Prabhu Vellaisamy

pvellais@andrew.cmu.edu | (609) 423-3147 | LinkedIn | GitHub | Google Scholar | ORCID | Web of Science

EDUCATION

Carnegie Mellon University — Pittsburgh, PA
Doctor of Philosophy in Electrical and Computer Engineering — Sep 2021 – Dec 2026 (Expected)

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Electrical and Computer Engineering — Jan 2020 – May 2021

Sri Ramaswamy Memorial (SRM) Institute of Science and Technology — Chennai, India
Bachelor of Technology in Electrical and Electronics Engineering — Jun 2014 – Jul 2018

RESEARCH INTERESTS

LLM inference optimization · Deep learning accelerators · Neuromorphic computing · Unary computing · VLSI/ASIC design · Hardware-software co-design · CPU-GPU coupled architectures · Edge AI

ACADEMIC AND INDUSTRY EXPERIENCE

Samsung Semiconductor Inc. — San Jose, CA
Artificial Intelligence (AI) Research Scientist Intern (Offer Accepted) — Jun 2026 – Sep 2026

NVIDIA Corporation — Santa Clara, CA
Silicon Solution Engineering Intern (Offer Accepted) — Mar 2026 – Jun 2026

Samsung Semiconductor Inc. — San Jose, CA
AI Characterization & Tight Coupling Analysis Intern — Jun 2024 – Aug 2024

- Built SKIP, a PyTorch Profiler framework that uncovered critical LLM inference bottlenecks, revealing GH200 suffers 2.8x higher prefill latency and 4x larger CPU-bounded regions vs. Intel x86+H100/AMD x86+A100.
- Spearheaded 5-person CMU-Samsung research collaboration from concept to publication; first-authored paper accepted at ISPASS 2025.

MediaTek USA Inc. — San Jose, CA
AI Architecture & Algorithm Intern — Jun 2022 – Dec 2022

- Developed TubGEMM (ISVLSI'23) and OzMAC (VLSI-SoC'24), novel compute units for edge AI that reduced power consumption by 40%+ while maintaining throughput on TSMC N5 process.

PHD RESEARCH

Carnegie Mellon University — Pittsburgh, PA
Advisor: Prof. J.P. Shen, Prof. S. Blanton | *CMU NCAL, CMU ACTL Research Groups*

- Collaborated with research team across 4 research groups (CMUNCAL, CMU-ACTL, UCF-UNARY, NEXUS), delivering 12 peer-reviewed publications and mentoring 15+ graduate and undergraduate students.

- Discovered and characterized performance bottlenecks in LLM inference on GH200 through systematic profiling of model configurations - research funded by Samsung Semiconductor ($100K+ project funding).
- Architected Tempus Core, an INT8 temporal-unary convolution accelerator that achieved 53% area reduction, 44% power savings, and 5x iso-area throughput improvement over baseline NVDLA convolution core - published at DATE 2025.
- Created TNNGen, an end-to-end automation framework that compiles PyTorch models to layout-ready netlists in validated across 7 modalities - published in ISCAS'24 and selected for TCAS-II'24 publication.
- Developed TNN7, an open-source set of 9 optimized macros for 7nm PDK extension to ASAP7 with 9 optimized hard macros that reduced energy-delay product (EDP) by 45% against baseline design.

PUBLICATIONS

Peer-Reviewed Publications

1. **Vellaisamy, P.**, Tripathi, S., Natarajan, V., Thenarasu, S. "TaxBreak: Unmasking the Hidden Costs of LLM Inference Through Overhead Decomposition." *ISPASS 2026* [Accepted].
2. Price, D., **Vellaisamy, P.**, Shen, J.P., Wu, D. "Mugi: Value Level Parallelism for Efficient LLMs." *ASPLOS 2026*.
3. Lister, D., **Vellaisamy, P.**, Shen, J.P., Wu, D. "Catwalk: Unary Top-K for Efficient Ramp-No-Leak Neuron Design for Temporal Neural Networks." *ISVLSI 2025*. [**Amar Mukherjee Best Paper Award**].
4. **Vellaisamy, P.**, Labonte, T., Chakraborty, S., Turner, M., Sury, S., Shen, J.P. "Characterizing and Optimizing LLM Inference Workloads on CPU-GPU Coupled Architectures." *ISPASS 2025*.
5. **Vellaisamy, P.**, Nair, H., Kang, T., Ni, Y., Fan, H., Qi, B., Hung, H.F., Chen, J., Blanton, R.D.S., Shen, J.P. "Tempus Core: Area-Power Efficient Temporal-Unary Convolution Core for Low-Precision Edge DLAs." *DATE 2025*.
6. Nair, H., **Vellaisamy, P.**, Lin, T.H., Wang, P., Blanton, R.D.S., Shen, J.P. "OzMAC: An Energy-Efficient Sparsity-Exploiting Multiply-Accumulate-Unit Design for DL Inference." *IEEE VLSI-SoC 2024*.
7. **Vellaisamy, P.**, Nair, H., Wu, D., Blanton, R.D.S., Shen, J.P. "Exploration of Unary Arithmetic-Based Matrix Multiply Units for Low Precision DL Accelerators." *ISVLSI 2024*.
8. Venkatachelam, S., Nair, H., **Vellaisamy, P.**, Zhou, Y., Youssfi, Z., Shen, J.P. "Realtime Person Identification via Gait Analysis using IMU Sensors on Edge Devices." *ICONS 2024*.
9. **Vellaisamy, P.**, Nair, H., Gupta, D., Ratnakaram, V., Shen, J.P. "TNNGen: Automated Design of Neuromorphic Sensory Processing Units for Time-Series Clustering." *ISCAS 2024* and *TCAS-II 2024*.
10. **Vellaisamy, P.**, Nair, H., Finn, J., Trivedi, M., Chen, A., Li, A., Lin, T.H., Wang, P., Blanton, R.D.S., Shen, J.P. "tubGEMM: Energy-Efficient and Sparsity-Effective Temporal-Unary-Binary Based Matrix Multiply Unit." *ISVLSI 2023*.

11. Nair, H., **Vellaisamy, P.**, Chen, A., Finn, J., Li, A., Trivedi, M., Shen, J.P. "tuGEMM: Area-Power-Efficient Temporal Unary GEMM Architecture for Low Resolution Edge AI." *ISCAS 2023*.
12. Nair, H., **Vellaisamy, P.**, Bhasuthkar, S., Shen, J.P. "TNN7: A Custom Macro Suite for Implementing Highly Optimized Designs of Neuromorphic TNNs." *ISVLSI 2022*.

Workshop Papers (5)

1. Price, D., **Vellaisamy, P.**, Shen, J.P., Wu, D. "Mugi: Value Level Parallelism For Nonlinear Operations in LLMs." *Workshop on Unary Computing (WUC), ASPLOS 2026*.
2. Price, D., **Vellaisamy, P.**, Shen, J.P., Wu, D. "Agraph: A Unified Graph Representation for At-Will Simulation of Emerging Stacks." *Workshop on Unary Computing (WUC), ASPLOS 2026*.
3. **Vellaisamy, P.**, Nair, H., Wu, D., Shen, J.P. "Exploration of Unary Based GEMM Designs for Conventional AI/DL Accelerators." *2nd Workshop on Unary Computing (WUC), ASPLOS 2024*.
4. Xi, Q., **Vellaisamy, P.**, Wu, D. "xBrain: Brain-Like Computing for Explainable Brain-Computer Interfaces." *Young Architect Workshop (YArch), ASPLOS 2024*.
5. **Vellaisamy, P.**, Shen, J.P. "Towards a Design Framework for TNN-Based Neuromorphic Sensory Processing Units." *Young Architect Workshop (YArch), ASPLOS 2022*.

PRESENTATIONS

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

FELLOWSHIPS, AWARDS, AND HONORS

- Amar Mukherjee Best Paper Award, ISVLSI 2025.
- ISVLSI 2024 Travel Grant.
- CMU GSA Conference Grant.
- Qualcomm Innovation Fellowship 2023.
- DAC 2022 Young Fellow.

- Young Architect 2022, ASPLOS.
- Carnegie Institute of Technology Dean's Fellowship.

TEACHING EXPERIENCE

Carnegie Mellon University Pittsburgh, PA

Department of Electrical and Computer Engineering

1. **18-340/640: Hardware Arithmetic for Machine Learning**

   Teaching Instructor 4 semesters; approximately 50 students per semester

2. **18-743: Neuromorphic Computer Architecture and Processor Design**

   Teaching Instructor 5 semesters; approximately 20 graduate students per semester

3. **18-740: Modern Computer Architecture**

   Teaching Instructor 1 semester; approximately 100 students

TECHNICAL SKILLS

- **Tools:** vLLM, TensorRT, NVIDIA Nsight Systems, Nsight Compute, nvprof, Synopsys VCS, Design Compiler, Cadence Xcelium, Genus, Innovus, AMD Vivado, Intel Quartus Prime.
- **Programming:** Python, PyTorch, SystemVerilog, Verilog, C++, Tcl.
- Architectures **Languages:** English (Fluent), Hindi (Fluent), Tamil (Fluent), Japanese (Basic).

PROFESSIONAL SERVICE

Peer Reviewer

- *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems* (IEEE TVLSI).
- *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits* (IEEE JXCDC).

PROFESSIONAL MEMBERSHIPS AND HONOR SOCIETIES

- IEEE-Eta Kappa Nu (HKN).
- Sigma Xi Scientific Research Honor Society.

RELEVANT COURSEWORK

Large Language Models: Methods and Applications · 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