

2023 IEEE Hot Chips 35 Symposium (HCS 2023)

**Palo Alto, California, USA
27-29 August 2023**



**IEEE Catalog Number: CFP23HCS-POD
ISBN: 979-8-3503-3908-6**

**Copyright © 2023 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:	CFP23HCS-POD
ISBN (Print-On-Demand):	979-8-3503-3908-6
ISBN (Online):	979-8-3503-3907-9
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

Shaheen: An Open, Secure, and Scalable RV64 SoC for Autonomous Nano-UAVs.....	1
<i>L. Valente, A. Veeran, M. Sinigaglia, Y. Tortorella, A. Nadalini, N. Wistoff, B. Sá, A. Garofalo, R. Psiakis, M. Tolba, A. Kulmala, N. Limaye, O. Sinanoglu, S. Pinto, D. Palossi, L. Benini, B. Mohammad, D. Rossi</i>	
Driving Compute Scale-Out Performance with Optical I/O Chiplets in Advanced System-In-Package Platforms	7
<i>Mark Wade, Chen Sun, Matt Sysak, Vladimir Stojanovic, Pooya Tadayon, Ravi Mahajan, Babak Sabi</i>	
HyperAccel Latency Processing Unit (LPUTM) Accelerating Hyperscale Models for Generative AI.....	8
<i>Seungjae Moon, Junsoo Kim, Jung-Hoon Kim, Junseo Cha, Gyubin Choi, Seongmin Hong, Joo-Young Kim</i>	
MLSoC™ - An Overview.....	9
<i>Srivi Dhruvanarayan, Victor Bittorf</i>	
An Open-Source 130-nm Fusion-Enabled Deconvolution Kernel Generator IC for Real-Time mmWave Radar Platform Motion Compensation.....	16
<i>Nikhil Poole, Priyanka Raina, Amin Arbabian</i>	
TrustForge: A Cryptographically Secure Enclave	17
<i>Todd Austin, Valeria Bertacco, Alex Kisil</i>	
PHEP: Paillier Homomorphic Encryption Processors for Privacy-Preserving Applications in Cloud Computing.....	18
<i>Guiming Shi, Yi Li, Xueqiang Wang, Zhanhong Tan, Dapeng Cao, Jingwei Cai, Yuchen Wei, Zehua Li, Wuke Zhang, Yifu Wu, Wei Xu, Kaisheng Ma</i>	
A Scalable Multi-Chiplet Deep Learning Accelerator with Hub-Side 2.5D Heterogeneous Integration	28
<i>Zhanhong Tan, Yifu Wu, Yannian Zhang, Haobing Shi, Wuke Zhang, Kaisheng Ma</i>	
A Heterogeneous SoC for Bluetooth LE in 28nm	37
<i>Felicia Guo, Nayiri Krzysztofowicz, Alex Moreno, Jeffrey Ni, Daniel Lovell, Yufeng Chi, Kareem Ahmad, Sherwin Afshar, Josh Alexander, Dylan Brater, Cheng Cao, Daniel Fan, Ryan Lund, Jackson Paddock, Griffin Prechter, Troy Sheldon, Shreesh Sreedhara, Anson Tsai, Eric Wu, Kerry Yu, Daniel Fritchman, Aviral Pandey, Ali Niknejad, Kristofer Pister, Borivoje Nikolic</i>	
A Heterogeneous RISC-V SoC for ML Applications in Intel 16 Technology	43
<i>Yufeng Chi, Franklin Huang, Raghav Gupta, Ella Schwarz, Jennifer Zhou, Reza Sajadiany, Animesh Agrawal, Max Banister, Michelle Boulos, Jason Chandran, Jessica Dowdall, Leena Elzeiny, Claire Gantan, Anthony Han, Roger Hsiao, Chadwick Leung, Edwin Lim, Jose Rodriguez, Tushar Sondhi, Mitchell Twu, Rongyi Wang, Mike Xiao, Ruohan Yan, Paul Kwon, Zhaokai Liu, Jerry Zhao, Bob Zhou, Ali Niknejad, Kristofer Pister, Borivoje Nikolic</i>	
Veyron V1 Data Center-Class RISC-V Processor	44
<i>N/A</i>	
Arm Neoverse V2 Platform: Leadership Performance and Power Efficiency for Next-Generation Cloud Computing, ML and HPC Workloads	52
<i>Magnus Bruce</i>	

AMD Next Generation “Zen 4” Core and 4th Gen AMD EPYC™ 9004 Server CPU	65
<i>Kai Troester, Ravi Bhargava</i>	
P870 High-Performance RISC-V Processor.....	78
<i>N/A</i>	
The Next Generation of High Performance, Energy-Efficient Computing: Intel® Xeon® Processors Built on Efficient-Core	88
<i>Don Soltis, Stephen Robinson</i>	
Memory-Centric Computing with SK Hynix's Domain-Specific Memory	96
<i>Yongkee Kwon, Guhyun 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, Chanwook Park, Yosub Song, Byeongsu Yang, Hyeongdeok Lee, Seungyeong Park, Wonjun Lee, Seongju Lee, Kyuyoung Kim, Daehan Kwon, Chunseok Jeong, John Kim, Euicheol Lim, Junhyun Chun</i>	
Samsung PIM/PNM for Transfmer Based AI : Energy Efficiency on PIM/PNM Cluster	109
<i>Jin Hyun Kim, Yuhwan Ro, Jinin So, Sukhan Lee, Shin-Haeng Kang, Yeongon Cho, Hyeonsu Kim, Byeongho Kim, Kyungsoo Kim, Sangsoo Park, Jin-Seong Kim, Sanghoon Cha, Won-Jo Lee, Jin Jung, Jong-Geon Lee, Jieun Lee, Joonho Song, Seungwon Lee, Jeonghyeon Cho, Jaehoon Yu, Kyomin Sohn</i>	
Architecting for Flexibility and Value with Next Gen Intel® Xeon® Processors	125
<i>Chris Gianos</i>	
Caliptra	133
<i>N/A</i>	
Intel® Energy Efficiency Architecture	138
<i>Efraim Rotem</i>	
CSS N2: Arm Neoverse N2 Platform, Delivered to Partners as a Fully Verified, Customizable Subsystem.....	147
<i>Anitha Kona</i>	
The First Direct Mesh-To-Mesh Photonic Fabric	159
<i>Jason Howard</i>	
Hardware for Deep Learning	168
<i>Bill Dally</i>	
Supercharged AI Inference on Modern CPUs	197
<i>Lawrence Spracklen, Subutai Ahmad</i>	
Qualocmm® Hexagon™ NPU	208
<i>Eric Mahurin</i>	
Moffett Antoum®: A Deep-Sparse AI Inference System-On-Chip for Vision and Large-Language Models.....	218
<i>Zhibin Xiao</i>	
Inside the Cerebras Wafer-Scale Cluster: Cerebras Systems	235
<i>Sean Lie</i>	
Intel Agilex® 9 Direct RF-Series FPGAs with Integrated 64 Gbps Data Converters.....	256
<i>Ben Esposito</i>	

FABRIC8LABS: Electrochemical Additive Manufacturing (ECAM) for Cooling High Performance ICs	274
<i>Ian Winfield, Joseph Madril, Tim Ouradnik, Michael Matthews, Guillermo Romero</i>	
AMD Next-Generation FPGA Built from Chiplets	283
<i>Dinesh Gaitonde</i>	
Hummingbird™ Low-Latency Computing Engine	297
<i>Maurice Steinman</i>	
IBM NorthPole Neural Inference Machine	307
<i>Dharmendra S. Modha, Filipp Akopyan, Alexander Andreopoulos, Rathinakumar Appuswamy, John V. Arthur, Andrew S. Cassidy, Pallab Datta, Michael V. Debole, Steven K. Esser, Carlos Ortega Otero, Jun Sawada, Brian Taba, Arnon Amir, Deepika Bablani, Peter J. Carlson, Myron D. Flickner, Rajamohan Gandhasri, Guillaume J. Garreau, Megumi Ito, Jennifer L. Klamo, Jeffrey A. Kusnitz, Nathaniel J. McClatchey, Jeffrey L. McKinstry, Yutaka Nakamura, Tapan K. Nayak, William P. Risk, Kai Schleupen, Ben Shaw, Jay Sivagnaname, Daniel F. Smith, Ignacio Terrizzano, Takanori Ueda</i>	
NVIDIA's Resource Transmutable Network Processing ASIC	336
<i>Kevin Deierling</i>	
AMD Ryzen™ 7040 Series: Technology Overview	343
<i>Mahesh Subramony, David Kramer, Indrani Paul</i>	
A Machine Learning Supercomputer with an Optically Reconfigurable Interconnect and Embeddings Support	357
<i>Norman P. Jouppi, Andy Swing</i>	
Exciting Directions for ML Models and the Implications for Computing Hardware	369
<i>Jeff Dean, Amin Vahdat</i>	

Author Index