

Vu Le

Amherst, MA, USA

✉: vdle@umass.edu ♦ [in](https://www.linkedin.com/in/vule20): vule20 ♦ [G](https://github.com/vule20): vule20 ♦ [🎓](#): Scholar

EDUCATION

University of Massachusetts Amherst — Ph.D. Computer Science
Teaching Assistant, COMPSCI 335 Computer Architecture, UMass Amherst

Sep 2024 – May 2029

EXPERIENCE

University of Massachusetts Amherst

Ph.D. Student

Aug 2024 – present

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](#)

- 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
- Developing a hardware-grounded oracle that predicts latency for arbitrary matrix sizes and sequence lengths and searches for the schedule with the lowest end-to-end latency under a target latency and resource budget for a large spatial AI Engine array and FPGA like Versal AI
- Validated it on the XDNA2 NPU with hardware-aware GEMM and FlashAttention kernels, [merged](#) into AMD's MLIR-AIR compiler for SmolVLA, and a tile-selection policy; together they cut vision-stage latency by **53.0%** and end-to-end latency by **43.0%** over AMD's original flow, and by **42.9%** and **55.3%**, respectively, over AMD's Ryzen AI Software

Phenikaa University / Phenikaa-X

Research Associate / Engineer

Jul 2022 – Oct 2023

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

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 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

- **Hardware Systems & Architecture** – FPGA (Vivado, HLS), AMD Versal AIE/NPU, system integration and bring-up (multi-camera MIPI-to-DPU pipeline on Kria KV260, Jetson), embedded firmware, MLIR, hardware and performance modeling, Chisel, Verilog
- **Programming & Tools** – Python, C++, C, Bash, CUDA, PyTorch, HPC, Git, Docker, Linux, $\text{\LaTeX}$

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