Seung Won Min

Website: www.davidmin.net | Google Scholar | LinkedIn

Summary

- GPU architect and systems researcher specializing in GPU performance optimization and data communication in heterogeneous CPU-Accelerator systems.
- Hands-on experience across: CUDA kernel development, PyTorch internals modification, and performance profiling.

Work Experience

GPU Architect 2022 – Current
NVIDIA Corporation *Santa Clara, CA*

Research Assistant 2016 – 2022
University of Illinois at Urbana-Champaign *Urbana, IL*

- Designed GPU-centric data communication architectures for scalable multi-GPU training with reduced CPU overhead.
- Developed seamless hardware/software frameworks to enable efficient accelerator-oriented data access while hiding complexity from end-users.
- Built FPGA-based hardware prototypes for realistic performance measurement and low-level interconnect analysis.

Research Internship – Fine-Grained CNN Accelerator Design 05/2019 – 08/2019
IBM Research *Yorktown Heights, NY*

- Developed a communication library enabling fine-grained convolutional neural network (CNN) layer offloading to accelerator.

Research Internship – Near-Memory Accelerator Design 06/2017 – 08/2017
IBM Research *Yorktown Heights, NY*

- Designed experimental near-memory accelerator on IBM POWER8 server with Ubuntu Linux.
- Implemented a proprietary shared memory architecture enabling low-latency, DMA-less data communication between host CPU and accelerator.

Technical Skills

Languages: Python, C/C++, CUDA, Verilog
ML Frameworks: PyTorch (internal modifications), Deep Graph Library (upstream contributor)
Hardware & Tools: GPU architecture, NVIDIA Nsight, Intel VTune, FPGA acceleration, Linux kernel/driver development

Research Projects

PyTorch-Direct: GPU-Centric Data Access Architecture for Irregular Data Accesses

- Identified CPU-induced data communication bottlenecks in multi-GPU GNN training and designed a GPU-centric communication architecture where GPUs directly access host memory over PCIe, eliminating CPU orchestration overhead.
- Modified PyTorch framework internals to natively support GPU-initiated memory access; work adopted by AWS and upstreamed to Deep Graph Library, providing 2x performance in 100M node graph training.

GNN Training with Multi-GPU Data Tiering

- Developed a memory tiering strategy for multi-GPU GNN training that uses weighted reverse PageRank to statistically predict frequently accessed nodes and places hot data across a collective NVLink-connected GPU memory pool, reducing CPU-GPU PCIe traffic by 87–95%.
- Designed an interleaved multi-GPU tensor distribution scheme over NVLink that balances memory and interconnect bandwidth consumption across GPUs, enabling training on 350GB datasets with eight NVIDIA H100 GPUs.

EMOGI: Efficient Out-of-Memory Graph Traversal on GPUs

- Developed optimized GPU kernels for out-of-memory graph traversal with aggressive memory access alignment and request merging to maximize PCIe bandwidth utilization, achieving 1.4x–4.7x speedups across diverse graph workloads.
- Demonstrated a GPU-centric approach that eliminates CPU overhead in existing methods, enabling efficient processing of graph datasets exceeding GPU memory capacity.

Near-Memory Accelerator Design on IBM POWER8 Systems

- Designed and prototyped a near-memory accelerator spanning Linux kernel driver development and system-level integration on IBM POWER8 server.
- Implemented an SoC-like architecture with embedded CPU, scratchpad memory, and DDR3 controller for efficient near-memory computation with minimal data movement overhead.

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

- “GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture", Zaid Qureshi, Vikram Sharma Mailthody, Isaac Gelado, **Seung Won Min**, Amna Masood, Jeongmin Park, Jinjun Xiong, CJ Newburn, Dimitri Vainbrand, Wen-mei Hwu, *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2023
- “FSSD: FPGA-Based Emulator for SSDs", Luyi Yu, Yuebin Lu, Mohit Mandava, Eric Richter, Vikram Sharma Mailthody, **Seung Won Min**, Wen-mei Hwu, Nam Sung Kim, *33rd International Conference on Field-Programmable Logic and Applications (FPL)*, 2023
- “Graph Neural Network Training with Data Tiering", **Seung Won Min**, Kun Wu, Mert Hidayetoğlu, Jinjun Xiong, Xiang Song, Wen-mei Hwu, In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (*KDD*), 2022
- **[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