2017 IEEE International Symposium on High Performance Computer Architecture (HPCA 2017)

Austin, Texas, USA
4 – 8 February 2017

IEEE Catalog Number: CFP17013-POD
ISBN: 978-1-5090-4986-8

Copyright © 2017 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:	CFP17013-POD
ISBN (Print-On-Demand):	978-1-5090-4986-8
ISBN (Online):	978-1-5090-4985-1
ISSN:	1530-0897

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

2017 IEEE International Symposium on High Performance Computer Architecture

HPCA 2017

Table of Contents

Message from the Program Chair.....xi
Committees.....xiii
Program Committee.....xiv
External Review Committee.....xvi
Sponsors.....xix

HPCA Session 1: Lightning Rounds

HPCA Session 2: Best Paper Nominees

Towards Pervasive and User Satisfactory CNN across GPU Microarchitectures 1
Mingcong Song, Yang Hu, Huixiang Chen, and Tao Li

Near-Optimal Access Partitioning for Memory Hierarchies with Multiple Heterogeneous Bandwidth Sources 13
Jayesh Gaur, Mainak Chaudhuri, Pradeep Ramachandran, and Sreenivas Subramoney

NCAP: Network-Driven, Packet Context-Aware Power Management for Client-Server Architecture 25
Mohammad Alian, Ahmed H. M. O. Abulila, Lokesh Jindal, Daehoon Kim, and Nam Sung Kim

Supporting Address Translation for Accelerator-Centric Architectures 37
Yuchen Hao, Zhenman Fang, Glenn Reinman, and Jason Cong

HPCA Session 3A: Industrial Session

Vulnerabilities in MLC NAND Flash Memory Programming: Experimental Analysis, Exploits, and Mitigation Techniques 49
Yu Cai, Saugata Ghose, Yixin Luo, Ken Mai, Onur Mutlu, and Erich F. Haratsch

Defect Analysis and Cost-Effective Resilience Architecture for Future DRAM Devices 61
Sanguhn Cha, Seongil O, Hyunsung Shin, Sangjoon Hwang, Kwangil Park, Seong Jin Jang, Joo Sun Choi, Gyo Young Jin, Young Hoon Son, Hyunyoon Cho, Jung Ho Ahn, and Nam Sung Kim

Architecting an Energy-Efficient DRAM System for GPUs 73
Niladrish Chatterjee, Mike O'Connor, Donghyuk Lee, Daniel R. Johnson, Stephen W. Keckler, Minsoo Rhu, and William J. Dally

Design and Analysis of an APU for Exascale Computing 85
Thiruvengadam Vijayaraghavany, Yasuko Eckert, Gabriel H. Loh, Michael J. Schulte, Mike Ignatowski, Bradford M. Beckmann, William C. Brantley, Joseph L. Greathouse, Wei Huang, Arun Karunanithi, Onur Kayiran, Mitesh Meswani, Indrani Paul, Matthew Poremba, Steven Raasch, Steven K. Reinhardt, Greg Sadowski, and Vilas Sridharan

BRAVO: Balanced Reliability-Aware Voltage Optimization 97
Karthik Swaminathan, Nandhini Chandramoorthy, Chen-Yong Cher, Ramon Bertran, Alper Buyuktosunoglu, and Pradip Bose

HPCA Session 3B: Cache

Maximizing Cache Performance Under Uncertainty 109
Nathan Beckmann and Daniel Sanchez

SWAP: Effective Fine-Grain Management of Shared Last-Level Caches with Minimum Hardware Support 121
Xiaodong Wang, Shuang Chen, Jeff Setter, and José F. Martínez

A Split Cache Hierarchy for Enabling Data-Oriented Optimizations 133
Andreas Sembrant, Erik Hagersten, and David Black-Schaffer

Fast and Accurate Exploration of Multi-level Caches Using Hierarchical Reuse Distance 145
Rafael K. V. Maeda, Qiong Cai, Jiang Xu, Zhe Wang, and Zhongyuan Tian

HPCA Session 4A: Power, Energy & Large-Scale Computing

Enabling Effective Module-Oblivious Power Gating for Embedded Processors 157
Hari Cherupalli, Henry Duwe, Weidong Ye, Rakesh Kumar, and John Sartori

Application-Specific Performance-Aware Energy Optimization on Android Mobile Devices 169
Karthik Rao, Jun Wang, Sudhakar Yalamanchili, Yorai Wardi, and Ye Handong

Fast Decentralized Power Capping for Server Clusters 181
Reza Azimi, Masoud Badiei, Xin Zhan, Na Li, and Sherief Reda

Random Folded Clos Topologies for Datacenter Networks 193
Cristóbal Camarero, Carmen Martínez, and Ramón Beivide

HPCA Session 4B: Memory

Tiny Directory: Efficient Shared Memory in Many-Core Systems with Ultra-Low-Overhead Coherence Tracking 205
Sudhanshu Shukla and Mainak Chaudhuri

Partial Row Activation for Low-Power DRAM System 217
Yebin Lee, Hyeonggyu Kim, Seokin Hong, and Soontae Kim

Understanding and Optimizing Power Consumption in Memory Networks 229
Xun Jian, Pavan Kumar Hanumolu, and Rakesh Kumar

SoftMC: A Flexible and Practical Open-Source Infrastructure for Enabling Experimental DRAM Studies 241
Hasan Hassan, Nandita Vijaykumar, Samira Khan, Saugata Ghose, Kevin Chang, Gennady Pekhimenko, Donghyuk Lee, Oguz Ergin, and Onur Mutlu

HPCA Session 5A: NOC

Static Bubble: A Framework for Deadlock-Free Irregular On-chip Topologies 253
Aniruddh Ramrakhyani and Tushar Krishna

Designing Low-Power, Low-Latency Networks-on-Chip by Optimally Combining Electrical and Optical Links 265
Sebastian Werner, Javier Navaridas, and Mikel Luján

Near-Ideal Networks-on-Chip for Servers 277
Pejman Lotfi-Kamran, Mehdi Modarressi, and Hamid Sarbazi-Azad

Design and Evaluation of AWGR-Based Photonic NoC Architectures for 2.5D Integrated High Performance Computing Systems 289
Paolo Grani, Roberto Proietti, Venkatesh Akella, and S. J. Ben Yoo

HPCA Session 5B: Security

Secure Dynamic Memory Scheduling Against Timing Channel Attacks 301
Yao Wang, Benjamin Wu, and G. Edward Suh

Cold Boot Attacks are Still Hot: Security Analysis of Memory Scramblers in Modern Processors 313
Salessawi Ferede Yitbarek, Misiker Tadesse Aga, Reetuparna Das, and Todd Austin

Cooperative Path-ORAM for Effective Memory Bandwidth Sharing in Server Settings 325
Rujia Wang, Youtao Zhang, and Jun Yang

Camouflage: Memory Traffic Shaping to Mitigate Timing Attacks 337
Yanqi Zhou, Sameer Wagh, Prateek Mittal, and David Wentzlaff

HPCA Session 6A: Emerging Storage

SILC-FM: Subblocked InterLeaved Cache-Like Flat Memory Organization 349
Jee Ho Ryoo, Mitesh R. Meswani, Andreas Prodromou, and Lizy K. John

ATOM: Atomic Durability in Non-volatile Memory through Hardware Logging 361
Arpit Joshi, Vijay Nagarajan, Stratis Viglas, and Marcelo Cintra

KAML: A Flexible, High-Performance Key-Value SSD 373
Yanqin Jin, Hung-Wei Tseng, Yannis Papakonstantinou, and Steven Swanson

Balancing Performance and Lifetime of MLC PCM by Using a Region Retention Monitor 385
Mingzhe Zhang, Lunkai Zhang, Lei Jiang, Zhiyong Liu, and Frederic T. Chong

HPCA Session 6B: Scheduling

Reliability-Aware Scheduling on Heterogeneous Multicore Processors 397
Ajeya Naithani, Stijn Eyerman, and Lieven Eeckhout

Hipster: Hybrid Task Manager for Latency-Critical Cloud Workloads 409
Rajiv Nishtala, Paul Carpenter, Vinicius Petrucci, and Xavier Martorell

Cooper: Task Colocation with Cooperative Games 421
Qiuyun Llull, Songchun Fan, Seyed Majid Zahedi, and Benjamin C. Lee

MemPod: A Clustered Architecture for Efficient and Scalable Migration in Flat Address Space Multi-level Memories 433
Andreas Prodromou, Mitesh Meswani, Nuwan Jayasena, Gabriel Loh, and Dean M. Tullsen

HPCA Session 7A: Novel Architectures

Exploring Hyperdimensional Associative Memory 445
Mohsen Imani, Abbas Rahimi, Deqian Kong, Tajana Rosing, and Jan M. Rabaey

GraphPIM: Enabling Instruction-Level PIM Offloading in Graph Computing Frameworks 457
Lifeng Nai, Ramyad Hadidi, Jaewoong Sim, Hyojong Kim, Pranith Kumar, and Hyesoon Kim

High-Bandwidth Low-Latency Approximate Interconnection Networks 469
Daichi Fujiki, Kiyo Ishii, Ikki Fujiwara, Hiroki Matsutani, Hideharu Amano, Henri Casanova, and Michihiro Koibuchi

Compute Caches 481
Shaizeen Aga, Supreet Jeloka, Arun Subramaniyan, Satish Narayanasamy, David Blaauw, and Reetuparna Das

HPCA Session 7B: Control-Flow and Microarchitecture

Boomerang: A Metadata-Free Architecture for Control Flow Delivery 493
Rakesh Kumar, Cheng-Chieh Huang, Boris Grot, and Vijay Nagarajan

PABST: Proportionally Allocated Bandwidth at the Source and Target 505
Derek R. Hower, Harold W. Cain, and Carl A. Waldspurger

SOUP-N-SALAD: Allocation-Oblivious Access Latency Reduction with Asymmetric DRAM Microarchitectures 517
Yuhwan Ro, Hyunyoon Cho, Eojin Lee, Daejin Jung, Young Hoon Son, Jung Ho Ahn, and Jae W. Lee

Transparent and Efficient CFI Enforcement with Intel Processor Trace 529
Yutao Liu, Peitao Shi, Xinran Wang, Haibo Chen, Binyu Zang, and Haibing Guan

HPCA Session 8A: Accelerators

PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning 541
Linghao Song, Xuehai Qian, Hai Li, and Yiran Chen

FlexFlow: A Flexible Dataflow Accelerator Architecture for Convolutional Neural Networks 553
Wenyan Lu, Guihai Yan, Jiajun Li, Shijun Gong, Yinhe Han, and Xiaowei Li

Needle: Leveraging Program Analysis to Analyze and Extract Accelerators from Whole Programs 565
Snehasish Kumar, Nick Sumner, Vijayalakshmi Srinivasan, Steve Margerm, and Arrvindh Shriraman

Radiation-Induced Error Criticality in Modern HPC Parallel Accelerators 577
Daniel Alfonso Goncalves De Oliveira, Laercio Lima Pilla, Mauricio Hanzich, Vinicius Fratin, Fernando Fernandes, Caio Lunardi, José María Cela, Philippe Olivier Alexandre Navaux, Luigi Carro, and Paolo Rech

HPCA Session 8B: Best of CAL

HPCA Session 9A: GPU Power & Energy

Pilot Register File: Energy Efficient Partitioned Register File for GPUs 589
Mohammad Abdel-Majeed, Alireza Shafaei, Hyeran Jeon, Massoud Pedram, and Murali Annavaram

G-Scalar: Cost-Effective Generalized Scalar Execution Architecture for Power-Efficient GPUs 601
Zhenhong Liu, Syed Gilani, Murali Annavaram, and Nam Sung Kim

Dynamic GPGPU Power Management Using Adaptive Model Predictive Control 613
Abhinandan Majumdar, Leonardo Piga, Indrani Paul, Joseph L. Greathouse, Wei Huang, and David H. Albonesi

HPCA Session 9B: GPU

Efficient Sequential Consistency in GPUs via Relativistic Cache Coherence 625
Xiaowei Ren and Mieszko Lis

Processing-in-Memory Enabled Graphics Processors for 3D Rendering 637
Chenhao Xie, Shuaiwen Leon Song, Jing Wang, Weigong Zhang, and Xin Fu

Controlled Kernel Launch for Dynamic Parallelism in GPUs 649
Xulong Tang, Ashutosh Pattnaik, Huaipan Jiang, Onur Kayiran, Adwait Jog, Sreepathi Pai, Mohamed Ibrahim, Mahmut T. Kandemir, and Chita R. Das

Author Index 661