

2021 IEEE Hot Chips 33 Symposium (HCS 2021)

**Palo Alto, California, USA
22 – 24 August 2021**



**IEEE Catalog Number: CFP21HCS-POD
ISBN: 978-1-6654-1398-5**

**Copyright © 2021 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:	CFP21HCS-POD
ISBN (Print-On-Demand):	978-1-6654-1398-5
ISBN (Online):	978-1-6654-1397-8
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

AQUABOLT-XL: SAMSUNG HBM2-PIM WITH IN-MEMORY PROCESSING FOR ML ACCELERATORS AND BEYOND	1
<i>Jin Hyun Kim; Shin-Haeng Kang; Sukhan Lee; Hyeonsu Kim; Woongjae Song; Yuhwan Ro; Seungwon Lee; David Wang; Hyunsung Shin; Bengseng Phuah; Jihyun Choi; Jinin So; Yeongon Cho; Joonho Song; Jangseok Choi; Jeonghyeon Cho; Kyomin Sohn; Youngsoo Sohn; Kwangil Park; Nam Sung Kim</i>	
ARM NEOVERSE N2: ARM'S 2ND GENERATION HIGH PERFORMANCE INFRASTRUCTURE CPUS AND SYSTEM IPS	14
<i>Andrea Pellegrini</i>	
QUALCOMM® CLOUD AL 100 : 12TOPS/W SCALABLE, HIGH PERFORMANCE AND LOW LATENCY DEEP LEARNING INFERENCE ACCELERATOR	28
<i>Karam Chatha</i>	
AI COMPUTE CHIP FROM ENFLAME	38
<i>Ryan Liu; Chuang Feng</i>	
WORLD LARGEST MOBILE IMAGE SENSOR WITH ALL DIRECTIONAL PHASE DETECTION AUTO FOCUS FUNCTION	52
<i>Sukki Yoon; Jungbin Yun; Yunhwan Jung; Ilyun Jeong; Junghee Choi; Wooseok Choi; Jeongguk Lee; Hansoo Lee; Juhyun Ko</i>	
ARCHITECTURAL CHALLENGES: AI CHIPS, DECISION SUPPORT AND HIGH PERFORMANCE COMPUTING	63
<i>Dimitri Kusnezov</i>	
NEW VALUE CREATION BY NANO-TACTILE SENSOR CHIP EXCEEDING OUR FINGERTIP DISCRIMINATION ABILITY	74
<i>Hidekuni Takao; Kazuki Watatani; Kazutami Arimoto</i>	
SOT-MRAM – THIRD GENERATION MRAM MEMORY OPENS NEW OPPORTUNITIES : HOT CHIPS CONFERENCE AUGUST 2021	92
<i>Barry Hoberman; Jean-Pierre Nozieres</i>	
ALDER LAKE ARCHITECTURE	97
<i>Efraim Rotem; Yuli Mandelblat; Vadim Basin; Eli Weissmann; Arik Gihon; Rajshree Chabukswar; Russ Fenger; Monica Gupta</i>	
NEXT GENERATION “ZEN 3” CORE	109
<i>Mark Evers; Leslie Barnes; Mike Clark</i>	
REAL-TIME AI FOR ENTERPRISE WORKLOADS: THE IBM TELUM PROCESSOR	125
<i>Christian Jacobi</i>	
SAPPHIRE RAPIDS	136
<i>Arijit Biswas</i>	
MULTI-MILLION CORE, MULTI-WAFER AI CLUSTER	147
<i>Sean Lie</i>	
A PHOTONIC NEURAL NETWORK USING < 1 PHOTON PER SCALAR MULTIPLICATION	168
<i>Tianyu Wang; Shi-Yuan Ma; Logan G. Wright; Tatsuhiko Onodera; Brian C. Richard; Peter L. McMahon</i>	
THE ANTON 3 ASIC: A FIRE-BREATHING MONSTER FOR MOLECULAR DYNAMICS SIMULATIONS	176
<i>Peter J. Adams; Brannon Batson; Alistair Bell; Jhanvi Bhatt; J. Adam Butts; Timothy Correia; Bruce Edwards; Peter Feldmann; Christopher H. Fenton; Anthony Forte; Joseph Gagliardo; Gennette Gill; Maria Gorlatova; Brian Greskamp; J. P. Grossman; Jeremy Hunt; Bryan L. Jackson; Mollie M. Kirk; Jeffrey S. Kuskin; Roy J. Mader; Richard McGowen; Adam McLaughlin; Mark A. Moraes; Mohamed Nasr; Lawrence J. Nociolo; Lief O'Donnell; Andrew Parker; Jon L. Peticolas; Terry Quan; T. Carl Schwink; Keun Sup Shim; Naseer Siddique; Jochen Spengler; Michael Theobald; Brian Towles; William Vick; Stanley C. Wang; Michael Wazlowski; Madeleine J. Weingarten; John M. Williams; David E. Shaw</i>	
DYNAMIC NEURAL ACCELERATOR FOR RECONFIGURABLE & ENERGY-EFFICIENT NEURAL NETWORK INFERENCE	187
<i>Nikolay Nez; Antonio N. Vilchez; Hamid R. Zohouri; Oleg Khavin; Sakyasingha Dasgupta</i>	
EDGE Q: 5G WITH AN EDGE	198
<i>Sriram Rajagopal</i>	
ENIAD: A RECONFIGURABLE NEAR-DATA PROCESSING ARCHITECTURE FOR WEB-SCALE AI-ENRICHED BIG DATA SERVICE	205
<i>Jialiang Zhang; Jingjane Li</i>	

ACCELERATING ML RECOMMENDATION WITH OVER A THOUSAND RISC-V/TENSOR PROCESSORS ON ESPERANTO'S ET-SOC-1 CHIP	209
<i>Dave Ditzel; Roger Espasa; Nivard Aymerich; Allen Baum; Tom Berg; Jim Burr; Eric Hao; Jayesh Iyer; Miquel Izquierdo; Shankar Jayaratnam; Darren Jones; Chris Klingner; Jin Kim; Stephen Lee; Marc Lupon; Grigorios Magklis; Bojan Maric; Rajib Nath; Mike Neilly; Duane Northcutt; Bill Orner; Jose Renau; Gerard Reves; Xavier Reves; Tom Riordan; Pedro Sanchez; Sri Samudrala; Guillem Sole; Raymond Tang; Tommy Thorn; Francisco Torres; Sebastia Tortella; Daniel Yau</i>	
VIDEO CODING UNIT (VCU).....	221
<i>Aki Kuusela; Clint Smullen</i>	
GRAPHCORE COLOSSUS MK2 IPU	236
<i>Simon Knowles</i>	
SM6: A 16NM SYSTEM-ON-CHIP FOR ACCURATE AND NOISE-ROBUST ATTENTION-BASED NLP APPLICATIONS	249
<i>Thierry Tambe; En-Yu Yang; Glenn G. Ko; Yuji Chai; Coleman Hooper; Marco Donato; Paul N. Whatmough; Alexander M. Rush; David Brooks; Gu-Yeon Wei</i>	
XEHPC: PONTE VECCHIO	256
<i>David Blythe</i>	
AN ENERGY-EFFICIENT FLOATING-POINT DNN PROCESSOR USING HETEROGENEOUS COMPUTING ARCHITECTURE WITH EXPONENT-COMPUTING-IN-MEMORY	273
<i>Juhyoung Lee; Jihoon Kim; Wooyoung Jo; Sangyeob Kim; Sangjin Kim; Donghyeon Han; Jinsu Lee; Hoi-Jun Yoo</i>	
OMNIDRL: AN ENERGY-EFFICIENT MOBILE DEEP REINFORCEMENT LEARNING ACCELERATORS WITH DUAL-MODE WEIGHT COMPRESSION AND DIRECT PROCESSING OF COMPRESSED DATA	283
<i>Juhyoung Lee; Sangyeob Kim; Ji-Hoon Kim; Sangjin Kim; Wooyoung Jo; Donghyeon Han; Hoi-Jun Yoo</i>	
PNNPU: A FAST AND EFFICIENT 3D POINT CLOUD-BASED NEURAL NETWORK PROCESSOR WITH BLOCK-BASED POINT PROCESSING FOR REGULAR DRAM ACCESS.....	294
<i>Sangjin Kim; Juhyoung Lee; Dongseok Im; Hoi-Jun Yoo</i>	
PHOTONIC CO-PROCESSORS IN HPC: USING LIGHTON OPUS FOR RANDOMIZED NUMERICAL LINEAR ALGEBRA	306
<i>Daniel Hesslow; Alessandro Cappelli; Igor Carron; Laurent Daudet; Raphaël Lafargue; Kilian Müller; Ruben Ohana; Gustave Pariente; Iacopo Poli</i>	
NVIDIA DATA CENTER PROCESSING UNIT (DPU) ARCHITECTURE	311
<i>Idan Burstein</i>	
RDNA™ 2 GAMING ARCHITECTURE	321
<i>Andrew Pomianowski</i>	
EXYNOS 1080: HIGH-PERFORMANCE, LOW-POWER CPU AND GPU WITH AMIGO.....	330
<i>Taehee Lee; Dongkeun Kim; Joonseok Kim</i>	
SAMSUNG NEURAL PROCESSING UNIT: AN AI ACCELERATOR AND SDK FOR FLAGSHIP MOBILE AP.....	337
<i>Jun-Seok Park; Heonsoo Lee; Dongwoo Lee; Jewoo Moon; Suknam Kwon; Sanghyuck Ha; Minseong Kim; Junghun Park; Jihoon Bang; Sukhwan Lim Inyup Kang</i>	
ELPIS: HIGH PERFORMANCE LOW POWER CONTROLLER FOR DATA CENTER SSDS	348
<i>Seungwon Lee; Juhyung Hong; Chulseung Lee; Sandeep Vallabhaneni; Soon-Gon Kim; Yong Jin Shin; Mijung Noh; Soon-Jae Won; Soonbok Jang; Yong Ho Song</i>	
INDUSTRY'S FIRST 7.2 GBPS 512GB DDR5 MODULE	352
<i>Sung Joo Park; Jonghoon J. Kim; Kun Joo; Young-Ho Lee; Kyoungsun Kim; Young-Tae Kim; Woo-Jin Na; Ikjoon Choi; Hye-Seung Yu; Wonyoung Kim; Juyeon Jung; Jaejun Lee; Dohyung Kim; Young-Uk Chang; Gong Heum Han; Hangi Jung; Sunwon Kang; Jeonghyeon Cho; Hoyoung Song; Tae-Young Oh; Young-Soo Sohn; Sangjoon Hwang; Jooyoung Lee</i>	
A PLUG-AND-PLAY UNIVERSAL PHOTONIC PROCESSOR FOR QUANTUM INFORMATION PROCESSING	358
<i>Taballione</i>	
EDGE INFERENCE ENGINE FOR DEEP & RANDOM SPARSE NEURAL NETWORKS WITH 4-BIT CARTESIAN-PRODUCT MAC ARRAY AND PIPELINED ACTIVATION ALIGNER	360
<i>Kota Ando; Jaehoon Yu; Kazutoshi Hirose; Hiroki Nakahara; Kazushi Kawamura; Thiem Van Chu; Masato Motomura</i>	
A CORDIC-BASED TRIGONOMETRIC HARDWARE ACCELERATOR WITH CUSTOM INSTRUCTION IN 32-BIT RISC-V SYSTEM-ON-CHIP	371
<i>Khai-Duy Nguyen; Dang Tuan Kiet; Trong-Thuc Hoang; Nguyen Quang Nhu Quynh; Cong-Kha Pham</i>	
SYSTEM-ON-CHIP IMPLEMENTATION OF TRUSTED EXECUTION ENVIRONMENT WITH HETEROGENEOUS ARCHITECTURE.....	378
<i>Trong-Thuc Hoang; Cristian Duran; Ronaldo Serrano; Marco Sarmiento; Khai-Duy Nguyen; Akira Tsukamoto; Kuniyasu Suzuki; Cong-Kha Pham</i>	

MOZART: DESIGNING FOR SOFTWARE MATURITY AND THE NEXT PARADIGM FOR CHIP ARCHITECTURES.....	386
<i>Karu Sankaralingam; Tony Nowatzki; Greg Wright; Poly Palamuttam; Jitu Khare; Vinay Gangadhar; Preyas Shah</i>	
HETEROGENEOUS COMPUTING TO ENABLE HIGHEST LEVEL OF SAFETY	396
<i>Ramanujan Venkatadri</i>	
LIGHTON OPTICAL PROCESSING UNIT: SCALING-UP AI AND HPC WITH A NON VON NEUMANN CO-PROCESSOR.....	416
<i>Charles Brossollet; Alessandro Cappelli; Igor Carron; Charidimos Chaintoutis; Amélie Chatelain; Laurent Daudet; Sylvain Gigan; Daniel Hesslow; Florent Krzakala; Julien Launay; Safa Mokaadi; Fabien Moreau; Kilian Muller; Ruben Ohana; Gustave Pariente; Iacopo Poli; Giuseppe L. Tommasone</i>	
MORPHEUS II: A RISC-V SECURITY EXTENSION FOR PROTECTING VULNERABLE SOFTWARE AND HARDWARE	422
<i>Todd Austin; Austin Harris; Tarunesh Verma; Shijia Wei; Alex Kisil; Misiker Aga; Valeria Bertacco; Baris Kasikci; Mohit Tiwari</i>	
INTEL’S HYPERSCALE-READY INFRASTRUCTURE PROCESSING UNIT (IPU)	431
<i>Brad Burres; Dan Daly; Mark Debbage; Eliel Louzoun; Christine Severns-Williams; Naru Sundar; Nadav Turbovich; Barry Wolford; Yadong Li</i>	
QUANTUM COMPUTING WITH ATOMS.....	439
<i>Christopher Monroe</i>	
MOJO LENS - AR CONTACT LENSES FOR REAL PEOPLE	452
<i>Mike Wiemer; Renaldi Winoto</i>	
SAMBA NOVA SN10 RDU:ACCELERATING SOFTWARE 2.0 WITH DATAFLOW	480
<i>Raghu Prabhakar; Sumti Jairath</i>	
SKYDIO AUTONOMY ENGINE: ENABLING THE NEXT GENERATION OF AUTONOMOUS FLIGHT	499
<i>Abraham Bachrach</i>	
XILINX EDGE PROCESSORS	521
<i>Ross Freeman; James V. Barnett; Bernard V. Vonderschmitt</i>	
Author Index	