YUAN MENG

(+1) 518 961 3768 ⋄ ymeng643@usc.edu
Personal Website: https://meng-yuan-usc.com/

EDUCATION

University of Southern California *August 2019 - May 2024 (expected)*
Ph.D. Candidate, Computer Engineering Advisor: Viktor Prasanna
Ming Hsieh Department of Electrical and Computer Engineering

Rensselaer Polytechnic Institute *September 2015 - May 2019*
B.S., Dual Major: Electrical Engineering & Computer and Systems Engineering Overall GPA: 3.8/4.0
Department of Electrical, Computer, and Systems Engineering

RESEARCH INTERESTS

- **System software and libraries for portable machine learning on heterogeneous data centers (CPU, GPU, and FPGA)**
- **Algorithm hardware co-optimizations for deep learning**
- **Parallel and distributed reinforcement learning**

RESEARCH EXPERIENCE

Acceleration and Design Automation of Deep Learning Inference Ongoing
Graduate Research Assistant *@FPGA/Parallel Computing Lab, USC*

- Designed a unified hardware accelerator supporting dynamic mapping of different convolution algorithms across CNN layers for low-latency inference. (FPGA '21)
- Proposed a polynomial-time PBQP optimal solution to the algorithm mapping problem on series-parallel CNN graphs, constructed with ONNX for interfacing deep learning libraries. (FPGA '21, IC3 '22)
- Proposed a compute-saving kn2row-based algorithm hardware co-design methodology that reduced zero-computation by more than 90% for fractionally-strided convolution in upsampling and generative CNNs. (HiPC '21)

High-Performance Reinforcement Learning (RL) Ongoing
Graduate Research Assistant *@FPGA/Parallel Computing Lab, USC*

- Proposed the first CPU-FPGA system design for dynamic Monte-Carlo Tree Search applications, and achieved 3× higher throughput than state-of-the-art CPU-only implementations. (FPL '22, FPGA '23)
- Proposed a novel mapping framework for Deep RL on CPU-GPU-FPGA heterogeneous platforms with on-chip Replay Management to alleviate memory bottleneck, and achieved up to 4 times higher throughput than existing CPU-GPU frameworks. (CF '22, TPDS)
- Implemented a pipelined architecture on FPGA for Table-based Q learning that improves resource efficiency by $12\times \sim 90\times$ compared to state-of-the-art hardware designs (IPDPSW '20)
- Designed and developed a systolic-array-based architecture to accelerate Proximal Policy Optimization on CPU-FPGA heterogeneous platforms. (FCCM '20, TPDS)

Mesh Adaptation using FPGAs Fall 2018
Undergraduate Research Assistant *@Scientific Computation Research Center, RPI*

- Implemented and optimized a parameterized 2-D mesh adaptation accelerator on FPGAs with High Level Synthesis, evaluated algorithm efficiency and optimization methods.
- Developed and explored parameterized OpenCL kernels implementing tensor operations, including array reductions and sliding average on cloud FPGA nodes.

PUBLICATIONS

Journal Papers:

Chi Zhang, **Yuan Meng** and Viktor Prasanna “A Framework for Mapping DRL Algorithms with Prioritized Replay Buffer onto Heterogeneous Platforms.” IEEE Transactions on Parallel and Distributed Systems (TPDS)

Yuan Meng, Sanmukh Kuppannagari, Rajgopal Kannan and Viktor Prasanna “PPOAccel: A High-Throughput Acceleration Framework for Proximal Policy Optimization.” IEEE Transactions on Parallel and Distributed Systems (TPDS)

Seleced Conferernce Papers:

Yuan Meng, Rajgopal Kannan and Viktor Prasanna “A Framework for Monte-Carlo Tree Search on CPU-FPGA Heterogeneous Platform via Dynamic Tree Management.” The 2023 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2023)

Yuan Meng, Hongjiang Men and Viktor Prasanna “Accelerator Design and Exploration for Deformable Convolution Networks.” The 36th IEEE Workshop on Signal Processing Systems (SiPS 2022)

Yuan Meng, Rajgopal Kannan and Viktor Prasanna “Accelerating Monte-Carlo Tree Search on CPU-FPGA Heterogeneous Platform.” The 22nd International Conference on Field Programmable Logic & Applications (FPL 2022)

Yuan Meng, Chi Zhang, and Viktor Prasanna, “FPGA acceleration of deep reinforcement learning using on-chip replay management.” Proceedings of the 19th ACM International Conference on Computing Frontiers (CF 2022). **(Best Paper Award)**

Yuan Meng, Sanmukh Kuppannagari, Rajgopal Kannan and Viktor Prasanna “How to Avoid Zero-Spacing in Fractionally-Strided Convolution? A Hardware-Algorithm Co-Design Methodology.” High Performance Computing, Data, and Analytics (HiPC 2021).

Yuan Meng, Sanmukh Kuppannagari, Rajgopal Kannan and Viktor Prasanna, “DYNAMAP: Dynamic Algorithm Mapping Framework for Low Latency CNN Inference.” The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2021).

Yuan Meng, Sanmukh Kuppannagari, Rajgopal Kannan and Viktor Prasanna, “How to Efficiently Train Your AI Agent? Characterizing and Evaluating Deep Reinforcement Learning on Heterogeneous Platforms.” 24th IEEE High Performance Extreme Computing Conference (HPEC 2020). **(Outstanding Student Paper Award)**

Yuan Meng, Sanmukh Kuppannagari, and Viktor Prasanna. “Accelerating Proximal Policy Optimization on CPU-FPGA Heterogeneous Platforms.” 28th IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM 2020).

Yuan Meng, Sanmukh Kuppannagari, Rachit Rajat, Ajitesh Srivastava, Rajgopal Kannan, Viktor Prasanna “QTAccel: A Generic FPGA based Design for Q-Table based Reinforcement Learning Accelerators.” 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW 2020).

AWARDS & HONOURS

- Finalist, Ming Hsieh Institute Scholarship 2022-2023
University of Southern California

- Best Paper Award 2022
ACM International Conference on Computing Frontiers

- Outstanding Student Paper Award 2020
IEEE High Performance Extreme Computing Conference

- Annenberg Fellowship 2019-2020
University of Southern California

- Dean's Honor List 2015-2019
Rensselaer Polytechnic Institute

WORK EXPERIENCE

Graduate Teaching Assistant Fall 2020 - Spring 2021
University of Southern California

- Parallel and Distributed Computing
- Accelerated Computing using FPGAs
- Parallel Programming

Electronics Engineer Co-op January 2018 - June 2018
Hasbro, Inc.

- Prototyping for Animatronics and games
- Research on embedded voice recognition and computer vision applications in toys

Undergraduate Teaching Assistant/Mentor Spring 2017 - Spring 2019
Rensselaer Polytechnic Institute

- Embedded Control
- Foundation of Computer Science

TECHNICAL STRENGTHS

Programming Languages	Python, C/C++, PostgreSQL Scripting
Parallel Programming	CUDA, OpenMP, MPI, SYCL-oneAPI
Hardware Design	VIVADO & VITIS HLS, OpenCL, Verilog
Embedded Prototyping	Arduino, stm32 microprocessors, Raspberry Pi
Other Software & Tools	PyTorch, PyOpenCL, PyBind, CAD (NX, AutoCAD)

MENTORING

- Tiffany Liu - Parallel algorithm and implementation of multi-agent Q learning (QRTS)
- Peter Wang, Tianxin Zhu (Undergraduate Student) - Multi-Core acceleration of AlphaZero using adaptive parallelism [Submitting to IA3, SC 2023]
- Samuel Wiggins (PhD Student) - Acceleration of communication in multi-agent reinforcement learning [Accepted to DATA 2023]
- Haomei Liu (Undergraduate Student) - End to End Framework for CNN Acceleration on FPGAs with Dynamic Algorithm Mapping [Accepted to IC3 2022]
- Hongjiang Men (Master Student) - Accelerator Design For Deformable Convolution Networks using high-bandwidth memory [Accepted to SiPS 2022]
- Nathaniel Peura (Master Student) - FGYM: Framework for benchmarking FPGA-accelerated Reinforcement Learning under VITIS software development flow [Accepted to FPL 2021 DEMO]
- Sarah Chow (Undergraduate Student) - HLS-accelerated Deep Neural Network Inference