

Zhuohang Bian

Peking University | School of Integrated Circuits | Ph.D. Student



I Profile

Phone / WeChat: 13656160098

Email: bianzhuohang26@stu.pku.edu.cn

B.Eng.: Beihang University - School of Computer Science and Engineering - Computer Science and Technology (2022-2026)

Ph.D.: Peking University - School of Integrated Circuits - Electronic Design Automation and Computing Systems (2026-)

I Summary

Research focus on LLM inference systems and computer architecture, with current work on KV cache management and scheduling for multi-agent applications. Experienced across workload analysis, system design, implementation, and evaluation.

- **Undergraduate major rank: 19/239 (top 10%); average score: 92.54; GPA: 3.86/4**
- **Selected core courses:** Object-Oriented Programming 100, Intelligent Computing Architecture 99, Probability and Statistics 99, Data Mining 97, Computer Organization 95...
- Advanced TA for **Computer Organization** and TA for **Operating Systems** at BUAA School of Computer Science; compiler course project adopted as teaching example in national compiler-course teacher training.

I Research Experience

Primary Focus: Multi-Agent LLM Inference Systems Peking University • Zhuo Lab • Feb 2025–Present

I TokenCake: KV Cache Scheduling for Multi-Agent Applications

First Author • EuroSys 2027

- Identified **temporal idleness** during function calls and **spatial contention** under concurrent agents, including critical-path caches being displaced by non-critical tasks.
- Designed agent-aware **Time/Space Schedulers** with proactive offload, predictive upload, and dynamic memory partitioning; reduced end-to-end latency by **47.06%** over vLLM and improved effective GPU KV cache utilization by **16.9%**.

I TokenDance: Collective KV Cache Reuse and Compression

First Author • Preprint

- Characterized the All-Gather pattern in synchronous multi-round interactions, where position shifts prevent request-level prefix caching from exploiting shared outputs and create redundant computation and memory use.
- Designed a **KV Collector** and **Diff-Aware Storage** for round-level collective reuse and block-sparse differential storage; achieved **11–17×** compression, **2.7×** more concurrent agents, and **1.9×** faster prefill.

I TokenSim: Hardware/Software Simulator for LLM Inference

Second Author • APPT 2025

- Co-developed a modular simulator for dynamic real-world workloads with extensible **scheduling, memory management, and operator-level performance models**, reporting latency distributions and time-varying memory usage.
- Validated simulation error at **below 1%** against real systems and enabled design-space studies of P/D disaggregation, heterogeneous hardware selection, memory quotas, and long-conversation memory pools.

State Key Laboratory of Complex & Critical Software Environment, Beihang University Prof. Xianglong Liu's Group Safe Reinforcement Learning

Jul 2024-Jan 2025

- Built **AMB**, a robustness benchmark for multi-agent reinforcement learning, covering **4 real-world environments, 13 uncertainty types, and 15 key hyperparameters** and supporting **82,620** experiments. Hyperparameter optimization alone improved cooperation by **52.60%**, robustness by **34.78%**, and resilience by **60.34%** on average.

- Implemented attack algorithms and experiment pipelines; contributed to result analysis and paper writing, especially on the relationship between cooperation and resilience.
- NeurIPS 2025 paper: [arXiv:2510.11824](https://arxiv.org/abs/2510.11824)

Sino-German Joint Software Institute, Beihang University Prof. Zhongzhi *Sep 2023-May 2024*
Luan's Group High-Performance Computing

- Optimized Multi-Head Attention in PyTorch for the **MT-3000 CPU-DSP heterogeneous processor** on the new-generation Tianhe supercomputer; implemented DSP-specific optimizations for core operators such as **Linear / Softmax** with matrix tiling, DMA double buffering, communication optimization, and head-level operator scheduling.
- Handled operator registration with PyTorch 1.8 C++ extensions and performance tuning/evaluation on MT-3000; completed the Linear operator implementation.
- Outstanding Paper, HPC China 2024: Multi-Head Attention Optimization on the MT-3000 Processor.

I Projects

SenseTime Internship - LightLLM LLM Inference Acceleration Framework *Jun 2025-Oct 2025*

Developed algorithms and framework components for LLM inference systems on [LightLLM](#); merged **6 public PRs**, focusing on **prefill-decode disaggregated scheduling, KV cache migration, and multimodal model support**.

- Refactored the PD Master node-selection path, designed an extensible selector mechanism, and added node-load-prediction-based **memory-aware routing** with load information reporting to mitigate Prefill / Decode imbalance.
- Extended KV cache migration for P/D disaggregation: added **timeout retry and KV return** for chunked decode, and implemented **CUDA/NCCL-based kv_trans_v2** cross-GPU migration in multi-DP settings.
- Integrated the **MinerU2** multimodal model into LightLLM, covering tokenizer, image processor, vision modules, visual RPC, and server APIs; added observability for token usage and server ID.
- Links: repository [LightLLM](#); PRs [#944](#) / [#970](#) / [#999](#) / [#1011](#) / [#1024](#) / [#1034](#)

2024 National Computer System Design Competition, Loongson Cup CPU Track - *Jul 2024-Aug 2024*
4th Place / Second Prize

- Implemented a **13-stage out-of-order dual-issue superscalar processor** supporting LoongArch32R except floating point, with branch prediction, L1 Cache, and TLB; booted Linux, reached **80 MHz** on FPGA, achieved average IPC of **1.02**, and placed 4th nationally.
- Owned key backend modules and integration in a four-person team, including **ALU/LSU/MDU Issue Queues, Commit logic, and memory-access state machines**.
- Repository: [BOOM_chip](#)

I Honors and Awards

- | | |
|---|---|
| • Outstanding Graduate of Beijing (2026) | • AVIC Scholarship ; Qisan Medal Scholarship |
| • First Prize , CUMCM (National, 2024) | • 4th Place / Second Prize , Loongson Cup (2024) |
| • Meritorious Winner , MCM (2024) | • Second Prize , NSCSCC (2025) |
| • First Prize , Beijing Physics Competition (2023) | • Outstanding Student Leader , BUAA (2023–2025) |
| • BUAA Academic Excellence First Prize; Competition Top Prize | • BUAA Innovation Second Prize; Outstanding League Cadre; Outstanding Student |

I Languages and Skills

English: CET-6: 589

Homepage: zhhangbian.github.io

Programming: Python / C++ / CUDA / SystemVerilog

Systems: LLM serving, distributed inference, processor design, reinforcement-learning platforms; maintained the BUAA Computer Organization teaching system.