Martin Nguyen

(+1) 872-294-1416 | manhntm3@gmail.com | Linkedin: manhntm3 | Github: manhntm3 | Chicago, IL

A dedicated and results-driven software engineer with a system-oriented mindset and expertise in performance engineering. Knowledgeable about machine learning research areas. Experienced in lock-free concurrency, event-driven architectures, kernel-level networking (eBPF), low-level optimization of GPU model inference, and GPU memory models.

Education

The University of Illinois at Chicago
Master of Science in Computer Science, GPA: 3.9/4.0 — *Aug. 2024 – May. 2026*

FPT University
Bachelor of Computer Science, Full-ride scholarships — *Aug. 2017 – Aug 2021*

Professional Experience

Founding Software Engineer — Jul. 2021 – Aug. 2024
Vizgard Ltd — *London, United Kingdom*

- Contributed to the company's growth from early prototype to securing over **$2.5M in venture funding**.
- Led the design and development of a real-time computer vision system for surveillance and unmanned system automation. Architected an event-driven, lock-free MPMC double buffer pipeline to optimize latency and minimize memory locality issues, achieving 6 simultaneous 1080p camera streams on NVIDIA Jetson and 40 streams across 4 NVIDIA Ada RTX6000.
- Built and optimized pipelines for multiple deep learning models, including object detection(DETR/YOLO), object tracking (SiameseRPN++ and DeepSORT), pose estimation(AlphaPose) and face redaction. All models are trained using PyTorch/TensorFlow on GCP; deployed with TensorRT for high-speed inference.
- Led performance optimization across the stack (profiling, memory locality, batching, async execution), accelerated critical paths by $\sim$**3**$\times$ (measured in Nsight System/Remotery) by moving CPU stages to CUDA kernels without compromising model accuracy .
- Developed a low-latency WebRTC streaming server using GStreamer and Node.js to deliver real-time HD video to browsers and RTSP endpoints.

Software Research Engineer — Jan. 2020 – Jul. 2021
VinAI Research (acquired by Qualcomm Research)

- Designed a two-stage infrared anti-spoofing system, achieving $< 3\%$ false acceptance rate (FAR).
- Improved the face recognition model using knowledge distillation and network optimization, delivering a 4$\times$ speed improvement with equivalent accuracy.

Projects

Weak Memory Behavior on GPUs — *PTX, CUDA, C++*
Investigated weak memory consistency behavior on NVIDIA GPUs using PTX and CUDA, identifying synchronization patterns to improve kernel reliability and performance. (Github: https://github.com/manhntm3/CS554FinalProject)

Dynamic eBPF Firewall for DDoS Mitigation in Rust — *eBPF, Rust, Linux Network*
Developed a Linux kernel module as a dynamic eBPF-based network firewall in Rust to block traffic and mitigate DDoS attacks via real-time IP filtering. (Github: https://github.com/manhntm3/cs594-sp25-ebpf)

Conversational Agent using AWS Bedrock — *EC2, Scala, Go, gRPC, Python, AWS Lambda*
Developed a distributed LLM training pipeline using Apache Spark and implemented both RESTful and gRPC servers for cloud integration. Deployed services on AWS EC2 to route requests to AWS Lambda and Bedrock, powering an agent built on LLaMA. (Github: https://github.com/manhntm3/ConversationAgent)

Skills Summary

Programming Languages:	C/C++, Python, Java, JavaScript, Scala
DL Accelerators & Inference:	CUDA, TensorRT, Triton, ONNX, GPU/TPU
ML/DL Framework:	PyTorch, JAX, TensorFlow, Scikit-Learn
Systems:	Linux, eBPF/XDP, memory models, perf/Remotery/Nsight profiling
Generative AI / Multimodal:	LLMs, VLMs, Transformers, HuggingFace, LoRA/QLoRA, RAG
Databases & Big Data:	MySQL, MongoDB, Redis, Spark, Hadoop
MLOps:	Docker, Kubernetes, Jenkins, AWS, GCP, Azure