Dibakar Gope

☎ +1(979) 229-6899
✉ gope@wisc.edu
☞ https://dibakar.github.io/
☞ https://github.com/Dibakar/

Senior Research Engineer
Arm Research
Austin, Texas, USA

Education

University of Wisconsin, Madison, Wisconsin USA
Ph.D., Electrical Engineering, August 2017, GPA 3.95
Dissertation: "Architectural Support for Scripting Languages"
Advisor: Professor Mikko H. Lipasti
Minor: Computer Sciences

Texas A&M University, College Station, Texas USA
M.S., Computer Engineering, August 2011, GPA 4.0
Thesis: "Maximizing Crosstalk-Induced Slowdown During Path Delay Test"
Advisor: Professor Duncan M. (Hank) Walker

Birla Institute of Technology & Science, Pilani, India
B.E. (Hons.), Electrical and Electronics Engineering, August 2008, GPA 9.6

Research and Professional Experience Summary

- 6 years of research experience on variety of topics of Computer Architecture, and 2 years of research experience on Digital Circuit design; authored 10+ research publications in top-tier conferences
- 2+ years of industrial experience on architecting and compressing neural networks for highly constrained systems; authored 3 research publications, and 3 publications are currently under submission
- Good knowledge across various sub-domains of Machine Learning ranging from computer vision, natural language processing, linear algebra to bayesian learning to reinforcement learning by taking 12+ courses on online Coursera platform; see https://github.com/Dibakar/

Publications (Peer Reviewed)

1. Ternary Hybrid Neural-Tree Networks for Highly Constrained IoT Applications
   Dibakar Gope, Ganesh Dasika, and Matthew Mattina
   In *2nd Conference on Systems and Machine Learning* (**SysML**), March 2019.
2. Run-Time Efficient RNN Compression for Inference on Edge Device
   Urmish Thakker, Jesse Beu, Dibakar Gope, Ganesh Dasika, and Matthew Mattina
   In *4th Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications*, *In conjunction with* **ISCA**, June 2019.
3. RNN Compression using Hybrid Matrix Decomposition
   Urmish Thakker, Ganesh Dasika, Jesse Beu, Dibakar Gope, and Matthew Mattina
   In *tinyML Summit*, March 2019.
4. A Case for Scoped Persist Barriers in GPUs
   Dibakar Gope, Arkaprava Basu, Sooraj Puthoor, and Mitesh Meswani
   In *11th Workshop on General Purpose Processing using GPU* (**GPGPU**), *In conjunction with ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (**PPoPP**), February 2018.
5. The CURE: Cluster Communication Using Registers
   Vignyan Reddy Kothinti Naresh, Dibakar Gope, and Mikko H. Lipasti
   In *International Conference on Compilers, Architecture, and Synthesis for Embedded Systems* (**CASES**), October 2017.

6. Architectural Support for Server-Side PHP Processing
   Dibakar Gope, David J. Schlais, and Mikko H. Lipasti
   In *Proceedings of the 44th ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), June 2017.
7. Hash Map Inlining
   Dibakar Gope, and Mikko H. Lipasti
   In *Proceedings of the 25th IEEE International Conference on Parallel Architectures and Compilation Techniques* (**PACT**), September 2016.
8. Statement-Level Parallelism for Scripting Languages
   Dibakar Gope, and Mikko H. Lipasti
   In *1st Workshop on the High Performance Scripting Languages, In conjunction with* **PPoPP**, February 2015.
9. Bias-Free Branch Predictor
   Dibakar Gope, and Mikko H. Lipasti
   In *Proceedings of the 47th IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), December 2014.
10. Bias-Free Neural Predictor
    Dibakar Gope, and Mikko H. Lipasti
    In *Proceedings of the 4th JILP Workshop on Computer Architecture Competitions (JWAC-4): Championship Branch Prediction* (**CBP**), June 2014.
11. Atomic SC for Simple In-order Processors
    Dibakar Gope, and Mikko H. Lipasti
    In *Proceedings of the 20th IEEE International Symposium on High Performance Computer Architecture* (**HPCA**), February 2014.
    Nominated for Best Paper Award
12. Maximizing Crosstalk-Induced Slowdown during Path Delay Test
    Dibakar Gope, and Duncan M. (Hank) Walker
    In *Proceedings of the 30th IEEE International Conference on Computer Design* (**ICCD**), September 2012.
13. Exploring a Circuit Design Approach based on One-hot Multi-valued Domino Logic
    Dibakar Gope, Pey-Chang K. Lin, and Sunil P. Khatri
    In *Proceedings of the 53rd IEEE International Midwest Symposium on Circuits & Systems* (**MWSCAS**), August 2010.
14. Detection of High Resistance Bridge Defects using Slack-Based Dynamic Bridging Fault Model
    Dibakar Gope, Srinivasulu Alampally, Srinivas K. Vooka, and Rubin A. Parekhji
    In *Proceedings of the Synopsys Users Group India Conference* (**SNUG**), 2008.

Publications (In Submission)

1. Ternary MobileNets via Per-Layer Hybrid Filter Banks
   Dibakar Gope, Jesse Beu, and Matthew Mattina
   Submitted to the *8th International Conference on Learning Representations* (**ICLR**), 2020.
2. Compressing Recurrent Neural Networks without Compromising Inference Runtime
   Urmish Thakker, Jesse Beu, Dibakar Gope, Ganesh Dasika, and Matthew Mattina
   Submitted to the *3rd Conference on Systems and Machine Learning* (**SysML**), 2020.
3. Compressing RNNs for IoT Devices by 15-38x using Kronecker Products
   Urmish Thakker, Jesse Beu, Dibakar Gope, Chu Zhou, Igor Fedorov, Ganesh Dasika, and Matthew Mattina
   Submitted to the *3rd Conference on Systems and Machine Learning* (**SysML**), 2020.

Journal Publications

1. The CURE: Cluster Communication Using Registers
   Vignyan Reddy Kothinti Naresh, Dibakar Gope, and Mikko H. Lipasti
   In *ACM Transactions on Embedded Computing Systems* (**TECS**), October 2017.

Patent Applications

1. Apparatus and Method for Bias-Free Branch Prediction
   Mikko Lipasti, and Dibakar Gope
   U.S. Patent number 9,952,870, issued on April 24, 2018.
2. Scoped Persistence Barriers for Non-Volatile Memories
   Arkaprava Basu, Mitesh R. Meswani, Dibakar Gope, and Sooraj Puthoor
   U.S. Patent pending, filed on September 23, 2016.

Industrial Experience

- Senior Research Engineer, Machine Learning & AI, Arm Inc., Austin, TX (07/10/2017-Present)

  Ternary Hybrid Neural-Tree Networks for Highly Constrained IoT Applications (SysML 2019 conf.)
  - Proposed a hybrid network which combines the strengths of current neural and tree-based learning techniques in conjunction with ternary quantization for running highly constrained IoT applications efficiently on Arm Cortex-M microcontrollers
  - Obtained significant reduction in computations, model size and overall memory footprint over the state-of-the-art keyword-spotting neural network with negligible loss in accuracy

  Ternary MobileNets via Per-Layer Hybrid Filter Banks (under submission)
  - Maintained precision critical convolutional filters in full-precision values and applied ternary quantization (StrassenNets) only to easy-to-quantize and mutually similar filters at each layer
  - Obtained significant energy savings, model size reduction and runtime speedups on custom hardware over full-precision MobileNets trained on ImageNet with negligible loss in accuracy

  RNN Compression using Hybrid Matrix Decomposition and Kronecker Product (tinyML Summit 2019, under submission)
  - Proposed a hybrid matrix decomposition technique for RNNs that divides the weight matrix into two parts - an unconstrained upper half and a constrained lower half composed of rank-1 blocks
  - Leveraged Kronecker product to compress RNN cells even further
  - Achieved significant compression for RNNs while being faster than pruning and more accurate than a traditional matrix factorization technique

  The State of Binary / Ternary Quantization in Recurrent Neural Networks (Ongoing)
  - Study the potential of three recent techniques for quantizing weights in the state-of-the-art language modeling network (AWD-LSTM) trained on Penn Treebank to binary/ternary bits
  - Combine their strengths to derive a better method for quantizing AWD-LSTMs to binary/ternary bits

  Machine Learning Inference
  - Performed microarchitectural characterization of the execution of key machine learning applications on existing processors
  - Implemented microarchitectural timing models of processors for exploring a variety of configurations to perform machine learning inference
  - Recommended changes for efficiently running key applications

- Co-Op Engineer, Advanced Micro Devices, Inc., Austin, TX (06/08/2015 - 12/11/2015)

A Case for Scoped Persist Barriers in GPUs (GPGPU 2018 workshop)
- Introduced scoped persist barriers for GPUs to correctly and efficiently manipulate persistent data structures residing in non-volatile memory technologies (NVRAM).
- Proposed three flavors of scoped persist barriers, demonstrated trade-offs between programmability, performance and recoverability of them through a novel parallel B+tree insertion kernel on GPU.

- Co-Op Engineer, Advanced Micro Devices, Inc., Sunnyvale, CA (05/24/2010 - 08/25/2010)
- Developed gate-level ATPG models of mixed-signal macros (MISC I/Os, GDDR5) from Spice netlist using Conformal LEC tool and hand-modeling techniques.

- FSL Design Engineer I, Freescale Semiconductor India Pvt. Ltd., India (07/08/2008 - 07/24/2009)
- Designed the I/O padring of a cellular processor considering placement, routing and ESD constraints.
- Performed RTL logic, timing verification on few modules of a cellular processor.

- Project Trainee, Texas Instruments India Pvt. Ltd., India (01/07/2008 - 06/27/2008)
- Developed a novel timing-driven ATPG technique targeting high-resistance bridge faults through long paths in deep-submicron designs using slack based bridging fault model. This invention resulted in a conference paper publication.
- Developed BIST pattern optimization flow targeting multiple fault models without any loss of test coverage and without adding extra design complexity.

Academic Experience

- Fellowship, University of Wisconsin, Madison, Wisconsin (Sep. 2011 - Aug. 2012)
Research Assistant, University of Wisconsin, Madison, Wisconsin (Sep. 2012 - Jun. 2017)
I conducted original research in improving the execution efficiency of PHP scripting language through hardware accelerators and compiler optimizations, designing highly accurate branch prediction, and efficient memory consistency model for modern processor architecture.

Hardware Accelerators for Real-World Server-Side PHP Web Applications (ISCA 2017 conf.)
- Performed detailed microarchitectural characterization of popular large-scale PHP web applications used to generate dynamic web content, found little or no opportunity for conventional microarchitectural enhancements
- Identified targeted acceleration of four fine-grained PHP activities – hash table accesses, heap management, string manipulation, and regular expression handling as an answer to improve their performance and energy efficiency
- Exploited PHP application-intrinsic characteristics and regular expression patterns to design novel, inexpensive hardware accelerators for these activities and support rich PHP language features

Inlining Hash Map Accesses in Server-Side PHP Applications (PACT 2016 conf.)
- Current state-of-the-art JIT compilers can not inline accesses to hash maps in real-world server-side PHP applications accessing databases and other middleware
- Identified the root cause hampering inlining of such hash maps and provided compiler enhancements to mitigate those in the HipHop VM infrastructure

Designing Highly Accurate Branch Predictor by Capturing Useful Information (MICRO 2014 conf., ISCA CBP 2014 workshop)
- Biased branches resolve as either taken or not-taken virtually every time, hence including them in a branch predictor's history does not directly provide any useful information
- Exploited this observation to improve prediction accuracy than state-of-the-art predictors through learning correlations only with non-biased branches

- Such bias-free neurally-inspired perceptron predictors reach very deep into the execution history, outperforming state-of-the-art neural predictors

Efficient Sequential Consistency (SC) Support for In-Order Processors (HPCA 2014 conf.)
- State-of-the-art SC enforcing hardware proposals for conventional out-of-order processors are not efficient for many-core chips with simple, in-order cores
- Designed a light-weight scheme for enforcing mutual exclusion to maintain proper SC order for reordered references
- Proposed solution mitigates most of the performance loss caused by cache misses in in-order multiprocessors and works without requiring any speculation, rollback support or any alteration to the coherence protocol

- Graduate Asst Researcher, Texas A&M University, College Station, Texas (09/01/2010 - 07/05/2011)
I conducted original research in design for testing, and dynamic CMOS circuits, designed and implemented a path delay test generator maximizing crosstalk-induced slowdown in modern digital circuits, and performed a detailed evaluation of a circuit and layout fabric based on dynamic circuits.

Developing ATPG Algorithm to Maximize Crosstalk-Induced Slowdown (ICCD 2012 conf.)
- Proposed a timing-driven test generation algorithm to sensitize multiple aligned aggressors coupled to a delay-sensitive victim path to detect the combination of a delay spot defect and crosstalk-induced slowdown

Exploring a Circuit Design Approach based on One-hot Multi-valued Domino Logic (MWSCAS 2010 conf.)
- Demonstrated improvement in delay, area and power in comparison to a standard cell based realization and avoided the problems encountered by traditional domino logic approaches

COURSEWORK (MACHINE LEARNING)

- Introduction to Deep Learning (Advanced ML Specialization on Coursera)
- Bayesian Methods for Machine Learning (Advanced ML Specialization on Coursera)
- Practical Reinforcement Learning (Advanced ML Specialization on Coursera)
- Natural Language Processing (Advanced ML Specialization on Coursera)
- Deep Learning in Computer Vision (Advanced ML Specialization on Coursera)
- Neural Networks and Deep Learning (Deep Learning Specialization on Coursera)
- Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization (Deep Learning Specialization on Coursera)
- Structuring Machine Learning Projects (Deep Learning Specialization on Coursera)
- Convolutional Neural Networks (Deep Learning Specialization on Coursera)
- Sequence Models (Deep Learning Specialization on Coursera)
- Linear Algebra (Mathematics for ML Specialization on Coursera)
- Multivariate Calculus (Mathematics for ML Specialization on Coursera)
- Principal Component Analysis (PCA) (Mathematics for ML Specialization on Coursera)
- Introduction to Recommender Systems: Non-Personalized and Content-Based (Recommender Systems Specialization on Coursera)

- See my GitHub repositiry https://github.com/Dibakar/ for the course certificates, and my Jupyter notebooks for various assignments and projects of these courses

Skills

- Machine Learning Frameworks: TensorFlow, MXNet, PyTorch, Keras
- Computer Languages: C, C++, Python, PHP, JavaScript, MATLAB, Perl, OpenCL, CUDA, SQL
- Architecture Evaluation Simulators/Tools: Gem5, AMD Gem5-GPU, Pin
- Compilers: HipHopVM (PHP JIT compiler from Facebook), LLVM, Trimaran
- Hardware Description Language: Verilog

Professional Activities

- External Review Committee Member:
  - Int'l Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2020
  - Int'l Symposium on Computer Architecture (ISCA), 2019
  - Int'l Conference on Parallel Architectures and Compilation Techniques (PACT), 2019

- Program Committee Member:
  - Int'l Symposium on Workload Characterization (IISWC) 2019
  - Int'l Conference on Computer Design (ICCD), 2019
  - Int'l Conference on Computing Frontiers (CF), 2019
  - Int'l Conference on Compilers, Architecture, and Synthesis of Embedded Systems (CASES), 2019

- Presenter for numerous conference publications

References

Professor Mikko H. Lipasti
University of Wisconsin, Madison
Dept. of Electrical and Computer Engineering
1415 Engineering Drive
Madison, WI 53706, USA
☎ +1(608) 265-2639
✉ mikko@engr.wisc.edu

Professor Gurindar S. Sohi
University of Wisconsin, Madison
Dept. of Computer Sciences
1210 West Dayton Street
Madison, WI 53706, USA
☎ +1(608) 262-7985
✉ sohi@cs.wisc.edu

Professor Duncan M. (Hank) Walker
Texas A&M University
Dept. of Computer Science and Engineering
College Station, TX 77843, USA
☎ +1(979) 862-4387
✉ walker@cse.tamu.edu

Professor Arkaprava Basu
Indian Institute of Science
Dept. of Computer Science and Automation
Bangalore, Karnataka 560012, India
☎ +91 80 2293 3560
✉ arkapravab@iisc.ac.in