Manish Gupta

San Jose, CA, USA
mgupta[dot]iitr[AT]gmail[dot]com • http://cseweb.ucsd.edu/~m7gupta

Summary

I am currently working at NVIDIA. I have completed Ph.D. in computer science from UC San Diego. My expertise is in computer architecture, compilers, and GPUs.

Education

- **University of California, San Diego (UCSD)**
  Ph.D. in Computer Science and Engineering, Powell Fellow — Sep 2010 – Sep 2017
  M.S. in Computer Science and Engineering, GPA 3.8/4.0 — Sep 2010 – Jun 2012
- **Indian Institute of Technology, Roorkee, India (IIT-R)**
  B.Tech. in Electrical Engineering, Gold Medalist, GPA 9.2/10.0 — May 2004 – Jun 2008

Experience (Industry & Academia)

- **Sr. Deep Learning Software Engineer, NVIDIA, Santa Clara, CA** Dec 2017 – Present
  - General matrix multiply (Gemm) kernel development in CUTLASS, an open source project
  - WMMA CUDA C++ API targeting NVIDIA's Volta and Turing tensor cores (source code)
  - Designed and implemented double precision complex and gaussian complex Gemms
  - LLVM-IR inspired representation of Gemm and Convolution operations (source code)
  - More CUTLASS source code listings I implemented: cutlass profiler, Gemm pipeline
  - Heuristic-based kernel selection for CUDA Deep Neural Network (cuDNN)
- **Research Assistant, UCSD, San Diego, CA** Sep 2010 – Sep 2017
  - Static taint analysis for CUDA kernels using LLVM
  - Software-based recovery mechanisms for code executing on unreliable hardware
  - Compiler infrastructure to analyze and generate reliable code
  - Big Data analytic and improving Hadoop performance via reuse
- **Software Engineer, AMD Research, Austin, TX** Jan 2016 – Sep 2016
  - Compiler passes using LLVM for GPU kernels written in C++AMP
  - Reliability vs. performance trade-offs study for memories
  - Worked with the U.S. Department of Energy (DoE) on Fast Forward exascale computing project
- **Firmware Engineer, Qualcomm, Hyderabad, India** Sep 2008 – Aug 2010
  - Coded and integrated features such as Dolby Digital decoder, echo-canceller & noise suppressor, automatic gain control, and IIR filters on Qualcomm ADSP
  - Acquired hands-on knowledge of processor architecture, memory management, OS, assembly & C language, RTOS, firmware, system simulation, Tcl programming, and JTAGs
  - Worked on audio codecs as a part of DSP firmware team

Publications (Selected)

- **Reliability-aware Data Placement for Heterogeneous Memory Architecture**
  Manish Gupta, Vilas Sridharan, David Roberts, Andreas Prodromou, Ashish Venkat, Dean Tullsen and Rajesh Gupta in HPCA 2018
- **Compiler Techniques to Reduce the Synchronization Overhead of GPU RMT**
  Manish Gupta, Daniel Lowell, John Kalamatianos, Steven Raasch, Vilas Sridharan, Dean Tullsen, Rajesh Gupta in DAC 2017
- **ASAR: Application-Specific Approx Recovery to Mitigate Hardware Variability**
  Manish Gupta, Abbas Rahimi, Daniel Lowell, John Kalamatianos, Dean Tullsen, Rajesh Gupta in SELSE 2017

- **Verifying GPU Kernels by Test Amplification**
  Alan Leung, Manish Gupta, Yuvraj Agarwal, Rajesh Gupta, Ranjit Jhala, and Sorin Lerner in PLDI 2012

Patents (Selected)

- **Paired Value Comparison for Redundant Multi-Threading Operations**
- **Reliability-aware control of Heterogeneous Memories**
- **Hybrid RMT: Synchronization-free Software-based RMT with Hardware assisted Instruction Lock-stepping**
- **Programmable Polling Comparison-Store and Hardware Assist for Flexible Redundant Multithreading**
- **Reducing GPU RMT Overhead via FIFO Store Buffering**
- **Programmable Paired Comparison-Store with Hardware Assist**
- **Shared-private FIFO for Wait Free Synchronization**
- **Scrubber Workgroup for Synchronization Avoiding Redundant Multithreading**
- **Memory Reliability Management System**
- **Bufferless Communication for Redundant Multithreading using Register Permutation**
- **Waterfall Counters with Application to AVF Estimation**

Ph.D. Dissertation

- **Software Techniques to Enhance Reliability of Memory and Compute Units**
  Written by: Manish Gupta
  Advised by: Prof. Rajesh Gupta and Prof. Dean Tullsen
  University of California, San Diego, 2017

Talks and Posters

- Reliability-aware Data Placement for Heterogeneous Memory Architecture at International Symposium on High-Performance Computer Architecture, *Vienna, Austria* in Feb 2018 (Selected Talk)
- Software Techniques to Enhance Reliability of Memory and Compute Units at Indian Institute of Technology, *Kanpur, India* in Nov 2017 (Invited Talk)
- Software Techniques to Enhance Reliability of Memory and Compute Units at Indian Institute of Technology, *Roorkee, India* in Dec 2017 (Invited Talk)
- Reliability and Performance Trade-off Study of Heterogeneous Memories at Workshop on Memory and Storage Systems, *Gandhi Nagar, India* in Dec 2017 (Invited Speaker)
- Software Techniques to Enhance Reliability of Memory and Compute Units at University of California, *Los Angeles, CA, U.S.A.* in July 2017 (Invited Talk)
- Compiler Techniques to Reduce the Synchronization Overhead of GPU RMT at Design Automation Conference, *Austin, TX, U.S.A.* in June 2017 (Selected Talk)
- ASAR: Application-Specific Approx Recovery to Mitigate Hardware Variability at Silicon Errors in Logic – System Effects, *Boston, MA, U.S.A.* in March 2017 (Selected Talk)
- Reliability and Performance Trade-off Study of Heterogeneous Memories at International Symposium on Memory Systems, *Alexandria, VA, U.S.A.* in October 2016 (Selected Talk)
- Variability's Effect on Program Analysis and Compilers at University of California, *San Diego, CA, U.S.A.* in Oct 2012 (Poster)
- Verifying GPU Kernels by Test Amplification at Programming Language Design and Implementation *Beijing, China* in June 2012 (Poster)

Skills

- **Programming Languages:** C++, C, Python, Shell, Java, MATLAB, Haskell
- **Parallel Programming:** CUDA, C++AMP, OpenCL
- **Compiler Frameworks:** LLVM, Clang, GCC
- **Simulators and Tools:** Ramulator, FaultSim, LaTeX, PinTool, Hadoop, Gem5, McPAT, QEMU
- **Benchmark Suites:** SPEC, PARSEC, Proxy Apps, HPC workloads
- **New Interests (Self-learning):** Deep Learning concepts, TensorFlow (basics)

Teaching Experience

- Basic Data Structure & Object Oriented Design, Teaching Assistant, Fall 2015
- Software for Embedded Systems, Teaching Assistant, Spring 2014
  Please find my teaching reviews at links: Teaching Eval 1 & Teaching Eval 2

Courses (Selected)

Artificial Intelligence ∘ Software for Embedded Systems ∘ Algorithms (I & II)
Operating Systems ∘ Computability & Complexity ∘ Advanced Compiler Design

Awards & Honors

- Gold Medal (Department) for highest GPA in B.Tech EE at IIT-R in a batch 80 students 2008
- Silver Medal (Institute) at IIT-R 2008
- Late Smt. Kailash Devi Goyal Fellowship at IIT-R 2008
- QualStar for contributions to audio firmware at Qualcomm 2009
- QualStar for contributions to SBC encoder at Qualcomm 2010
- Awarded with three years Powell Fellowship at UCSD 2010

Languages

- English (proficient), Hindi (native), and Russian (basic)