Zhen Peng

CA, USA | hi.pengzhen@gmail.com | (510) 931-8704 | johnpzh.github.io | zhen-peng-48930a63 | johnpzh

Research Interests

Zhen is a postdoc in the Future Computing Technologies Group (formerly known as the High-Performance Computing Group) at Pacific Northwest National Laboratory (PNNL). His research interests include high-performance computing, compilers, heterogeneous architectures, and continuum computing.

Education

Ph.D. **College of William & Mary**, Computer Science — Williamsburg, VA, USA
- Advisor: Dr. Bin Ren — Aug 2016 – Jan 2023
- Dissertation Title: Exploring Multi-Level Parallelism for Graph-Based Applications via Algorithm and System Co-Design

M.S. **Huaqiao University**, Computer Software and Theory — Xiamen, Fujian, China
- Advisor: Dr. Tian Wang — Sept 2013 – July 2016
- Dissertation Title: Research on Target Tracking in Wireless Sensor Networks with Mobility Elements

B.E. **Huaqiao University**, Computer Science and Technology — Xiamen, Fujian, China
- Dissertation advisor: Dr. Weibo Xie — Sept 2009 – June 2013
- Dissertation Title: Popular Hot News Website Subscription

Experience

Pacific Northwest National Laboratory (PNNL), Post Doctorate Research Associate — Richland, WA, USA (remote)
compilers, disaggregated memory, I/O orchestration, generative AI — Apr 2023 – present — 3 years 2 months
- Optimize distributed scientific workflows by prioritizing critical data flow
- Extend an MLIR-based compiler to support Fabric Attached Memory (FAM) through CXL
- Develop MLIR-based compiler for sparse computation on heterogeneous hardware
- Use generative AI for efficient prediction of protein redox potentials

Pacific Northwest National Laboratory (PNNL), PhD Intern — Richland, WA, USA
compilers, code generation — June 2022 – Apr 2023 — 11 months
- Extend MLIR-based compiler for redundancy-aware code optimization

Department of CS, College of William & Mary (W&M), Research Assistant — Williamsburg, VA, USA
graph algorithms, parallelization, multi-threading, vectorization/SIMD, distributed — Aug 2017 – June 2022 — 4 years 11 months
- Optimize deep neural network inference on edge devices
- Parallelize and optimize graph-based Approximate Nearest Neighbors Search (ANNS)
- Parallelize Pruned Landmark Labeling algorithm for the shortest path problem
- Optimize parallel graph processing on emerging many-core architectures

Kuaishou, US R&D Center, Machine Learning Research Intern — Palo Alto, CA, USA
machine learning, operator fusion, inference acceleration — Apr 2021 – Sept 2021 — 6 months
- Automate the model implementation to TensorRT
- Accelerate inference through operator fusion in convolutional neural networks (CNNs)

Skills

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

Frameworks: MLIR, LLVM, OpenMP, AVX-512, MPI, Nextflow, TVM, TensorFlow Lite, TensorRT

Publications

CCom: A Compiler Framework for Linear Algebra on CXL-Attached Memory 2026
Zhen Peng, Roberto Gioiosa, Gokcen Kestor, Andres Marquez
(HCDS 2026, the 5th Workshop on Heterogeneous Composable and Disaggregated Systems)

FastFlow: Rapid Workflow Response By Prioritizing Critical Data Flows and their Interactions 2025
Jesun Sahariar Firoz, Hyungro Lee, Luanzheng Guo, Meng Tang, Nathan R. Tallent, **Zhen Peng**
10.1145/3733723.3733735 (SSDBM 2025, the 37th International Conference on Scalable Scientific Data Management)

LiteForm: Lightweight and Automatic Format Composition for Sparse Matrix-Matrix Multiplication on GPUs 2025
Zhen Peng, Polykarpos Thomadakis, Jacques Pienaar, Gokcen Kestor
10.1145/3731545.3731574 (HPDC 2025, the 34th ACM International Symposium on High-Performance Parallel and Distributed Computing)

Towards Recognizing Food Types for Unseen Subjects 2025
Jiexiong Guan, Junjie Wang, Wei Niu, **Zhen Peng**, Shuangquan Wang, Zhenming Liu, Gang Zhou, Bin Ren
10.1145/3696424 (ACM Transactions on Computing for Healthcare, Volume 6, Issue 1, No. 1, pp. 1-21, January 2025)

Automatic Code Generation for High-Performance Graph Algorithms 2023
Zhen Peng, Rizwan A. Ashraf, Luanzheng Guo, Ruiqin Tian, Gokcen Kestor
10.1109/PACT58117.2023.00010 (PACT 2023, the 32nd International Conference on Parallel Architectures and Compilation Techniques)

iQAN: Fast and Accurate Vector Search with Efficient Intra-Query Parallelism on Multi-Core Architectures 2023
Zhen Peng, Minjia Zhang, Kai Li, Ruoming Jin, Bin Ren
10.1145/3572848.3577527 (PPoPP 2023, the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming)

Speed-ANN: Low-Latency and High-Accuracy Nearest Neighbor Search via Intra-Query Parallelism 2022
Zhen Peng, Minjia Zhang, Kai Li, Ruoming Jin, Bin Ren
10.48550/arXiv.2201.13007 (arXiv 2022)

MemHC: An Optimized GPU Memory Management Framework for Accelerating Many-body Correlation 2022
Qihan Wang, **Zhen Peng**, Bin Ren, Jie Chen, Robert G. Edwards
10.1145/3506705 (ACM Transactions on Architecture and Code Optimization, Volume 19, Issue 2, No. 24, pp 1-26, June 2022)

Multi-Level Intermediate Representation Decoder For Heterogeneous Platforms 2024
Zhen Peng, Yang Liu, Hanxian Huang, Yongxiong Ren, Jishen Yang, Lingzhi Liu, Xin Chen
patents.google.com/patent/US11928446B2/en (U.S. Patent No. 11928446)

Parallelizing Pruned Landmark Labeling: Dealing with Dependencies in Graph Algorithms 2020
Ruoming Jin, **Zhen Peng** (equal contribution), Wendell Wu, Feodor Dragan, Gagan Agrawal, Bin Ren
10.1145/3392717.3392745 (ICS 2020, the 34th ACM International Conference on Supercomputing)

ATMem: Adaptive Data Placement in Graph Applications on Heterogeneous Memories 2020
Yu Chen, Ivy Peng, **Zhen Peng**, Xu Liu, Bin Ren
10.1145/3368826.3377922 (CGO 2020, International Symposium on Code Generation and Optimization)

Pruned Landmark Labeling Meets Vertex Centric Computation: A Surprisingly Happy Marriage! 2019
Ruoming Jin, **Zhen Peng**, Wendell Wu, Feodor Dragan, Gagan Agrawal, Bin Ren
10.48550/arXiv.1906.12018 (arXiv 2019)

GraphPhi: Efficient Parallel Graph Processing on Emerging Throughput-oriented Architectures 2018
Zhen Peng, Alexander Powell, Bo Wu, Tekin Bicer, Bin Ren
10.1145/3243176.3243205 (PACT 2018, International Conference on Parallel Architectures and Compilation Techniques)

Reliable Wireless Connections for Fast-Moving Rail Users Based on a Chained Fog Structure 2017
Tian Wang, **Zhen Peng**, Sheng Wen, Weijia Jia, Yiqiao Cai, Hui Tian, Yonghong Chen
10.1016/j.ins.2016.06.031 (Information Sciences, Volume 379, pp 160-176, 2017)

Extracting Target Detection Knowledge Based on Spatio-temporal Information in Wireless Sensor Networks 2016
Tian Wang, **Zhen Peng**, Chen Wang, Yiqiao Cai, Yonghong Chen, Hui Tian, Junbin Liang, Bineng Zhong
10.1155/2016/5831471 (International Journal of Distributed Sensor Networks, Volume 12, No. 2, 2016)

Following Targets for Mobile Tracking in Wireless Sensor Networks 2016
Tian Wang, **Zhen Peng**, Junbin Liang, Sheng Wen, Md Zakirul Alam Bhuiyan, Yiqiao Cai, Jiannong Cao
10.1145/2968450 (ACM Transactions on Sensor Networks, Volume 12, Issue 4, No. 31, pp 1-24, 2016)

Dependable Cascading Target Tracking in Heterogeneous Mobile Camera Sensor Networks 2015
Zhen Peng, Tian Wang, Md Zakirul Alam Bhuiyan, Xiaoqiang Wu, Guojun Wang
10.1007/978-3-319-27161-3_48 (ICA3PP 2015, International Workshops and Symposiums on Algorithms and Architectures for Parallel Processing)

Detecting Targets Based on a Realistic Detection and Decision Model in Wireless Sensor Networks 2015
Tian Wang, **Zhen Peng**, Junbin Liang, Yiqiao Cai, Yonghong Chen, Hui Tian, Bineng Zhong
10.1007/978-3-319-21837-3_82 (WASA 2015, Wireless Algorithms, Systems, and Applications)

Continuous tracking for mobile targets with mobility nodes in WSNs 2014
Tian Wang, **Zhen Peng**, Yonghong Chen, Yiqiao Cai, Hui Tian
10.1109/SMARTCOMP.2014.7043867 (SmartComp 2014, International Conference on Smart Computing)

Service

Program Committee Member: ExHetAI 2025

Conference Reviewer: ICPP (2020, 2024, 2025), IPDPS (2021, 2022, 2025), HPDC-2026, HiPC (2018, 2019, 2023, 2024), ICAT-2023, PPoPP-2021, ICS-2021, NPC (2018, 2019, 2020), Bench (2019, 2020), BIGCOM-2019, ICCCN-2019, UIC-2018, SCS-2017

Journal Reviewer: ACM Transactions on Architecture and Code Optimization (TACO), IEEE Transactions on Cloud Computing (TCC), Expert Systems with Applications (ESWA)

Artifact Evaluation Committee Member: SC (2023, 2024, 2025, 2026), ALENEX (2024, 2025)