Eiman Ebrahimi

Phone No.: 512-293-6198

eiman.ebrahimi@gmail.com
hps.ece.utexas.edu/people/ebrahimi/

MAIN INTERESTS

My background is primarily in computer system architecture, especially in the cooperation of hardware with system-software, the runtime system, and the compiler to improve performance, energy-efficiency, and quality of service.

More recently I have become interested and involved in the design and performance analysis of platforms targeted at deep learning.

EDUCATION

The University of Texas at Austin — PhD., Dec 2011
Supervisor: Yale N. Patt

The University of Texas at Austin — MSc., May 2007
GPA: 3.97 / 4.0

University of Tehran, BSc., May 2005
GPA: 16.9 / 20

PUBLICATIONS

Ugljesa Milic, Oreste Villa, Evgeny Bolotin, Akhil Arunkumar, Eiman Ebrahimi, Aamer Jaleel, Alex Ramirez, David Nellans, "**Beyond the socket: NUMA-Aware GPUs,**" *The 50th Annual IEEE/ ACM International Symposium on Microarchitecture (MICRO-50),* Boston, Oct 2017

Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, David Nellans,"**MCM-GPU: Multi-Chip-Module GPUs for Continued Performance,**" T*he 44th International Symposium on Computer Architecture (ISCA-44),* Toronto, June 2017.

Kevin Hsieh, Eiman Ebrahimi, Gwangsun Kim, Niladrish Chatterjee, Mike O'Conner, Nandita Vijaykumar, Onur Mutlu, and Stephen W. Keckler, "**Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems**," *The 43rd International Symposium on Computer Architecture (ISCA-43)*, Seoul, June 2016.

Milad Hashemi, Khubaib, Eiman Ebrahimi, Onur Mutlu, and Yale N. Patt, "**Accelerating Dependent Cache Misses with an Enhanced Memory Controller,**" *The 43rd International Symposium on Computer Architecture (ISCA-43)*, Seoul, June 2016.

Neha Agarwal, David Nellans, Eiman Ebrahimi, Thomas F. Wenisch, John Danskin, and Stephen W. Keckler, "**Selective GPU Caches to Eliminate CPU-GPU Hardware Cache Coherence**," T*he 22nd International Symposium on High-Performance Computer Architecture (HPCA)*, Barcelona, March 2016.

Mark Stephenson, Siva Kumar Sastry Hari, Yunsup Lee, Eiman Ebrahimi, Daniel R. Johnson, David Nellans, Mike O'Connor, and Stephen W. Keckler, "**Flexible Software Profiling of GPU Architectures,"** *The 42nd International Symposium on Computer Architecture (ISCA-42),* Portland, June 2015.

Rustam Miftakhutdinov, Eiman Ebrahimi, Yale N. Patt, **"Predicting Performance Impact of DVFS for Realistic Memory Systems,"** *The 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-45),* Vancouver, December 2012.

Marco A. Z. Alves, Khubaib, Eiman Ebrahimi, Veynu T. Narasiman, Carlos Villavieja, Phillipe O. A. Navaux, and Yale N. Patt, **"Energy Savings via Dead Sub-Block Prediction,"** *The 24th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD),* New York, October 2012.

Eiman Ebrahimi, Chang Joo Lee, Onur Mutlu, and Yale N. Patt, **"Fairness via Source Throttling: A Configurable and High-Performance Fairness Substrate for Multi-Core Memory Systems,"** *ACM Transactions on Computer Systems (TOCS),* April 2012.

Eiman Ebrahimi, Rustam Miftakhutdinov, Chris Fallin, Chang Joo Lee, Jose A. Joao, Onur Mutlu, and Yale N. Patt, **"Parallel Application Memory Scheduling,"** *The 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-44),* Porto Alegre, December 2011.
Acceptance Rate: 21%

Eiman Ebrahimi, Chang Joo Lee, Onur Mutlu, and Yale N. Patt, **"Prefetch-Aware Shared Resource Management for Multi-Core Systems,"** *The 38th International Symposium on Computer Architecture (ISCA-38),* San Jose, June 2011.
Acceptance Rate: 19.2%

Eiman Ebrahimi, Chang Joo Lee, Onur Mutlu, and Yale N. Patt, **"Fairness via Source Throttling: A Configurable and High-Performance Fairness Substrate for Multi-Core Memory Systems,"** *The 15th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-15),* Pittsburgh, March 2010.
Best Paper Award.
Acceptance Rate: 17.7%

Eiman Ebrahimi, Onur Mutlu, Chang Joo Lee, and Yale N. Patt, "**Coordinated Control of Multiple Prefetchers in Multi-Core Systems,"** *The 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-42)*, New York, December 2009.
Acceptance Rate: 24.8%

Eiman Ebrahimi, Onur Mutlu, Yale N. Patt, **"Techniques for Bandwidth-Efficient Prefetching of Linked Data Structures in Hybrid Prefetching Systems,"** *The 15th International Symposium on High-Performance Computer Architecture (HPCA-15)*, Raleigh, February 2009.
One of three papers nominated for Best Paper Award by the Program Committee.
Acceptance Rate: 19%

Hadi Esmailzadeh, Amir Moghimi, Eiman Ebrahimi, Caro Lucas, Zainalabedin Navabi, Sied Mehdi Fakhraie, **"DCim++: a C++ library for object-oriented hardware design and distributed Simulation," IEEE** *International Symposium on Circuits and Systems (ISCAS)*, Island of Kos, Greece, 2006.

Hadi Esmaeilzadeh, Neda Shahidi, Eiman Ebrahimi, Amir Moghimi, Caro Lucas, Zainalabedin Navabi, "**Cim++: A C++ Library for Object Oriented Hardware Design,"** *International Journal of Science and Information Technology (IJIST) Lecture Notes of 1st International Conference on Informatics, vol. 1, no. 2, pp. 35-41*, September 2004.

WORK EXPERIENCE

Senior Research Scientist, NVIDIA, Austin USA Sep 2014 - Present
Main Projects:
Exploring Multi-GPU RNN/LSTM Training: In this project, we explored the performance and algorithmic perspectives of RNN training in multi-GPU systems. There are two main end-to-end training characteristics that we focus on: (1) final accuracy, and (2) accuracy improvement over time. We note that there are three main factors, which determine these two end-to-end training characteristics: (1) utilization (throughput), (2) learning ability, and (3) final accuracy. These factors tightly interplay with each other. We empirically show that focusing on any one factor does not provide a complete picture of end-to-end behavior. Based on our observations we study a hybrid training scheme that dynamically takes advantage of the throughput and learning ability characteristics of different distributed training models.

Exploiting Data Locality in GPUs via Expressive Cross-layer Abstractions and Architectural Optimizations: A number of different forms of data locality exist in CUDA programs that are not directly expressed or exploited. Our experiments show that leveraging this data locality to get performance is a challenging task, requiring coordination of multiple architectural techniques such as CTA scheduling, prefetching, caching prioritization and bypassing, etc. While software and architectural optimizations exist to help exploit this data locality, they are unable to fully leverage this data locality because: a) coordinating multiple architectural techniques in hardware is either not possible or a very challenging task and b) programmer-transparent architectural techniques simply miss the key semantics behind the data locality required to effectively exploit locality. We can do better if programmer intent about data locality were conveyed to the underlying architecture. We propose a flexible cross-layer abstraction, which we refer to as the "Locality Descriptor", to more explicitly express data locality in CUDA programs, and allow hardware to use this information to effectively exploit the benefits of data locality.

GPU Memory System Bottlenecks of RNN/LSTM Training and Inference: With the large improvements in compute speeds on modern GPUs, memory system bottlenecks become more relevant in workloads that use highly optimized compute kernels. RNN/LSTM training and inference is an important example. We study what parts of the memory system become the main bottleneck, and explore what combination of software optimizations and architectural mechanisms can resolve these bottlenecks most effectively for these networks.

Multi-Chip-Module GPUs for Continued Performance Scalability: demonstrated that package-level integration of multiple GPU modules to build larger logical GPUs can enable continuous performance scaling beyond Moore's law. Evaluated the feasibility of a basic Multi-Chip-Module GPU (MCM-GPU) design and designed three architectural optimizations that significantly improve data locality for the GPU modules and minimize the negative aspects of limited bandwidth between them.

Selective GPU Caches to Eliminate CPU-GPU Coherence: designed several architectural improvements to offset performance penalty of selective caching. These optimizations bring a selective caching GPU to within 4% of a fully cache coherent implementation without the need to integrate CPUs and GPUs under a single coherence protocol.

Transparent Offloading and Mapping (TOM) - Enabling Programmer-Transparent Near-Data Processing in GPU Systems: designed an offload candidate selection mechanism that can be implemented as a static compiler pass that balances the bandwidth costs and benefits of offloading. Also designed a reactive data mapping mechanism that remaps pages across the stacks to increase the stack-level locality of the offloaded computation and its data.

Computer Architect, NVIDIA, Austin USA - Apr 2012 - Sep 2014
Main roles: High-performance and Quality-of-Service aware memory controllers for Tegra SoCs.

I contributed to the development and correlation of a highly accurate memory controller/memory system simulator for NVIDIA's Tegra SoCs.
Owned the design and implementation of the API used by multiple teams (software and SoC design teams), that sets up the Tegra memory controller for quality-of-service guarantees and performance.
Worked on NVIDIA discrete GPU memory controllers at NVIDIA. My contribution there was the evaluation and design of address mapping policies across the entire memory system. Address mapping is a heavily optimized part of NVIDIAs designs targeted at distributing memory requests between multiple similar resources at each level of the memory system hierarchy.

Postdoctoral Researcher, The University of Texas at Austin --- Dec 2011 - Apr 2012
Researching management of shared resources for high performance, quality of service, and energy efficiency in big-data systems.
I was also involved in mentoring a couple of PhD students, helping them track and make continuous progress on milestones towards their thesis.

Research Intern, IBM Austin Research Lab, Austin USA — May 2010 - Aug 2010
Researching high-performance and fair memory subsystems for multi-core systems.

Research Intern, Microsoft Research, Redmond USA — May 2008 - Aug 2008
Researching techniques for bandwidth-efficient prefetching of linked data structures in hybrid prefetching systems. The results of this research were later published in the HPCA 2009 paper listed in the Publications section below.

Senior Design Engineer, PA Semi. Santa Clara USA— May 2007 - Aug 2007
Performance evaluation/analysis, identifying performance bottlenecks.

TALKS

- International Conference on High-Performance Computer Architecture, 2016. Title: Selective GPU Caches to Eliminate CPU-GPU Coherence.
- ECE NYU Faculty Candidate Talk, 2014. Title: Rethinking Computer Architectures in the Age of Big-Data.
- ECE Illinois Faculty Candidate Talk, 2012. Title: Fair and High Performance Shared Resource Memory Resource Management for Chip Multiprocessors.
- International Symposium on Computer Architecture, 2011. Title: Prefetch-Aware Shared-Resource Management for Multi-Core Systems.
- IBM Research, Austin, 2011. Title: Managing memory system inter-thread interference.
- International Conference on Architectural Support for Programming Languages and Operating Systems, 2010. Title: Fairness via Source Throttling: A Configurable and High-Performance Fairness Substrate for Multi-Core Memory Systems
- International Symposium on Microarchitecture, 2009. Title: Coordinated Control of Multiple Prefetchers in Multi-Core Systems.
- International Conference on High-Performance Computer Architecture, 2009. Title: Techniques for Bandwidth-Efficient Prefetching of Linked Data Structures in Hybrid Prefetching Systems.

TECHNICAL REPORTS

Chang Joo Lee, Eiman Ebrahimi, Veynu Narasiman, Onur Mutlu, and Yale N. Patt, **"DRAM-Aware Last-Level Cache Replacement,"** *HPS Technical Report,* TR-HPS-2010-007, December 2010.

Eiman Ebrahimi, Chang Joo Lee, Onur Mutlu, and Yale N. Patt, **"Prefetch-Aware Shared Resource Management for Multi-Core Systems,"** *HPS Technical Report,* TR-HPS-2010-005, December 2010.

Chang Joo Lee, Veynu Narasiman, Eiman Ebrahimi, Onur Mutlu, and Yale N. Patt, **"DRAM-Aware Last-Level Cache Writeback: Reducing Write-Caused Interference in Memory Systems,"** *HPS Technical Report,* TR-HPS-2010-002, April 2010.

Eiman Ebrahimi, Onur Mutlu, Yale N. Patt, **"Techniques for Bandwidth-Efficient Prefetching of Linked Data Structures in Hybrid Prefetching Systems,"** *HPS Technical Report*, TR-HPS-2008-004, October 2008.

ACADEMIC RESEARCH EXPERIENCE

Research Assistant, The University of Texas at Austin, HPS Research Group — Sep 2007 - Dec 2011
Investigating high-performance and fair memory system designs for CMP systems.
High-performance and bandwidth efficient prefetching for CMP systems.
Bandwidth-efficient prefetching techniques for linked data structures in hybrid prefetching systems.

Research Assistant, The University of Texas at Austin, EDGE Group — May 2006 - Aug 2006
Investigating scratchpad memories for modern dataflow processors.

Research Assistant, University of Tehran, VLSI Group — Apr 2004 - Jul 2005
Design and implementation of a distributed simulation environment for the Cim++ hardware description language.

SKILLS

Programming: C/C++, Java, C#, Verilog, Cuda, Python
Tools: NvProf, vTune, Oprofile, Pin, Cadence (Verilog-XL, Virtuoso, Spectre), Synopsys (VCS/ Virsim, DesignVision)

Employability Status: U.S. Citizen