Shaojie Tan

943648187@qq.com | shaojiemike.top | GitHub | Google Scholar

Huawei Ascend training systems engineer with an M.S. in Computer Science from USTC. Focused on large-model training and inference optimization on Ascend NPUs, multimodal migration, distributed parallelism, RL post-training, and quantization, backed by computer-architecture and multi-node CPU/GPU/NPU optimization experience.

Education

University of Science and Technology of China | *M.S. in Computer Science* — Sep 2021–Jun 2024

University of Science and Technology of China | *B.S. in Computer Science* — Sep 2017–Jun 2021

Experience & Research

Huawei Ascend | *AI Systems Engineer* — 2024 H2–Present

- Optimized PyTorch / torch_npu training paths via CPU affinity, operator Lazy Init, parameter tuning, and L2 cache; adapted OpenSoraPlan, DeepSeek-VL2, and GLM-4.1V with Megatron / FSDP.
- Contributed to DeepSeek-R1 and Qwen3 large-EP inference and Attention INT8 quantization; adapted UniVL, verl + Qwen2.5-VL, and DanceGRPO, raising a generative-model RL score from 0.4 to 0.8.
- Doubled Intern-S1 Pro inference through efficient GMM and post-processing computation (report); optimized GDN Triton kernels and the framework to improve Intern-S2 Preview (Qwen3.5-35B) SFT from 0.12x to 0.72x H200 (report).

Huawei 2012 Laboratories | *Research Intern · PIM Task Scheduling* — Jul 2023–Sep 2023

- Built a compile-time performance model and reduced data movement through locality and connectivity clustering.
- Selected execution between PIM and CPU using parallelism and memory-port pressure, achieving near-optimal scheduling with static analysis alone.

USTC × Huawei | *Kunpeng 920 Basic-Block Throughput Modeling* — Sep 2021–Nov 2022

- Developed a throughput prediction framework and benchmark suite using instruction throughput, latency, and port-pressure data.
- Improved llvm-mca accuracy on AArch64 by modeling the Kunpeng 920 microarchitecture; published at IEEE HPCC 2022.

Selected Projects

Pujiang Lab · Intern-S2 Preview SFT Optimization | *Qwen3.5-35B* — Apr 2026–Jun 2026

- Profiled GDN at 70%+ of out-of-box runtime, removed transpose overhead in causal_conv1d / KKT, replaced or optimized key Triton paths with Ascend C, and added chunk-size 128 support.
- Optimized solve-tril, KKT, RMSNormGated, CV-core synchronization, CPU affinity, and SP2 communication, improving SFT from 0.12x to 0.72x H200.

Multi-GPU Quantum Circuit Simulation with QuEST | *ASC20-21 First Prize* — Nov 2020–May 2021

- Used CUDA-aware MPI, streams / buffers, and cuBLAS to scale from 30 to 34 qubits on 8 A100 GPUs across 4 nodes, saving 200+ GB versus the CPU system.

Multi-node CPU Optimization | *IPCC 2022 Third Prize · 5/130+* — Jul 2022–Nov 2022

- Combined unrolling, vectorization, OpenMP, and MPI with data-layout, mixed-precision, and PSRS optimizations; reached 10,000x speedup on two AMD EPYC 7452 nodes and up to 2,800,000x on one node.

Technical Skills

- **AI frameworks:** PyTorch, torch_npu, Megatron, FSDP, verl, vLLM. **Kernels & systems:** Triton, Ascend C, Ascend training/inference, multimodal models, RL, Expert Parallel, Attention INT8 quantization.
- **Programming & parallel:** C, C++, Python, Shell, MPI, OpenMP, CUDA, cuBLAS. **Architecture:** AArch64, CPU microarchitecture, PIM, performance modeling, cache and bandwidth optimization.

Honors & Publications

- **Outstanding Graduate, USTC (2024)**; First Prize, ASC20-21 and ISC21 Student Supercomputer Challenge Finals.
- **Publications:** **S. Tan** et al., "Uncovering the Performance Bottleneck of Modern HPC Processor," **CCF Trans. HPC 2024**; Q. Jiang, **S. Tan**, et al., "A^{3}PIM," **DATE 2024**; "Quantifying Throughput of Basic Blocks," **IEEE HPCC 2022**.

AI Practice & Engineering Automation

Embrace AI through hands-on projects: AIHisTrendFuture (live) maps AI evolution; DevTracking tracks delivery; IPD-Teams organizes stateful agents. AutoResearch and ProgramModeling explore performance-model-driven optimization automation.

Career Vision

See the essence through first principles; amplify the value of intelligence through systems innovation.