John Kloosterman — Research Statement

jkloosterman.net
jklooste@umich.edu

As machine learning analyzes larger data sets and phones process higher-resolution images and videos, applications are demanding more computational power. At the same time, this computation is being done in energy-constrained environments such as battery-powered mobile devices or data centers where cost is driven by electricity consumption and cooling. As a result, computer systems must become both faster and more energy efficient, requiring a change from past methods which focused on performance, often at the expense of efficiency. My research creates the designs and tools needed to reach the difficult goal of achieving performance and efficiency at the same time.

My past research in computer architecture has improved the efficiency of graphics processing units (GPUs). GPUs are faster and more energy-efficient than traditional CPUs but suitable for a smaller set of workloads. My future work, in the setting of my job as lecturer, will continue the focus on performance and efficiency with work involving undergraduate researchers. One aspect of that work will be finding ways to better utilize the powerful GPUs found in recent smartphones, making a new class of computationally intensive mobile applications viable. Another direction of my research, in computer science education, will quantify the effectiveness of my teaching and the impact of curriculum changes. Involving students in my research and collecting data on my teaching will allow me to integrate my research and teaching as an educator and scholar.

Past Work: Energy-Efficient Computer Architecture

My past work in energy-efficient computing has focused on GPU architecture. GPUs can perform many computationally intensive tasks like neural network training, image processing, and large-scale simulations more efficiently than CPUs. This is possible because they are specialized for applications that consist of many identical, independent tasks. My research leverages patterns that emerge as GPUs execute these tasks to unlock additional energy efficiency.

For instance, my WarpPool work, published in MICRO 2015, found that because the tasks running on the GPU are identical, they often access memory in regular ways. Current GPU architectures can use this regularity to merge memory requests together when tasks adjacent in hardware access similar memory locations. My work identified the same locality patterns between tasks father away in the GPU and proposed hardware that could merge these requests as well. This increased energy efficiency as fewer memory requests need to be issued, and increased performance as the additional merges resolved bottlenecks in the memory system.

My later RegLess work, published in MICRO 2017, increased energy efficiency by reducing the size of GPU register files. Because so many tasks execute simultaneously on a GPU, the register file must be large to maintain the execution state of all those tasks, making it power-hungry. To shrink the register file, RegLess stores registers in main memory when possible, moving them into the register file just before they will be accessed. Since the bandwidth of the memory system is smaller than that of the register file, hardware must precisely manage which registers to transfer and when in order to avoid performance loss. Using this system, I reduced the register file to 25% of its size without affecting the average performance of a common GPU benchmark suite.

Future Direction: Mobile GPU Applications

My future work will continue in the same direction as my architecture research, discovering ways to increase both energy efficiency and performance at the same time. Mobile devices like smartphones are one domain where this is crucial. Mobile devices are built around a system-on-chip that integrates the main CPU with other components, including a GPU. Current mobile GPUs are very powerful, but have only recently started to be used for non-graphics workloads like desktop and server GPUs. Making fuller use of the efficient GPU will enable new kinds of applications to run on mobile devices by unlocking additional performance.

Mobile GPU applications face two obstacles that current server and desktop applications do not. The first is fragmentation – there are many different mobile GPU manufacturers which each use very different architectures in their devices, as opposed to the smaller number of designs in desktops and servers. This makes it difficult to write GPU kernels that achieve high performance on every device. The second challenge is interactivity – mobile applications must do their computation while the user has an application open, which means that overloading the GPU causes visual artifacts for users.

My research will enable new classes of applications on mobile devices that leverage the GPU while addressing these challenges. One example project for an undergraduate researcher would be implementing a feature to search on a phone for photos that contain a given person's face. Currently, this feature can only be implemented on servers, which raises privacy concerns as the photos need to be sent to a third party. My research contributions would bridge the gaps needed to bring the same features to the more constrained mobile environment, making headway on the fundamental fragmentation and interactivity challenges. To address fragmentation, the algorithms used would be selected to perform well on a large percentage of mobile GPUs. To address interactivity, the amount of work done per photo would need to be small.

Future Direction: Computer Science Education

Alongside this research in efficient systems, I will also pursue a research program in computer science education. My past involvement in education research included a collaboration with a medical student introducing a video game as an instructional tool. Medical students learn to construct differential diagnoses, which are rankings of the most likely conditions given a patient's symptoms. In the video game, students developed these skills by interacting with a virtual patient. They could ask the patient questions, perform labs, and request imaging studies, with the game rewarding them for gathering the data useful for their differential diagnosis. The game was used in an obstetrics and gynecology class at the University of Michigan Medical School, where students indicated in a survey that the game complemented the in-class discussions and found it valuable as a study tool.

Going forward, I will build computer science education research into my courses as a method to continuously improve them. I learned at the University of Michigan to give students a course entry and exit survey for a small amount of credit. By asking the same questions at the beginning and end of the semester, it is possible to track outcomes that are not reflected in grades, such as whether students feel they could succeed in a career in computer science. Comparing data across semesters makes it possible to track the effects of changes in course methods and material. For instance, if I were introducing a new programming project into a course, I would use this data to quantify its impact on student learning. Projects should motivate students, should show how their computer science skills can be used outside of the classroom, and should avoid frustrating students with setup steps and difficult debugging. Collecting data about student success and feelings would allow me to judge whether the new project met these goals. As I gather data and experience about teaching methods and effective assignments, I plan to share this knowledge by publishing at venues such as SIGCSE.

Conclusion

My research program will extend my past work in computer architecture towards a broader vision of efficient, high-performance systems. Undergraduate research will be a key part of this project, beginning with work on better utilizing the powerful GPUs in mobile phones. In addition, my computer science education research will allow me to assess the changes I make to my courses. These projects in both systems and education will broaden my impact as a scholar and teacher.