# Joseph Li

Email: joeyorjoey@gmail.com

## Education

**Carnegie Mellon University** Pittsburgh, PA
B.S. in Computer Science; GPA: 3.91 Aug 2019 - May 2023

- **Courses**: Parallel Computer Architecture & Programming (TA), Compiler Design & Implementation
  Deep Learning, ML Systems, Data Structures & Algorithms, Digital Systems

## Experience

**NVIDIA Deep Learning Algorithms Intern** May 2022 - Aug 2022 (Remote)

- Integrated 8-bit floating point (FP8) and BFloat16 (BF16) arithmetic into BERT for faster training
- Developed and evaluated FP8 training recipe for PyTorch's automatic mixed precision extension
- Added BFloat16 support to experimental LAMB optimizer using CUDA

**IBM Systems Research Intern** May 2021 - Aug 2021 (Remote)

- Designed an ML pipeline scheduler for parallel cloud cluster training to improve training speed
- Implemented a prefix tree-based asynchronous work-stealing system in Python and Ray
- Resulted in 3.5x speedup in execution time and 2x improvement in memory usage compared to naive algorithm

**Bear Robotics Intern** Jul 2019 - Aug 2019 (Redwood City, CA)

- Optimized 3D vision pipeline using CUDA for uniform voxel downsampling on GPU in order to reduce CPU load
- Improved global path planning algorithm for autonomous robot navigation in C++ and ROS

## Projects

**C0 to x86 SSA Compiler with Fork-Join Parallelization Extension** Aug 2021 - Dec 2021

- Designed and implemented compiler front-end, middle-end, and back-end in Rust.
- Added a compiler extension that implements spawn/sync multithreaded parallelism through pthreads and cactus stack management routine
- Optimizations: SSA, copy propagation, dead code elimination, constant folding, graph-coloring register allocator

**OCaml LLVM Backend** Feb 2022 - Present

- Writing Cmm to LLVM pass for OCaml. Collaboration with Jane Street

**Pystreaming: A Lightweight Python Package for Audio and Video Streaming** Dec 2019 - Present

- Developed Python package for low-latency distributed inference across computer clusters to facilitate real-time AI tasks
- Leverages ZMQ messaging libraries, multi-core parallelism, and implements safe failure patterns
- Available for download through PyPI: https://pypi.org/project/pystreaming

**Warehouse Robots: GPU-Accelerated Training for Reinforcement Learning** Mar 2021 - May 2021

- Wrote custom CUDA kernels to simulate a multi-agent grid-world environment
- Achieved 17x speedup in overall training, 71x speedup in rendering, 103x speedup in environment simulation

## Leadership

**Roboclub: Robobuggy Chairman** Mar 2022 - Present

- Managed mechanical, electrical, and software teams to create an autonomous vehicle for CMU's annual buggy race
- Coordinated with Carnegie Mellon Sweepstakes to plan new rules, practice times, and event logistics

**Teaching Assistant for Parallel Computer Architecture and Programming** Fall 2021, Fall 2022

- Wrote and tested a new two-part VLSI-routing assignment for over 100 students
- Held office hours, recitations, and exam reviews to teach CUDA, OpenMP, MPI, and various parallel architecture concepts

## Skills

**Languages:** C++, C, Python, Rust, SystemVerilog, **Python Frameworks**: PyTorch, PySpark, Ray, Sklearn, ZMQ