Anirudh Mohan Kaushik

amkaushi@uwaterloo.ca

Education

- **University of Waterloo** Waterloo, Canada
  PhD candidate, Electrical and Computer Engineering; GPA:92.50/100 September 2015 – current
  - Supervisor: Dr. Hiren D.Patel, Associate Professor of Electrical and Computer Engineering
  - Courses: Database Systems Implementation, Computer-aided reasoning for Software Engineering, FPGA Based ReconFigurable Computation
    * Research on predictable real-time systems and novel computer architectures for graph analytics. Currently have one conference paper and one journal article under submission.
    * Organized the Computer Architecture Reading Group (CARG), a weekly discussion on new and classic publications in the area of computer architecture.

- **University of Waterloo** Waterloo, Canada
  MASc, Electrical and Computer Engineering; GPA:81.17/100 September 2012 – April 2014
  - Supervisor: Dr. Hiren D.Patel
  - Courses: Algorithm Design and Analysis, Methods and tools for Software Engineering, Introduction to Optimization, Computer Organization, Real-time Embedded Software, Register-transfer-level Digital Systems
    * Master's thesis: Accelerating Mixed-Abstraction SystemC Models on Multi-Core CPUs and GPUs Research on accelerating SystemC (hardware design language in C) simulations on multi-core CPUs and GPUs. Designed a compiler front-end in LLVM/Clang to parse input SystemC models and identify efficient partitioning of SystemC design to execute on CPUs and GPUs. Resulted in three publications at peer-reviewed conferences.

- **Vellore Institute of Technology University** Vellore, India
  Bachelor of Technology, Electronics and Communications; GPA:82/100 June 2008 – June 2012
  - Final year project: Efficient multi-ported memories for FPGAs
    * Implemented a multi-ported memory module for FPGAs to increase instruction-level parallelism and reduce contention among simultaneous processors accessing the memory.
    * The design goals met were low resource utilization and scalability.

Experience

- **CAESR, University of Waterloo** Waterloo, Canada
  Research graduate student (PhD/MASc) Sept. 2012 – April 2014, Sept. 2015 – Present
  - Research in electronic design automation (EDA), real-time systems, and computer architecture. Worked on four projects, each resulting in top-tier conference and peer-reviewed journal submissions.
  - Teaching assistant for mandatory undergraduate course in computer hardware (ECE 222). Responsible for tutorials and labs.

- **IBM Toronto Software Lab, IBM Canada** Toronto, Canada
  Software optimization performance analyst, POWER performance team May 2014 – August 2015
  - Worked with the POWER performance team to identify optimization opportunities for proprietary IBM software and open-source software on IBM power platforms.
  - Optimizing real-world transaction benchmarks (TPC-DS) on IBM DB2 using IBM POWER 8 platform. Worked on identifying specific tuning opportunities to exploit POWER 8 hardware features.
  - Optimizing Spark on IBM POWER 8. Worked on understanding Spark internals and identifying performance opportunities to tune Spark applications for IBM POWER 8 systems. Also carried out competitive analysis of Spark on POWER and Intel x86 using machine learning benchmarks based on Spark.

Skills

Languages: C/C++, Python, SQL, Unix shell scripting, Verilog, VHDL, SystemC, x86 assembly, ARM assembly

Technologies: clang for LLVM, Xilinx tool chain for FPGAs, Micro-architectural simulators: Sniper, gem5, PostgreSQL, Git, Linux, Perf, Intel VTune

Awards and Honors

Ontario Gradute Scholarship	Spring 2017 – Winter 2018
President's Graduate Scholarship	Spring 2017 – Winter 2018
University of Waterloo Graduate Scholarship	Winter 2014, Spring 2017

Projects

Predictable cache coherence for real-time systems [1, 2]: Investigated the applicability of conventional bus based cache coherence protocols for real-time systems that require strict latency guarantees for memory requests. Observed various sources of unpredictability, and designed a new predictable bus based cache coherence protocol for real-time systems. Prototyped the protocol using gem5, a cycle accurate simulator. This work has one accepted paper at 23rd Real-time and Embedded Technology and Applications Symposium (RTAS) 2017 [1]. A follow up technical report [2] investigates cache coherence protocols for mixed-critical real-time systems.

Hardware prefetching for Graph Analytics [3]: Designed a novel hardware prefetcher for graph analytics that takes into account different memory access patterns, and effects of graph execution on underlying micro-architecture. Novelty of the prefetcher is that it utilizes the effects of graph execution on underlying micro-architecture to identify phases of graph execution that impacts performance most, and identifies different data-dependent access patterns. This work is under submission.

Reverse-engineering DRAM memory controller through latency-based analysis [4, 5]: This work identified and developed inference techniques to reverse-engineer the features of the DRAM memory controller through latency based analysis. Features such as the page policy, address mapping, and command arbitration schemes were targets of identification using inference algorithms. Some use cases for identifying features of the memory controller are better memory allocators for performance, and side-channel attack countermeasures. The inference algorithms were tested using a micro-architectural simulator along with a detailed DRAM simulator (MacSim + DRAMSim2) and using a Xilinx FPGA board with an on-board DDR2 memory module. This work resulted in a conference publication at RTAS 2015 [4] and a journal submission is under review.

Accelerating mixed-abstraction SystemC models on multi-core CPUs and GPUs [6]: This thesis focused on identifying opportunities for accelerating simulation of hardware models described in SystemC, which is a hardware descriptive language (HDL), on multi-core CPUs and GPUs. I designed a novel partitioning algorithm that identifies segments of the hardware model to be executed on the CPU and GPU to minimize resource contention, and expensive memory transfer overheads between CPU and GPU. The identification and code transformation from SystemC to CUDA for GPU identified segments of the model are automated through a compiler framework called SystemC-clang [7]. This thesis is available at [6], and has two invited publications [8, 9] on the same theme.

SystemC-clang: A SystemC front-end in clang [7]: SystemC-clang is an open-source SystemC front-end written in clang, and is based on the LLVM compiler framework. It parses SystemC models written in different abstraction levels such as register-transfer-level (RTL) and transaction level (TL) models, and converts the models into an intermediate representation amenable for different types of analysis. SystemC-clang also performs code translation from SystemC to CUDA for accelerating SystemC models on GPUs. This project resulted in a conference submission at Forum of Design and Languages (FDL), 2013 [7], and is currently an open-source project on GitHub (https://github.com/anikau31/systemc-clang).

Publications

[1] Mohamed Hassan, **Kaushik, Anirudh M**, and Hiren Patel. Predictable cache coherence for multi-core real-time systems. In *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2017 IEEE*, pages 235–246. IEEE, 2017.

[2] Nivedita Sritharan, **Kaushik, Anirudh M**, Mohamed Hassan, and Hiren Patel. HourGlass: Predictable Time-based Cache Coherence Protocol for Dual-Critical Multi-Core Systems. 2017. Technical report.

[3] **Kaushik, Anirudh M** and Hiren Patel. Gretch: A Hardware Prefetcher for Graph Analytics. Submitted to 55th Design and Automation Conference, 2018.

[4] Mohamed Hassan, **Kaushik, Anirudh M**, and Hiren Patel. Reverse-engineering embedded memory controllers through latency-based analysis. In *21st IEEE Real-Time and Embedded Technology and Applications Symposium*, pages 297–306. IEEE, 2015.

[5] Mohamed Hassan, **Kaushik, Anirudh M**, and Hiren Patel. Exposing Implementation Details of Embedded DRAM Memory Controllers through Latency-based Analysis. Submitted to ACM Transactions on Embedded Computing Systems, 2018.

[6] **Kaushik, Anirudh Mohan**. Accelerating mixed-abstraction SystemC models on multi-core CPUs and GPUs. 2014. MASc thesis.

[7] **Kaushik, Anirudh M** and Hiren D Patel. SystemC-clang: An open-source framework for analyzing mixed-abstraction SystemC models. In *Specification & Design Languages (FDL), 2013 Forum on*, pages 1–8. IEEE, 2013.

[8] Valeria Bertacco, Debapriya Chatterjee, Nicola Bombieri, Franco Fummi, Sara Vinco, **Kaushik, Anirudh M**, and Hiren D Patel. On the use of GP-GPUs for accelerating compute-intensive EDA applications. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 1357–1366. EDA Consortium, 2013.

[9] Mahesh Nanjundappa, **Kaushik, Anirudh**, Hiren D Patel, and Sandeep K Shukla. Accelerating SystemC simulations using GPUs. In *High Level Design Validation and Test Workshop (HLDVT), 2012 IEEE International*, pages 132–139. IEEE, 2012.