

Jing Fang

(+1) 425-463-7874

jing.fang.zju@gmail.com

Redmond, US

OBJECTIVE

AI framework engineer with proven track record in machine learning framework, model performance tuning and distributed systems. Leveraging expertise in linear algebra kernel implementation, model quantization and distributed system design, I seek to contribute to machine learning framework, model optimization and model serving pipeline.

WORKING EXPERIENCE

Microsoft

Senior Software Engineer, AI Framework | US, Redmond

May 2024 - Jul. 2025

Owner of mlas (linear algebra subroutine) library and CPU kernel optimization for ONNX runtime (ORT).

- Added Q/DQ format support and quantization kernel fusion. Improved memory allocation during Q/DQ model loading, resulting in **42% memory reduction and 45% latency reduction**. (C++, Python)
- Added various 4-bit quantized kernels (**word embedding**, **MatMul**, etc.). Achieved ~85% model size reduction for a RoBERTa-based model. (C++, CUDA, ARM NEON, x86 AVX2)
- Collaborated with the research team and quantized a Roberta-based model using mixed precision. Achieved 8.2% model size reduction, 60% single thread CPU usage reduction and 23% latency reduction, enabling a \$15M contract signoff.
- Added vectorized CPU kernels (**rotary positional embedding**, **MatMul**, **softmax**, **tanh**, etc.) to optimize block sparse attention and group query attention. **Boosted 48% perfill throughput and 67% decoding speed** on Phi-3-4K-Instruct model. (C++, ARM NEON, x86 AVX2)

Senior Software Engineer, Azure Resource Graph | US, Redmond

Mar. 2023 - May 2024

Software Engineer 2, Azure Resource Graph | China -> US

Nov. 2020 - Feb. 2023

Owner of ingestion data pipeline. Relocated from Shanghai to Redmond in Nov. 2021.

- Owner of the ingestion data pipeline that served 53.7 billion write transactions per day, ensured eventual consistency with 500 ms P95 latency and 99.9% availability.
- Led a three-member team in the design and development of a data pipeline to replicate RBAC datasets. Attained 100% availability in a 6 months time window using automatic dataset failover. Achieved eventual consistency with 10 seconds P99 SLA. Resolved the single point of failure for the RBAC datasets and unblocked the query traffic to scale out unlimitedly. (C#)
- Innovated an algorithm for monitoring message broker health. Based on it built a retry mechanism and an automatic failover solution across multiple brokers. This work improved per hop availability from 99% to 99.99% and expanded traffic capacity from 40MB per second to boundless, establishing it as a fundamental component for the team's resilient data pipelines. (C#)

Software Engineer 2, Azure Stack Storage | China, Shanghai

Oct. 2019 - Nov. 2020

- Re-designed and implemented the garbage collection (GC) for block blob storage. Achieved 80% GC throughput improvement, and reduced P99 transaction latency from >10s to <100ms. (C++)
- Designed and implemented the **batch mode CancellationTokenSource**. Resolved the lock contention issue induced by Timer. Achieved 10% CPU usage reduction and 4x upload throughput increase. (C#)

Software Engineer 2, O365 Monitoring | China, Suzhou

Apr. 2019 – Sep. 2019

KLA-Tencor

Algorithm Engineer | China, Shanghai

May 2017 – Mar. 2019

Worked on software performance tuning, machine learning algorithm, and image processing algorithm.

- Added 10+ AVX2 kernels. Improved the memory access pattern for 6+ functions. Resolved multiple intra-loop dependency issues. Achieved 20% e2e throughput increase in two months. (C/C++, AVX2)
- Designed and developed a profiling framework and boosted ~5x testing throughput. (bash)
- Developed a hotspot-based defect detection algorithm with a colleague. Achieved 15x capture rate vs. the baseline in an in-house demo and secured a multi-million purchase order. (C/C++)

Application Engineer | China -> Singapore

May 2013 – May 2017

Worked on designing, prototyping, and application of image-based defect inspection solutions. Relocated from Shanghai to Singapore in Aug 2015.

- Designed and developed a neural network based defect classifier from scratch using Numpy and Scipy and won the Best Execution Award in the Hackathon. (Python)
- Independently designed a vector-based pattern search algorithm that improved the flexibility of the original solution, and co-filed a patent with five other colleagues. Patent number: [US 10,018,571 B2](#).

EDUCATION

M.Sc. in Applied Physics | University of Twente, Netherlands

Aug. 2010 – Sep. 2012

Received University of Twente Scholarship

B.Eng. in Information Engineering | Zhejiang University, China

Sep. 2006 – Jun. 2010

Overall GPA 3.87/4.0, major GPA 3.91/4.0

Received Outstanding Graduate of Zhejiang University Award and multiple other awards

TECHNICAL SKILLS

Performance Tuning	ARM NEON, x86 SSE/AVX2/AVX512, memory blocking/buffering, loop optimization, perf, VTune Amplifier, PerfView, CMake.
Programming	C/C++ (9+ years), C# (5+ years), Python (3+ years), CUDA, Matlab, bash.
AI/ML	linear/logistic regression, random forest, convolutional neural network models, transformer models, model quantization.
Distributed System	multi-threading, data partitioning, caching, asynchronous workflow, data replication, Azure.