max Rmax*)

Delving deeper into the history of the top HPC systems development from the point of view of the Top500 list, the historical technological advancement can be represented using the maximum Rmax values as given in Fig. 3. For reflective comparison purposes the Rmax data points in this figure are additionally annotated with the ranking of the corresponding Top1 system on the Green500 list. It may be quite unexpected, yet very encouraging, that the Top1 HPCs belong to the top Green65 ever since the start of the green initiative. This means that all Top1 HPC system designers so far have been very good supporters of the energy-efficient trend setting the HPC community into a state where the rest should strive to follow the examples set by the leader. What is even more significant is the tremendous leap in Rmax values since the previously pinpointed November 2010. This major improvement in the HPC system design belongs to the custom build Sparc-based K-computer at RIKEN AICS and is followed with a corresponding, although smaller, leap in the needed power. This consequentially leads towards greener high performing systems that are well within the first 50 on the green list. The historically largest jump in Rmax occurred in the first half of 2013 resulting with an around twice more powerful system (today's famous TH-2) than the last one previously built. Of course, when comparing with Fig. 2, this move is accompanied with the steepest rise in maximum power consumption ever recorded which again belongs to TH-2. Still, the overall green performances of the new system are among the Green50. However, this brings up the question whether the rise from 17.6 to 33.8 petaflops justifies the extra 5 MW power needed, which are comparable to the peak power of todays' modern high-powered electric locomotives. The trends of the last two years show that while holding the primacy in the Top500, Tianhe-2 slowly falls down on the Green500 list leaving way to other more green friendly solutions.

The historical lists analysis shows that decreasing in the Top ranking goes hand in hand with the increased number of systems that belong to the same performances range. Cutting edge technology over time becomes mainstream implementation towards which the other HPC systems are converging to, so that after more than 5 years the same architectural design starts to leave the field of top HPC systems. It should be expected that todays top HPCs will be 'replicated' with exponentially increasing frequency in the next years, which means that in the next few years a large body of the Top500 list will exhibit comparable energy-efficiency as the one represented by the top of the top HPCs today.

One very pronounced problem with the power consumption in HPC systems today is the practically inefficient use of the energy consumption due to the reduced performance when compared to the theoretical peak. The measured performance obtained with HPL (Rmax) is drastically different across systems (ranging from 28% to 81% of the projected theoretical peak), while the real performances of the scientific applications drop to only 10% of Rpeak [14]. Another problem is that even when working with Rmax, the distribution of the performances of the Top500 HPC is not even close to normal, exhibiting a standard deviation that is 2 to 3 times larger than the average. As it is presented in Fig. 3, the gap between the max and average performance is extremely pronounced, thus making this two metrics unrealistic and possibly misguiding representations of the current status of the top HPC systems lists.

3.1. Performance vs. Power Distributions Analysis

Since the results have shown that the typical average, max-min analysis can sometimes be inconclusive due to the extreme deviation in the observed values, in order to uncover the true nature of the HPC systems we also analyzed the distributions of the performances and power consumption for the 500 lists.

In Fig. 4 the Rmax cumulative distribution function (CDF) for the performances of the top HPC systems is presented in the full historically available period for the Top500 ranking. For increased readability only every third year distribution is presented on a log-log axis. The first evident conclusion from the presented distributions [15] is that they all follow a power-law like distribution that ends with a very heavy tail (representing the Top5-10 systems of the time period). The power-law exponent of the corresponding fitted power-law functions falls into the typical [2, 3] interval that is the most commonly found among other power-law distributions of natural and man-made systems. This proves the previously discussed problem of analyzing the top lists using average and max-min values only, since these measurements do not capture the behavior of a large number of the HPCs systems. On the other hand, as previously discussed, the distributions confirm the exponential increase of the number of similarly performing systems over time.

Another noticeable result that is revealed by analyzing the Rmax distribution functions over time is the trend of relatively regular increase in performances. The distance between the distributions is comparably similar for each represented three-year interval. This trend depicts how the lowest performing HPCs are regularly being removed off the Top500 list over time. We would like to stress that the same results can be obtained when all available distributions are represented over the whole time period,

but the full extent of these results is not presented graphically in order to avoid overcrowding the figure.

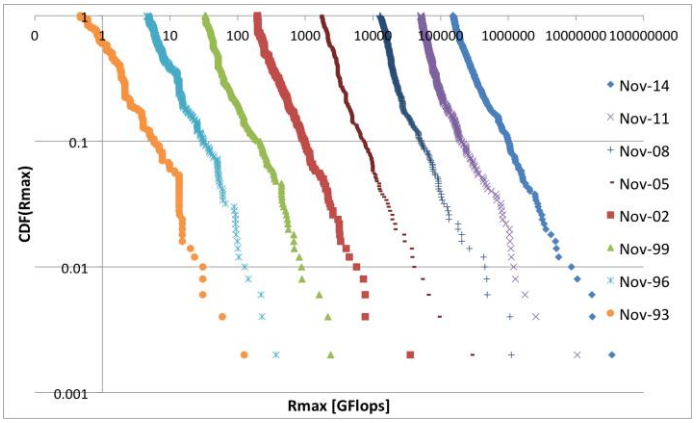


Fig. 4. Rmax cumulative distribution function snapshots taken in three-year intervals

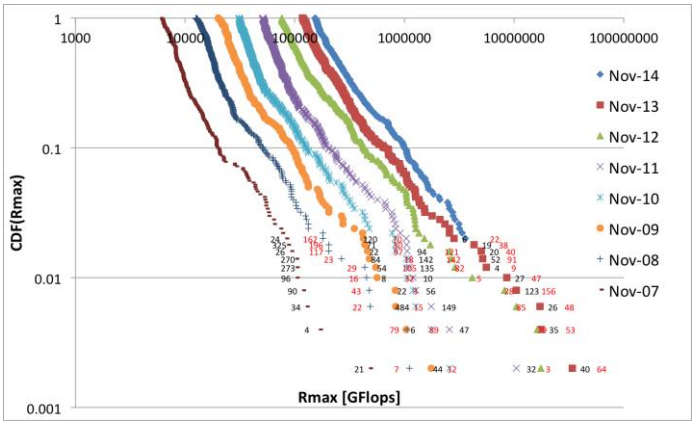


Fig. 5. Rmax cumulative distribution function snapshots taken from the green initiative start (2007-2014)

Additionally, we decided to more closely observe the conduct of the performance distributions in the latest 2007-2014 interval in which the green energy-efficiency initiative is active. The Rmax CDFs from this time period are presented in Fig. 5 where the Top10 systems are annotated with their rankings according to the Green500 list. As it was previously discussed, the power-law like heavy tail distributions retain their main characteristics while the overall performance grows constantly over the years. Since the time interval is smaller in this case, we can observe an overlap of the Top10 systems in the cumulative distributions, which is especially evident in the last 3 years. It is very

encouraging to note that in the last 5 years all Top10 HPC systems belong to the upper Green150 list making them part of the energy-efficient elite according to the MFlops/W metrics.

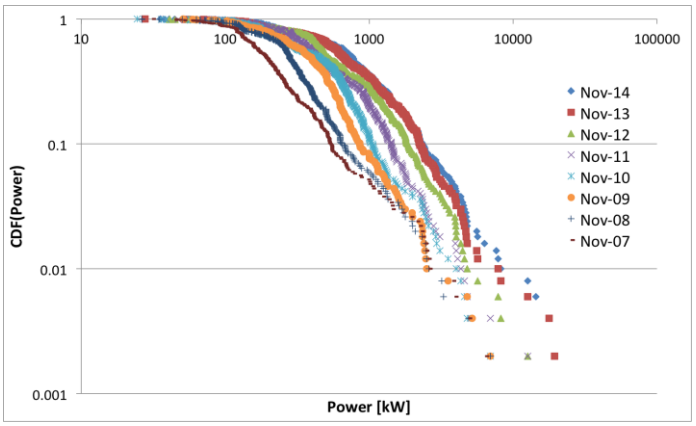


Fig. 6. Power consumption cumulative distribution function snapshots taken from the green initiative start (2007-2014)

When comparing these results with the CDFs for the HPC systems power consumption in the same time period, see Fig. 6, we observe a similar less prominent power-law like distribution that is much more clustered compared to the performances distributions. These distributions start with a pronounced normal distribution for the lowest performing HPC systems, and then turn into power-law scaling ending again with heavy tails. This means that if we examine the low power inhabitants of the Top500 list, their power consumption values will be normally distributed thus making any HPC system that falls in the last 150-200 systems group, a significant power representative of the low power set. However, as we move upward on the list, the difference in power consumption from one HPC system to another become significant thus disallowing for any trivial comparisons. The presented overlapping and clustering of the CDFs shows that the power consumption rise of the top lists main body follows a slower rhythm compared to the performances. This is clear evidence that the green initiative is actively influencing the decisions of the HPC designers that work hard towards big leaps in performances while trying (and succeeding) in lowering the rising rate of power consumption.

3.2. Rank Correlation

The existence of more then one metrics, hence more than one ranking of the HPC systems opens the issue of their correlation. Rank correlation statistics try to determine the correspondence between two measurements. There are many different types of correlation coefficients [16] that reflect somewhat different aspects of a monotone association and are interpreted differently in statistical analysis, such as the Pearson *r* correlation, Spearman's rank-order correlation and Kendall's *tau* correlation. Pearson's

correlation measures the degree of the relationship between linear related variables. Since, in our case, we are comparing rankings, the more suitable choice would be the Spearman's rank-order. The problem with this rank-order is that it does not handle ties well, while ties are very often found in our analyzed lists (there are always a number of very similar machines on the lists, that have the exact same values for Rmax, total cores, power, etc.). Thus, we chose the Kendall's *tau-b* correlation, due to the fact that it is especially suited when ties in the lists are frequent. For comparison purposes, in the bottom table in Fig. 7 the values of the two other correlation coefficients are also given, and it can be seen that they correspond to the Kendall's coefficient leading towards the same, but stronger, conclusions since in these cases the ties are implying additional correlation.

Another important relationship between these lists is that the Green500 list is a permutation of the Top500 list. The Top500 list is the primary list that uses the maximum performance as a single filtering and ordering parameter. The Green500 list then calculates the MFlops/W ratio for the same systems of the Top500 list and produces a permutation ranking according to this metric.

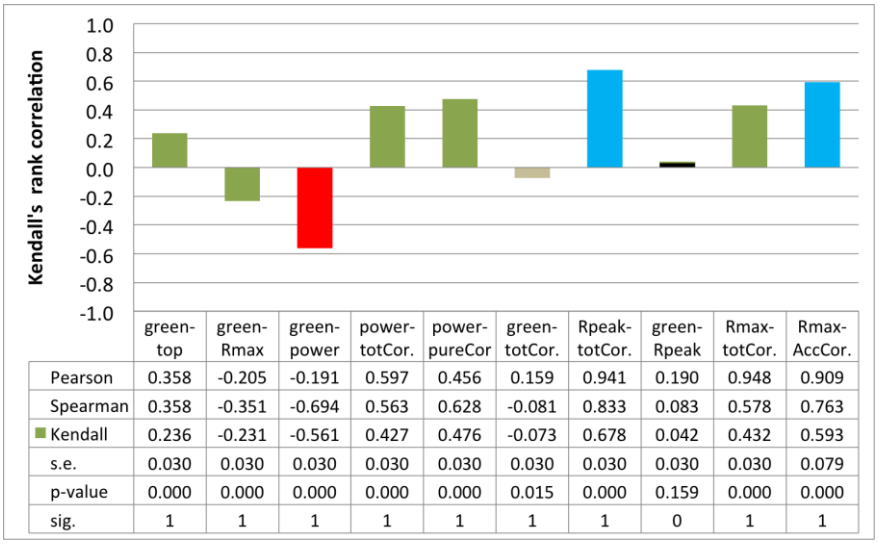
Using a 95% confidence interval, we investigated the correlation between different metrics from the lists and also performed two tails tests based on the *p*-value in order to prove the significance of the obtained correlation results. The value of the correlation coefficients reflects the strength of the correlation between the observed rankings according to the two chosen metrics. Thus, values closer to +1 represent strong positive correlation. In other words, higher positive values signify that the lists coincide, i.e. if an element is higher on the first list, it is expected to be so on the second as well. Negative coefficients (closer to -1) indicate strong negative correlation, such that if an element is ranked higher on the first list, it is expected to be lower on the second. Values around 0 indicate that the rankings on the two lists are not correlated. When testing the significance of the values for the correlation coefficient, values for *p*-value higher than 0.05 imply that the two observed metrics are statistically independent in which case the value of the correlation coefficient has no significance.

The obvious starting point of the correlation analysis was to identify the correlation of the two types of ranking of the HPC systems: Green500 and Top500. As it is presented in Fig. 7, both rankings show moderate correlation, with a *tau-b* of 0.236 meaning that in almost ¼ of the lists rankings the positioning in the two lists concur.

Since the metrics used in the ordering of the Green500 list is the ratio of maximum performance Rmax and the total power of the system, we made a detailed investigation of the correlation between the Green500 ranking and the Rmax of the HPC systems. The Kendall's rank correlation for these values indicates light negative correlation, justifying the Green500 motivation, which is that performance is not the only significant metric. It is more obvious when correlating the green rank with the system's power requirements. This is one of the strongest correlations, naturally negative (Kendal's *tau-b* = -0.561), confirming that the Green500 is not led by high-energy consumers.

Most of the power used by a HPC system, usually goes to the CPU itself, guiding us to the next correlation investigation between the power of the system and the total number of cores. The value of 0.427 illustrates a moderate correlation, which signifies that the rise of total cores will in many cases result with comparative rise in power ranks. As mentioned before, starting from 2010 some of the HPC systems started to use accelerator cores, as a massive, low power and high performance alternative to standard

CPU cores. Taking this into consideration, the correlation between the power and the number of pure CPU cores is even stronger, pointing to the conclusion that most of the power is consumed by the complex CPU cores rather than the accelerator cores. On the other hand, the positioning on the Green500 lists is not at all correlated to the total number of cores, as demonstrated by the value of only -0.073.



	green-top	green-Rmax	green-power	power-totCor.	power-pureCor	green-totCor.	Rpeak-totCor.	green-Rpeak	Rmax-totCor.	Rmax-AccCor.
Pearson	0.358	-0.205	-0.191	0.597	0.456	0.159	0.941	0.190	0.948	0.909
Spearman	0.358	-0.351	-0.694	0.563	0.628	-0.081	0.833	0.083	0.578	0.763
Kendall	0.236	-0.231	-0.561	0.427	0.476	-0.073	0.678	0.042	0.432	0.593
s.e.	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.079
p-value	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.159	0.000	0.000
sig.	1	1	1	1	1	1	1	0	1	1

Fig. 7. Rank correlation coefficients for different pairs of observed ranking metrics (*the Kendall's correlation coefficient is presented with the columns where red represents strong negative correlation, light blue strong positive correlation, green light to moderate correlation, gray represents absence of correlation and black failed significance tests*)

The strongest positive correlation in our analysis is the one between the theoretical performances of the system - Rpeak and the total number of cores. This is expected, since the addition of more cores raises the bar of the theoretical performance of an HPC system. The only correlation that had the *p*-value above the given significance level was the one between the Green500 ranking and the theoretical peak performance, which combined with the result of very weak correlation with the number of total cores, leads to the conclusion that the green ranking can not be related to the ranking according to the theoretical performances of the HPC systems.

If instead of the theoretical peak, we observe the actual performances using Rmax, as expected, the significant positive correlation between the performance metrics and the total number of cores shows that higher performance is usually achieved using more cores. Even stronger is the correlation between the performance and the number of accelerator cores, which explains the increasing penetration of HPC systems with high number of accelerator cores in the Top500 list, since these heterogeneous architectures prove to be a "straightforward" solution for obtaining high Rmax values that enables them high placement among the Top500.

4. Cross Comparison of Various HPC Design Options

The analyzed timelines combined with the results from the presented rank correlation both suggest the existence of notable trends in the HPC evolving architecture. This global performance versus power consumption overview can be used as a starting point for a thorough analysis of the influence of each part of the HPC architecture on the system green performances. This analysis can serve as a guideline for deciding on the future HPC architectures as well as point out topics for further improvement in the research field. Thus, in the following subsections we investigate the influence of each significant part of the HPC system design options from the performance and power consumption point of view.

4.1. Processor Family

One of the first decisions made when designing an HPC system is the processor family and generation. Today's top HPC systems are mostly based on the well-established Intel SandyBridge (46%) with the Intel Xeon 5 processor generation, followed by the newer Intel IvyBridge (27%) that promises significantly lower CPU power consumption due to the decreased feature size to 22 nm. They are tailed by the older Intel Nehalem (7%), and AMD x86_64 (6%) that is shoulder to shoulder with the newest newcomer since late 2013, Intel Haswell, that stays at 22 nm, but boasts new microarchitecture that is more power efficient. Almost all of the SandyBridge implementations come with 8 cores per processor socket, while IvyBridge increases to 10 and 12, and Haswell leads towards 12, 14 and 16. The AMD processor family, on the other hand, is based on Opterons (mostly 6100 or 6200 series) with 12 or 16, and the PowerPC family with PowerBQC processor utilizes exclusively 16 cores per socket.

When comparing their performance/power ratio, as given in Fig. 8, it can be seen that the PowerPC processor family offers the best green performances and is outperformed only by Intel SandyBridge and IvyBridge in the cases when a huge number of co-processor cores are used. These results confirm that the design of IBM's PowerPC is made with two goals in mind: low energy consumption and high performance. Another interesting observation is the leap from the older Power to the new PowerPC family is tremendous. One of the major changes towards this goal is the reduced processor frequency from 3.8 to 1.6 GHz. These simple, power-efficient processors originally developed for embedded systems are the basis of the top performing IBM BlueGene/Q. They also include several task specific acceleration engines. Although the PowerPC is the most powerful and green friendly processor architecture so far, its major drawback is the custom interconnection and design that creates a software and hardware "isolation" effect many are trying to get away from. Thus, the most general choice is the Intel based alternative. Fortunately, it seems that Intel's latest fourth generation, Haswell, is finally stepping up and coming close to the PowerPC performances which proves the published characteristics of this new architecture and is expected to lead towards major improvements as it spreads on other systems in the future lists. On the other hand, the different levels of performances that can be spotted at the most common SandyBridge

are mainly due to the different interconnection types used (namely Gigabit Ethernet, 10G Ethernet and InfiniBand), whose influence is discussed later on.

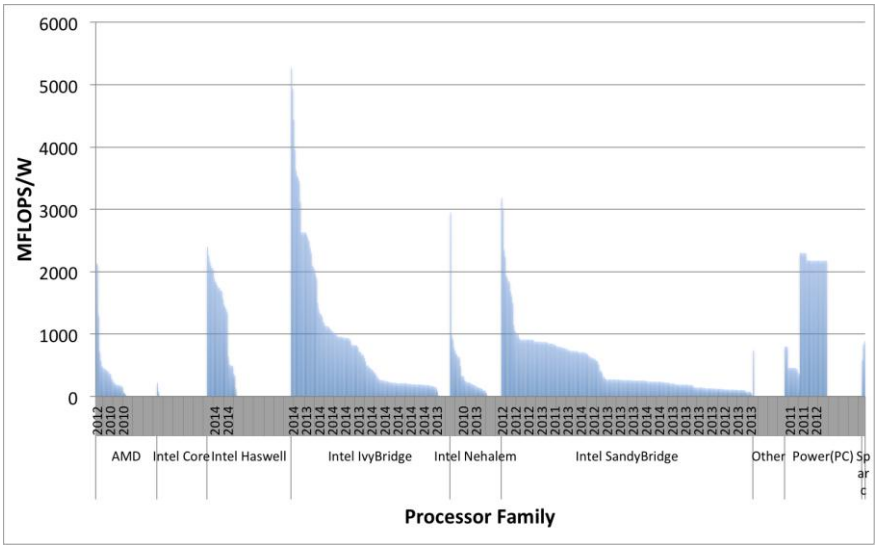


Fig. 8. Performance/power categorization of different HPC processor families

Another important observation can be made by following the performances of one particular Xeon processor generation with a different number of cores per socket, or different frequency. Comparative analysis has shown that as the number of cores per socket rises, the performance/power ratio for the system is expected to rise as well, while as the frequency rises, the performance/power ratio falls. These two trends will pave the way for future processor generations that are expected to decrease their frequency and increase the number of cores per socket in order to achieve better performances for lower energy consumption. This is already somewhat the case with the Intel IvyBridge with an increased number of cores per socket and lowered frequency, continued by the newest Haswell. Thus, almost half of the IvyBridge representatives have a performance/power ratio above 1000, and the Haswell's average is close to 2400.

As for the HPCs based on the AMD processor family, they are usually designed using the MPP architecture, unlike the typical Intel Cluster. Increasing the number of cores per socket increases their performance/power ratio as well. In order to reach a higher-performance/power ratio (above 1300) they need to be combined with co-processors. They are also usually seen in combination with a Cray interconnect which provides better performances than other types of interconnect for this family.

4.2. Interconnect Type

As already noted, another important characteristic of a HPC system is the type of interconnection used. Today's top performing systems are mainly based on the InfiniBand technology (45%), followed by Gigabit Ethernet (20%) and 10G Ethernet (18%). The rest of the interconnect types are mainly custom based for the specific systems like Cray interconnect (3%), or the custom one in the IBM BlueGenes (13%).

The performances of different interconnect type families as seen from the green perspective are presented in Fig. 9. The results show that IBM's Custom interconnect is the most performance/power efficient solution. However, this type of interconnect is used only in combination with a Power-family processor, which is its major compatibility drawback. Also, due to its torus-like topology, this interconnect requires more programming effort.

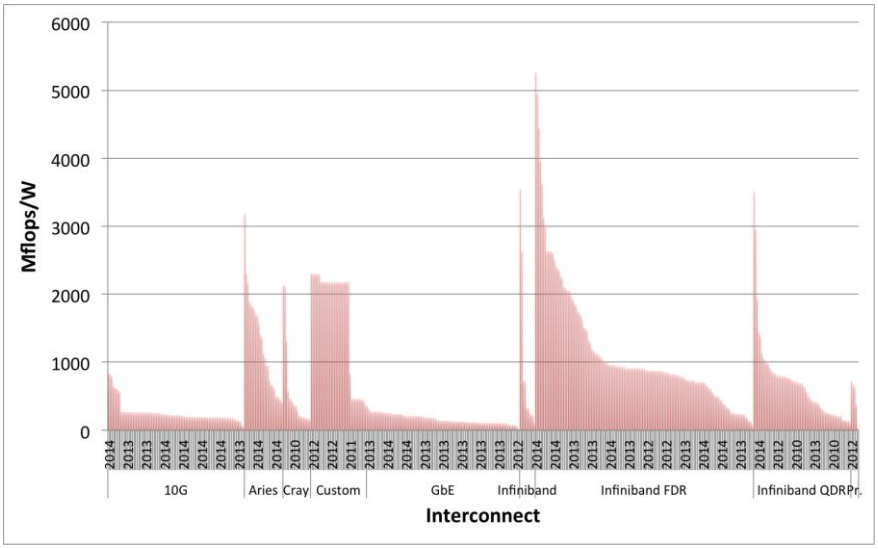


Fig. 9. Interconnect type influence on the supercomputer performances

The second best performing interconnect type is the InfiniBand with its different implementations (DDR, FDR, and QDR [17]). InfiniBand has gained wide acceptance in HPC systems mainly due to its high bandwidth and in particular due to its low latency and high flexibility. The InfiniBand technology is seen as the successor of the common Gigabit and 10G Ethernet [18] and, as it can be seen in the figure, highly outperforms its predecessors in terms of green performance. Unfortunately, it is always found in combination with the Intel Xeon processor family, while some of the Ethernet based solutions are designed using AMD Opteron. InfiniBand seems to be significantly better than 10G or Gigabit Ethernet from the energy efficiency perspective. For instance, in the case of systems based on the Xeon E5-2680 8C 2.700GHz processor, InfiniBand gives an average of 3.5 times better performance/power ratio compared to 10G, and is 2.7 times better compared to Gigabit Ethernet. However, this is not always the case. Among

the top HPC systems there are a number of examples that show that InfiniBand does not always combine well with NVidia co-processors.

For an example, when comparing two systems that differ only in the use of accelerator cores: the first one is designed with and the second without co-processors, it turns out that the performance/power ratio drops rapidly (around 3.5 times) when co-processors are introduced, mainly due to the lower performances of the systems that drop significantly below the peak. This example must raise a flag of careful inspection of the system since, despite expectations, adding co-processors into the system will not always boost the performances and green behavior. We must note however, that there is an example (namely, the CSIRO GPU Cluster), which is a successful example of mixing InfiniBand and NVidia co-processors. Thus, in the given figure, the best and worst performing InfiniBand based examples are the ones with co-processors.

The performance interconnection between InfiniBand and GPUs is just becoming a hot topic in the research community [19]. It is important to note that the Ethernet based interconnections never exhibited this problem when co-processors are introduced into the system. On the contrary, examples show that systems based on 10G Ethernet interconnection perform as well as InfiniBand based solutions in the cases when the 10G based system is built using a great number of cores. This is another important remark regarding InfiniBand, namely the InfiniBand based systems are usually built using a smaller number of cores with rare examples of systems with a great number of cores, which is mainly due to the complexities of its flat fabric. Also, with InfiniBand Remote DMA the cores are free from overseeing the network data read/write, which boosts the system performances without the need to add more cores [20].

4.3. Accelerators / Co-processors

Adding accelerators, or co-processors, is the current trend for achieving green HPC performances that started in November 2009 when the first heterogeneous HPC system that included PowerxCell co-processor, initially built in 2008, was transformed into the Top500 Roadrunner. The year 2009 has seen the first heterogeneous systems built using NVIDIA and ATI HD accelerators. In the next year the trend has gotten momentum with more than 10 systems on the list, while a fast rise of this architecture has been foreseen [21] and confirmed so far.

The architectures and programming models of co-processors may differ from CPUs and vary among different co-processor types. This heterogeneity leads to challenging problems in implementing and porting application operations when striving to obtain best performances. In Nov. 2014, the heterogeneous HPC systems constitute 15% of the top 500, the worst one of which has maximum performance that is only 14% of its theoretical peak. The current average efficiency in terms of sustained versus peak performance of the top supercomputers that are constructed using NVIDA GPUs is around 0.60, which is still behind the average of 0.7 of the ones without any accelerators. However, the top supercomputers that utilize the newest NVIDIA K40xx accelerators seem to be catching up with an average efficiency of 0.67. Interestingly, the unique specific case of the PEZY-SC exascaler system with almost 99.8% of its cores being ARM based co-processors [22] manages to reach the efficiency of only 47% of its theoretical capabilities yet proudly standing on the 2nd place in the Green500 list. The

Green1 place is reserved for another custom build system by ASUS with 90% of its cores being AMD FirePro accelerator cores that enable effective usage of only 50% of the theoretical peak.

One must always bear in mind that these differences are a lot more pronounced when considering real workloads. Still, we must not forget that out of the total of 75 such systems, 18 are placed on the currently highest top positions according to the Green500 list followed by the design-wise opposite IBM's BlueGene/Q architecture.

It is of practical interest to analyze how the amount of share of co-processors in the number of total cores impacts the performance/power ratio of the system, as well as compare the performances of the main competitive accelerator technologies. This analysis is presented in Fig. 10. It is clear that the accelerator share in almost all of the systems is well above 50% (average 66%), with a typical 60% for NVidia and 88% for Xeon Phi. As it is presented, with careful systems design the expected achieved green performance can be above 1000 Mflops/W, with potential to reach staggering 5300 Mflops/W. Note that the PEZY-SC system is not presented in this figure due to its uniqueness and thus inability for cross comparison.

There seem to be two actual choices for a co-processor in the today's systems: the Intel many integrated core (MIC) Xeon Phi and NVidia's most popular K20xx and newer K40xx options that show improved performances. It is evident that NVidia's new Kepler architecture improves the GPU's performance significantly, mainly due to the new streaming multiprocessor SMX [23]. However, experimental cross comparison shows that different types of co-processors are more appropriate for specific data access patterns and types of parallelism. The MIC's performance compares well with that of the GPU when regular operations and computation patterns are used [24]. The GPU is more efficient for those operations that perform irregular data access and heavily use atomic operations. The programming tools and languages employed for code development for a MIC are the same as those used for CPUs. This is a significant advantage as compared to GPUs. For the MIC, auto vectorization is performed by the compiler, which however needs additional guidance when complex pointer manipulations are used [24].

With the GPUs carrying out a substantial portion of the calculations, host memory, PCIe bus, and network interconnect performance characteristics need to be matched with the GPU performance in order to maintain a well-balanced system [25]. InfiniBand QDR interconnect is highly desirable to match the GPU-to-host bandwidth. Host memory needs to at least match the amount of memory on the GPUs in order to enable their full utilization [26]. Yet, many challenges remain open so that accelerated HPC systems are truly energy efficient with practical performances a lot closer to their theoretical peak compared to today, especially since accelerator technologies have proven to be difficult to program and unsuitable for some workloads [27]. Even more important is the inability of many applications to efficiently map to accelerator architectures. At the high end, suitability for a wide range of applications is a must. These issues call into question accelerator viability in the largest exascale machines.

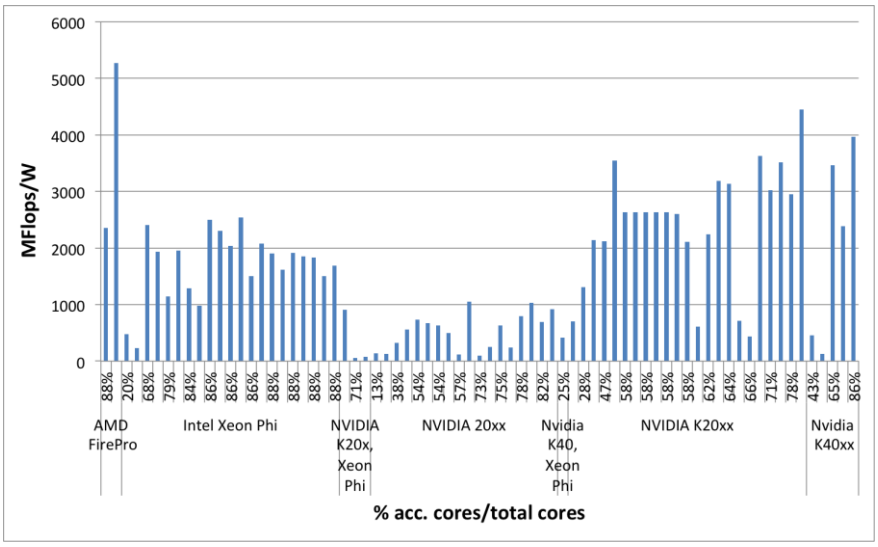


Fig. 10. Green performances of different heterogeneous HPC systems

4.4. System family impact

In Fig. 11 the power consumption related to the number of total cores for the most prominent system families of today's HPC supercomputers is presented. The results indicate that there are three current trends that depict the green status and scaling of the different system families. The most prominent example are the HP Cluster Platform system families that are all consistently following a linear increase in the total power with the rising number of total cores. This example is also the least performing one since the toll of more power needed for increasing the number of cores (and thus performances) is the highest of all compared. However, there are two members of this group that show low power consumption in combination with a large number of cores (encircled on the figure). This "out of normal" behavior is due to the fact that these systems are supported by a great number of co-processors, which on the other hand require a lot less power per core compared to a "pure" core in the system.

Another obvious trend that strongly relates to the system family are the IBM BlueGenes. The figure clearly shows that IBM BlueGenes scale extremely well with only slight increase of power demand for a great increase of total number of cores, which further accentuates the excellent properties of this system family since it never relies on increasing its performances by adding accelerators or co-processors. Furthermore, the Mflops/W ratio for this system family is consistently rising over the years with around 370 for the systems using PowerPC 4C processors in 2008, 450 when using Power7 8C in 2011, to a staggering 2300 when using the PowerBQC in 2012-13. It is also of great importance that these systems are scaling with the same Mflops/W ratio when keeping all of the parameters the same and simply increasing only the number of total cores. However, there is only one BlueGene newcomer in 2014, and the

BlueGenes have fallen out of the Green20 list in 2014, which gives the impression that these systems are losing the battle in favor of the heterogeneous systems.

The rest of the system families seem to fall somewhere between the worst and best extremes. Here we find other IBMs, the Bullxs, as well as the Crays and the SGI ICEs.

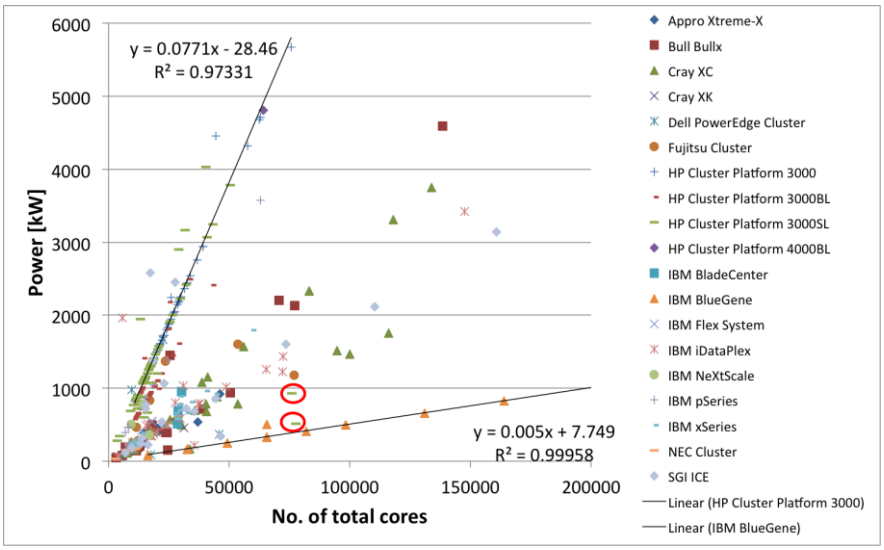


Fig. 11. Power consumption related to number of total cores for different system families

To establish the level of impact the difference in system family has over the rest of the system parameters (like processor, interconnection, co-processors) we made a comparison of three different supercomputers that differ in the system family only. This effectively means that the design difference of these systems is in the enclosure, which defines the physical placement of the cores, as well as fans and cooling among other parameters. Our analysis shows that direct liquid-cooling system [28] of the electronic components more than doubles the Mflops/W compared to other similar configuration. The method of implementation of the internal air-cooling system also strongly influences the efficiency. Systems with shared chassis fans show less efficiency than ones with tightly coupled fans. Thus, the enclosure type has great impact on the overall system performances and has to be chosen very carefully in order to minimize the power consumption while providing maximum system performances. The results also show that thermal aware schedulers are very important for achieving the green goal. Thus, major future efforts should be focused on this challenge.

5. Conclusion

The analysis of green efficiency of HPC systems presented in this paper has pointed to several important conclusions toward the critical decisions in HPC architectural design that pave the way for achieving best performances for minimum power consumption. In

order to ensure green performance of the future HPC mainstream technologies, designers should focus on building heterogeneous systems that will close the existing gap between the theoretical and achieved performance. The feasibility of this approach is confirmed with our distribution analysis showing that ever since the start of the green initiative the rate of growth in performances is higher compared to the increase in power consumption.

Because the power consumption strongly influences the green ranking, while it itself is mainly due to the number of "pure" cores, the computing power of future green HPC systems should be achieved using mainly accelerator cores (above 90%) that will positively induce high Rmax only. However, in order to ensure that the accelerator cores will provide the pursued performances, the rest of the system components must be carefully chosen. The cross comparison of the various HPC design options directs towards Haswell low power high performing processor family combined with InfiniBand interconnection in a carefully designed system family chassis that employs liquid based cooling of the individual elements together with thermal balanced schedulers. However, there are still other open issues that must be resolved with the main problem being simplified application development and efficient porting to a heterogeneous environment.

6. References

1. [Online] Available: http://top500.org (current November 2014)
2. [Online] Available: http://www.netlib.org/benchmark/hpl (current November 2014)
3. Sharma, S., Hsu, C.-H., Feng, W.: Making a Case for a Green500 List. 20th IPDPS (2006)
4. Kindratenko, V., Trancoso, P.: Trends in High-Performance Computing. Novel Architectures. IEEE Computing in Science and Engineering. (2011)
5. Feng, W., Cameron, K. W.: The Green500 List: Encouraging Sustainable Supercomputing. IEEE Computer, 50-55. (2007)
6. Feng, W., Lin, H.: The Green500 List: Year Two. www.green500.org. (2009)
7. Subramaniam, B., Saunders, W., Scogland, T., Feng, W.: Trends in energy-efficient computing: A perspective from the Green500. IGCC. (2013)
8. Hsu, C-H., Kuehn, J. and Poole, S. Towards efficient supercomputing: searching for the right efficiency metric. Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering (ICPE '12). ACM, New York, NY, USA, 157-162. (2012)
9. Subramaniam, B.; Feng, W.: Understanding Power Measurement Implications in the Green500 List. GreenCom, IEEE/ACM ICPSCom, 245-251. (2010)
10. Lange, K-D., Tricker, M., Arnold, J., Block, H., and Sharma, S.: SPECpower_ssj2008: driving server energy efficiency. In Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering (ICPE '12). ACM, NY, USA, 253-254. (2012)
11. Chung-Hsing Hsu; Poole, S.W.: Power signature analysis of the SPECpower_ssj2008 benchmark. 2011 IEEE ISPASS , 227-236. (2011)
12. Subramaniam, B., Feng, W.: The Green Index: A Metric for Evaluating System-Wide Energy Efficiency in HPC Systems. 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum, 1007–1013. (2012)
13. Filiposka, S., Mishev, A., Juiz, C.: Opportunities and Challenges for Green HPC. ICT Innovations 2014, Advances in Intelligent Systems and Computing, 311, 45-54. (2015)
14. Liu, Y., Zhu, H.: A survey of the research on power management techniques for high-performance systems. Soft. Pract. Exper. 40:943-964. (2010)

15. Clauset, A., Shalizi, C. R., Newman, M. E.: Power-law distributions in empirical data. SIAM review, 51(4), 661-703. (2009)
16. Hauke, J., Kossowski, T.: Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones geographicae, 30(2), 87-93. (2011)
17. Vienne, J., Chen, J., Wasi-ur-Rahman, Md., Islam, N. S., Subramoni, H., Panda, D. K.: Performance Analysis and Evaluation of InfiniBand FDR and 40GigE RoCE on HPC and Cloud Computing Systems. IEEE 20th Annual Symposium on High-Performance Interconnects, 48–55. (2012)
18. Bortolotti, D., Carbone, A., Galli, D., Lax, I., Marconi, U., Peco, G., Perazzini, S., Vagnoni, V.M., Zangoli, M.: Comparison of UDP Transmission Performance Between IP-Over-InfiniBand and 10-Gigabit Ethernet. IEEE Tran. on Nuc. Sci. 58:4:1606-1612. (2011)
19. Reano, C., Mayo, R., Quintana-Orti, E.S., Silla, F., Duato, J., Pena, A.J.: Influence of InfiniBand FDR on the performance of remote GPU virtualization. IEEE International Conference on Cluster Computing (CLUSTER), 1-8. (2013)
20. Reano, C., Mayo, R., Quintana-Orti, E., Silla, F., Duato, J., Pena, A.: Influence of InfiniBand FDR on the Performance of Remote GPU Virtualization. IEEE Cluster 2013. (2013)
21. Krieder, S.J., Raicu, I.: An Overview of Current and Future Computing Accelerator Architectures, 1st Greater Chicago Area System Research Workshop Poster Session. (2012)
22. Rajovic, N., Vilanova, L., Villavieja, C., Puzovica, N., Ramireza, A.: The low power architecture approach towards exascale computing. Journal of Comp. Sci., 4, 439–443 (2013)
23. Jeong, H., at al: Performance of Kepler GTX Titan GPUs and Xeon Phi System. Journal of Computational Physics, PoS (LATTICE 2013) 423. (2013)
24. Teodoro, G., Kurc, T., Kong, J., Cooper, L., Saltz, J.: Comparative Performance Analysis of Intel Xeon Phi, GPU, and CPU. Distributed, parallel and cluster computing, Cornell. (2013)
25. Kindratenko, V. V., Enos, J. J., Shi, G., Showerman, M. T., Arnold, G. W., Stone, J. E., Phillips, J. C., Hwu,W.: GPU Clusters for High-Performance Computing. IEEE International Conference on Cluster Computing and Workshops. (2009)
26. Elnashar, A., Aljahadli, S.: Experimental and Theoretical Speedup Prediction of MPI-Based Applications. Computer Science and Inf. Systems, Vol. 10, No. 3, 1247-1267. (2013)
27. Hermmert, S.: Green HPC From Nice to Necessity. IEEE Comp. in Sci. and Eng. (2010)
28. Loken, C., et al.: SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. J. Phys.: Conf. Ser. 256 012026. (2010)

Sonja Filiposka received her PhD in technical sciences in 2009 and currently is an associated professor at the Faculty of Computer Science and Engineering. Her main research interests include complex systems, optimization, network architectures and technologies and cloud computing. She is author of over 100 scientific papers published in international journals and conference proceedings and has been a member of many national and international research, as well industry oriented projects. She is also actively involved in women in engineering and girls in ICT projects.

Anastas Mishev, associated professor at the Faculty of Computer Science and Engineering, UKIM, Skopje. Holds a PhD in computer science since 2009. In the focus of his research are infrastructures for collaborative computing and research, primarily Grid and High Performance Computing systems. His aim is to get these systems closer to all potential users, mainly the research communities in order to fully use their potential. He researched in the areas of computer architectures and networks, cloud computing, HPC and software engineering. He participated in the implementation of over 30 international projects funded by TEMPUS, PHARE, DAAD and FP programs,

targeting the development of IT infrastructure and IT education, with the most important being FP: HP-SEE as national coordinator, SEERA-EI, EGI-InSPIRE, GN3, GN3+, GN4. He is author of over 50 scientific papers published in international journals and proceedings of conferences and held over 30 presentations at scientific conferences and workshops, mostly in the fields of grid computing, HPC, computational chemistry, software engineering, etc.

Carlos Juiz is an associate professor at the University of the Balearic Islands, Palma de Mallorca, Spain, and leads the ACSIC research group (acsic.uib.es). His research interests are mainly Performance Engineering, IT Governance and Green Computing. He is senior member of ACM, senior member of IEEE and member of the Domain Committee on Cloud Computing at IFIP.

Received: February 28, 2015; Accepted: November 8, 2015.