Position Paper: Cross-Layer Large-Scale Control of Warehouse Scale Computing Systems

Arun Ravindran and Bharat Joshi

Department of Electrical and Computer Engineering
University of North Carolina at Charlotte
{*aravindr, bsjoshi*}*@uncc.edu*

Dinesh Mehta and Tyrone Vincent

Department of Electrical Engineering and Computer Science
Colorado School of Mines
{*dmehta, tvincent*}*@mines.edu*

October 30, 2014

Research Summary

The goal of our research is to develop a cross-layer architecture for energy and performance management of tens of thousands of computing nodes in a warehouse-scale computing system (WSC). We hypothesize that a cost-effective power-performance optimal operation of WSCs is possible if they are treated as large-scale dynamical systems subjected to unexpected disturbances including variations in the workload, cost of energy, and hardware failures. Such large-scale dynamical systems in other engineering fields such as process engineering, aerospace, and electric power systems are successfully managed using the rich theory of complex systems and control. Similar to these systems, WSCs are strongly interconnected with interacting subsystems exchanging information and energy with the environment.

Although some work has been done in applying control theory to computing, existing work is mostly focused on single servers. Further, while we recognize the importance of abstraction as a means of building complex systems, we advocate the cross-layer bidirectional flow of information regarding

the dynamic state of the system. Each layer would have a set of parameters tunable at run-time that can be used to guide its dynamical behavior. Additionally, each state will have well defined interfaces for sending and receiving state information with its neighbors. Key challenges include how best to realize a scalable resource allocation and control framework that combines open loop optimization and closed loop control, and to provide experimental demonstration for big data applications on a heterogeneous cluster with multiple parallel programming frameworks.

Experimental needs

The control framework requires stochastic performance models of the WSC hardware, system software and application layers. We plan on using a machine learning based approach to map hardware counter measurements to performance and energy consumption metrics. Note that such measurements, required to obtain the model training data, needs to be carried on a subset generated by sampling tuning parameters space for multiple benchmarks. Examples of tunable parameters include: core idling, DVFS, multiple disk speeds, memory bank idling, all at the hardware layer; DFS block size, sort buffer size, replication factor, all at the system software layer; task parallelism, QoS, deadline miss tolerance, algorithm choice, all at the application software layer. Virtualization and lack of direct access to hardware on public cloud services, such as Amazon Web Services, precludes such measurements. In our research labs, we have deployed a small scale cluster with specialized hardware for detailed subsystem level power measurements. However, a small scale experimental set up does not allow us to perform scaling studies (both in the number of nodes and problem size). We feel that lack of access to massive clusters is a serious impediment to research on warehouse scale computing in the academia.

The experimental facilities to be developed under the proposed NSF Cloud program can potentially meet our research needs. While we only need access to the hardware for short periods of time, we will need root access to all or part of the cluster to perform the above described measurements. A key challenge for the NSF Cloud program is how such access could be provided without adversely affecting other users.