# Yunkai Wang

Shanghai China  yk.wang@sjtu.edu.cn  (+86)159-7988-5716  runoobb

## Education

**Shanghai Jiao Tong University** *Sept 2022 – Jun 2026*
*BS in Microelectronics Science and Engineering*

- GPA: **3.9/4.3**; Ranking: **6/74**
- **Selected Courses:** Data Structure (93/100), Discrete Mathematics (97/100), Design of Digital Integration Circuits (93/100), Design of Artificial Intelligence Chip (91/100), Computer Architecture (Ongoing)
- **Self-study Courses**: Introduction to Computer System, Computer Architecture, Parallel Computing and Architecture Deep Learning Systems,

## Research Experience

**Model Compression and Domain Specific Accelerator** *Feb 2024 – Oct 2024*

- Learn Pytorch and Deploy LLM profiling attention score distribution
- Conduct model compression on LLM
- Implement mixed precision PE array in RTL(code)

**Accelerate LLM Inference on GPGPU** *Nov 2024 – March 2025*

- Implement and profile common LLM CUDA kernels (code)
- Research on sparse attention pattern and acceleration

## Projects

**NEMU(NJU Emulator) (code)** *Sept 2023 - Jun 2024*

- Build a system emulator to support riscv32e architecture using C and create functions to run PC console games like Super Mario
- Implement a CPU core capable of executing riscv32e instructions, functions to emulate I/O devices, interrupts, OS booting

**GPGPUSim** *Dec 2024 - March 2025*

- Manipulate GPGPUSim v4.0.0 and an extended version modeling multi-GPU and nvlink system
- Distributed LLM inference acceleration on multi-GPU system(Computation communication overlapping)
- Reproduce microarchitecture papers in the future (Ongoing)

**CNN Accelerator(Course Project)** *April 2025 - June 2025*

- Estimation bit-serial EDDO architecture performance and energy on DnnWeaver2 (Ongoing)
- RTL implementation of CNN accelerator based on Eyeriss (Ongoing)

## Honors and Awards

**SJTU Third-Class Scholarship** *Sept 2023 – Jun 2024*

## Technologies

**Languages:** C++/C, CUDA/Triton, Python/Pytorch, (System)Verilog

**Tools:** GPGPU-Sim, HuggingfaceLLM, Synopsys Design Compiler, Vivado, Verilator, Modelsim