

Thursday, September 10Scaife Hall Auditorium
Room 1254:30 p.m.
Refreshments at 4:00 p.m.**José Martínez**

Associate Professor

Electrical & Computer Engineering Department
Cornell University
Ithaca, NY

José Martínez (Ph.D. '02 UIUC) is associate professor of electrical and computer engineering at Cornell. His research work in computer architecture has earned several awards; among them: two IEEE Micro Top Picks in Computer Architecture papers; a HPCA Best Paper Award; a NSF CAREER Award; and two IBM Faculty Awards. He has been recognized with a Kenneth A. Goldman '71 Excellence in Teaching Award, and as a Merrill Presidential Teacher. He is a member of the Computer Systems Laboratory and the Intelligent Information Systems Institute at Cornell, as well as the ACM, the IEEE, and the SHPE societies. José Martínez is currently on sabbatical leave at Carnegie Mellon.

Coordinated Management of Multiple Interacting Resources in Chip Multiprocessors: A Machine Learning Approach

Efficient sharing of system resources is critical to obtaining high utilization and enforcing system-level performance objectives on chip multiprocessors (CMPs). Although several proposals that address the management of a single microarchitectural resource have been published in the literature, coordinated management of multiple interacting resources on CMPs is a much harder problem.

We propose a framework that manages multiple shared CMP resources in a coordinated fashion to enforce higher-level performance objectives. We formulate global resource allocation as a machine learning problem. At runtime, our resource management scheme monitors the execution of each application, and learns a predictive model of system performance as a function of allocation decisions. By learning each application's performance response to different resource distributions, our approach makes it possible to anticipate the system-level performance impact of allocation decisions at runtime with little runtime overhead. As a result, it becomes possible to make reliable comparisons among different points in a vast and dynamically changing allocation space, allowing us to adapt our allocation decisions as applications undergo phase changes.

Our evaluation concludes that a coordinated approach to managing multiple interacting resources is key to delivering high performance in multiprogrammed workloads, but this is possible only if accompanied by efficient search mechanisms. We also show that it is possible to build a single mechanism that consistently delivers high performance under various important performance metrics.

ECE Seminar Hosts

| | |
|--------------------|----------------------|
| Jeyanandh Paramesh | paramesh@ece.cmu.edu |
| Onur Mutlu | onur@cmu.edu |
| Gabriela Hug | ghug@ece.cmu.edu |