Leveraging Network Technology for Power-Efficient Computer Architecture

Jiang Hu

Professor
Dept. of Electrical Engineering
Texas A&M University

Thursday, September 24th 4:30 PM Scaife Hall 125

Abstract:
The fundamental limit of chip power density has been and continues to be a grand challenge to the silicon-based computing technology. This talk presents a new approach to processor power management that leverages network technology. In modern chip multiprocessors, resources shared by processor cores, such as on-chip communication and shared memory, play an increasingly critical role in determining the overall performance. Our key observation is that per-core dynamic voltage/frequency scaling (DVFS) can be used as a client regulation mechanism for Quality-of-Service (QoS) of the shared resources. As a result, power savings are feasible with little performance loss or even performance gain due to the better allocation of shared resources. Based on this observation, we propose a new DVFS technique based on TCP Vegas, a congestion control protocol from the networking domain. We also propose an uncore (communication + last level cache) DVFS technique that is applied in conjunction with the QoS-driven core DVFS. For single-application cases, our techniques achieve 53% energy savings with less than 1% performance degradation on average. Some applications actually show a small performance improvement while still saving significant energy, indicating the power of utilizing DVFS as a means of QoS. We demonstrate that these results also hold for multi-application workloads.

Bio:
Jiang Hu received the B. S. degree in optical engineering from Zhejiang University, China, in 1990, the M. S. degree in physics in 1997, and the Ph. D. degree in electrical engineering from the University of Minnesota in 2001. He was with IBM Microelectronics from January 2001 to June 2002.
Currently, he is a professor in the Department of Electrical and Computer Engineering at the Texas A&M University. His research interests include large scale optimization, low power system design, design for robustness/reliability, hardware acceleration for machine learning and hardware security.