Energy Demand Response Modeling for High Performance Computing Systems

Kishwar Ahmed and Jason Liu
Florida International University

Demand response aims at energy reduction and power stability during peak load periods by providing financial incentives to participants to reduce their energy consumption. Various demand response programs are offered by energy service providers to encourage energy reduction from participants. There are two types of demand response programs: economic demand response and emergency demand response. In the former, participants voluntarily enroll in the programs (without the need of prior commitment) and willingly reduce the load based on incentives offered by the supplier. In the latter, prior agreement from the participants are necessary, possibly in exchange of a reduced energy price. Once enrolled, the participants are obliged to reduce the energy consumption upon requests from the suppliers when supply shortages or emergency conditions occur.

High Performance Computing (HPC) systems can consume an enormous amount of energy during their operation. The energy cost is a major component of the overall cost of operation of an HPC system. Any reduction in the electricity bill can be a significant benefit for HPC facilities. In this work, we attempt to address the following question: *Can HPC systems reduce the energy consumption and energy cost through participation in emergency and economic demand response?*

To enable HPC system's emergency demand response participation, we first developed a detailed emergency demand response participation model based on frequency scaling [1]. The challenge is to balance the potential loss of performance against the possible gain in power system stability and energy reduction. We developed job scheduling and resource allocation models to balance application power and performance to enable demand response participation. The job scheduling model requires facilities such as job eviction and restart in response to the reduced power level during emergency demand response periods. We have also explored different resource allocation capabilities, such as power capping and job scaling to reduce energy consumption during the demand response periods [2]. In both cases, a carefully designed resource provisioning model is required in order to achieve optimal energy conservation and power stability during the demand response periods. To facilitate power and performance prediction of HPC applications with unknown input parameters, our model includes various regression models, using processor frequency, processor power-cap level, and job size as control parameters. We also developed a simulator for job scheduling and resource provisioning to study the effect of demand response. The simulator is built upon a parallel discrete-event simulation engine capable of handling large-scale models of HPC architectures and applications [3].

We next developed an economic demand response participation model for HPC systems to allow both HPC operators and HPC users to jointly reduce the energy cost [4]. The challenge is to properly incentivize the HPC users to participate in the demand response program. Since participation in the demand response program can potentially prolong the application's execution, a proper rewarding mechanism is necessary to enable willing participation. Such rewarding mechanism should reward users equivalent to their contributions—an HPC user with more contributions toward demand response should receive more rewards compared to those with less contributions. To achieve this, we resort to the contract theory from the field of microeconomics. According to contract theory, an employer (in this case, the HPC operator) offers contracts to the employees (the HPC users) based on the employees preference and yet under information asymmetry (e.g., the HPC operator may not have the complete information of the HPC users willingness to participate). We proposed an economic demand-response model, where the HPC operator offers a set of contracts to the HPC users based on a design to encourage the willing participation of the users in demand response. HPC users then can voluntarily choose either none or one of the contracts to accept to run the jobs. A contract is similar to a service agreement with incentives under which the user's application may run under reduced capacity to allow energy savings.

In our studies, we use real-life measurements and trace data in simulation to show the effectiveness of the proposed methods. We show that the demand response models can benefit all involved participant. HPC systems can reduce the energy cost in operation. HPC users can earn rewards from participation. Energy service provider can achieve overall energy reduction and power grid stability during emergency events.

References

[1] K. Ahmed, J. Liu, and X. Wu, "An energy efficient demand-response model for high performance computing systems," in *Proceedings of the 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2017)*, 2017.

[2] K. Ahmed, J. Liu, and K. Yoshii, "Enabling demand response for hpc systems through power capping and node scaling," in *Proceedings of the 20th IEEE International Conference on High Performance Computing and Communications (HPCC 2018)*, 2018.

[3] N. Santhi, S. Eidenzenz, and J. Liu, "The Simian concept: parallel discrete event simulation with interpreted languages," in *Proceedings of the 2015 Winter Simulation Conference (WSC)*, 2015.

[4] K. Ahmed, J. Bull, and J. Liu, "Contract-based demand response model for hpc systems," Submitted for publication.