Shengjie Hu

Tel: +86 136-4123-2881 | Email: shengjie.hu@ia.ac.cn

Research Interests：Distributed Computing, AI Computing, Big Data Computing

EDUCATION

Beihang University Sep 2017 - Jan 2020

Master of Computer Science | 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 Sep 2013 - July 2017

Bachelor of Computer Science | GPA:3.72/4.0
Relevant Courses: Advanced Mathematics(91), Advanced Language Programing(93), Data Structures and Algorithm(83), Principles of Database System(97), Principles of Compiling(93)，Principles of Computer Organization(97), Computer Graphics(96),

PUBLICATION

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 Sub-stage 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.

RESEARCH EXPERIENCE

Scalable Heterogeneous Computing Architecture CASIA 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 implemented a flexible PCIe Switch-based hardware platform, enabling dynamic GPU resource allocation based on task requirements.
- The research achievement, **"XIANGXUE-3B"** server, achieved two championships and three runner-ups in the **MLPerf** v2.0 in 2022.

Coflow Scheduling in Data-Parallel Clusters Beihang University 2019 - 2020

- Addressed application-level communication performance bottlenecks of distributed computing in Spark by leveraging deep reinforcement learning based coflow (semantic-related flow groups) 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 Beihang University 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.

IPsec VPN Gateway Performance Optimization（SDN） Beihang University 2018 - 2019

- Developed a high-performance IPsec VPN gateway capable of efficiently managing heavy traffic loads. Designed a multi-stage parallel pipeline architecture featuring fine-grained flow scheduling and dynamic multi-core load balancing.
- Achieved 10Gbps throughput (on a 8 cores cpu platform) with microsecond-level processing latency.

WORK EXPERIENCE

Institute of Automation, Chinese Academy of Sciences (CASIA) Apr 2020 - Present

Position: Assistant Researcher

Key Responsibility:

- 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'**, 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 the 100GB scale per second. This system has been utilized in the **Tianlai Project** and the **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 *Oceanic* distributed **AI application platform**, optimized for deploying and training AI algorithms in edge scenarios. The platform includes features such as data annotation, dataset management, algorithm training, deployment, and compatibility with multiple inference chips. It has been successfully applied across various industrial and agricultural AI services.

MANAGEMENT EXPERIENCE

Guangdong Institute of Artificial Intelligence and Advanced Computing (Branch of CASIA) Assistant Dean 2023 - Present

- Assisted the Dean in facilitating the **transfer of technology from research to industry**.
- Oversee **research project management and administrative operations.**

SKILLS

Programming: C/C++, Python, Golang

Frameworks & Tools: Spark, Hadoop, Kubernetes (K8S), OpenStack, DPDK

AI & Machine Learning: Deep Reinforcement Learning, Few-Shot Learning