Yuhao Zhu

The University of Texas at Austin
Electrical and Computer Engineering
yzhu@utexas.edu
http://yuhaozhu.com/

RESEARCH INTERESTS

I am interested in building better hardware and software to make next-generation Client and Cloud computing faster, more energy-efficient, and deliver better Quality of Experience.

In my dissertation work, I am designing a mobile computing substrate to enable future Web applications. It addresses the fundamental trade-off between performance, energy efficiency, and user QoS experience via synergistic cross-layer optimizations across the processor architecture, Web browser runtime, programming language, and application layers.

EDUCATION

The University of Texas at Austin *2010-present*
Ph.D. candidate, Electrical and Computer Engineering
M.S.E.E, 2015
Advisor: Vijay Janapa Reddi

Beihang University, Beijing, China *2006-2010*
B.S., Computer Science and Engineering
Advisor: Yangdong Deng

HONORS & RECOGNITIONS

- Google Ph.D. Fellowship, 2016
- IEEE MICRO Top Picks of Computer Architecture (Honorable Mention), 2016
- Invitation to Google Mobile Faculty Summit, 2015
- Qualcomm Innovation Fellowship Finalist, 2015
- Best of Computer Architecture Letters Award, 2014
- UT Austin Microelectronics and Computer Development Fellowship, 2011-2012
- Outstanding Undergraduate Thesis Award, 2010

PUBLICATIONS

Journal Articles

- Yuhao Zhu, Matthew Halpern, Vijay Janapa Reddi
  The Role of the CPU in Energy-Efficient Mobile Web Browsing
  IEEE MICRO Mobile Special Issue, Jan/Feb 2015, 35(1):26-33

- Yuhao Zhu, Aditya Srikanth, Jingwen Leng, Vijay Janapa Reddi
  Exploiting Webpage Characteristics for Energy-Efficient Mobile Web Browsing
  Computer Architecture Letters (CAL), Oct 2012, 13(1):33-36
  One of Three Papers Awarded Best of CAL for 2014

- Yuhao Zhu, Bo Wang, Yangdong Deng
  Massively Parallel Logic Simulation with GPUs
  ACM Transactions on Design Automation of Electronic Systems (TODAES), June 2011, 16(3):29

Conference Papers

- Yuhao Zhu, Vijay Janapa Reddi
  GreenWeb: Language Extensions for Energy-Efficient Mobile Web Computing
  PLDI 2016

- Matthew Halpern, Yuhao Zhu, Vijay Janapa Reddi
  Mobile CPU's Rise to Power: Quantifying the Impact of Generational Mobile CPU Design Trends on Performance, Energy, and User Satisfaction
  HPCA 2016

- Yuhao Zhu, Daniel Richins, Matthew Halpern, Vijay Janapa Reddi
  Microarchitectural Implications of Event-driven Server-side Web Applications
  MICRO 2015
  IEEE MICRO Top Picks of Computer Architecture (Honorable Mention) in 2016
- Yuhao Zhu, Matthew Halpern, Vijay Janapa Reddi
  Event-based Scheduling for Energy-Efficient QoS (eQoS) in Mobile Web Applications
  HPCA 2015
- Matthew Halpern, Yuhao Zhu, Vijay Janapa Reddi
  Mosaic: Cross-Platform User-Interaction Record and Replay Tool for the Fragmented Android Ecosystem
  ISPASS 2015
- Yuhao Zhu, Vijay Janapa Reddi
  WebCore: Architectural Support for Mobile Web Browsing
  ISCA 2014
- Chen Zhou, Xiaofei Wang, Weichao Xu, Yuhao Zhu, Vijay Janapa Reddi, Chris Kim
  Estimation of Instantaneous Frequency Fluctuation in a Fast DVFS Environment Using an Empirical BTI Stress-Relaxation Model
  IRPS 2014
- Yuhao Zhu, Vijay Janapa Reddi
  High-Performance and Energy-Efficient Mobile Web Browsing on Big/Little Systems
  HPCA 2013
- Yuhao Zhu, Yangdong Deng, Yubei Chen
  Hermes: An Integrated CPU/GPU Microarchitecture for IP Routing
  DAC 2011
- Bo Wang, Yuhao Zhu, Yangdong Deng
  Distributed Time, Conservative Parallel Logic Simulation on GPUs
  DAC 2010

Book Chapters

- Yangdong Deng, Yuhao Zhu, Bo Wang
  Asynchronous Parallel Logic Simulation on Modern Graphics Processors
  GPU Solutions to Multi-scale Problems in Science and Engineering, 2013
- Yangdong Deng, Xiaomemg Jiao, Shuai Mu, Kang Kang, Yuhao Zhu
  NPGPU: Network Processing on Graphics Processing Units
  Theoretical and Mathematical Foundations of Computer Science, 2011

Miscellaneous

- Daniel Richins, Yuhao Zhu, Matthew Halpern, Vijay Janapa Reddi
  Locality Lost: Unlocking the Performance of Event-driven Servers
  Summary of the MICRO 2015 paper
- Yuhao Zhu, Vijay Janapa Reddi
  WebCore: An Architectural Substrate for Enabling High-Performance and Energy-Efficient Mobile Web
  Summary of the ISCA 2014 paper
- Yuhao Zhu, Vijay Janapa Reddi
  Energy-Conscious QoS in the Mobile Web with Hardware Heterogeneity
  Summary of the HPCA 2013 paper
- Yuhao Zhu, Yubei Chen, Yangdong Deng
  A New Heterogeneous Packet Processing Architecture and Its Analytical Performance Model
  Extended version of the DAC 2011 paper

ACADEMIC TALKS

- The Human Processing Unit (HPU) as a New Approximate Computing Substrate
  WAX 2016 co-located with ASPLOS 2016, April 2016, Atlanta, GA
- Scalable End-to-end Quality Control in Approximate Computing
  WAX 2016 co-located with ASPLOS 2016, April 2016, Atlanta, GA
- Microarchitectural Implications of Event-driven Server-side Web Applications
  MICRO 2015, December 2015, Waikiki, HI (lightening version)
- Exploiting Webpage Characteristics for Energy-Efficient Mobile Web Browsing
  Best of CAL 2014 presented at HPCA 2015, February 2015, San Francisco, CA
- Event-Based Scheduling for Energy-Efficient QoS (eQoS) in Mobile Web Applications
  HPCA 2015, February 2015, San Francisco, CA
- WebCore: Architectural Support for Mobile Web Browsing
  ISCA 2014, June 2014, Minneapolis, MN (lightening version)
  Intel invited talk, July 2014, Austin, TX
- High-Performance and EnergyEfficient Mobile Web Browsing on Big/Little Systems
  HPCA 2013, February 2013, Shenzhen, China
  UT Austin PL Lunch, September 2012, Austin, TX
  AMD Research Lab, August 2012, Austin, TX
- Hermes: An Integrated CPU/GPU Microarchitecture for IP Routing
  DAC 2011, June 2011, San Diego, CA

RESEARCH EXPERIENCE

The University of Texas at Austin (*Graduate*)

High-Performance and Energy-Efficient Web Computing *8/2011-present*

Programming Language Support for Energy-Efficient Mobile Web Computing (PLDI 2016)

- Proposed two fundamental programming abstractions that capture critical aspects of user QoS experience in mobile Web
- Designed and implemented GreenWeb, a Web language extension that express the two QoS abstractions and enable runtime energy-efficiency optimization
- Designed AutoGreen, an automatic annotation framework that improves developer productivity by applying GreenWeb annotations without developers intervention.

Server Microarchitecture Implications of Event-driven Web Applications (MICRO 2015)

- Identified single-thread performance as a major bottleneck for event-driven server workloads due to front-end inefficiencies in microarchitecture
- Correlated the root-cause of front-end inefficiencies to the event-driven programming model: little intra-event code reuse coupled with heavy inter-event code reuse
- Developed a microarchitecture mechanism that coordinates the instruction cache replacement policy with instruction prefetcher to significantly mitigate front-end bottlenecks.

Understanding the Role of CPU in Mobile Web Browsing (*IEEE MICRO 2015*)

- Instrumented six top Android smartphones from 2009 to 2014 to quantify how webpage load time and energy consumption vary with CPU performance and network latency
- Proposed a new CPU DVFS governor that coordinates CPU performance with net-work latency for energy savings with little performance impact.

Event-based Scheduling for Energy-Efficient QoS (*HPCA 2015*)

- Introduced the notion of "energy-efficient QoS" (eQoS). It is a framework to systematically reason about energy-QoS trade-offs in mobile Web applications
- Proposed and developed event-based scheduling (EBS) as a new method to achieve eQoS. It trades off performance with energy consumption without hurting user QoS experience.

WebCore: Specialized Processor Architecture for Mobile Web Applications (*ISCA 2014*)

- Performed microarchitecture design space explorations to identify the proper baseline architecture for energy-efficient mobile Web browsing
- Developed specialized hardware accelerators and cache structures that mitigate instruction delivery and data supply bottlenecks in the general purpose design.

Energy-Efficient Mobile Web Browsing on Big/Little Systems (*CAL 2014, HPCA 2013*)

- Instrumented the Pandaboard (ARM Cortex-A9 based) and Beagleboard (Cortex-A8 based) development boards for hardware power/energy analysis
- Built regression models for predicting webpage load time and energy consumption
- Implemented a scheduling mechanism in Mozilla Firefox that leverages big/little heterogeneous systems and outperforms conventional OS DVFS in both performance and energy.

CPU Microarchitecture (with Prof. Yale N. Patt) *2/2011-7/2011*

Microarchitecture Support for Speculative Loop-level SIMDization

- Designed Pintools to characterize multiple benchmarks (SPEC2006, SPLASH, etc.) in order to identify the upper bound of loop level parallelization
- Proposed hardware mechanisms for dynamically vectorizing programs on a cycle-accurate x86 microarchitecture simulator.

Tsinghua University (*Undergraduate*)

GPU Programming and Microarchitecture (with Prof. Yangdong Deng) *6/2009-7/2010*

Hermes: An Integrated CPU/GPU Microarchitecture for IP Routing (*DAC 2011*)

- Proposed a CPU+GPU heterogeneous IP router microarchitecture to adaptively balance packet delay and overall throughput
- Developed analytical models to estimate the delay and throughput of IP packets.

Parallel Gate Level Logical Simulator on GPUs (*DAC 2010, TODAES 2011*)

- Proposed and implemented dynamic paging into GPGPU programming to mitigate the limited memory capacity of GPUs
- Achieved a performance again of 48X for gate-level simulation compared with Synopsys' sequential simulator.

SOFTWARE ARTIFACTS

GreenWeb (https://codereview.chromium.org/1835303002/)

- GreenWeb is a set of Web language extensions defined as Cascading Style Sheet (CSS) rules that let developers express user QoS expectations at an abstract level.
- The Web runtime dynamically determines how to deliver the target QoS experience while minimizing the energy consumption based on programmer-assisted QoS information.

AutoGreen (https://github.com/yuhao/AutoGreen)

- An automatic annotation framework to annotate (mobile) Web applications with GreenWeb language extensions to enable energy-efficiency optimizations by the Web runtime.

Node Benchmark (https://github.com/nodebenchmark/benchmarks)

- Event-driven server workload suite along with a client-side load generator
- The ultimate goal is to enable research on designing better server systems in light of the event-driven execution model that is increasingly prevalent in today's cloud environment.

COMMUNITY SERVICES

Tiny Transactions on Computer Science (TinyToCS) (http://tinytocs.org/)
- Program Committee Chair, Volume IV
- Program Committee Member, Volume III

MobiTools (http://mobitools.ece.utexas.edu/)
- Tutorial on *Infrastructure and Tools for Mobile Computer Architecture Research with an Emphasis on Real System Measurement*, co-located with ISCA 2016
- Co-founder and organizer

Emerging Topics in Computer System Design Reading Group (https://goo.gl/gzT4Fw)
- Co-founder and regular discussion moderator.
- Organize the weekly discussion seminar on emerging trends in computer system research within UT ECE (~15 graduate students and faculty members.)

Women in Engineering Program (WEP) (http://www.engr.utexas.edu/wep)
- Mentor in the Graduates Linked with Undergraduates in Engineering (GLUE) Program.
- Currently mentoring two female sophomores on crowdsourcing-based approximate computing and Web evolution analysis. Meet weekly to direct their research.

Artifacts Evaluation CGO-PPoPP
- Program Committee Member, 2015
- Program Committee Member, 2016

Solicited Reviewer
- DAC 2011, 2012
- TODAES 2011
- ESL 2015

INDUSTRIAL EXPERIENCE

Google Inc. Summer 2015
- Led the efforts of Flow API Version 2 as part of Trace Event, which is a programming model for Chrome code instrumentation and analysis. It has been integrated into Chrome.
- Developed analysis infrastructure that associates Chrome runtime behaviors with user inputs to enable critical path analysis. It has been integrated into Catapult, which is the official performance analysis tool for Chrome.

AMD Research Lab Summer 2013
- Studied the Web browser caching mechanism in Chromium
- Developed performance analytical models for Web browser caching.

AMD Research Lab Summer 2012
- Analyzed microarchitecture bottlenecks of Web browsing on AMD Trinity APUs
- Added support for AMD Trinity APU in libpfm and PAPI performance counter library.

STMicroelectronics Summer 2011
- Fully integrated an ARM System-on-Chip (SoC) functional simulator with a set of embedded benchmarks (e.g., EEMBC, Whetstone, etc.)
- Designed a timing simulator for the cache hierarchy in an ARM SoC.

TEACHING EXPERIENCE

Teaching Assistant
- Lower Division, Introduction to Embedded Systems, with Prof. Jonathan W. Valvano
- Upper Division, Computer Architecture, with Prof. Yale N. Patt
- Graduate Level, Dynamic Compilation, with Prof. Vijay Janapa Reddi