Research Statement:
Increasing Performance and Power Efficiency Through Smart Programming Languages

Bilge Acun

As the size of the data centers and High Performance Computing (HPC) centers keeps growing, their power consumption grows as well. Current systems, such as Blue Waters supercomputer at UIUC, consume tens of megawatts of power leading to millions of dollars in energy bills, significant power strains on local, state, or regional energy grid systems, and the environmental impact on natural resources that provide the power for the supercomputer. As the scale grows, extrapolating the current trends, each data-center may need their own nuclear power plants for their energy needs in near future. Therefore, power and energy efficient system design, and energy efficient computing has become an important research challenge. Not only the hardware, but also the software can be improved to increase the power and energy efficiency of these systems. My research develops smart, dynamic software techniques such as load balancing and machine learning to improve the performance, power, and energy consumption of HPC data centers. The techniques I developed have been published in top journal and conferences including a cover-featured article in the prestigious *IEEE Computer Special Issue on Energy Efficienct Computing* [1] and articles in *International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing-SC)* [2, 3], *International Conference on Supercomputing (ICS)* [4], *IEEE International Conference on Cluster Computing (CLUSTER)* [5] and others [6, 7]. In addition to my articles that total 95 citations, I am also the the lead inventor of two patent-pending technologies that make large-scale data centers more energy-efficient (*currently under filing process by IBM*). My research have used thousands of processors of top supercomputers in the world including Edison and Cori at National Energy Research Scientific Computing Center (NERSC), Stampede at Texas Advanced Computing Center (TACC), Cab at Lawrence Livermore National Laboratory (LLNL), Blue Waters at University of Illinois at Urbana-Champaign (UIUC), Minsky at IBM T.J. Watson Research Center.

Data center and supercomputer system power costs has two major parts: machine power and cooling costs. My research helps reduce both of the costs using smart programming languages, also called dynamic, adaptive runtime systems.

Machine power efficiency: High Performance Computing systems have thousands of processors. The design and manufacture of processors in these system causes inherent variation in supercomputer architectures such as variation in power and temperature of the chips. Such variation also manifests itself as frequency differences among processors and can lead to unpredictable and sub-optimal performance in tightly coupled HPC applications. In my paper published at the 2016 *International Conference on Supercomputing (ICS'16)*, a top highly competitive conference in this field, we analyze performance, power, and temperature variation among the processors of four different supercomputers; Edison, Cab, Stampede, and Blue Waters using up to 16,384 processing cores [4]. We analyze measurements from temperature and power instrumentation and find that intrinsic differences (the differences that occur in the chip manufacturing process) in the chips' power efficiency is the culprit behind the frequency variation. We propose a novel programming language, Charm++ [2], which uses dynamic load balancing technique to cope with this variation and to increase the performance efficiency of the HPC applications up to 16% [4, 7]. My thesis work is focusing on how to increase not only performance but also the power, or energy efficiency in such variable enviroments.

Cooling power efficiency: My work also addresses the high cooling power costs of supercomputers using machine learning approaches. The processors in supercomputers have varying temperatures. Some processors can have higher temperatures than normal (or the average). These processors are called *hot-spots*. The *hot-spots* can be caused by the intensive computation workloads running on these processors as well as the uneven cooling infrastructure. The processors that are closer to the cooling source tend to have lower temperatures than others. Hot-spots cause the cooling system to work inefficiently and increase the power consumption of not just the processor and but also the cooling system as well. In my paper [5], we propose a thermal-aware programming language approach to control the temperatures of the processors automatically and to avoid the hot-spots in HPC data centers. This control is done by Dynamic Voltage and Frequency Scaling (DVFS) of the processor. By avoiding the hot-spots with this technique, the cooling costs can be reduced by 12% with minimal timing penalty. This novel technique implemented in Charm++ runtime system, does these controls transparently from the application so it does not require any additional effort from the programmer.

An accurate temperature prediction model is necessary to understand the temperature variations in large scale and to further mitigate them. For this purpose, I implemented a neural network-based modeling approach for predicting core temperatures of different workloads, under different core frequencies, fan speed levels, and ambient temperature [3, 8]. The model provides guidance for cooling control (i.e. fan speed control) as well as cooling-aware algorithms such as frequency control algorithms and thermal-aware load balancing. The proactive fan control mechanism that I propose can reduce the maximum cooling power by 53% on average by predicting core temperatures. Moreover, through decoupled fan control methods and thermal-aware load balancing algorithms, I show that temperature variations in large scale platforms can be reduced from 25 C to 2 C, making cooling systems more efficiencycient with minimal performance overhead [3, 8].

To summarize, my research shows how programming languages can be smart and to be used to increase the performance and power efficiency of the supercomputers. I use many novel software techniques such as load balancing and machine learning methods to enable this with minimal timing penalty and demonstrate them with Charm++. These techniques are published in top conferences in the HPC field. I believe future programming languages will be power and thermal aware. Using my software skills and computer science background to help overcome the global energy consumption challenge motivated me to pursue my research in this field.

Publications

[1] **Bilge Acun**, Akhil Langer, Harshitha Menon, Osman Sarood, Ehsan Toton, and Laxmikant V. Kalé. "Power, Reliability, Performance: One System to Rule Them All". In: *IEEE Computer, Energy Efficient Computing Special Issue*. 2016.

[2] **Bilge Acun**, Abhishek Gupta, Nikhil Jain, Akhil Langer, Harshitha Menon, Eric Mikida, Xiang Ni, Michael Robson, Yanhua Sun, Ehsan Totoni, Lukasz Wesolowski, and Laxmikant Kalé. "Parallel Programming with Migratable Objects: Charm++ in Practice". In: *High Performance Computing, Networking, Storage and Analysis, SC14: International Conference for (SC)*. 2014.

[3] **Bilge Acun**, Eun Kyung Lee, Yoonho Park, and Laxmikant V. Kalé. "Neural Network-Based Task Scheduling with Preemptive Fan Control". In: *International Workshop on Energy Efficient Supercomputing (E2SC)*. ACM, 2016.

[4] **Bilge Acun**, Phil Miller, and Laxmikant V. Kalé. "Variation Among Processors Under Turbo Boost in HPC Systems". In: *International Conference on Supercomputing (ICS)*. 2016.

[5] Harshitha Menon, **Bilge Acun**, Simon Garcia De Gonzalo, Osman Sarood, and Laxmikant Kalé. “Thermal Aware Automated Load Balancing for HPC Applications”. In: *Cluster Computing (CLUSTER), 2013 IEEE International Conference on*. IEEE. 2013, pp. 1–8.

[6] Abhishek Gupta, **Bilge Acun**, Osman Sarood, and Laxmikant V. Kalé. “Towards Realizing the Potential of Malleable Parallel Jobs”. In: *Proceedings of the IEEE International Conference on High Performance Computing (HiPC)*. 2014.

[7] **Bilge Acun** and Laxmikant V. Kalé. “Mitigating Processor Variation Through Dynamic Load Balancing”. In: *IEEE International Workshop on Variability in Parallel and Distributed Systems (VarSys, IPDPS)*. IEEE, 2016.

[8] **Bilge Acun**, Eun Kyung Lee, Yoonho Park, and Laxmikant V. Kalé. “Support for Proactive Cooling Decisions with Neural Network-Based Temperature Prediction”. In: *High Performance Computing, Networking, Storage and Analysis, SC14: International Conference for (SC)*. ACM, 2017 (In Submission).