Bing Han

(631) 479-9014
Stony Brook, NY
bingshiunhan@gmail.com
Website | LinkedIn | Github

Education

Ph.D candidate in Computer Science, *Stony Brook University*, GPA 3.89/4.00 — 09.2022 — 05.2027(Anticipated)
Bachelor of Electrical Engineering, *National Taiwan University*, GPA 3.85/4.30 — 09.2015 — 01.2020

Selected publications

- **[SIGMetrics'25]** Energy-efficient GPU SM allocation, **B.Han**, W.Lin, K.Parekh, T.Paul, A.Gandhi, Z.Liu (Workshop)
- **[SoCC'24]** KACE: Kernel-Aware Colocation for Efficient GPU Spatial Sharing, **B.Han**, T.Paul, Z.Liu, A.Gandhi

Professional Experience

Applied Scientist Intern — **05.2025 — 08.2025**
Amazon — *Palo Alto, CA*

- Reduced LLM deployment costs **43%** by selecting cost-efficient AWS GPU instances via latency predictions.
- Found 3× faster LLM parallelism configs in 30s with simulation and one-time device profiling, avoiding GPU benchmarks.

Research Assistant — **07.2023 — Present**
Stony Brook University, Advisor: Prof.Anshul Gandhi, Prof.Zhenhua Liu — *Stony Brook, NY*

- **Project: GPU performance analysis and prediction on DL serving**
- Designed an AI workload-aware colocation strategy using fine-grained GPU kernel profiles from **NVIDIA Nsight Compute**. Reduced total job completion time by 36%.

- **Project: energy-efficient GPU sharing**
- Predicted optimal SM compute isolation ratios for colocated workloads using **NVIDIA MPS** and **DCGM** metrics, improving energy efficiency by **35%**.
- Extended the framework under 60/100/200W power caps, achieving **97%** of optimal throughput across all settings.

- **Project: Optimize DL scheduling with Kubernetes**
- Optimized AI system scheduling and built an end-to-end ML deployment pipeline in **Kubernetes**, enabling efficient resource allocation and **shortest-job-first** scheduling for colocated tasks, improving performance and reducing task completion time by **20%**.

Data Engineer Intern — **12.2018 — 07.2019**
Cathay Financial Holdings — *Taipei, Taiwan*

- Developed scalable machine learning pipelines using **Hadoop**, **Spark**, and **Kafka** microservices, leveraging Docker to ensure efficient distributed computing for high-volume data processing.
- Deployed an **automation pipeline** for configuration tuning, reducing configuration time by 50% in **Proof-of-Concepts**.

Technical sales Intern — **04.2021 — 04.2022**
Intel — *Taipei, Taiwan*

- Led **Xeon E server launch program** in Asia ($300M data center business). Strengthened cross-geographical **market relations** and engaged with 20+ **ODM supply manufacturers** to resolve platform enablement challenges.

Selected projects

Find Yourbike – a shared bike tracking website — **[MongoDB/Flask/Nginx/React/Docker]**
Cloud Computing and Cyber Security — *Taipei, Taiwan*

- Accomplished **full-stack web development**, with a backend composed of **MongoDB**, 2 **Flask** API servers, and **Nginx** as reverse proxy and load-balancer. Frontend designed using **React** and **Node.js**.
- Integrated **Google Maps JavaScript API** in the frontend to display nearby station recommendations. Enabled live location detection and station navigation, features unsupported by the official rental website.

AICUP 2021 - Chinese Medical Dialogue Analysis Competition — **[Pytorch/NLP]**
1st place, 81 teams in total — *Taipei, Taiwan*

- Trained **deep learning BERT** models to complete reading comprehension tasks based on medical dialogues of over 2000+ words. Utilized **BM25** to rank word cosine similarity under BERT's input length constraints.
- Implemented the **XLNet model** to assess patient risk levels, achieving 92% accuracy.

Skills

Languages(#years)	Python(>5), C++(4), JavaScript(4), Go(1)
Frameworks and tools	**Machine Learning Pytorch**, Keras, Nsight | **Cluster** Kubernetes | **Web** Node.js, React, Nginx, Flask | **Database** SQL, MongoDB | **Tools** Docker,Linux, Hadoop, AWS Lambda/EC2

Honors and Awards

- **Chairman's Fellowship**, 2022-2024
- **OSDI Travel Award**, 2024
- **AICUP - 1st place among 174 competitors** ,2021
- **Dean's List**, 2016