HONGHAO JIA

+86 13069804665 | 1390751361@qq.com | Beijing, China | github.com/LoveLonelyTime

EDUCATION

Institute of Computing Technology, Chinese Academy of Sciences — Beijing, China
Jointly-Trained Master in State Key Lab of Processors; Research: AI Infra — *Sep. 2025 – Expected Jun. 2027*

University of Science and Technology of China — Hefei, China
Master of Microelectronics; GPA: 3.59 / 4.3 — *Sep. 2024 – Expected Jun. 2027*

Harbin Engineering University — Harbin, China
Bachelor of Computer Science and Technology; GPA: 3.82 / 4.3 — *Sep. 2020 – Jun. 2024*

PROFESSIONAL EXPERIENCE

AI Accelerators for Efficient Quantized Inference of LLMs | Research Project — *Sep. 2025 – Present*

- The research project is to design a **predictive quantization** with hardware-software Co-Design for on-device inference, aiming to develop a novelty predictive quantization hardware to balance accuracy and memory footprint reduction.
- Reproduced and evaluated SOTA LLM quantization algorithms, including **KIVI**, **GPTQ**, and **AWQ**, …
- Deployed and benchmarked LLMs (**BERT**, **Qwen**, **Llama**, …) using **PyTorch** and **Hugging Face Transformers**, covering model loading, quantization, inference execution, and performance evaluation.
- Investigated and implemented **low-level GPU operators** for LLM inference, including **GEMM**, **GEMV**, and **QUANT-PACK**, … with practical experience in **CUDA** and **Triton** kernel development and performance tuning.

Cambricon Technologies Co., Ltd. | IC Backend Engineer Intern — *Sep. 2025 – Dec. 2025*

- Designed and optimized a **12 nm CMOS standard cell library** focusing on area efficiency, using **Cadence Virtuoso**, **Synopsys DC**, **ICC** across the full custom digital design flow.

Bergamot: A Superscalar RISC-V RV32GC Processor | School-level Project — *Nov. 2023 – Jun. 2024*

- Implemented and validated Bergamot, **a full-scale RISC-V RV32GC processor** capable of booting and running mainstream **Linux kernel** on **Xilinx FPGA**.
- Implemented the full RISC-V instruction processing pipeline (Frontend, Execute, Backend) using **Chisel HDL**, achieving **~74%** reduction in code size compared to Verilog, while improving design modularity and reusability.
- Designed a multi-level, dynamic, **2-way superscalar** microarchitecture with **4 parallel functional sub-pipelines** (Float, ALU, Branch, Memory), enabling execution of up to a **9-level pipeline** and sustaining a peak throughput of **2 IPC**.
- Implemented a **MMU** supporting the **Sv32 virtual memory scheme**, including fully functional ITLB and DTLB, enabling Linux virtual memory, process isolation, and address translation.
- Designed and integrated a **branch prediction** component with a **BTB-based** structure, achieving an average branch prediction accuracy of **~91%**, significantly reducing control hazards and pipeline flushes.
- Implemented a **hierarchical cache component**, including L1 instruction and data caches (2-way set associative, 32 KiB I-cache / 32 KiB D-cache) and a 4-way set associative 512 KiB L2 cache, achieving **~67%** hit rate for local accesses.
- Developed an **interrupt handling** and **MMIO** framework, supporting the **AXI4 bus protocol**, allowing flexible integration of external peripherals (SPI, I2C, GPIO, …).
- Achieved **160+ stars** on GitHub [github.com/LoveLonelyTime/Bergamot].

Software Research Institute, China Unicom Co., Ltd. | Software Developer Intern — *Jun. 2022 – Sep. 2022*

- Implemented a user service order (eSIM) information management system to support internal experimental operations, enabling efficient order tracking, lifecycle management, and real-time service status inquiry for end users.
- Developed a front-end application for clients using the **WeChat Mini Program framework**, leveraging **WXML**, **JavaScript**, and **WXSS** to deliver a responsive and lightweight WeChat Mini program across mobile devices.
- Redesigned and normalized database schemas, optimizing table relationships and reducing **~23%** user-related data redundancy.
- Applied **MyBatis** advanced mappings (association, collection, cache) to optimize database access and significantly improve **3.4×** SQL query efficiency.
- Implemented a multi-level caching architecture (**Redis + MyBatis cache**) to accelerate access to high-frequency temporal data, reducing average response latency by **4.2×**.
- Implemented an event-based service system using **Kafka + MQTT** to support **IoT** device message queue.
- Participated in **CI/CD pipeline** maintenance, supporting automated build, testing, and deployment.

SKILLS

- **Languages:** Chinese (Native), English (CET6: 512, TOEFL: 88, TOEIC: 790), Japanese (JLPT N1: 109)
- **Programming:** Python, Java, Web Development (JS/TS, H5, CSS, Vue), Spring Framework, C/C++, Rust, Verilog
- **Software:** VS Code, Linux, MySQL, Docker, Git, CI/CD, PyTorch, Hugging Face Transformers
- **Soft Skills:** Independent Research, Teamwork, Academic Writing

AWARDS & HONORS

Bronze Award in the ACM/ICPC (International Collegiate Programming Contest), Asia Regional Contest. — *2022, 2023*
Gold Award in the CCPC (China Collegiate Programming Contest), Heilongjiang Provincial Contest. — *2023*
Silver Award in the CCPC, Northeast Three Provinces Contest. — *2023*
First-Class Academic Scholarship at Harbin Engineering University. — *2022*
First-Class Academic Scholarship at the University of Science and Technology of China. — *2024*