Energy and Power Aware Job Scheduling and Resource Management: Global Survey — Initial Analysis

Conference Paper · May 2018

DOI: 10.1109/IPDPSW.2018.00111

9 authors, including:

Matthias Maiterth
Ludwig-Maximilians-University of Munich
7 PUBLICATIONS **217** CITATIONS
SEE PROFILE

Siddhartha Jana
Intel
17 PUBLICATIONS **139** CITATIONS
SEE PROFILE

Andrea Borghesi
University of Bologna
44 PUBLICATIONS **397** CITATIONS
SEE PROFILE

Andrea Bartolini
University of Bologna
147 PUBLICATIONS **1,720** CITATIONS
SEE PROFILE

2018 IEEE International Parallel and Distributed Processing Symposium Workshops

Energy and Power Aware Job Scheduling and Resource Management: Global Survey — Initial Analysis

Matthias Maiterth*†, Gregory A. Koenig§, Kevin Pedretti‡ Siddhartha Jana*, Natalie Bates§, Andrea Borghesi¶, Dave Montoya‖, Andrea Bartolini¶, Milos Puzovic**

Energy Efficient HPC Working Group
Energy and Power Aware Job Scheduling and Resource Management Team

*Intel Corp., †LMU Munich, ‡SNL, §EE HPC WG, ¶University of Bologna, ‖LANL, **The Hartree Centre

***Abstract*—This work describes the motivation and methodology of a first-of-its-kind global survey of HPC centers actively employing Energy and Power Aware Scheduling and Resource Management solutions for their production systems. The Energy-Efficient High-Performance-Computing Working-Group (EE HPC WG) Energy and Power Aware Job Scheduling and Resource Management (EPA JSRM) team conducted comprehensive interviews over the course of 2016 and 2017. In this work, we present the selection of participating sites, the motivation behind the survey, a detailed description of the questionnaire, and illustrate why getting a global view of the ongoing efforts is a major step towards more efficient systems. Job Scheduling and Resource Management is being tackled using new approaches regarding Power and Energy and has important implications for achievable center strategies. With this survey, we are laying foundations necessary to give insights in how problems and respective solutions are approached across sites and centers to allow to identify differences, similarities, solutions, and possible technology transfer across sites and centers. Upcoming work will focus on the survey responses and the analysis thereof. At the point of writing, the EPA JSRM team is in the major analysis phase of the centers' responses. By splitting the work in this fashion we achieve increased clarity in presentation and have the opportunity to generate more detailed analysis in benevolence of the community and reader.**

***Index Terms*—power, energy, performance, power-aware, computing, scheduling**

I. Introduction

The power and energy demands of high-performance computing (HPC) centers are growing due to an increasing number of systems on site, an increase in individual system size and an increase in both the rate of change and magnitude of system power fluctuations. With these factors in mind, the Energy Efficient High-Performance Computing Working Group (EE HPC WG) [15] assembled a team to investigate how HPC centers are using energy and power aware job scheduling and resource management capabilities. The EE HPC WG Energy and Power Aware Job Scheduling and Resource Management (EPA JSRM) team identified and surveyed major HPC sites that have either actively deployed or are engaged in technology development with the intention to deploy large-scale EPA JSRM technologies in a production environment.

With a geographically diverse high-performance-computing landscape, the group tried to identify similarities and differences among the centers. The survey spans from motivation to implementation, looking at system components impacted, user interaction and level of automation of the solution. The questions try to be comprehensive in nature and span a wide range of information. The intent is to give a clear picture of what exists and to allow for improved decision making for other centers when trying to advance or introduce EPA JSRM solutions.

Each of the surveyed sites is unique in terms of funding structure, geographical as well as geopolitical situation. The combination of these factors has a strong impact on the possibilities and approaches the sites are willing to take and can take. Without such a survey, the agendas and conclusions that can be drawn stay hidden.

The survey conducted does not advocate a definite way for how HPC centers should approach energy management from a job scheduling and resource management standpoint, but rather gives a comprehensive overview on information provided by centers actively pursuing this approach in production sites backed by technology, research efforts and collaborations.

With a broad interest in EPA solutions, a lot of research is conducted and published in this area whereas, in contrast, publications are rare about technology that makes it into real production systems. Since the technologies deployed in production systems are mature, relevant research questions should already be answered by then. There may, however, still be opportunities for research and the results of this survey may be of interest to the research and academic community. Understanding what is used and tried in production, and why, is a valuable point of information, to drive future directions. Active vendor engagement is apparent in EPA solutions as well, since energy efficiency is a key concern. Most HPC centers are concerned with operational costs and energy efficiency and some even have a broader focus on sustainability.

978-1-5386-5555-9/18/$31.00 ©2018 IEEE
DOI 10.1109/IPDPSW.2018.00111

A lot of vendors try to incorporate EPA solutions into at least part of their current portfolio, all emphasizing different aspects. With funding agencies trying to keep low operational expenses, the introduction of limits to power and energy consumption is sometimes even labeled with a strict number. This can be seen by initial exascale system forecasts, as well as current and upcoming procurement announcements. To a lesser degree, this already started to become apparent in the petascale procurements (e.g. [31]). All these factors show a heightened interest in EPA solutions in general, and with EPA JSRM in particular.

To provide a context for the motivation of the EPA JSRM Team, a similar analysis done by the EE HPC WG in different subfields should be highlighted. To better understand the interaction of electrical grid integration and supercomputing centers, Bates et al. [6] carried out an analysis of anticipated usage patterns of supercomputing centers and how these can be safely integrated into power grid management for better cost and risk management. The paper highlighted possible partnerships and interactions of Electricity Service Providers (ESPs) and Supercomputing Centers (SCs). An extension to the European area was conducted in a subsequent study by Patki et al. [36], focused on the similarities and differences of Electricity Service Provider-Supercomputing Center (ESP-SC) relationship based on the geographic locations in Europe and the United States. The focus of the cited work is again the interaction of demand management and SC coordination, showing potentials, but also necessary regulatory and technological steps.

The focus of this work is on steps and technological advances and approaches taken by the supercomputing centers for job scheduling and resource management systems to monitor, control and steer power and energy consumption at the SCs.

The remainder of this paper is structured as follows: Section II introduces a detailed insight on the motivation of the EE HPC WG EPA JSRM team on the topic and why the survey was conducted. Section III gives an insight on selection criteria, the sites selected, and the selection process, to clarify representation. Section IV is the second main focus of this paper highlighting the questions posed and explaining the reasoning behind these. Section V links the planned next steps of the EE HPC WG for analysis and gives a short synopsis of the centers activities. Section VI gives a short overview of related work to make clear how to position this work in the literature. Finally, Section VII concludes the paper.

II. Motivation for the Survey

The investigations of SCs and their ESPs in both the United States and Europe suggested that fine and course grained power management as well as job scheduling were approaches that sites might use to manage their power in response to requests from their electricity service providers [6]. In response to this insight and a general recognition of a strong interest in predicting and managing power usage for HPC data centers for many reasons, the EE HPC WG has formed an EPA JSRM team. It is recognized that lessons learned and best practices are being gained in sites that have deployed energy aware resource management and job scheduling software that relies on power monitoring hardware. This team is chartered to capture those lessons and best practices.

A. Job Schedulers and Resource Manager

Throughout this paper we refer to two types of system software that we specifically define here. Job schedulers allow high-performance computing users to efficiently share the computing resources that comprise an HPC system. Users submit batch jobs into one or more batch queues that are defined within the job scheduler. The job scheduler examines the overall set of pending work waiting to run on the computer and makes decisions about which jobs to place next onto the computational nodes within the computer. Generally speaking, the job scheduler attempts to optimize some characteristic such as overall system utilization or fast access to resources for some subset of batch jobs within the computing center's overall workload. The various queues that are defined within the job scheduler may be designated as having higher or lower priorities and may be restricted to some subset of the center's users, thus allowing the job scheduler to understand distinctions of importance of certain jobs within the overall workflow.

To carry out its work, a job scheduler typically interacts with one or more resource managers. A resource manager is a piece of system software that has privileged ability to control various resources within a datacenter. These resources can include things such as the physical nodes that make up a high-performance computer's computational resources; disks, disk channels, or burst buffer hardware that comprise I/O resources; or network interfaces, network channels, or switches that comprise interconnect resources. For example, a job scheduler might use resource management software to configure the processing cores, memory, disk, and networking resources within one or more computational nodes in accordance with the requested resources for a specific batch job prior to launching that job onto the allocated computational nodes.

Finally, in some cases, resource management software might have the ability to actuate pieces of the physical plant that are responsible for delivering electricity to the datacenter or cooling the datacenter.

This paper considers the synthesized use of job schedulers and resource managers to provide energy and power aware job scheduling and resource management capabilities within a high-performance computing datacenter. Figure 1 presents an overview of the different components that may participate in such a solution. As shown, depending on the complexity of the implementation, the tasks of an EPA JSRM solution can be divided into four functional categories - the monitoring and control of energy/power consumed by the resources, and their availability. Energy/Power 'monitoring' techniques complement traditional resource management of processors, memory, nodes, disks, and networks. The 'control' of energy/power is heavily dependent on telemetry sensors that are responsible

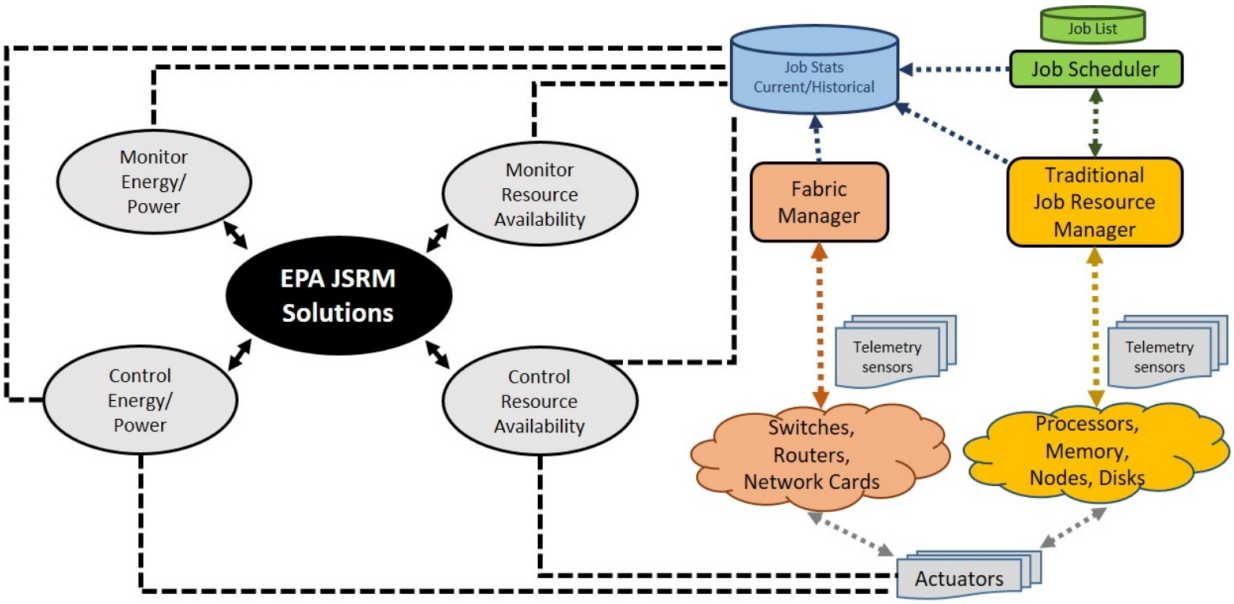

Fig. 1. Interactions among multiple components that make up a typical EPA JSRM (Energy and Power Aware Job Scheduling and Resource Management) solution.

for constantly monitoring the activity of the system resources. Examples of such control techniques could range from simple human-controlled actuation of processor dynamic voltage and frequency settings to reduce power to much more complex scenarios where the job scheduler has detailed historical knowledge of job characteristics and schedules multiple jobs simultaneously in a way that optimizes for certain energy- or power-specific objectives. Because system-wide software agents like the job scheduler have access to details of a supercomputing center's entire workload, and can potentially apply advanced data analytics to the problem, they have the potential for improving the energy and power consumption of supercomputers in ways that are unlikely to be possible for human-controlled processes. Accordingly, we expect that a trend in coming years will be to have system-wide techniques play an increasing role in these endeavors. Due to the previous absence of a general overview of EPA solutions employed, actual usage and benefits were only known within individual centers.

B. New dedicated EPA JSRM team.

In mid-2016, the EE HPC WG formed a team focused on energy and power management through the use of job scheduling, resource management, and associated tools. The Energy and Power Aware Job Scheduling and Resource Management (EPA JSRM) team is comprised of approximately 70 members from supercomputing centers, various academic and laboratory research centers, and the vendor community, particularly focusing on job scheduling and resource management software vendors and system integrators. Most of the members are from North America and Europe, however there are members from Asia as well.

During its work, the team identified a number of supercomputing centers that have developed, or are currently developing, technologies that use EPA JSRM techniques on one or more large-scale systems. Overall, eleven sites were identified and nine sites agreed to participate in a survey that asked questions about each site's supercomputer installation, typical utilization metrics and the types of jobs the site typically runs, and details of the use of EPA JSRM techniques employed by the site. After examining responses to the survey questionnaire from each site, a three-person subteam interviewed personnel from the site to clarify details in their written survey responses or to ask for further technical details of responses that seemed especially noteworthy. The goal is to present a high-level evaluation of these survey responses, including unique characteristics of individual sites as well as common characteristics across sites. Based on this evaluation, the objective is to present recommendations for system software researchers and scheduler vendors who are working in this area in an effort to help guide these endeavors.

III. Center Selection

In this section we describe the requirements that were used to identify which centers to target in our survey. A three-part test was utilized:

1) The center should be representative of a high performance computing center and have at least one system that is

in the Top500 list of fastest compute systems in the world [42].
2) The center should have either actively deployed or are engaged in technology development with the intention to deploy large-scale EPA JSRM technologies in a production environment.
3) The center's leadership was willing to participate and openly talk about their efforts in the area.

With this criteria, the diverse membership of the EE HPC WG and the broader HPC community was consulted to identify canidate centers to target. Item 2 of the selection criteria was one of the main elimination factors for the survey. Many of the sites identified were conducting exploratory research activities, but had not yet deployed anything in production and had no expectation to do so in the near future. Ultimately, a list of eleven centers was identified. To the best of our knowledge, this list comprised the entire set of supercomputing centers meeting our criteria, however the EE HPC WG team is always open to including additional sites based on feedback. The team contacted each of the eleven centers, of which nine elected to participate in the survey.

The centers interviewed were:

1) RIKEN, Japan
2) Tokyo Institute of Technology, Japan
3) CEA, France
4) KAUST, Saudi Arabia
5) LRZ, Germany
6) STFC, United Kingdom
7) Trinity (LANL+Sandia), United States
8) CINECA, Italy
9) JCAHPC, Japan

These span the geographic regions of Asia, Europe and the United States. The centers span academic institutions and national research laboratories with different foci. The differences also become apparent in funding strategy and sources, energy billing, energy service provider environment, but also geographic and thus thermal environment. The geographic location of the centers can be seen in Figure 2.

After identification of the participants, the interviews of the sites themselves spanned eleven months from initial request to final review of the responses (September 2016 to August 2017).

IV. Description of questionnaire

This section describes the rationale behind the choice of the questions designed for the survey. In addition to getting a high-level overview of each center's EPA JSRM solution, the goal was also to obtain a deeper understanding of the technical, engineering, management, and logistic decisions that drive these efforts. The following is the full listing of the complete questions:

Q1: What motivated your site's development and implementation of energy or power aware job scheduling or resource management capabilities?

Q2: Please describe your data center and major high-performance computing system or systems where energy or power aware job scheduling and resource management capabilities have been deployed in a way that covers some or all of the following points of interest.
   (a) Total site power budget or capacity in watts.
   (b) Total site cooling capacity.
   (c) Major high-performance computing system or systems in terms related to: number of cabinets, nodes, and cores; peak performance; node architecture, high-speed network type, memory; peak, average, and idle power draw.

   Other information to help describe site/system level drivers for energy or power aware job scheduling and resource management.

Q3: Describe the general workload on your high-performance computing system or systems. Specifically, any or all of the following would be useful:
   (a) What is running right now, or what does a typical snapshot look like? How many jobs are running? What sizes are these jobs? Generally how long do jobs run?
   (b) What does the backlog of queued jobs look like? How many jobs are currently waiting? What are the sizes of waiting jobs? How long will they run?
   (c) What is the throughput of your system? Approximately how many jobs per month?
   (d) In simple terms, describe your main scheduling goal. Possible examples of scheduling goals might include priority, turn-around time, fairness, efficiency, or system utilization. What percentage of your systems use would you consider to be capability (using the maximum computing power to solve a single large problem in the shortest amount of time) or capacity (using efficient cost-effective computing power to solve a few somewhat large problems or many small problems)?
   (e) If you have statistical information available, what is the minimum, median, maximum, and 10th, 25th, 75th, and 90th percentile job size and wallclock time?

Q4: Describe the energy and power aware job scheduling and resource management capabilities of your large-scale high-performance computing system or systems.

Q5: List and briefly describe all of the elements that comprise your energy and power aware job scheduling and resource management capabilities.
   (a) Include an implementation time component to your answer (this is, when was it implemented?).
   (b) Are these elements commercially available supported products?
   (c) Has there been much non-portable/non-product work done to implement your capabilities?

Q6: Do you have application/task level joint optimization, such as topology-aware task allocation, as a way of directly improving energy consumption or indirectly improving energy consumption (for example, by improving application performance, resulting in reduced wallclock

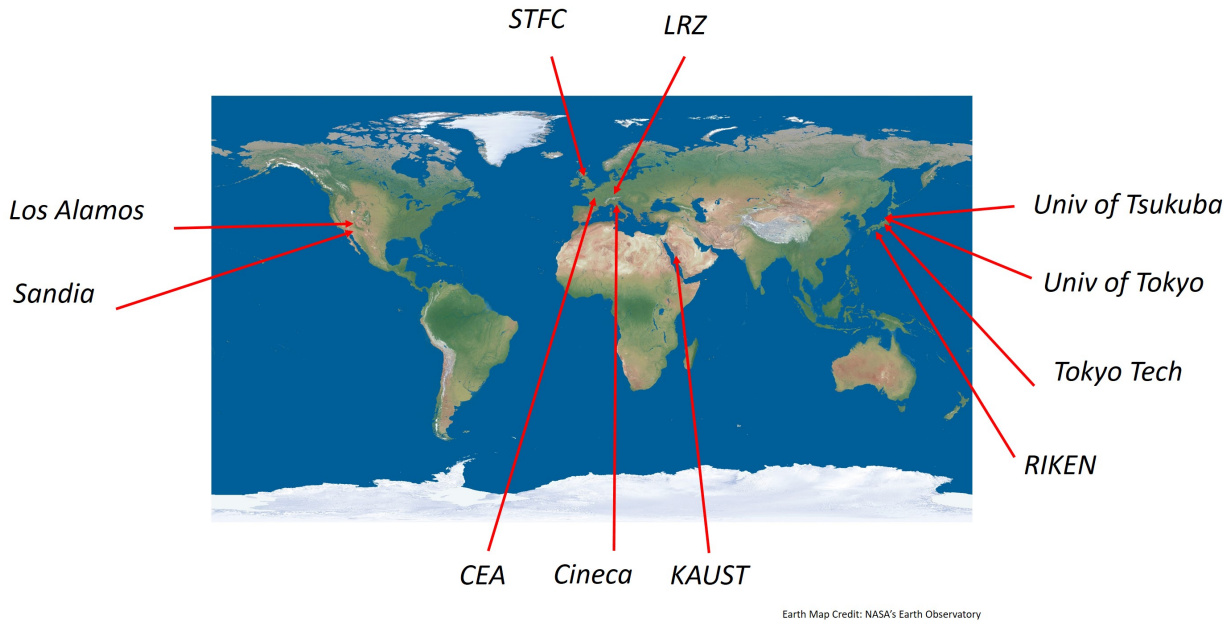

Fig. 2. Map of the geographic location of the particiapting centers. [27]

time)? Did you engage software development communities to improve your energy and power aware job scheduling and resource management solution for this capability?

Q7: How well does your solution work? What are the advantages and disadvantages of your implementation? Describe any results, benefits, or unintended consequences of your implementation.

Q8: What are the next steps for the energy or power aware job scheduling and resource management capability you have developed?

(a) Do you intend to continue site development and/or product deployment?

(b) Will your planned next steps drive new requirements in procurement documents, NRE funding, etc.?

In the following we go into the rationale of the questions. The purpose of the first question is to determine each center's motivations behind pursuing energy and power aware job scheduling and resource management in an attempt to identify motives common among multiple centers. The purpose of the second and third questions is to determine each center's hardware environment (question 2) and the typical workloads running on that hardware (question 3). Any energy and power aware job scheduling and resource management approach needs to take into consideration the hardware and workload characteristics of the given center, so understanding these characteristics is critical for evaluating each center's approach. Similarly, question 4 and question 5 ask about details of each center's EPA JSRM solution, including asking sub-questions that focus on whether the solution is built using vendor-supplied elements and/or elements that have been created custom by the center. Question 4 is the specific point of the questionnaire while question 5 seeks to identify (1) how involved vendors are in helping centers build EPA JSRM solutions, and (2) how heavily centers are using one-off homegrown control systems that might be interesting to study in more detail. Similarly, question 6 asks about very advanced job scheduling techniques such as topology-aware placement of a job's processing elements onto computational resources. A positive response to this question would likely indicate a very high level of sophistication in EPA JSRM techniques that would definitely be interesting to report about in this paper. Further, such sophisticated techniques would likely involve at least some understanding of each application's internal structure and, therefore, require assistance from application developers. The question specifically asks whether such interaction takes place within each center. Finally, the motivation behind question 7 and question 8 are to get a qualitative assessment of each center's EPA JSRM solution as well as potential next steps. Each center is the subject matter expert for their unique solution, so allowing the center an opportunity to openly assess the efficacy of the solution is important.

The answers were first requested in written form and then followed up by a telephone interview to clarify answers and get deeper insight into the responses. This was especially important when sites answered widely different, omitted responses or were going into extensive detail. The total number of pages for responses ranged from 8 to 17 pages per center.

V. Prelude to Survey Analysis

In this section we describe a high level summary of the answers given by the HPC centers.

This section presents a high-level summary of the site responses to the survey, categorized into capabilities that each site is considering in the context of research, technology development with the intent to eventually deploy into production, and those that are actively deployed into each site's production computing environment. Due to the selection process described above, some sites may not have research or technology development efforts, however all sites have some type of production deployment of energy and power aware job scheduling and resource management in place.

The summary of the answers shown in Tables I and II provide a short overview of the activities at the interviewed centers. This work will be followed by an in-depth analysis of the interviews, going into detail in regards to their solutions, the similarities, and applicability. The reader should be reminded that the in depth analysis of the survey is not the primary focus of the work at hand. As mentioned in the abstract this division of scope allows for more in depth coverage of the results.

It should be mentioned that the undertaking of the survey was also presented as a poster submission at SC18 [27]. Additionally a short insight into points from Question 4 were given in an invited article by Siddhartha Jana on the insideHPC website [26]. The full analysis will be synthesised from the raw material of the interview and whitepaper [16] in an upcoming document with consensus of the EE HPC WG EPA JSRM team, as soon as the full analysis is finished.

VI. Related Work

Solutions for Energy and Power Aware Job Scheduling and Resource Management have been approached from several sides, the related work section will look at Academic Research Projects. There are also solutions provided by vendors, but these will not be reviewed in this paper.

Several research avenues have been explored in order to curtail the power and energy consumption of HPC systems.

TABLE I
PART 1 OF THE SUMMARY OF THE ANSWERS FROM EACH CENTER.

Center	Research Activities	Technology Development with Intent to Deploy	Production Development
RIKEN	Integrating job scheduler info with decision to use grid vs. gas turbine energy	Power-aware job scheduling for Post-K, with Fujitsu	3 days for large jobs each month
			Automated emergency job killing if power limit exceeded
			Pre-run estimate of power usage of each job, based on temperature
Tokyo Tech	Activities to facilitate Production Development	Inter-system power capping. TSUBAME2 and TSUBAME3 will need to share the facility power budget.	Resource manager dynamically boots or shuts down nodes to stay under power cap (summer only, enforced over ˜ 30 min window). Interacts with job scheduler to avoid killing jobs. NEC implemented, works cooperatively with PBS Pro.
			Resource manager shuts down nodes that have been idle for a long time.
			Uses virtual machines to split compute nodes. (Complicates physical node shutdown.)
CEA	Analyze collected power and energy info archived long term and use for EPA scheduling	Gives users mark on how well they used power and energy	Energy use provided to users at end of every job
	Investigating how to use and apply `mpi_yield_when_idle`	Together with BULL developing power adaptive scheduling in SLURM	Manually shutting down nodes to shift power budget between systems
	Investigating with BULL power capping and DVFS	Developing 'layout logic' in SLURM, be able to tell what PDUs/Chillers a node or rack depends on and avoid scheduling jobs on them when maintenance	
KAUST	Monitoring and managing power usage under data center power and cooling limits	Analyzing and detecting most power hungry applications in production. Developing optimal power limit constraint strategy for users on Shaheen Cray XC40, while maintaining several HPC systems in production (BG/P and clusters)	Static power capping via Cray CAPMC. 30% of nodes run uncapped, 70% run with 270 W power cap.
			Using SLURM Dynamic Power Management (SDPM) that interfaces with Cray CAPMC (KAUST worked with SchedMD to develop SDPM)
LRZ	Investigating merging SLURM and GEOPM for system energy & power control.	Working on adding energy-aware scheduling capabilities to SLURM, similar to what they have with LoadLeveler today.	First time new app runs: characterized for frequency, runtime and energy.
	Investigating scheduling for power instead of energy		Administrator selects job scheduling goal, energy to solution or best performance.
	Linking job scheduler with IT infrastructure + cooling; scheduler may delay jobs when IT infrastructure is particularly inefficient		LRZ worked with IBM on energy-aware scheduling support in LoadLeveler, now ported to LSF.

TABLE II
PART 2 OF THE SUMMARY OF THE ANSWERS FROM EACH CENTER.

Center	Research Activities	Technology Development with Intent to Deploy	Production Development
STFC	IBM/LSF energy-aware scheduling is experimented with on small-scale (360 node) system	Deployment of reporting tool for user power consumption at the job level. (Fine as well as coarse granularity)	Continuously collecting power and energy system monitoring info, data center, machine, and job levels
	Programmable interface (PowerAPI-based) for application power measurements of code segments (with interface to JSRM)		
	Investigation of power aware policies using higher level abstract e.g., GEOPM and Job Scheduler.		
LANL + Sandia	Analyzing power system monitoring info to assess potential of EPA scheduling, gather traces for evaluating EPA approaches.	EPA job scheduling support developed with Adaptive Inc. for MOAB/Torque, interfaces with Cray CAPMC and Power API. Trinity is now using SLURM, but MOAB work remains available for future use.	Cray CAPMC power capping infrastructure, out-of-band control, administrator ability to set system-wide and node-level power caps (available on all Cray XC systems).
		Developed Power API implementation with Cray, utilized by MOAB/Torque for EPA job scheduling.	
CINECA	Scalable power monitoring, used to predict per-job power use and used to generate predictive models for node power and temperature evolution (with University of Bologna)	Developing together with E4 EPA job scheduling support in SLURM. Also tracking EPA SLURM work being done by BULL and SchedMD.	EPA job scheduling on Eurora system (now decommissioned) using PBSPro, collaboration with Altair
HCAHPC (University of Tsukuba and the University of Tokyo)	Activities to facilitate Production Development.	–	Ability to set power caps for groups of nodes via the resource manager (Fujitsu proprietary product)
			Manual emergency response, admin sets power cap.
			Delivering post-job energy use reports to users.

In the rest of this section we are going to present a brief overview of the techniques found in the literature and current state-of-the-art.

A detailed survey of the research on power management techniques for high performance systems can be found in [12], [30], [32]. As clearly pointed out by Hsu et al. [22], a change of the current perspective is required: the focus must shift from performance-based metrics (such as the performance-power ration) to new ones which take into account different aspects of the problem, i.e. integrating the notions of total cost of ownership, productivity and reliability.

Several works aim at curbing the overall power consumption at the supercomputer level., i.e. power capping. Sarood et al. [38] describe an integer linear programming model to enforce power capping in an HPC cluster through over-provisioning. Their approach combines over-provisioning with a power-aware scheduler. Mammela et al. [33] present an energy-aware scheduler that turns off idle nodes every time the scheduler detects that no activity can be scheduled for a sufficiently long time on a certain node.

Many approaches take advantage of "moldable jobs", i.e., jobs which can run with different configurations (number of nodes, cores or threads) [5], [35], [37]. Given the current power consumption and power budget, the best configuration is chosen for each job before its start. Other authors tried to exploit the power and performance variability among nodes and components within the same system [25], [39].

Dynamic Voltage and Frequency Scaling (DVFS) allows one to exchange processor performance for lower power consumption and has been explored as a means to increase the supercomputer's energy-efficiency [4], [20], [21]. However, reducing the operating clock may increase the duration of the applications which run on the now more energy efficient resources [4], [20], [23]. Approaches to overcome this issue take advantage of compute, memory, communication phases [21] or heterogeneous nodes [20]. An extension of these approaches into job scheduling is discussed by Etinski et al. [18], [19], which extends the standard job scheduling algorithm with power budgeting capability through DVFS.

An alternative to the direct control of frequency scaling is Intel's Running Average Power Limit (RAPL) [13], which provides a software configurable and hardware enforced power cap. Several works have combined Intel's RAPL feature with job scheduling algorithms [8], [17]. These approaches rely on allocating a reduced power budget to each node and sockets. Ellsworth et al. [17] dynamically share the budget between nodes aiming to give more power to the nodes which run critical jobs and processes.

An orthogonal approach to achieving a system level power budget does not limit the performance of the processing elements, but limits the jobs concurrently running on the computational resources [9]–[11].

Other work focuses on minimizing the energy consumption and/or the related energy costs [4], [7], [28], [29]. These energy aware schedulers and resource managers act on the

job execution order alone, without requiring any hardware modification nor any change in the operational frequencies of the computing nodes.

A very important aspect for energy and power aware job schedulers and resource managers is knowledge of an application's features before its execution; this allows for the JSRM technologies to make EPA informed decisions. Application features may include the application's tag, historical data and model regression [40], user's meta-information, such as a tag identifying similar jobs [4], machine learning techniques and job submission information [9], [41].

From the above discussion, we can see that in recent years several techniques, algorithms and strategies have been proposed in the state-of-the-art for energy and power aware job schedulers and resource managers in the high performance computing domain. The findings of the survey described in this paper show early attempts at deploying EPA JSRM technologies on large scale production deployments in supercomputing centers. Understanding this gap between research and current practice trends can be very important in setting future research agendas.

VII. Next Steps in Survey Analysis

This paper presents an overview and initial summary of a comprehensive survey of Energy and Power Aware Job Scheduling and Resource Management techniques employed in nine Top 500 high-performance computing centers in the United States, Europe, and Asia. The paper explains the center selection criteria and the specific questions asked during the survey process along with the motivations behind these questions.

Each center presents a unique combination of characteristics related to funding structure, geopolitical situation, and geographic circumstances. Accordingly, each center approaches EPA JSRM in a slightly different way. This paper presents a prelude to the survey analysis, giving a general overview of the techniques in research, deployment, and assessment toward production.

The EE HPC WG EPA JSRM team is currently developing a detailed analysis of the survey results. This analysis will not only explore each site's response to each question in greater depth than the current paper, but will also identify common themes in the responses as well as identify any particularly noteworthy approaches or techniques employed at specific sites.

Finally, since the EE HPC WG works within the broad high-performance computing community, future work will seek to identify general recommendations based on our observations from the survey. We expect these recommendations to include guidance to hardware and software vendors about potential product features that would likely be useful to sites developing EPA JSRM solutions. We also expect these recommendations to include guidance to high-performance computing centers regarding EPA JSRM tools and techniques that have been particularly useful to contemporary centers.

Acknowledgment

The EE HPC WG EPA JSRM team would like to thank the participating HPC Centers to provide us with their time and knowledge for the interviews, and reviews thereof. We would especially like to thank Keiji Yamamoto, Fumiyoshi Shoji, Kazuo Minami, Toshio Endo, Francis Belot, Gilles Wiber, Matthieu Hautreux, Jean-Marc Ducos, Bilel Hadri, Torsten Wilde, Michael Rudgyard, Neil Morgan, Stephen Hill, James Laros, Josip Loncaric, Carlos Cavazzoni, Toshihiro Hanawa, Masaaki Kondo and all others participating in the interviews.

References

[1] *19th International Parallel and Distributed Processing Symposium (IPDPS 2005), CD-ROM / Abstracts Proceedings, 4-8 April 2005, Denver, CO, USA*. IEEE Computer Society, 2005.

[2] I. Adaptive Computing Enterprises. Moab workload manager administrator guide 9.0.3. http://docs.adaptivecomputing.com/9-0-3/MWM/Moab-9.0.3.pdf.

[3] Altair, PBS Works. PBS Professional 13.0 Users Guide. http://www.pbsworks.com/pdfs/PBSUserGuide13.0.pdf.

[4] A. Auweter, A. Bode, M. Brehm, L. Brochard, N. Hammer, H. Huber, R. Panda, F. Thomas, and T. Wilde. A case study of energy aware scheduling on supermuc. In J. M. Kunkel, T. Ludwig, and H. W. Meuer, editors, *Supercomputing*, pages 394–409, Cham, 2014. Springer International Publishing.

[5] P. E. Bailey, A. Marathe, D. K. Lowenthal, B. Rountree, and M. Schulz. Finding the limits of power-constrained application performance. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '15, pages 79:1–79:12, New York, NY, USA, 2015. ACM.

[6] N. Bates, G. Ghatikar, G. Abdulla, G. A. Koenig, S. Bhalachandra, M. Sheikhalishahi, T. Patki, B. Rountree, and S. Poole. Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges. *Informatik-Spektrum*, 38(2):111–127, Apr 2015.

[7] A. A. Bhattacharya, D. Culler, A. Kansal, S. Govindan, and S. Sankar. The need for speed and stability in data center power capping. *Sustainable Computing: Informatics and Systems*, 3(3):183 – 193, 2013. Selected papers from the 2012 IEEE International Green Computing Conference (IGCC 2012).

[8] D. Bodas, J. Song, M. Rajappa, and A. Hoffman. Simple power-aware scheduler to limit power consumption by hpc system within a budget. In *Proceedings of the 2Nd International Workshop on Energy Efficient Supercomputing*, E2SC '14, pages 21–30, Piscataway, NJ, USA, 2014. IEEE Press.

[9] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini. Predictive modeling for job power consumption in hpc systems. In J. M. Kunkel, P. Balaji, and J. Dongarra, editors, *High Performance Computing*, pages 181–199, Cham, 2016. Springer International Publishing.

[10] A. Borghesi, F. Collina, M. Lombardi, M. Milano, and L. Benini. Power capping in high performance computing systems. In G. Pesant, editor, *Principles and Practice of Constraint Programming*, pages 524–540, Cham, 2015. Springer International Publishing.

[11] A. Borghesi, C. Conficoni, M. Lombardi, and A. Bartolini. Ms3: A mediterranean-stile job scheduler for supercomputers - do less when it's too hot! In *2015 International Conference on High Performance Computing Simulation (HPCS)*, pages 88–95, July 2015.

[12] K. W. Cameron, R. Ge, and X. Feng. High-performance, power-aware distributed computing for scientific applications. *IEEE Computer*, 38(11):40–47, 2005.

[13] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, and C. Le. Rapl: Memory power estimation and capping. In *Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design*, ISLPED '10, pages 189–194, New York, NY, USA, 2010. ACM.

[14] J. Eastep, S. Sylvester, C. Cantalupo, B. Geltz, F. Ardanaz, A. Al-Rawi, K. Livingston, F. Keceli, M. Maiterth, and S. Jana. Global extensible open power manager: A vehicle for hpc community collaboration on co-designed energy management solutions. In J. M. Kunkel, R. Yokota, P. Balaji, and D. Keyes, editors, *High Performance Computing*, pages 394–412, Cham, 2017. Springer International Publishing.

[15] EE HPC WG. Energy efficient high performance computing working group (ee hpc wg) website. https://eehpcwg.llnl.gov/.

[16] EE HPC WG - EPA JSRM team. Whitepaper, 'energy and power aware job scheduling and power management'. https://eehpcwg.llnl.gov/documents/conference/sc17/sc17_bof_epa_jsrm_whitepaper_110917_rev_1.pdf.

[17] D. A. Ellsworth, A. D. Malony, B. Rountree, and M. Schulz. Dynamic power sharing for higher job throughput. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '15, pages 80:1–80:11, New York, NY, USA, 2015. ACM.

[18] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Optimizing job performance under a given power constraint in hpc centers. In *International Conference on Green Computing*, pages 257–267, Aug 2010.

[19] M. Etinski, J. Corbalan, J. Labarta, and M. Valero. Parallel job scheduling for power constrained hpc systems. *Parallel Comput.*, 38(12):615–630, Dec. 2012.

[20] F. Fraternali, A. Bartolini, C. Cavazzoni, and L. Benini. Quantifying the impact of variability and heterogeneity on the energy efficiency for a next-generation ultra-green supercomputer. *IEEE Transactions on Parallel and Distributed Systems*, PP(99):1–1, 2017.

[21] V. W. Freeh, D. K. Lowenthal, F. Pan, N. Kappiah, R. Springer, B. L. Rountree, and M. E. Femal. Analyzing the energy-time trade-off in high-performance computing applications. *IEEE Transactions on Parallel and Distributed Systems*, 18(6):835–848, June 2007.

[22] C. Hsu, W. Feng, and J. S. Archuleta. Towards efficient supercomputing: A quest for the right metric. In *19th International Parallel and Distributed Processing Symposium (IPDPS 2005), CD-ROM / Abstracts Proceedings, 4-8 April 2005, Denver, CO, USA* [1].

[23] C.-h. Hsu and W.-c. Feng. A power-aware run-time system for high-performance computing. In *Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*, SC '05, pages 1–, Washington, DC, USA, 2005. IEEE Computer Society.

[24] IBM. IBM Spectrum LSF Energy Aware Scheduling, LSF10.1.0. https://www.ibm.com/support/knowledgecenter/en/SSWRJV_10.1.0/lsf_welcome/lsf_kc_eas.html, online.

[25] Y. Inadomi, T. Patki, K. Inoue, M. Aoyagi, B. Rountree, M. Schulz, D. Lowenthal, Y. Wada, K. Fukazawa, M. Ueda, M. Kondo, and I. Miyoshi. Analyzing and mitigating the impact of manufacturing variability in power-constrained supercomputing. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '15, pages 78:1–78:12, New York, NY, USA, 2015. ACM.

[26] S. Jana. From sc17: Energy efficiency in the software stack cross community efforts. https://insidehpc.com/2017/12/sc17-energy-efficiency-software-stack-cross-community-efforts/.

[27] S. Jana, G. A. Koenig, M. Maiterth, K. T. Pedretti, A. Borghesi, A. Bartolini, B. Hadri, and N. J. Bates. P90: Global survey of energy and power-aware job scheduling and resource management in supercomputing centers. http://sc17.supercomputing.org/presentation/?id=post236&sess=sess293.

[28] B. Khemka, R. Friese, S. Pasricha, A. A. Maciejewski, H. J. Siegel, G. A. Koenig, S. Powers, M. Hilton, R. Rambharos, and S. Poole. Utility maximizing dynamic resource management in an oversubscribed energy-constrained heterogeneous computing system. *Sustainable Computing: Informatics and Systems*, 5:14 – 30, 2015.

[29] K. Leal. Energy efficient scheduling strategies in federated grids. *Sustainable Computing: Informatics and Systems*, 9(Complete):33–41, 2016.

[30] Y. Liu and H. Zhu. A survey of the research on power management techniques for high-performance systems. *Softw. Pract. Exper.*, 40(11):943–964, Oct. 2010.

[31] LRZ. Decision criteria and benchmark description for the acquisition of the european high performance computer SuperMUC at LRZ, March 2010. https://www.lrz.de/wir/berichte/TB/LRZ-Bericht-2010-03.pdf.

[32] M. Maiterth, T. Wilde, D. Lowenthal, B. Rountree, M. Schulz, J. Eastep, and D. Kranzlmueller. *Power aware high performance computing: Challenges and opportunities for application and system developers - Survey & tutorial*, pages 3–10. Institute of Electrical and Electronics Engineers Inc., United States, 9 2017.

[33] O. Mämmelä, M. Majanen, R. Basmadjian, H. De Meer, A. Giesler, and W. Homberg. Energy-aware job scheduler for high-performance computing. *Computer Science - Research and Development*, 27(4):265–275, Nov 2012.

[34] S. Morris Jette. Technical: Slurm power management support. https://slurm.schedmd.com/SLUG15/Power_mgmt.pdf, technical presentation.

[35] A. W. Mu'alem and D. G. Feitelson. Utilization, predictability, workloads, and user runtime estimates in scheduling the ibm sp2 with backfilling. *IEEE Trans. Parallel Distrib. Syst.*, 12(6):529–543, June 2001.

[36] T. Patki, N. Bates, G. Ghatikar, A. Clausen, S. Klingert, G. Abdulla, and M. Sheikhalishahi. Supercomputing centers and electricity service providers: A geographically distributed perspective on demand management in europe and the united states. In J. M. Kunkel, P. Balaji, and J. Dongarra, editors, *High Performance Computing*, pages 243–260, Cham, 2016. Springer International Publishing.

[37] T. Patki, D. K. Lowenthal, A. Sasidharan, M. Maiterth, B. L. Rountree, M. Schulz, and B. R. de Supinski. Practical resource management in power-constrained, high performance computing. In *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '15, pages 121–132, New York, NY, USA, 2015. ACM.

[38] O. Sarood, A. Langer, A. Gupta, and L. Kale. Maximizing throughput of overprovisioned hpc data centers under a strict power budget. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 807–818, Nov 2014.

[39] H. Shoukourian. Adviser for energy consumption management: green energy conservation.

[40] H. Shoukourian, T. Wilde, A. Auweter, and A. Bode. Predicting the energy and power consumption of strong and weak scaling hpc applications. *Supercomput. Front. Innov.: Int. J.*, 1(2):20–41, July 2014.

[41] A. Sîrbu and O. Babaoglu. Power consumption modeling and prediction in a hybrid cpu-gpu-mic supercomputer. In *Proceedings of the 22Nd International Conference on Euro-Par 2016: Parallel Processing - Volume 9833*, pages 117–130, New York, NY, USA, 2016. Springer-Verlag New York, Inc.

[42] TOP500.org. Top500 list of fastest supercomputers in the world. https://www.top500.org/.