Vu Le

211, Lederle Graduate Research Center, 740 N Pleasant St, Amherst, MA 01003
: vdle@umass.edu ◇ : vule20 ◇ : vule20 ◇ : Scholar

EDUCATION

University of Massachusetts Amherst *Sep 2024 – May 2029 (expected)*
Doctor of Philosophy in Computer Science

Vietnam National University, Hanoi (VNU) *Aug 2018 – Jun 2022*
Bachelor of Engineering in Mechatronics Engineering

RESEARCH EXPERIENCE

University of Massachusetts Amherst Aug 2024 – present
Ph.D. Student — Research Amherst, MA, USA

Co-advised by Prof. Deepak Ganesan and Prof. VP Nguyen

- **Project 1: Hardware-Aware Recurrent Inference for Real-Time Qubit Readout**
  - Conventional discriminators integrate the I/Q readout record into a single point, discarding the evidence a relaxing qubit leaves mid-readout; designed a hardware-aware LSTM accelerator on AMD Versal (AIE+FPGA) that preserves full temporal resolution within a 1μs real-time budget
  - On the qubit most affected by mid-readout relaxation, recovered full-resolution feed-forward accuracy (96.59% vs. 96.37%) at 403× fewer parameters and ∼103× fewer DSPs, cutting classification error by 68.2% over GMM, 43.6% over QubiCML, and 16.82% over MF-NN
  - Deployed on just 4 of 400 AIE tiles and 17 DSP58 blocks, 2.2–8.2× fewer DSP58 than QubiCML & MF-NN per qubit, at 854 ns latency (vs. 1,082 ns pure-AIE) and 96.78% mean accuracy across eight qubits, the strongest hardware-deployed operating point reported; submitted to ASPLOS
- **Project 2: Heterogeneous Computing for Emerging Physical AI Workloads (ongoing)**
  - Developing a hardware-aware method to map AI models onto heterogeneous compute blocks (AI Engines, FPGAs, CPUs) with guaranteed timing and minimal resource use, for compute & memory bound workloads such as real-time physical AI (VLAs) and scientific computing

Lawrence Berkeley National Laboratory — ATAP Division Dec 2024 – present
Research Affiliate — hosted by Dr. Yilun Xu Berkeley, CA, USA

- Serves as a Student Affiliate, bridging UMass spatial-hardware co-design research with LBNL's superconducting quantum control hardware and readout infrastructure (see project above)

George Mason University Sep 2023 – Dec 2023
Research Collaborator — advised by Prof. Bo Han Fairfax, VA, USA

- Investigated bandwidth- and compute-efficient neural avatar rendering for immersive telepresence; proposed patch-aware data transmission and edge-assisted patch-based inference techniques, leading to a co-authored publication at ACM SenSys 2024

Phenikaa Group Jul 2022 – Oct 2023
Research Engineer / Associate Hanoi, Vietnam

- Developed a quantized CNN pipeline for real-time solar-panel defect detection from drone IR imagery, deployed on Jetson Nano at 1.04 ms/image (∼960 FPS) with 85.35% accuracy across 11 anomaly types, using HPC-scale distributed training to accelerate the design cycle
- Developed and characterized a real-time hardware-accelerated computer vision pipeline, mapping multi-stream object detection models to a Xilinx FPGA DPU to satisfy latency constraints

Center for IC Design and Applications, VNU ITI Aug 2020 – Jun 2022
Research Intern Hanoi, Vietnam

- Investigated hardware-accelerated reinforcement learning inference: implemented a Deep Q-Network (DQN) in Chisel HDL and deployed on FPGA
- Designed a hardware camera interface for handwritten digit recognition as part of a Toshiba-sponsored project; wrote bare-metal C kernels, used the SiFive Freedom-E-SDK to boot the RISC-V chip, and integrated the full system into the Chipyard SoC framework using UC Berkeley's Rocket Chip Generator; performed remote FPGA debugging

PEER-REVIEWED PUBLICATIONS

1. **Vu Le**, Neel Vora, Devanshu Brahmbhatt, Yilun Xu, Gang Huang, and Phuc Nguyen. "Computing Systems for Superconducting Qubits: Challenges and Opportunities." In *QSys '25: Workshop on Quantum Systems, ACM MobiSys*, 2025. DOI: 10.1145/3711875.3737657
2. Tasnim Azad Abir, **Vu Le**, Endrowednes Kuantama, Pranjol Sen Gupta, Austin Copley, Judith Dawes, Mohammad Islam, Richard Han, and Phuc Nguyen. "Detection and Tracking of Drone Swarms using LiDAR." In *Proceedings of the 23rd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2025.
3. Ruizhi Cheng, Nan Wu, **Vu Le**, Eugene Chai, Matteo Varvello, and Bo Han. "MagicStream: Bandwidth-conserving Immersive Telepresence via Semantic Communication." In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pp. 365–379, 2024.
4. Hai Phan, Cindy Le, **Vu Le**, Yihui He, and Anh Totti Nguyen. "Fast and Interpretable Face Identification for Out-Of-Distribution Data Using Vision Transformers." In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

PREPRINTS & MANUSCRIPTS UNDER REVIEW

1. **Vu Le**, Yilun Xu, Neel Vora, Gang Huang, Deepak Ganesan, and VP Nguyen. "Hardware-Aware Recurrent Inference for Real-Time Qubit Readout." Manuscript under review, ASPLOS 2027.

TALKS & PRESENTATIONS

1. **Vu Le**. "Hardware-Software Co-Design with Spatial Accelerators and FPGA for Real-Time Qubit Readout." Presented to LBNL and AMD. Apr 2026, Berkeley, CA, USA.
2. **Vu Le**. "Opportunities in Computer Systems Research for Quantum Computing." *QSys '25: Workshop on Quantum Systems, ACM MobiSys.* Jun 2025. Anaheim, CA, USA.

TECHNICAL SKILLS

- **Spatial & FPGA Architecture** – Kernel development, tile placement and scheduling, vector intrinsics, cycle-accurate profiling, HLS; spatial/dataflow AI accelerators, FPGA (Vivado)
- **Machine Learning & Co-Design** – Hardware-software-algorithm co-design for constrained inference, model profiling; PyTorch; HPC (distributed multi-node training)
- **Programming** – Python, C++, low-level C, Bash, Chisel, Verilog
- **Tools** – Linux, Git, Docker, LaTeX

AWARDS & FELLOWSHIPS

- James Kurose Scholarship in Computer Science, University of Massachusetts Amherst, 2025
- Best Paper Runner-up Award, ACM SenSys 2024
- Fighting Spirit Prize, 25th LSI Design Contest, Okinawa, Japan, 2022