

**CS COLLOQUIUM SERIES - FALL 2006 (CS 688)
DEPARTMENT OF COMPUTER SCIENCE**

**Friday, September 22, noon to 1 pm
ENGB N-25**

ALL WELCOME !

**Resource Amplification:
Efficient Performance Scaling for Future Superscalars**

**Amir Roth
Computer and Information Sciences Department
University of Pennsylvania**

Continued performance improvement is the engine of the computer industry. For years, this engine has been fueled primarily by increases in clock frequency based on a combination of transistor miniaturization and deeper pipelines. But this fuel source is running out. Future improvements, if they come at all, will have to come from increased parallelism. Certainly explicit thread-level parallelism will be a big part of the solution and explicit data-level parallelism will pitch in where possible. But implicit fine-grain parallelism (i.e., good-old-fashion ILP) will contribute as well. Many applications and application regions will not be explicitly parallelized, and even the performance of parallel code is the sum of the performance of sequential threads.

This talk will describe a set of techniques we have developed for improved single-thread performance and performance efficiency. All of these techniques operate on the same basic principle: resource amplification. They increase the effective capacity and bandwidth (but not the physical capacity and bandwidth) of key pipeline structures like the register file and bypass network, issue queue, data cache and load/store queue, and even in-order structures like the instruction cache and register renaming logic. The different techniques achieve amplification in different ways. SVW and NoSQ amplify the data cache and load store queue by exploiting memory dependence prediction and filtered-replay verification, RENO uses dynamic instruction optimization to amplify the register file and issue queue, mini-graphs amplify all structures using aggressive instruction aggregation. Together these techniques allow a superscalar with bandwidth X and capacity Y to effectively perform like a superscalar with bandwidth $1.4X$ and capacity $1.4Y$. This amplification can be used for additional performance, for lower energy, or for a combination of both. This is joint work with Vlad Petric, Anne Bracy, Tingting Sha, Andrew Hilton and Milo Martin.

About the speaker:

Amir Roth is an assistant professor at the Computer and Information Sciences Department at the University of Pennsylvania, where he co-leads the architecture and compilers group (ACG). His current research focuses on scalable microarchitectures and memory systems for multi-core processors. His current projects are dataflow-minigraphs (with Anne Bracy), RENO (with Vlad Petric), and ON-Core (with Tingting Sha, Andrew Hilton and Prof. Milo Martin).

Prof. Roth received a BS in physics from Yale University in 1994 and a PhD in computer science from the University of Wisconsin-Madison in 2001. He received the NSF CAREER award in 2002.