Prabhu Vellaisamy

pvellais@andrew.cmu.edu - (609) 423-3147 – LinkedIn - Google Scholar - GitHub

EDUCATION

Carnegie Mellon University Pittsburgh, PA

Doctor of Philosophy in Electrical and Computer Engineering September 2021 - December 2026

Master of Science in Electrical and Computer Engineering January 2020 - May 2021

Sri Ramaswamy Memorial (SRM) Institute of Science and Technology Chennai, India

Bachelor of Technology in Electrical and Electronics Engineering June 2014 - July 2018

SKILLS

Tools: vLLM, NVIDIA Nsight Systems, Nsight Compute, nvprof, Synopsys VCS, Design Compiler, Cadence Xcelium, Genus, Innovus, AMD Vivado, Intel Advisor, Quartus Prime

Programming Languages: HDLs (SystemVerilog/Verilog), Software (Python/C++/Tcl), ML (PyTorch).

Spoken Languages: English (Fluent), Hindi (Fluent), Tamil (Fluent), Japanese (Basic)

WORK EXPERIENCE

Samsung Semiconductor Inc. San Jose, CA

1. *Artificial Intelligence (AI) Research Scientist Intern (Offer Accepted)* May 2026 – August 2026 (Expected)
   - Incoming AI Research Scientist Intern – profiling and optimizing large-scale LLM inference across heterogeneous CPU–GPU clusters and identifying system architecture bottlenecks.
2. *AI Characterization & Tight Coupling Analysis Intern* June 2024 – August 2024
   - Created SKIP, a novel PyTorch-based profiling tool for analyzing operator-kernel dynamics for large language model (LLM) inference workloads and demonstrated that GH200 incurs 2.8x more prefill latency over traditional PCIe-connected Intel x86 CPU - H100 GPU.
   - Demonstrated that GH200 exhibits a 4x larger CPU-bounded region for the same workload over PCIe-connected CPU-GPU systems due to Grace CPU inefficiencies.
   - Led group of 5 researchers on a joint CMU-Samsung paper submission and acceptance in ISPASS 2025.

MediaTek USA Inc. San Jose, CA

1. *AI Architecture & Algorithm Intern (Part-Time; Remote)* August 2022 – December 2022
   - Devised *Tub*GEMM (ISVLSI'23), an ultra low-power hybrid temporal-unary-binary general matrix multiplication (GEMM) unit, and OzMAC (VLSI-SoC'24), a novel bit-serial MAC unit for edge AI inference using TSMC N5 process technology.
2. *AI Architecture & Algorithm Intern* June 2022 – September 2022
   - Enhanced a UVM verification framework for a novel DLA by architecting and implementing new test sequences and updating deprecated source code to support new functional units.

PHD RESEARCH

Carnegie Mellon University Pittsburgh, PA

- Collaborated with 10+ researchers in CMU-NCAL, CMU-ACTL, and UCF-UNARY on research projects spanning unary computing for artificial intelligence, deep learning (DL) architecture designs, and neuromorphic computing.
- Directed a team of 9 undergraduate and graduate research assistants in designing unary-based deep learning architectures (DLAs), contributing to development of design methodologies and automation flows.
- Investigated performance bottlenecks in large language model inference on cutting-edge hardware (H100 GPUs, GH200), driving optimizations that enhanced performance (project funded by Samsung Semiconductor Inc.).
- Designed Tempus Core [DATE 2025], an INT8 temporal-unary convolution core for NVDLA, reducing area/power by 53%/44% and achieving 5x iso-area throughput.
- Developed TNNGen [ISCAS'24, TCAS-II'24], for translation of Temporal Neural Networks (TNNs) from PyTorch models to post-layout netlist, showcasing automated design generation for 7 different modalities and accelerating design process.
- Devised TNN7 (ISVLSI'22), a custom predictive 7nm open-source PDK for TNNs (as extension to ASAP7) consisting of nine custom hard macros, achieving 14% power, 16% delay, 28% area, and 45% EDP reduction over baseline design.

RESEARCH PUBLICATIONS

1. Price, D., **Vellaisamy, P.**, Shen, JP., Wu, D., "Mugi: Value Level Parallelism for Efficient LLMs." [ASPLOS 2026].
2. Lister, D., **Vellaisamy, P.**, Shen, JP, Wu, D. "Catwalk: Unary Top-K for Efficient Ramp-No-Leak Neuron Design for Temporal Neural Networks." [**Best Paper Award,** ISVLSI 2025].
3. **Vellaisamy P.**, Labonte, T., Chakraborty, S., Turner, M., Sury, S., and Shen, JP "Characterizing and Optimizing LLM Inference Workloads on CPU-GPU Coupled Architectures." [ISPASS 2025].

4. **Vellaisamy, P.**, Nair, H., Kang, T., Ni, Y., Fan, H., Qi, B., Hung, H.F., Chen, J., Blanton, RDS., and Shen, JP. "Tempus Core: Area-Power Efficient Temporal-Unary Convolution Core for Low-Precision Edge DLAs." [DATE 2025].
5. Nair, H., **Vellaisamy, P.**, Lin, TH., Wang, P., Blanton, RDS., and Shen, JP. "OzMAC: An Energy-Efficient Sparsity-Exploiting Multiply-Accumulate-Unit Design for DL Inference." [VLSI-SoC 2024].
6. **Vellaisamy, P.**, Nair, H., Wu, D., Blanton, RDS., and Shen, JP. "Exploration of Unary Arithmetic-Based Matrix Multiply Units for Low Precision DL Accelerators." [ISVLSI 2024].
7. Venkatachelam, S., Nair, H., **Vellaisamy, P.**, Zhou, Y., Youssfi, Z., and Shen, JP. "Realtime Person Identification via Gait Analysis using IMU Sensors on Edge Devices" [ICONS 2024].
8. **Vellaisamy, P.**, Nair, H., Gupta, D., Ratnakaram V., and Shen, JP. "TNNGen: Automated Design of Neuromorphic Sensory Processing Units for Time-Series Clustering." [ISCAS 2024, Selected and published in TCAS-II 2024].
9. **Vellaisamy, P.**, Nair, H., Finn, J., Trivedi, M., Chen, A., Li, A., Lin, TH., Wang, P., Blanton, RDS., and Shen, JP. "tubGEMM: Energy-Efficient and Sparsity-Effective Temporal-Unary-Binary Based Matrix Multiply Unit." [ISVLSI 2023].
10. Nair, H., **Vellaisamy, P.**, Chen, A., Finn, J., Li, A., Trivedi, M., and Shen, JP. "tuGEMM: Area-Power-Efficient Temporal Unary GEMM Architecture for Low Resolution Edge AI." [ISCAS 2023].
11. Nair, H., **Vellaisamy, P.**, Bhasuthkar, S. and Shen, JP. "TNN7: A Custom Macro Suite for Implementing Highly Optimized Designs of Neuromorphic TNNs." [ISVLSI 2022].

PRESENTATIONS

1. Vellaisamy, Prabhu. “Characterizing and Optimizing LLM Inference Workloads on CPU-GPU Coupled Architectures.” Jülich Supercomputing Center (JSC), Jülich, Germany (Remote), 20 May 2025. Invited Talk.
2. Vellaisamy, Prabhu. “Characterizing and Optimizing LLM Inference Workloads on CPU-GPU Coupled Architectures.” ISPASS 2025, Ghent, Belgium, 12 May 2025. Conference Presentation.
3. Vellaisamy, Prabhu. “Tempus Core: Area-Power Efficient Temporal-Unary Convolution Core for Low-Precision Edge DLAs.” DATE 2025, Lyon, France, 1 April 2025. Conference Presentation.
4. Vellaisamy, Prabhu. “OzMAC: An Energy-Efficient Sparsity-Exploiting Multiply-Accumulate-Unit Design for DL Inference.” VLSI-SoC 2024, Tangier, Morocco, 7 October 2024. Conference Presentation.
5. Vellaisamy, Prabhu. “Exploration of Unary Arithmetic-Based Matrix Multiply Units for Low Precision DL Accelerators” ISVLSI 2024, Knoxville, TN., 2 July 2024. Conference Presentation.
6. Vellaisamy, Prabhu. “TNNGen: Automated Design of Neuromorphic Sensory Processing Units for Time-Series Clustering” ISCAS 2024, Singapore, 21 May 2024. Conference Presentation.

WORKSHOP PAPERS

1. **Vellaisamy, P.**, Nair, H., Wu, D., and Shen, JP., “Exploration of Unary Based GEMM designs for Conventional AI/DL Accelerators.” 2nd Workshop on Unary Computing (WUC), ASPLOS 2024.
2. Xi, Q., **Vellaisamy, P.**, and Wu, D., “xBrain: Brain-Like Computing for Explainable Brain-Computer Interfaces.” Young Architect Workshop (YArch), ASPLOS 2024.
3. **Vellaisamy, P.**, and Shen, JP. “Towards a Design Framework for TNN-Based Neuromorphic Sensory Processing Units.” Young Architect Workshop (YArch), ASPLOS 2022.

FELLOWSHIPS & AWARDS

- Amar Mukherjee Best Paper Award of ISVLSI 2025
- ISVLSI 2024 Travel Grant
- **2023 Qualcomm Innovation Fellowship - North America Winner**
- Carnegie Institute of Technology's Dean Fellow
- CMU GSA Conference Grant
- DAC 2022 Young Fellow
- Young Architect 2022, ASPLOS

TEACHING EXPERIENCE

- Hardware Arithmetic for Machine Learning (18-340/640) – Fall 2025, Fall 2024, Fall 2023, Fall 2021.
- Neuromorphic Computer Architecture and Processor Design (18-743) – Spring 2025, Spring 2024, Spring 2023, Spring 2022, Spring 2021.
- Modern Computer Architecture (18-740) – Fall 2022.

RELEVANT COURSEWORK

Large Language Models: Methods and Applications (Current), Neuromorphic Computer Architecture, Modern Computer Architecture, Introduction to Machine Learning, Hardware Arithmetic for Machine Learning, Introduction to Embedded Deep Learning, Advanced Digital Integrated Circuit Design, Applied Cryptography, Fundamentals of Computational Biology.

EXTRACURRICULAR ACTIVITIES

Honor societies: IEEE-Eta Kappa Nu, Sigma Xi