Hardware and Software Techniques to Scale the Memory Wall

With Moore's Law coming to an end, architects must find ways to sustain performance growth without technology scaling. The most promising path is to either increase the number of cores or to specialize processors to important computations. But neither approach will be successful unless architects drastically reduce the time and energy processors spend accessing data. Even on current systems, processors often spend the majority of their time and energy moving data, not processing it. For example, a single main memory access takes hundreds of cycles and costs the energy of a thousand floating-point operations. What's worse, these costs are increasing as processors scale to larger core counts.

I will present new architectural techniques that drastically reduce data movement. In the first part of the talk, I will describe how to place data in distributed on-chip cache banks to minimize both cache misses and on-chip network traffic. To tackle this NP-hard problem, I have developed simple hardware mechanisms and efficient placement algorithms that perform within 1% of idealized solutions. In the second part of the talk, I will describe how to make better use of scarce cache capacity through analytical cache replacement policies. The key challenge faced by practical replacement policies is the inherent uncertainty of memory references. I will show how to tackle this uncertainty using Markov decision processes and present an efficient hardware implementation.

These techniques yield significant performance improvements and energy savings that increase with system size. Unlike many prior techniques that rely on best-effort heuristics, my work follows an analytical design approach that yields well-behaved systems that are easy to reason about and simple to manage.

Bio:
Nathan Beckmann earned his PhD from MIT in Fall 2015 under the supervision of Prof. Daniel Sanchez, where he continues as a postdoctoral researcher. Nathan's work aims to bridge theory and practice by building practical systems with a firm analytical foundation. His PhD dissertation received the George M. Sprowls Award for an outstanding thesis in CS at MIT. He received his SM from MIT, which won the William A Martin award, and his BS from UCLA.

Thursday, March 17
10:00 a.m. GHC 6115
Host: Todd Mowry