Zhen Peng

Postdoctoral Research Associate
High-Performance Computing Group
Pacific Northwest National Laboratory (PNNL)

Email: hi.pengzhen@gmail.com
Phone: +1 (510) 931-8704
https://johnpzh.github.io/

Research Interests

High-performance computing, sparse computing, compilers

EDUCATION

Ph.D. in Computer Science 08/2016 – 01/2023
Department of Computer Science, College of William & Mary, Williamsburg, VA
Advisor: Dr. Bin Ren

M.S. in Computer Software and Theory 09/2013 – 06/2016
Department of Computer Science, Huaqiao University, Xiamen, China
Advisor: Dr. Tian Wang

B.E. in Computer Science and Technology 09/2009 – 06/2013
Department of Computer Science, Huaqiao University, Xiamen, China

RESEARCH EXPERIENCE

Post Doctorate RA 04/2023 – Present
Pacific Northwest National Laboratory (PNNL), Richland, WA

Automatic Code Generation for Graph Algorithms in Linear Algebra Expressions

- Design and extend intermediate representation to support parallel graph kernel in compiler COMET.
- Optimize code transformation for sparse computation, such as SpGEMM.

Research Assistant 08/2017 – 06/2022
Department of Computer Science, College of William & Mary, VA

Accelerate Deep Neural Network Inference on Edge Devices

- Analyze inference procedure of TensorFlow Lite for Micro on microcontroller units (MCU).
- Speed up the inference procedure by tuned loop unrolling and customized quantization method.

Efficient Parallelization of Graph-based Approximate Nearest Neighbors Search (ANNS)

- Analyze and parallelize the best-first search algorithm for ANNS on the graph-based index.
- Reduce the intra-query latency on CPUs by a tailored parallelism scheme and synchronization mechanism.

Parallelizing Pruned Landmark Labeling: Dealing with Dependencies in Graph Algorithms

- Analyze and parallelize the sequential 2-hop labeling for shortest distance queries in large graphs.
- Reduce the query latency on CPUs using the parallel algorithm that breaks the dependency.

Efficient Parallelization of Graph Processing on Emerging Many-core Architectures

- Design the graph processing system for typical graph algorithms such as BFS to tap into many-core CPUs.
- Achieve good performance and scalability by be aware of data locality, load balance, and update conflicts.

INTERNSHIP EXPERIENCE

PhD Intern 06/2022 – 04/2023
Pacific Northwest National Laboratory (PNNL), Richland, WA

Redundancy-Aware Code Generation for Sparse Tensor Expressions

- Detect the redundancy in code generation for sparse tensor expressions.
- Implement partial fusion algorithm in the MLIR-based compiler COMET.

ML Research Intern 04/2021 – 09/2021
Kuaishou, US R&D Center, Palo Alto, CA

Automate the Model Implementation to TensorRT

- Translate models from TVM Relay IR to TensorRT Python code.
- Speed up the model deployment procedure and reduce the labor costs.

Accelerate Inference Through Operation Fusion in Convolutional Neural Network (CNN)

- Try to add customed operation fusion pass in TVM.
- Transform fused computational graph to TensorRT C++ code to accelerate the inference.

PATENT

[1] *"Multi-Level Intermediate Representation Decoder for Heterogeneous Platforms*," U.S. Patent Application No. 17/524,619, Filing date: November 11, 2021

PUBLICATIONS

[1] **Zhen Peng**, Rizwan A. Ashraf, Luanzheng Guo, Ruiqin Tian, and Gokcen Kestor, "Automatic Code Generation for High-Performance Graph Algorithms," *The 32nd International Conference on Parallel Architectures and Compilation Techniques (PACT 2023)*, October 21-25, 2023, Vienna, Austria.

[2] **Zhen Peng**, Minjia Zhang, Kai Li, Ruoming Jin, and Bin Ren, "iQAN: Fast and Accurate Vector Search with Efficient Intra-Query Parallelism on Multi-Core Architectures," *The 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP 2023)*, February 25-March 1, 2023, Montreal, Canada.

[3] **Zhen Peng**, Minjia Zhang, Kai Li, Ruoming Jin, and Bin Ren, "Speed-ANN: Low-Latency and High-Accuracy Nearest Neighbor Search via Intra-Query Parallelism," *arXiv:2201.13007*, 2022.

[4] Qihan Wang, **Zhen Peng**, Bin Ren, Jie Chen, and Robert G. Edwards, "MemHC: An Optimized GPU Memory Management Framework for Accelerating Many-body Correlation," *ACM Transactions on Architecture and Code Optimization (TACO)*, Volume 19, Issue 2, No. 24, pp 1-26, June 2022.

[5] Ruoming Jin*, **Zhen Peng***, Wendell Wu, Feodor Dragan, Gagan Agrawal, and Bin Ren, "Parallelizing Pruned Landmark Labeling: Dealing with Dependencies in Graph Algorithms," *The 34th ACM International Conference on Supercomputing (ICS 2020)*, June 29-July 2, 2020, Online. (* Equal contribution)

[6] Yu Chen, Ivy Peng, **Zhen Peng**, Xu Liu, and Bin Ren, "ATMem: Adaptive Data Placement in Graph Applications on Heterogeneous Memories," *International Symposium on Code Generation and Optimization (CGO 2020)*, February 22-26, 2020, San Diego, CA, USA.

[7] Ruoming Jin, **Zhen Peng**, Wendell Wu, Feodor Dragan, Gagan Agrawal, and Bin Ren, "Pruned Landmark Labeling Meets Vertex Centric Computation: A Surprisingly Happy Marriage!" arXiv:1906.12018, 2019.

[8] **Zhen Peng**, Alexander Powell, Bo Wu, Tekin Bicer, and Bin Ren, "GraphPhi: Efficient Parallel Graph Processing on Emerging Throughput-oriented Architectures," *International conference on Parallel Architectures and Compilation Techniques (PACT 2018)*, November 1-4, 2018, Limassol, Cyprus.

PUBLICATIONS BEFORE PH.D.

[1] Tian Wang, **Zhen Peng**, Sheng Wen, Weijia Jia, Yiqiao Cai, Hui Tian, and Yonghong Chen. "Reliable Wireless Connections for Fast-Moving Rail Users Based on a Chained Fog Structure." *Information Sciences (Inf. Sci.)*, 379: 160-176, 2017.

[2] Tian Wang, **Zhen Peng**, Chen Wang, Yiqiao Cai, Yonghong Chen, Hui Tian, Junbin Liang, and Bineng Zhong. "Extracting Target Detection Knowledge Based on Spatio-temporal Information in Wireless Sensor Networks." *International Journal of Distributed Sensor Networks (IJDSN)*, 2016 (doi:10.1155/2016/5831471), 2016.

[3] Tian Wang, **Zhen Peng**, Junbin Liang, Sheng Wen, Md Zakirul Alam Bhuiyan, Yiqiao Cai, and Jiannong Cao. "Following Targets for Mobile Tracking in Wireless Sensor Networks." *ACM Transactions on Sensor Networks (TOSN)*, 12(4): 31:1-31:24, 2016.

[4] **Zhen Peng**, Tian Wang, Md Zakirul Alam Bhuiyan, Xiaoqiang Wu, and Guojun Wang. "Dependable Cascading Target Tracking in Heterogeneous Mobile Camera Sensor Networks." *Springer International Publishing, Algorithms and Architectures for Parallel Processing (ICA3PP Workshops and Symposium)*, 2015: 531-540.

[5] Tian Wang, **Zhen Peng**, Junbin Liang, Yiqiao Cai, Yonghong Chen, Hui Tian, and Bineng Zhong. "Detecting Targets Based on a Realistic Detection and Decision Model in Wireless Sensor Networks." *Springer International Publishing, Wireless Algorithms, Systems, and Applications (WASA)*, 2015: 836-844.

[6] Tian Wang, **Zhen Peng**, Yonghong Chen, Yiqiao Cai, and Hui Tian. "Continuous tracking for mobile targets with mobility nodes in WSNs." *International Conference on Smart Computing (SMARTCOMP)*, Hong Kong, pp. 261-268, November 3-5, 2014.

TECHNICAL SKILLS

Programming Languages: C++, C, Python, Bash

Frameworks and Tools: AVX-512, OpenMP, MPI, MLIR, TensorFlow Lite for Micro, TVM, TensorRT

AWARDS & HONORS

Student Travel Grands, PACT '18 2018

Student Travel Awards, ASPLOS '18 2018

TEACHING EXPERIENCE

College of William & Mary

Grading TA: CSCI 304 Computer Organization 02/2018 – 05/2018

Grading TA: CSCI 304 Computer Organization 09/2017 – 01/2018

Grading TA: CSCI 243 Discrete Structures 02/2017 – 05/2017

Grading TA: CSCI 141 Computational Problem Solving 09/2016 – 01/2017

PROFESSIONAL SERVICES

Conference Reviewer:

IPDPS-2025, ICPP-2024, HiPC-2024, ICAT-2023, HiPC-2023, SC-2023, IPDPS-2022, PPoPP-2021, ICS-2021, IPDPS-2021, ICPP-2020, NPC-2020, Bench-2020, HiPC-2019, Bench-2019, NPC-2019, BIGCOM-2019, ICCCN-2019, NPC-2018, HiPC-2018, UIC-2018, SCS-2017

Journal Reviewer:

TACO