

Gennady Pekhimenko

CONTACT INFORMATION	University of Toronto Computer Science Dept. (BA 5232) 40 St. George Street Toronto ON, M5S2E4	Work: (+1) 416-946-0250 Mobile: (+1) 647-916-6900 E-mail: pekhimenko@cs.toronto.edu Webpage: www.cs.toronto.edu/~pekhimenko/
RESEARCH INTERESTS	Systems, Computer Architecture, Machine Learning, Data Compression, Approximate Computing, Compilers, Bioinformatics	
EDUCATION	Carnegie Mellon University, USA <i>PhD in Computer Science, Computer Science Dept.</i> Thesis: “Practical Data Compression for Modern Memory Hierarchies” Advisors: Todd C. Mowry and Onur Mutlu <i>July 2016</i>	
	University of Toronto, Canada <i>MS in Computer Science</i> Department of Computer Science Thesis: “Machine Learning Algorithms for Choosing Compiler Heuristics” Advisor: Angela Demke Brown <i>Jan 2008</i>	
	Taurida National University, Ukraine <i>BS in Biology, Biochemistry (Part-time)</i> Faculty of Biology and Chemistry <i>May 2006</i>	
	Moscow State University, Russia <i>Diploma (5-year program) in Applied Mathematics & Computer Science</i> Faculty of Computational Mathematics and Cybernetics, Department of System Programming Thesis: “Performance Analysis of MPI-Programs” Advisor: Victor A. Krukov <i>May 2004</i>	
APPOINTMENT	Vector Institute, Canada <i>Faculty Member, Vector Institute</i> <i>Aug 2019 – present</i> <i>Faculty Affiliate, Vector Institute</i> <i>May 2018 – Aug 2019</i>	
	University of Toronto, Canada <i>Assistant Professor, Computer Science Department</i> <i>June 2017 – present</i> <i>Assistant Professor, Electrical & Computer Engineering Dept.</i> <i>Jan 2018 – present</i>	
AWARDS & HONORS	<ul style="list-style-type: none">◇ Connaught New Researcher Award \$ 10000 CAD award <i>2018 – 2020</i>◇ NVIDIA Graduate Fellowship 5 winners nation-wide <i>2015 – 2016</i>◇ First place in ACM SRC (Student Research Competition) Energy-Efficient Data Compression for GPU Memory Systems @ ASPLOS’15 <i>Mar 2015</i>◇ Qualcomm Innovation Fellowship Finalist Together with Nandita Vijaykumar. Selected as one of 35 out of 146 teams <i>2015 – 2016</i>◇ Facebook Fellowship Finalist \$500 cash prize <i>2015 – 2016</i>	

- ◇ **Microsoft Research Fellowship** 2013 – 2015
12 winners nation-wide
- ◇ **Qualcomm Innovation Fellowship** 2013 – 2014
Together with Chris Fallin. 10 winner teams nation-wide
- ◇ **First Heidelberg Laureate Forum Invitation** Sep 2013
Young researcher of the US delegation
- ◇ **Second place in ACM SRC (Student Research Competition)** Sep 2012
Linearly Compressed Pages: A Main Memory Compression Framework
with Low Complexity and Low Latency @ PACT'12
- ◇ **Alexander Graham Bell Canada Graduate Scholarship** 2012 – 2013
NSERC (Canada's NSF) CGS-D2 Scholarship
- ◇ **IBM First Patent Application Award Achievement** Jan 2010
\$2000 cash prize
- ◇ **Wolfond Scholarship** 2006 – 2007
University of Toronto Scholarship for high academic achievements
- ◇ **Best Student Award** Apr 2003
Selected as the best student in MSU, CS Department

PEER-REVIEWED
CONFERENCE
PUBLICATIONS

24. Sihang Liu, Korakit Seemakhupt, Gennady Pekhimenko, Aasheesh Kolli, and Samira Khan.
Janus: Optimizing Memory and Storage Support for Non-Volatile Memory Systems. ACM/IEEE International Symposium on Computer Architecture (**ISCA'19**). June 2019.
23. Hongyu Miao, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S. McKinley, and Felix Xiaozhu Lin.
StreamBox-HBM: Stream Analytics on High Bandwidth Hybrid Memory. ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS'19**). April 2019.
22. Anand Jayarajan, Jinliang Wei, Garth Gibson, Alexandra Fedorova, and Gennady Pekhimenko.
Priority-based Parameter Propagation for Distributed DNN Training. SysML Conference (**SysML'19**). April 2019.
21. Hongyu Zhu, Mohamed Akrouf, Bojian Zheng, Andrew Pelegris, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko.
Benchmarking and Analyzing Deep Neural Network Training. IEEE International Symposium on Workload Characterization (**IISWC'18**). October 2018.
20. Gennady Pekhimenko, Chuanxiong Guo, Myeongjae Jeon, Ryan Huang, and Lidong Zhou.
TerseCades: Efficient Data Compression in Stream Processing. USENIX Annual Technical Conference (**ATC'18**). July 2018.
19. Animesh Jain, Amar Phanishayee, Jason Mars, Lingjia Tang, and Gennady Pekhimenko.
Gist: Efficient Data Encoding for Deep Neural Network Training. International Symposium on Computer Architecture (**ISCA'18**). June 2018.
18. Nandita Vijaykumar, Abhilasha Jain, Diptesh Majumdar, Kevin Hsieh, Gennady Pekhimenko, Eiman Ebrahimi, Nastaran Hajinazaran, Phillip B. Gibbons, and Onur Mutlu .
A Case for Richer Cross-layer Abstractions: Bridging the Semantic Gap to Enhance Memory Optimization. International Symposium on Computer Architecture (**ISCA'18**). June 2018.
17. Hongyu Zhu, Bojian Zheng, Amar Phanishayee, Bianca Schroeder, and Gennady Pekhimenko.
DNN-Train: Benchmarking and Analyzing DNN Training. SysML Conference (**SysML'18**). February 2018.
16. Hongyu Miao, Heejin Park, Myeongjae Jeon, Gennady Pekhimenko, Kathryn S. McKinley, and Felix Xiaozhu Lin.
StreamBox: Modern Stream Processing on a Multicore Machine. USENIX Annual Tech-

- nical Conference (**ATC'17**). July 2017.
15. Donghyuk Lee, Samira Khan, Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Gennady Pekhimenko, Vivek Seshadri, and Onur Mutlu.
Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms. ACM SIGMETRICS / IFIP Performance (**SIGMETRICS'17**). June 2017.
 14. Hasan Hassan, Nandita Vijaykumar, Samira Khan, Saugata Ghose, Kevin Chang, Gennady Pekhimenko, Donghyuk Lee, Oguz Ergin, and Onur Mutlu.
SoftMC: A Flexible and Practical Infrastructure for Enabling Experimental DRAM Studies. International Symposium on High-Performance Computer Architecture (**HPCA'17**). February 2017
 13. Nandita Vijaykumar, Kevin Hsieh, Gennady Pekhimenko, Samira Khan, Ashish Shrestha, Saugata Ghose, Adwait Jog, Phillip B. Gibbons, Onur Mutlu.
Zorua: A Holistic Approach to Resource Virtualization in GPUs. International Symposium on Microarchitecture (**MICRO'16**). October 2016.
 12. Kevin Chang, Abhijith Kashyap, Hasan Hassan, Saugata Ghose, Kevin Hsieh, Donghyuk Lee, Tianshi Li, Gennady Pekhimenko, Samira Khan, Onur Mutlu.
Understanding Latency Variation in Modern DRAM Chips: Experimental Characterization, Analysis, and Optimization. ACM SIGMETRICS / IFIP Performance (**SIGMETRICS'16**). June 2016.
 11. Gennady Pekhimenko, Evgeny Bolotin, Nandita Vijaykumar, Onur Mutlu, Todd C. Mowry, Stephen W. Keckler.
Toggle-Aware Bandwidth Compression for GPUs. International Symposium on High-Performance Computer Architecture (**HPCA'16**). March 2016.
 10. Hasan Hassan, Gennady Pekhimenko, Nandita Vijaykumar, Vivek Seshadri, Donghyuk Lee, Oguz Ergin, Onur Mutlu.
ChargeCache: Reducing DRAM Latency by Exploiting Row Access Locality. International Symposium on High-Performance Computer Architecture (**HPCA'16**). March 2016.
 9. Vivek Seshadri, Gennady Pekhimenko, Olatunji Ruwase, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, Todd C. Mowry, Trishul Chilimbi.
Page Overlays: An Enhanced Virtual Memory Framework to Enable Fine-grained Memory Management. International Symposium on Computer Architecture (**ISCA'15**). June 2015.
 8. Nandita Vijaykumar, Gennady Pekhimenko, Adwait Jog, Abhishek Bhowmick, Rachata Ausavarungnirun, Onur Mutlu, Chita R. Das, Mahmut T. Kandemir, Todd C. Mowry.
A Case for Core-Assisted Bottleneck Acceleration in GPUs: Enabling Efficient Data Compression. International Symposium on Computer Architecture (**ISCA'15**). June 2015.
 7. Gennady Pekhimenko, Dimitrios Lymberopoulos, Oriana Riva, Karin Strauss, Doug Burger.
PocketTrend: Architecting Search Engines for Trending Topics. International World Wide Web Conference (**WWW'15**). May 2015.
 6. Gennady Pekhimenko, Tyler Hubery, Rui Cai, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, Todd C. Mowry.
Exploiting Compressed Block Size as an Indicator of Future Reuse. International Symposium on High-Performance Computer Architecture (**HPCA'15**). February 2015.
 5. Donghyuk Lee, Yoongu Kim, Gennady Pekhimenko, Samira Khan, Vivek Seshadri, Kevin Chang, Onur Mutlu.
Adaptive-Latency DRAM: Optimizing DRAM Timing for the Common-Case. International Symposium on High-Performance Computer Architecture (**HPCA'15**). February 2015.
 4. Bradley Thwaites, Gennady Pekhimenko, Amir Yazdanbakhsh, Girish Mururu, Jongse Park, Hadi Esmaeilzadeh, Onur Mutlu, Todd C. Mowry.

Rollback-Free Value Prediction with Approximate Loads. International Conference on Parallel Architectures and Compilation Techniques (**PACT'14, Short Paper**). August 2014.

3. Gennady Pekhimenko, Vivek Seshadri, Yoongu Kim, Hongyi Xin, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, Todd C. Mowry.
Linearly Compressed Pages: A Low-Complexity, Low-Latency Main Memory Compression Framework. International Symposium on Microarchitecture (**MICRO'13**). December 2013.
2. Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Yixin Luo, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, Todd C. Mowry.
RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization. International Symposium on Microarchitecture (**MICRO'13**). December 2013.
1. Gennady Pekhimenko, Vivek Seshadri, Onur Mutlu, Phillip B. Gibbons, Michael A. Kozuch, Todd C. Mowry.
Base-Delta-Immediate Compression: Practical Data Compression Mechanism for On-Chip Caches. International Conference on Parallel Architectures and Compilation Techniques (**PACT'12**). September 2012.
8. Donghyuk Lee, Samira Manabi Khan, Lavanya Subramanian, Saugata Ghose, Rachata Ausavarungnirun, Gennady Pekhimenko, Vivek Seshadri, Onur Mutlu.
Design-Induced Latency Variation in Modern DRAM Chips: Characterization, Analysis, and Latency Reduction Mechanisms. **POMACS: Proceedings of the ACM on Measurement and Analysis of Computing Systems**. 2017.
7. Hongyi Xin, Richard Zhu, Sunny Nahar, John Emmons, Gennady Pekhimenko, Carl Kingsford, Can Alkan, Onur Mutlu.
Optimal Seed Solver: Optimizing Seed Selection in Read Mapping. **Oxford Bioinformatics**. 2016.
6. Amir Yazdanbakhsh, Gennady Pekhimenko, Bradley Thwaites, Girish Mururu, Jongse Park, Hadi Esmaeilzadeh, Onur Mutlu, Todd C. Mowry.
RFVP: Rollback-Free Value Prediction with Approximate Memory Loads. ACM Transactions on Architecture and Code Optimization (**TACO'16**). 2015.
5. Donghyuk Lee, Saugata Ghose, Gennady Pekhimenko, Samira Khan, Onur Mutlu.
Simultaneous Multi Layer Access: A High Bandwidth and Low Cost 3D-Stacked Memory Interface. ACM Transactions on Architecture and Code Optimization (**TACO'16**). 2015.
4. Amir Yazdanbakhsh, Gennady Pekhimenko, Bradley Thwaites, Girish Mururu, Jongse Park, Hadi Esmaeilzadeh, Onur Mutlu, Todd C. Mowry.
Mitigating the Bandwidth Bottleneck with Approximate Load Value Prediction. **IEEE Design & Test**. 2016.
3. Gennady Pekhimenko, Evgeny Bolotin, Mike O'Connor, Onur Mutlu, Todd C. Mowry, Stephen W. Keckler.
Toggle-Aware Compression for GPUs. IEEE Computer Architecture Letters (**CAL'15**). May 2015.
2. Hongyi Xin, John Greth, John Emmons, Gennady Pekhimenko, Carl Kingsford, Can Alkan, Onur Mutlu.
Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter for Local Alignment in Read Mapping. **Oxford Bioinformatics**. January 2015.
1. Gennady Pekhimenko, Angela Demke Brown.
Software Automatic Tuning: From Concepts to State-of-the-Art Results, Chapter 19. **Springer**. September 2010.

JOURNALS &
BOOK CHAPTERS

OTHER
PEER-REVIEWED
PUBLICATIONS

4. Bojian Zheng and Gennady Pekhimenko.
EcoRNN: Efficient Computing of LSTM RNN on GPUs. Student Research Competition at IEEE/ACM International Symposium on Microarchitecture (**SRC@MICRO'18**). October 2018.
3. Gennady Pekhimenko, Evgeny Bolotin, Mike O'Connor, Onur Mutlu, Todd C. Mowry, Stephen W. Keckler.
Energy-Efficient Data Compression for GPU Memory Systems. Student Research Competition at International Conference on Architectural Support for Programming Languages and Operating Systems (**SRC@ASPLOS'15**). March 2015.
2. Gennady Pekhimenko, Todd C. Mowry, Onur Mutlu.
Linearly Compressed Pages: A Main Memory Compression Framework with Low Complexity and Low Latency. Student Research Competition at International Conference on Parallel Architectures and Compilation Techniques (**SRC@PACT'12**). September 2012.
1. Gennady Pekhimenko, Angela Demke Brown.
Efficient Program Compilation through Machine Learning Techniques. International Workshop on Automatic Performance Tuning (**iWAPT'09**). October 2009

PATENTS, THESES

5. Gennady Pekhimenko.
Practical Data Compression for Modern Memory Hierarchies. PhD Thesis, Carnegie Mellon University. July 2016.
4. Dimitrios Lymberopoulos, Oriana Riva, Karin Strauss, Doug Burger, Gennady Pekhimenko.
Trend Response Management. Patent No. 20150227517. August 2015.
3. Yaoqing Gao, Tong Chen, Zehra Sura, Gennady Pekhimenko, Kevin O'Brien, Khaled Mohammed, Roch Archambault, Raul Silvera.
Managing Speculative Assist Threads. Patent No. 20110093838. October 2010.
2. Gennady Pekhimenko.
Machine Learning Algorithms for Choosing Compiler Heuristics. MS Thesis, University of Toronto. January 2008.
1. Gennady Pekhimenko.
Performance Analysis of MPI-Programs. Diploma Thesis, Moscow State University, Russia. May 2004.

GRANTS

- ◇ Mitacs, Accelerate, "Next Generation AI Accelerator Algorithm Hardware Co-Optimization", **\$30,000**. 2019–2020
- ◇ Huawei, Research Grant, "Efficient Data Compression/Deduplication for Persistent Memory and DRAM", **\$426,000**. 2019–2022
- ◇ IBM Canada, CAS Program, "Efficient Compiler-Driven Pointer Compression", Award #1112, **\$102,000**. 2019–2022
- ◇ Huawei, Research Grant, "Compiler Infrastructure for Optimizing DNN Workloads", **\$289,300**. 2019–2022
- ◇ NSERC Discovery Grant (Increase) "Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications", **\$12,500 total**. 2018–2023
- ◇ NSERC CRD, "Efficient Memory Footprint Reduction for Java Performance", **\$204,000**. 2019–2022
- ◇ Huawei, Research Grant, "Efficient Distributed DNN Training", **\$199,546**. 2018–2021
- ◇ NSERC UTEA "Parallelism and Hardware Heterogeneity Support in Modern Compilers", UTEA: Qiongsi Wu, **\$4,875 total**. 2018–2018
- ◇ Connaught New Researcher Award "Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications", Connaught Fund, **\$10,000 total**. 2018–2019
- ◇ NSERC Discovery Accelerator Supplement Grant "Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications", NSERC (522575), **\$120K total**.

2018–2021

- ◇ NSERC Discovery Grant “Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications”, NSERC (RGPIN-2018-06514), **\$140K total.** 2018–2023
- ◇ IBM Canada, CAS Program, “Efficient Memory Footprint Reduction for Java Performance”, Award #1063, **\$102,000.** 2018–2021
- ◇ Huawei, HiRP Open Program, “Hardware/Software Optimization and Compiler Support for Heterogeneous Systems”, **\$87,044.** 2017–2019
- ◇ Canada Foundation for Innovation (CFI), John Evans Leaders Fund program “Heterogeneous Systems Laboratory”, CFI (Award #36585), **\$240K total.** 2017–2020

ADVISING

Current:

- ◇ Hongyu Zhu, PhD Student. DNN Profiling and Analysis.
- ◇ Pavel Klishin, PhD student. DNN training acceleration with FPGAs.
- ◇ Bojian Zheng, PhD student. Hardware acceleration for LSTM-based RNNs.
- ◇ Alexandra Tsvetkova, PhD Student. Software support for GPU virtualization.
- ◇ James Gleeson, PhD Student (co-advised with Eyal de Lara). Optimizing reinforcement learning training.
- ◇ Anand Jayarajan, PhD Student.
- ◇ Shang (Sam) Wang, MSc. student. DNN training parallelization.
- ◇ Geoffrey Yu, MSc. student. ML auto-tuning.
- ◇ Xiaodan (Serina) Tan, MASc. student. Hardware support for GPU virtualization.
- ◇ Qionsi Wu, MSc. student. Compiler support for multi-threading with OpenMP.
- ◇ Hanjie Qiu, MSc. student. Efficient data compression/deduplication for non-volatile memories.
- ◇ Jiacheng Yang, MASc. student.
- ◇ Izaak Niksan, BSc. student. Memory profiler for DNN training.
- ◇ Jiahuang (Jacob) Lin, MScAC student. Speech recognition using DeepSpeech2 model for inference.
- ◇ Pavel Golikov, BSc. student. DNNs performance modeling on modern GPUs.

Graduated:

- ◇ Andrew Pelegris, MSc. student. Binarized DNNs acceleration.
- ◇ Mohamed Akrouf, MScAC student (2018). Reinforcement learning profiling and training.
- ◇ Ming (Michael) Yang, BSc. student (2019). New simulator infrastructure for GPUs.
- ◇ Yifan Bai, BSc. student (2019). Jacobian-based approach for DNN training.
- ◇ Kuei-Fang (Albert) Hsueh, BSc. student (2019). Machine translation using Transformer model for inference.
- ◇ Akshay Nair, BSc. student (2018). Simulation infrastructure for GPUs.

MENTORING

CMU (PhD, Masters and undergraduate):

- ◇ Amir Yazdanbakhsh, PhD Student. Research Project: Rollback-Free Value Prediction with Approximate Loads.
- ◇ Hasan Hassan, Masters student. Research Project: Reducing DRAM Latency by Exploiting Row Access Locality.
- ◇ Mahmoud Khairy, Masters student. Research Project: Efficient DRAM Refresh for GPUs.
- ◇ Arthur Perais, PhD Student. Research Project: Synergy Analysis Between Value Prediction and Data Compression.
- ◇ Hongyi Xin, PhD Student. Research Project: Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter for Local Alignment in Read Mapping.
- ◇ Nandita Vijaykumar, PhD Student. Research Project: Core-Assisted Bottleneck Acceleration.
- ◇ Abhishek Bhowmick, Undergraduate student (currently Masters student at CMU). Research Project: GPU Main Memory Compression and Prefetching.

- ◇ Tyler Huberty and Rui Cai, Undergraduate students (currently at Apple and Microsoft). Research Project: CARP: Compression-Aware Replacement Policies.
- ◇ Jason Lin and Brian Osbun, Undergraduate students (currently at Microsoft and CMU). Research Project: Bandwidth-Optimized Prefetching.
- ◇ Martyn Romanko and Lei Fan, Masters students (currently at Intel). Research Project: Implementation and Energy Analysis of Base-Delta-Immediate Compression.

WORK
EXPERIENCE

- ◇ Assistant Professor at the **University of Toronto**, CS Department, *July 2017 – Present*
- ◇ Assistant Professor at the **University of Toronto**, ECE Department (by courtesy), *January 2018 – Present*
- ◇ Researcher at **Microsoft Research** with Systems Research Group, *July 2016 – Aug 2017*
- ◇ Graduate Student Researcher at **Carnegie Mellon University** with Prof. Todd C. Mowry and Prof. Onur Mutlu, *Sep 2010 – August 2016*
- ◇ Research Consultant at **Microsoft** with Dr. Marc Tremblay, *Feb 2015 – Jul 2015*
- ◇ Research Intern at **NVIDIA Research** with Dr. Stephen Keckler and Dr. Evgeny Bolotin, *Summer 2014*
- ◇ Research Intern at **Microsoft Research** with Dr. Karin Strauss, Dr. Dimitrios Lybmeropoulos, Dr. Oriana Riva, *Summer 2013*
- ◇ Research Intern at **Microsoft Research** with Dr. Ella Bounimova, Dr. Patrice Godefroid, and Dr. David Molnar, *Summer 2012*
- ◇ Compiler Engineer/Researcher (Full-time) at **IBM** with Raul Silvera and Yaoging Gao, *May 2007 – Jun 2010*
- ◇ Graduate Student Researcher at the **University of Toronto** with Prof. Angela Demke Brown, *Sep 2006 – Jan 2008*
- ◇ Compiler Engineer (Full-time) at **Elbrus, Moscow, Russia** with Dr. Vladimir Volkonskii, *May 2004 – Aug 2006*
- ◇ System Programmer at **Intel-MSU Lab, Moscow, Russia** with Prof. Viktor Krukov, *May 2003 – Jun 2004*

TEACHING
EXPERIENCE

- ◇ **Instructor** at the University of Toronto, CSC D70H: Compiler Optimization, Undergraduate *Winter 2019, 2018*
- ◇ **Instructor** at the University of Toronto, CSC 2224H: Parallel Computer Architecture and Programming, Graduate *Fall 2018, 2017*
- ◇ **Teaching Assistant** at Carnegie Mellon University, Optimizing Compilers, Graduate *Spring 2012*
- ◇ **Teaching Assistant** at Carnegie Mellon University, Introduction to Computer Systems, Undergraduate *Fall 2011*
- ◇ **Teaching Assistant** at the University of Toronto, Operating Systems, Undergraduate *Fall 2007, Spring 2008*
- ◇ **Teaching Assistant** at the University of Toronto, Computer Programming, Undergraduate *Spring 2007*
- ◇ **Teaching Assistant** at the University of Toronto, Software Engineering, Undergraduate *Fall 2006*

INVITED TALKS

- 51. *ML Performance: Benchmarking Deep Learning Systems* ISCA'19 Tutorial, Phoenix, Ar. *June 2019*
- 50. *ML Performance: Benchmarking Deep Learning Systems* ASPLOS'19 Tutorial, Providence, RI. *April 2019*
- 49. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations* Google Platform Team, Sunnyvale, CA. *April 2019*
- 48. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations* FastPath'19 Workshop Keynote, Madison, WI. *March 2019*

47. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Apple, Cupertino, CA. December 2018
46. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Facebook, Menlo Park, CA. December 2018
45. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Google, Mountain View, CA. December 2018
44. *Algorithms vs. Architectures: Rivals or Partners in Pushing AI Boundaries?*
Huawei AI Workshop, Shanghai, China. October 2018
43. *TerseCades: Efficient Data Compression in Stream Processing*
USENIX ATC'18, Boston, MA. July 2018
42. *Benchmarking and Analyzing DNN Training*
Vector Institute, Toronto, ON. May 2018
41. *Benchmarking and Analyzing DNN Training*
SysML'18, Stanford, CA. Feb 2018
40. *Practical Data Compression for Memory Hierarchy and DNNs*
Yandex, Moscow, Russia. Nov 2017
39. *A Case for Toggle-Aware Compression for GPU Systems*
HPCA-22, Barcelona, Spain. Mar 2016
38. *RFVP: Rollback-Free Value Prediction with Safe-to-Approximate Loads*
HiPEAC16, Prague, Czech Republic. Jan 2016
37. *Linearly Compressed Pages*
University of Texas at Austin, Austin, TX. Nov 2015
36. *Exploiting Compressed Block Size as an Indicator of Future Reuse*
PDL Retreat, Bedford, PA. Oct 2015
35. *Linearly Compressed Pages*
University of Illinois at Urbana-Champaign, Urbana, IL. Oct 2015
34. *Base-Delta-Immediate Compression*
University of Alberta, Edmonton, Canada. Sep 2015
33. *PocketTrend: Timely Identification and Delivery of Trending Search Content to Mobile Users*
WWW-24, Florence, Italy. May 2015
32. *Exploiting Compressed Block Size as an Indicator of Future Reuse*
MIT, Boston, MA. May 2015
31. *Energy-Efficient Data Compression for Modern Memory Systems*
QInF Finals, San Diego, CA. Mar 2015
30. *Energy-Efficient Data Compression for GPU Memory Systems*
SRC@ASPLOS'15, Istanbul, Turkey. **First Place in ACM SRC Competition** Mar 2015
29. *Exploiting Compressed Block Size as an Indicator of Future Reuse*
Intel Atom Group, Hillsboro, OR. Feb 2015
28. *Exploiting Compressed Block Size as an Indicator of Future Reuse*
HPCA-21, Bay Area, CA. Feb 2015
27. *Energy-Efficient Data Compression*
Qualcomm, San Diego, CA. Sep 2014
26. *Energy-Efficient Data Compression For GPU Memory Systems*
NVIDIA Research, Santa Clara, CA. Sep 2014
25. *Linearly Compressed Pages*
Intel Labs, Santa Clara, CA. Sep 2014
24. *Energy-Efficient Data Compression*
UC Berkeley, ASPIRE Lab, Berkeley, CA. Sep 2014
23. *Linearly Compressed Pages*
Huawei R&D, Santa Clara, CA. Aug 2014
22. *Energy-Efficient Data Compression*
NVIDIA Research, Santa Clara, CA. Jul 2014
21. *Guest Lecture on Cache Compression*

- 18447: Introduction to Computer Architecture, Pittsburgh, PA. Apr 2014
20. *Linearly Compressed Pages*
CMU Cloud Workshop, Pittsburgh, PA. Apr 2014
 19. *Linearly Compressed Pages*
Samsung Research, San Jose, CA. Dec 2013
 18. *Linearly Compressed Pages*
Oracle Labs, Belmont, CA. Dec 2013
 17. *Linearly Compressed Pages: A Low-Complexity, Low-Latency Main Memory Compression Framework*
MICRO-46, Davis, CA. Dec 2013
 16. *Linearly Compressed Pages*
Stanford Cloud Workshop, Mountain View, CA. Dec 2013
 15. *Linearly Compressed Pages*
NVIDIA Research, Santa Clara, CA. Dec 2013
 14. *Main Memory Compression and Low-Cost Compression Algorithms*
PDL Retreat, Bedford, PA. Oct 2013
 13. *In-Memory Optimizations: Efficient Compression and Data Movement*
Heidelberg Laureate Forum, Heidelberg, Germany. Sep 2013
 12. *PocketTrend: Efficient Trend Detection for Mobile Devices*
Microsoft Research, Redmond, WA. Aug 2013
 11. *Base-Delta-Immediate Compression*
Microsoft Research, Redmond, WA. Jul 2013
 10. *Base-Delta-Immediate Compression*
University of Toronto, Ontario, Canada. Mar 2013
 9. *Heterogeneous Block Architectures*
Qualcomm, San Diego, CA. Mar 2013
 8. *Linearly Compressed Pages*
Intel, Hillsboro, OR. Feb 2013
 7. *Base-Delta-Immediate Compression*
Intel Labs, Hillsboro, OR. Feb 2013
 6. *Guest Lecture on Caching in Multi-Core Systems*
18742: Parallel Computer Architecture, Pittsburgh, PA. Oct 2012
 5. *Base-Delta-Immediate Compression: Practical Data Compression Mechanism for On-Chip Caches*
PACT, Minneapolis, MN. Sep 2012
 4. *Linearly Compressed Pages: A Main Memory Compression Framework with Low Complexity and Low Latency*
SRC@PACT, Minneapolis, MN. **Second Place in ACM SRC Competition** Sep 2012
 3. *Guest Lecture on Dynamic Compilation*
15745: Optimizing Compilers, Pittsburgh, PA. Feb 2012
 2. *Assist Threads for Data Prefetching in IBM XL Compilers*
CASCON, Toronto, ON. Nov 2009
 1. *Efficient Program Compilation through Machine Learning Techniques*
International Workshop on Automatic Performance Tuning, Tokyo, Japan. Oct 2009

SERVICE

Program and Organization Committees

- ◇ **PC Member**, SysML 2020 2019–2020
- ◇ **PC Member**, EuroSys 2020 2019–2020
- ◇ **Co-Chair**, Artifact Evaluation at SysML 2020 2019–2020
- ◇ **Publicity Co-Chair**, HPCA 2020 2019–2020
- ◇ **PC Member**, CGO 2020 2019–2020
- ◇ **PC Member**, MICRO 2019 2019
- ◇ **PC Member**, ICS 2019 2018–2019
- ◇ **Tutorial Organizer**, MLPerfBench at ISCA 2019 2019

	◇ PC Member , ISCA 2019	2018–2019
	◇ PC Member , SysML 2019	2018–2019
	◇ Co-Chair , Artifact Evaluation at SysML 2019	2018–2019
	◇ ERC (External Review Committee) Member , ASPLOS 2019	2018–2019
	◇ ERC (External Review Committee) Member , HPCA 2019	2018–2019
	◇ Program Co-Chair , Compiler-Driven Performance Workshop	2018
	◇ PC Member , MICRO 2018	2018
	◇ PC Member , ICS 2018	2017–2018
	◇ Publicity Co-Chair , ASPLOS 2018	2017–2018
	◇ ERC (External Review Committee) Member , MICRO 2017	2017
	◇ ERC (External Review Committee) Member , ISCA 2017	2016–2017
	◇ Web Chair , ISCA 2017	2016–2017
	◇ PC Member , ICWE 2017	2016–2017
	◇ ERC (External Review Committee) Member , ISCA 2016	2015–2016
	◇ PC Member , WWW 2016	2015–2016
	◇ Web Chair , ASPLOS 2016	2015–2016
	◇ Publicity Chair , HiPEAC 2015	2015
	◇ Information Director , Transactions on Computer Systems (TOCS)	2013–2017
	Reviewer	
	◇ ISCA	2011–2016
	◇ MICRO	2011–2015
	◇ ASPLOS	2012, 2016, 2017
	◇ HPCA	2012–2017
	◇ PACT	2013, 2014
	◇ DAC	2014–2015
	◇ DATE	2016
	◇ IISWC	2014
	◇ ICCD	2014
	◇ NOCS	2012
	◇ MICRO Top Picks	2012, 2013, 2015
	◇ Transactions on Architecture and Code Optimization (TACO)	2015
	◇ Computer Architecture Letters (CAL)	2015, 2016
	◇ Transactions on Parallel and Distributed Systems (TPDS)	2014–2015
	◇ Transactions on Computers (TC)	2013, 2015
	◇ Transactions on Very Large Scale Integration Systems (TVLSI)	2011–2012
	◇ ICAC	2013
ACTIVITIES	◇ CSD Speaking Skills Committee	2013–2016
	◇ Student Volunteer for CS Open House	2012–2013
	◇ CALCM Seminar Czar	2013
PROFESSIONAL MEMBERSHIPS	◇ IEEE Computer Society	2014–present
	◇ Association of Computing Machinery (ACM)	2012–present
	◇ ACM SIGARCH	2012–present
CITIZENSHIP STATUS	◇ Russian Citizenship	
	◇ Canadian Permanent Resident	