Bing Han

(631) 479-9014
Stony Brook, NY
bingshiunhan@gmail.com linkedin.com/in/bing-shiun-han | github.com/nba556677go

Education

PhD candidate in Computer Science, *Stony Brook University* 09.2022 — 05.2027(Anticipated)
Bachelor of Electrical Engineering, , *National Taiwan University* 09.2015 — 01.2020

Skills

Languages	Python, C++, JavaScript
Frameworks and tools	**Machine Learning/GPU** Pytorch, Keras, Nsight | **Cluster** Kubernetes | **Web** Node.js, React, Nginx, Flask | **Database** SQL, MongoDB | **Tools** Docker, Linux, AWS Lambda/EC2, Git

Professional Experience

Research Assistant **07.2023 — Present**
Stony Brook University *Stony Brook, NY*

- **Project: GPU performance analysis and prediction on DL serving**
- Improved GPU utilization by predicting and minimizing job interference under colocation. Reduced 36% completion time.
- Predicted optimal job colocation using fine-grained GPU kernel profiles from **NVIDIA Nsight Compute**. Analyzed over 20 GPU metrics to colocate workloads based on compute, memory, and cache usage.
- Trained a regression model with kernel metrics. Achieved 90% prediction accuracy with 30% of data as training set.
- Leveraged **NVIDIA MPS** for efficient job sharing with compute isolation. Achieved 1.5x increase in throughput.

- **Project: Optimize DL scheduling with Kubernetes**
- Optimized GPU scheduling policies for ML tasks placement to maximize resource allocation for colocating latency and throughput-sensitive tasks, such as chatbot and document-retrieval.
- Constructed an end-to-end ML deployment pipeline using **Kubernetes**. Modified K8S scheduler source code to enable **shortest-job-first** scheduling. Achieved a 20% reduction in total completion time.
- Implemented an ML profiler for accurate prediction of GPU memory and time usage. Predicted task completion time within 4% error rate.

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

- Engineered **Hadoop**, **Spark**, **Hive**, and **Kafka** microservices using **Docker**, leading the team's first Docker-based project and involving 33% of team members.
- 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

Alcohol Advisor - Alcohol Consumption Analysis **[D3/JavaScript/Flask]**
*Star project (**15 out of 58 teams**), Visualization* *Stony Brook, NY*

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.
- Performed **data augmentation** by including additional Chinese dialogues, improving accuracy by 20%.
- Implemented the **XLNet model** to assess patient risk levels, achieving 92% accuracy.

Selected publications

- KACE: Kernel-Aware Colocation for Efficient GPU Spatial Sharing, **B.Han**, T.Paul, A.Gandhi, Z.Liu(Socc24 submitted)