# 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* | : GPGPU-Sim, Gem5, OMNET++, OpenCV, Nvidia Nvprof, Intel VTune |
| *Hardware Definition Language* | : Verilog |
| *CAD Tools* | : HSPICE, Synopsys VCS, Synopsys Design Compiler, Synopsys PrimeTime PX, Cadence Virtuoso, Xilinx Vivado, Altera Quartus |
| *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

### *On-Chip Memory Organization for Deep Learning Training Accelerators*

We are exploring ***on-chip memory organizations of accelerators*** 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, 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. Our exploration is driven by analytical models that I developed.

### *In Memory Architectures for Reconfigurable Accelerators*

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.

### *Coarse Grained Logic Folding for FPGA Accelerators*

We are exploring the use of coarse-grained logic folding, to ***fit large machine learning 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.

### *Programmable Near Memory Accelerators*

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.

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

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.

## 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 Technical Conference (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.