2026 IEEE/ACM International Symposium on Code Generation and Optimization (CGO 2026)

Sydney, Australia
31 January - 4 February 2026

IEEE

IEEE Catalog Number: CFP26CGO-POD
ISBN: 979-8-3315-9289-9

Copyright © 2026 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: CFP26CGO-POD
ISBN (Print-On-Demand): 979-8-3315-9289-9
ISBN (Online): 979-8-3315-9288-2

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



Contents

Frontmatter

Welcome from the General Chair . . . iii
Welcome from the Program Chairs . . . iv
CGO 2026 Organization . . . v
CGO 2026 Sponsors and Supporters . . . viii

Compiling for ML 1

Enabling Spill-Free Compilation via Affine-Based Live Range Reduction Optimization
Prasanth Chatarasi, Alex Gatea, Wei Wang, Chris Bowler, Shubham Jain, Masoud Ataei Jaliseh, Nicole Khoun, Alberto Mannari, Bardia Mahjour, Viji Srinivasan, and Swagath Venkataramani — *IBM Research, USA; IBM, Canada; IBM, Switzerland* . . . 1

GRANII: Selection and Ordering of Primitives in GRAph Neural Networks using Input Inspection
Damitha Lenadora, Vimarsh Sathia, Gerasimos Gerogiannis, Serif Yesil, Josep Torrellas, and Charith Mendis — *University of Illinois at Urbana-Champaign, USA; NVIDIA, USA* . . . 14

Fast Autoscheduling for Sparse ML Frameworks
Bobby Yan, Alexander J Root, Trevor Gale, David Broman, and Fredrik Kjolstad — *Stanford University, USA; KTH Royal Institute of Technology, Sweden* . . . 28

Eliminating Redundancy: Ultra-compact Code Generation for Programmable Dataflow Accelerators
Prasanth Chatarasi, Alex Gatea, Bardia Mahjour, Jintao Zhang, Alberto Mannari, Chris Bowler, Shubham Jain, Masoud Ataei Jaliseh, Nicole Khoun, Kamlesh Kumar, Viji Srinivasan, and Swagath Venkataramani — *IBM Research, USA; IBM, Canada; IBM, Switzerland* 44

Security

PriTran: Privacy-Preserving Inference for Transformer-Based Language Models under Fully Homomorphic Encryption
Yuechen Mu, Guangli Li, Shiping Chen, and Jingling Xue — *UNSW, Australia; Institute of Computing Technology at Chinese Academy of Sciences, China; CSIROs Data61, Australia* . . . 57

FHEFusion: Enabling Operator Fusion in FHE Compilers for Depth-Efficient DNN Inference
Tianxiang Sui, Jianxin Lai, Long Li, Peng Yuan, Yan Liu, Qing Zhu, Xiaojing Zhang, Linjie Xiao, Mingzhe Zhang, and Jingling Xue — *Ant Group, China; UNSW, Australia* . . . 70

Towards Path-Aware Coverage-Guided Fuzzing
Giacomo Priamo, Daniele Cono D'Elia, Mathias Payer, and Leonardo Querzoni — *Sapienza University of Rome, Italy; EPFL, Switzerland* . . . 84

SecSwift, a Compiler-Based Framework for Software Countermeasures in Cybersecurity
François de Ferrière, Yves Janin, and Sirine Mechmech — *STMICROELECTRONICS, France; Grenoble INP, France* . . . 98

Abstractions

Partial-Evaluation Templates: Accelerating Partial Evaluation with Pre-compiled Templates
Florian Huemer, Aleksandar Prokopec, David Leopoldseder, Raphael Mosaner, and Hanspeter Mössenböck — *JKU Linz, Austria; Oracle Labs, Zurich, Switzerland; Oracle Labs, Vienna, Austria; Oracle Labs, Linz, Austria* . . . 109

Pyls: Enabling Python Hardware Synthesis with Dynamic Polymorphism via LCRS Encoding
Bolei Tong, Yongyan Fang, Chaorui Wang, Qingan Li, Jingling Xue, and Yuan Mengting — *Wuhan University, China; UNSW, Australia* 123

SkeleShare: Algorithmic Skeletons and Equality Saturation for Hardware Resource Sharing
Jonathan Van der Cruysse, Tzung-Han Juang, Shakiba Bolbolian Khah, and Christophe Dubach — *McGill University, Canada; Mila, Canada* . . . 136

Ember: A Compiler for Embedding Operations on Decoupled Access-Execute Architectures
Marco Siracusa, Olivia Hsu, Víctor Soria-Pardos, Joshua Randall, Arnaud Grasset, Eric Biscondi, Doug Joseph, Randy Allen, Fredrik Kjolstad, Miquel Moretó Planas, and Adrià Armejach — *Barcelona Supercomputing Center, Spain; Stanford University, USA; Carnegie Mellon University, USA; Arm, USA; Universitat Politècnica de Catalunya, Spain* . . . 150

Memory

Flow-Graph-Aware Tiling and Rescheduling for Memory-Efficient On-Device Inference
Yeonoh Jeong, Taehyeong Park, and Yongjun Park — *Yonsei University, Republic of Korea* . . . 164

VFlatten: Selective Value-Object Flattening using Hybrid Static and Dynamic Analysis
Arjun H. Kumar, Bhavya Hirani, Hang Shao, Tobi Ajila, Vijay Sundaresan, Daryl Maier, and Manas Thakur — *IIT Mandi, India; Sardar Vallabhbhai National Institute of Technology, Surat, India; IBM, Canada; IIT Bombay, India* 176

FRUGAL: Pushing GPU Applications beyond Memory Limits
Lingqi Zhang, Tengfei Wang, Jiajun Huang, Chen Zhuang, Ivan R. Ivanov, Peng Chen, Toshio Endo, and Mohamed Wahib — *RIKEN RCCS, Japan; Google Cloud, Japan; University of South Florida, USA; Institute of Science Tokyo, Japan* 188

Automatic Data Enumeration for Fast Collections
Tommy McMichen and Simone Campanoni — *Northwestern University, USA; Google, USA* 202

DSLs

FORTE: Online DataFrame Query Optimizer
Yoonho Choi, Kyoungtae Lee, Minji Kim, Hyungsoo Jung, and Hyojin Sung — *POSTECH, Republic of Korea; Seoul National University, Republic of Korea; Ewha Womans University, Republic of Korea* 216

LEGO: A Layout Expression Language for Code Generation of Hierarchical Mapping
Amir Mohammad Tavakkoli, Cosmin E. Oancea, and Mary Hall — *University of Utah, USA; University of Copenhagen, Denmark* . 228

Pushing Tensor Accelerators beyond MatMul in a User-Schedulable Language
Yihong Zhang, Derek Gerstmann, Andrew Adams, and Maaz Bin Safeer Ahmad — *University of Washington, USA; Adobe, USA* . . 242

Tawa: Automatic Warp Specialization for Modern GPUs with Asynchronous References
Hongzheng Chen, Bin Fan, Alexander Collins, Bastian Hagedorn, Evghenii Gaburov, Masahiro Masuda, Matthew Brookhart, Chris Sullivan, Jason Knight, Zhiru Zhang, and Vinod Grover — *Cornell University, USA; NVIDIA, USA; NVIDIA, UK; NVIDIA, Germany* 255

Quantum / HLS

Dependence-Driven, Scalable Quantum Circuit Mapping with Affine Abstractions
Marouane Benbetka, Merwan Bekkar, Riyadh Baghdadi, and Martin Kong — *NYU Abu Dhabi, United Arab Emirates; Ohio State University, USA* 268

Space-Time Optimisations for Early Fault-Tolerant Quantum Computation
Sanaa Sharma and Prakash Murali — *University of Cambridge, UK* 281

OpenQudit: Extensible and Accelerated Numerical Quantum Compilation via a JIT-Compiled DSL
Ed Younis — *Lawrence Berkeley National Laboratory, USA* 295

Selene: Cross-Level Barrier-Free Pipelining for Irregular Nested Loops in High-Level Synthesis
Sungwoo Yun, Seonyoung Cheon, Dongkwan Kim, Heelim Choi, Kunmo Jeong, Chan Lee, Yongwoo Lee, and Hanjun Kim — *Yonsei University, Republic of Korea; DGIST, Republic of Korea* 306

Parallelization / Vectorization

Enabling Automatic Compiler-Driven Vectorization of Transformers
Shreya Alladi, Alberto Ros, and Alexandra Jimborean — *University of Murcia, Spain* 319

Unlocking Python Multithreading Capabilities using OpenMP-Based Programming with OMP4Py
César Piñeiro and Juan C. Pichel — *University of Santiago de Compostela, Spain* 334

The Parallel-Semantics Program Dependence Graph for Parallel Optimization
Yian Su, Brian Homerding, Haocheng Gao, Federico Sossai, Yebin Chon, David I. August, and Simone Campanoni — *Northwestern University, USA; Princeton University, USA* 348

From Threads to Tiles: T2T, a Compiler for CUDA-to-NPU Translation via 2D Vectorization
Shuaijiang Li, Jiacheng Zhao, Ying Liu, Shuoming Zhang, Lei Chen, Yijin Li, Yangyu Zhang, Zhicheng Li, Runyu Zhou, Xiyu Shi, Chunwei Xia, Yuan Wen, Xiaobing Feng, and Huimin Cui — *Institute of Computing Technology at Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; University of Leeds, UK; University of Aberdeen, UK; XCORESIGMA, China* . . 362

Binary / JIT

Binary Diffing via Library Signatures
Andrei Rimsa, Anderson Faustino da Silva, Camilo Santana, and Fernando Magno Quintão Pereira — *CEFET-MG, Brazil; State University of Maringá, Brazil; Federal University of Minas Gerais, Brazil* 375

BIT: Empowering Binary Analysis through the LLVM Toolchain
Puzhuo Liu, Peng Di, Jingling Xue, and Yu Jiang — *Ant Group, China; Tsinghua University, China; UNSW, Australia* 390

Dr.avx: A Dynamic Compilation System for Seamlessly Executing Hardware-Unsupported Vectorization Instructions
Yue Tang, Mianzhi Wu, Yufeng Li, Haoyu Liao, Jianmei Guo, and Bo Huang — *East China Normal University, China* 403

Practical: Are Abstract-Interpreter Baseline JITs Worth It? An Empirical Evaluation through Metacompilation
Nahuel Palumbo, Guillermo Polito, Stéphane Ducasse, and Pablo Tesone — *Univ. Lille - Inria - CNRS - Centrale Lille - UMR 9189 CRIStAL, France* 416

Code Generation

TPDE: A Fast Adaptable Compiler Back-End Framework
Tobias Schwarz, Tobias Kamm, and Alexis Engelke — *TU Munich, Germany* 427

Synthesizing Instruction Selection Back-Ends from ISA Specifications Made Practical
Florian Drescher and Alexis Engelke — *TU Munich, Germany* 440

SparseX: Synergizing GPU Libraries for Sparse Matrix Multiplication on Heterogeneous Processors
Ruifeng Zhang, Xiangwei Wang, Ang Li, and Xipeng Shen — *North Carolina State University, USA; Pacific Northwest National Laboratory, USA; University of Washington, USA* 453

Compilation of Generalized Matrix Chains with Symbolic Sizes
Francisco López, Lars Karlsson, and Paolo Bientinesi — *Umeå University, Sweden* 466

Profiling / Instrumentation

TRACE4J: A Lightweight, Flexible, and Insightful Performance Tracing Tool for Java
Haide He and Pengfei Su — *University of California at Merced, USA* 479

Proton: Towards Multi-level, Adaptive Profiling for Triton
Keren Zhou, Tianle Zhong, Hao Wu, Jihyeong Lee, Yue Guan, Yufei Ding, Corbin Robeck, Yuanwei Fang, Jeff Niu, and Philippe Tillet — *George Mason University, USA; OpenAI, USA; University of Virginia, USA; University of California at San Diego, USA; Meta, USA* 493

On the Precision of Dynamic Program Fingerprints Based on Performance Counters
Anderson Faustino da Silva, Marcelo Borges Nogueira, Sérgio Queiroz de Medeiros, Jeronimo Castrillon, and Fernando Magno Quintão Pereira — *State University of Maringá, Brazil; Federal University of Rio Grande do Norte, Brazil; TU Dresden, Germany; Federal University of Minas Gerais, Brazil* 507

PASTA: A Modular Program Analysis Tool Framework for Accelerators
Mao Lin, Hyeran Jeon, and Keren Zhou — *University of California at Merced, USA; George Mason University, USA; OpenAI, USA* 520

Analysis

PIP: Making Andersens Points-to Analysis Sound and Practical for Incomplete C Programs
Håvard Rognebakke Krogstie, Helge Bahmann, Magnus Själander, and Nico Reissmann — *NTNU, Norway; Independent Researcher, Switzerland; Independent Researcher, Norway* 535

Thinking Fast and Correct: Automated Rewriting of Numerical Code through Compiler Augmentation
Siyuan Brant Qian, Vimarsh Sathia, Ivan R. Ivanov, Jan Hückelheim, Paul Hovland, and William S. Moses — *University of Illinois at Urbana-Champaign, USA; Institute of Science Tokyo, Japan; RIKEN RCCS, Japan; Argonne National Laboratory, USA* 548

PolyUFC: Polyhedral Compilation Meets Roofline Analysis for Uncore Frequency Capping
Nilesh Rajendra Shah, M V V S Manoj Kumar, Dhairya Baxi, and Ramakrishna Upadrasta — *IIT Hyderabad, India* 563

Accelerating App Recompilation across Android System Updates by Code Reusing
Hongtao Wu, Yu Chen, Mengfei Xie, Futeng Yang, Jun Yan, Jiang Ma, Jianming Fu, Chun Jason Xue, and Qingan Li — *Wuhan University, China; Guangdong OPPO Mobile Telecommunications, China; Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates* 577

Compiling for ML 2

QIGen: A Kernel Generator for Inference on Nonuniformly Quantized Large Language Models
Tommaso Pegolotti, Dan Alistarh, and Markus Püschel — *ETH Zurich, Switzerland; IST Austria, Austria* 589

DyPARS: Dynamic-Shape DNN Optimization via Pareto-Aware MCTS for Graph Variants
Hao Qian, Guangli Li, Qiuchu Yu, Xueying Wang, and Jingling Xue — *UNSW, Australia; Institute of Computing Technology at Chinese Academy of Sciences, China; Beijing University of Posts and Telecommunications, China* 603

Compiler-Runtime Co-operative Chain of Verification for LLM-Based Code Optimization
Hyunho Kwon, Sanggyu Shin, Ju Min Lee, Hoyun Youm, Seungbin Song, Seongho Kim, Hanwoong Jung, Seungwon Lee, and Hanjun Kim — *Yonsei University, Republic of Korea; SAIT, Republic of Korea* 617

Hexcute: A Compiler Framework for Automating Layout Synthesis in GPU Programs
Xiao Zhang, Yaoyao Ding, Bolin Sun, Yang Hu, Tatiana Shpeisman, and Gennady Pekhimenko — *University of Toronto, Canada; NVIDIA, Canada; Vector Institute, Canada* 630

Tensor Optimization

Multidirectional Propagation of Sparsity Information across Tensor Slices
Kaio Henrique Andrade Ananias, Danila Seliayeu, J. Nelson Amaral, and Fernando Magno Quintão Pereira — *Federal University of Minas Gerais, Brazil; University of Alberta, Canada* 644

Synthesizing Specialized Sparse Tensor Accelerators for FPGAs via High-Level Functional Abstractions
Hamza Javed and Christophe Dubach — *McGill University, Canada; Mila, Canada* 657

Progressive Low-Precision Approximation of Tensor Operators on GPUs: Enabling Greater Trade-Offs between Performance and Accuracy
Fan Luo, Guangli Li, Zhaoyang Hao, Xueying Wang, Xiaobing Feng, Huimin Cui, and Jingling Xue — *Institute of Computing Technology at Chinese Academy of Sciences, China; University of Chinese Academy of Sciences, China; UNSW, Australia; Beijing University of Posts and Telecommunications, China* 670

Tensor Program Superoptimization through Cost-Guided Symbolic Program Synthesis
Alexander Brauckmann, Aarsh Chaube, José Wesley de Souza Magalhães, Elizabeth Polgreen, and Michael F. P. OBoyle — *University of Edinburgh, UK* 683

Optimization

A Reinforcement Learning Environment for Automatic Code Optimization in the MLIR Compiler
Mohammed Tirichine, Nassim Ameur, Nazim Bendib, Iheb Nassim Aouadj, Djad Bouchama, Rafik Bouloudene, and Riyadh Baghdadi — *NYU Abu Dhabi, United Arab Emirates; École Nationale Supérieure dInformatique, Algeria; University of Science and Technology Houari Boumediene, Algeria* 696

Towards Threading the Needle of Debuggable Optimized Binaries
Cristian Assaiante, Simone Di Biasio, Snehasish Kumar, Giuseppe Antonio Di Luna, Daniele Cono D'Elia, and Leonardo Querzoni — *Sapienza University of Rome, Italy; Google, USA* 711

Compiler-Assisted Instruction Fusion
Ravikiran Ravindranath Reddy, Sawan Singh, Arthur Perais, Alberto Ros, and Alexandra Jimborean — *University of Murcia, Spain; Univ. Grenoble Alpes - CNRS - Grenoble INP - TIMACNRS, France* 726

LLM-VeriOpt: Verification-Guided Reinforcement Learning for LLM-Based Compiler Optimization
Xiangxin Fang, Jiaqin Kang, Rodrigo Rocha, Sam Ainsworth, and Lev Mukhanov — *Queen Mary University of London, UK; University of Edinburgh, UK* 740

Author Index 756