

LIZHENG CHEN

✉ lizhengchen@smail.nju.edu.cn · ☎ (+86) 180-650-57797 · 🍪 Cookiecoolkid

🎓 EDUCATION

Nanjing University (NJU), Nanjing, China 2022 – Present
Incoming Ph.D. Student (Direct Ph.D. Track), NASA Lab Start: Sep. 2026
B.S. in Computer Science and Technology, Kuang Yaming Honor School. (Expected Graduation: June. 2026)

📄 RESEARCH

The Long and Winding Tail: Navigating Skewed Data in Out-of-Core Vector Search

Under Review at OSDI Co-first author; student first author Mar. 2025 – Present

Motivation. Large-scale **out-of-core** vector search amplifies SSD I/O under **semantic skewness**.

- **Scale mismatch:** Uniform local indexing fails, while out-of-core **rebalancing** is prohibitively expensive.
- **Routing drift:** Coarse routing over-probes irrelevant clusters when **hot regions** become pronounced.
- **Limited filter:** Incomplete cluster-level pruning and **lossy** vector-level pruning trigger wasted re-ranking.

Design. GANNS enforces **unified I/O governance** across **storage, routing, and filtering** layers.

- **Hybrid indexing:** Offline auto-profiler selects per-cluster index types under a global DRAM budget.
- **Query-aware routing:** Dynamic in-memory graph abstraction tracking hot regions from query trajectories.
- **Prune-before-fetch:** Multi-level pruning bounds strategies prune candidates before SSD reads.

Results. At matched recall, GANNS achieves $>4.7\times$ higher QPS than DiskANN/Starling and up to $5.2\times$ higher QPS than PipeANN; compared with SPANN, it delivers up to $17.2\times$ higher QPS and $25.0\times$ lower latency.

GPU-Accelerated Graph Index Construction for ANNS via Compute–Memory Disaggregation

Under Review at SIGMOD Contributing author (non-lead) Feb. 2025 – Present

CMANNS is a GPU-centric framework that accelerates graph index construction via *compute–memory disaggregation* and out-of-core pipelining, enabling fast, memory-efficient builds while preserving high recall.

⚙️ PROJECTS

BusTub (CMU 15-445): Disk-Oriented Single-Node RDBMS

Jan. 2025 – Mar. 2025

C++, Concurrency, Storage & Index GitHub: Private (course policy)

Implemented core components of a disk-backed relational database system. Built an **LRU-K buffer pool manager** with *RAII page guards* for safe page management and efficient disk I/O. Implemented an **extendible hash index** supporting dynamic CRUD via directory growth and on-demand bucket splits. Developed the query engine with **execution operators & optimizer rules** (e.g., aggregation, predicate pushdown). Implemented transactional support with **MVCC**, including *undo logging and garbage collection* for concurrent execution.

ICS-PA (NJU): Software-Simulated Computer System

Sep. 2023 – Jan. 2024

C, ISA, OS, Emulator GitHub: Cookiecoolkid/ICS-PA

Built an end-to-end interactive computer system stack in software, from ISA-level emulation to a minimal OS. Implemented the **NEMU emulator** with instruction decode/execute, memory mapping, and port I/O as the execution foundation. Developed **AbstractMachine (AM)** to abstract *context switching, device access, and event handling* behind standardized interfaces. Built **NanoOS-lite** with *process scheduling, syscall handling, ELF loading*, and a simple file-system abstraction that maps file descriptors to disk-backed files.

★ EXPERIENCE & HONORS

Tencent Technology, Summer Intern (Developer), Shenzhen, China

Jul. 2025 – Sep. 2025

CET-4: 580 CET-6: 537 CCF-CSP: 290 (Top 5% overall)

Nanjing University People Scholarship, Recipient

2024, 2025