ARDAVAN PEDRAM

531 Lassen St,
Los Altos, CA 94022
http://web.stanford.edu/~perdavan/
Cell: 512-217-1945
Perdavan@gmail.com

Education

The University of Texas at Austin — Aug 2013
PhD, Computer Engineering

University of Tehran — Sept 2006
MSc, Computer Engineering

Awards & Grants

- IEEE Micro Top Picks 2018.
- Best Paper Award IEEE ASAP 2017.
- *National Science Foundation* **(NSF)** awarded a **$1M grant** to my initiated, authored, and conducted proposal in Stanford "PRISM: Platform for Rapid Investigation of efficient Scientific-computing & Machine-learning".
- *National Science Foundation* **(NSF)** awarded a **$500K grant** to my PhD research project: "Algorithm/Architecture Co-Design of Low Power and High-Performance Linear Algebra Compute Fabrics" (2012-2015).
- Special Interest Group, SIGDA-DAC PhD Forum, 2013.
- TCPP Best Poster Award, IEEE IPDPS PhD Forum, 2013.
- Best Paper Runner-Up in IESS, Germany, 2009.

Industry

Samsung America,
Director of DataCenter Distributed ML Training System — May 2020 – Present

- In charge of initiation, definition, technical management, roadmap, and team alignment for next generation algorithm/architecture codesign of the datacenter deep learning training system.

Cerebras Systems,
Next Gen Architecture Lead — Dec 2018 – Jan 2020

- In charge of technical management, roadmap, and team alignment for next generation algorithm/architecture codesign of computer core, memory hierarchy, and interconnect components.
- Architected the memory hierarchy of processor core and respected mapping of DNN kernels.
- Architected the datapath of the processor core and respected mapping of various micro-kernels on it.
- Proposed major improvements on the current Waferscale architecture to increase utilization and save power for the next generation.
- Designed and implemented an enriched roofline model in Python with several memory components for workload characterization on Waferscale compute fabric.
- Workload characterization versus off-chip memory bandwidth, latency, and capacity for modern DNNs.
- Investigated HBM memory design and integration into Waferscale system with respected programming model.

Cerebras Systems,
Research Scientist Nov 2016 – Nov 2018

- Analyzed and successfully proved the feasibility of mapping FFT kernel on Waferscale integrated system.
- Published the first paper of the company on pipelined backpropagation architecture design exploration.
- In charge of recruiting and university relation.
- Evaluated the feasibility and proposed the Polyhedral method for distributed task generation.
- Presented a tutorial on training deep learning in the cloud at Hotchips30.
- Implemented, and tested all of the special math functions for the internal low-level C++ language using piecewise-polynomial approximation and Newton-Raphson methods.
- Created a generalized Python test framework for multi-input multi-precision evaluation of floating-point arithmetic implementation of special functions and their relative error.
- Performed aggressive performance optimization techniques to reduce the latency of sequential C level code for math function.

Movidius Inc, Nov 2015 – Sep 2016
Consulting Scientist and Principal Research Scientist

- Analyzed and optimized linear algebra and CNN library implementations for custom processors.
- Researched and implemented fast (Winograd) convolution methods and provided cost analyses.
- Researched error analyses and testing of Convolutional Neural Network implementations.

Texas Advanced Computing Center (TACC), Austin, Texas June 2013 – Aug 2013
Summer Intern

- Researched on Fast Fourier Transform for multi-core architectures that achieve orders of magnitude better energy efficiency in signal processing applications resulted into a published journal paper.

Butterfly Network Inc, Guilford, Connecticut May 2012 – Aug 2012
Summer Intern

- Worked directly in a startup with the image-processing team on bio-imaging hardware pipeline. Implemented and verified the Matched Filter and Peak Detector blocks on Xilinx Virtex-VI FPGA.
- Evaluated the power, performance, and area trade-offs of various Xilinx FFT and FIR IP cores to assess their impact on the resource usage within the image-processing pipeline.

Academia

Teaching

Stanford University, Department of Computer Science
Adjunct Professor of Computer Science
CS217 Hardware Accelerators for Machine Learning Jan 2020 – Apr 2020
CS217 Hardware Accelerators for Machine Learning Sep 2018 – Dec 2018
https://cs217.stanford.edu

- Proposed the first course on designing hardware accelerators for machine learning and deep learning.
- Created the syllabus and timeline of the course.
- Taught in all lectures except the guest lectures.
- Organized and scheduled the guest lecture topics and invited all of the guest lecturers.

Stanford University, Department of Computer Science and Electrical Engineering
Pervasive Parallelism Lab (PPL), Professor **Kunle Olukotun**
Research Associate Sept 2014 – present

- Director of the PRISM Project. Leading a team of several graduate student researching on codesign for deep learning applications
- Pursued on automating and porting compute intensive kernels on Coarse Grain Reconfigurable Architectures (CGRAs)
- Co-wrote the PRISM NSF proposal. PRISM is a Platform for Rapid Investigation of efficient Scientific-computing & Machine-learning accelerators.
- Assessed the algorithmic attributes of Convolutional Neural Networks (CNNs) targeting improved energy efficiency for machine learning data classification applications. Recommended and developed a new framework to mathematically formulate the loop ordering, scheduling, blocking, and parallelism in the memory hierarchy of CNNs to optimize the design.

The University of Texas at Austin, Department of Electrical Engineering Sept 2013 – Sept 2014
FLAME Group, Professor **Robert van de Geijn** and Professor **Andreas Gerstlauer**
Postdoctoral Fellow

- Task scheduling of Direct Acyclic Graphs (DAGs) on Heterogeneous platforms.
- Cycle accurate simulation of the Linear Algebra Processor (LAP) and its integration into MARSSx86 cycle accurate simulator to evaluate system level trade-offs.
- Researched on Singular Value Decomposition (SVD) and its parallel implementation algorithms.

The University of Texas at Austin, Department of Electrical Engineering Aug 2008 – May 2013
FLAME Group, Professor **Robert van de Geijn** and Professor **Andreas Gerstlauer**
Research Assistant
Dissertation:
"Algorithm/Architecture Codesign of Low Power and High Performance Linear Algebra Compute Fabrics"

- Introduced an interdisciplinary approach to low-power high-performance hardware design and memory hierarchy of a flexible scientific accelerator ("LAP") with orders of magnitude improved efficiency for level-3 BLAS and FFTs, as well as Cholesky, LU, and QR matrix factorizations.
- Investigated upper limits on performance/power ratios that can be achieved for dense linear algebra and FFTs, and demonstrated the sources of inefficiency in the current systems by creating analytical models to benchmark state of the art GPUs and CPUs using Wattch and MCPAT methodology.
- Implemented a cycle accurate micro-architectural functional simulator for the linear algebra accelerator with C++ released under The Free BSD license.
- Performed state of the art micro-architectural level research on Floating-Point Multiply Accumulate (FP-MAC) units and their extensions resulting better efficiency for the domain of linear algebra.

Tutorials in Major Computer Systems Conferences

- **Hot chips 30:** "Architectures for Accelerating Deep Neural Nets. Part 3: Accelerating Training in the Cloud", August 2018.
- **ISCA 2019, 2020, and 2021: "**Hardware Accelerators for Training Deep Neural Networks".

Conference Papers

1- H. Abdelaziz, A. Shafie, J. Shin, **A. Pedram,** J. Hassoun, “Rethinking Floating Point Overheads for Mixed Precision DNN Accelerators” MLSys 2021.
2- D. Koeplinger, M. Feldman, R. Prabhakar, Y. Zhang, S. Hadjis, R. Fiszel, T. Zhao, L. Nardi, **A. Pedram**, C. Kozyrakis, K. Olukotun, “Spatial: A Language and Compiler for Application Accelerators” PLDI 2018.
3- Y. Li, **A. Pedram**, “CATERPILLAR: Coarse Grain Reconfigurable Architecture for Accelerating the Training of Deep Neural Networks” ASAP 2017. **(Best paper award)**
4- R. Prabhakar, Y. Zhang, D. Koeplinger, M. Feldman, T. Zhao, S. Hadjis, **A. Pedram**, C. Kozyrakis, K. Olukotun, “Plasticine: A Reconfigurable Architecture for Parallel Patterns” ISCA 2017. **(IEEE Micro Top Picks 2018)**
5- A. Vasilyev, N. Bhagdikar, **A. Pedram**, S. E. Richardson, S. Kvatinsky, M. Horowitz, “Evaluating Programmable Architectures for ISP and Computer Vision” MICRO-49, 2016.
6- H. Ha, **A. Pedram**, S. E. Richardson, S. Kvatinsky, M. Horowitz, “Improving Energy Efficiency of DRAM by Exploiting Half Page Row Access” MICRO-49, 2016.
7- M. Asri, **A. Pedram**, L. K. John, A. Gerstlauer, “Simulator Calibration for Accelerator-Rich Architecture Studies,” SAMOS 2016.
8- S. Han, X. Liu, H. Mao, J. Pu, **A. Pedram**, M. Horowitz, B. Dally, “EIE: Efficient Inference Engine on Compressed Deep Neural Network,” ISCA 2016.
9- **A. Pedram**, J. McCalpin, A. Gerstlauer, “Transforming a Linear Algebra Core to an FFT Accelerator” ASAP2013.
10- **A. Pedram**, A. Gerstlauer, R. van de Geijn, “Floating Point Architecture Extensions for Optimized Matrix Factorization,” ARITH21, 2013.
11- **A. Pedram**, A. Gerstlauer, R. van de Geijn, “On the Efficiency of Register File versus Broadcast Interconnect for Collective Communications in Data-Parallel Hardware Accelerators,” SBAC-PAD 2012.
12- **A. Pedram**, S. Gilani, N. S. Kim, R. van de Geijn, M. Schulte, A. Gerstlauer, “A Linear Algebra Core Design For Efficient Level-3 BLAS,” ASAP2012.
13- **A. Pedram**, A. Gerstlauer, R. van de Geijn, “Towards a High-performance, Low-power Linear Algebra Core,” ASAP2011.
14- **A. Pedram**, D. Craven, A. Gerstlauer, “Modeling Cache Effects at the Transaction Level,” IESS2009. **(Best paper runner-up)**
15- M. Daneshtalab, **A. Pedram**, M. H. Neishaburi, M. Riazati, A. Afzali-Kusha, S. Mohammadi, “Distributing Congestions in NoCs through a Dynamic Routing Algorithm based on Input and Output Selections,” VLSID'07.
16- **A. Pedram**, M. Daneshtalab, S. M. Fakhraie, “An Efficient Parallel Architecture for Matrix Computations,” NORCHIP2006.
17- **A. Pedram**, M. Daneshtalab, S. M. Fakhraie, “A High-Performance Memory-Efficient Parallel Hardware for Matrix Computations in Signal Processing Applications,” ISCIT2006.
18- **A. Pedram**, M. R. Jamali, C. Lucas, S. M. Fakhraie, “Reconfigurable Parallel Hardware for Computing Local Linear Neuro-Fuzzy Model,” PARELEC06.
19- M. R. Jamali, **A. Pedram**, M. R. Milasy, C. Lucas, “Design and Implementation of BELBIC Pattern,” ICEE2006

Journal Papers

1- R. Prabhakar, Y. Zhang, D. Koeplinger, M. Feldman, T. Zhao, S. Hadjis, **A. Pedram**, C. Kozyrakis, K. Olukotun, "Plasticine: A Reconfigurable Architecture for Parallel Patterns," IEEE Micro Top Picks, 2018.
2- **A. Pedram,** S. Richardson, Sameh Galal, S. Kvatinsky, M. Horowitz, "Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era," IEEE Design and Test, 2017.
3- **A. Pedram**, J. McCalpin, A. Gerstlauer, "A Highly Efficient Multicore Floating-Point FFT Architecture Based on Hybrid Linear Algebra/FFT Cores," Journal of Signal Processing Systems, Springer, June, 2014.
4- **A. Pedram**, A. Gerstlauer, R. van de Geijn, "Algorithm, Architecture, and Floating-Point Codesign of a Matrix Factorization Accelerator," IEEE Transactions on Computers Special Section on Computer Arithmetic, August 2014.
5- **A. Pedram**, R. van de Geijn, A. Gerstlauer, "Codesign Tradeoffs for High-Performance, Low-Power Linear Algebra Architectures," IEEE Transactions on Computers Special Issue on Energy Efficient Computing, Volume 61, Issue 2, pp 1724-1736, December 2012.
6- **A. Pedram**, M. R. Jamali, C. Lucas, S. M. Fakhraie, "Local Linear Model Tree (LOLIMOT) Reconfigurable Parallel Hardware," Trans, Engineering, Computing and Technology, Volume 13, pp 96-101, May 2006.

Technical Reports

1- N. Gamboah, K Kudrolli, A. Dhoot, **A. Pedram**, "Campfire: Compressible, Regularization-Free, Structured Sparse Training for Hardware Accelerators" arXiv, 2020.
2- X. Yang, J. Pu, B. B. Rister, N. Bhagdikar, S. Richardson, S. Kvatinsky, J. Ragan-Kelley, **A. Pedram,** M. Horowitz, "A Systematic Approach to Blocking Convolutional Neural Networks" arXiv, 2016

Invited Talks

- **SIAM Parallel Processing for Scientific Computing** "Principles of Hardware Accelerator Design for Machine Learning" Feb, 2020.
- **ValleyML State of AI** "Principles of Hardware Accelerator Design for Machine Learning" Aug 2019.
- **University Washington at Seattle 2017** "PRISM: Platform for Rapid Investigation of Efficient Scientific-computing & Machine-learning" July 2017
- **AMD Research 2017** "Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era" Santa Clara, California, March 31, 2017.
- **Nvidia Research**, "PRISM: Platform for Rapid Investigation of Efficient Scientific-computing & Machine-learning" Santa Clara, California, Oct 12, 2015.
- **Qualcomm Inc.** Bay Area Research and Development Center (**BARD**), "Algorithm/Architecture Codesign for Extremely Efficient Compute Fabrics," Santa Clara, California, July 1, 2015.
- **Barcelona Supercomputing Center (BSC)**, "Algorithm/Architecture Codesign for Extremely Efficient Compute Fabrics," Barcelona, Spain, March 24, 2015.
- **IBM**, "A High Performance Low-Power Linear Algebra Processor", Austin, Texas, Nov 4, 2011.

Professional Activities

Core Program Panelist [2017, 2018]

- National Science Foundation (NSF) Division of Computing and Communication Foundations (CCF)

Technical Program Committee

- The 27th and 28th International Conference on Application-specific Systems, Architectures and Processors **[ASAP 2016, 2017, 2018]**
- International Conference on Compilers, Architecture, and Synthesis for Embedded Systems **[CASES 2016, 2017, 2018, 2019]**
- The 28th IEEE International Parallel & Distributed Processing Symposium **[IPDPS 2014]**

Journal Reviewer

- IEEE Transactions on Embedded Computing Systems (TECS)
- IEEE Transactions on Very Large Scale Integrated Circuits (TVLSI)
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
- ACM Transactions on Architecture and Code Optimization (TACO)
- IEEE Journal of Solid State Circuits (JSSC)
- IEEE Transactions on Emerging Topics in Designs for Application-Specific Computing (TETC)
- ACM Transactions on Reconfigurable Technology and Systems (TRETS)
- Journal of Design Automation for Embedded Systems (DAEM)
- IEEE Computer Architecture Letters (CAL)
- IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)
- IEEE Micro Special Issue on Heterogeneous Computing
- IEEE Transactions on Circuits and Systems Part-II (TCAS-II)
- ACM Transactions on Parallel Computing (TOPC)
- IEEE Transactions on Parallel and Distributed Systems (TPDS)

Conference Reviewer:

- 21st International Conference on Parallel and Distributed Computing, Euro-Par 2015
- 32nd IEEE International Conference on Computer Design (ICCD 2014)
- The 2014 International Symposium on Code Generation and Optimization (CGO 2014)
- The 15th IEEE International Conference on Computational Science and Engineering (CSE 2012)
- 3rd ACM/SPEC International Conference on Performance Engineering (ICPE 2012)
- IEEE International Symposium on High Performance Computer Architecture (HPCA)
- IEEE International Conference on Parallel and Distributed Systems (ICPADS)

Workshops and PhD Forums

1- **Supercomputing 2013,** "Algorithm/Architecture Codesign of Low Power and High Performance Linear Algebra Compute Fabrics," PhD Dissertation Showcase
2- **DAC 2013**, "Algorithm/Architecture Codesign of Low Power and High Performance Linear Algebra Compute Fabrics," ACM SIGDA PhD Forum. **(Travel grant award)**
3- **IPDPS 2013**, "Algorithm/Architecture Codesign of Low Power and High Performance Linear Algebra Compute Fabrics," PhD Forum, **(Best poster award)**
4- **HPCA 2012**, "Overcoming Register File Inefficiencies by Using 2D Broadcast Bus Interconnects in Linear Algebra Accelerators," SOCs, Heterogeneous Architectures and Workloads (SHAW-3), HPCA 2012