NANDITA VIJAYKUMAR

Electrical and Computer Engineering Department
Phone: (678) 296 5354
Email: nandita@cmu.edu
Web: http://www.cs.toronto.edu/~nandita/

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213

RESEARCH INTERESTS

Computer systems and architecture with a focus on the interaction between programming models, compilers, systems, and architecture; Compilers; Memory systems; GPUs; Heterogeneous Systems

In my dissertation work, I developed powerful, yet practical, cross-layer abstractions in CPUs and GPUs that *bridge the semantic gap* between the application and the underlying system and hardware. Leveraging these abstractions, I demonstrated significantly enhanced performance, productivity, and portability by making higher-level program information available to the compiler, OS, drivers, and hardware architecture. My research provides a unifying and highly practical approach to enabling cross-layer optimizations.

EDUCATION

Carnegie Mellon University — **Ph.D.** in Electrical and Computer Engineering
Aug 2013 – Oct 2019
Advisors: Prof. Onur Mutlu, Prof. Phillip B. Gibbons
Thesis: *Rethinking cross-layer abstractions to enhance productivity, portability, and performance.*

Carnegie Mellon University — **Masters** in Electrical and Computer Engineering
Aug 2013 – Aug 2019 (*Expected*) — Current GPA: 3.90/4.00
Advisors: Prof. Onur Mutlu, Prof. Phillip B. Gibbons

PES Institute of Technology — **B.E. Electrical Engineering**
Undergraduate Thesis Advisor: Prof. B. K. Arunkumar
Aug 2007 – May 2011 — GPA: 9.68/10.00
Undergraduate Research: *Neural networks and fuzzy logic in designing control systems for motor drives.*

PUBLICATIONS

SMASH: Co-designing Software Compression and Hardware-Accelerated Indexing for Efficient Sparse Matrix Operations [MICRO 2019]
Konstantinos Kanellopoulos, Nandita Vijaykumar, Christina Giannoula, Roknoddin Azizi, Skanda Koppula, Nika Mansouri Ghiasi, Taha Shahroodi, Juan Gomez-Luna, and Onur Mutlu

CROW: A Low-Cost Substrate for Improving DRAM Performance, Energy Efficiency, and Reliability [ISCA 2019]
Hasan Hassan, Minesh Patel, Jeremie S. Kim, A. Giray Yaglikci, Nandita Vijaykumar, Nika Mansouri Ghiasi, Saugata Ghose, and Onur Mutlu

A Case for Richer Cross-layer Abstractions: Bridging the Semantic Gap with Expressive Memory [ISCA 2018]
Nandita Vijaykumar, Abhilasha Jain, Diptesh Majumdar, Kevin Hsieh, Gennady Pekhimenko, Eiman Ebrahimi, Nastaran Hajinazar, Phillip B. Gibbons, Onur Mutlu

The Locality Descriptor: A Holistic Cross-Layer Abstraction to Express Data Locality in GPUs [ISCA 2018]
Nandita Vijaykumar, Kevin Hsieh, Eiman Ebrahimi, Phillip B. Gibbons, Onur Mutlu

Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds [NSDI 2017]
Kevin Hsieh, Aaron Harlap, Nandita Vijaykumar, Dimitris Konomis, Greg Ganger, Phillip B. Gibbons, Onur Mutlu

SoftMC: A Flexible and Practical Infrastructure for Enabling Experimental DRAM Studies [HPCA 2017]
Hasan Hassan, Nandita Vijaykumar, Samira Khan, Saugata Ghose, Kevin Chang, Gennady Pekhimenko, Oguz Ergin,

Onur Mutlu

Zorua: A Holistic Approach to Resource Virtualization in GPUs [MICRO 2016]
Nandita Vijaykumar, Kevin Hsieh, Gennady Pekhimenko, Samira Khan, Ashish Shrestha, Saugata Ghose, Adwait Jog, Phillip B. Gibbons, Onur Mutlu

Accelerating Pointer Chasing in 3D-Stacked Memory: Challenges, Mechanisms, Evaluation [ICCD 2016]
Kevin Hsieh, Samira Khan, Nandita Vijaykumar, Kevin K. Chang, Amirali Boroumand, Saugata Ghose, Onur Mutlu

Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems [ISCA 2016]
Kevin Hsieh, Eiman Ebrahimi, Gwangsun Kim, Niladrish Chatterjee, Mike O'Connor, Nandita Vijaykumar, Onur Mutlu, Stephen W. Keckler

Toggle-Aware Bandwidth Compression for GPUs [HPCA 2016]
Gennady Pekhimenko, Evgeny Bolotin, Nandita Vijaykumar, Mike O'Connor, Onur Mutlu, Todd C. Mowry, Stephen W. Keckler

ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality [HPCA 2016]
Hasan Hassan, Gennady Pekhimenko, Nandita Vijaykumar, Vivek Seshadri, Donghyuk Lee, Oguz Ergin, Onur Mutlu

A Case for Core-Assisted Bottleneck Acceleration in GPUs: Enabling Flexible Data Compression with Assist Warps [ISCA 2015]
Nandita Vijaykumar, Gennady Pekhimenko, Adwait Jog, Abhishek Bhowmick, Rachata Ausavarungnirun, Chita Das, Mahmut Kandemir, Todd C. Mowry, Onur Mutlu

Book Chapters

Decoupling the Programming Model from Resource Management in Throughput Processors [Many Core Computing: Hardware and Software, IET, 2019]
Nandita Vijaykumar, Kevin Hsieh, Gennady Pekhimenko, Samira Khan, Ashish Shrestha, Saugata Ghose, Adwait Jog, Phillip B. Gibbons, Onur Mutlu

A Framework for Accelerating Bottlenecks in GPU Execution with Assist Warps [Advances in GPU Research and Practice, Morgan Kaufmann, 2016]
Nandita Vijaykumar, Gennady Pekhimenko, Adwait Jog, Saugata Ghose, Abhishek Bhowmick, Rachata Ausavarungnirun, Chita Das, Mahmut Kandemir, Todd C. Mowry, Onur Mutlu

Papers Under Submission/Preparation

Towards Practical, Efficient, and Realizable Hardware-Software Interfaces to Enhance Application Expressivity
Nandita Vijaykumar, Mehrshad Lotfi, Konstantinos Kanellopoulos, Ataberk Olgun, Nisa Bostanci, Hasan Hassan, Phillip B. Gibbons, Onur Mutlu

EcoRNN: Efficient Computing of LSTM RNN Training on GPUs
Bojian Zheng, Abhishek Tiwari, Nandita Vijaykumar, Gennady Pekhimenko

Open-Source Tools and Infrastructure

Expressive Memory: A Full-System Cross-Layer Interface in CPUs
A cross-layer interface implemented in RISC-V cores on an FPGA with full-stack support. It enables compiler, OS, and architecture research in hardware-software codesigns by supporting efficient communication of higher-level program information to hardware components (ISCA 2018).
`(Artifact analysis work in preparation for submission.)`

SoftMC: Software Memory Controller
An FPGA-based testing platform that can control and test memory modules designed for the commonly-used DDR interface with a C++-based API (HPCA 2017).
https://github.com/CMU-SAFARI/SoftMC

IMPICA: In-Memory Pointer Chasing Accelerator
A gem5-based simulator that models an in-memory pointer chasing accelerator, its corresponding driver, and its applications (ICCD 2016). The simulator has been used as a starting point for PIM (processing-in-memory) research. https://github.com/CMU-SAFARI/IMPICA

Professional Experience

ETH Zurich, Visiting Student in the Systems Group [April 2018 - present]
Rich Cross-Layer Abstractions for Specialized Architectures

Nvidia Research, Graduate Intern with Dave Nellans and Eiman Ebrahimi [Jun 2017 - Aug 2017]
A Holistic Cross-Layer Abstraction to Express and Exploit Data Locality in GPUs

Microsoft Research, Graduate Intern with Olatunji Ruwase and Trishul Chilimbi [Jun 2016 - Aug 2016]
Compressed and Optimized Models for Deep Neural Network Training

Intel, Graduate Intern with Chris Wilkerson (Intel Labs) and Kingsum Chow (Intel SSG) [Jun 2014 - Dec 2014]
Architectural Support for Managed Languages

Advanced Micro Devices, Design Engineer [July 2011 - July 2013]
Architecture/Performance Modeling

ABB, Undergraduate Intern [Jan 2011 - May 2011]
Design and Verification of Low Power Control Products

BEML, Bangalore, Summer Intern [Jun 2009 - Aug 2009]
Embedded and Control Systems in Metro Trains.

Advanced Micro Devices, Undergraduate Intern [Aug 2009 - Dec 2009]
Architecture/Performance Modeling

Awards and Honors

Invited to Rising Stars in Computer Architecture [2018]
Georgia Tech

Invited to Rising Stars in EECS [2017]
Stanford University

Qualcomm Innovation Fellowship Finalist [2015 – 2016]
Qualcomm, USA

Benjamin Garver Lamme/Westinghouse Fellowship [2013 – 2014]
Carnegie Mellon University

Prof. MRD Merit Scholarship [2007 – 2011]
PES Institute of Technology

Spotlight Award for Outstanding achievement in deploying clustering algorithms for workload organization [2012]
Advanced Micro Devices, India

Spotlight Award for Specialized and customer specific workload analysis [2012]
Advanced Micro Devices, India

Distinction Awards for Academic Excellence [2007 - 2011]
PES Institute of Technology

Invited Talks and Posters

Rethinking the Hardware-Software Contract: Enabling practical and general cross-layer optimizations

- ◊ VMware Research, Palo Alto, CA [July 2019]
- ◊ AMD Research, Santa Clara, CA [May 2019]
- ◊ Penn State University, State College, PA [April 2019]

◊ Simon Fraser University, Vancouver, BC [April 2019]
◊ University of Chicago, Chicago, IL [April 2019]
◊ University of Southern California, Los Angeles, CA [April 2019]
◊ University of Waterloo, Waterloo, ON [March 2019]
◊ Rutgers University, New Brunswick, NJ [March 2019]
◊ Duke University, Durham, NC [March 2019]
◊ University of Toronto, Toronto, ON [March 2019]
◊ University of British Columbia, Vancouver, BC [March 2019]
◊ Boston University, Boston, MA [March 2019]
◊ University of Pennsylvania, Philadelphia, PA [February 2019]
◊ University of California, Santa Barbara, CA [February 2019]
◊ University of Texas at Austin, TX [February 2019]
◊ PDL, Carnegie Mellon University, PA [January 2019]
◊ MSR India, Bangalore, India [January 2019]

Expressive Memory: Rethinking the Hardware-Software Contract with Rich Cross-Layer Abstractions
◊ Penn State University, State College, PA [November 2018]
◊ University of Illinois at Urbana-Champaign, Urbana-Champaign, IL [November 2018]
◊ PDL Retreat, Carnegie Mellon University, Bedford Springs, PA [October 2018]
◊ Massachusetts Institute of Technology, Cambridge, MA [October 2018]
◊ CALCM Seminar, Carnegie Mellon University, Pittsburgh, PA [October 2018]
◊ Rising Stars in Computer Architecture, Georgia Tech, Atlanta, GA [October 2018]
◊ EPFL, Lausanne, Switzerland [September 2018]
◊ ETH Zurich, Switzerland [September 2018]

Towards Practical and Powerful Hardware-Software Interfaces to Bridge the Semantic Gap
◊ Poster at CWWMCA Workshop at MICRO-51, Fukuoka, Japan [October 2018]

A Case for Richer Cross-layer Abstractions: Bridging the Semantic Gap with Expressive Memory
◊ ISCA-45, Los Angeles, CA [June 2018]

The Locality Descriptor: A Holistic Cross-Layer Abstraction to Express Data Locality in GPUs
◊ ISCA-45, Los Angeles, CA [June 2018]

A Rich Cross-Layer Interface to Enhace Application Expressivity
◊ Poster at Intel Science and Techology Center (ISTC) Retreat, Santa Clara, CA [October 2017]

Cross-Layer Compute and Memory Abstractions for Enhanced Programmability, Portability, and Performance
◊ Poster at Rising Stars in EECS, Stanford University [November 2017]

Zorua: A Holistic Approach to Resource Virtualization in GPUs
◊ MICRO-49, Taipei, Taiwan [October 2016]
◊ CALCM Seminar, Carnegie Mellon University, Pittsburgh, PA [October 2016]

A Framework for Accelerating Bottlenecks in GPU Execution with Assist Warps
◊ ETH Zurich, Switzerland [January 2016]

A Case for Core-Assisted Bottleneck Acceleration in GPUs: Enabling Flexible Data Compression with Assist Warps
◊ ISCA-42, Portland, OR [June 2015]
◊ Penn State University, State College, PA [June 2015]

Energy-Efficient Data Compression for Modern Memory Systems
◊ Qualcomm Innovative Fellowship Finals, San Diego, CA [Mar 2015]

TEACHING

Carnegie Mellon University, Teaching Assistant with Prof. Phil Gibbons [Spring 2017]
Optimizing Compilers, Graduate

Carnegie Mellon University, Teaching Assistant with Prof. Onur Mutlu [Fall 2015]
Computer Architecture, Graduate

PES Institute of Technology, Teaching Assistant with Prof. Abha Tripathi [Fall 2010]
Power Systems Analysis, Undergraduate

PES Institute of Technology, Teaching Assistant with Prof. S. Venkatesh [Spring 2010]
Digital Signal Processing, Undergraduate

PES Institute of Technology, Teaching Assistant with Prof. Gayathri Devi [Spring 2009]
Linear Integrated Circuits, Undergraduate

STUDENTS SUPERVISED

Hasan Hassan [2015-present]
PhD Research, ETH Zurich. *DRAM testing infrastructures (HPCA 2017) and efficient DRAM substrates (HPCA 2016, ISCA 2019)*.

Konstantinos Kanellopoulos [2018-present]
Research Internship, ETH Zurich. *Hardware-Software codesign for sparse linear algebra (MICRO 2019)*.

Nika Mansouri [2018-present]
Masters Research, ETH Zurich, *Automatic code offload for PIM architectures.*

Bojian Zheng [2018-present]
Masters Research, University of Toronto. *EcoRNN: Fused LSTM RNN Implementation with Data Layout Optimization*

Mehrshad Lotfi [2018]
Research Internship, ETH Zurich. *Towards Practical and Realizable Interfaces to Enhance Application Expressivity*

Abhilasha Jain [2017]
Masters Research, CMU. *Cross-layer Interfaces for Efficient Caching (ISCA 2018)*.

Diptesh Majumdar [2017]
Masters Research, CMU. *Cross-layer Interfaces for Data Placement in Heterogeneous Memories (ISCA 2018)*.

Ashish Shrestha [2016]
Masters Research, CMU. *Zorua: A Holistic Approach to Resource Virtualization in GPUs (MICRO 2016)*.

Mahmoud Khairy [2015]
Research Internship, CMU. *Efficient DRAM Refreshes for GPUs.*

Madhav Iyengar [2015]
Graduate Research Project, CMU. *Introducing Heterogeneity in GPU Architectures.*

Jonathan Leung [2015]
Graduate Research Project, CMU. *Introducing Heterogeneity in GPU Architectures.*

Gaurav Srivastava [2015]
Graduate Research Project, CMU. *Improving Warp Scheduling in GPUs*

Elliot Rosen [2015]
Graduate Research Project, CMU. *Improving Warp Scheduling in GPUs*

Abhishek Bhowmick [2013]
Undergraduate Internship, CMU. *Enabling Flexible Data Compression in GPUs (ISCA 2015)*

GRANTS

NSF Award, *CSR-Core: Effective Data Compression for Modern Memory Systems*, National Science Foundation (Award #1423172). *Contributed to writing and ideas.* [2014–2017]

NSF Expeditions Collaborative Proposal, *Prescriptive Memory: Razing the Semantic Wall between Applications and Computer Systems*. *Contributed to writing and ideas.* [2018]

SERVICE

Program Committee member: IISWC 2019, GPGPU 2019

Reviewer: ICS 2018, MICRO 2017, ICS 2017, PLDI 2017, ISCA 2014-2017, MICRO 2014-2015, HPCA 2014-2017, PACT 2014, DAC 2014-2015, IISWC 2014, ICCD 2014, MICRO Top Picks 2015.

RELEVANT GRADUATE COURSEWORK

Visual Computing Systems, Visual Learning and Recognition, Advanced Operating Systems, Graduate Computer Architecture, Embedded Real-Time Systems, How to Write Fast Code, Advanced Storage Systems, Optimizing Compilers, Special Topics in Computer Systems: Parallel, Heterogeneous, and Emerging Architectures

SKILLS

Programming Languages/Frameworks: C/C++, C#, Python, Perl, OpenMP, CUDA, OpenCL, Hadoop, Assembly (ARM, x86), Html, Halide, TensorFlow, Caffe

Tools/Simulators: GPGPUSim, Simics, PIN, CUDA Nsight/nvprof, Vtune, QEMU, MATLAB, PSPICE, Code Analyst, FUSE, MiPower, LLVM