Xiaowei Ren

Senior Deep Learning Architect
NVIDIA Corporation
2578B West Mall, Vancouver
BC Canada V6T 2J9

Cell: (+1)778-893-7616
Homepage: https://ericrxw.github.io/xiaoweiren/
Email: xren@nvidia.com
renxiaowei66@gmail.com

Education

Sept. 2015 – Oct. 2020	PhD in Computer Engineering	University of British Columbia, Canada
	Dissertation: Efficient Synchronization Mechanisms for Scalable GPU Architectures	
Sept. 2012 – Jun. 2015	M.Sc. in Computer Engineering	Xi'an Jiaotong University, China
	Thesis: Parallel Acceleration Algorithms and FPGA Implementation for KLMS and KAP	
Sept. 2008 – Jun. 2012	B.Sc. in Electronic Engineering	Xi'an Jiaotong University, China

Professional Experience

Jun. 2021 – Present	Senior Architect	NVIDIA, Canada
Nov. 2020 – Jun. 2021	Postdoc Research Fellow	University of British Columbia, Canada
Sept. 2015 – Oct. 2020	Research Assistant	University of British Columbia, Canada
Sept. 2019 – Nov. 2019	Research Intern	Max Planck Institute for Software Systems, Germany
Aug. 2018 – Nov. 2018	Research Intern	NVIDIA, USA
May. 2017 – Aug. 2017	Research Intern	NVIDIA, USA
Sept. 2012 – Jun. 2015	Research Assistant	Xi'an Jiaotong University, China
Jul. 2011 – Sept. 2011	Undergraduate Intern	ICT, Chinese Academy of Science, China

Publications

- **Xiaowei Ren**, and Mieszko Lis. "CHOPIN: Scalable Graphics Rendering in Multi-GPU Systems via Parallel Image Composition", *27th International Symposium on High Performance Computer Architecture (HPCA)*, Seoul, South Korea, February 2021. (acceptance rate: 63/258 = 24.4%)

- Dingqing Yang, Amin Ghasemazar*, **Xiaowei Ren***, Maximilian Golub, Guy Lemieux, and Mieszko Lis. "Procrustes: a Dataflow and Accelerator for Sparse Deep Neural Network Training", *53rd International Symposium on Microarchitecture (MICRO)*, Athens, Greece, October 2020. (acceptance rate: 82/424 = 19.3%, *equal contribution)

- **Xiaowei Ren**, Daniel Lustig, Evgeny Bolotin, Aamer Jaleel, Oreste Villa, and David Nellans. "HMG: Extending Cache Coherence Protocols Across Modern Hierarchical Multi-GPU Systems", *26th International Symposium on High Performance Computer Architecture (HPCA)*, San Diego, USA, February 2020. (acceptance rate: 48/248 = 19.4%)

- **Xiaowei Ren**, and Mieszko Lis. "High-Performance GPU Transactional Memory via Eager Conflict Detection", *24th International Symposium on High Performance Computer Architecture (HPCA)*, Vienna, Austria, February 2018. (acceptance rate: 54/260 = 20.8%)

- **Xiaowei Ren**, and Mieszko Lis. "Efficient Sequential Consistency in GPUs via Relativistic Cache Coherence", *23rd International Symposium on High Performance Computer Architecture (HPCA)*, Austin, USA, February 2017. (acceptance rate: 50/224 = 22.3%)

- Pengju Ren, **Xiaowei Ren**, Sudhanshu Sane, Michel A. Kinsy, and Nanning Zheng. "A Deadlock-Free and Connectivity-Guaranteed Methodology for Achieving Fault-tolerance in On-Chip Networks", *IEEE Transactions on Computers (TC)*, 2016.

- **Xiaowei Ren**, Qihang Yu, Badong Chen, Nanning Zheng, and Pengju Ren. "A Reconfigurable Parallel Accelerator for the Kernel Affine Projection Algorithm", *IEEE International Conference on Digital Signal Processing (DSP)*, Singapore, July 2015.

- **Xiaowei Ren**, Qihang Yu, Badong Chen, Nanning Zheng, and Pengju Ren. "A 128-way FPGA Platform for the Acceleration of KLMS Algorithm", *Asia and South Pacific Design Automation Conference (ASP-DAC)*, Tokyo, Japan, January 2015. (University LSI Design Contest)

- **Xiaowei Ren**, Qihang Yu, Badong Chen, Nanning Zheng, and Pengju Ren. "A Real-time Permutation Entropy Computation for EEG Signals", *Asia and South Pacific Design Automation Conference (ASP-DAC)*, Tokyo, Japan, January 2015. (University LSI Design Contest)

- **Xiaowei Ren**, Pengju Ren, Badong Chen, Jose C. Principe, and Nanning Zheng. "A Reconfigurable Parallel Acceleration Platform for Evaluation of Permutation Entropy", *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Chicago, USA, August 2014.

- **Xiaowei Ren**, Pengju Ren, Badong Chen, Tai Min, and Nanning Zheng. "Hardware implementation of KLMS Algorithm using FPGA", *International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, July 2014.

- Pengju Ren, Qingxin Meng, **Xiaowei Ren**, and Nanning Zheng. "Fault-tolerant Routing for On-chip Network without Using Virtual Channel", *ACM/EDAC/IEEE Design Automation Conference (DAC)*, San Francisco, USA, June 2014. (acceptance rate: 3150/10963 = 29%)

Awards

2016 – 2020	UBC Graduate Support Initiative (GSI) Awards
2012 – 2015	National Master Scholarship (honors top 5% students)
2013 – 2014	Suzhou Industrial Park Scholarship
2010 – 2011	CASC Secondary Class Scholarship
2009 – 2010	National Encouragement Scholarship (honors top 3% students)
2008 – 2009	Siyuan Scholarship

Research Projects

- **Architectural Support for Efficient GPU Synchronizations**

  - **Extending Cache Coherence Across Hierarchical Multi-GPUs** **NVIDIA Research**
    Identified necessity of coherence hierarchy for performance scaling in multi-GPU systems. By leveraging the latest scoped GPU memory model, proposed, implemented, and evaluated HMG, a hier-

archical hardware cache coherence protocol for multi-GPUs. HMG not only avoids full cache invalidation in software coherence protocol, but also filters out write invalidation acknowledgments and transient coherence states. Despite minimal hardware overheads, HMG can achieve 97% of the performance of an idealized caching system. Published in the HPCA-2020.

  - **GPU Hardware Transactional Memory (TM)** **University of British Columbia**
    Identified excessive conflict detection latency as a key limitation of performance and available parallelism in prior GPU TM proposals. Designed, implemented, and evaluated GETM, the first GPU hardware TM with eager conflict detection. GETM relies on a novel logical-timestamp-based conflict detection mechanism: conflicts are detected eagerly when the initial memory access is made. Performance is up to 2.1× better than the best prior work WarpTM (1.2× gmean). Area overheads are 3.6× lower and power overheads are 2.2× lower. Published in the HPCA-2018.
  - **Efficient Sequential Consistency (SC) in GPUs** **University of British Columbia**
    Identified acquisition of write permissions as the main source of inefficiency in coherence protocol design for SC in GPUs. Developed, implemented, and evaluated RCC, an invalidate-free coherence protocol which grants write permissions without stalling but can still enforce SC. RCC is 29% faster than the best prior SC proposal for GPUs, and within 7% of the best non-SC design. Additionally, RCC allows for switching strong and weak consistency models at runtime with best-in-class performance and no hardware overhead. Published in the HPCA-2017.
  - **Light-weight GPU Cache Coherence Protocol** **NVIDIA Research**
    Implemented, and evaluated a light-weight cache coherence protocol for NVIDIA GPUs to avoid the expensive cost of entire cache invalidation in software coherence protocol.

- **Scalable Graphics Rendering in Multi-GPU Systems** **University of British Columbia**
  Identified redundant computation and sequential inter-GPU synchronization as critical performance bottlenecks of existing SFR (Split Frame Rendering) implementations. To eliminate the overheads, designed, implemented, and evaluated CHOPIN, a sort-last rendering scheme by taking advantage of mathematical properties of image composition to compose pixels in parallel. Developed a draw command scheduler and an image composition scheduler to address the problems of load imbalance and network congestion. Compared to the best prior SFR implementation, CHOPIN can offer speedups of up to 1.56× (1.25× gmean). Published in the HPCA-2021.

- **Domain-Specific Accelerators**

  - **Sparse Training Accelerator** **University of British Columbia**
    With tow other students, identified problems of designing a hardware accelerator for training sparse neural network directly, including load imbalance, dataflow etc. Proposed, designed, and evaluated a hardware-friendly sparse training algorithm, an accelerator architecture to enable sparse training, and a novel mechanism for load-balancing without complicating the on-chip network. Our accelerator can substantially reduce the time and energy required to train deep neural networks without losing accuracy, several time more efficient than prior proposals. Published in the MICRO-2020.
  - **FPGA Accelerators for Computation-intensive Algorithms** **Xi'an Jiaotong University**
    Designed, and implemented several FPGA accelerators for kernel-based machine learning and signal processing algorithms. Published in three conference papers and two student design contest papers.

- **Fault-tolerant Network-on-Chip Routing Algorithm** **Xi'an Jiaotong University**
  Proposed, implemented, and evaluated a deadlock-free, fault-tolerant routing algorithm for on-chip net-

work that guarantees maximal connectivity without using virtual channels. With 40% link damage, the algorithm guarantees 98% reliability with only 1% hardware overhead. Published in the Design Automation Conference (DAC) 2014 and the IEEE Transactions on Computers.

- **Memory Model Checking Algorithm** **Max Planck Institute for Software Systems**
  Proposed, and implemented efficient handling for spin-loops and barriers in stateless model checking.
- **Cache Performance Modelling** **NVIDIA Research**
  Contributed to the development of a new NVIDIA architecture simulator. Accurately modelled GPU cache performance and correlated to a real GPU system.

Talks & Presentations

- Oral, "CHOPIN: Scalable Graphics Rendering in Multi-GPU Systems via Parallel Image Composition", *HPCA*, Global Virtual Event, February 2021.
- Oral, "HMG: Extending Cache Coherence Protocols Across Modern Hierarchical Multi-GPU Systems", *HPCA*, San Diego, USA, February 2020.
- Oral, "High-Performance GPU Transactional Memory via Eager Conflict Detection", *HPCA*, Vienna, Austria, February 2018.
- Oral, "Efficient Sequential Consistency in GPUs via Relativistic Cache Coherence", *HPCA*, Austin, USA, February 2017.
- Oral and Poster, "A 128-way FPGA Platform for the Acceleration of KLMS Algorithm", *ASP-DAC*, Tokyo, Japan, January 2015.
- Oral and Poster, "A Real-time Permutation Entropy Computation for EEG Signals", *ASP-DAC*, Tokyo, Japan, January 2015.
- Poster, "A Reconfigurable Parallel Acceleration Platform for Evaluation of Permutation Entropy", *EMBC*, Chicago, USA, August 2014.
- Poster, "Hardware implementation of KLMS Algorithm using FPGA", *IJCNN*, Beijing, China, July 2014.

Professional Service

- External Reviewer Committee, MICRO, 2021
- External Reviewer Committee, ISCA, 2021
- Artifact Evaluation Committee, ASPLOS, 2021
- Reviewer, ACM TECS, 2020
- Reviewer, ACM TACO, 2020
- Reviewer, IEEE TPDS, 2021
- Reviewer, IEEE TC, 2020

- Reviewer, IEEE TCAD, 2020, 2021
- Reviewer, IEEE CAL, 2020, 2021
- Reviewer, JPDC, 2020
- Subreviewer, HPCA, 2021
- Shadow Program Committee, EuroSys, 2021

Teaching Experience

Jan. 2017 – Apr. 2017	Teaching Assistant, University of British Columbia, Canada EECE527: Advanced Computer Architecture (Instructor: Mieszko Lis)
Sept. 2016 – Dec. 2016	Teaching Assistant, University of British Columbia, Canada CPEN411: Computer Architecture (Instructor: Mieszko Lis)
Sept. 2015 – Dec. 2015	Teaching Assistant, University of British Columbia, Canada CPEN211: Introduction to Microcomputers (Instructor: Tor Aamodt)