Chenxu Niu

(806) 783-7599 | ncxhxgtg@gmail.com | linkedin.com/in/chenxu | github.com/chenxuniu

Education

Ph.D., Computer Science, Texas Tech University — Lubbock, TX
HPC, Sustainable AI, Inference Framework, Metadata Management — *06/2019 – 12/2025*

- Dissertation: "Accelerating Scientific Dataset Discovery with Semantic Metadata Search and Large Language Models"

M.S., Information Science, University of Science and Technology of China — Hefei, China
GPA: 3.95 out of 4.0 — *09/2015 - 06/2018*

B.E., Mathematics, University of Science and Technology of China — Hefei, China
GPA: 3.65 out of 4.0 — *09/2011 - 06/2015*

Experience

Solutions Architect — 01/2026 – Now
NVIDIA — *Santa Clara, CA*

- Working with NVIDIA AI Native, Consumer Internet and Enterprise customers on large data center GPU server and networking system deployments as Solution Architect Engineer
- Guide customer discussions on network design, compute/storage and support bring up of server/network/cluster deployments.
- Analyze and debug compute/network configuration, performance issues to deliver performant clusters

HPC Systems Administrator & Research Assistant — 06/2019 – 12/2025
Texas Tech University — *Lubbock, TX*

- **Served as the primary cluster administrator**, responsible for configuring, maintaining, and troubleshooting the SLURM workload manager and software stack to support hundreds of users
- Designed **real-time monitoring infrastructure** tracking power consumption, thermal management, and resource utilization across entire data center with iDrac, Redfish and Grafana
- **Automated deployment, configuration, and system updates** of cluster services using Ansible
- Installed, configured, and optimized **Finite Element Analysis software** and other engineering/scientific applications for NSF ACCESS users on HPC systems
- Deployed and optimized **distributed LLM training and inference** on GPU Clusters
- Accelerated semantic metadata search on scientific datasets with **RAG**

Technical Skills

Programming & DevOps: Python, C/C++, CUDA, Docker, Kubernetes, Terraform, Ansible, Git
HPC & Infrastructure: Warewulf, Slurm, Spack, InfiniBand, Lustre/GPFS, Linux Admin
AI/ML Frameworks: PyTorch, vLLM, TensorRT-LLM, Ray, DeepSpeed, Hugging Face, CUDA programming
Distributed Systems: MPI, OpenMP, Apache Spark, Redis, message queues, microservices architecture
Cloud & Monitoring: AWS/GCP, Prometheus, Grafana, ELK stack, power management APIs, thermal monitoring
Databases & Storage: PostgreSQL, MongoDB, HDF5, NetCDF, Distributed filesystems, object storage

Key Projects

HPC Data Center Design & Implementation | *NSF REPACSS Project* — 09/2023 – Present

- Led end-to-end design and deployment of renewable energy-powered HPC facility
- Implemented comprehensive monitoring stack (iDrac/Redfish/Grafana/TimeScaleDB) tracking real-time PUE, carbon footprint, and thermal management across 1000+ sensors
- Worked on a high-performance network topology with 200Gbps InfiniBand HDR fabric and 100Gbps Ethernet switching, optimizing MPI communication patterns for AI/ML workloads

Distributed LLM Inference Monitoring and Optimization | *vLLM, TensorRT-LLM, Ray* — 09/2023 – 06/2025

- Built real-time monitoring dashboard tracking GPU utilization, memory bandwidth, power draw, and thermal throttling across distributed inference clusters

- Created automated benchmark suite comparing vLLM, TensorRT-LLM, and Ray Serve performance across different hardware configurations and model sizes
- Developed a benchmark named **TokenPowerBench** to measure the energy efficiency and performance of open-source large language models across different inference engines and deployment configurations

Semantic Search and RAG application on Scientific Datasets | *RAG, HDF5, netCDF* 09/2019 – 06/2023

- Designed intelligent query engine processing 10TB+ scientific datasets (climate, genomics, astronomy) using RAG framework with 85% query accuracy
- Implemented distributed metadata indexing system for self-describing file formats (HDF5, netCDF, Zarr) supporting 100M+ scientific data objects
- Built vector embedding pipeline using transformer models to enable semantic search across heterogeneous scientific metadata and attributes
- Developed high-performance parallel query service using MPI and OpenMP, achieving 5x speedup over traditional metadata search approaches

Selected Publications

1. **Chenxu Niu**, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang and Yong Chen. **TokenPowerBench: Benchmarking the Power Consumption of LLM Inference.** In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI '26)*, 2026. (Acceptance Rate: 17.6%)

2. Gwok-Waa Wan, SamZaak Wong, Shengchu Su, **Chenxu Niu**, Ning Wang, Xinlai Wan, Qixiang Chen, Mengnv Xing, Jingyi Zhang, Jianmin Ye, Yubo Wang, RongChang Song, Tao Ni, Qiang Xu, Nan Guan, Zhe Jiang, Xi Wang, Yong Chen, Jun Yang. **FIXME: Towards End-to-End Benchmarking of LLM-Aided Design Verification.** In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI '26)*, 2026. (Acceptance Rate: 17.6%)

3. **Chenxu Niu**, Wei Zhang, Yongjian Zhao and Yong Chen. **Energy Efficient or Exhaustive? Benchmarking Power Consumption of LLM Inference Engines.** In *ACM SIGENERGY Energy Informatics Review, Volume 5 Issue 2*, 2025.

4. **Chenxu Niu**, Wei Zhang, Mert Side and Yong Chen. **ICEAGE: Intelligent Contextual Exploration and Answer Generation Engine for Scientific Data Discovery.** In *Proceedings of the 37th ACM International Conference on Scalable Scientific Data Management (SSDBM'25)*, 2025.

5. **Chenxu Niu**, Wei Zhang, Suren Byna and Yong Chen. **PSQS: Parallel Semantic Querying Service for Self-describing File Formats.** In *Proceedings of the 13th IEEE International Conference on Big Data (BigData'23)*, 2023.

6. **Chenxu Niu**, Wei Zhang, Suren Byna and Yong Chen. **Kv2vec: A Distributed Representation Method for Key-value Pairs from Metadata Attributes.** In *Proceedings of the 27th IEEE High Performance Extreme Computing Conference (HPEC'22)*, 2022.

7. Wei Zhang, Suren Byna, **Chenxu Niu** and Yong Chen. **Exploring Metadata Search Essentials for Scientific Data Management.** In *Proceedings of 26th IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC '19)*, 2019.

Professional Service

Program Committee Member: **AAAI '26, PEARC '26**
Reproducibility Committee Member: **SC '25**
Paper Reviewer: **SC '25, AAAI '26, BigData '25, PEARC '26**
Conference Volunteer: **SC '21 and SC '24**

Teaching Experience

Graduate Teaching Assistant 2020 – 2022
Texas Tech University (TTU) *Lubbock, Texas*

- **Computational Thinking with Data Science** (Fall 2021 - Fall 2022)
- **Advanced Operating System Design** (Spring 2021)
- **Analysis of Algorithms** (Spring 2020)

Invited Talks & Presentations

Semantic Search and Natural Language Query over HDF5. *2024 HDF5 User Group Meeting (HUG24)*, 2024.