NEERAJ KULKARNI

PhD Candidate, Computer Systems Laboratory, Cornell University nsk49@cornell.edu,+1(607)379-8492

Research Interests

High-performance Computer Architecture, Datacenters, System Resource Management

Education

August 2014 - Current: Ph.D in Electrical and Computer Engineering, **Cornell University**
GPA - 4.0/4.0 Area: Computer Architecture
Committee: Dr. Dave Albonesi, Dr. Christina Delimitrou, Dr. Jose Martinez

July 2014: B.Tech-M.Tech (Dual Degree) in Electrical Engineering, **IIT Kanpur**, India
Master's CGPA - 10.0/10.0, Bachelor's CGPA - 9.0/10.0

Work Experience

June 2017 - Dec 2017: HPC Architect Engineer - Intern at Intel Federal LLC, Hudson, MA
Manager: Olivier Franza, Mentor: Elliot Fleming
Developed distributed and pro-active micro-architectural techniques to mitigate on-chip supply voltage variations induced by di/dt (current spikes) in large CGRAs.

Publications

Neeraj Kulkarni, Christina Delimitrou, Christine Shoemaker, David Albonesi, "Rapid Power Management in Latency-Critical Systems using Reconfigurable Cores" is *under submission*.

Neeraj Kulkarni, Feng Qi, Christina Delimitrou, "Pliant: Leveraging Approximation to Improve Datacenter Resource Efficiency" published at the 25^{th} IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2019.

Neeraj Kulkarni, Feng Qi, Christina Delimitrou, "Leveraging Approximation to Improve Datacenter Resource Efficiency" published in IEEE Computer Architecture Letters (CAL), 2018.

Research Projects

Power Management in Servers using Reconfigurable Cores *(Under Submission)* May. '18 - Current
Advisors- Dr. Christina Delimitrou, Dr. Dave Albonesi

- Tackled low utilization (by co-scheduling of latency-critical & batch apps) and poor energy proportionality (by dynamically tailoring static power usage) in servers using reconfigurable cores.
- Formulated it as an optimization problem that involves appropriately reconfiguring cores, to maximize throughput of batch apps and meet QoS of latency-critical apps while operating under server node's power budget.
- Employed collaborative filtering & heuristic search to rapidly characterize the system and find near-optimal solution.
- Met QoS of latency-sensitive apps & achieved up to 3.5x speed-up on batch apps compared to core-level gating strategy.

Exploring Accuracy, Performance and Resource usage Trade-offs in Machine Learning Algorithms Mar. '19 - Current
Advisor- Dr. Christina Delimitrou

- Deployment of ML apps difficult - i) large computations at training, ii) real-time latency & limited resources for inference.
- Employed approximations to explore accuracy and performance & resource requirements trade-off.
- Working on exploring possibilities to dynamically sacrifice accuracy during high load and limited resources scenarios.

Using Approximate Computing to Improve Datacenter Utilization Sept. '16 - Dec. '18
Advisor- Dr. Christina Delimitrou

- Co-scheduled approximate apps with latency-sensitive apps to boost utilization and mitigated interference in shared resources by employing approximation & resource reclamation from approximate apps.
- Leveraged approximation to a) directly reduce interference, and b) preserve performance of approximate apps when executing with reduced resources, at the cost of some accuracy loss.
- Developed runtime system that switches approximation degree and relocates cores, caches & memory among apps.
- Preserved QoS for all evaluated co-scheduled workloads, while incurring a 2.1% loss in output quality, on average.

Power Management in Heterogeneous Reconfigurable Multi-Core Architectures Feb. '15 - May '18
Advisors- Dr. Dave Albonesi, Dr. Christine Shoemaker

- Built an energy-efficient multicore system to tackle the dark silicon problem using asymmetric & reconfigurable CPU cores that operate by turning on/off lanes in sections of the superscalar out-of-order pipeline.
- Formulated it as an optimization problem that involves mapping applications onto different core types & reconfiguring the cores, to maximize the system throughput while operating under a power budget.
- Developed techniques to solve problem at millisecond-scale using machine learning & global optimization algorithms.
- Outperformed core-level gating and state-of-the-art by up to 30% and 15% respectively across SPEC2006 benchmarks.

Design of Flexible Associative Cache based on Way-Prediction Feb. '16 - May '16
Advisor- Dr. Christopher Batten

- Developed RTL implementation of flexible associative cache (using way-prediction) having performance similar to set-associative cache but energy consumption similar to direct-mapped cache.
- Developed cycle-level C model of way-prediction cache to analyze performance and guide RTL design decisions.
- Designed the flexible associative cache in Verilog; pushed it through the ASIC design tool flow to analyze energy & area.
- Demonstrated 28% energy savings at cost of only 2% performance loss compared to set-associative cache.

Master's Thesis - **Acceleration of Critical Sections in Multi-Threaded Programs** Feb. '13 - June '14
Advisors- Dr. Mainak Chaudhuri, Dr. SSK Iyer

- Demonstrated that load/stores are majorly responsible for stalls in critical sections of multithreaded programs.
- Accelerated loads/stores in critical sections by prioritizing them in Network-On-Chip switches which reduced the network buffer latency of critical requests & subsequently stall times of critical load/stores.
- Achieved speed-up of up to 13% over baseline (round-robin scheduling in switches) across parallel benchmark suites.

Nano-Satellite Project Jugnu, IIT Kanpur with Indian Space Research Organization (ISRO) Dec. '09 - Dec. '11
Advisor- Dr. Nalinaksh Vyas

- Headed the Motor-Driver subsystem, co-headed the Inertial Measurement Unit (IMU) subsystem and worked in the On-board Communication and Ground Station team over the 2 years of my involvement in the project.
- Designed & implemented power efficient control system driving 4 DC motors (Actuators of Attitude Control algorithm)
- Programmed the IMU board involving data acquisition from sensors, flash memory management, interface with OBC.
- Worked at Indian Satellite Centre, Bangalore for conducting various tests during flight model fabrication of JUGNU.

Awards

2017	Selected to participate in 2017 January Colman Leadership Program.
2014	Awarded H.C. Torng Fellowship at Cornell University.
2012 & 13	Academic Excellence Award for being in top 7% of the institute by IIT Kanpur for 2 years.
2009	All India Rank (AIR) 380 in IIT Joint Entrance Exam amongst 0.4 million entrants.
2009	National Top 1% & selected for Indian National Physics Olympiad & Chemistry Olympiad among 35K students.
2008 & 09	National Top 1% and selected for Indian National Astronomy Olympiad at Senior Level among 30K applicants.
2007	National Talent Search(NTS) Scholarship awarded to top 500 students in India.

Technical Skills

Programming:	C, C++, Python, Perl, MATLAB.
Digital Design:	Verilog, BlueSpec Verilog, Synopsys- Design Compiler, PrimeTime, PrimePower, Xilinx ISE, ModelSim.
Architectural Simulators:	sesc, gem5, zsim, McPAT.

Relevant Coursework

Postgraduate/Senior Undergraduate Level

Systems:	Datacenter Computing, Computer Architecture, Advanced Computer Architecture, Operating Systems, Microprocessor Design Methodology, High Level Design Automation
VLSI/Circuits:	Complex ASIC Design, VLSI System Design, Digital/Analog VLSI Circuits, Semiconductor Device Modeling, Organic Electronics, Solid State Devices, Microelectronics
CS (Cornell):	Analysis of Algorithms, Machine Learning for Data Science, Advanced Machine Learning

Teaching Experience

- Head Teaching Assistant for Digital Logic & Computer Organization (ECE2300) at Cornell University in Fall '18 & '19
- Teaching Assistant for Digital Logic & Computer Organization (ECE2300) at Cornell University in Fall '15
- Teaching Assistant for Embedded Systems (ECE3140) at Cornell University in Spring '15.
- Teaching Assistant, Digital Electronics, Fall '13. Awarded best TA award for the year 2013-14 at IIT Kanpur.

Academic Projects

Implementation of multi-core PARC Processor Sept. '14 - Dec. '14
Advisor- Dr. Chritopher Batten

- Implemented a 5-stage pipelined core running PARC ISA with aggressive bypassing and a 2-way set associative FSM write-back write-allocate cache.
- Implemented 8-node bi-directional ring network to support 4-core system with private i-cache and shared d-cache.

Six-stage Pipelined SMIPS Processor and Branch Prediction Schemes Jan. '13 – April. '13
Advisors- Dr. Arvind (MIT), Dr. Ameya Karkare

- Developed an inorder 6-stage pipelined SMIPS processor from a single cycle processor in BlueSpec Verilog (BSV).
- Implemented Set Associative Branch Table Buffer, Two Level Direction Predictors (GAg, SAg and Gshare), Tournament Predictor (GAg/Gshare + SAg), and PPM like Tag-based Predictor (variant of TAGE).
- Coupled the predictors with the SMIPS processor & evaluated prediction accuracies using typical C programs.

CORDIC Processor Design Feb. '13 – April. '13
Advisor- Dr. S. Qureshi

- Implemented 16-bit CORDIC core which has the ability to compute sin, cos, arctan, sinh, cosh, sqrt and ln functions using Verilog in 3 styles of hardware architecture: Combinational, Sequential and Pipelined.
- Developed scaling-free algorithm for sin & cos, which allowed skipping of around 50% of the iterations.
- Obtained power savings of upto 20% in scaling free algorithm over the normal sequential implementation.

Thread Criticality Predictors for Load Balancing Aug. '12 – Nov. '12
Advisor- Dr. Mainak Chaudhuri

- Implemented hardware architecture for the thread criticality prediction (TCP) based on the measurements of L1 and L2 cache misses from each core in a CMP as proposed by Bhattacharjee et.al.
- Applied the TCP hardware to guide task stealing in Intel's TBB library which initially followed random policy.
- Evaluated the performance gains of the TCP method using PARSEC benchmarks to get speed-up as high as 10%.

Characterization of PARSEC Benchmarks Sept. '12
Advisor- Dr. Mainak Chaudhuri

- Instrumented PARSEC apps using PIN tool to find out re-use distance and sharing profiles of accessed memory blocks.
- Simulated 3-level cache hierarchy to calculate the misses for different cache policies and re-use distances of L3 misses.