

Hsu-Tzu (Irene) Ting

☎ +886-966-202-083 | ✉ hsutzu.ting@salab.cs.nthu.edu.tw |  Hsu-Tzu Ting

AREAS OF INTEREST

Distributed Systems, Cloud Computing, GPU Architecture

EDUCATION

- **M.S. in Computer Science** 09/2023 - 07/2025 (Expected)
National Tsing Hua University (NTHU) Hsinchu, Taiwan
 - Large-scale System Architecture Lab Led by Jerry Chou
 - Courses: Advanced High Performance Computing Cluster Practice, Virtualization Technology and its Applications
- **Exchange Student in Computer Science** 03/2023 - 08/2023
Dresden University of Technology (TUD) Saxony, Germany
 - Courses: Distributed Operating Systems, Foundation of Concurrent and Distributed Systems
- **B.S. in Computer Science** 09/2019 - 08/2023
National Tsing Hua University (NTHU) Hsinchu, Taiwan
 - GPA: 4.2/4.3, Top 5% Graduate
 - Courses: Data Structures, Operating Systems, Parallel Programming, Introduction to Machine Learning

PUBLICATIONS

- **KubeComp: Resource-centric Composable Container Orchestrator for GPU Pooling**
**Hsu-Tzu Ting, Jerry Chou, Ming-Hung Chen, I-Hsin Chung, Huaiyang Pan*
UCC '24: Proceedings of the IEEE/ACM 17th International Conference on Utility and Cloud Computing
 - Developed a Kubernetes-based framework for efficient resource allocation on top of the composable infrastructure (PCIe fabric GPU chassis).
 - Implemented GPU pooling with dynamic reallocation mechanism to optimize resource utilization.
 - Achieved up to 80% improvement in job waiting time on the testbed and simulation.
- **PCIe Bandwidth-Aware Scheduling for Multi-Instance GPU**
*Yan-Mei Tang, *Hsu-Tzu Ting, Jerry Chou, Wei-Fang Sun, Ming-Hung Chen, I-Hsin Chung*
Under Review
 - Recognized PCIe bandwidth contention in real-world tasks, leading to increased job execution times.
 - Proposed an online PCIe bandwidth-aware scheduler to mitigate the PCIe bandwidth contention among NVIDIA Multi-instance GPU (MIG) instances and leveraged MIG reconfiguration.
- **Reproducing Performance of Data-Centric Python by SCC Team From National Tsing Hua University**
*Fu-Chiang Chang, En-Ming Huang, Pin-Yi Kuo, Chan-Yu Mou, *Hsu-Tzu Ting, Pang-Ning Wu, Jerry Chou*
IEEE Transactions on Parallel and Distributed Systems
 - Reproducibility challenge is one of the applications in the student cluster competition.

AWARDS

- **ISC23 Student Cluster Competition: Second Place** Hamburg, Germany
Issued by HPC-AI Advisory Council 05/2023
- **SC22 Student Cluster Competition: Overall Winner** Dallas, USA
Issued by SCC 2022 Committee 11/2022
- **IEEE TCHPC Student Travel Award** Dallas, USA
Issued by IEEE Computer Society 11/2022
- **Academic Achievement Awards (Top 5% in Class): 5 Semesters** Hsinchu, Taiwan
Issued by National Tsing Hua University

EXPERIENCE

- **International Research Collaboration**

RIKEN Center for Computational Science

Upcoming

- Aims to address scientific challenges in I/O and data management for AI science workflows

IBM Thomas J. Watson Research Center

03/2023 - Present

- Investigated NVIDIA multi-instance GPU job scheduling for GPU utilization maximization.
- Explored the strength of PCIe fabric GPU chassis and incorporated it into Kubernetes.

- **Research Assistant | National Tsing Hua University**

Distributed Training and Elastic Training Scheduling in Kubernetes

03/2022 - 11/2022

- Conducted distributed elastic training using the Horovod framework within Kubernetes.
- Monitored cluster resource utilization with Prometheus and triggered automatic scaling.
- Improved throughput by 35% through dynamic scaling compared to non-scaling.

- **Teaching Assistant | National Tsing Hua University**

Distributed Systems Course

02/2024 - 06/2024

- Delivered Kubernetes course on cluster setup and load-balanced inference service deployment.
- Designed an assignment on a custom scheduler following the Kubernetes scheduling framework.

Operating Systems Course

09/2022 - 01/2024

- Supported students with assignments covering the implementation of system calls, memory management, CPU scheduling, and file systems.

- **Student Cluster Competition**

ISC High Performance 2023

02/2023 - 05/2023

- Observed FluTAS application bottleneck through profiling (using Nsight Compute and Nsight System).
- Improved performance by 6% by binding the GPU processes with the corresponding NUMA nodes.

Supercomputing Conference 2022

07/2022 - 11/2022

- Accelerated PHASTA application by porting to PETSc solver and enabled vectorization.
- Compared the application's performance and scalability on both our AMD and Intel nodes.

- **Software Engineer Internship**

Google | Pixel Watch MCU GPU Command Buffer Debugger

06/2024 - 09/2024

- Enabled the command buffer dump feature and managed inter-chip communication through RPC calls.
- Designed a parser to reverse engineer command buffer binaries into human-readable instructions.

Garmin | Smart WorkStation

07/2022 - 12/2022

- Fine-tuned object detection models and integrated an object tracking feature into the smart workstation.

SKILLS

- **Programming Languages:** C/C++, Python, Go
- **Parallel Programming:** MPI, Pthread, CUDA
- **Container Orchestration:** Docker, Kubernetes
- **Profiling Tools:** Intel VTune, Nsight System, Nsight Compute