

<b>CONTACT INFORMATION</b>	poant@nvidia.com	<a href="https://research.nvidia.com/person/po-an-tsai">https://research.nvidia.com/person/po-an-tsai</a>
<b>RESEARCH INTERESTS</b>	Computer system and architecture. Domain-specific (ML/DL) architecture. Accelerator modeling and prototyping. Memory hierarchy design. SW/HW co-optimization.	
<b>EDUCATION</b>	<p><b>Ph.D. in Computer Science</b>, June 2015 - June 2019 <span style="float: right;"><i>Massachusetts Institute of Technology</i></span></p> <ul style="list-style-type: none"> <li>• Advisor: Professor Daniel Sanchez</li> <li>• Thesis: Redesigning the Memory Hierarchy to Exploit Static and Dynamic Application Information</li> <li>• Minor: Optimization Methods</li> </ul> <p><b>S.M. in Computer Science</b>, June 2015 <span style="float: right;"><i>Massachusetts Institute of Technology</i></span></p> <ul style="list-style-type: none"> <li>• Advisor: Professor Daniel Sanchez</li> <li>• Thesis: Reducing Data Movement in Multicore Chips with Computation and Data Co-scheduling</li> </ul> <p><b>B.S. in Electrical Engineering</b>, June 2012 <span style="float: right;"><i>National Taiwan University (NTU), Taiwan</i></span></p>	
<b>HONORS AND AWARDS</b>	<p>Distinguished Artifact Award, MICRO-55, 2022</p> <p>IEEE Micro Top Picks Award, 2021</p> <p>Best Paper Nominee, HPCA-21, 2015</p> <p>Best Poster Award, MIT Industry-Academia Partnership Workshop – MIT, 2014</p> <p>Jacobs Presidential Fellowship – MIT, 2013</p> <p>Valedictorian – NTUEE, 2012</p> <p>Presidential Award – NTU, 2010, 2011, 2012</p> <p>Second Prize, NTUEE Undergraduate Research Award – NTU, 2012</p> <p>Star Futures Award, Altera International FPGA Design Contest – China, 2011</p>	
<b>WORK EXPERIENCE</b>	<p><b>Sr. Research Scientist</b>, July 2022 – Current <span style="float: right;"><i>Architecture Research Group, NVIDIA</i></span></p> <p><b>Research Scientist</b>, July 2019 – June 2022</p> <ul style="list-style-type: none"> <li>• Contribute to multiple generations of the hardware and software design for NVIDIA Tensor Cores</li> <li>• Co-lead research activities on hardware acceleration for autonomous machines</li> <li>• Worked on a flexible tensor accelerator that leverages a hierarchical and configurable data delivery network to adapt to a variety of GEMM and CNN workloads (<b>US Patent 17/343582, 17/343597</b>).</li> <li>• Contributed to an open-sourced tool (Timeloop+Acceleergy) for rapid evaluation of DNN accelerators.</li> <li>• Worked on analytical modeling methodology for sparse tensor accelerators (<b>MICRO-55</b>).</li> <li>• Worked on DARPA SDH program – Developing Software-Defined Hardware that enables near ASIC performance with high programmability for data-intensive algorithms.</li> </ul> <p><b>Research Assistant</b>, September 2013 – May 2019 <span style="float: right;"><i>Computation Structure Group, MIT, Cambridge MA</i></span></p> <p><b>Object-based memory hierarchies:</b></p> <ul style="list-style-type: none"> <li>• Hotpads (<b>MICRO-51</b>): designed an object-based memory hierarchy designed from the ground up for modern, memory-safe languages. Hotpads reduces memory hierarchy energy by 2.6×.</li> <li>• Zippads (<b>ASPLOS-24</b>): designed a compressed memory hierarchy for object-based programs. Zippads reduces main memory footprint by 2× while improving performance by 30%.</li> </ul> <p><b>Software-defined memory hierarchies:</b></p> <ul style="list-style-type: none"> <li>• Jenga (<b>ISCA-44</b>): designed a software-defined, heterogeneous memory hierarchy that adapts to the need of applications. Jenga improves full-system EDP by 23% on average and by up to 85%.</li> <li>• AMS (<b>MICRO-51</b>): proposed an analytical model and scheduling algorithms for systems with near-data processing (NDP) capabilities. AMS improves performance by up to 37%.</li> <li>• Nexus (<b>PACT-26</b>): developed an asymptotically better data replication policy for distributed shared caches. Nexus improves performance by 23% on average for replication-sensitive workloads.</li> </ul> <p><b>Ph.D. Intern</b>, Summer 2015 <span style="float: right;"><i>Distributed Resource Management Team, VMware, Palo Alto CA</i></span></p> <ul style="list-style-type: none"> <li>• Manager: Lan Gao    Mentor: Rean Griffith and Saham Gamage</li> <li>• Developed and prototyped a VM scheduler that performs multi-dimensional resource balancing and traffic engineering which reduces the runtime overhead by 10× while improving utilization by 5%.</li> <li>• The proposed algorithm is implemented in the 2016 release and filed as a US patent (US 15283274).</li> </ul>	

## PUBLICATIONS

### **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.

### **SIMD<sup>2</sup>: 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 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. **Nominated for the best paper award**

**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

**SKILLS  
AND TOOLS**

Programming languages and projects:

- **C, C++:** Analytical modeling tool for DNN accelerators (Timeloop), Event-driven multicore-processor simulator (Zsim)
- **Python, PyTorch:** Fine-tuning and accelerating sparse DNNs
- **Verilog:** FPGA-accelerated augmented-reality system, FPGA-accelerated medical image processing
- **CUDA, OpenCL:** GPGPU-assisted ultra-sonic array imaging system
- **Java:** Extending Maxine VM, a meta-circular JVM implementation for research
- **Python, Matlab:** Data analysis for predicting earning potential on an adult dataset
- **Python, HTML, Javascript:** Searchable and encrypted remote file system

**Libraries:** CUDA, OpenCL, matplotlib, Pytorch, TensorFlow

**Tools:** Git, Intel Pin, Timeloop, Zsim, ModelSim, Altera Quantus II, IC Encounter

**SERVICE**

- Program Committee Member, ISMM'20, HPCA'21, MICRO'22.
- External Review Committee Member, PACT'20, MICRO'20, MICRO'21, ASPLOS'22, ISCA'22, ISCA'23.
- Artifact Evaluation Co-Chair, ISCA'23.
- Organizer, Workshop on Democratizing Domain-Specific Accelerators @ MICRO'22
- Workshop/Tutorial Co-Chair, HPCA'21.
- Tutorial Organizer for Timeloop/Accelergy Tutorial: Tools for Evaluating Deep Neural Network Accelerator Designs, MICRO'19, ISCA'20, ISPASS'20.
- Submissions Co-Chair, MICRO-50, 2017
- President, 2014-2015, Taiwanese student association at MIT.