

Jun Kong

Email: jkong123@asu.edu | Websites: [Homepage](#) [Github](#)

RESEARCH INTERESTS

- Database and Storage Systems: indexing structures and storage engines,
- LLM Inference — KV cache management and offloading for efficient LLM serving.

EDUCATIONS

Arizona State University

Arizona, US

Ph.D in Computer Science, advised by [Prof. Zhichao Cao](#)

July 2025 - PRESENCE

Lanzhou University

Lanzhou, China

Bachelor in Computer Science

SEP 2019 - JUN 2023

- Outstanding Graduation Thesis Award (Top 10%)
- First-Class Scholarship (Top 5%)

ONGOING RESEARCHES

Index Design for Large-Scale KV Cache Storage in LLM Inference

May 2025 - PRESENT

- Conducting research on large-scale KV cache storage systems for LLM inference, where persistence and index scalability are becoming critical challenges.
- Investigating efficient on-SSD index acceleration techniques to mitigate latency bottlenecks when index structures exceed main memory capacity.
- Designing and evaluating prototype systems to explore performance trade-offs between memory usage, throughput, and storage efficiency.

Optimizing MVCC Performance in FoundationDB Storage Servers

AUG 2025- PRESENT

- Profiling CPU overhead in FoundationDB storage servers caused by PTreap-based multi-version data maintenance during MVCC.
- Designing index structures and evaluating optimizations to reduce computational overhead.

INDUSTRY EXPERIENCE

[Megvii Technology](#)

JUN 2023 – JUL 2025

Software Development Engineer

I worked on large-scale storage systems at [MEGVII](#)(A leading Chinese AI company), focusing on the intersection of storage and AI workloads.

Overlay: An Self-Developed Object Storage System (OSS) with 100PiB data in production.

- Diagnosed and resolved numerous deep concurrency and performance issues in a 100+ PiB scale object storage system, involving race conditions, broken heartbeats.
- Developed an asynchronous I/O abstraction layer using `io_uring` to address the Go runtime's over-spawning of OS threads under high-concurrency file I/O.
- Migrated the storage system deployment to Kubernetes platform.

Nori: small-file accelerating system for ML training with 600GB/s and 500k QPS.

- Implemented priority-based scheduler, reducing waiting time for most tasks from hours to seconds.
- Redesigned database access patterns and improves system capacity from 800K tasks to 3M tasks.

UNDERGRADUATE PROJECTS

LZU OS: A 64-bit education-oriented OS kernel from scratch [LINK](#)

NOV 2020 - APR 2022

Lanzhou University, Advised by Associate Prof. Li Liu

- Wrote code of device tree, inline assembly, virtual memory, process management, memory allocator.
- Ported buddy-system allocator and SLAB aKubernetesllocator from Linux 2.6 to this system.
- Wrote tutorials to teach how to write an OS step by step.
- Won Excellence Award, China National College Student Computer System Ability Contest 2021.

Simplified Ibex RISC-V Core Re-implementation [LINK](#)

JUN 2021 - MAR 2022

Lanzhou University, Advised by Associate Prof. Anping He

- Engineered a simplified RISC-V (RV32I) processor core from scratch using the Chisel Hardware Description Language, leveraging the open-source Ibex CPU architecture as a reference model.
- Verified functionality through simulation and FPGA synthesis.

INTERNSHIP EXPERIENCE

[Megyii Technology](#)

JUN 2022 - OCT 2022

Software Development Intern

I worked on a Kubernetes orchestration and package management system similar to Helm.

- Added support for OCI(*Open Container Initiative*) standard and transformed packages to OCI artifact format.

AWARDS AND HONORS

- Lanzhou University Second-Class Scholarship (Top 10%)
- Excellence Award, China National College Student Computer System Ability Contest 2021.

SKILLS

- Proficient in C, Go, Python and Bash.
- Experienced in Kubernetes.