

# **2022 IEEE Hot Chips 34 Symposium (HCS 2022)**

**Cupertino, California, USA  
21 – 23 August 2022**



**IEEE Catalog Number: CFP22HCS-POD  
ISBN: 978-1-6654-6029-3**

**Copyright © 2022 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:	CFP22HCS-POD
ISBN (Print-On-Demand):	978-1-6654-6029-3
ISBN (Online):	978-1-6654-6028-6
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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

CURRAN ASSOCIATES INC.  
**proceedings**  
.com

# TABLE OF CONTENTS

AMD Ryzen™ 6000 Series for Mobile: Technology Overview .....	1
<i>Jim Gibney</i>	
AMD 400G Adaptive SmartNIC SoC: Technology Preview .....	13
<i>Jaideep Dastidar, David Riddoch, Jason Moore, Steve Pope, Jim Wesselkamper</i>	
AMD Instinct™ MI200 Series Accelerator and Node Architectures .....	29
<i>Alan Smith, Norman James</i>	
Arm Morello Evaluation Platform -Validating CHERI-Based Security in a High-Performance System .....	41
<i>Richard Grisenthwaite</i>	
壁仞™ BR100 GPGPU: Accelerating Datacenter Scale AI Computing.....	52
<i>Mike Hong, Lingjie Xu</i>	
Cerebras Architecture Deep Dive: First Look Inside the HW/SW Co-Design for Deep Learning: Cerebras Systems .....	63
<i>Sean Lie</i>	
Kraken: A Direct Event/Frame-Based Multi-Sensor Fusion SoC for Ultra-Efficient Visual Processing in Nano-UAVs .....	80
<i>Alfio Di Mauro, Moritz Scherer, Davide Rossi, Luca Benini</i>	
Intel's Ponte Vecchio GPU: Architecture, Systems & Software.....	90
<i>Hong Jiang</i>	
HNPU-V2: A 46.6 FPS DNN Training Processor for Real-World Environmental Adaptation Based Robust Object Detection on Mobile Devices .....	105
<i>Donghyeon Han, Dongseok Im, Gwangtae Park, Youngwoo Kim, Seokchan Song, Juhyoung Lee, Hoi-Jun Yoo</i>	
DSPU: A 281.6mW Real-Time Deep Learning-Based Dense RGB-D Data Acquisition with Sensor Fusion and 3D Perception System-on-Chip .....	114
<i>Dongseok Im, Gwangtae Park, Zhiyong Li, Junha Ryu, Sanghoon Kang, Donghyeon Han, Jinsu Lee, Wonhoon Park, Hankyul Kwon, Hoi-Jun Yoo</i>	
Trinity: End-to-End In-Database Near-Data Machine Learning Acceleration Platform for Advanced Data Analytics .....	127
<i>Ji-Hoon Kim, Seunghee Han, Kwanghyun Park, Soo-Young Ji, Joo-Young Kim</i>	
Large-Scale Graph Neural Network Services Through Computational SSD and In-Storage Processing Architectures .....	135
<i>Miryeong Kwon, Donghyun Gouk, Sangwon Lee, Myoungsoo Jung</i>	
Neuro-CIM: A 310.4 TOPS/W Neuromorphic Computing-in-Memory Processor with Low WL/BL Activity and Digital-Analog Mixed-Mode Neuron Firing.....	148
<i>Sangyeob Kim, Sangjin Kim, Soyeon Um, Soyeon Kim, Kwantae Kim, Hoi-Jun Yoo</i>	
DFX: A Low-Latency Multi-FPGA Appliance for Accelerating Transformer-Based Text Generation .....	161
<i>Seongmin Hong, Seungjae Moon, Junsoo Kim, Sungjae Lee, Minsub Kim, Dongsoo Lee, Joo-Young Kim</i>	

An Efficient High-Quality FHD Super-Resolution Mobile Accelerator SoC with Hybrid-Precision and Energy-Efficient Cache.....	170
<i>Zhiyong Li, Sangjin Kim, Dongseok Im, Donghyeon Han, Hoi-Jun Yoo</i>	
Heterogenous Integration Enables FPGA Based Hardware Acceleration for RF Applications .....	183
<i>Sergey Shumarayev, Allen Chan, Tim Hoang, Robert Keller</i>	
Dimensity 9000 – A Flagship Smartphone SoC.....	193
<i>E. Wang, Stefan Rosinger, Saurabh Pradhan</i>	
NODAR 3D Vision System: Enabling Mass Production of Autonomous Vehicles.....	205
<i>N/A</i>	
VTA-NIC: Deep Learning Inference Serving in Network Interface Cards .....	215
<i>Kenji Tanaka, Yuki Arikawa, Kazutaka Morita, Tsuyoshi Ito, Takashi Uchida, Natsuko Saito, Shinya Kaji, Takeshi Sakamoto</i>	
Nvidia Hopper GPU: Scaling Performance.....	223
<i>Jack Choquette</i>	
Nvidia Grace.....	246
<i>Jonathon Evans</i>	
NVIDIA ORIN System-on-Chip.....	256
<i>Michael Ditty</i>	
From High-Level Frameworks to Custom Silicon with SODA.....	265
<i>Serena Curzel, Nicolas Bohm Agostini, Reece Neff, Ankur Limaye, Jeff Zhang, Vinay Amatya, Marco Minutoli, Vito Giovanni Castellana, Joseph Manzano, David Brooks, Gu-Yeon Wei, Fabrizio Ferrandi, Antonino Tumeo</i>	
Enabling Scalable Application-Specific Optical Engines (ASOE) by Monolithic Integration of Photonics and Electronics.....	272
<i>Christoph Schullien</i>	
LightTrader: World’s First AI-Enabled High-Frequency Trading Solution with 16 TFLOPS / 64 TOPS Deep Learning Inference Accelerators .....	288
<i>Hyunsung Kim, Sungyeob Yoo, Jaewan Bae, Kyeongryeol Bong, Yoonho Boo, Karim Charfi, Hyo-Eun Kim, Hyun Suk Kim, Jinseok Kim, Byungjae Lee, Jaehwan Lee, Myeongbo Shim, Sungho Shin, Jeong Seok Woo, Joo-Young Kim, Sunghyun Park, Jinwook Oh</i>	
Scaling of Memory Performance and Capacity with CXL Memory Expander.....	293
<i>S. J. Park, H. Kim, K.-S. Kim, J. So, J. Ahn, W.-J. Lee, D. Kim, Y.-J. Kim, J. Seok, J.-G. Lee, H.-Y. Ryu, C. Y. Lee, J. Prout, K.-C. Ryoo, S.-J. Han, M.-K. Kook, J. S. Choi, J. Gim, Y. S. Ki, S. Ryu, C. Park, D.-G. Lee, J. Cho, H. Song, J. Y. Lee</i>	
System Architecture and Software Stack for GDDR6-AiM.....	307
<i>Yongkee Kwon, Kornijcuk Vladimir, Nahsung Kim, Woojae Shin, Jongsoon Won, Minkyu Lee, Hyunha Joo, Haerang Choi, Guhyun Kim, Byeongju An, Jeongbin Kim, Jaewook Lee, Ilkon Kim, Jaehan Park, Chanwook Park, Yosub Song, Byeongsu Yang, Hyungdeok Lee, Seho Kim, Daehan Kwon, Seongju Lee, Kyuyoung Kim, Sanghoon Oh, Joonhong Park, Gimoon Hong, Dongyoon Ka, Kyudong Hwang, Jeongje Park, Kyeongpil Kang, Jungyeon Kim, Junyeol Jeon, Myeongjun Lee, Minyoung Shin, Minhwan Shin, Jaekyung Cha, Changson Jung, Kijoon Chang, Chunseok Jeong, Euicheol Lim, Il Park, Junhyun Chun, Sk Hynix</i>	

Amber: Coarse-Grained Reconfigurable Array-Based SoC for Dense Linear Algebra Acceleration .....	320
<i>Kathleen Feng, Alex Carsello, Taeyoung Kong, Kalhan Koul, Qiaoyi Liu, Jackson Melchert, Gedeon Nyengele, Maxwell Strange, Keyi Zhang, Ankita Nayak, Jeff Setter, James Thomas, Kavya Sreedhar, Po-Han Chen, Nikhil Bhagdikar, Zachary Myers, Brandon D'Agostino, Pranil Joshi, Stephen Richardson, Rick Bahr, Christopher Torng, Mark Horowitz, Priyanka Raina</i>	
Vision Perception Unit: Next-Generation Smart CMOS Image Sensor .....	335
<i>Wenqi Ji, Yuxing Han, Jiangtao Wen, Yubin Hu, Futang Wang, Yuze He, Xi Li, Jun Zhang</i>	
NOEMA: A Massive-Scale Brain Activity Decoding Chip .....	342
<i>Ameer Abdelhadi, Eugene Sha, Andreas Moshovos</i>	
A 7-nm FinFET 1.2-TB/s/mm <sup>2</sup> 3D-Stacked SRAM with an Inductive Coupling Interface Using Over-SRAM Coils and Manchester-Encoded Synchronous Transceivers.....	375
<i>Kota Shiba, Mitsuji Okada, Atsutake Kosuge, Mototsugu Hamada, Tadahiro Kuroda</i>	
A 13.7μJ/prediction 88% Accuracy CIFAR-10 Single-Chip Wired-Logic Processor in 16-nm FPGA Using Non-Linear Neural Network .....	382
<i>Yao-Chung Hsu, Atsutake Kosuge, Rei Sumikawa, Kota Shiba, Mototsugu Hamada, Tadahiro Kuroda</i>	
Accelerating Graphic Rendering on Programmable RISC-V GPUs .....	389
<i>Blaise Tine, Varun Saxena, Santosh Srivatsan, Joshua R. Simpson, Fadi Alzammar, Liam Paul Cooper, Sam Jijina, Swetha Rajagoplan, Tejaswini Anand Kumar, Jeff Young, Hyesoon Kim</i>	
Built for the Edge: The Next-Generation Intel® Xeon D 2700 & 1700 Processors.....	397
<i>Praveen Mosur</i>	
HALO: A Flexible and Low Power Processing Fabric for Brain-Computer Interfaces .....	408
<i>Abhishek Bhattacharjee, Rajit Manohar</i>	
Lightmatter .....	427
<i>N/A</i>	
The Groq Software-Defined Scale-Out Tensor Streaming Multiprocessor: From Chips-to-Systems Architectural Overview .....	440
<i>Dennis Abts, John Kim, Garrin Kimmell, Matthew Boyd, Kris Kang, Sahil Parmar, Andrew Ling, Andrew Bitar, Ibrahim Ahmed, Jonathan Ross</i>	
Beyond Compute: Enabling AI Through System Integration.....	475
<i>N/A</i>	
Untether AI: Boqueria .....	519
<i>Robert Beachler, Martin Snelgrove</i>	
Semiconductors Run the World.....	529
<i>Pat Gelsinger</i>	
DOJO: The Microarchitecture of Tesla's Exa-Scale Computer .....	539
<i>Emil Talpes, Douglas Williams, Debjit Das Sarma</i>	
DOJO: Super-Compute System Scaling for ML Training.....	553
<i>Bill Chang, Rajiv Kurian, Doug Williams, Eric Quinnell</i>	

Meteor Lake and Arrow Lake: Intel Next-Gen 3D Client Architecture Platform with Foveros ..... 576  
*Wilfred Gomes, Slade Morgan, Boyd Phelps, Tim Wilson, Erik Hallnor*

The NVLink-Network Switch: nVIDIA’s Switch Chip for High Communication-Bandwidth  
Superpods ..... 596  
*Alexander Ishii, Ryan Wells*

Juniper’s Express 5: A 28.8Tbps Network Routing ASIC and Variations..... 608  
*Chang-Hong Wu*

**Author Index**