

CONTACT	poant@nvidia.com	https://research.nvidia.com/person/po-an-tsai
EDUCATION	Massachusetts Institute of Technology <i>Ph.D.</i> in Computer Science, minor in <i>Optimization Methods</i> <i>S.M.</i> in Computer Science <i>Advisor: Professor Daniel Sanchez</i>	September 2013 – June 2019 June 2015
	National Taiwan University <i>B.Sc.</i> in Electrical Engineering	June 2012
SKILLS AND TOOLS	Languages: C, C++, Python, Java, Verilog, Matlab, bash, SQL Libraries: CUDA, OpenCL, matplotlib, Pytorch, TensorFlow Tools: Git, Intel Pin, Timeloop, Zsim, ModelSim, Altera Quantus II, IC Encounter	
WORK EXPERIENCE	NVIDIA Research, Westford MA <i>Sr. Research Scientist</i> <i>Research Scientist</i>	July 2022 – Current July 2019 – June 2022
	<p>I develop architectures to address the emerging demands of computer vision and machine learning algorithms. This task requires understanding and analyzing the interplay between hardware, software, and algorithms. Specifically, I work on developing future tensor accelerator that accelerates a wider range of tensor algorithms than conventional accelerators. To evaluate designed accelerators, I use and contribute to an open-source analytical modeling tool (Timeloop+Accelergy) for rapid evaluation of DNN accelerators. My research has influenced how future generations of HW and SW systems will be designed at NVIDIA.</p> <p>I also collaborate with teams across the company, spanning software, research, engineering, and product groups, and publish original research and speak at conferences and events.</p>	
	MIT Computer Science & Artificial Intelligence Lab, Cambridge MA <i>Research Assistant</i>	September 2013 – June 2019
	<p>My Ph.D. research focuses on reducing data movement in computer systems to improve their performance and energy efficiency. I designed new memory hierarchies, developed algorithms for data placement and workload scheduling, and co-designed hardware/software to optimize systems.</p> <p>Across my projects, I prototyped ideas extending Zsim, a C++, Intel Pin-based open-sourced multicore simulator. I leveraged latest commodity hardware features (e.g., Intel CAT) and profiled workloads using hardware performance counters. I made essential changes throughout the software stack, including applications (e.g., key-value store, graph analytics) and runtime/compiler in Maxine, a Java-based research JVM.</p>	
	VMware, Palo Alto CA <i>Ph.D. Intern</i>	June 2015 – August 2015
	<p>Distributed Resource Management Team. Worked on a VM scheduler that performs multi-dimensional resource balancing and traffic engineering. Proposed a randomized and graph-clustering-based algorithm and evaluated it using a trace-driven simulator written in Python. My algorithm reduces the runtime overhead by 10× while improving utilization by 5% and was publicly released in 2016 and filed as a US patent.</p>	
PATENT	Resource-Based Virtual Computing Instance Scheduling Po-An Tsai, Sahan Gamage, and Rean Griffith.	US 15283274
	Flexible Accelerator for a Tensor Workload Po-An Tsai, Neal Crago, Angshuman Parashar, Joel Springer Emer, Stephen William Keckler	US Patent 17343597, 17343582
RECENT PUBLICATIONS	Sparseloop: An Analytical Approach To Sparse Tensor Accelerator Modeling Yannan Nellie Wu, Po-An Tsai , Angshuman Parashar, Vivienne Sze, Joel S Emer.	MICRO-55
	SIMD²: A Generalized Matrix Instruction Set for Accelerating Tensor Computation beyond GEMM Yunan Zhang, Po-An Tsai , Hung-Wei Tseng.	ISCA-49