Global Survey of Energy and Power-aware Job Scheduling and Resource Management in Supercomputing Centers

Siddhartha Jana, Gregory A. Koenig, Matthias Maiterth, Kevin T. Pedretti
Andrea Borghesi, Andrea Bartolini, Bilel Hadri, Natalie J. Bates
Intel Corporation, Energy Efficient HPC Working Group, Sandia National Laboratories, University of Bologna
ETH Zurich, KAUST Supercomputing Lab, Energy Efficient HPC Working Group

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Software and its engineering** → *Software libraries and repositories*;

KEYWORDS

HPC, Energy efficiency, Power Management, Job Scheduling, Resource Management

ACM Reference format:
Siddhartha Jana, Gregory A. Koenig, Matthias Maiterth, Kevin T. Pedretti and Andrea Borghesi, Andrea Bartolini, Bilel Hadri, Natalie J. Bates. 2017. Global Survey of Energy and Power-aware Job Scheduling and Resource Management in Supercomputing Centers. In *Proceedings of SC'17 Conference, Denver, CO, USA, July 2017 (SC'17),* 3 pages.
https://doi.org/

1 BACKGROUND

One of the major challenges that supercomputing centers face in building systems for high performance computing involves issues of energy consumption. Contemporary petascale systems can have peak power demands that exceed 20 megawatts and instantaneous power fluctuations of 8 megawatts. Despite ongoing improvements in microarchitectures and the use of high degrees of parallelization found in accelerator-based systems, the expectation is that the energy draw of large-scale systems will continue to increase as the community moves toward exascale systems.

2 INTRODUCTION

In the past, the Energy Efficient High Performance Computing Working Group (EEHPCWG) has published work in the past that surveys how supercomputing centers in the United States [1] and Europe[2] have been approaching problems related to energy consumption.

SC'17, July 2017, Denver, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/

This study is one of the first efforts within the HPC community that attempts to bring together representatives of multiple supercomputing centers across the globe with the sole purpose of discussing energy and power-aware job scheduling and resource management (EPA-JSRM). As preparation for this work, the authors have conducted an extensive survey of several Top500 supercomputing sites from around the world: CEA (Alternative Energies and Atomic Energy Commission, France), Cineca (Italy), KAUST (King Abdullah University of Science and Technology, Saudi Arabia), LRZ (Leibniz Supercomputing Centre, Germany), RIKEN (Japan), STFC (Science and Technology Facilities Council, United Kingdom), Tokyo Institute of Technology (Japan), University of Tokyo (Japan), and University of Tsukuba (JCAHPC, Japan), Los Alamos and Sandia National Laboratories (Trinity, United States).The goal of this survey has been to collect and analyze information regarding:

- Motivation behind investing in EPA-JSRM related activities
- Target infrastructure that is expected to be controlled by JSRM frameworks (e.g. site-wide power budget, cooling capacity, etc.)
- Workload characteristics of the systems
- Adopted design for EPA-JSRM
- Implementation details
- Application/task level and topology-aware solutions
- Results and challenges
- Next steps including system procurement

The purpose of this poster is to investigate the state of the practice in energy and power aware job scheduling techniques, not to advocate for any particular technique or techniques

3 LESSONS LEARNED

In this poster, we plan to tackle each of the questions listed above by highlighting the key takeaways from the sites' responses. We present an overview below. For brevity, in this extended abstract, we focus more on the responses, and less on mapping them to the names of the actual sites.

3.1 Motivation for the sites

One of the major concerns observed among the sites, was the anticipation of inevitable shortage of power and energy available for running future systems. Another strong motivation has been the enforcement of hard upper limits on power consumption, set by contractual agreements between the site management and funding sources.

Some supercomputing centers are facing the challenge of limited power and cooling capacity available on-site. The potential for significant cost-savings associated with higher energy efficiency is another strong motivation. Other sites are pursuing the goal of "being green" and ecologically responsible by investing in such JSRM techniques. For some sites, achieving a low PUE target is the primary goal behind their investments in energy and power-aware techniques. Another interesting motivation for JSRM is to ration the availability of cooling infrastructure so that the thermal capacity can be directed from workloads with low-demand towards those with higher power-signatures.

In short, all the challenges currently faced by the centers, relate to reducing operational costs in one form or the other.

3.2 JSRM solutions adopted by the sites

When it comes to adopting JSRM approaches within supercomputing centers, two different focus areas arise - monitoring and management of energy/power based metrics.

- Monitoring:
  - There have been efforts to measure energy/power continuously, both in-band as well as out-of-band.
  - While some centers rely on sensors that make direct real-time measurements of energy/power consumed, others use thermal-based sensors together with a model to indirectly derive power and energy metrics.
  - Some centers are working on exposing high-level APIs and feedback mechanisms to enable end-users and resource managers to monitor the power consumed by various system components. The goal is to provide an interface that can be used by future JSRM solutions for making informed decisions while allocating power budgets.
  - Some sites have also been investing in implementing statistical approaches for predicting energy/power demands of a give job based on its user-id, resource request, job-size, and job-length. Such cases are useful for sites where energy/power requirements by the applications can be mapped to specific users submitting the jobs.
- Power Management:
  - One notable approach being adopted by supercomputing sites is dynamically shutting down jobs as a response to system power reaching the total budget. The decision on what specific jobs to shut down depends on a number of factors like the job-size, application power, job-length, etc. This is a reactive approach.
  - A proactive solution to the budget overflow problem is to configure the resource manager to reduce the number of nodes available for job allocation. This approach reduces the theoretical maximum power that can be consumed by the system, but at the cost of reducing system utilization.
  - Another notable approach being adopted is leveraging power-capping mechanisms supported by CPU and system vendors to ensure that the total budget remains below a specific limit. This hardware enforced limit is usually applied over a specific time-window (in the order of minutes).
  - An alternative approach to using power-capping is triggering the system processing units to operating at specific operating frequencies (hardware p-states). Some sites are working on designing portable APIs to enable users and system admins to list the actual value of the frequencies.
  - It has been observed that there is very little 'intuition' among end users when it comes to mapping application behavior to specific p-states. And so, many sites are working towards system-wide frameworks (job schedulers, resource managers, etc.) that attempt at using static prediction models to map applications to their ideal p-states. While this approach gives promising power efficiency metrics, the variation in application performance during repeated runs is being investigated.
  - One feedback-driven approach being explored is "tagging" applications based on their power efficiency attained over past job runs. It is expected that these "tags" will then be used by future resource managers while allocating resources during future job requests. This directly addresses the challenge in having the job scheduler be responsible to identify and maintain a historical record of unique set of job characteristics.

4 CONCLUSION AND FUTURE WORK

This poster takes a significant step towards highlighting the state-of-art practices for energy and power aware job scheduling and resource management (EPA-JSRM). The results discussed are taken from a recent survey that interviewed ten participating centers across the globe. In fact, most of the approaches discussed in this poster have been used for driving new requirements in system procurement documents by those sites.

The EPA-JSRM solutions discussed have already been adopted - at least, in parts by these sites. The next phase of the road-map for some of these sites, is to continue working on more stable designs of system-wide frameworks that allocate resources in a power-aware manner. Job schedulers and resource managers are some examples of such frameworks. Other sites plan to invest in robust energy/power predictors that rely on statistical modeling of historical records of job runs. Power-capping mechanisms exposed by vendors is another promising approach that is being actively investigated.

The EEHPC WG solicits feedback from the HPC community to improve upon this survey efforts. We plan on drafting a white paper that describes the survey results in greater detail. Additional details at https://eehpcwg.llnl.gov/

REFERENCES

[1] Natalie Bates, Girish Ghatikar, Ghaleb Abdulla, Gregory A. Koenig, Sridutt Bhalachandra, Mehdi Sheikhalishahi, Tapasya Patki, Barry Rountree, and Stephen Poole. 2015. Electrical Grid and Supercomputing Centers: An Investigative Analysis of Emerging Opportunities and Challenges. *Informatik-Spektrum* 38, 2 (01 Apr 2015), 111–127. https://doi.org/10.1007/s00287-014-0850-0

[2] Tapasya Patki, Natalie Bates, Girish Ghatikar, Anders Clausen, Sonja Klingert, Ghaleb Abdulla, and Mehdi Sheikhalishahi. 2016. *Supercomputing Centers and Electricity Service Providers: A Geographically Distributed Perspective on Demand Management in Europe and the United States*. Springer International Publishing, Cham, 243–260. https://doi.org/10.1007/978-3-319-41321-1_13