Shengze (William) Wang

Ph.D. Candidate, Computer Science and Engineering

Homepage: https://shengze.io
Email: shengze@ucsc.edu

Technical Skills

- **Academic**: Computer Networks & Distributed Systems and their applications in AI Infrastructure (Systems & Networking for AI), GPU/TPU communication, distributed LLM Training & Inference Efficiency
- **Languages & Databases**: C/C++, Rust, Python, eBPF, Redis, Spanner, Bigtable, DynamoDB, Milvus, HDFS, SQL
- **Platforms & Tools**: Linux, ROS, Ray, Kubernetes - k8s (Borg), RDMA, NCCL, ICI, CUDA, GPU/TPU clusters, vLLM, SGLang, JAX/XLA, PyTorch, Transformers, ns-3, P4 Switch, TORs, GCP, AWS, Azure, Docker, Slurm, Git
- **Highlights**: ML flow scheduling, GPU/TPU communication, Distributed training & Inference, Scalable Key-Value Storage, Applied AI for Systems & Databases, Advanced Network Protocols & Security, Workload Characterization & load Balancing, Content Delivery & High-Performance Networks, Hashing Algorithms, MLOps & Production Deployment, Edge-Cloud Systems

Education

- **University of California, Santa Cruz (UCSC)** — San Francisco Bay Area
  Ph.D. in Computer Science and Engineering; Regents Fellowship; BE Dean's Fellowship — *2023 - present*
- **University of North Texas (UNT)** — Dallas - Fort Worth
  B.S. in Computer Science; GPA: 4.0; Outstanding Award (Top 1 of Class 2023); President's List — *2020 - 2023*
- **King's College London (KCL)** — London, United Kingdom
  Visiting Student in Computer Science and Technologies; Scored: 95/100 — *2019*

Working Experience

- **Student Researcher** — Google AI2 — Sept. 2026 - present
- **Software Engineering Intern** — Google AI2 — June. 2026 - Sept. 2026
- **Graduate Student Researcher** — Baskin School of Engineering, UCSC — June. 2023 - present
- **Undergraduate Research Assistant** — Department of Computer Science, UNT — Sept. 2021 - May. 2023

Selected Projects

- **ML flow scheduling, collective communication optimization for large-scale ML clusters** — 2026 - present
  Google AI2
  - Building and researching scalable networking systems for AI infrastructure, focusing on GPU/TPU communication and distributed LLM Training & Inference efficiency.
- **Scalable AI Infrastructure for High-Performance LLM Serving — C++, Python** — 2023 - present
  Qian Lab, https://users.soe.ucsc.edu/~qian/; Center for Research in Systems and Storage (CRSS)
  - Designed methods to **provide user anonymity and low-overhead encryption for queries/responses in large-scale overlay networks**, blending failover resilience with robust privacy guarantees. Developed **load balancing mechanisms with distributed key management** to support a decentralized and fault-tolerant LLM system.
  - Integrate Vector Databases to **accelerate retrieval-augmented generation over billions of embeddings**, achieving efficient approximate nearest neighbor lookups for knowledge-intensive LLM applications. Investigate novel indexing and query pipelines to enhance retrieval accuracy and throughput.
  - Devised a **draft-then-filter** mechanism to generate candidate tokens, then selectively offload low-confidence tokens to the full-scale LLM, achieving up to **2× speedup in real-time inference** without sacrificing output quality.
  - Implemented **NLL-based confidence scoring** to dynamically filter high-quality drafts, preventing unnecessary requests to the target LLM and thus lowering GPU consumption. Enhanced MLOps with continuous integration and robust monitoring, **streamlining the model lifecycle from draft-model updates to large-scale deployments.**
  - Proposed **"CALID"**: a novel inference framework that integrates **speculative decoding** principles to **boost throughput and reduce computational overhead** for large language models.
  - Proposed **Span-Level Fine-Tuning with unlikelihood training**: a novel approach that leverages annotated unfaithful spans in LLM-generated summaries to **reduce hallucinations and improve factual accuracy.**
  - Proposed **"GenTorrent"**: a decentralized overlay network to **enhance Large Language Model (LLM) serving scalability and cost-efficiency** by harnessing computing resources from distributed contributors. It addresses fundamental challenges in decentralized LLM serving, including **overlay network organization, anonymous communication for privacy, efficient overlay forwarding for load balancing and cache reuse, and decentralized verification of model serving quality**. GenTorrent aims to **democratize AI innovation**, significantly **reduce serving latency, and improve user privacy**, offering a novel approach to future AI deployment.
- **Resource Storage and Discovery in Network & Database Systems —C++, Rust** — 2023 - present
  Qian Lab, https://users.soe.ucsc.edu/~qian/; NSF Center for Systems and Storage, https://ssrc.us/
  - Investigate fundamental problems in emerging networks, emphasizing **efficient data placement, fault tolerance, and high-throughput designs**. (e.g., datacenter networks, CDNs, and quantum networking)
  - Architect and refine critical components—network protocols, routing algorithms, hashing strategies, and load balancers—for **enterprise-scale deployments** (e.g., HPC clusters, IoT networks, programmable switches).
  - Implement and evaluate prototypes using **event-driven simulators** (e.g., ns-3, p2psim) and cloud-based testbeds (AWS, Lambda Labs, CloudLabs, Supercomputers), leveraging asynchronous I/O and concurrency.

- Proposed **"LEAD"**: A novel Distributed Learned Hash Table that embeds machine-learned models within Distributed hash table structures to **significantly optimize range query performance** for distributed networked systems. LEAD outperforms existing range-query solutions by demonstrating **superior scalability, reduced latency, and robustness against network churns**. LEAD opens a completely new field for further research on integrating learned models with distributed systems.(https://github.com/ShengzeWang/LEAD; https://github.com/ShengzeWang/RM)
- Proposed **"Vortex":** A fully decentralized, planet-scale Vector overlay. Designed Distributed Learned Hashing (DLH) for locality and load balance, DHT routing for fault-tolerant lookup under churn, and per-peer D-HNSW for high-recall local search. Results match SOTA centralized systems' accuracy/latency while reducing per-peer index memory by $\sim 100\times$. (Learned Hash Function Library: https://github.com/ShengzeWang/LearnedHash)

- **Vehicular Edge Computing and Connected Autonomous Vehicles — Python, ROS** 2021 - 2023
  NSF Center for Electric, Connected and Autonomous Technologies, https:// ecat.center/ , http:// veclab.org/
  - **Profiled hardware resource usage (GPU, CPU, Memory)** for real-time object detectors (YOLO, Faster R-CNN, SSD) deployed in ROS-based CAV perception pipelines.
  - Investigated model optimizations, including **quantization, architectural pruning, and mixed-precision**, achieving a measurable trade-off between inference speed and detection accuracy under edge-device constraints.
  - Characterized **memory contention and identified performance bottlenecks**, enabling targeted optimization strategies (e.g., improved scheduling, memory partitioning) that enhanced detection throughput.
  - Developed **workload models** reflecting diverse edge-device configurations (e.g., NVIDIA Jetson, Intel CPUs, Raspberry Pi), facilitating **informed resource allocation and adaptive scheduling** across heterogeneous deployments.
  - Implemented and validated **Vehicle-to-Edge (V2X) communication frameworks** using AWS Edge Services, resulting in reduced latency and improved real-time responsiveness

- **False Discovery Rates (FDR) Control in Metaproteomics Search — C++, Python** 2021 - 2023
  Center for Computational Epidemiology and Response Analysis (CeCERA), https:// cerl.unt.edu/
  - Addressed systematic FDR biases in metaproteomics pipelines by incorporating **probabilistic modeling and statistical corrections**, reducing false-positive identifications across large proteomic datasets.
  - Proposed **"FineFDR"**: an open-source, fine-grained FDR assessment framework that seamlessly integrates with Comet and Percolator outputs at multiple taxonomic ranks. (https://github.com/Biocomputing-Research-Group/FDR)
  - Implemented the **Expectation-Maximization General-Mixture Model** for clustering proteomic samples based on abundance profiles, substantially enhancing the detection sensitivity for lower-abundance peptides.
  - Benchmarked six FDR control solutions (including Comet, Percolator, and Tailor) on ten diverse datasets, demonstrating notable gains in **precision and increased peptide/protein identifications** compared to state-of-the-art approaches.

Selected Publications

- **Semord: Learned Semantic-Preserving Placement and Low-Fanout Routing for Distributed Vector Search** Sept. 2026
  ArXiv preprint arXiv:2609.25514. Submitted. *First Author*
- **A Distributed Learned Hash Table** Feb. 2026
  IEEE/ACM Transactions on Networking (TON) *First Author*
- **Enhancing Faithfulness in Abstractive Summarization via Span-Level Fine-Tuning** Feb. 2026
  TrustNLP, ACL 2026; ArXiv preprint arXiv:2510.09915. *Co-Author*
- **PlanetServe: A Decentralized, Scalable, and Privacy-Preserving Overlay for Democratizing Large Language Model Serving** Sep. 2025
  2026 USENIX Symposium on Networked Systems Design and Implementation (NSDI) *Co-primary Author*
- **LEAD: A Distributed Learned Hash Table** Aug. 2025
  2025 IEEE International Conference on Network Protocols (ICNP) *First Author & Oral*
- **Vortex: Efficient Decentralized Vector Overlay for Similarity Search and Delivery** Aug. 2025
  *2025 IEEE International Conference on Network Protocols (ICNP); *BEST POSTER AWARD** *First Author*
- **Characterizing Perception Deep Learning Algorithms and Applications for Vehicular Edge Computing** Jan. 2025
  Algorithms 2025, 18(1), 31; Special Issue: Machine Learning for Edge Computing *Co-Author*
- **CALID: Collabrative Accelerate LLM Inference with Draft Model with Filter Decoding** May. 2024
  Poster at 2024 BayLearn - Machine Learning Symposium (Apple) *Co-Author*
- **Distributed Learned Hash Table** Sept. 2024
  2024 IEEE International Conference on Network Protocols (ICNP) *First Author & Poster*
- **Perception Workload Characterization and Prediction on the Vehicular Edges** Jul. 2023
  2023 IEEE International Conference on Edge Computing (EDGE) *Co-primary Author*
- **Fine-grained Taxonomy-specific False Discovery Rates Control in Metaproteomics** Nov. 2022
  2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) *First Author & Oral*

PROFESSIONAL SERVICES

- **Reviewer/AE Committee** IEEE/ACM TON, IEEE TDSC, ACM SIGCOMM, MLSys
- **Teaching Assistant** CSE 13S: Computer Systems and C Programming, CSE 121 Embedded System Design
- **Mentor** NSF Research Experiences for Undergraduates (REU) in Vehicular Edge Computing and Security