Hsien-Cheng Huang (黃咸誠)

ryankert01@gmail.com - linkedin.com/in/ryankert01 - github.com/ryankert01

TECHNICAL SKILLS

Programming Languages: Python, C++, CUDA, Golang, Rust, TS/JS
LLM Serving & Inference: vLLM, SGLang, Triton (GPU kernels), paged attention / KV-cache, continuous batching, FlashAttention, speculative decoding
Distributed & GPU Infra: Ray, KubeRay, GPU scheduling & autoscaling, Docker
ML & Tooling: PyTorch, Gin, PostgreSQL, Git, GitHub Actions, OpenCV

EXPERIENCES

AI Infra Open-Source Contributions - Ray / KubeRay *2025 - 2026*

- **Ray (ray-project/ray) – 10 PRs merged** into the distributed-compute framework underpinning Ray Train / Ray Serve. Fixed a **V2 autoscaler** bug that blocked **scaling GPU worker nodes from zero** and helped deprecate the legacy V1 autoscaler – core to **elastic GPU scheduling**; added map-namespace support to Ray Data's expression engine (+356 LOC).
- **KubeRay – 2 PRs merged:** built end-to-end auth-token test coverage for **RayService** (the LLM-serving CRD that runs serving workloads on **Kubernetes**) and hardened the history-server log collector – hands-on with the Ray-on-Kubernetes serving control plane.

LLM Inference Serving on Tenstorrent NPU - Lab Project *Jan 2026 - present*

- Building a high-throughput LLM serving path on **Tenstorrent NPUs** – bringing up **paged attention** (paged KV-cache), **FlashAttention**-style fused attention, and **continuous (in-flight) batching** to maximize accelerator utilization, porting vLLM/SGLang-style serving techniques to non-GPU hardware.
- Working at the **tt-metal** kernel / tt-transformers layer; upstreamed a prefill-kernel prefix-caching fix to tenstorrent/tt-metal – **3.8x faster wall-time and 4.9x lower mean TTFT** (goodput 42→159 tok/s) on a 95%-cache-hit serving workload.

Speculative Decoding Research - NYCU *2025 - present*

- Researching **speculative decoding** (draft-and-verify) to cut LLM decode latency while preserving output quality – inference-optimization work directly applicable to vLLM / SGLang serving.

Efficient LLM Training OSS - Liger-Kernel (Triton) *May 2024 - Dec 2024*

- Authored **Triton GPU kernels** (FusedLinearCrossEntropy for Mixtral) for LinkedIn's Liger-Kernel, a kernel library that raises **multi-GPU training throughput by 20%** and **cuts memory use by 60%**; adopted by HuggingFace.
- Wrote upstream HuggingFace docs and fixed correctness bugs in the fused cross-entropy path.

Golang Backend Intern *Aug 2024 - Dec 2024*

- Raised API throughput up to 10x by introducing fasthttp and automatic Redis pipelining.
- Brought versioned migrations (goose) and an efficient log pipeline (WebSocket, MQTT) feeding our log-storage system.

Apache Mahout Committer *Nov 2025 - present*

- **Elected committer Jan 2026** on Apache Mahout (top-level ASF project); contributing since 2025 with 62 PRs merged at 91% acceptance.
- Own **CUDA kernels and GPU throughput benchmarking** for Mahout's QDP encoding pipeline – authored IQP encode/batch CUDA kernels with multi-arch (sm86) builds and CUDA/Torch-tensor bindings; benchmarked throughput against PennyLane and AMD-GPU baselines, and an FWT kernel optimization cut encoding cost from $O(4^n)$ to $O(n \cdot 2^n)$ (up to 53x faster).

ACHIEVEMENTS

- **English Proficiency: C1 Level (CEFR - Advanced/Proficient User)** - demonstrated by TOEIC score of 930/990
- **Silver Award - 2023 ICPC Taiwan Private University Programming Contest**, Won Silver Medals in a coding competition with 200+ participants. (news)
- **CPE (Collegiate Programming Examination)**, score 6/7, ranking: 50 / 2811 (top 1.8%)

EDUCATION

National Yang Ming Chiao Tung University **Hsinchu, Taiwan**

MS, Computer Science and engineering *Aug 2025 - June 2027*
GPA to date: 4.1/4.3

Yuan Ze University **Taoyuan, Taiwan**

BS, Computer Science and Engineering *Aug 2021 - June 2025*