Zhufeng (Zephyr) Qiu

+1 (206) 670-8645 | zhufqiu@gmail.com | zhufqiu.com | github.com/Zhufeng-Qiu | Seattle, WA, US

RESEARCH INTERESTS

GPU Systems | High-Performance Computing | Compression-Accelerated Collective Communication | Distributed & Parallel Systems | ML Systems | High-Performance LLM Training & Inference

RESEARCH SUMMARY

I am interested in building efficient and scalable systems for large-scale AI workloads, with a focus on GPU optimization, collective communication, and distributed execution. My experience spans parallel data processing, production-scale distributed systems, and performance engineering across serial CPU, OpenMP, MPI, CUDA, and NCCL backends, including lossless compression of collective payloads and a controlled study of when communication–computation overlap pays. These experiences motivate my pursuit of Ph.D. research at the intersection of HPC and ML systems, particularly GPU runtime optimization, communication-efficient execution, and high-performance LLM training and inference.

EDUCATION

Northeastern University, Khoury College of Computer Sciences **Jan. 2023 – May 2025**
- M.S. in Computer Science, GPA: 4.0/4.0
- Relevant coursework: Parallel Data Processing, Algorithms, Computer Networking, Cloud Computing, Database Management Systems; Teaching Assistant for Algorithms and Parallel Data Processing.

University of Southern California, Viterbi School of Engineering **Aug. 2019 – May 2021**
- M.S. in Applied Data Science, GPA: 3.83/4.0
- Relevant coursework: Scientific Computing and Visualization, Analysis of Algorithms, Machine Learning, Data Mining, Database Systems.

Wuhan University, School of Resource and Environmental Sciences **Sep. 2014 – Jun. 2018**
- B.S. in Geographical Information Science, GPA: 3.58/4.0
- Relevant coursework: Data Structures, Linear Algebra, Discrete Mathematics, Probability and Mathematical Statistics, C Programming.

SELECTED SYSTEMS, HPC & ML PROJECTS

Multi-GPU Similarity Engine — Lossless Collective Compression & Overlap GitHub **Jul.2026 - Aug. 2026**
Extension of USC INF 553 Data Mining / CSCI 596 Scientific Computing Projects | C++, CUDA, NCCL, MPI, OpenMP, PySpark, CMake, Nsight Systems
- Re-engineered a Yelp collaborative-filtering workload from Spark into serial C++, OpenMP, MPI, CUDA, and dual-GPU NCCL backends behind a frozen numerical contract; every backend bit-exact, processing **1.17M candidate pairs in 4.21 ms on two NVIDIA A100s — 312× serial, ~20× over 16-thread OpenMP**.
- Compressed the NCCL AllReduce payload **3× losslessly (56.2 → 18.8 MB)** by packing six integer sufficient statistics into two uint64 words with carry-free fields, letting NCCL **sum them in compressed form without decompressing**; verified bit-exact across 2.58M pairs.
- Isolated the interconnect by running the identical binary on **NVLink and PCIe**: communication is 11% vs 91% of the iteration, so the same compression buys **4% vs 55%**, and combined with overlap cuts the PCIe iteration **2.53×**.
- Derived and Nsight-verified the overlap ceiling min(compute, comm)/total — overlap pays when compute and communication are **comparable**, not when communication is large; a chunk-size sweep also exposed and fixed a double-buffering race.

Quantized LLM Inference Study on a LoRA-Fine-Tuned Llama 3.1 8B GitHub **Jul.–Sep. 2025; Jul.-Aug. 2026**
Independent inference-systems study extending an LLM-engineering course fine-tune | Python, PyTorch, CUDA, PEFT/LoRA, bitsandbytes, Hugging Face Transformers, Modal
- Fine-tuned Meta Llama 3.1 8B with QLoRA (rank 32 / α 64 on attention projections, **27M trainable parameters — 0.34% of base**) under completion-only supervision on a curated 400,000-item dataset, and served it 4-bit NF4 on an NVIDIA T4 via Modal.
- Built a CUDA-event inference benchmark across four quantization configurations on **Turing and Ampere**, decomposing latency into prefill and decode against an analytical roofline; showed **97% of the tokens are prefill but only 39% of the time is** — decode costs 54× more per token.
- Identified a **2.50× end-to-end regression in the deployed configuration**: a bf16 compute dtype carried over from Ampere training onto Turing, which has no bf16 tensor cores. Localized entirely to prefill (4.42×, MFU collapsing to 3.2%) and isolated by an unquantized control agreeing within 0.8%.
- Showed **4-bit quantization buys capacity, not speed** — NF4 decode runs 1.69× slower than fp16 while moving 3.9× fewer bytes, reaching 9.8% of peak bandwidth against fp16's 64.1%: dequantization-bound, not bandwidth-bound.

Distributed Apriori Mining on Yelp Dataset GitHub **Feb. 2024 – May 2024**
Northeastern CS 6240 Parallel Data Processing | Java, Hadoop MapReduce, HBase, Hive, AWS S3
- Built a distributed Apriori pipeline on Hadoop MapReduce over ~550K Yelp reviews and ~15K businesses, decomposing candidate generation, support counting, pruning, and iterative frequent-itemset discovery, with Hive external tables preprocessing raw JSON into per-user review baskets.
- Implemented two interchangeable storage backends for cross-pass state — an HBase path materializing each pass as a table with support filtering applied at write time, and a file-based path over HDFS/S3 with a runtime flag switching between local and S3 discovery of per-pass outputs.

Zephyr Manus — Agent Execution and Tool-Orchestration Platform Live Demo **Mar. 2026 – May 2026**
Independent Project | FastAPI, Redis Streams, PostgreSQL, Docker, Next.js, MCP/A2A
- Built a Planner–ReAct agent runtime supporting task decomposition, external-tool execution, persistent state, and human-in-the-loop wait/resume workflows.

- Developed Docker-isolated execution for Browser, Shell, and File tools and an event-driven FastAPI backend using Redis Streams, PostgreSQL, SSE, and MCP/A2A.

Train Ticket Flash-Sale System [GitHub] **Jan. 2025 – Sep. 2025**

Independent Project | Java, Spring Boot, Spring Cloud, Redis, RocketMQ, MySQL

- Built a high-concurrency ticket-purchasing service using Redis-based inventory control and RocketMQ-based asynchronous processing to decouple request handling from downstream order creation.
- Designed idempotency and consistency mechanisms for concurrent requests, mitigating duplicate purchases, overselling, and retry-related failures under flash-sale workloads.

RESEARCH EXPERIENCE

PM2.5 Inversion and Spatiotemporal Analysis for the Wuhan Metropolitan Area **Mar. 2016 – Oct. 2017**

National Undergraduate Innovation and Entrepreneurship Training Program | Core Team Member | Advisor: Prof. Huanfeng Shen

- Integrated ground-monitoring measurements, NASA MODIS/AOD satellite observations, and meteorological data, performing preprocessing and spatiotemporal matching across multiple data sources.
- Contributed to the development and evaluation of a mixed-effects model for PM2.5 inversion, using cross-validation and baseline comparisons to assess predictive performance.
- Used MATLAB/R for modelling and ArcGIS to generate spatially continuous PM2.5 concentration maps across the Wuhan "1+8" metropolitan area; the national program passed final review with a **Good** rating.

Web-Based Dynamic Thematic Mapping Technologies **Mar. 2018 – May 2018**

Undergraduate Thesis | SinoMaps Press Collaboration | Advisors: Prof. Haihong Zhu and Yuntao Long

- Surveyed web-based thematic-mapping methods and designed the overall, functional, and logical architecture for online visualization, interaction, and dynamic data updates.
- Built and validated a D3.js customs trade-map and map-service prototype supporting thematic visualization, statistical charts, interactive queries, and live data updates.

INDUSTRY EXPERIENCE

Beaconfire Solutions Inc. — Software Development Engineer **Jul. 2025 – Present**

Seattle, WA (Remote)

- Developed enterprise HR and digital-commerce microservices with Spring Boot/Spring Cloud, integrating authenticated APIs, relational/document data stores, AWS S3, and RabbitMQ-backed asynchronous workflows.

NSFOCUS Technologies Group Co., Ltd. — R&D Engineer **Jun. 2021 – Dec. 2022**

Wuhan, China

- Built and maintained distributed security-data pipelines using Kafka and Elasticsearch for multi-source log ingestion and normalization, and implemented Spark/MapReduce batch workflows serving 100+ enterprise clients and processing approximately 5M records per day.
- Developed host-anomaly detection features for illicit file transfers and backdoor processes using fuzzy inference, hierarchical clustering, and k-core graph mining over multidimensional security logs.
- Contributed to a production alerting workflow that achieved effective response for **100% of severe threats and 90% of high-risk threats** in a telecom deployment.

NSFOCUS Technologies Group Co., Ltd. — R&D Engineer Intern **Jul. 2018 – Mar. 2019**

Wuhan, China

- Processed and normalized multi-source security logs covering user operations, file access, and process behavior to support downstream host-anomaly analysis.

TECHNICAL SKILLS

HPC / GPU: CUDA, NCCL, MPI, OpenMP, Nsight Systems, CMake
Languages: C/C++, Python, Java, Scala, SQL, JavaScript/TypeScript
ML / LLM: PyTorch, Hugging Face Transformers, PEFT/LoRA, bitsandbytes, scikit-learn, LangGraph, ReAct, RAG, MCP/A2A
Distributed / Data: Spark/PySpark, Hadoop MapReduce, Hive, HBase, Kafka, Elasticsearch, Redis, PostgreSQL, MySQL, Chroma, AWS
Systems / Software: Docker, Modal, Spring Boot, Spring Cloud, Git, Jenkins, Maven

HONORS

- Outstanding Student Leader, Wuhan University, 2014–2015 and 2015–2016.
- Outstanding Individual in Theoretical Study, Wuhan University, 2014–2015.