

2025 34th International Conference on Parallel Architectures and Compilation Techniques (PACT 2025)

**Irvine, California, USA
3-6 November 2025**

IEEE Catalog Number: CFP25073-POD
ISBN: 979-8-3315-8296-8



Copyright © 2025, IEEE

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:	CFP25073-POD
ISBN (Print-On-Demand):	979-8-3315-8296-8
ISBN (Online):	979-8-3315-8295-1

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

TABLE OF CONTENTS

Scalable Processing-Near-Memory for 1M-Token LLM Inference: CXL-Enabled KV-Cache Management Beyond GPU Limits	1
<i>Dowon Kim, MinJae Lee, Janghyeon Kim, HyuckSung Kwon, Hyeonggyu Jeong, Sang-Soo Park</i> <i>Minyong Yoon, Si-Dong Roh, Yongsuk Kwon, Jinin So, Jungwook Choi</i>	
SPipe: Hybrid GPU and CPU Pipeline for Training LLMs under Memory Pressure	14
<i>Junyeol Ryu, Yujin Jeong, Daeyoung Park, Jinpyo Kim, Heehoon Kim, Jaejin Lee</i>	
ScaleMoE: A Fast and Scalable Distributed Training Framework for Large-Scale Mixture-of-Experts Models	30
<i>Seohong Choi, Huize Hong, Tae Hee Han, Joonsung Kim</i>	
LibraPIM: Dynamic Load Rebalancing to Maximize Utilization in PIM-Assisted LLM Inference Systems	43
<i>Hyeongjun Cho, Yoonho Jang, Hyungi Kim, Seongwook Kim, Keewon Kwon, Gwangsun Kim, Seokin Hong</i>	
Doppeladler: Adaptive Tensor Parallelism for Latency-Critical LLM Deployment on CPU-GPU Integrated End-User Device	57
<i>Jiazhi Jiang, Xiao Liu, Jiangsu Du, Dan Huang, Yutong Lu</i>	
Exploring Memory Tiering Systems in the CXL Era via FPGA-based Emulation and Device-Side Management	71
<i>Yiqi Chen, Xiping Dong, Zhe Zhou, Zhao Wang, Jie Zhang, Guangyu Sun</i>	
CPC: Coordinated Page Cache for Serverless Computing	84
<i>Keun Soo Lim, Yunjay Hong, Jongheon Jeong, Sam Son, Donguk Kim, Yeonhong Park, Jae W. Lee</i> <i>Jinkyu Jeong</i>	
SCREME: A Scalable Framework for Resilient Memory Design	97
<i>Fan Li, Mimi Xie, Yanan Guo, Huize Li, Xin Xin</i>	
Cache Miss Curve Analysis via Cardinality Domain	110
<i>Eishi Arima, Martin Schulz</i>	
EARTH: Efficient Architecture for RISC-V Vector Memory Access	122
<i>Hongyi Guan, Yichuan Gao, Chenlu Miao, Haoyang Wu, Hang Zhu, Mingfeng Lin, Huayue Liang</i>	
ANG: Accelerating NFA processing on GPUs via Exploring Multi-Level Fine-Grained Parallelism	135
<i>Yuguang Wang, Yunmo Zhang, Zeyu Liu, Junqiao Qiu, Zhenlin Wang</i>	
Accelerating DFS-based Subgraph Matching on GPU via Reusing Intersection	148
<i>Chen Chen, Shanzhi Gu, Junsheng Chang, Li Shen</i>	
Multiway Merge Partitioning for Sparse-Sparse Matrix Multiplication on GPUs	160
<i>Eric Lorimer, Ruobing Han, Sung Ha Kang, Hyesoon Kim</i>	
DMO-DB: Mitigating the Data Movement Bottlenecks of GPU-Accelerated Relational OLAP	172
<i>Chaemin Lim, Suhyun Lee, Jinwoo Choi, Joonsung Kim, Jinho Lee, Youngsok Kim</i>	
Agentic Auto-Scheduling: An Experimental Study of LLM-Guided Loop Optimization	186
<i>Massinissa Merouani, Islem Kara Bernou, Riyadh Baghdadi</i>	

LOOPer: A Learned Automatic Code Optimizer For Polyhedral Compilers	201
<i>Massinissa Merouani, Afif Boudaoud, Iheb Nassim Aouadj, Nassim Tchoulak, Islem Kara Bernou Hamza Benyamina, Fatima Benbouzid-Si Tayeb, Karima Benatchba, Hugh Leather, Riyadh Baghdadi</i>	
Guess, Measure & Edit: Using Lowering to Lift Tensor Code	216
<i>José Wesley de Souza Magalhães, Jackson Woodruff, Jordi Armengol-Estapé, Alexander Brauckmann Luc Jaulmes, Elizabeth Polgreen, Michael F.P. O'Boyle</i>	
Automatic Generation of Actor-based Parallelism from Shared-Memory Parallel Programs	229
<i>Jun Shirako, Vivek Sarkar</i>	
Automatic Code-Generation for Accelerating Structured-Mesh-Based Explicit Numerical Solvers on FPGAs	243
<i>Beniel Thileepan, Suhaib A. Fahmy, Gihan R. Mudalige</i>	
FLASH: An Abstract Machine for Modelling Fully Homomorphic Encryption Accelerators	257
<i>Alireza Tabatabaeian, Arrvindh Shriraman</i>	
Energy-Efficient Acceleration of Hash-Based Post-Quantum Cryptographic Schemes on Embedded Spatial Architectures	270
<i>Yanze Wu, Md Tanvir Arafin</i>	
Fine-Grained Fusion: The Missing Piece in Area-Efficient State Space Model Acceleration	281
<i>Robin Geens, Arne Symons, Marian Verhelst</i>	
Squire: A General-Purpose Accelerator to Exploit Fine-Grain Parallelism on Dependency-Bound Kernels	292
<i>Rubén Langarita, Jesús Alastruey-Benedé, Pablo Ibáñez-Marín, Santiago Marco-Sola, Miquel Moretó Adrià Armejach</i>	
Bancroft: Genomics Acceleration Beyond On-Device Memory	306
<i>Se-Min Lim, Seongyoung Kang, Sang-Woo Jun</i>	
Hera: A Heterogeneity-Aware Multi-Tenant Inference Server for Personalized Recommendations	320
<i>Yujeong Choi, John Kim, Minsoo Rhu</i>	
Salient Store: Enabling Smart Storage for Continuous Learning Edge Servers	333
<i>Cyan Subhra Mishra, Deeksha Chaudhary, Mahmut Taylan Kandemir, Chita R. Das</i>	
Bit-Level Semantics: Scalable RAG Retrieval with Neurosymbolic Hyperdimensional Computing	347
<i>Hyunsei Lee, Shinhyoung Jang, Jaewoo Gwak, Jongho Park, Yeseong Kim</i>	
Optimizing 3D Gaussian Splattering for Mobile GPUs	359
<i>Md Musfiqur Rahman Sanim, Zhihao Shu, Bahram Afsharmanesh, AmirAli Mirian, Jiexiong Guan, Wei Niu Bin Ren, Gagan Agrawal</i>	
GPU Stream-Aware Communication for Effective Pipelining	372
<i>Naveen Namashivayam, Krishna Kandalla, Pen-Chung Yew, Trey White, Larry Kaplan, Mark Pagel</i>	
TPE: XPU-Point: Simulator-Agnostic Sample Selection Methodology for Heterogeneous CPU-GPU Applications	385
<i>Alen Sabu, Harish Patil, Wim Heirman, Changxi Liu, Trevor E. Carlson</i>	
Generating Two-Level, GPU-Aware Mappings for Distributed Tensor Computations	401
<i>Botao Wu, Martin Kong</i>	

A Stable Marriage Requires a Shared Residence with Low Contention and Mutual Complementarity	416
<i>Jiaxin Liu, Rubao Lee, Cathy H. Xia, Xia Odong Zhang</i>	
Optimize Winograd Convolution for a Novel MIMD Many-core Architecture PEZY-SC3s	431
<i>Yi Zhou, Qinglin Wang, Lian Wang, Zhiyan Liu, Bingwei Wang, Feiming Liu, Xiangdong Pei, Jie Liu</i>	
CoroAMU: Unleashing Memory-Driven Coroutines through Latency-Aware Decoupled Operations	444
<i>Zhuolun Jiang, Songyue Wang, Xiaokun Pei, Tianyue, Mingyu Chen</i>	
POSTER: DaPPA: A Data-Parallel Programming Framework for Processing-in-Memory Architectures	458
<i>Geraldo F. Oliveira, Alain Kohli, David Novo, Ataberk Olgun, A.Giray Yağlıkçı, Saugata Ghose</i> <i>Juan Gómez-Luna, Onur Mutlu</i>	
Poster: HeteroSched: Co-Optimizing Scheduling and Parallelization for Deep Learning Workloads	461
<i>Bahram Afsharmanesh, Md Musfiqur Rahman Sanim, AmirAli Mirian, Gagan Agrawal</i>	
POSTER: IRISX: A Dynamic Trade-off System for Performance Portability on Multi-Accelerator Platforms	464
<i>Sanil Rao, Mohammad Alaul Haque Monil, Het Mankad, Narasinga Rao Miniskar, Keita Teranishi</i> <i>Jeffrey Vetter, Franz Franchetti</i>	
POSTER: PIMAP: Characterizing a Real Processing-in-Memory System for Analytical Data Processing	467
<i>Manos Frouzakis, Juan Gómez-Luna, Geraldo F. Oliveira, Mohammad Sadrosadati, Onur Mutlu</i>	
Poster: Value-Aware Scheduler for Energy Reduction	470
<i>Haiyue Ma, Kaifeng Xu, David Wentzlaff</i>	