Yen-Shi Wang

(+1) 412-218-9816 | yenshiw@gmail.com | yen-shi.github.io | yen-shi | yen-shi

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Electrical and Computer Engineering, GPA: 3.93/4.0 — Dec. 2020
Coursework: Foundations of Computer Systems, Optimizing Compilers, Cloud Computing, How to Write Fast Code

National Taiwan University — Taipei, Taiwan
Bachelor of Science in Computer Science and Information Engineering, GPA: 3.85/4.0 — Jan. 2019
Bachelor of Science in Electrical Engineering, GPA: 3.85/4.0 — Jan. 2019
Coursework: Algorithm Design and Analysis, System Programming, Operating Systems, Deep Learning, Multimedia Analysis

Experience

NVIDIA — TensorRT Team — Remote work from Pittsburgh, PA
Performance Software Engineering Intern — May. 2020 - Aug. 2020

- Improved C++11 multithreading server for MLPerf Inference BERT benchmark to scale linearly from 1- to 20-GPU machines.
- Optimized GPU utilization with CUDA streams and graphs, solved runtime bugs on CPU and GPU, boosted throughput by 25%.
- Actively updated internal documents, involved in group channels and discussions, and worked as a team in remote environment.

Carnegie Mellon University — Pittsburgh, PA
Teaching Assistant — Cloud Computing — Jan. 2020 - May. 2020

- Managed an AWS state machine to automatically generate similarity reports on student's submissions of 10 projects.
- Containerized frontend of quiz cheat checking system written with Django into Docker image and deployed to AWS ECS.
- Answered questions range from Linux, Hadoop, Spark, AWS Auto Scaling, MySQL, Azure Functions, Docker, to Kubernetes.

Skymizer — Taipei, Taiwan
C++ Developer — worked on Open Neural Network Compiler (ONNC) — Apr. 2019 - Jul. 2019

- Rewrote 21 optimizations for deep learning models from ONNX, added testing framework from scratch, and ported into ONNC.
- Initiated quantization flow in ONNC backend to perform 8-Bit quantization for NVIDIA Deep Learning Accelerator (NVDLA).
- Introduced per-channel symmetric quantization, resulted mean squared error is hundreds times smaller than per-layer method.

BravoAI Co., Ltd. — Taipei, Taiwan
Software Engineer — focused on Optical Character Recognition — Mar. 2018 - Sep. 2018

- Developed a system using Pytorch and Tensorflow to convert fields on medical certificate from paper into electronic forms.
- Deployed entire system with four Docker containers running Flask web service, operating at a speed of 0.5 image/sec.
- Obtained per-character accuracy of over 95% and sold to two biggest insurance companies in Taiwan.

Skills

Programming Languages	C++, C, Python, Java, Bash, Javascript
Tools and Others	Linux, Git, GDB, Make, Docker, CUDA, Terraform, AWS, Hadoop, Spark, MySQL

Projects

Distinctness Analysis in LLVM for C/C++ (final project in Optimizing Compilers at CMU) — Mar. 2020 - May. 2020

- Created an LLVM Module Pass to generate function call graphs and perform Andersen's pointer analysis.
- Read LLVM doxygen, grew familiar with LLVM Infrastructure and dealt with Functions, Loops and at least 10 Instructions.
- Distinctness can be used to identify distinct variables inside loops and hence parallelize computation in loops.

Cache, Malloc, and Shell Labs (projects in Computer Systems at CMU) — Sep. 2019 - Nov. 2019

- Deepened understanding of memory hierarchy, virtual pages, and best practices to write cache and memory efficient code.
- Implemented C function malloc with doubly linked segregated lists and first fit algorithm to achieve 74% memory utilization.
- Designed a simple Linux shell supporting background jobs, signals handling, and I/O redirection with command line parser.

Honors

2018	**Rank 116**, Google Code Jam 2018, Round 1C	
2017	**Silver Medal**, ACM-ICPC Asia Hua-Lien Regional Contest	Hua-Lien, Taiwan
2017	**Third Prize**, National Collegiate Programming Contest 2017	Taipei, Taiwan
2013	**Silver Medal**, 54th International Mathematical Olympiad (IMO)	Santa Marta, Colombia