Sirui (Ray) Chen

(802)-771-5288 | ray_chensirui@hotmail.com

PROFESSIONAL SUMMARY

Exploration-driven and technically versatile Computer Science and Mathematics graduate student with strong foundations in computer architecture, machine learning, and deep learning. Experienced in Python, C, C++, and parallel programming for high-performance computing. Hands-on with CUDA, TensorFlow, and PyTorch, as well as backend service development and front-end interface design. Adept at bridging research with real-world applications through performance optimization and architecture-aware AI implementations. Passionate about leveraging GPUs and advanced architectures to accelerate deep learning and high-performance workloads.

EDUCATION

Northwestern University, Evanston, IL, U.S Sep 2024 – Expected Dec 2025

*Master of **Computer Science***

- **GPA:** 4.0 / 4.0

Middlebury College, Middlebury, VT, U.S. Sep 2020 – May 2024

*Bachelor of Arts in **Computer Science** & **Mathematics***

- **GPA:** 3.72 / 4.0
- **Honors:** First-Class Honor Graduates, multiple Dean's List recognitions

WORK EXPERIENCE

TAL Education Group, Beijing, China Jun 2025 – Sep 2025

NLP Engineer Intern, AI Solution Team | **Python, C, C++**

- Developed large-scale LaTeX syntax detection and correction models for AI-based problem-solving systems, optimized data preprocessing and inference workflows to handle millions of records.
- Designed and implemented evaluation pipelines in Python and C++ to align model inference with human annotations, achieved >90% in accuracy, recall, and F1-score.
- Leveraged C++ for high-efficiency text parsing modules and integrated CUDA-based GPU acceleration in preprocessing tasks, improved throughput and reduced latency.
- Reduced false detection rate from 51.67% to 14%, with rendering success rate at 100%.

Middlebury College, VT, USA Mar 2023 – Dec 2023

Teaching Assistant and Grader of Computer Science Department | **Python, C++**

- Assisted in courses such as Computer Architecture, Theory of Computation, and Algorithms and Complexity.
- Designed five automated grading tools in Python and C++ to detect errors and standardize evaluation, reduced manual grading workload by 50%.
- Provided office-hour academic support, clarified complex topics in computational models, algorithmic optimization, and memory hierarchy design.

PROJECTS

Parallel Scan Algorithm Design and Performance Analysis in CUDA Mar 2025 – Jun 2025

Team Lead, Northwestern University | **CUDA, C++**

- Designed and implemented a work-efficient parallel scan (prefix sum) algorithm in CUDA to process 16M+ elements, achieving 15.1× speedup over a single-threaded CPU baseline.
- Applied advanced GPU optimization techniques, including a hierarchical multi-block kernel strategy and thread-level multi-element processing, increasing arithmetic intensity and throughput.
- Optimized memory access by leveraging shared memory to minimize global memory traffic and introducing conflict-free padding to eliminate bank conflicts, demonstrating strong understanding of GPU memory architecture.
- Performed detailed performance profiling and FLOPS analysis, comparing CPU (1.53 GFLOPS) and GPU (23.22 GFLOPS) execution, and identified key bottlenecks such as memory bandwidth and synchronization overhead.

AI-Powered Educational Game Recommender System Sep 2024 – Dec 2024

Lead Developer, Northwestern University | **Python, Neural Networks, CUDA**

- Processed 10,000+ game datasets with reviews and user feedback to train a FFNN for personalized recommendations.
- Applied CUDA-accelerated TensorFlow training to speed up experimentation cycles.
- Achieved 80% prediction accuracy and 95% recommendation precision.

Generative AI-Driven Learning Assistant Website Jan 2025 – Mar 2025

Team Lead, Northwestern University | **Generative AI, Firebase, React, JavaScript, HTML**

- Developed a quiz generation platform powered by OpenAI models with optimized inference logic for latency reduction.
- Integrated a Firebase backend with scalable data handling to ensure high availability during peak usage.

SKILLS

- **Programming Languages:** Python, Java, C/C++, JavaScript
- **Front-End Development:** React, HTML5, CSS3
- **Back-End Development:** Spring Boot, Node.js
- **Parallel & GPU Computing**: CUDA, OpenCL
- **Machine Learning:** TensorFlow, PyTorch
- **Database:** MySQL, MongoDB

PERSONAL INTERESTS

Badminton, Soccer, Snowboarding, Cooking, Traveling, Boardgames.