Zixuan Chen

zixuanc3@andrew.cmu.edu | (412) 980-3743 | github.com/Snowfall99 | linkedin.com/in/chenzx-isnowfall99

Education

Carnegie Mellon University **Pittsburgh, PA, USA**
Master of Science in Information Networking (MSIN) - Advanced Study August 2025 - May 2027 (expected)

- GPA: 3.95 / 4.00
- Selected Coursework: Storage Systems, Compiler Design, Database Systems, Machine Learning Systems
- Teach Assistant: 15-445/645 Database Systems (Fall 2026)

Shanghai Jiao Tong University **Shanghai, China**
Bachelor of Engineering, School of Computer Science September 2018 - June 2022

Work Experience

ByteDance **Beijing, China**
Senior Software Engineer, Cloud Computing Team July 2022 - April 2025

- **Designed and implemented topology-aware scheduling** for GPU and HPC compute nodes, reducing intra-node and inter-node communication latency; built resource reservation mechanism for VM migration and resize.
- **Maintained the production VM scheduler and resource controller**, handling millions of requests across multiple regions, supporting low-latency online provisioning and offline resource defragmentation.
- **Led end-to-end development of Nvidia L20 and A10 GPU instance types**, orchestrating component adaptations, compute node deployment workflows, hardware monitoring, and NCCL / vLLM performance evaluation.
- **Built event-driven failure handling and auto-scaling** for the VM service, including failure-handling workflows (migration, evacuation), event notification, horizontal/vertical scaling, and monitoring plugins for VM and GPU resources.

Projects

Cost-Efficient Multi-LLM Serving over Heterogeneous Cloud GPUs
CMU Catalyst Group, Advised by Professor Zhihao Jia and Rashmi Vinayak October 2025 - April 2026

- **Built the GPU profiling pipeline** feeding the offline Serving Template ILP, measuring per-batch latency across 5 GPU types, prefill/decode phases, and 6 models spanning dense, MoE, and hybrid-attention architectures.
- **Designed and implemented the runtime system**, a three-tier control plane (coordinator, per-instance schedulers, GPU workers) connected via ZeroMQ, supporting heterogeneous pipeline-parallel execution on vLLM, KV transfer between prefill and decode instances, and online instance life-cycle management.
- **Designed and performed end-to-end evaluation** on AWS EC2 platform across 6 models, demonstrating up to 2.79x cost reduction and 2.39x higher goodput compared to baselines.

Operator-Disaggregated LLM Serving over Heterogeneous GPUs
CMU Catalyst Group, Advised by Professor Zhihao Jia and Rashmi Vinayak October 2025 - present

- **Built the analytical profiling framework and cost model** that characterizes operator-level latency on heterogeneous GPUs across attention variants (full / sliding-window), GEMM/MoE operators, batch sizes, and parallelism strategies.
- **Developed the runtime system**, supporting operator-level partitions across heterogeneous device groups on vLLM, inter-stage activation transfer based on NCCL, and bipartite microbatch scheduling.

ThemiX: A Timing-balanced BFT Protocol (Middleware '23)
SJTU Decentralized Computing Lab, Advised by Professor Shengyun Liu August 2021 - May 2022

- **Co-designed ThemiX BFT protocol**, applying optimistic responsiveness and coin-bypassing optimizations, enabling fast path feature and allowing the protocol to proceed under real-world network latency.
- **Led the system development and evaluation**, implementing ThemiX protocol in Golang, and performing evaluations of fast-path performance and fault-tolerance behavior across 7 AWS regions worldwide.

Skills

Programming Languages: C/C++, Rust, Python, Golang
Tech Skills: vLLM, NCCL, CUDA, PyTorch, Triton, Kubernetes