Po-An Tsai

Curriculum Vitae, July 2024

CONTACT INFORMATION

Po-An Tsai
2 Technology Park Dr, Westford, MA 01886
poant@nvidia.com
https://research.nvidia.com/person/po-an-tsai

RESEARCH INTERESTS

Computer systems and architecture; machine learning and tensor algebra accelerator architectures.

EDUCATION

Ph.D. in Computer Science, June 2019 — *Massachusetts Institute of Technology*
- Advisor: Professor Daniel Sanchez
- Thesis: Redesigning the Memory Hierarchy to Exploit Static and Dynamic Application Information

S.M. in Computer Science, June 2015 — *Massachusetts Institute of Technology*

B.S. in Electrical Engineering, June 2012 — *National Taiwan University (NTU), Taiwan*

HONORS AND AWARDS

Best Paper Award Nominee, ISCA-51, 2024
Distinguished Artifact Award, MICRO-55, 2022
IEEE Micro Top Picks Award, 2021
Best Paper Award Nominee, HPCA-21, 2015
Best Poster Award, MIT Industry-Academia Partnership Workshop, 2014
Jacobs Presidential Fellowship, MIT, 2013
Valedictorian, NTUEE, 2012
Presidential Award, NTU, 2010, 2011, 2012
Second Prize, NTUEE Undergraduate Research Award, 2012
Star Futures Award, Altera International FPGA Design Contest, 2011

WORK EXPERIENCE

Sr. Research Scientist, July 2022 – Current
Research Scientist, July 2019 – June 2022 — *Architecture Research Group, NVIDIA*
Research Assistant, September 2013 – May 2019 — *Computation Structures Group, MIT*
Software Engineering Intern, Summer 2015 — *Distributed Resource Management Team, VMware*

PUBLICATIONS

SIDA: Sparse Inter-operator Dataflow Architecture with Cross-Iteration Reuse
Yunan Zhang, **Po-An Tsai**, Hung-Wei Tseng
The 57th Annual International Symposium on Microarchitecture (MICRO-57), November 2024.

Mind the Gap: Attainable Data Movement and Operational Intensity Bounds for Tensor Algorithms
Qijing (Jenny) Huang, **Po-An Tsai**, Joel S Emer, Angshuman Parashar
The 51st Annual International Symposium on Computer Architecture (ISCA-51), June 2024.
Best Paper Award Nominee.

SDQ: Sparse Decomposed Quantization for LLM Inference
Geonhwa Jeong, **Po-An Tsai**, Stephen W. Keckler, Tushar Krishna
arXiv preprint arXiv:2406.13868, June 2024.

Abstracting Sparse DNN Acceleration via Structured Sparse Tensor Decomposition
Geonhwa Jeong, **Po-An Tsai**, Abhimanyu R. Bambhaniya, Stephen W. Keckler, Tushar Krishna
arXiv preprint arXiv:2403.07953, March 2024.

Symphony: Orchestrating Sparse and Dense Tensors with Hierarchical Heterogeneous Processing
Michael Pellauer, Jason Clemons, Vignesh Balaji, Neal Crago, Aamer Jaleel, Donghyuk Lee, Mike O'Connor, Anghsuman Parashar, Sean Treichler, **Po-An Tsai**, Stephen W. Keckler, Joel S. Emer
ACM Transactions on Computer Systems, Vol. 41, December 2023.

RM-STC: Row-Merge Dataflow Inspired GPU Sparse Tensor Core for Energy-Efficient Sparse Acceleration
Guyue Huang, Zhengyang Wang, **Po-An Tsai**, Chen Zhang, Yufei Ding, Yuan Xie
The 56th IEEE/ACM International Symposium on Microarchitecture (MICRO-56), October 2023.

HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity
Yannan Nellie Wu, **Po-An Tsai**, Saurav Muralidharan, Angshuman Parashar, Vivienne Sze, Joel S. Emer

The 56th IEEE/ACM International Symposium on Microarchitecture (MICRO-56), October 2023.

Accelerating Sparse Data Orchestration via Dynamic Reflexive Tiling
Toluwanimi O. Odemuyiwa, Hadi Asghari-Moghaddam, Michael Pellauer, Kartik Hegde, **Po-An Tsai**, Neal C. Crago, Aamer Jaleel, John D. Owens, Edgar Solomonik, Joel S. Emer, Christopher W. Fletcher
The 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-28), March 2023.

Demystifying Map Space Exploration for NPUs
Sheng-Chun Kao, Angshuman Parashar, **Po-An Tsai**, Tushar Krishna
2022 IEEE International Symposium on Workload Characterization (IISWC), November 2022.

Sparseloop: An Analytical Approach To Sparse Tensor Accelerator Modeling
Yannan Nellie Wu, **Po-An Tsai**, Angshuman Parashar, Vivienne Sze, Joel S. Emer
The 55th IEEE/ACM International Symposium on Microarchitecture (MICRO-55), October 2022.
Distinguished Artifact Award.

SIMD^2: A Generalized Matrix Instruction Set for Accelerating Tensor Computation beyond GEMM
Yunan Zhang, **Po-An Tsai**, Hung-Wei Tseng
The 49th Annual International Symposium on Computer Architecture (ISCA-49), June 2022.

Ruby: Improving Hardware Efficiency for Tensor Algebra Accelerators Through Imperfect Factorization
Mark Horeni, Pooria Taheri, **Po-An Tsai**, Angshuman Parashar, Joel S. Emer, Siddharth Joshi
IEEE International Symposium on Performance Analysis of Systems and Software, May 2022.

Union: A Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators
Geonhwa Jeong, Gokcen Kestor, Prasanth Chatarasi, Angshuman Parashar, **Po-An Tsai**, Sivasankaran Rajamanickam, Roberto Gioiosa, and Tushar Krishna
The 30th International Conference on Parallel Architectures and Compilation Techniques (PACT-30), September 2021.

Leaking Secrets through Compressed Caches
Po-An Tsai, Andres Sanchez, Christopher W. Fletcher, and Daniel Sanchez.
IEEE Micro's Top Picks from the Computer Architecture Conferences, May/June 2021.

Mind Mappings: Enabling Efficient Algorithm-Accelerator Mapping Space Search
Kartik Hegde, **Po-An Tsai**, Sitao Huang, Vikas Chandram, Angshuman Parashar, and Christopher W. Fletcher.
The 25th International Conference on Architectural Support for Programming Languages and Operating Systems, (ASPLOS-26), April 2021.

Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators
Yannan Nellie Wu, **Po-An Tsai**, Angshuman Parashar, Vivienne Sze, Joel S. Emer
IEEE International Symposium on Performance Analysis of Systems and Software, March 2021.

Hardware Abstractions for Targeting EDDO Architectures with the Polyhedral Model
Angshuman Parashar, Prasanth Chatarasi, and **Po-An Tsai**.
International Workshop on Polyhedral Compilation Techniques (IMPACT), January 2021.

Safecracker: Leaking Secrets through Compressed Caches
Po-An Tsai, Andres Sanchez, Christopher W. Fletcher, and Daniel Sanchez.
The 25th International Conference on Architectural Support for Programming Languages and Operating Systems, (ASPLOS-25), March 2020.

Compress Objects, Not Cache Lines: An Object-Based Compressed Memory Hierarchy
Po-An Tsai and Daniel Sanchez.
The 24th International Conference on Architectural Support for Programming Languages and Operating Systems, (ASPLOS-24), April 2019.

Rethinking the Memory Hierarchy for Modern Languages
Po-An Tsai, Yee Ling Gan, and Daniel Sanchez.
The 51st International Symposium on Microarchitecture (MICRO-51), October 2018.

Adaptive Scheduling for Systems with Asymmetric Memory Hierarchies
Po-An Tsai, Changping Chen, and Daniel Sanchez.
The 51st International Symposium on Microarchitecture (MICRO-51), October 2018.

KPart: A Hybrid Cache Partitioning-Sharing Technique for Commodity Multicores

Nosayba El-Sayed, Anurag Mukkara, **Po-An Tsai**, Harshad Kasture, Xiaosong Ma, and Daniel Sanchez.
The 24th Intl. Symposium on High Performance Computer Architecture (HPCA-24), February 2018.

Nexus: A New Approach to Replication in Distributed Shared Caches
Po-An Tsai, Nathan Beckmann, and Daniel Sanchez.
The 26th International Conference on Parallel Architectures and Compilation Techniques (PACT-26), September 2017.

Jenga: Software-Defined Cache Hierarchies
Po-An Tsai, Nathan Beckmann, and Daniel Sanchez.
The 44th International Symposium on Computer Architecture (ISCA-44), June 2017.

Scaling Distributed Cache Hierarchies with Computation and Data Co-Scheduling
Nathan Beckmann, **Po-An Tsai**, and Daniel Sanchez.
The 21st International Symposium on High Performance Computer Architecture (HPCA-21), February 2015.
Best Paper Award Nominee.

Hybrid Path-Diversity-Aware Adaptive Routing with Latency Prediction Model in Network-on-Chip Systems
Po-An Tsai, Yu-Hsin Kuo, En-Jui Chang, and An-Yeu Wu.
International Symposium on VLSI Design, Automation & Test, (VLSI-DAT), March 2013.

Path-Diversity-Aware Adaptive Routing in Network-on-Chip Systems
Yu-Hsin Kuo, **Po-An Tsai**, Hao-Ping Ho, En-Jui Chang, Hsien-Kai Hsin, and An-Yeu Wu.
The 6th International Symposium on Embedded Multicore SoCs (MCSoC), September 2012.

PATENT

Resource-Based Virtual Computing Instance Scheduling
US Patent 15283274
Po-An Tsai, Sahan Gamage, and Rean Griffith.

Flexible Accelerator for a Tensor Workload
US Patent 17343597, 17343582 **Po-An Tsai**, Neal Crago, Angshuman Parashar, Joel Springer Emer, Stephen William Keckler

Pruning and accelerating neural networks with hierarchical fine-grained structured sparsity
US Patent 17681967
Yannan Wu, **Po-An Tsai**, Saurav Muralidharan, Joel Springer Emer

Generating sparse neural networks
US Patent 18222916
Geonhwa Jeong, **Po-An Tsai**, Jeffrey Michael Pool

SERVICE

- Program Committee Member: ISMM'20, HPCA'21, MICRO'22, DAC'24, MICRO'24
- External Review Committee Member: PACT'20, MICRO'20, MICRO'21, ASPLOS'22, ISCA'22, ISCA'23, ISCA'24
- Reviewer: ACM TACO, ACM TCAD, IEEE JSSC, IEEE CAL
- Artifact Evaluation Co-Chair: ISCA'23.
- Workshop/Tutorial Co-Chair: HPCA'21.
- Submissions Co-Chair: MICRO'17
- Organizer: Workshop on Democratizing Domain-Specific Accelerators @ MICRO'22, MICRO'23, MICRO'24
- Organizer: Tutorial on Sparse Tensor Accelerators: Abstraction and Modeling @ ISCA'21.
- Organizer: Tutorial on Timeloop/Accelergy: Tools for Evaluating Deep Neural Network Accelerator Designs @ MICRO'19, ISCA'20, ISPASS'20.