

2024 IEEE Hot Chips 36 Symposium (HCS 2024)

**Stanford, California, USA
25-27 August 2024**



**IEEE Catalog Number: CFP24HCS-POD
ISBN: 979-8-3503-8851-0**

**Copyright © 2024 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:	CFP24HCS-POD
ISBN (Print-On-Demand):	979-8-3503-8851-0
ISBN (Online):	979-8-3503-8850-3
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

IBM Telum [®] II Processor and IBM Spyre [™] Accelerator Chip for AI	1
<i>Chris Berry</i>	
SK Hynix AI-Specific Computing Memory Solution: from AiM Device to Heterogeneous AiMX-xPU System for Comprehensive LLM Inference	16
<i>Guhyun Kim, Jinkwon Kim, Nahsung Kim, Woojae Shin, Jongsoon Won, Hyunha Joo, Haerang Choi, Byeongju An, Gyeongcheol Shin, Dayeon Yun, Jeongbin Kim, Changhyun Kim, Ilkon Kim, Jaehan Park, Yosub Song, Byeongsu Yang, Hyeongdeok Lee, Seungyeong Park, Wonjun Lee, Seonghun Kim, Yonghoon Park, Yousub Jung, Gi-Ho Park, Euicheol Lim</i>	
Built for the Edge: The Intel [®] Xeon [®] 6 SoC	29
<i>Praveen Mosur</i>	
MN-Core 2: Second-Generation Processor of MN-Core Architecture for AI and General-Purpose HPC Application.....	43
<i>Jun Makino</i>	
Towards “True” GPU Performance Scaling for OpenGPU	54
<i>Blaise Tine, Hyesoon Kim</i>	
Tesla Transport Protocol Over Ethernet (TTPoE): A New Lossy, Exa-Scale Fabric for the Dojo AI Supercomputer.....	59
<i>Eric Quinnell</i>	
AMD Versal [™] AI Edge Series Gen 2 for Vision and Automotive	71
<i>Tomai Knopp, Jeffrey Chu, Sagheer Ahmad</i>	
AMD Instinct MI300X Generative AI Accelerator and Platform Architecture	85
<i>Alan Smith, Vamsi Alla</i>	
Next Generation “Zen 5” Core	96
<i>Brad Cohen, Mahesh Subramony, Mike Clark</i>	
Qualcomm Oryon [™] CPU	110
<i>Gerard Williams</i>	
Onyx: A Programmable Accelerator for Sparse Tensor Algebra	121
<i>Kalhan Koul, Maxwell Strange, Jackson Melchert, Alex Carsello, Yuchen Mei, Olivia Hsu, Taeyoung Kong, Po-Han Chen, Hui Feng Ke, Keyi Zhang, Qiaoyi Liu, Gedeon Nyengele, Akhilesh Balasingam, Jayashree Adivarahan, Ritvik Sharma, Zhouhua Xie, Christopher Torng, Joel Emer, Fredrik Kjolstad, Mark Horowitz, Priyanka Raina</i>	
XiangShan: An Open-Source Project for High-Performance RISC-V Processors Meeting Industrial-Grade Standards	167
<i>Kaifan Wang, Jian Chen, Yinan Xu, Zihao Yu, Zifei Zhang, Guokai Chen, Xuan Hu, Linjuan Zhang, Xi Chen, Wei He, Dan Tang, Ninghui Sun, Yungang Bao</i>	
LSPU: AF 20.7 Ms Low-Latency Point Neural Network-Based 3D Perception and Semantic LiDAR SLAM System-on-Chip for Autonomous Driving System.....	180
<i>Jueun Jung, Seungbin Kim, Bokyoung Seo, Wuyoung Jang, Sangho Lee, Jeongmin Shin, Donghyeon Han, Kyuho Jason Lee</i>	

SambaNova SN40L RDU: Breaking the Barrier of Trillion+ Parameter Scale Gen AI Computing.....	194
<i>Raghu Prabhakar</i>	
Lunar Lake Architecture Session.....	206
<i>Arik Gihon</i>	
Intel Gaudi 3 AI Accelerator: Architected for Gen AI Training and Inference	231
<i>Roman Kaplan</i>	
An AI Compute ASIC with Optical Attach to Enable Next Generation Scale-Up Architectures	239
<i>Manish Mehta</i>	
4 Tb/s Optical Compute Interconnect Chiplet for XPU-To-XPU Connectivity	254
<i>Saeed Fatholouloumi</i>	
NVIDIA Blackwell Platform: Advancing Generative AI and Accelerated Computing	263
<i>Ajay Tirumala, Raymond Wong</i>	
Sustainable Computing for AI & Cloud Native Workloads.....	280
<i>Matthew Erler</i>	
Wafer-Scale AI: GPU Impossible Performance.....	292
<i>Sean Lie</i>	
ACF-S: An 8-Terabit / Sec SuperNIC for High-Performance Data Movement in AI & Accelerated Compute Networks.....	328
<i>Shrijeet Mukherjee, Thomas Norrie</i>	
Inside Maia 100	341
<i>Sherry Xu, Chandru Ramakrishnan</i>	
Next Gen MTIA -Recommendation Inference Accelerator	350
<i>Mahesh Maddury, Pankaj Kansal, Olivia Wu</i>	
RNGD - Tensor Contraction Processor for Sustainable AI Computing.....	364
<i>June Paik</i>	
Blackhole & TT-Metalium: The Standalone AI Computer and Its Programming Model	381
<i>Jasmina Vasiljevic, Davor Capalija</i>	
The Journey to AI Pervasiveness.....	396
<i>Victor Peng</i>	
Predictable Scaling and Infrastructure.....	414
<i>Trevor Cai</i>	

Author Index