Yuhao Ge

Mobile: (217)926-3291 | Email: yuhaoge2@illinois.edu | Portfolio | LinkedIn

Education

University of Illinois at Urbana-Champaign M.Sc. Computer Science GPA: **4.0**/4.0 Aug. 2023 - May 2025

University of Illinois at Urbana-Champaign B.Sc. Computer Engineering GPA: **3.95**/4.0 Aug. 2019 - May 2023

- Honors: Highest Honors, Bronze Tablet (3%, 2023), Dean's List (2020&2022)

Zhejiang University B.Eng. Computer Engineering GPA: **3.93**/4.0 Aug. 2019 - May 2023

- Honors: Zhejiang Provincial Government Scholarship, First-Class Scholarship (3%, 2020&2022)

Skills

- **Programming:** Python, C/C++, JavaScript, SystemVerilog, Assembly, SQL
- **Frameworks & Tools:** CUDA, PyTorch, TensorFlow, TVM, Triton, PyG, Flask, React, Node.js, Kubernetes, Grafana, Airflow, AWS, GCP, MongoDB, Neo4j, Redis

Work Experience

Amazon, Annapurna Labs | *Software Engineer Intern* | *Compiler, ML* May 2024 - Present

- Neuron Compiler | *Optimize deep learning workloads on AWS Trainium and Inferentia chips*
  - Developed infrastructure for automatic kernel generation, compilation, profiling, and visualization with a defined sweep space
  - Collected data for **DMA pattern** analysis and optimized the **DMA latency model**
  - Build the first-generation **autotuning** framework from scratch to support continuous **compilation optimization**

University of Illinois, Urbana-Champaign | *Student Researcher* | *Compiler, GPU, MHSA* Aug. 2023 – Feb. 2024

- SPLAT | *Optimized GPU code generation framework for SParse reguLar Attention*
  - Developed SPLAT, an optimized framework for efficient **sparse-MHSA**, targeting moderate sparsity levels
  - Introduced **Affine Compressed Sparse-Row** (ACSR) format for **regular sparsity** patterns in MHSA
  - Engineered advanced **GPU code-generation algorithms** for ACSR, enhancing sparse-MHSA kernel performance
  - Achieved 2.05x and 4.05x speedups over **Triton** and **TVM** kernels with SPLAT implementation

TikTok | *Software Engineer Intern* | *C++, Lua, Game Engine, AR/VR* May 2022 - Aug. 2022

- Amazing Engine | *TikTok's Next-Generation 3D Game Engine for AR/VR Effects*
  - Collaborated in developing TikTok's **3D Game Engine**, which empowers users to create/use interactive **AR/VR** stickers
  - Implement a query-based animation system **Motion Matching** in **C++** for realistic and responsive avatar control
- Skeleton Retargeting System | *An SDK for Skeleton Retargeting*
  - Developed an **SDK** for **Skeleton Retargeting** in **C++** and **Lua**, supporting animation adaptation across character models
  - Integrated the cross-functional team's **Text-to-Animation** algorithm into our game engine using the developed SDK

University of California, Los Angeles | *Visiting Student Researcher* | *ML, RL, GNN, FPGA, EDA* June 2022 - Nov. 2022

- GNNDSE | *An automated design space exploration for automatic FPGA accelerator design* | *Advisor: Prof. Jason Cong*
  - Combined **GNN** with an **ML/RL**-based **Design Space Exploration** to achieve **FPGA** Accelerator **Design Automation**
  - Developed a learning-based **Cost Model** with **GNN** as a surrogate of the HLS tool for quick and accurate assessment
  - Optimized DSE by deploying heuristic algorithms such as **Genetic Algorithm** and **Simulated Annealing**
  - Used **Reinforcement-Learning** and **Bandits** for automatic algorithm selection, boosting exploration speed by 11%

Projects

An NFT pricing service powered by machine learning Feb. 2023 - June 2023

- Worked at **NFTGo startup** on ML-based NFT price prediction; developed the GoPricing service from scratch
- Created a **Regression Model** for NFT pricing and packaged services into APIs using the **FastAPI** web framework
- Streamlined periodic data processing, model training/updating, and monitoring with Apache **Airflow**
- Supported over 3000 collections, achieving Mean Absolute Percentage Errors (MAPE) of 3%-8%

Remote Car Control System with Real-time 3D Reconstruction Jan. 2023 - May 2023

- Developed a Raspberry Pi robot car with remote control via joysticks, utilizing PID control and STM32 microcontroller
- Implemented WiFi-based communication for transmission of commands and RGBD images between the car and server
- Implemented the **SLAM** framework RTAB-Map on the server for real-time **3D reconstruction**, achieving a 10Hz framerate

Linux-like OS Kernel Design Jan. 2022 - May 2022

- Designed a Linux-like operating system in **C** and **Assembly** with a GUI supporting multi-terminal and mouse control
- Support System calls, Memory paging, Scheduling, Interrupt handling, Device drivers, Signal, TCP connection, file system

Implement A Game Efficiently on the FPGA Board Sept. 2021 - Nov. 2021

- Ported the game "Doodle Dump" to FPGA with **SystemVerilog**, achieving low power consumption and high efficiency
- Implemented a SOC with NIOS II in **C** to manage complex tasks like USB protocol and memory I/O
- Consumed only 400KB memory, 0.5w power to achieve a 50hz frame rate, won the **Best Design** Prize