PERQ: Fair and Efficient Power Management of Power-Constrained Large-Scale Computing Systems

Tirthak Patel
Northeastern University
patel.ti@husky.neu.edu

Devesh Tiwari (advisor)
Northeastern University
tiwari@northeastern.edu

ABSTRACT

Background: High-performance computing (HPC) has enabled computational scientists to expedite the scientific discovery, but the exceedingly high power consumption of large-scale HPC systems is one of the top ten challenges for future exascale systems. Consequently, strict power constraints are placed on modern-day systems, which requires intelligent power management to obtain a high system throughput (jobs completed per unit time). Hardware over-provisioning has been shown as an effective technique to increase the efficiency of power-constrained large-scale systems [5, 9].

Traditionally, a system designer would fill a system with as many compute nodes as can be simultaneously powered up at their peak capacity (i.e., their thermal design power or TDP) under the given system power budget – referred to as *worst-case provisioning. However, HPC applications typically only consume 25%-70% of a node's TDP.* Thus, HPC systems can be over-provisioned with more nodes than the system's power budget can simultaneously accommodate at peak power. Over-provisioning enables the system to concurrently execute more jobs compared to a worst-case provisioned system, thereby, increasing the system throughput. Recognizing this opportunity, researchers have worked toward making over-provisioned systems more efficient and economical [1, 2, 5–10].

Limitations of Existing Approaches: One power-management approach for an over-provisioned system is to execute jobs on all the nodes and cap their power at the same level. This approach is promising since it can achieve higher job throughput by using more nodes in parallel. Moreover, it is also "fair" to different jobs since the power is distributed evenly across all nodes. However, prior works have shown that such a "fairness-oriented policy" does not yield sufficient improvement in system throughput to overcome the capital and operational cost of over-provisioning [1, 3, 4, 6]. Therefore, to outweigh these costs, many "throughput-oriented policies" have also been proposed. One such approach is to give maximum power to jobs which are the closest to completion and are running on fewest nodes. While effective in improving system throughput, such policies are unfair by design. *We note that although large-scale HPC systems are primarily designed for high performance, failure to integrate fairness into resource management can have undesirable adverse effects (e.g., unintended delay in scientific discoveries, inaccurate resource consumption accounting, unfair treatment of users).*

Goal: In summary, power management should meet two conflicting objectives simultaneously: (1) achieve high system throughput to outweigh the cost of over-provisioning, and (2) maintain fairness among jobs. A learning-based, rule-based, or ad-hoc power management strategy would not be able to provably fulfill both the objectives simultaneously because such approaches lack dynamic feedback and theoretical underpinnings.

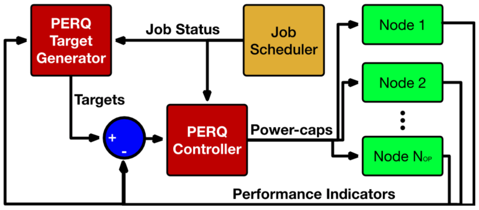

Figure 1: Conceptual overview of PERQ feedback control.

Our Contribution: PERQ: To this end, we propose PERQ, a multi-objective control-theoretic policy to allocate power in a way which achieves high system throughput while being fair to the jobs. As shown in Fig. 1, PERQ uses dynamic feedback to adjust to jobs with diverse characteristics and to assess the performance impact of its power-capping decisions. PERQ leverages the observation that HPC applications have different sensitivities to power-capping: some jobs perform as well at lower power as they do at higher power, while others are more sensitive. Using multi-objective control, PERQ carefully reduces the power allocation of certain jobs and increases the power allocation of other jobs to maximize system throughput. However, making these decisions requires an accurate estimation of the relationship between the power allocation and the performance for different jobs. This is the most challenging task toward providing fair and efficient power management. To overcome this challenge, PERQ builds a novel state-space system model, derived using system identification theory, that accurately estimates power allocation vs. performance relationship without over-fitting the model to training workloads.

PERQ Evaluation Results: PERQ's experimental and simulation-based evaluation shows that it leverages the differences in power-cap sensitivity of different jobs to achieve higher system throughput for over-provisioned systems while remaining fair to concurrently running jobs. PERQ's evaluation is driven by characteristics of real-world large-scale HPC systems and jobs. Although PERQ uses one set of benchmarks to build the system model, it is evaluated using a different set of applications, confirming that it is effective for a diverse set of applications. PERQ provides adaptive, stable, and fair treatment for jobs of different characteristics, and better system throughput for systems with different levels of over-provisioning. PERQ improves system throughput by up to 50% points, compared to the fairness-oriented allocation policy, while being up to 100% points more fair than a variety of throughput-oriented policies. PERQ has low overhead and works effectively across a wide range of control parameters. Overall, PERQ helps HPC systems be fair and achieve a higher profit over the capital and operational expenses of over-provisioning. PERQ experiments and measurement data are available at `https://github.com/GoodwillComputingLab/PERQ`.

REFERENCES

[1] Neha Gholkar, Frank Mueller, et al. 2016. Power Tuning HPC Jobs on Power-Constrained Systems. In *Proceedings of the 2016 International Conference on Parallel Architectures and Compilation*. ACM, 179–191.

[2] Neha Gholkar, Frank Mueller, et al. 2018. PShifter: Feedback-Based Dynamic Power Shifting within HPC Jobs for Performance. In *2018 HPDC*.

[3] Yanpei Liu, Guilherme Cox, et al. 2016. FastCap: An Efficient and Fair Algorithm for Power Capping in Many-Core Systems. In *Performance Analysis of Systems and Software (ISPASS), 2016 IEEE International Symposium on*. IEEE, 57–68.

[4] Frank Mueller, Barry Rountree, et al. 2016. *Power Tuning for HPC Jobs under Manufacturing Variations*. Technical Report. North Carolina State University. Dept. of Computer Science.

[5] Tapasya Patki et al. 2013. Exploring Hardware Overprovisioning in Power-Constrained, High Performance Computing. In *Proceedings of the International Conference on Supercomputing*. ACM, 173–182.

[6] Tapasya Patki et al. 2015. Practical Resource Management in Power-Constrained, High Performance Computing. In *Proceedings of the Symposium on High-Performance Parallel and Distributed Computing*. ACM, 121–132.

[7] Tapasya Patki et al. 2016. Economic Viability of Hardware Overprovisioning in Power-Constrained High Performance Computing. In *Proceedings of the Workshop on Energy Efficient Supercomputing*. IEEE Press, 8–15.

[8] Ryuichi Sakamoto, Tapasya Patki, et al. 2018. Analyzing Resource Trade-offs in Hardware Overprovisioned Supercomputers. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 526–535.

[9] Osman Sarood, Akhil Langer, et al. 2013. Optimizing Power Allocation to CPU and Memory Subsystems in Overprovisioned HPC Systems. In *Cluster Computing (CLUSTER), 2013 IEEE International Conference on*. IEEE, 1–8.

[10] Will Whiteside, Shelby Funk, et al. 2017. PANN: Power Allocation via Neural Networks Dynamic Bounded-Power Allocation in High Performance Computing. In *Proceedings of the 5th International Workshop on Energy Efficient Supercomputing*. ACM, 8.