Thierry Tambe

CONTACT

Paul G. Allen Building
330 Jane Stanford Way
Stanford, CA 94305

Web: https://thierrytambe.com
E-mail: ttambe@stanford.edu

RESEARCH INTERESTS

I work at the intersection of VLSI design, computer architecture, and machine learning to co-design solutions across the hardware-software computing stack with the goal of overcoming fundamental limitations we now face due to the end of Dennard Scaling — and validating these proof-of-concepts in ASIC chip tape-outs. I am interested in developing novel algorithms, memory systems, specialized hardwares, and scalable silicon systems for emerging computation- and memory-intensive applications, while tuning their designs and inter-dependencies to promote greater performance, efficiency, reliability, and TCO.

EDUCATION

Harvard University, Cambridge, MA
Ph.D., Electrical Engineering 2023
Thesis Title: *Architecting High Performance Silicon Systems for Accurate and Efficient On-Chip Deep Learning.*

Texas A&M University, College Station, TX
M.Eng., Electrical Engineering 2012
B.S., Electrical Engineering 2010

AWARDS AND HONORS

NVIDIA Graduate Fellowship 2021 - 2022
- *One of 5 honored out of 350+ applicants. Awarded $50K towards tuition and stipend.*

IEEE SSCS Predoctoral Achievement Award 2021 - 2022
- *$1K honorarium recognizing outstanding PhD students in the field of Solid-State Circuits.*

IEEE MICRO Top Picks Honorable Mention 2022
- *Top 24 across all papers published at top-tier computer architecture venues in 2021.*

Finalist for the Lemelson-MIT Student Prize 2021
- *One of 12 finalists out of 100+ student inventors across the United States.*

ACM SIGDA Research Highlights Nominee 2021
- *Nominee out of top 10 papers published in ACM SIGDA sponsored conferences in 2020.*

Best Paper Award at ACM/IEEE Design Automation Conference (DAC) 2020
- *Top honor out of 228 accepted papers published at DAC.*

PROFESSIONAL EXPERIENCE

Stanford University, Stanford, CA
Assistant Professor, Department of Electrical Engineering 2024 - Present

NVIDIA, Santa Clara, CA
Research Scientist, Accelerators & VLSI Research Group 2023 - Present

Harvard University, Cambridge, MA
Graduate Researcher, Harvard Architecture, Circuits, and Compilers Group 2018 - 2023
- Investigating and building cross-stack solutions (algorithms, hardware architectures, emerging memories, real-time systems, and silicon tape-outs) for on-chip machine learning and emerging computation- and memory-intensive applications.
- Advisors: Prof. Gu-Yeon Wei, Prof. David Brooks

NVIDIA, Virtual
Research Intern, ASIC & VLSI Research Group 2021 - 2021

- Investigated the software/hardware co-design space of various numerical data types for optimality in machine learning training.
- Proposed, and prototyped a novel custom data type for efficient deep learning training. (*patent application submitted*)

Intel, Hillsboro, OR
Senior Design Engineer, Scalable Perf. CPU Development Group 2012 - 2017

- Owned the path-finding, architecture and design of a portfolio of mixed-signal circuits (high-speed receivers and transmitters, voltage regulators, clocking) for High Bandwidth Memory (HBM), successfully proven on mass-produced 14nm Xeon and Xeon-Phi server CPU chips.
- Chaired the *Simulation and Methodology Tech. Work Group* which standardized circuit simulation methodologies across Intel for reliable PVT, timing, variation, and aging analyses.

Intel, Hillsboro, OR
Graduate Intern, Converged Core Development Organization 2011 - 2011

- Developed scripted utilities to automate the verification process of Haswell processor mixed-signal circuits.

Biotronik, Sugar Land, TX
IC Design Intern, Texas Design Center 2010 - 2010

- Designed ultra-low power analog circuits over and near subthreshold region for implantable cardiovascular devices.

Conference Publications

1. **A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management**
   Thierry Tambe, Jeff Zhang, Coleman Hooper, Tianyu Jia, Paul N. Whatmough, Joseph Zuckerman, Maico Cassel Dos Santos, Erik Jens Loscalzo, Davide Giri, Kenneth Shepard, Luca Carloni, Alexander Rush, David Brooks, Gu-Yeon Wei.
   International Solid-State Circuits Conference, 2023. (**ISSCC'23**).

2. **ASAP: Automatic Synthesis of Area-Efficient and Precision-Aware CGRAs**
   Chen Tan, **Thierry Tambe**, Jeff Zhang, Bo Fang, Tong Geng, Gu-Yeon Wei, David Brooks, Antonino Tumeo, Ganesh Gopalakrishnan, Ang Li.
   ACM International Conference on Supercomputing, 2022. (**ICS'22**).

3. **GoldenEye: A Platform for Evaluating Emerging Data Formats in DNN Accelerators**
   Abdulrahman Mahmoud, **Thierry Tambe**, Tarek Aloui, David Brooks, Gu-Yeon Wei.
   IEEE International Conference on Dependable Systems and Networks, 2022. (**DSN'22**).
   Code available on GitHub.

4. **EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference**
   Thierry Tambe, Coleman Hooper, Lillian Pentecost, Tianyu Jia, En-Yu Yang, Marco Donato, Victor Sanh, Paul Whatmough, Alexander Rush, David Brooks, Gu-Yeon Wei.
   International Symposium on Microarchitecture, 2021. (**MICRO'21**).

Artifact Badges: Available, and Functional

5. **A 25mm^2 SoC for IoT Devices with 18ms Noise Robust Speech-to-Text Latency via Bayesian Speech Denoising and Attention-Based Sequence-to-Sequence DNN Speech Recognition in 16nm FinFET.**
   Thierry Tambe, En-Yu Yang, Glenn G. Ko, Yuji Chai, Coleman Hooper, Marco Donato, Paul Whatmough, Alexander Rush, David Brooks, Gu-Yeon Wei.
   International Solid-State Circuits Conference, 2021. (**ISSCC'21**).
   Code available on GitHub.

6. **Robomorphic Computing: A Design Methodology for Domain-Specific Accelerators Parameterized by Robot Morphology**
   Sabrina M. Neuman, Brian Plancher, Thomas Bourgeat, **Thierry Tambe**, Srinivas Devadas, Vijay Janapa Reddi.
   International Conference on Architectural Support for Programming Languages and Operating Systems, 2021. (**ASPLOS'21**).
   IEEE MICRO Top Picks Honorable Mention

7. **A Scalable Bayesian Inference Accelerator for Unsupervised Learning**
   Glenn G. Ko, Yuji Chai, Marco Donato, Paul Whatmough, **Thierry Tambe**, Rob A. Rutenbar, David Brooks, Gu-Yeon Wei.
   IEEE Hot Chips Symposium, 2020. (**Hot Chips'20**).

8. **A 3mm2 Programmable Bayesian Inference Accelerator for Unsupervised Machine Perception using Parallel Gibbs Sampling in 16nm**
   Glenn G. Ko, Yuji Chai, Marco Donato, Paul Whatmough, **Thierry Tambe**, Rob A. Rutenbar, David Brooks, Gu-Yeon Wei.
   Symposia on VLSI Technology and Circuits, 2020. (**VLSI'20**).

9. **Algorithm-Hardware Co-design of Adaptive Floating-Point Encodings for Resilient Deep Learning Inference**
   Thierry Tambe, En-Yu Yang, Zishen Wang, Yuntian Deng, Vijay Janapa Reddi, Alexander Rush, David Brooks, Gu-Yeon Wei.
   ACM/IEEE Design Automation Conference, 2020. (**DAC'20**).
   Best Paper Award
   Nominee for ACM SIGDA Research Highlights

10. **MASR: A Modular Accelerator for Sparse RNNs**
    Udit Gupta, Brandon Reagen, Lillian Pentecost, Marco Donato, **Thierry Tambe**, Alexander Rush, Gu-Yeon Wei, David Brooks
    International Conference on Parallel Architectures and Compilation Techniques, 2019. (**PACT'19**).
    Best Paper Nominee

Journal Publications

1. **A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference With Bayesian Sound Source Separation and Attention-Based DNNs**
   Thierry Tambe, En-Yu Yang, Glenn G. Ko, Yuji Chai, Coleman Hooper, Marco Donato, Paul Whatmough, Alexander Rush, David Brooks, Gu-Yeon Wei.
   IEEE Journal of Solid-State Circuits, 2022. (**JSSC'22**).

Workshop Publications

1. **Learnings from a HLS-based High-Productivity Digital VLSI Flow**
   Thierry Tambe, David Brooks, Gu-Yeon Wei.
   Workshop on Languages, Tools, and Techniques for Accelerator Design, 2022. (**LATTE'22**).

2. **From DSLs to Accelerator-rich Platform Implementations: Addressing the Mapping Gap**
   Bo-Yuan Huang*, Steven Lyubomirsky*, **Thierry Tambe***, Yi Li, Mike He, Gus Smith, Gu-Yeon Wei, Aarti Gupta, Sharad Malik, and Zachary Tatlock.
   Workshop on Languages, Tools, and Techniques for Accelerator Design, 2021. (**LATTE'21**).

ArXiv Publications

1. **CAMEL: Co-Designing AI Models and Embedded DRAMs for Efficient On-Device Learning**
   Sai Qian Zhang*, **Thierry Tambe***, Nestor Cuevas, Gu-Yeon Wei, and David Brooks.
   arXiv:2305.03148, 2023.

2. **Specialized Accelerators and Compiler Flows: Replacing Accelerator APIs with a Formal Software/Hardware Interface**
   Bo-Yuan Huang, Steven Lyubomirsky, Yi Li, Mike He, **Thierry Tambe**, Gus Henry Smith, Akask Gaonkar, Vishal Canumalla, Gu-Yeon Wei, Aarti Gupta, Zachary Tatlock, Sharad Malik.
   arXiv:2203.00218, 2022.

3. **AdaptivFloat: A Floating-Point based Data Type for Resilient Deep Learning**
   Thierry Tambe, En-Yu Yang, Zishen Wang, Yuntian Deng, Vijay Janapa Reddi, Alexander Rush, David Brooks, Gu-Yeon Wei.
   arXiv:1909.13271, 2019.
   Code available on GitHub.

Chip Tapeouts

1. **A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management**
   A 4.60mm^2 sparse Transformer processor that dynamically tailors its energy and latency expenditures according to the complexity of the input query it processes.
   Process technology: GlobalFoundries 12LP
   Tapeout date: October 2021
   Publication: ISSCC 2023

2. **A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference With Bayesian Sound Source Separation and Attention-Based DNNs**
   A 25mm^2 many-accelerators IoT SoC with specialized processing of attention-based DNNs and Bayesian workloads.
   Process technology: TSMC 16FFC
   Tapeout date: June 2019
   Publication: JSSC 2022, ISSCC 2021

3. **A Scalable Bayesian Inference Accelerator for Unsupervised Learning**
   A 3mm^2 programmable processor for unsupervised probabilistic machine perception tasks.
   Process technology: TSMC 16FFC
   Tapeout date: May 2018
   Publication: Hot Chips 2020, VLSI 2020

Seminar and Invited Talks

1. *Effective SW/HW Co-Design of Specialized ML Accelerators using HLS*
   - Invited webinar, Siemens (*1500+ attendees, a Siemens webinar record!*) Feb 2022
2. *SM6: A 16nm System-on-Chip for Accurate and Noise-Robust Attention-Based NLP Applications*
   - Poster, Hot Chips' 33. Aug 2021
   - Poster, Arm Research Summit, Austin, TX. Sep 2019
   - Invited talk, Samsung Adv. Inst. of Tech., Suwon, South Korea Jul 2019
3. *Algorithms, Architectures, and Prototypes for Accurate and Noise-Robust Speech and Natural Language Processing Inference*
   - Invited talk, Cornell Computer Systems Laboratory (CSL) Apr 2021
   - Invited talk, IBM 5th Workshop on the Future of Computing Arch. Nov. 2020
4. *AdaptivFloat: A Data Type for Resilient Deep Learning Inference*
   - Invited talk, FPTalks Jun 2020
5. *Closing the algorithm/hardware design and verification loop with speed via high-level synthesis*
   - Invited talk, CHIPKIT Tutorial at ISCA'20 May 2020
6. *Open Edge Hardware and Software for Natural Language Translation and Understanding*
   - Invited talk, FOSDEM, Brussels, Belgium Feb 2020
7. *Adaptive Quantization of Deep Neural Networks*
   - Poster, Computing Research Assoc. URMD, Waikoloa, Hawaii Mar 2019

Students Mentored

- **Alicia Golden** (1st year PhD student) Sep 2022 - Present
  Evaluating scaling trends and timing critical paths in ARM CPUs.
- **Nestor Cuevas** (2nd year PhD student) Nov 2021 - Present
  Design and Characterization of Embedded DRAM Memories for Efficient ML Training.
- **Coleman Hooper** (4th year undergrad student) Feb 2020 - May 2022
  Hardware-Software Co-Design for Energy-Efficient Deployment of Transformer-Based Speech Recognition Models on Edge Devices.
- **Maria Sturzu** (4th year undergraduate student) Mar 2020 - Aug 2020
  FPGA Prototyping of HLS-based Machine Learning Accelerators.
- **Zishen Wan** (2nd year Master student) Jan 2019 - Dec 2019
  Study of Posit Numeric in Speech Recognition Neural Inference.

Teaching Experience

Graduate Teaching Fellow — Harvard University, Cambridge, MA

CS248 – Advanced Design of VLSI Circuits and Systems — Spring 2020

- *Designed lab materials on designing HLS-based AI hardware accelerators.*
- *Hosted recitation sections, office hours, and graded students' problem sets and lab assignments.*

MIT ESP Spark Educational Outreach — Mar 2021

- *Taught a course on the basics of chip development for AI to Greater Boston middle schoolers.*

Academic Service

Organizing Committee

- The NOPE Workshop @ASPLOS, 2022

Invited Reviewer

- Design Automation Conference (DAC), 2022
- IEEE Transactions on Neural Networks and Learning Systems, 2021

Professional Memberships

ACM, IEEE, Black in AI

Other Information

Languages: English (fluent), French (native), Spanish (conversational)