Xiaowei Ren

PhD at the University of British Columbia
471, 6335 Thunderbird Residence, UBC
Vancouver, BC Canada, V6T 2G9

Cell: (+1)778-893-7616
Email: renxiaowei66@gmail.com / xiaowei@ece.ubc.ca
URL: http://www.linkedin.com/in/xiaowei-ren-95a7318a

Education

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

Professional Experience

Sept. 2015 – Present	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 Architecture Research Group, USA
May. 2017 – Aug. 2017	Research Intern	NVIDIA Architecture Research Group, 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

Journal Articles (refereed)

- Pengju Ren, **Xiaowei Ren**, Sudhanshu Sane, Michel A. Kinsy and Nanning Zheng, "A Deadlock-Free and Connectivity-Guaranteed Methodology for Achieving Fault-tolerance in Direct Networks", *IEEE Transactions on Computers (TC)*, 2016. (acceptance rate: 30%)

Conference Articles (refereed)

- **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%)

- **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

2019 – 2020	UBC Graduate Student Initiative (GSI) Awards (CAD $6000)
2018 – 2019	UBC Graduate Student Initiative (GSI) Awards (CAD $8000)
2017 – 2018	UBC Graduate Student Initiative (GSI) Awards (CAD $6000)
2016 – 2017	UBC Graduate Student Initiative (GSI) Awards (CAD $4000)
2012 – 2015	National Master Scholarship (honors top 5% students)
2013 – 2014	Suzhou Industrial Park Scholarship
2010 – 2011	CASC Secondary Class Scholarship
2009 – 2010	National Motivational Scholarship (honors top 5% students)
2008 – 2009	Siyuan Scholarship

Research Projects

- **Architectural Support for Efficient GPU Synchronizations**
  - **Extending Cache Coherence Across Hierarchical Multi-GPUs** **NVIDIA Research**
    Identified the necessity of coherence hierarchy for performance scaling in multi-GPU systems. By leveraging the latest scoped GPU memory model, we proposed, implemented, and evaluated HMG, a hierarchical hardware cache coherence protocol for multi-GPUs. HMG not only avoided the full cache invalidation in software coherence protocol, but also filtered out the write invalidation messages and transient coherence states. With a small hardware cost, HMG can achieve 97% performance of an idealized caching system. Published in the HPCA-2020.

- **GPU Hardware Transactional Memory (TM)** **University of British Columbia**
  Identified excessive commit-time latency as a key limitation of performance and available parallelism in prior GPU TM proposals. Designed, implemented, and evaluated GETM, the first eager GPU hardware TM. GETM relies on a novel logical-timestamp-based eager conflict detection mechanism: conflicts are detected eagerly when the initial memory access is made. Performance is up to 2.1x better than the state-of-the art prior work WarpTM (1.2x on average). Area overheads are 3.6x lower and power overheads are 2.2x lower. Published in the HPCA-2018.
- **Efficient Sequential Consistency (SC) in GPUs** **University of British Columbia**
  Identified acquisition of coherence permissions as the main source of inefficiency in SC implementations proposed for GPUs. Developed, implemented, and evaluated RCC, a sequentially consistent memory system for GPUs based on an invalidate-free coherence protocol. RCC is 30% faster than the best prior SC proposal for GPUs, and within 7% of the best GPU solution with relaxed consistency. Additionally, the protocol allows for switching strong and weak consistency models at runtime with best-in-class performance and no hardware overhead. Published in the HPCA-2017.
- **Scalable Graphics Rendering in Multi-GPUs** **University of British Columbia**
  Conquer inter-GPU communication bottleneck by leveraging the potential parallelism of sort-last rendering. Propose a dynamic workload dispatcher for load-balancing across multiple GPUs, and a sub-image composition scheduler to mitigate network congestion. (Undergoing Project)
- **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.

- **Domain-Specific Accelerators**

  - **Sparse Training Accelerator** **University of British Columbia**
    Identified problems of designing a hardware accelerator for training sparse neural network directly, including load imbalance, dataflow etc. Propose, design, and evaluate a hardware-friendly sparse training algorithm, an Eyeriss-like accelerator architecture to enable sparse training, and a novel dataflow for load-balancing without complicating the on-chip network. (Undergoing Project)
  - **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 networks that guarantees maximal connectivity without using virtual channels. With 40% link damage, the algorithm guarantees 98% reliability with only 1% hardware overhead. Published in 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 Modeling** **NVIDIA Research**
  Contributed to the development of a new NVIDIA architecture simulator. Accurately modeled GPU cache performance and correlated to a real GPU system.

Talks & Presentations

- 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.

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)