Seung Won Min

Email: min16@illinois.edu | Website: www.davidmin.net

Research Interest

- Data communication optimization with complex data access patterns
- Close hardware interconnect analysis and optimization of CPU-Accelerator system

Education

University of Illinois at Urbana-Champaign — Urbana, IL
Doctor of Philosophy in Computer Engineering (Advisor: Prof. Wen-mei Hwu) — *May 2017 – May 2022*

University of Illinois at Urbana-Champaign — Urbana, IL
Master of Science in Computer Engineering (Advisor: Prof. Nam-Sung Kim) — *May 2015 – May 2017*

University of Illinois at Urbana-Champaign — Urbana, IL
Bachelor of Science in Electrical Engineering — *Aug 2009 – May 2010, Aug 2012 – May 2015*

Publications

- "Graph Neural Network Training with Data Tiering", **Seung Won Min**, Kun Wu, Mert Hidayetoğlu, Jinjun Xiong, Xiang Song, Wen-mei Hwu, *Preprint*
- **[Upstreamed to Deep Graph Library]** "Large Graph Convolutional Network Training with GPU-Oriented Data Communication Architecture", **Seung Won Min**, Kun Wu, Sitao Huang, Mert Hidayetoğlu, Jinjun Xiong, Eiman Ebrahimi, Deming Chen, Wen-mei Hwu, *Proceedings of the VLDB Endowment (VLDB)*. 14, 11 (July 2021), 2087-2100
- "Tearing Down the Memory Wall", Zaid Qureshi, Vikram Sharma Mailthody, **Seung Won Min**, I-Hsin Chung, Jinjun Xiong, Wen-mei Hwu, Deming Chen, Wen-mei Hwu, *TECHCON* 2020
- "EMOGI: Efficient Memory-access for Out-of-memory Graph-traversal In GPUs", **Seung Won Min**, Vikram Sharma Mailthody, Zaid Qureshi, Jinjun Xiong, Eiman Ebrahimi, Wen-mei Hwu, *Proceedings of the VLDB Endowment (VLDB)*, 14, 2 (October 2020), 114–127
- "Analysis and Optimization of I/O Cache Coherency Strategies for SoC-FPGA Device", **Seung Won Min**, Sitao Huang, Mohamed Aly, Jinjun Xiong, Deming Chen, Wen-mei Hwu, *International Conference on Field-Programmable Logic and Applications (FPL)*, 2019
- **[Best Paper Nominee]** "Application-Transparent Near-Memory Processing Architecture with Memory Channel Network", Mohammad Alian, **Seung Won Min**, Hadi Asgharimoghaddam, Ashutosh Dhar, Dong Kai Wang, Thomas Roewer, Adam McPadden, Oliver OHalloran, Deming Chen, Jinjun Xiong, Daehoon Kim, Wen-mei Hwu, Nam Sung Kim, *Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2018
- "Semi-Coherent DMA: An Alternative I/O Coherency Management for Embedded Systems", **Seung Won Min**, Mohammad Alian, Wen-mei Hwu, Nam Sung Kim, *IEEE Computer Architecture Letters (CAL)*, 2018

Talks

- "Unified Tensor: Enabling GPU-Centric Data Access for Efficient Large Graph GNN Training", Graph Neural Networks User Group Meeting, Sept. 2021
- "PyTorch-Direct: Introducing Deep Learning Framework with GPU-Centric Data Access for Faster Large GNN Training", NVIDIA GTC, 2021

Research Projects

Prototyping PCIe Switch on Xilinx Versal SoC-FPGA

- Built a fully functioning PCIe 3.0 switch on Xilinx Versal SoC-FPGA for PCIe packet manipulation/probing between CPU and PCIe devices (e.g., NVMe SSD, Gigabit Ethernet, and GPU).
- It is mainly used for debugging and analyzing performance bottlenecks in device-to-device interactions, which are generally impossible with the existing software-oriented debugging methods.

PyTorch-Direct: GPU-Centric Data Access for Efficient Sparse Data Access

- Identified excessive CPU-induced data communication overheads in a multi-GPU graph neural-network (GNN) training and proposed a new GPU-oriented data communication model, which releases GPUs from the CPU management. In the GPU-oriented method, GPUs directly access CPU memory over PCIe.
- Closely analyzed GPU-PCIe traffic behavior using FPGA to discover a potential performance drawback of the GPU-oriented data communication over PCIe and proposed a countermeasure to eliminate the drawback.

Analysis and Optimization of SoC-FPGA Cache Coherent Interconnect

- Extensively investigated available cache coherency options available in SoC-FPGA platform and provided detailed case study analysis to explain abnormal performance behaviors observed in SoC-FPGA designs.
- Performed in-depth analysis of how compiler/cache hierarchy/prefetcher policy can have different impacts on cache coherent data transfer performance and demonstrated how to control them.
- Provided case studies using several real-world applications such as machine learning and computer vision.

Near-Memory Accelerator Design on IBM POWER8 Systems

- Worked on various aspects of hardware prototyping, including RTL designing and Linux driver coding to enable the experimental near-memory accelerator – designing a fully functioning near-memory accelerator required a comprehensive understanding of both hardware and software.
- Proposed and implemented an SoC-like near memory accelerator design with embedded CPU, scratchpad memory, and DDR3 controller, so while the acceleration unit provides raw computation power, the embedded CPU can handle control details.

Work Experience

Research Assistant 01/01/2016 – Current

University of Illinoist at Urbana-Champaign *Urbana, IL*

- Building an accelerator-oriented communication architecture for better scalability and lower CPU overhead.
- Devising a seamless hardware/software support of the new data communication architecture to hide overwhelming details from the end-users.
- Demonstrating several proof-of-concept designs in FPGA for realistic performance measurements.

Research Internship - Fine-Grained CNN Accelerator Design 05/28/2019 – 08/15/2019

IBM Research *Yorktown Heights, NY*

- Developed a communication library enabling fine-grained CNN accelerations on FPGA.
- Proposed a design to utilize IBM's coherent accelerator processor interface (CAPI) to hide address translation details from FPGA and minimize the software overhead in data communication.

Research Internship - Near-Memory Accelerator Design 06/05/2017 – 08/25/2017

IBM Research *Yorktown Heights, NY*

- Designed, debugged, and evaluated experimental near-memory accelerator on IBM POWER8 server with Ubuntu Linux running.
- Designed and implemented a shared accelerator memory space between the host CPU and the accelerator for low latency data communication and a DMA-less design.

Technical Skills

Languages: Python, C/C++, CUDA, VHDL, Verilog
Developer Tools: PyTorch, TensorFlow, Xilinx Vivado

Other Experience

IBM Bluehack/Hackathon North America 2017 Nominee of TJ Bot Section Summer 2017

UIUC ECE 342 Electronic Circuits Head TA Fall 2015

- Responsible for coordinating class logistics and supervising five course TAs and three graders.

HKN Dr. Everitt's Neighborhood Editor Spring 2014

- Wrote, reviewed, edited, and updated course introduction articles for ECE students.