Po-Chi, Hsu

https://github.com/pchsu-hsupc

Email: hpc880214@gmail.com
Mobile: +886-912-129-381

Education

- **National Yang Ming Chiao Tung University** — Hsinchu, TW
  Master of Science in Computer Science — GPA 4.18 / 4.30 — *Sep 2022 – Now*
- **National Yunlin University of Science and Technology** — Yunlin, TW
  Bachelor of Science in Computer Science — GPA 3.65 / 4.0 — *Sep 2017 – Jun 2022*

Experience

- **Sciwork** — Hsinchu, TW
  Modmesh's Contributor — *April 2025 – Now*
  - **Array Searching Operations (argwhere, where, argmax, argmin)**
    Implemented high-performance C++ multi-dimensional array search primitives, and exposed them to Python via Pybind11, leveraging profiling-driven optimizations to achieve NumPy-comparable performance.
- **Mediatek** — ChuPei, TW
  Software Engineer Intern (Windows Wi-Fi Driver) — *July 2023 – Aug 2023*
  - **Driver Development - Efuse**
    Implemented driver operations for reading and writing efuse, configuring and managing internal circuit parameters. Patched event handling mechanisms to ensure process stability and accuracy.
  - **Driver Development - Sniffer Mode**
    Extracted information from RX descriptors. Incorporated Radiotap and PCAP headers for advanced packet analysis. Established robust packet receiving and transmission workflows in sniffer mode to aid in network monitoring and debugging.
- **National Yang Ming Chiao Tung University** — Hsinchu, TW
  Teaching Assistants — *Sep 2022 – Jun 2024*
  - **Embedded Systems Capstone, IoT Devices and Platforms**
  - **Discrete Mathematics - Tutor foreign student**

Thesis

- **Feature Alignment and Compositional Token for Human Pose Estimation**
  Video-based 2D human pose estimation struggles with motion blur, occlusions, and truncated body parts. While heatmap-based methods use temporal cues, they often miss the joint dependencies, causing unrealistic poses. We resolve this by combining spatiotemporal alignment with token representations to robustly capture motion and joint relations.

Projects

- **Ported a Modern C++ ray-tracing engine to CUDA**
  Refactoring CPU std::vector usage into custom GPU pointer-to-pointer structures and rewriting the recursive pixel-reflection routine as an iterative algorithm to prevent GPU stack overflows. Provides a **12x** speedup over a single-threaded CPU, and **6x** speedup over OpenMP/Pthreads in high-resolution (4K/8K) rendering workflows.
- **Event Management Platform**
  Developed a Flask-based user authentication microservice with RESTful APIs for an event management platform, enabling user registration, login, and logout. Integrated MongoDB for secure and scalable data persistence and Implemented an automated CI pipeline in GitHub Actions that tested and built the application. Implemented unit and integration tests using Python's unittest and JavaScript's Jest to ensure API reliability. Deployed and orchestrated microservices using Docker Swarm for efficient scalability and maintainability.
- **Large Language Model Acceleration**
  On NVIDIA T4 GPUs, we successfully accelerated inference of the LLaMA-3.2 3B model by combining quantization, FlashAttention, KV-cache optimization, and deploying it with a more efficient serving framework, llama.cpp. As a result, throughput improved from approximately 29 tokens/sec to over 73 tokens/sec, achieving a **2.7**× speed-up. Additionally, by incorporating QLoRA fine-tuning, we were able to further reduce perplexity, enhancing model quality.

Skills

- **Courses**
  Physical Design Automation (A+), Parallel Programming (A+), Software Testing (A+), Cloud Native Development (A+), Network Programming (A+), Edge AI
- **Keywords**
  C, C++, Python, CUDA, PyTorch, GDB, Wireshark, Git, WinDbg, GitHub Actions, Valgrind, LLVM, Linux, Unix, Socket, Shell, Makefile, Boost.Asio, OpenMP, Open MPI, OpenCL, Pthread, gprof, Selenium, Flask, RESTful, Jest, Docker, Docker Swarm, MongoDB, OpenCV, OpenGL