

Vu Le

211, Lederle Graduate Research Center, 740 N Pleasant St, Amherst, MA 01003

✉: vdle@umass.edu ◇ [in](https://www.linkedin.com/in/vule20): [vule20](#) ◇ [vule20](#) ◇ [Scholar](#) ◇ vule.us

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 Electronics 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**

- Designed a hardware-aware recurrent ML accelerator on a heterogeneous spatial architecture (AI Engine array + programmable logic) for latency-critical real-time inference
- Partitioned model computation across independent parallel and sequential dependency paths to meet a sub-microsecond latency budget
- Beats prior SOTA hardware-deployed ML classifiers at a fraction of the hardware cost

- **Project 2: Heterogeneous Computing for Emerging Physical AI Workloads (ongoing)**

- Developing a hardware-grounded oracle that predicts end-to-end latency for a given schedule, enabling both pre-implementation latency estimation and configuration search under a target latency and resource budget, for a large spatial AI Engine array and FPGA
- Reduced [FlashAttention](#) latency by **16–27%** and [LayerNorm](#) by **40–52%** vs. the reference kernels from AMD on XDNA2 NPU (32 tiles), together with a new schedule, it cuts end-to-end latency for [SmolVLA's vision encoder](#) by **11.4%**.

Lawrence Berkeley National Laboratory — ATAP Division

Dec 2024 – Sep 2026

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 an edge-assisted patch-based neural rendering technique that accelerated inference by $5.8\times$ (6.4 to 37 FPS) while preserving visual quality (SSIM 0.92), contributing to a system achieving up to $1195\times$ bandwidth reduction over state-of-the-art, in a co-authored paper at ACM SenSys

Phenikaa Group

Jul 2022 – Oct 2023

Research Associate / Engineer

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

- Designed and implemented the camera-to-NPU pipeline integration, including MIPI camera capture, for a multi-stream object detection system mapped to a Xilinx FPGA DPU on a Kria evaluation board, meeting the real-time latency requirement for a self-parking application

Center for IC Design and Applications, VNU ITI
Research Intern

Aug 2020 – Jun 2022
 Hanoi, Vietnam

- Built a RISC-V SoC from silicon up for a Toshiba-sponsored embedded vision project – custom hardware camera interface, bare-metal C firmware, and full Chipyard/Rocket-Chip integration; also designed a custom hardware array for accelerating DQN inference in Chisel HDL.

PEER-REVIEWED PUBLICATIONS

1. **Vu Le**, Neel Vora, Devanshu Brahmabhatt, 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](https://doi.org/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.

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 & Hardware Architecture** – Kernel development, scheduling, vector intrinsics, HLS, NPU, MLIR-AIE, register-allocation analysis, FPGA (Vivado)
- **Machine Learning & Co-Design** – Hardware-software-algorithm co-design for constrained inference, model profiling; PyTorch
- **Programming** – Python, C++, C, Bash, Chisel, Verilog, CUDA
- **Tools** – HPC, Linux, Git, Docker, L^AT_EX

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