I2PC

Illinois-
Intel
Parallelism
Center

Distinguished Speaker Series

Thomas Wenisch

University of Michigan

Efficiency Challenges in Warehouse-Scale Computers

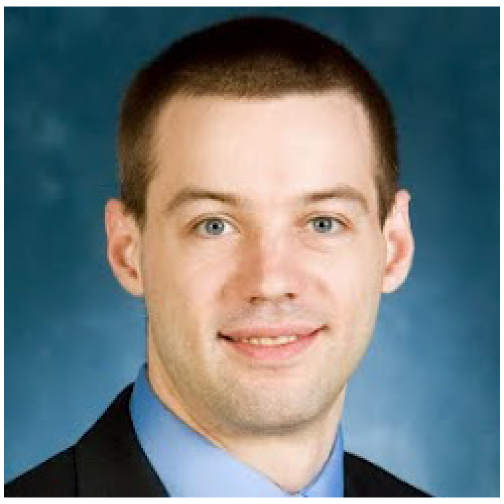
Wednesday, March 28, 2012

4 - 5 PM CST

3405 Siebel Center

For more information about I2PC research, education, and activities, visit:

i2pc.cs.illinois.edu



Abstract

Architects and circuit designers have made enormous strides in managing the energy efficiency and peak power demands of processors and other silicon systems. Sophisticated power management features and modes are now myriad across system components, from DRAM to processors to disks. And yet, despite these advances, typical data centers today suffer embarrassing energy inefficiencies: it is not unusual for less than 20% of a data center's multi-megawatt total power draw to flow to computer systems actively performing useful work. Managing power and energy is challenging because individual systems and entire facilities are conservatively provisioned for rare utilization peaks, which leads to energy waste in underutilized systems and over-provisioning of physical infrastructure. Power management is particularly challenging for Online Data Intensive (OLDI) services---workloads like social networking, web search, ad serving, and machine translation that perform significant computing over massive data sets for each user request but require responsiveness in sub-second time scales. These inefficiencies lead to worldwide energy waste measured in billions of dollars and tens of millions of metric tons of CO2.

In this talk, I discuss what, if anything, can be done to make OLDI systems more energy-proportional. Specifically, through a case study of Google's Web Search application, I will discuss the applicability of existing and proposed active and idle low-power modes to reduce the power consumed by the primary server components (processor, memory, and disk), while maintaining tight response time constraints, particularly on 95th-percentile latency. Then, I will briefly discuss our work on Power-Routing, a proposal to dynamically switch servers among redundant power feeds to reduce overprovisioning in data center power delivery infrastructure. Finally, I will close with brief comments on our new and ongoing work in power, performance, and thermal management at the other end of the computing spectrum, namely Smart Phone devices.

Bio

Thomas Wenisch is the Morris Wellman Faculty Development Assistant Professor of Computer Science and Engineering at the University of Michigan, specializing in computer architecture. Tom's prior research includes memory streaming for commercial server applications, store-wait-free multiprocessor memory systems, memory disaggregation, and rigorous sampling-based performance evaluation methodologies. His ongoing work focuses on data center architecture, energy-efficient server design, smartphone architecture, and multi-core / multiprocessor memory systems. Tom received an NSF CAREER award in 2009, two papers selected in IEEE Micro Top Picks, and a Best Paper Award at HPCA 2012. Prior to his academic career, Tom was a software developer at American Power Conversion, where he worked on data center thermal topology estimation. He is co-inventor on six patents. Tom received his Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN