ZHILIN TANG

Chicago, IL +1 (312) 919-9003 tangzhilinnz@gmail.com
linkedin.com/in/tangzhilin tangzhilinnz.github.io

SUMMARY

Software Engineer with 5+ years of industry experience specializing in Real-Time Systems and GPU Programming. Deep expertise in C/C++, CUDA, Python, Triton, and GPU Shader programming. Skilled in optimizing Deep Learning algorithms and leveraging GPU architecture for maximum performance. Strong background in software design patterns, multithreading optimization, and large-scale software architecture.

PROFESSIONAL EXPERIENCE

C/C++ Software Engineer — Jan 2021 – Aug 2024

Chengdu Denglin Technology Co., Ltd (AI Computing Semiconductor) — *Chengdu, China*

- **PyTorch Integration (System Architecture):** Spearheaded the end-to-end integration of the PyTorch ecosystem onto proprietary hardware; resolved complex compilation, linking, and ABI compatibility challenges to enable native execution of AI models, directly expanding the company's compatible model library by 50+ models.
- **Kernel Optimization (Performance):** Engineered high-performance C++/CUDA implementations of Deep Learning operators (e.g., Attention, Convolution, GEMM); utilized Nsight Compute to diagnose bottlenecks and drive optimizations in memory coalescing and thread parallelism, resulting in a 60% reduction in inference latency and a 1.5x increase in training throughput.
- **Testing Infrastructure (Quality):** Architected a hybrid C++/Python unit test framework for CUDA-compatible libraries (cuDNN/cuBLAS APIs); automated the validation of 100+ distinct operators, reducing regression testing cycles from 2 days to 3 hours and substantially improving release stability.
- **Cross-Functional Debugging:** Partnered with hardware architects and compiler teams to investigate critical silicon-software interface bugs; diagnosed and patched complex root causes in the AI compiler stack, resolving major blockers that prevented the successful execution of key AI models.

Embedded Software Engineer in Test — May 2020 – Jan 2021

Wind River Software Technology (RTOS & Embedded Systems) — *Chengdu, China*

- **RTOS Compliance & Safety:** Validated and debugged OS-level C APIs for the VxWorks RTOS to meet stringent aviation-grade certification standards; refactored legacy code to resolve compliance violations, ensuring 100% adherence to safety and reliability requirements for mission-critical deployments.
- **Simulation & Verification:** Engineered comprehensive unit test suites using C and Python to verify software features; leveraged Simics for full-system hardware simulation and Wassp for coverage analysis, which accelerated the validation process and guaranteed rigorous adherence to aviation software quality standards.

ACADEMIC RESEARCH & DEVELOPMENT

Graduate Research Assistant (Efficient Transformer Neural Networks) — Jun 2025 – Present

DePaul University Jarvis College of Computing and Digital Media — *Chicago, IL*

- **Algorithmic Innovation:** Developed a novel tree-based hierarchical attention mechanism to address the quadratic bottleneck in standard Transformers; successfully reduced memory complexity to O(n) and computation to O(n log n), enabling the efficient processing of long-sequence data.
- **Model Implementation:** Engineered a robust prototype of the attention mechanism in PyTorch and integrated it into Transformer architectures; validated model efficacy across NLP, classification, and time-series forecasting, proving the algorithm's generalization capabilities and competitive accuracy against standard benchmarks.

- **HPC Acceleration:** Designed and implemented custom CUDA and Triton kernels to accelerate the hierarchical attention mechanism; conducted deep performance profiling and low-level optimization (memory coalescing, shared memory usage, kernel fusion), achieving maximal hardware utilization compared to standard PyTorch implementations.

EDUCATION

DePaul University **Sep 2024 – Jun 2026**
Master of Science in Software Engineering | *Chicago, IL*
GPA: 4.0/4.0 | Software Architecture, C++ Real-Time Systems, AI systems, Deep Learning, CUDA, Triton

Chengdu University of Information Technology **Sep 2006 – Jul 2010**
Bachelor of Engineering in Microelectronics | *China*
Embedded Systems & RTOS, Circuit Design, Digital Signal Processing, Chip Architecture

CERTIFICATION

Waikato Institute of Technology **Apr 2019 – Nov 2019**
Graduate Diploma in Applied Information Technology – Software Engineering
Data Structures & Algorithms, Linear Algebra, Python, C++ Object-Oriented Programming

TECHNICAL SKILLS

Programming Languages:	C, C++, C#, Python, Java, Shell
Graphics APIs:	DirectX, OpenGL
GPU & HPC:	CUDA, Triton, Shader
AI & Machine Learning:	PyTorch, TensorFlow
Build & Tools:	CMake, Make, Ninja, Bazel
Operating Systems:	Windows, Linux, RTOS, VxWorks
Version Control & CI/CD:	Git, GitLab, Perforce, Docker, Jenkins