Harideep Nair

Mountain View, CA
Personal Website: *hpnair.github.io*
Mobile: +1-408-326-9548
Email: *harideepnr7@gmail.com*
LinkedIn: *linkedin.com/in/harideepnr7*

Research: Neuromorphic processor design and silicon implementation targeting extremely energy-efficient acceleration of diverse AI/ML workloads, along with automated design tool development spanning PyTorch models to chip layout.

Key Highlights besides PhD Research:

- 2 years industry experience on DL Accelerator SW-HW codesign at MediaTek (Exemplary Performance Award).
- Research collaboration with MediaTek on energy-efficient, sparsity-exploiting low-precision MAC designs synthesized in TSMC N5 using industry-standard EDA tools.
- Co-created two graduate courses on Computer Architecture (18-740 and 18-743) at CMU.
- 5 years experience as *Head Teaching Assistant* delivering lectures and developing lab assignments exploring CPU/GPU/NPU cores in flagship mobile SoCs from Qualcomm (Snapdragon) and MediaTek (Dimensity).

Education

- **Carnegie Mellon University** — Pittsburgh, PA
  Ph.D. - Electrical and Computer Engineering — Aug'18 - Aug'24
  Thesis: Cortical Columns Computing Systems - Microarchitecture, Functional Building Blocks and Design Toolsuite
  Advisor: Prof. John Paul Shen
  Thesis Committee: John Paul Shen (Distinguished Service Professor, CMU), James E. Smith (Emeritus Professor, UW-Madison), Brandon Lucia (Kavčić-Moura Professor, CMU), and *Perry Wang (Director, MediaTek)*
- **Indian Institute of Technology (IIT) Bombay** — Mumbai, India
  B.Tech. + M.Tech. - Electrical Engineering | **Minor** - Computer Science — Jul'13 - Jul'18
- **National University of Singapore** — Singapore
  Semester Exchange Program - Electrical and Computer Engineering — Fall'16

Professional Experience

- **MediaTek Inc., San Jose, CA | AI Computer Architecture Research Intern** — Jan'22 - Dec'22
  - Developed **architectural simulator** to assess design tradeoffs in future generations of MediaTek AI accelerator.
  - Added **MLIR compiler** support for variety of neural network operators for mapping AI workloads to hardware.
  - Performed research on novel **matrix multiplication units** implementing brain-inspired temporal arithmetic.
- **MediaTek Inc., San Jose, CA | AI Computer Architecture Research Intern** — Jan'21 - Dec'21
  - Worked on **ISA** and microarchitecture design for next-gen AI accelerator in production mobile SoCs.
  - Designed novel **RTL** blocks within convolution engine from scratch and implemented in System Verilog.
  - Performed full functional verification and obtained post-synthesis power, performance and area in **TSMC N5**.
  - Developed neuromorphic algorithms for **AI super resolution** targeting low-power mobile deployment.
- **MediaTek Inc., San Jose, CA | AI Computer Architecture Research Intern** — May'20 - Aug'20
  - Developed hardware-efficient **Computer Vision** algorithms for edge inferencing on Dimensity SoCs.
  - Performed feature testing of **NeuroPilot** AI software ecosystem and worked on its official documentation.
  - Helped kickstart **MediaTek-CMU collaboration** to support course on Modern Computer Architecture.
- **Ushva Clean Technology, Mumbai, India | Embedded System Engineer Intern** — Nov'15 - Jan'16
  - Part of the team building a "Smart Home Solar Power System" with wireless load control and data monitoring.
  - Created a **wireless** central hub and six mini hubs using PIC MCUs, and RF/Wi-Fi modules (**IoT** system).

Book Chapters

- John Paul Shen and **Harideep Nair**. "*Cortical Columns Computing Systems: Microarchitecture Model, Functional Building Blocks, and Design Tools*". *Neuromorphic Computing*, Ch. 8, IntechOpen, 15 Nov. 2023. Crossref, doi:10.5772/intechopen.110252.

Conference & Journal Publications

- **Harideep Nair**, Prabhu Vellaisamy, Tsung-Han Lin, Perry Wang, Shawn Blanton, and John Paul Shen. "Commercial Evaluation of Zero-Skipping MAC Design for Bit Sparsity Exploitation in DL Inference", Accepted In *2024 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, IEEE, 2024.

- Shanmuga Venkatachalam, **Harideep Nair**, Prabhu Vellaisamy, Yongqi Zhou, Ziad Youssfi, and John Paul Shen. "Realtime Person Identification via Gait Analysis using IMU Sensors on Edge Devices", Accepted In *2024 International Conference on Neuromorphic Systems (ICONS)*, 2024.

- Prabhu Vellaisamy, **Harideep Nair**, Di Wu, Shawn Blanton, and John Paul Shen. "Exploration of Unary Arithmetic-Based Matrix Multiply Units for Low Precision DL Accelerators", Accepted In *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, IEEE, 2024.

- **Harideep Nair**, William Leyman, Agastya Sampath, Quinn Jacobson, and John Paul Shen. "NeRTCAM: CAM-Based CMOS Implementation of Reference Frames for Neuromorphic Processors", Accepted In *Proceedings of the 2024 Annual Neuro-Inspired Computational Elements Conference (NICE)*, 2024.

- **Harideep Nair**, David Barajas-Jasso, Quinn Jacobson, and John Paul Shen. "TNN-CIM: An In-SRAM CMOS Implementation of TNN-Based Synaptic Arrays with STDP Learning", Accepted In *2024 IEEE 6th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, IEEE, 2024.

- Prabhu Vellaisamy*, **Harideep Nair***, Vamsikrishna Ratnakaram, Dhruv Gupta, and John Paul Shen. "TNNGen: Automated Design of Neuromorphic Sensory Processing Units for Time-Series Clustering", In *IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS-II)*, IEEE, 2024 [Published by Invitation after Acceptance in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*].

- Prabhu Vellaisamy, **Harideep Nair**, Joseph Finn, Manav Trivedi, Albert Chen, Anna Li, Tsung-Han Lin, Perry Wang, Shawn Blanton, and John Paul Shen. "*tubGEMM: Energy-Efficient and Sparsity-Effective Temporal-Unary-Binary Based Matrix Multiply Unit*", In *2023 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 1-6. IEEE, 2023.

- **Harideep Nair**, Prabhu Vellaisamy, Albert Chen, Joseph Finn, Anna Li, Manav Trivedi, and John Paul Shen. *"tuGEMM: Area-Power-Efficient Temporal Unary GEMM Architecture for Low-Precision Edge AI"*, In *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5. IEEE, 2023.

- **Harideep Nair**, Prabhu Vellaisamy, Santha Bhasuthkar, and John Paul Shen. "*TNN7: A Custom Macro Suite for Implementing Highly Optimized Designs of Neuromorphic TNNs*", In *2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 152-157. IEEE, 2022.

- Shanmuga Venkatachalam, **Harideep Nair**, Ming Zeng, Cathy Tan, Ole Mengshoel, and John Paul Shen. "*SemNet: Learning Semantic Attributes for Human Activity Recognition with Deep Belief Networks*", *Frontiers in Big Data*, Vol. 5, 2022. doi:10.3389/fdata.2022.879389.

- **Harideep Nair**, John Paul Shen, and James E. Smith. "*A Microarchitecture Implementation Framework for Online Learning with Temporal Neural Networks*", In *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 266-271. IEEE, 2021.

- Shreyas Chaudhari, **Harideep Nair**, José M.F. Moura, and John Paul Shen. "*Unsupervised Clustering of Time Series Signals using Neuromorphic Energy-Efficient Temporal Neural Networks*", In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7873-7877. IEEE, 2021.

- **Harideep Nair**, Cathy Tan, Ming Zeng, Ole J. Mengshoel, and John Paul Shen. "*AttriNet: Learning Mid-Level Features for Human Activity Recognition with Deep Belief Networks*", In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) and Proceedings of the 2019 ACM International Symposium on Wearable Computers (ISWC)*, pp. 510-517. 2019.

- Raj Kumar Choudhary, Newton Singh, **Harideep Nair**, Rishabh Rawat, and Virendra Singh. "*Freeflow Core: Enhancing Performance of In-order Cores with Energy Efficiency*", In *2019 IEEE 37th International Conference on Computer Design (ICCD)*, pp. 702-705. IEEE, 2019.

*Both co-first authors contributed equally to this work.

Invited Talks

- "Brain-Emulating and Silicon-Optimized Neuromorphic Sensory Processing Units", ***AMD Research***, Santa Clara, CA. *June 2024*.
- "Brain-Emulating and Silicon-Optimized Neuromorphic Sensory Processing Units", ***Numenta, Inc.***, Redwood City, CA. *May 2024*.
- "Introduction to Neuromorphic Computing", Course on *Embedded Deep Learning*, **Carnegie Mellon University**, Pittsburgh, PA. *Dec. 2023*.
- "Neuromorphic Cortical Column-Based Edge-AI Sensory Processors for Beam Prediction", *AI Research Team,* ***Ericsson, Inc.***, Santa Clara, CA. *Aug. 2023*.
- "Building a Silicon Neocortex in CMOS", *Alternative Computing Group*, *National Institute of Standards and Technology (**NIST**)*, Gaithersburg, MD. *Dec. 2020*.

Submissions Currently Under Review

- **Harideep Nair**, Anand Raju, Swamynathan Siva, Akshunna Vaishnav, Shanmuga Venkatachalam, Quinn Jacobson, and John Paul Shen. "*ReFCAM: CMOS Implementation of Complete Reference Frames for Cortical Columns Computing*", Under Review In *IEEE International Conference on Computer Design (ICCD)* 2024.
- **Harideep Nair**, Prabhu Vellaisamy, Wei-Che Huang, YoungSeok Na, Yuyang Kang, and John Paul Shen. "*C3SGen: An Automated PyTorch-to-Layout Design Toolsuite for Cortical Columns Computing Systems*", Under Review In *IEEE International Conference on Computer Design (ICCD)* 2024.
- Prabhu Vellaisamy, **Harideep Nair**, Thomas Kang, Yichen Ni, Haoyang Fan, Bin Qi, Jeff Chen, Shawn Blanton, and John Paul Shen. "*Tempus-Core: : Evaluation of Temporal-Unary-Binary MAC Units for Deep Learning Accelerators*", Under Review In *IEEE International Conference on Computer Design (ICCD)* 2024.

Selected Honors and Awards

Institute Level:

- **Exemplary Performance Award** by **MediaTek** for impactful innovative contribution during 2-year internship.
- **Qualcomm Innovation Fellowship** and **Dean's Fellowship** for pursuing PhD at Carnegie Mellon University.
- **Rank 1** in Mathematics Olympiad conducted by IIT Bombay Mathematics Association.

State Level:

- Ranked among **State Top 10** in Maths Talent Search Examination **four times**.
- Obtained **99.98 percentile** in Maharashtra Talent Search Examination **thrice** consecutively.

National Level:

- **Gold Medal** in National Chemistry Olympiad (top 40 students in India); selected for the IChO Training Camp.
- **All India Rank 3** in All India Open Mathematics Scholarship Examination.
- **INSPIRE scholarship** by the Government of India for being among top 1% in 12^{th} grade examination.
- Prestigious **National Talent Search Examination (NTSE) scholarship** by the Government of India.

International Level:

- **TF LEaRN Scholarship** (1/53 recipients all over Asia) for semester exchange at National University of Singapore.
- **99.95 percentile** in Science Olympiad conducted by Science Olympiad Foundation among 0.1 million candidates.

Skills

- **Programming**: C/C++/System C, Python, SystemVerilog/Verilog/Verilog-A, VHDL, MATLAB, Java, HTML, SQL
- **ML Frameworks**: PyTorch, TensorFlow, Edge Impulse, TensorFlow Lite, Tensorboard, Keras, MLOps, SNPE, NeuroPilot
- **Software Tools**: Gem5, Snipersim, Ramulator, Synopsys Design Compiler, Cadence Genus, Virtuoso, Xilinx Vivado, HFSS

Course Design & Teaching/Leadership Experience (11 Semesters)

- **Carnegie Mellon University | 18-740 (Modern Comp. Arch. & Design)** Fall'19, '20, '21, '22, '23
  - Played instrumental role in **creating** the course at CMU and establishing **industry collaboration**.
  - Led over 10 TAs across five offerings of the course, managing course logistics and delivering several lectures.
  - Designed lab assignments implementing branch predictors (inc. TAGE) in C and ROB/IQ/Renaming in Verilog.
  - Designed lab assignments exploring CPU/GPU/NPU cores inside Qualcomm/MediaTek's SOTA mobile SoCs.
- **Carnegie Mellon University | 18-743 (Neuromorphic Comp. Arch.)** Spring'19, '20, '21, '22, '23, '24
  - Played instrumental role in **creating** the course at CMU, closely related to *NCAL* research.
  - Led over 12 TAs across six offerings of the course, managing course logistics and delivering several lectures.
  - Mentored over 35 teams of graduate students in research projects (S/W algorithms & H/W implementations).
- **IIT Bombay | EE-309 (Microprocessors)** Fall'17
  - Helped with design, proctoring and grading of examinations for a class of 100 undergraduate students.
  - Held regular office hours to help clarify conceptual questions, encouraging interaction among students.

Selected Technical Reports, Workshops & Posters

- **Exploring Unary MACs for DLAs | *ASPLOS 2nd Workshop on Unary Computing*** Apr'24
  Programming/Tools: PyTorch, System Verilog, Cadence EDA Tools
  - Position paper on preliminary exploration comparing recently proposed unary arithmetic-based MAC units.
  - Performed power-performance-area (PPA) and DL workload sparsity analysis across multiple bit precisions.
- **Cortical Columns Computing System | *IBM IEEE CAS/EDS AI Compute Symposium*** Nov'23
  Programming/Tools: PyTorch, System Verilog, Cadence EDA Tools
  - Position paper on prior research accomplishments and future roadmap for Cortical Columns Computing System.
  - Presented poster to diverse audience from academia/industry at IBM Watson Research Center, NY.
- **DL-Based Gait Recognition on Arduino | CMU** Aug'23 - Dec'23
  Programming/Tools: TensorFlow, Edge Impulse, Arduino Nano 33 BLE Sense
  - Efficient 4-layer CNN design for classifying IMU sensor data; demonstrated on Arduino Nano TinyML kit.
  - Achieved 96.7% accuracy on 24 classes with 125 mW power and 70 ms inference latency.
- ***OzMAC - Zero Omitting MAC Architecture* | MediaTek, CMU** Aug'22 - Dec'22
  Programming/Tools: PyTorch, System Verilog, Synopsys EDA Tools
  - Designed zero-skipping MAC architecture that leverages dynamic data sparsity to achieve high energy efficiency.
  - Profiled and demonstrated high data sparsity for eight state-of-the-art INT8 mobile DNN workloads.
  - Achieved 21%, 70% and 28% improvements in area, power, and energy relative to conventional MACs.
- **Neuromorphic Custom Macro Cells | WIP Workshop, Design Automation Conference (DAC)** Jul'22
  Programming/Tools: System Verilog, Cadence EDA Tools, Virtuoso
  - Developed highly optimized custom hard macros for efficient CMOS implementation of neuromorphic designs.
  - Achieved 27%, 17% and 55% improvements in area, power and energy-delay product (EDP).
- **Facial Emotion Recognition using Efficient Deep Neural Networks | CMU** Aug'19 - Dec'19
  Programming/Tools: PyTorch, Qualcomm SNPE, Snapdragon 855 HDK
  - Hardware-efficient attention-based CNN design; demonstrated on Qualcomm Snapdragon 855 mobile platform.
  - Top 10 accuracy performance in ICML'13 FER Challenge with 3x/8x less power/latency than VGG-19.
- ***Hardware-Aware Neural Network Architectures Using FBNet* | CMU** Jan'19 - May'19
  Programming/Tools: PyTorch, Tensorboard, Raspberry Pi
  - Developed NAS methodology based on FBNets with combined loss-latency-energy optimization.
  - Generated architectures achieved similar performance as MobileNetV2 with 3.8x/2.5x less energy/latency.
- **Computer Systems | CMU** Aug'18 - Dec'18
  Programming/Tools: C

- Implemented dynamic memory allocator for C and optimized its space utilization, resulting in 2x throughput.
- Designed interactive shell command-line interpreter for running built-in as well as user programs.

- ***Masters Thesis - Energy-Efficient Microarch. with Dynamic Renaming*** | **IIT Bombay** May'17 - Jul'18
  Programming/Tools: C, Gem5, SniperSim
  - Freeflow frontend with inorder backend along with dynamic renaming optimizations for IPC improvement.
  - Achieved 150% higher energy-efficiency relative to out-of-order superscalar CPU with just 3.5% drop in IPC.
- **Hardware Acceleration of AES Decryption** | **National University of Singapore** Aug'16 - Dec'16
  Programming/Tools: VHDL, C, Xilinx Zynq-7000 FPGA
  - Implemented AES decryption engine and TFT display controller on FPGA to decrypt encrypted image inputs.
  - Developed its equivalent software simulator in C to perform algorithmic validation.
- **Electromagnetically Secure Integrated Circuits** | **Purdue University** May'16 - Aug'16
  Programming/Tools: Ansys HFSS, MATLAB
  - Modeled multi-layered IC stack and determined layers responsible for significant EM signal leakage.
  - Simulated EM side channel attack and successfully extracted correct key byte using correlation analysis.

Selected Graduate Coursework

CMU:

- Machine Learning by *Tom Mitchell* & *Matt Gormley*
- Deep Learning by *Bhiksha Raj*
- Embedded Deep Learning by *Ziad Youssfi*
- Systems and Toolchains for AI Engineers by *Mohamed Farag* & *Guannan Qu*
- Neural Computation by *Tai Sing Lee*
- Foundations of Computer Systems by *John P. Shen* & *Saugata Ghose*
- Hardware Architectures for Machine Learning by *Diana Marculescu*

NUS:

- Embedded Hardware System Design by *Rajesh Panicker* & *Ha Yajun*
- Computer Vision and Image Processing by *Sim-Heng Ong*

IITB:

- Advanced Topics in Computer Architecture by *Virendra Singh*
- Advanced Processor Design by *Virendra Singh*
- Operating Systems by *Dhananjay M. Dhamdhere*
- Computer and Network Security by *Kameswari Chebrolu*

Extra-Curricular Services & Activities

- Reviewer for IEEE International Symposium on Circuits and Systems (**ISCAS**) 2024.
- Reviewer for IEEE Transactions on Multimedia (**TMM**) 2024.
- **Eleven** Teaching Assistantships at CMU (**led 10 as head TA**) and one at IIT Bombay.
- Invited for **Young Asian Leaders Global Conference** at Nanyang Technological University, Singapore.
- **Winning team** in Intra-Institute Electrical Circuit Designing Competition at IIT Bombay.
- Awarded Hostel **Technical Special Mention** for excellent contribution towards technical activities at IIT Bombay.
- Secured **First Prize** in District-Level Chess Tournament conducted by Shiv Sena, Mumbai.
- Publicity Coordinator at *MCYC Community Services*, Singapore for spreading awareness about Fostering.
- Fundraising as part of *Sing Youth Hub* for charity towards an Old Age Home in Singapore.
- Finished a basic course in German, and learning Spanish and French at a beginner level.
- Possess a foreign currency collection of coins (72 countries) and notes (15 countries).