ASHUTOSH DHAR

Telephone: 217-417-3649 | Email: ashutosh.h.dhar@gmail.com
https://www.linkedin.com/pub/ashutosh-dhar/17/473/21
ashutoshdhar.com

SUMMARY

I'm a PhD candidate in the Electrical and Computer Engineering department at the University of Illinois, Urbana-Champaign, where my primary area of research is in Computer Architecture, with a focus on Reconfigurable and Heterogeneous architectures. My research explores the application of reconfiguration in conventional architectures. I work with Professor Deming Chen.
My research focuses on GPU and multi-core architectures, with an emphasis on providing micro-architecture support for reconfiguration and is informed by a strong background in computer architecture and digital circuit design as well as accelerator development using GPUs, FPGAs and CGRAs. My recent work looks at deep learning and its acceleration as the motivation.

EDUCATION

PhD in Electrical and Computer Engineering (PhD Candidate)
University of Illinois- Urbana Champaign
Masters of Science (MS) in Electrical and Computer Engineering (2014)
University of Illinois- Urbana Champaign — *GPA: 3.96/4.0*
Bachelor of Technology (B.Tech) in Electrical and Electronics Engineering (2012)
National Institute of Technology – Tiruchirappalli (NIT – Trichy), India — *First Class with Distinction*

SKILLS

Programming Languages	: C, C++, Python, CUDA, Shell Script, Perl, MATLAB, SPICE
Tools and Packages	: OpenCV, Nvidia Nvprof, Intel VTune, GPGPU-Sim, Gem5, OMNET++
Hardware Definition Language	: Verilog, Vivado C/C++ HLS
CAD Tools	: HSPICE, Synopsys VCS, Synopsys Design Compiler, Synopsys PrimeTime PX, Cadence Virtuoso, Xilinx Vivado, Altera Quartus, Xilinx Vivado HLS
Undergraduate Course Work	: VLSI Systems, Computer Architecture, Digital Signal Processing, Operating Systems
Graduate Course Work	: Computer Architecture, Parallel Computer Architecture, Artificial Intelligence, Computer Security, Computer Vision, System on Chip Design, Digital IC Design, Design of Fault Tolerant Digital Systems, Analog IC Design, Designing and Building Applications for Extreme Scale Systems, Applied Parallel Programming

WORK EXPERIENCE

Research Intern (Architecture Research Group), NVIDIA Research

(May 2018 – August 2018): Austin, TX

Worked on ***optimizing on-chip memory organizations for deep learning accelerators***. My work focused on finding an organization that would be suitable for a range deep learning models. We explored the trade-offs between capacity, performance, and organizations for a variety of memory structures, as well as hierarchies in the memory-system, on a case-by-case basis to understand the needs of individual DL models. In addition, we explored the impact of different compute organizations on memory-system design

Research Intern, IBM Research

(June 2017 – August 2017): IBM T J Watson Research Center, Yorktown Heights, NY

Worked on ***acceleration and optimization of massively parallel and distributed training of deep networks***. My work focused on compression algorithms for accelerating distributed training, with an emphasis on reducing the communication overhead involved in large scale distributed training. I studied compression techniques on a variety of deep networks that could be deployed in a scalable and GPU-friendly fashion. The work was integrated into a proprietary deep learning infrastructure toolchain.

GPU Architecture Intern, NVIDIA

(May 2016 – August 2016): Santa Clara, CA

Worked on ***performance modeling of new features in next generation GPUs*** and systems. My work focused on building a new performance model and simulation infrastructure for newly added features. The simulator was developed from scratch and will serve as the base infrastructure for future architectures. The simulation infrastructure was developed to be highly scalable, fast and cycle accurate.

Graduate Intern, Cisco Systems Inc.

(May 2013 – August 2013): Silicon Engineering, Enterprise Networking Group, San Jose, CA

Worked on the ***power analysis of Cisco's next-generation switching ASIC***. I focused on developing a power analysis flow using Synopsys's PrimeTime PX. My work involved selecting cases and running tests/simulations to stress blocks, synthesizing blocks-under-test and analyzing the power and clock gating effectiveness under stress as well as under nominal and idle states. I worked on integrating hooks into the verification environment to enable this, along with integrating vendor power libraries and constraints, along with developing scripts to automate tasks and create reports. Analysis was done for actual power on gate-level netlists, with a comparative analysis between netlists before and after place-route, DFT insertion and clock tree creation.

GRADUATE RESEARCH EXPERIENCE

Reconfigurable Compute-Memory Organization for Deep Learning Accelerators

We are exploring ***a reconfigurable compute-memory fabric*** that would be suitable for a range deep learning models. We explore the trade-offs between capacity, performance, and organizations for a variety of memory structures, on a case-by-case basis to understand the needs of individual DL models. In addition, we explored the impact of different compute organizations on memory-system design.

In Memory Architectures for Reconfigurable Accelerators

(MICRO 2020)

We are exploring the use of ***in memory computing units***, embedded in the last level cache of chip multiprocessors for accelerating machine learning algorithms. Our goal is to provide the best of two worlds – in memory computing and reconfigurable computing. We will explore task mapping and scheduling, along with memory consistency.

Work involves:

1. Workload characterization and profiling
2. Architectural modeling of accelerator and components via cycle accurate CPU simulator
3. Building circuit models of memory blocks and RTL models of additional blocks
4. Synthesis and power analysis using 45nm libraries and Synopsys CAD tools.

Coarse Grained Logic Folding for FPGA Accelerators

(VLSID 2020 Best Paper Award)

We are exploring the use of coarse-grained logic folding, to ***fit large workloads on to the FPGAs*** with limited capacity and to improve resource utilization efficiency. The key idea is to partition circuits and demonstrate sharing of hardware resources, across time, to fit the large workloads. As a part of the project we explore customized overlay architectures and explore optimal task scheduling with an ILP based solution, along with workload partitioning techniques to derive solutions.

Work involves:

1. Building analytical models of the workloads and the reconfiguration overhead.
2. Designing an ILP based scheduler.
3. Building circuits and protocols to effectively mask the overhead of reconfiguration.

Machine Learning Based AutoTuners for Coarse Grained Reconfigurable Architectures (CGRA)

As a part of the DARPA software defined hardware (SDH) project, we are working on ***developing auto-tuners for CGRAs.*** Our goal is to monitor the target hardware in real-time and recommend changes to the code and configuration of the CGRA for optimal performance. The overall goal is to specialize hardware and software to be cognizant of the data.

Programmable Near Memory Accelerators

(MICRO 2018 Best Paper Award Nomination)

We explored architectures and programming models that ***incorporate reconfigurable accelerators adjacent to DRAM memory***, as a part of the system. Our goal is to demonstrate seamless integration and transparent programming models, that allow developers to leverage existing distributed/parallel programming tools to leverage near memory acceleration (NMA), and to minimize the need to modify existing processor and memory system architecture.

Work involves:

1. Development of an NMA framework on FPGA based platform
2. Integration of specialized driver sets to support NMA

Addressing GPGPU Inefficiencies with a Cross-Core Sharing and Core-Reorganization

(FCCM 2017)

Explored a new ***GPU architecture capable of reconfiguring itself to better suit application characteristics***. We propose sharing resources between GPU cores to improve performance and energy efficiency of GPGPU applications. As a part of this work, coarse-grained resource sharing is used to exploit to reorganize the GPU's resources. We explore a variety of architectural trade-offs and demonstrate energy savings.

The work involves:

1. Workload characterization and profiling of applications.
2. Architectural modeling of the GPU via a cycle accurate simulator and power simulations.
3. RTL models of GPU data-path and control-paths components and design of reconfigurable circuits
4. Synthesis and power analysis of components using 45nm libraries with Synopsys CAD tools

A Scalable and High-Density Phase Change Memory based FPGA Architecture

(DATE 2015)

Development of a ***new Phase Change Memory based FPGA architecture*** that exploits the characteristics of PCM cells for improved logic density and latency. Work involved the design of a new Look Up Table (LUT) architecture, as well as Logic Blocks (CLB) using PCM cells. Our novel approach was able to double logic density, and significantly improve the area-delay characteristics of the FPGA.
Work Involved:

1. Transistor level design of LUTs in 22nm technology.
2. SPICE modeling and simulations.
3. Evaluation of LUTs through complete FPGA CAD flow – technology mapping, place-route, using data derived from SPICE simulations.
4. Architecture modeling.

Scaling Convolution Neural Networks to Large Compute Clusters

Supervised two undergraduate researchers as a part of an NSF REUs – Passionate on Parallel. The task involved studying the Torch 7 framework for implementing convolution neural networks (CNNs). The primary goal was to parallelize the framework, via MPI, to allow it to be run on large computing clusters via the use of data-parallelism and model-parallelism.

RESEARCH PUBLICATIONS

1. Xinheng Liu, Cong Hao, Yao Chen, Ashutosh Dhar, and Deming Chen, "Wino-SA: Efficient Systolic Architecture for Winograd Convolution," Proceedings of SRC TECHCON, September 2020.
2. Ashutosh Dhar, Xiaohao Wang, Hubertus Franke, Jinjun Xiong, Jian Huang, Wen-mei Hwu, Nam Sung Kim, Deming Chen, “FReaC Cache: Folded Logic Reconfigurable Computing in the Last Level Cache”, 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2020
3. Ashutosh Dhar, Mang Yu, Wei Zuo, Xiaohao Wang, Nam Sung Kim and Deming Chen., “Leveraging Dynamic Partial Reconfiguration with Scalable ILP Based Task Scheduling,” Proceedings of IEEE 2020 33nd International Conference on VLSI Design and 2020 19th International Conference on Embedded Systems (VLSID). **(Best Paper Award)**
4. Cong Hao, Yao Chen, Xinheng Liu, Atif Sarwari, Daryl Sew, Ashutosh Dhar, Bryan Wu, Dongdong Fu, Jinjun Xiong, Wen-mei Hwu, Junli Gu, and Deming Chen, “NAIS: Neural Architecture and Implementation Search and its Applications in Autonomous Driving,” Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD), November 2019. (Invited)
5. Ashutosh Dhar, Sitao Huang, Jinjun Xiong, Damir Jamsek, Bruno Mesnet, Jian Huang, Nam Sung Kim, Wen-mei Hwu, and Deming Chen, “Near-Memory and In-Storage FPGA Acceleration for Emerging Cognitive Computing Workloads,” Proceedings of IEEE Computer Society Annual Symposium on VLSI (ISVLSI), July 2019. (Invited)
6. M. Alian, S. Min, H. Asgharimoghaddam, A Dhar, et. al., "Application-Transparent Near-Memory Processing Architecture with Memory Channel Network", 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2018 **(Nominated for Best Paper Award)**
7. A. Dhar and D. Chen, “Efficient GPGPU Computing with Cross-Core Resource Sharing and Core Reconfiguration”, Proceedings of IEEE International Symposium on Field-Programmable Custom Computing Machines, (FCCM) 2017
8. A. Dhar and D. Chen, "Neuromorphic Architecture Inspired Fast, Efficient and Configurable On-Chip Learning Via In-Memory Computing and RRAM”, Poster, 2015 Workshop on Hardware and Algorithms for Learning On-a-chip (HALO), (ICCAD) 2015
9. C. Wei, A. Dhar and D. Chen, “A Scalable and High-Density FPGA Architecture with Multi-Level Phase Change Memory”, Proceedings of Design, Automation and Test in Europe, (DATE) 2015
10. J. Wang, A. Dhar, D. Chen, Y. Liang, Y. Wang, and B. Guo, "Workload Allocation and Thread Structure Optimization for MapReduce on GPUs," Proceedings of SRC Technical Conference (TECHCON), September 2014.

GRADUATE COURSE PROJECTS

CS598 - Designing and Building Applications for Extreme Scale Systems

(Spring 2015): Prof William Gropp

A graduate level course project, in which we ***built performance models for GPGPU workloads.*** We built expectation models based on the architecture of the GPU and the algorithm to predict the performance of the application. Models were compared to data from application runs on target GPU and used to further refine the models.

CS533 - Parallel Computer Architecture

(Spring 2013): Prof Joseph Torellas

A graduate level course project, in which we ***explored techniques to improve the power efficiency of Network-On-Chips (NOCs)*** for many-core systems. We studied the network traffic of a multi-processor system for a variety of NOC topologies and memory models and analyzed the various facets of the system to reduce the power consumption of across the NOC. We presented optimal network design choices based on application workloads and processor architecture.

ECE549 - Computer Vision

(Spring 2013): Prof Svetlana Lazebnik

As a graduate level course project, we developed ***a gesture recognition system for a web-cam based system***. We designed the system using a Mixture-of-Gaussian model for background subtraction and a convex hull estimator. We implemented the recognition with an empirical model instead of using trained classifiers.

CS461 - Computer Security

(Fall 2012): Prof Susan Hinrichs

As a graduate level course project, I developed a ***Verilog implementation of the SHA-1 hash algorithm.*** The implementation was designed to be modular and hierarchical. The core was designed to function in a stand-alone fashion as well as for SoCs. The core was verified using Synopsys VCS and was fully synthesizable. The implementation adhered to the NIST SHA-1 standard.

UNDERGRADUATE RESEARCH

Undergraduate Final Thesis, National Institute of Technology- Trichy, India

(January 201 – May 2012): Dr Sishaj P Simon (Department of Electrical and Electronics Engineering)

Development of a ***swarm intelligence based robotic system for grid exploration.*** The framework developed consists of an innovative modified Ant Colony Optimization algorithm applied to path planning of autonomous drones as well as a communication protocol. To demonstrate the effectiveness of the system, we created a distributed robotic system and operated it within the framework.

MISCELLANEOUS

1. VLSID 2020 Best Paper Award
2. Included in List of Teachers Ranked as Excellent. (Spring 2013, Fall 2013, Spring 2017, Fall 2019).
3. Reviewer for ACM Transactions on Design Automation of Electronic Systems (TODAES).
4. Reviewer for ACM Transactions on Reconfigurable Technology and Systems (TRETS).
5. Reviewer for IEEE Transactions on Computer Aided Design (TCAD).
6. Reviewer for IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS).
7. Reviewer for SCIENCE CHINA Information Sciences.
8. Secondary reviewer for ACM/EDAC/IEEE Design and Automation Conference (DAC).
9. Secondary reviewer for IEEE/ACM International Conference on Computer Aided Design (ICCAD).
10. Secondary reviewer for IEEE/ACM International Symposium on Low Power Electronics and Designs (ISLPED).
11. Secondary reviewer for ACM/DIGDA International Symposium on Field Programmable Gate Arrays (FPGA).

GRADUATE TEACHING EXPERIENCE

ECE498ICC – Internet of Things

(Spring 2019): Prof Deming Chen, Prof Wen-mei Hwu, Prof Jinjun Xiong

As the graduate teaching assistant for the class, I am responsible for holding regular office hours as well lab sections. In my role as a TA, I assist in the creation and grading of homeworks, labs, and exams. I helped develop lab material relating to concepts in machine learning, deep learning, Python, IOT devices, and accelerators.

ECE527 – System on Chip Design

(Fall 2015, Fall 2017): Prof Deming Chen

As the graduate teaching assistant for the class, I am responsible for assisting the Professor with course logistics as well as helping students grasp key concepts. I helped develop a whole new set of machine problems and teaching material for the course centered around the Xilinx Zynq SoC platform. I helped develop material that taught students concepts relating to High Level Synthesis, SoC Design, DMA based transfers, Accelerator development and Hardware-Software co-design.

ECE110 – Introduction to Electronics

(Spring 2013 – Spring 2015, Spring 2016, Spring 2018, Fall 2019): Dr. Patricia Franke, Dr. Christopher Schmitz
I was a graduate teaching assistant for the course, focusing on the laboratory portion of the course. My responsibilities included providing brief lectures on key concepts relating to the lab and supervising and assisting students with lab work. The course focused on introductory concepts of Electrical Engineering from basic circuit analysis to logic design.