2025 IEEE Hot Chips 37 Symposium (HCS 2025)

Stanford, California, USA
24-26 August 2025

IEEE Catalog Number: CFP25HCS-POD
ISBN: 979-8-3315-0301-7

**Copyright © 2025 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP25HCS-POD
ISBN (Print-On-Demand):	979-8-3315-0301-7
ISBN (Online):	979-8-3315-0300-0
ISSN:	2573-203X

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings.com

TABLE OF CONTENTS

AMD RDNA 4 Radeon 9000 Series GPU 1
Andy Pomianowski, Laks Pappu

Clearwater Forest the Next Generation Intel® Xeon® Processor with Efficiency Cores 13
Don Soltis, Stephen Robinson

NVIDIA GB10 SoC: AI Supercomputer on Your Desk 21
Andi Skende

PEZY-SC4s: The Fourth Generation MIMD Many-core Processor with High Energy Efficiency and Flexibility for HPC and AI Applications 28
Naoya Hatta, Shuntaro Tsunoda, Kouhei Uchida, Taichi Ishitani, Toru Koizumi, Ryota Shioya, Kei Ishii

IBM's Power11 Processor 49
William Starke

CORSAIR: An In-Memory Computing ChipletArchitecture for Inference-Time Compute Acceleration 64
Sudeep Bhoja

RTX 5090: Designed for the Age of Neural Rendering 83
Marc Blackstein

Everactive Self-Powered SoC with Energy Harvesting, Wakeup Receiver, and Energy-Aware Subsystem 93
Benton H. Calhoun, David D. Wentzloff, Kuo-Ken Huang, Kyle Craig

Memory: Almost the Only Thing that Matters: A Revolution in Memory Architecture for the Data Center 109
Mark Kuemerle

Cuzco: A High-Performance RISC-V RVA23 Compatible CPU IP 124
Ty Garibay, Shashank Nemawarkar

Passage M1000: A 3D Photonic Interposer for AI 135
Darius Bunandar

Ironwood: Delivering Best in Class Perf, perf/TCO and perf/Watt for Reasoning Model Training and Serving 152
Norman P. Jouppi, Sridhar Lakshmanamurthy

FABRIC8LABS Electrochemical Additive Manufacturing ECAM Enabled Thermal Solutions for the Al Data Center 165
Michael Matthews, Ian Winfield, Joseph Madril, Douglas De Aquino Castro, Charles Biset

UB-Mesh: An New Interconnection Technology for Large AI SuperNode 173
Heng Liao

Presto: A Unified RISC-V-Compatible SoC for Multi-Scheme FHE Acceleration Over Module Lattice 180
Luchang Lei, Yu Duan, Cheng Peng, Yongqing Zhu, Gangfeng Du, Zhenyu Guan, Huazhong Yang, Yongpan Liu, Zhen Gu, Song Bian, Hongyang Jia

ConnectX-8 SuperNIC 200
Idan Burstein

A 4.69mW LLM Processor with Binary/Ternary Weights for Billion-Parameter Llama Model 211
Sangyeob Kim, Jungwan Lee, Byeongju Kim, Hoi-Jun Yoo

Azure Secure Hardware Architecture: A Robust Security Foundation for Cloud Workloads 220
Bryan Kelly

A UCIe Optical I/O Retimer Chiplet for AI Scale-Up 232
Vladimir Stojanovic

Taping Out Three Class Chips Per Semester in Intel 16 Technology 243
Lucy Revina, Ethan Gao, Ken Ho, Daniel Lovell, Kristofer Pister, Borivoje Nikolic

Photonic Interconnect for Accelerated Computing Celestial AI Photonic Fabric Module – The World's First SoC with In-Die Optical IO 270
Phil Winterbottom

Co-Packaged Silicon Photonics Switches for Gigawatt AI Factories 277
Gilad Shainer

Basilisk: A 34mm² End-To-End Open-Source 64-bit Linux-Capable RISC-V SoC in 130nm BiCMOS: Integrated Systems Laboratory (ETH Zürich) 289
Philippe Sauter, Thomas Benz, Paul Scheffler, Martin Poviser, Frank K. Gürkaynak, Luca Benini

Intel® IPU E2200: Second Generation Infrastructure Processing Unit (IPU) 296
Pat Fleming, Chihjen Chang, Derek Collier, Anjali Singhai, Stephen Doyle, Eliel Louzoun, David Lee, Vetrivel Ayyavu, Sarig Livne, Robert Hathaway, Tony Hurson, Jackson Ellis, Tamar Bar-Kanarik, Jonathan Kenny, Cristine Dumitrescu, Yaron Wolberger

Bit-Separable Transformer Accelerator Leveraging Output Activation Sparsity for Efficient DRAM Access 304
Seunghyun Park, Daejin Park

Specialized IC for World-Lock Rendering in Augmented and Mixed Reality Devices 309
Ohad Meitav, Jay Tsao

KLIMA: Low-Latency Mixed-signal In-Memory Computing Accelerator for Solving Arbitrary-order Boolean Satisfiability 318
Tinish Bhattacharya, Dongseok Kwon, George Hutchinson, Xiangyi Zhang, Giacomo Pedretti, Fabian Bohm, John Paul Strachan, Thomas Van Vaerenbergh, Ray Beausoleil, Ignacio Rozada, Dmitri Strukov

4th Gen AMD CDNA™ Generative AI Architecture Powering AMD Instinct M350 Series GPUs and Platforms 322
Michael Steffen, Michael Floyd

AMD Pensando™ Pollara 400 AI NIC Architecture and Application 341
Kevin Chu

Hot ChipsKeynote 359
Noam Shazeer

Up and Running with Rapidus: How Japan and Cutting-Edge Technologies Are Transforming Semiconductor Manufacturing 382
Atsuyoshi Koike

ENABLING AI Infrastructure: Tomahawk Ultra - Ultra Low Latency, High Bandwidth Ethernet Switch for HPC & AI/ML Applications 407
Mohan Kalkunte, Asad Khamisy

Author Index