Nineteenth Conference for African American Researchers in the Mathematical Sciences
CAARMS19 - University of California, San Diego, La Jolla CA, July 24-27, 2013

Scaling Computation to Petabytes of Data and Beyond

James Mickens
Microsoft Research
mickens@microsoft.com

Computer programs are data-driven—a program consumes data, manipulates that data in some way, and then generates output data. In various scientific disciplines, the size of this input data and output data is growing at a rapid rate. For example, biologists are now analyzing genomes which contain terabytes of information; speech recognition software develops phonetic models by extracting common structures from vast corpi of preexisting sound recordings. In these situations, the amount of data to analyze is far too vast to be stored on a single machine, or even a hundred machines. Instead, the information must be stored in a datacenter, which is a large building containing tens of thousands (or hundreds of thousands) of machines. These machines collaborate to solve large computational problems that would be intractable for a single computer to solve.

In this talk, I describe how to design a fast, datacenter-scale computation platform for handling massive amounts of information. I explain how optimal performance is achieved through careful co-design of hardware and software. By designing the hardware and the software at the same time, we ensure that 1) all hardware components can communicate with each other at maximum speeds, and 2) any piece of software can communicate with any piece of hardware at that hardware's maximum speed. This results in a storage system that can execute large-scale computations which were previously infeasible.

Biography

James Mickens is a researcher in the Distributed Systems group at Microsoft's Redmond lab. His current research focuses on datacenter storage systems for large-scale computation. Mickens investigates ways to make these systems faster and more robust to hardware failure. He also works on client-side web applications, designing new ways to diagnose and fix bugs without requiring assistance from end-users. Mickens received a bachelor's degree in computer science from Georgia Tech, and a PhD in computer science from the University of Michigan. During his stay at Michigan, he was notorious for scheduling his thesis defense in the early hours of the morning so that nobody would attend it. This technique, popularly called "The Mickens Gambit," is now banned in 36 states.