Power Management

ANL: Sridutt Bhalachandra, Kamil Iskra, Swann Perarnau, Valentin Reis, Kazutomo Yoshii
LLNL: Tapasya Patki

Argo Project

The goal of Argo is to augment and optimize existing OS/R components for use in production HPC systems, providing portable, open source, integrated software that improves the performance and scalability of and that offers increased functionality to exascale applications and runtime systems.

For an overview of Argo, please refer to https://www.argo-osr.org/.

Power Management Overview

Dynamic power management is expected to be critical in the exascale era, both in terms of not exceeding the overall available power budget and in terms of utilizing the available power to make the most application progress. In the Argo project, we employ hierarchical power management that includes both system-global and node-local mechanisms and policies.

Recently, PowerStack has emerged as a new community-wide effort for power management in HPC systems. We—together with our partners from industry, academia, and other research labs across the United States, Europe, and Asia—are the founding members of this initiative. PowerStack is expected to leverage some of the tools we have developed; parts of the Argo power management infrastructure will also serve as first prototypes.

Global Power Management

Global management of power in Argo is handled by the Global Resource Manager (GRM). As part of this effort we added dynamic power awareness to the popular batch scheduler SLURM, as shown in the figure below. Our collaborators from the University of Tokyo tested it on 960 nodes of the HA8K system, analyzing idle power and throughput. The results in the figure on the right show the benefits of hardware overprovisioning at scale as well as sensitivity to the degree of overprovisioning.

We also extended the Flux scheduler with power awareness addressing static manufacturing variation; in an experiment on a 1,200-node system at LLNL this resulted in a 2.3× reduction in rank-to-rank performance variation under a power cap.
Node Power Management

Power management at the node level is built into the Argo Node Resource Manager (NRM). As depicted on the right, it works in a closed control loop, obtaining goals (power limit) from the GRM, acting on application workloads launched within containers by, for example, adjusting the CPU p-states or changing the power cap via the Intel RAPL mechanism, and getting feedback through the monitoring of hardware sensors measuring, for example, power draw, temperature, fan speed, and frequency.

Critical to the success of this approach is the self-reporting by applications. We provide a simple API that application processes can use to periodically update the NRM on their progress. This gives NRM reliable feedback on the efficacy of its power policies, and it can also be used for a more robust identification of the critical path, rather than relying on heuristics based on performance counters.

The figure on the right shows the results from early experiments with a bulk-synchronous parallel application-aware power policy on a number of benchmarks and applications. NRM identifies the critical path across application phases to dynamically adapt power, showing an energy reduction of up to 15%.

Planned Improvements

GRM:
- Integration of the capabilities of GRM with those of the power-aware runtime developed in Intel's GEOPM and refined as part of the ECP ST 2.3.1.12 PowerSteering project

NRM:
- Exploiting multiple power controls for improved energy efficiency in presence of workload imbalance
- Improvements to energy efficiency in the presence of memory bottlenecks
- Exploring energy efficiency opportunities on future heterogeneous architectures

Status

Our software is open source. GRM can be found at https://github.com/tpatki/power-slurm. Node-local components can be found at https://xgitlab.cels.anl.gov/argo (spread across the nrm and libnrm repositories, with containers being a dependency). We are always looking for additional collaborators with more workloads. If you think that your project could benefit from power management, we encourage you to get in touch with us.

Contact

For more information, please contact Tapasya Patki patkil@llnl.gov or Swann Perarnau swann@anl.gov.