# Zhen Peng

hi.pengzhen@gmail.com (510) 931-8704 johnpzh.github.io 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, compiler, continuum computing, machine learning optimization.

## Education

**Ph.D.** **College of William & Mary**, Computer Science — Williamsburg, VA, USA
Aug 2016 – Jan 2023
- Advisor: Dr. Bin Ren
- 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
Sept 2013 – July 2016
- Advisor: Dr. Tian Wang
- Dissertation Title: Research on Target Tracking in Wireless Sensor Networks with Mobility Elements

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

## Experience

**Pacific Northwest National Laboratory (PNNL)**, Post Doctorate Research Associate — Richland, WA, USA (remote)
Apr 2023 – present
- Optimize distributed scientific workflows through prioritizing critical data flow
- Extend 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
June 2022 – Apr 2023
- Extend MLIR-based compiler for redundancy-aware code optimization

**Department of CS, College of William & Mary (W&M)**, Research Assistant — Williamsburg, VA, USA
Aug 2017 – June 2022
- Optimize deep neural network inference on edge devices
- Parallelize and optimize graph-based Approximate Nearest Neighbors Search (ANNS)
- Parallelize Pruned Landmark Labeling algorithm for 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
Apr 2021 – Sept 2021
- Automate the model implementation to TensorRT
- Accelerate inference through operation fusion in convolutional neural networks (CNNs)

## Projects

**FastFlow** — PNNL
Optimization for Distributed Scientific Workflows Through Prioritizing Critical Data Flow — Feb 2025 – present

- Implement the optimization method and pipeline to monitor and construct the critical data flow paths of given distributed workflows, using Nextflow and NetworkX
- Apply space folding and time folding for a given DAG workflow to model its parallelism and iterations, which is used to infer analytical rules to explain substructure scaling and predict edge properties

**AMAIS** PNNL
Compiler Support to Fabric Attached Memory (FAM) for Artificial Intelligence Aug 2024 – present
- Extend the compiler to support FAM that is enabled through Compute Express Link (CXL) protocol
- Optimize performance of AI applications that take advantage of a large memory pool through CXL

**COMET** PNNL
MLIR-Based Compiler for Computational Kernels on Heterogeneous Hardware June 2022 – Sept 2025
- Design and extend intermediate representation to support parallel graph kernel in compiler COMET
- Optimize code transformation for sparse computation, such as SpGEMM

**RedoxAI** PNNL
Generative AI for Efficient Prediction of Protein Redox Potentials June 2024 – Dec 2024
- Design and implement the LLMs-involved pipeline for training and predicting protein redox potentials
- For given magnitude and coordinates of a charge, predict its redox potential value. Conversely, for given redox potential, predict the possible magnitude and coordinate of a charge

**EdgeML** W&M
Accelerate Deep Neural Network Inference on Edge Devices June 2021 – June 2022
- Analyze inference procedure of TensorFlow Lite for Micro on microcontroller units (MCU)
- Speed up the inference procedure by tuning loop unrolling and customized quantization methods

**Speed-ANNS** W&M
Efficient Parallelization of Graph-based Approximate Nearest Neighbors Search (ANNS) Sept 2019 – Sept 2022
- 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

**Auto-TensorRT** Kuaishou, US R&D Center
Automate the Model Implementation to TensorRT Apr 2021 – Sept 2021
- Translate models from TVM Relay IR to TensorRT Python code
- Speed up the model deployment procedure and reduce the labor costs

**Fusion-TVM** Kuaishou, US R&D Center
Accelerate Inference Through Operation Fusion in Convolutional Neural Network (CNN) Apr 2021 – Sept 2021
- Try to add custom operation fusion pass in TVM
- Transform fused computational graph to TensorRT C++ code to accelerate the inference

**Parallel-PLL** W&M
Parallelizing Pruned Landmark Labeling – Dealing with Dependencies in Graph Algorithms Sept 2018 – Jan 2020
- 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

**GraphPhi** W&M
Efficient Parallelization of Graph Processing on Emerging Many-core Architectures July 2017 – Nov 2018

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

## Skills

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

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

## Publications

**FastFlow: Rapid Workflow Response By Prioritizing Critical Data Flows and their Interactions**
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**
**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**
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**
**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**
**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**
**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**
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**
**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**
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**

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!**
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**
**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**
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**
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**
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**
**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**
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**
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), 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, IEEE Transactions on Cloud Computing, Expert Systems with Applications

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