

Vu Le

Amherst, MA, USA

✉: vdle@umass.edu ◊ : [vule20](#) ◊ : [vule20](#) ◊ : Scholar

EDUCATION

University of Massachusetts Amherst — Ph.D. Computer Science

Sep 2024 – May 2029

Teaching Assistant, COMPSCI 335 Computer Architecture, UMass Amherst

EXPERIENCE

University of Massachusetts Amherst

Aug 2024 – present

Ph.D. Student

Amherst, MA, USA

Co-advised by [Prof. Deepak Ganesan](#) and [Prof. VP Nguyen](#); also a Research Affiliate at Lawrence Berkeley National Laboratory, hosted by [Dr. Yilun Xu](#)

- Developing a hardware abstraction and analytical latency model for spatial AI accelerators that predicts latency for arbitrary matrix sizes and sequence lengths, and selects tilings, placements, and schedules that minimize end-to-end latency under latency and resource budgets
- Currently validating on the AMD XDNA2 NPU with SmolVLA: model-guided schedules reach **195 ms** end to end (latest, unpublished), **4.0×** faster than Ryzen AI Software, **2.1×** than FastFlowLM, and **1.9×** than MLIR-AIR main ([public](#): **54.8%** and **12.5%** lower end to end than Ryzen AI SW and FastFlowLM; **52.8%** lower vision latency than the original MLIR-AIR); resulting high performance kernels and compiler support [merged](#) into AMD's MLIR-AIR
- Designed a hardware-aware LSTM accelerator on AMD Versal (AIE+FPGA) for real-time qubit readout (submitted, ASPLOS 2027), diagnosing a register-spill-induced pipeline stall (minimum II: 61→103 cycles) via LLVM register-allocator analysis and resolving it by offloading the fused-gate computation to FPGA at its dependency boundary, cutting exposed tail latency by **60.3%** vs. pure AIE; achieved the highest accuracy among real-time hardware SOTA, cutting relative error by **68.2%** over GMM, **43.6%** over QubiCML, and **16.82%** over MF-NN, at **854 ns** latency, with **2.2–8.2×** lower DSP58 usage than QubiCML/MF-NN

Phenikaa University / Phenikaa-X

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; also built a 3-camera streaming pipeline (MIPI to VPU to DPU) on a Xilinx Kria KV260 for a self-parking system, including model quantization and deployment to the DPU

Center for IC Design and Applications, VNU ITI

Aug 2020 – Jun 2022

Research Intern

Hanoi, Vietnam

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

TECHNICAL SKILLS

- **ML Performance & Architecture** – Low-latency kernel optimization, analytical latency modeling, tiling and scheduling on spatial accelerators (AI Engine, NPU), accelerator compilers (MLIR), FPGA design and HW/SW partitioning (Vivado, HLS, Verilog, Chisel), RISC-V SoC bring-up
- **Programming & Tools** – C++, Python, C, CUDA, PyTorch, AI coding agents, hardware profiling, Git, Docker, Linux

PUBLICATIONS & AWARDS

6 peer-reviewed papers, including ACM MobiSys 2025, SenSys 2024, and WACV 2023; 1 manuscript under review (ASPLOS 2027) – : Scholar

James Kurose Scholarship in Computer Science, UMass Amherst, 2025 · Best Paper Runner-up, ACM SenSys 2024 · Fighting Spirit Prize, 25th LSI Design Contest, Okinawa, 2022