
Washington, DC, USA
16 – 20 February 2019

HPCA 2019

Table of Contents

Message from the General Chairs xiii
Message from the Program Chairs xiv
Organizing Committee xvi
Program Committee xviii
Industry Session Program Committee xx
External Review Committee xxi
Keynote Abstracts xxiv
Sponsors xxvii

Session 1: Best Paper Nominees

The Accelerator Wall: Limits of Chip Specialization 1

Adi Fuchs (Princeton University) and David Wentzlaff (Princeton University)

Stretch: Balancing QoS and Throughput for Colocated Server Workloads on SMT Cores 15

Artemiy Margaritov (University of Edinburgh), Siddharth Gupta (EPFL), Rekai Gonzalez-Alberquilla (Arm Ltd.), and Boris Grot (University of Edinburgh)

CIDR: A Cost-Effective In-Line Data Reduction System for Terabit-Per-Second Scale SSD Arrays 28

Mohammadamin Ajdari (POSTECH), Pyeongsu Park (Seoul National University), Joonsung Kim (Seoul National University), Dongup Kwon (Seoul National University), and Jangwoo Kim (Seoul National University)

Composite-ISA Cores: Enabling Multi-ISA Heterogeneity Using a Single ISA 42

Ashish Venkat (University of Virginia), Harsha Basavaraj (University of California, San Diego), and Dean M. Tullsen (University of California, San Diego)
**Session 2A: Accelerators for DNNs**

HyPar: Towards Hybrid Parallelism for Deep Learning Accelerator Array 56

Linghao Song (Duke University), Jiachen Mao (Duke University), Youwei Zhuo (University of Southern California), Xuehai Qian (University of Southern California), Hai Li (Duke University), and Yiran Chen (Duke University)

E-RNN: Design Optimization for Efficient Recurrent Neural Networks in FPGAs 69

Zhe Li (Syracuse University), Caiwen Ding (Northeastern University), Siyue Wang (Northeastern University), Wujie Wen (Florida International University), Youwei Zhuo (University of Southern California), Chang Liu (Carnegie Mellon University), Qianru Qiu (Syracuse University), Wenyao Xu (University at Buffalo (SUNY)), Xue Lin (Northeastern University), Xuehai Qian (University of Southern California), and Yanzhi Wang (Northeastern University)

Bit Prudent In-Cache Acceleration of Deep Convolutional Neural Networks 81

Xiaowei Wang (University of Michigan), Jiecao Yu (University of Michigan), Charles Augustine (Intel Corporation), Ravi Iyer (Intel Corporation), and Reetuparna Das (University of Michigan)

Shortcut Mining: Exploiting Cross-Layer Shortcut Reuse in DCNN Accelerators 94

Arash Azizimazreah (Oregon State University) and Lizhong Chen (Oregon State University)

**Session 2B: Power Efficiency**

Fine-Tuning the Active Timing Margin (ATM) Control Loop for Maximizing Multi-core Efficiency on an IBM POWER Server 106

Yazhou Zu (The University of Texas at Austin), Daniel Richins (The University of Texas at Austin), Charles Lefurgy (IBM Research), and Vijay Reddi (Harvard University)

μDPM: Dynamic Power Management for the Microsecond Era 120

Chih-Hsun Chou (University of California, Riverside), Laxmi N. Bhuyan (University of California, Riverside), and Daniel Wong (University of California, Riverside)

Adaptive Voltage/Frequency Scaling and Core Allocation for Balanced Energy and Performance on Multicore CPUs 133

George Papadimitriou (University of Athens), Athanasios Chatzidimitriou (University of Athens), and Dimitris Gizopoulos (University of Athens)

Resilient Low Voltage Accelerators for High Energy Efficiency 147

Nandhini Chandramoorthy (IBM T.J.Watson Research Center), Karthik Swaminathan (IBM T.J.Watson Research Center), Martin Cochet (IBM T.J.Watson Research Center), Arun Paidimarri (IBM T.J.Watson Research Center), Schuyler Eldridge (IBM T.J.Watson Research Center), Rajiv Joshi (IBM T.J.Watson Research Center), Matthew Ziegler (IBM T.J.Watson Research Center), Alper Buyuktosunoglu (IBM T.J.Watson Research Center), and Pradip Bose (IBM T.J.Watson Research Center)
Session 3A: Datacenter

Pliant: Leveraging Approximation to Improve Datacenter Resource Efficiency ............................. 159
Neeraj Kulkarni (Cornell University), Feng Qi (Cornell University),
and Christina Delimitrou (Cornell University)

Kelp: QoS for Accelerated Machine Learning Systems ................................................................. 172
Haishan Zhu (The University of Texas at Austin), David Lo (Google),
Liqun Cheng (Google), Rama Govindaraju (Google), Parthasarathy
Ranganathan (Google), and Mattan Erez (The University of Texas at
Austin)

Enhancing Server Efficiency in the Face of Killer Microseconds .................................................. 185
Amirhossein Mirhosseini (University of Michigan), Akshitha Sriraman
(University of Michigan), and Thomas F. Wenisch (University of
Michigan)

Poly: Efficient Heterogeneous System and Application Management for Interactive Applications .......... 199
Shuo Wang (Peking University), Yun Liang (Peking University), and Wei
Zhang (Hong Kong University of Science and Technology)

Session 3B: Emerging Technologies

The What's Next Intermittent Computing Architecture ................................................................. 211
Karthik Ganesan (University of Toronto), Joshua San Miguel (University
of Wisconsin - Madison), and Natalie Enright Jerger (University of
Toronto)

eQASM: An Executable Quantum Instruction Set Architecture ...................................................... 224
X. Fu (Delft University of Technology), L. Riesebos (Delft University
of Technology), M. A. Rol (Delft University of Technology), Jeroen van
Straten (Delft University of Technology), J. van Someren (Delft
University of Technology), N. Khammassi (Delft University of
Technology), I. Ashraf (Delft University of Technology), R. F. L.
Vermeulen (Delft University of Technology), V. Newsum (Netherlands
Organisation for Applied Scientific Research (TNO); Delft University
of Technology), K. K. L. Loh (Netherlands Organisation for Applied
Scientific Research (TNO); Delft University of Technology), J. C. de
Sterke (Delft University of Technology), W. J. Vlothuizen (Netherlands
Organisation for Applied Scientific Research (TNO); Delft University
of Technology), R. N. Schouten (Delft University of Technology), C. G.
Ahmudever (Delft University of Technology), L. DiCarlo (Delft
University of Technology), and K. Bertels (Delft University of
Technology)

Reliability Evaluation of Mixed-Precision Architectures ............................................................... 238
Fernando Fernandes dos Santos (UFRGS), Caio Lunardi (UFRGS), Daniel
Oliveira (UFRGS), Fabiano Libano (UFRGS), and Paolo Rech (UFRGS)
Architecting Waferscale Processors - A GPU Case Study

Saptadeep Pal (University of California, Los Angeles), Daniel Petrisko
(University of Illinois at Urbana-Champaign), Matthew Tomei
(University of Illinois at Urbana-Champaign), Puneet Gupta (University of California, Los Angeles), Subramanian S. Iyer (University of California, Los Angeles), and Rakesh Kumar (University of Illinois at Urbana-Champaign)

Session 4A: Security

Conditional Speculation: An Effective Approach to Safeguard Out-of-Order Execution Against Spectre Attacks

Peinan Li (State Key Laboratory of Information Security, Institute of Information Engineering, CAS and University of Chinese Academy of Sciences), Lutan Zhao (State Key Laboratory of Information Security, Institute of Information Engineering, CAS and University of Chinese Academy of Sciences), Rui Hou (State Key Laboratory of Information Security, Institute of Information Engineering, CAS and University of Chinese Academy of Sciences), Lixin Zhang (Institute of Computing Technology, CAS), and Dan Meng (State Key Laboratory of Information Security, Institute of Information Engineering, CAS and University of Chinese Academy of Sciences)

FPGA Accelerated INDEL Realignment in the Cloud

Lisa Wu (University of California, Berkeley), David Bruns-Smith (University of California, Berkeley), Frank A. Nothaft (Databricks), Qijing Huang (University of California, Berkeley), Sagar Karandikar (University of California, Berkeley), Johnny Le (University of California, Berkeley), Andrew Lin (University of California, Berkeley), Howard Mao (University of California, Berkeley), Brendan Sweeney (University of California, Berkeley), Krste Asanovi (University of California, Berkeley and SiFive, Inc.), David A. Patterson (University of California, Berkeley and Google, Inc.), and Anthony D. Joseph (University of California, Berkeley)

POWERT Channels: A Novel Class of Covert Communication Exploiting Power Management Vulnerabilities

S. Karen Khatamifard (University of Minnesota), Longfei Wang (University of South Florida), Amitabh Das (University of South Florida), Selcuk Kose (University of Rochester), and Ulya R. Karpuzcu (University of Minnesota)

Session 4B: Industry Session 1: Mobile & Low Power

Killi: Runtime Fault Classification to Deploy Low Voltage Caches without MBIST


Gables: A Roofline Model for Mobile SoCs

Mark Hill (University of Wisconsin—Madison) and Vijay Janapa Reddi (Harvard University)
Machine Learning at Facebook: Understanding Inference at the Edge 331
  Carole-Jean Wu (Facebook), David Brooks (Facebook), Kevin Chen (Facebook), Douglas Chen (Facebook), Sy Choudhury (Facebook), Marat Dukhan (Facebook), Kim Hazelwood (Facebook), Eldad Isaac (Facebook), Yangqing Jia (Facebook), Bill Jia (Facebook), Tommer Leyvand (Facebook), Hao Lu (Facebook), Yang Lu (Facebook), Lin Qiao (Facebook), Brandon Reagen (Facebook), Joe Spisak (Facebook), Fei Sun (Facebook), Andrew Tulloch (Facebook), Peter Vajda (Facebook), Xiaodong Wang (Facebook), Yanghan Wang (Facebook), Bram Wasti (Facebook), Yiming Wu (Facebook), Ran Xian (Facebook), Sungjoo Yoo (Facebook), and Peizhao Zhang (Facebook)

Session 5A: Accelerators for Emerging Applications

VIP: A Versatile Inference Processor 345
  Skand Hurkat (Microsoft) and José F. Martínez (Cornell University)

Darwin-WGA: A Co-processor Provides Increased Sensitivity in Whole Genome Alignments with High Speedup 359
  Yatish Turakhia (Stanford University), Sneha D. Goenka (Stanford University), Gill Bejerano (Stanford University), and William J. Dally (Stanford University, NVIDIA Research)

Analysis and Optimization of the Memory Hierarchy for Graph Processing Workloads 373
  Abanti Basak (University of California, Santa Barbara), Shuangchen Li (University of California, Santa Barbara), Xing Hu (University of California, Santa Barbara), Sang Min Oh (University of California, Santa Barbara), Xinfeng Xie (University of California, Santa Barbara), Li Zhao (Alibaba, Inc.), Xiaowei Jiang (Alibaba, Inc.), and Yuan Xie (University of California, Santa Barbara)

FPGA-Based High-Performance Parallel Architecture for Homomorphic Computing on Encrypted Data 387
  Sujoy Sinha Roy (University of Birmingham and KU Leuven, imec-COSIC, Belgium), Furkan Turan (KU Leuven, imec-COSIC, Belgium), Kimmo Jarvinen (University of Helsinki), Frederik Vercauteren (KU Leuven, imec-COSIC, Belgium), and Ingrid Verbauwhede (KU Leuven, imec-COSIC, Belgium)

Session 5B: Memory Hierarchy Management

Bingo Spatial Data Prefetcher 399
  Mohammad Bakhshalipour (Sharif University of Technology and Institute for Research in Fundamental Sciences (IPM)), Mehran Shakerinava (Sharif University of Technology), Pejman Lofit-Kamran (Institute for Research in Fundamental Sciences (IPM)), and Hamid Sarbazi-Azad (Sharif University of Technology and Institute for Research in Fundamental Sciences (IPM))

NoMap: Speeding-Up JavaScript Using Hardware Transactional Memory 412
  Thomas Shull (University of Illinois at Urbana-Champaign), Jiho Choi (University of Illinois at Urbana-Champaign), Maria J. Garzaran (Intel), and Josep Torrellas (University of Illinois at Urbana-Champaign)
FUSE: Fusing STT-MRAM into GPUs to Alleviate Off-Chip Memory Access Overheads 426
Jie Zhang (Yonsei University), Myoungsoo Jung (Yonsei University), and Mahmut Kandemir (Pennsylvania State University)

Featherlight Reuse-Distance Measurement 440
Qingsen Wang (College of William & Mary), Xu Liu (College of William & Mary), and Milind Chabbi (Scalable Machines Research)

Session 6A: Industry Session 2: Microarchitecture

Efficient Load Value Prediction Using Multiple Predictors and Filters 454
Rami Sheikh (Qualcomm Technologies, Inc.) and Derek Hower (Qualcomm Technologies, Inc.)

BRB: Mitigating Branch Predictor Side-Channels 466
Ilias Vougioukas (Arm Research / University of Southampton), Nikos Nikoleris (Arm Research), Andreas Sandberg (Arm Research), Stephan Diestelhorst (Arm Research), Bashir M. Al-Hashimi (University of Southampton), and Geoff V. Merrett (University of Southampton)

Elastic Instruction Fetching 478

Session 6B: Best of CAL (Computer Architecture Letters)

The Best of IEEE Computer Architecture Letters in 2018 491
Paul Gratz (Texas A&M University)

Session 7A: GPUs/Modeling

Poise: Balancing Thread-Level Parallelism and Memory System Performance in GPUs Using Machine Learning 492
Saumay Dubish (University of Edinburgh), Vijay Nagarajan (University of Edinburgh), and Nigel Topham (University of Edinburgh)

A Hybrid Framework for Fast and Accurate GPU Performance Estimation through Source-Level Analysis and Trace-Based Simulation 506
Xiebing Wang (Technical University of Munich), Kai Huang (Sun Yat-Sen University), Alois Knoll (Technical University of Munich), and Xuehai Qian (University of Southern California)

Understanding the Future of Energy Efficiency in Multi-Module GPUs 519
Akhil Arunkumar (Arizona State University), Evgeny Bolotin (NVIDIA), David Nellans (NVIDIA), and Carole-Jean Wu (Arizona State University)
Session 7B: Microarchitecture

R3-DLA (Reduce, Reuse, Recycle): A More Efficient Approach to Decoupled Look-Ahead Architectures 533
Sushant Kondguli (University of Rochester) and Michael Huang (University of Rochester)

Recycling Data Slack in Out-of-Order Cores 545
Gokul Subramanian Ravi (University of Wisconsin - Madison) and Mikko Lipasti (University of Wisconsin - Madison)

Freeway: Maximizing MLP for Slice-Out-of-Order Execution 558
Rakesh Kumar (Norwegian University of Science and Technology (NTNU)), Mehdi Alipour (Uppsala University), and David Black-Schaffer (Uppsala University)

Session 8A: Memory

Enabling Transparent Memory-Compression for Commodity Memory Systems 570
Vinson Young (Georgia Institute of Technology), Sanjay Kariyappa (Georgia Institute of Technology), and Moinuddin Qureshi (Georgia Institute of Technology)

D-RaNGe: Using Commodity DRAM Devices to Generate True Random Numbers with Low Latency and High Throughput 582
Jeremie S. Kim (Carnegie Mellon University; ETH Zürich), Minesh Patel (ETH Zürich), Hasan Hassan (ETH Zürich), Lois Orosa (ETH Zürich), and Onur Mutlu (ETH Zürich; Carnegie Mellon University)

PageSeer: Using Page Walks to Trigger Page Swaps in Hybrid Memory Systems 596
Apostolos Kokolis (University of Illinois at Urbana-Champaign), Dimitrios Skarlatos (University of Illinois at Urbana-Champaign), and Josep Torrellas (University of Illinois at Urbana-Champaign)

Session 8B: Accelerators for Graphics/VR

PIM-VR: Erasing Motion Anomalies In Highly-Interactive Virtual Reality World with Customized Memory Cube 609
Chenhao Xie (University of Houston), Xingyao Zhang (University of Houston), Ang Li (Pacific Northwest National Laboratory), Xin Fu (University of Houston), and Shuaiwen Song (Pacific Northwest National Laboratory)

Rendering Elimination: Early Discard of Redundant Tiles in the Graphics Pipeline 623
Martí Anglada (Universitat Politècnica de Catalunya), Enrique de Lucas (Semidynamics Technology Services), Joan-Manuel Parcerisa (Universitat Politècnica de Catalunya), Juan L. Aragón (Universidad de Murcia), Pedro Marcuello (Semidynamics Technology Services), and Antonio González (Universitat Politècnica de Catalunya)

Early Visibility Resolution for Removing Ineffectual Computations in the Graphics Pipeline 635
Martí Anglada (Universitat Politècnica de Catalunya), Enrique de Lucas (Semidynamics Technology Services), Joan-Manuel Parcerisa (Universitat Politècnica de Catalunya), Juan L. Aragón (Universidad de Murcia), and Antonio González (Universitat Politècnica de Catalunya)
Session 9A: Emerging Memory Technologies

String Figure: A Scalable and Elastic Memory Network Architecture 647
Matheus Ogleari (University of California, Santa Cruz), Ye Yu
(University of Kentucky), Chen Qian (University of California, Santa Cruz), Ethan Miller (University of California, Santa Cruz), and Jishen Zhao (University of California, San Diego)

NAND-Net: Minimizing Computational Complexity of In-Memory Processing for Binary Neural Networks 661
Hyeonuk Kim (KAIST), Jaehyeong Sim (KAIST), Yeongjae Choi (KAIST), and Lee-Sup Kim (KAIST)

Active-Routing: Compute on the Way for Near-Data Processing 674
Jiayi Huang (Texas A&M University), Ramprakash Reddy Puli (NVIDIA Corporation), Pritam Majumder (Texas A&M University), Sungkeun Kim (Texas A&M University), Rahul Boyapati (Intel Corporation), Ki Hwan Yum (Texas A&M University), and Eun Jung Kim (Texas A&M University)

Session 9B: Industry Session 3: Servers

Understanding the Impact of Socket Density in Density Optimized Servers 687
Manish Arora (University of California, San Diego), Matt Skach (University of Michigan), Wei Huang (AMD Research, Advanced Micro Devices, Inc.), Xudong An (AMD Research, Advanced Micro Devices, Inc.), Jason Mars (University of Michigan), Lingjia Tang (University of Michigan), and Dean M. Tullsen (University of California, San Diego)

A Scalable Priority-Aware Approach to Managing Data Center Server Power 701
Yang Li (IBM/Carnegie Mellon University), Charles R. Lefurgy (IBM), Karthick Rajamani (IBM), Malcolm S. Allen-Ware (IBM), Guillermo J. Silva (IBM), Daniel D. Heimsoth (IBM), Saugata Ghose (Carnegie Mellon University), and Onur Mutlu (ETH Zurich/Carnegie Mellon University)

Power Aware Heterogeneous Node Assembly 715
Bilge Acun (IBM T. J. Watson Research Center), Alper Buyuktosunoglu (IBM T. J. Watson Research Center), Eun Kyung Lee (IBM T.J. Watson Research Center), and Yoonho Park (IBM T.J. Watson Research Center)

Author Index 729