Pengcheng Wang

West Lafayette, US wang4495@purdue.edu

Website LinkedIn Google Scholar GitHub

Summary

Researcher with a decade of experience in algorithm-system co-design for efficient and adaptive AI computation across server GPUs, AI accelerators, and embedded edge platforms. Current work focuses on compute-adaptive vision-language models and foundation model deployment for intelligent systems, combining system-level optimization with content-, contention-, and noise-aware adaptation to sustain accuracy and throughput under tight latency, energy, and resource budgets. Earlier research spans high-dimensional numerical methods in computational electromagnetics and physical-layer algorithms for national 5G infrastructure, reflecting a consistent focus on closing the gap between computational demand and real-world hardware constraints.

Education

Ph.D. Candidate - Purdue University **Aug 2019 - Present**

Electrical and Computer Engineering. Advisors: Prof. Somali Chaterji, Prof. Saurabh Bagchi *West Lafayette, IN, US*

M.S. with Excellent Graduate Honor - Tongji University **Sep 2014 - Apr 2017**

Electronic Science and Technology. Advisor: Prof. Meisong Tong *Shanghai, China*

B.E. with Excellent Graduate Honor - Tongji University **Sep 2010 - Jun 2014**

Electronic Science and Technology. Advisor: Prof. Meisong Tong *Shanghai, China*

Research and Professional Experience

Machine Learning Engineer Intern, EmbodyX **Sep 2025 - Present**

Foundation model optimization for robotic systems *West Lafayette, IN*

- Investigating systems-level optimization techniques to accelerate training and inference of large language models (LLMs) and vision-language models (VLMs) for foundation model development.
- Exploring model compression and parameter-efficient fine-tuning methods to enable scalable deployment of foundation models on compute-constrained platforms.
- Designing system-level solutions that ensure robust inference performance of large-scale models in real-world deployment environments.

Adaptive Vision-Language Models (VLMs) under Resource Constraints **Jan 2025 - Present**

Purdue University. Advisors: Prof. Somali Chaterji, Prof. Yin Li, Prof. Saurabh Bagchi

- Characterizing the inference design space of vision-language models across token, width, and depth dimensions to identify fundamental efficiency–accuracy trade-offs.
- Developing a noise-resilient and compute-adaptive VLM framework that meets specified computational budgets while maximizing inference accuracy under degraded input conditions.
- Designing a learning-based inference scheduler that is jointly noise- and latency-aware, dynamically selecting execution paths to optimize per-sample resource allocation during inference.

Software Engineer Intern – AI Accelerator Toolchain, Sunlune Corp. **Jan 2025 - Aug 2025**

AI chip software infrastructure *Santa Clara, CA*

- Developed and validated kernel, runtime, and driver frameworks for domain-specific AI accelerators, integrating automated configuration systems and evaluating performance on cycle-accurate simulation platforms.
- Integrated large language model inference kernels and optimized runtime scheduling workflows to enable efficient execution of Llama-family models on custom AI hardware.
- Conducted systematic performance profiling and cross-platform debugging to identify computational bottlenecks and improve kernel execution efficiency on novel architectures.

Generative AI for Chip Design, Sunlune Corp. **May 2024 - Jan 2025**

AI-driven electronic design automation *Santa Clara, CA*

- Developed AI-driven design automation flows for high-performance digital ASICs, achieving significant reductions in design iteration time and improvements in power, performance, and area (PPA) metrics.
- Designed deep reinforcement learning models for logic synthesis and technology mapping, yielding a 10% improvement in delay-power product with enhanced scalability across diverse circuit designs.
- Integrated reinforcement learning with human feedback (RLHF) to encode IC design domain expertise into optimization models, enabling more adaptive and interpretable design space exploration.
- Developed heuristic-based post-processing techniques that produced substantial reductions in both power consumption and critical path delay.

Content-Aware Adaptive 3D Object Detection for Embedded GPUs **Aug 2022 - Jan 2025**

Purdue University. Advisors: Prof. Somali Chaterji, Prof. Yin Li, Prof. Saurabh Bagchi

- Proposed the first adaptive 3D object detection system for LiDAR point cloud data, capable of meeting runtime latency targets (35–200 ms) while maximizing detection accuracy on NVIDIA embedded GPUs.
- Conducted a systematic architectural analysis of 3D detection pipelines, identifying key computational bottlenecks in voxelization, voxel feature encoding, and 3D spatial feature extraction.
- Introduced five adjustable control knobs spanning the 3D encoder, CNN backbone, and detection head, enabling fine-grained accuracy–latency trade-offs through multiple execution branches.
- Deployed and evaluated on NVIDIA Jetson AGX Orin and Xavier; demonstrated 2–5% higher mean Average Precision than baselines at over 20 FPS, significantly outperforming CenterPoint and DSVT.

Adaptive and Efficient Video Object Detection on Edge Devices — Aug 2019 - Dec 2022

Purdue University. Advisors: Prof. Somali Chaterji, Prof. Yin Li, Prof. Saurabh Bagchi

- Performed comprehensive energy and latency benchmarking of 10+ object detection models (YOLO, EfficientDet, Faster R-CNN) on embedded platforms under varying power and thermal constraints.
- Evaluated 20+ classification and detection architectures on the ILSVRC Video dataset, systematically identifying optimal model configurations for resource-limited inference scenarios.
- Developed a synthetic 3D resource contention generator (CPU, memory bandwidth, GPU) to enable controlled evaluation of model robustness in multi-tenant edge computing environments.
- Established and maintained a heterogeneous embedded testbed comprising NVIDIA Jetson platforms (Nano, TX2, Xavier, Orin) with a complete software stack for reproducible algorithm development and evaluation.

Low-Power Wide-Area IoT Network for Smart Agriculture — Aug 2019 - Aug 2022

Project: Purdue-WHIN (Wabash Heartland Innovation Network) — *West Lafayette, IN*

- Investigated anomaly detection on large-scale soil and water sensor data using statistical and machine learning methods, leveraging AWS IoT Greengrass for edge-based inference.
- Engineered an automated network monitoring system providing daily anomaly detection alerts for 60+ environmental sensors deployed across four counties in Indiana.
- Deployed and managed a wide-area IoT network integrating 60+ sensors and 30+ Raspberry Pi edge nodes, ensuring robust database and web service infrastructure for real-time data access.

5G New Radio System Algorithm Engineer, ZTE Corporation — Mar 2017 - Jul 2019

5G NR Uplink Physical Layer R&D — *Shenzhen, China*

- Developed and integrated Tap Delay Line (TDL) and Cluster Delay Line (CDL) wireless channel models into the 5G NR system simulation platform, leveraging expertise in electromagnetic propagation modeling to improve channel characterization fidelity.
- Designed and implemented Windowed-OFDM signal processing algorithms to reduce spectral leakage and inter-carrier interference, improving the spectral efficiency of 5G uplink transmission.
- Formulated the Physical Random Access Channel (PRACH) algorithm for efficient large-scale initial network access, and conducted systematic performance validation under realistic deployment scenarios.
- Applied the K-Nearest Neighbors algorithm to optimize initial Outer Loop Power Control, demonstrating early integration of machine learning techniques into communication system optimization.
- Contributed to updates of the 3GPP 5G NR uplink physical channel protocol and participated in field coverage testing and data analysis to inform national-scale 5G deployment strategies.

Computational Electromagnetics Research, Tongji University — Sep 2013 - Mar 2017

Advisor: Prof. Meisong Tong. Research Assistant and Graduate Researcher — *Shanghai, China*

- Developed numerical solvers based on Volume Integral Equations (VIE) and Combined Field Integral Equations (CFIE) for large-scale electromagnetic scattering analysis of conducting, dielectric, and composite structures.
- Proposed a stable time-domain solution framework using the Nyström discretization scheme and Laguerre temporal basis functions for transient electromagnetic scattering by composite objects. Published in *IEEE Transactions on Antennas and Propagation* (2016).
- Designed mixed volume-surface integral equation formulations for microstrip antenna analysis and electromagnetic modeling of conducting-dielectric composite structures, with results presented at IEEE AP-S/URSI and IEEE ICCEM.
- Investigated the treatment of hypersingular integrals in volume integral equations using meshless methods, advancing the accuracy and computational efficiency of numerical electromagnetic solvers.

Teaching Experience

Teaching Assistant - Purdue University — Jan 2024 - May 2025

Department of Agricultural and Biological Engineering — *West Lafayette, IN*

- ABE 591: Machine Learning for IoT and Computer Systems, Spring 2025.
- ABE 591: Machine Learning for IoT and Computer Systems, Spring 2024.

Teaching Assistant - Tongji University **Sep 2014 - Jan 2017**

Department of Electronic Science and Technology *Shanghai, China*

- Semiconductor Physics, Fall 2016.
- Electromagnetic Fields and Waves, Spring 2016.
- Semiconductor Physics, Fall 2015.
- Electronics and Digital Technology, Spring 2015.
- Semiconductor Physics, Fall 2014.

Publications

- **Pengcheng Wang**, ZhuoMing Liu, Shayok Bagchi, Ran Xu, Saurabh Bagchi, Yin Li, and Somali Chaterji. "Agile3D: Adaptive Contention- and Content-Aware 3D Object Detection for Embedded GPUs." *The 23rd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), 2025*.
- **Pengcheng Wang**, Shayok Bagchi, Yin Li, and Somali Chaterji. "Adapt3D: Adaptive 3D Object Detection System for Embedded GPUs." *The 17th International Conference on COMmunication Systems & NETworkS (COMSNETS), 2025*.
- Lee, Jayoung, **Pengcheng Wang**, Ran Xu, Sarthak Jain, Venkat Dasari, Noah Weston, Yin Li, Saurabh Bagchi, and Somali Chaterji. "Virtuoso: Energy- and Latency-Aware Streamlining of Streaming Videos on SOCs." *ACM Transactions on Design Automation of Electronic Systems (TODAES), 2023*.
- Mousoulis, Christos, **Pengcheng Wang**, Nguyen Loc Do, Joseph F. Waimin, Nithin Raghunathan, Rahim Rahimi, et al. "An Open Dataset of Sensor Data from Soil Sensors and Weather Stations at Production Farms." *arXiv preprint arXiv:2302.09072, 2023*.
- Xu, Ran, Jayoung Lee, **Pengcheng Wang**, Saurabh Bagchi, Yin Li, and Somali Chaterji. "LiteReconfig: Cost and Content Aware Reconfiguration of Video Object Detection Systems for Mobile GPUs." *In Proceedings of the Seventeenth European Conference on Computer Systems (EuroSys), 2022*.
- Xu, Ran, Rakesh Kumar, **Pengcheng Wang**, Peter Bai, Ganga Meghanath, Somali Chaterji, Subrata Mitra, and Saurabh Bagchi. "ApproxNet: Content and Contention-Aware Video Object Classification System for Embedded Clients." *ACM Transactions on Sensor Networks (TOSN), 2021*.
- **Pengcheng Wang**, Jayoung Lee, Ran Xu, Venkat Dasari, Noah Weston, Yin Li, Saurabh Bagchi, and Somali Chaterji. "Benchmarking Video Object Detection Systems on Embedded Devices under Resource Contention." *In Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning (EMDL), 2021*.
- **Pengcheng Wang**, Edgardo Barsallo Yi, Tomas Ratkus, and Somali Chaterji. "ORPHEUS: Living Labs for End-to-End Data Infrastructures for Digital Agriculture." *arXiv preprint arXiv:2111.09422, 2021*.
- Xu, Ran, Chen-lin Zhang, **Pengcheng Wang**, Jayoung Lee, Subrata Mitra, Somali Chaterji, Yin Li, and Saurabh Bagchi. "ApproxDet: Content and Contention-Aware Approximate Object Detection for Mobiles." *In Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys), 2020*.
- Shankar, Karthick, **Pengcheng Wang**, Ran Xu, Ashraf Mahgoub, and Somali Chaterji. "Janus: Benchmarking Commercial and Open-Source Cloud and Edge Platforms for Object and Anomaly Detection Workloads." *In 2020 IEEE 13th International Conference on Cloud Computing (CLOUD), 2020*.
- **Pengcheng Wang**, Hao Sun, and Meisong Tong. "Numerical Solution of Electromagnetic Scattering by Very Slim Conducting Structures." *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (AP-S/URSI), 2017*.
- Tong, Meisong, and **Pengcheng Wang**. "Stable Solution of Time-Domain Combined Field Integral Equations for Transient Electromagnetic Scattering by Composite Structures Based on Nyström Scheme and Laguerre Function." *IEEE Transactions on Antennas and Propagation (TAP)*, vol. 64, no. 7, pp. 3239–3244, 2016.
- **Pengcheng Wang** and Meisong Tong. "Transient Analysis for Electromagnetic Scattering by Dielectric Objects Based on PMCHWT Equations." *Progress in Electromagnetics Research Symposium (PIERS), 2016*.
- **Pengcheng Wang**, Chunxia Yang, and Meisong Tong. "A Mixed Scheme for Analyzing Microstrip Antennas by Volume-Surface Integral Equations." *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (AP-S/URSI), 2016*.
- **Pengcheng Wang**, Xinzhou Zhao, Wenjie Chen, and Meisong Tong. "A Mixed Scheme for Solving Volume-Surface Integral Equations with Conducting-Dielectric Media." *IEEE International Conference on Computational Electromagnetics (ICCEM), 2016*.
- Chen, Wenjie, **Pengcheng Wang**, and Meisong Tong. "Transient Analysis for Electromagnetic Scattering by Composite Objects Based on Combined Field Integral Equations." *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (AP-S/URSI), 2016*.
- Chen, Wenjie, Guochun Wan, **Pengcheng Wang**, and Meisong Tong. "Accurate Solution of Time-Domain Electric Field Integral Equations for Transient Electromagnetic Scattering by Dielectric Objects." *IEEE International Conference on Computational Electromagnetics (ICCEM), 2016*.
- Zhang, J., R. P. Chen, **Pengcheng Wang**, and Meisong Tong. "Electromagnetic Modeling for a Miniaturized Patch Antenna with Thin Ferrite Films." *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (AP-S/URSI), 2015*.

• Tong, Meisong, Jie Zhang, **Pengcheng Wang**, and Jian Zhang. "Electromagnetic Analysis for Conductive Media Based on Volume Integral Equations." *IEEE Transactions on Antennas and Propagation (TAP)*, vol. 62, no. 12, pp. 6228–6235, 2014.

• **Pengcheng Wang**, Z. G. Zhou, J. H. Zhou, X. F. Yin, and Meisong Tong. "On the Treatment of Hypersingularity for Solving Volume Integral Equations." *Progress in Electromagnetics Research Symposium (PIERS), 2014*.

• **Pengcheng Wang**, Z. G. Zhou, J. X. Hong, Meisong Tong, and D. B. Miron. "Evaluation of Hypersingular Volume Integrals over a Cylindrical Domain in Meshless Methods." *IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting (AP-S/URSI), 2014*.

Honors and Awards

• NVIDIA Academic Grant Program, 2026

• Selected Mentee, GradBridge Program (Purdue & UC Berkeley), 2026

• Outstanding International Student Alumni Award, Tongji Foundation, 2025

• Dependable Computing Systems Laboratory Group Champion, Purdue University, 2025

• NSF Workshop on Grand Challenges in Resilience Poster Winner, Purdue University, 2024

• NSF Workshop on Grand Challenges in Resilience Poster Winner, Purdue University, 2023

• Excellent Graduate, Tongji University, 2017

• National Scholarship for Graduate Students, Ministry of Education of China, 2016

• Excellent Master Scholarship, Tongji University, 2015

• Excellent Graduate, Tongji University, 2014

• Scholarship for Good Academic Performance, Tongji University, 2013

• Social Activities Scholarship, Tongji University, 2012

Academic Service

• Program Committee, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), AI4Sciences Track, 2026

• Reviewer, Journal of Systems Architecture (Elsevier), 2026

• Shadow Program Committee, ACM SIGMETRICS, 2026

• Artifact Evaluation Committee, ACM European Conference on Computer Systems (EuroSys), 2026

• Artifact Evaluation Committee, ACM International Conference on Mobile Systems, Applications, and Services (MobiSys), 2025

• Shadow Program Committee, ACM European Conference on Computer Systems (EuroSys), 2024

• Artifact Evaluation Committee, ACM Conference on Embedded Networked Sensor Systems (SenSys), 2024

• Artifact Evaluation Committee, USENIX Symposium on Operating Systems Design and Implementation (OSDI) and USENIX Annual Technical Conference (ATC), 2022

Professional Development

• Building RAG Agents with LLMs, NVIDIA Deep Learning Institute

• Introduction to Transformer-Based Natural Language Processing, NVIDIA Deep Learning Institute

• Generative AI with Diffusion Models, NVIDIA Deep Learning Institute

• Certificate of Excellence for Deep Reinforcement Learning, Hugging Face

• Fundamentals of Accelerated Computing with CUDA Python, NVIDIA Deep Learning Institute

• Fundamentals of Accelerated Computing with CUDA C/C++, NVIDIA Deep Learning Institute

• Building Real-Time Video AI Applications, NVIDIA Deep Learning Institute

• Building Video AI Applications at the Edge on Jetson Nano, NVIDIA Deep Learning Institute

• Disaster Risk Monitoring Using Satellite Imagery, NVIDIA Deep Learning Institute

• Getting Started with AI on Jetson Nano, NVIDIA Deep Learning Institute

• Fundamentals of Deep Learning, NVIDIA Deep Learning Institute

• Fundamentals of Accelerated Data Science with RAPIDS, NVIDIA Deep Learning Institute