Shengjie Hu

github.com/pHantomU94 phantomu94.github.io shengjie.horizon@gmail.com

RESEARCH INTERESTS

System-Level optimization for efficient training and inference, Edge-cloud co-inference, Distributed computing and optimization

EDUCATION

Beihang University **Beijing, CHN**
M.S. in Computer Science *Sept. 2017 - Jan. 2020*

- GPA: 3.58/4.0
- Relevant Courses: Design and Analysis of Algorithms(91), Distributed Operating System(85), Advanced Computer Network(90), High Performance Computer Architecture(85).

Beijing University of Technology **Beijing, CHN**
B.E. in Computer Science *Sept. 2013 - Jun. 2017*

- GPA: 3.72/4.0
- Relevant Courses: Advanced Mathematics(91), Advanced Language Programing(93), Data Structures and Algorithm(83), Principles of D atabase System(97), Principles of Compiling(93) Principles of Computer Organization(97), Computer Graphic(96),

RESEARCH EXPERIENCE

Cross-architecture Optimization and Inference of Machine Learning Models *2023 - 2025*

- We developed an inference platform named **Ocean**, which effectively leverages the architectural characteristics of heterogeneous computing chips to achieve software–hardware co-accelerated inference and deploy commonly used industrial neural network models across heterogeneous computing architectures (e.g., the Orin and RK3588 series), thereby enhancing their applicability in diverse application scenarios; in edge computing environments, the platform achieves up to **100% higher** inference efficiency compared with NVIDIA Triton.

Scalable Heterogeneous Computing Architecture *2021 - 2022*

- Developed highly scalable and adaptive computing resource scheduling techniques that can dynamically adjust resource allocation based on the scale of AI tasks.
- Designed and built a flexible PCIe switch–based platform that supports dynamic GPU resource allocation according to varying task demands.
- The research achievement, **"XIANGXUE-3B" server**, achieved **two championships** and three runner-ups in the **MLPerf v2.0** in 2022.

Coflow Scheduling in Distributed Computing Clusters *2019 - 2020*

- Addressed application-level communication performance bottlenecks of distributed computing in Spark by leveraging deep reinforcement learning based coflow scheduling.
- Optimized multi-level queue thresholds on switches, reducing average completion time for **coflow (2x faster than per-flow fairness)** and enhancing overall application efficiency.

Network Simulation Techniques for Spark *2018 - 2019*

- Analyzed the execution trace of Spark and then established a task execution model to predict task execution time and shuffle data volumes.
- Developed a tool that enables the simulation of Apache Spark cluster and the emulation of **shuffle processes** within the cluster.

PUBLICATIONS

Sun, Y., Hao, J., Zou, Z., Shu, L., & **Hu, S.** (2022). *Hierarchy SeparateEMD for Few Shot Learning.* In *Asian Simulation Conference*. Singapore: Springer Nature Singapore.

Li, W., **Hu, S.**, Wang, D., Chen, T., & Li, Y. (2020). *SPM: Modeling Spark Task Execution Time from the Substage Perspective.* In *Algorithms and Architectures for Parallel Processing (ICA3PP 2019)*. Melbourne, VIC, Australia: Springer International Publishing.

Li, W., **Hu, S.**, Sun, G., & Li, Y. (2018). *Adaptive Load Balancing on Multi-core IPsec Gateway.* In *Algorithms and Architectures for Parallel Processing (ICA3PP 2018)*. Guangzhou, China: Springer International Publishing.

WORK EXPERIENCE

Institute of Automation, Chinese Academy of Sciences (CASIA) *Apr. 2020 - Present*
Position: Assistant Researcher

- Designed and implemented a large scale distributed AI training system, including device selection, networking topology, virtualization (OpenStack), and container management (K8S). This infrastructure **supportted the training of 'Zidong Taichu' Taichu'**, a multimodal pre trained model with 100 billion parameters.
- Contributed to the development of a **cross-regional** heterogeneous distributed AI training platform, ranking 17th in the AIPerf500 ChinaSC2023 benchmark (11th in 2022).
- Contributed to the development of a **digital backend acquisition system** for large radio telescope arrays, enabling precise real-time data collection and processing at a rate of **100 GB per second**. The system has been deployed in **the Tianlai Project and Chinese Meridian Project–Phase II**.
- Developed a control system and driver library for **distributed quantum computing measurement and control**, facilitating precise instruction distribution and bidirectional communication with devices. This system has been deployed at **ZJU** and **BAQIS**.
- Spearheaded the development of the **Ocean distributed AI application platform**, optimized for deploying and training AI algorithms in edge scenarios. The platform integrates functions such as data annotation, dataset management, algorithm training, deployment, and multi-chip inference compatibility. It has been successfully applied in industrial and agricultural AI services, including applications in construction inspection and agricultural monitoring.

MANAGEMENT EXPERIENCE

Institute of Automation, Chinese Academy of Sciences (CASIA) *2023 - Present*
Position: Assistant Researcher

- Responsible for grant applications and full-cycle project management.
- Facilitating the transfer of projects from research to industry.
- Managed an AI application product development team of over 10 members.

SKILLS

Programming	C/C++, Python, Golang
Frameworks & Tools	Spark, Hadoop, Kubernetes (K8S), OpenStack, DPDK
AI & Machine Learning	Deep Reinforcement Learning, Few Shot Learning
Research Management Skills	Grant Application & Project Management Skills
Others	Strong Goal-achievement Capability & Diligence