

**EDUCATION****PEKING UNIVERSITY**

Beijing, China

09/2019–present

*School of Electronic Engineering and Computer Science*

- Major in Computer Science (Turning Class), Overall GPA: **3.82**/ 4.0 (Ranking 7/200);
- Programming & Framework: Expert in C++, Python; TVM, Dynet, CUDA, MPI, OpenMP
- Research Interest: Machine Learning System, Compiler Optimizations
- English Proficiency: TOEFL 109 (Speaking 23), GRE Verbal 160, Quantitative 169, Writing 3.0

*Awards and Honors*

- 1<sup>st</sup> Prize in ASC22, 2022
- Jiukun Scholarship (Academic Excellence Scholarship), 2021
- Shenzhen Stock Exchange Scholarship (Academic Excellence Scholarship), 2020
- Merit Student, Peking University, twice
- 1<sup>st</sup> Prize in Chinese University Mathematical Competition, 2021

**PUBLICATION**

Size Zheng, **Siyuan Chen**, Yun Liang. “COMB: Memory and Computation Coordinated Mapping of DNNs onto Complex Heterogeneous SoC”, in *the proceedings of the Design Automation Conference (DAC-60)*, July 2023.

Size Zheng\*, **Siyuan Chen\***, Pedi Song, Renze Chen, Xiuhong Li, Shengen Yan, Dahua Lin, Jingwen Leng, Yun Liang. “Chimera: An Analytical Optimizing Framework for Effective Compute-intensive Operators Fusion”, in *Proceedings of the 29th international symposium on High Performance Computer Architecture (HPCA-29)*, February 2023. (\*Equal Contribution)

**PREPRINT**

**Siyuan Chen**, Pratik Fegade, Tianqi Chen, Phillip B. Gibbons, Todd C. Mowry. "ED-Batch: Efficient Automatic Batching of Dynamic Deep Neural Networks via Finite State Machine". arXiv preprint.

**RESEARCH****Mapping Heterogeneous Neural Networks onto Heterogeneous SoC**

10/2022-11/2022

*Cooperated Research, Supervised by Prof. Yun (Eric) Liang, Depart. of EECS, Peking University*

- Observed current mapping framework does not consider the time sharing within one accelerator and treats communication between accelerators of SoC equally.
- Proposed a mapping framework to map heterogeneous neural network onto heterogeneous SoC that considers both DNN operators' time sharing, resource sharing, and SoC's heterogeneity in computation and communication.
- Designed and implemented a genetic algorithm to explore the combinational search space.
- Achieved 30% speedup and 25% reduction in energy consumption compared to state-of-the-art mappers.
- Accepted to DAC'23 as the co-first author with a PhD student.

**Efficient Automatic Batching of Dynamic Neural Networks via Learned Finite State Machines and batching-aware memory planning**

07/2022-1/2023

*Individual Research, Supervised by Prof. Phillip B. Gibbons, Todd C. Mowry and Tianqi Chen, Depart. of EECS, CMU*

- Observed current techniques to exploit batched parallelism for dynamic neural networks is suboptimal.
- Proposed a reinforcement learning based dynamic batching algorithm to minimize kernel launches and memory transfer.
- Proposed a memory allocation algorithm based on PQ-Tree to alleviate the memory transfer overhead.
- Achieved 1.15x, 1.39x, 2.45x end2end speedup for chain-based, tree-based, and lattice-based DNNs across CPU/GPU compared to SOTA frameworks.
- Submitted to ICML' 23 as the first author.

**Analytical Model on Kernel Fusion for Compute-intensive Operators on CPU and GPU**

10/2021-6/2022

*Cooperated Research, Supervised by Prof. Yun (Eric) Liang, Depart. of EECS, Peking University*

- Observed that compute-intensive operators chains (GEMM, CONV chains) become memory intensive in DNN workloads due to the gap between computing performance and memory bandwidth.
- Applied kernel fusion, loop tiling, loop reorder to improve the memory locality for compute-intensive operators chains.
- Formulate an analytical model to select best cache-level for kernel fusion, tile-sizes and loop permutation.
- Designed and implemented low complexity solver to solve for the best loop transformation configuration with minimized data movement volume under the memory capacity constraint.
- Implemented an end2end auto-scheduler and solver and achieve 1.5~2x speedup on CPU and GPU compared to vendor libraries and state-of-the-art tensor compilers.
- Accepted to HPCA' 23 as the co-first author with a PhD student.

**SELECTED COURSE PROJECT****Optimization of Gauss Seidel algorithm on GPU | HPC**

- Coded in CUDA to accelerate the 3D stencil operator.
- Customized with GPU optimization techniques like kernel fusion and avoiding warp divergence.
- Achieved 40x speedup (36s to 0.9s)

**ClaviCode | Online IDE**

- An online website served as a replacement for offline IDE like VScode.

- Led the team through the design, develop, and testing of the website.
- Assisted computer courses at PKU.

#### **C to RISC-V Compiler** | Compiler

- Design and implement a end2end compiler from subset of C to RISC-V.