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

My research lies in the general area of computer architecture, systems, and applied machine learning with the research focus on efficient memory systems, systems for machine learning, stream processing, machine learning for systems, compilers, and hardware acceleration.
Keywords: Systems, Computer Architecture, Applied Machine Learning, Stream Processing, Compilers

Section 1: Academic History and Professional Experience

EDUCATION

Carnegie Mellon University, *USA*

PhD in Computer Science, Computer Science Dept. *July 2016*
Thesis: “Practical Data Compression for Modern Memory Hierarchies”
Advisors: Todd C. Mowry and Onur Mutlu

University of Toronto, *Canada*

MS in Computer Science
Department of Computer Science *Jan 2008*
Thesis: “Machine Learning Algorithms for Choosing Compiler Heuristics”
Advisor: Angela Demke Brown

Moscow State University, *Russia*

Diploma (5-year program) in Applied Mathematics & Computer Science
Faculty of Computational Mathematics and Cybernetics,
Department of System Programming *May 2004*
Thesis: “Performance Analysis of MPI-Programs”
Advisor: Victor A. Krukov

CURRENT APPOINTMENTS

University of Toronto, *Canada*

Assistant Professor, Computer Science Department *June 2017 – present*
Assistant Professor, Electrical & Computer Engineering Dept. *Jan 2018 – present*

I’m currently holding a tenure-track position in the Department of Computer Science at University of Toronto at the rank of Assistant Professor. The University of Toronto Department of Computer Science is ranked as the top CS department in Canada and 9th globally, and computer architecture is ranked 8^{t}h in the world (based on csrankings.org). I’m also cross-appointed in the Department of Electrical and Computer Engineering. My responsibilities include pursuing innovative research at the highest international level; establishing a strong, externally funded independent research program; having a strong commitment to undergraduate and graduate teaching; and contributing to the enrichment of both undergraduate and graduate programs in the computer science department.

Vector Institute, *Canada*

Faculty Member, CIFAR AI Chair, Vector Institute *Aug 2019 – present*

Faculty Affiliate, Vector Institute *May 2018 – Aug 2019*

The Vector Institute is an independent, not-for-profit corporation dedicated to research in the field of artificial intelligence (AI), excelling in machine and deep learning. The Vector Institute is launched to drive excellence and leadership in Canada's knowledge, creation, and use of artificial intelligence (AI) to foster economic growth and improve the lives of Canadians. My role as a Canada CIFAR AI Chair and faculty member is to conduct research on building efficient systems to aid the cutting-edge AI research happening in Vector Institute, assist researchers at Vector in optimizing their workloads, and also help in selecting the best equipment for large model training.

CentML, *Toronto, ON*

CEO and Co-Founder *Mar 2022 – Present*

Together with my students and colleagues from industry created a new company, CentML, that offers an optimization platform for ML Training and Inference.

Past appointments

Amazon/AWS, *Toronto, ON*

Amazon Scholar *Sep 2020 – July 2022*

I consulted several teams at Amazon/AWS that develop ML hardware, compilers, frameworks, and new ML applications. I collaborated with both research and development teams in the USA and Canada.

Microsoft Research, *Redmond, WA*

Research Consultant	*July 2017 – June 2018*
Researcher, MSR Systems Research Group	*July 2016 – Aug 2017*
Research Consultant	*Feb 2015 – Jul 2015*
Research Intern	*Summer 2012, 2013*

Microsoft Research is the leading industry research lab, with the major office in Redmond, WA. I performed research in the systems research group full-time, part-time as a consultant, and twice as an intern when doing my PhD at CMU. This work resulted in multiple conference publications and patents.

NVIDIA Research, *Santa Clara, CA*

Research Intern *Summer 2014*

I did an internship at Nvidia Research working with Evgeny Bolotin, Group manager: Stephen Keckler. The work was focused on reducing the GPU bandwidth consumption using specialized compression algorithms and the results were published at HPCA 2016.

IBM, *Toronto, ON*

Compiler Engineer/Researcher (Full-time) *Feb 2007 – Jun 2010*

I worked as a software developer in the XLC/XLF compiler, and contributed to the PERCS project.

Elbrus, *Moscow, Russia*

Compiler Engineer (Full-time) *May 2004 – Aug 2006*

I worked as a software developer on the binary compiler for VLIW-based chips.

Intel-MSU Lab, *Moscow, Russia*

System Programmer *May 2003 – Jun 2004*

Performed research in the areas of parallel computing and optimization with Prof. Viktor Krukov.

Awards and Honors

- ◇ VMware Early Career Faculty Grant/Award, **USD$50,000**. *2022–2023*

- ◇ Google Scholar Research Award, **USD$60,000**. *2022–2023*
  "Holistic Systems Techniques for Efficient Training of Deep Learning Models"

- ◇ **ISCA Hall of Fame** *June 2021*
  at least 8 ISCA papers as an author

- ◇ **IEEE MICRO Top Picks Award** *2020*
  For "MLPerf Inference Benchmark" ISCA 2020 paper

- ◇ **HiPEAC Paper Award** *2020*
  For "MLPerf Inference Benchmark" ISCA 2020 paper

- ◇ Amazon, AWS Machine Learning Research Award. *2020–2021*
  USD$40,000 cash and USD$80,000 in AWS cloud credits
  "Efficient DNN Training at Scale: from Algorithms to Hardware"

- ◇ Facebook, Faculty Research Award, **USD$49,500**. *2020–2021*
  AI Systems HW/SW Co-Design
  "Efficient DNN Training at Scale: from Algorithms to Hardware"

- ◇ **IEEE MICRO Top Picks Honorable Mention** *2019*
  For "Janus: Optimizing Memory and Storage Support for Non-Volatile Memory Systems", ISCA 2019 paper

- ◇ CIFAR, AI Chair, **$1,000,000**. *2019–2024*
  "Systems for Machine Learning"

- ◇ Connaught New Researcher Award, Connaught Fund, **$10,000 total**. *2018–2019*
  "Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications"

- ◇ **NVIDIA Graduate Fellowship** *2015 – 2016*
  5 winners nation-wide

- ◇ **First place in ACM SRC (Student Research Competition)** *Mar 2015*
  Energy-Efficient Data Compression for GPU Memory Systems @ ASPLOS'15

- ◇ **Qualcomm Innovation Fellowship Finalist** *2015 – 2016*
  Together with Nandita Vijaykumar. Selected as one of 35 out of 146 teams

- ◇ **Facebook Fellowship Finalist** *2015 – 2016*
  $500 cash prize

- ◇ **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

Research Funding

- ◇ Amazon, AWS Research Gift, **USD$50,000**. *2022–2023*
  “Holistic Systems Techniques for Efficient Training of Machine Learning Models”
- ◇ VMware, VMware Early Career Faculty Grant, **USD$50,000**. *2022–2023*
- ◇ Mitacs, Accelerate, **$30,000 total**. *2022–2023*
  “Modeling Application Performance under Multi-Instance (Multi-Stream) Execution Scenarios”
- ◇ Google, Google Scholar Research Award, **USD$60,000**. *2022–2023*
  “Holistic Systems Techniques for Efficient Training of Deep Learning Models”
- ◇ Huawei, Research Grant, **$336,000 total**. *2022–2025*
  “Optimizing DNN Training and Emerging Applications”
- ◇ Amazon, AWS Research Gift, **USD$50,000**. *2021–2022*
  “DietCode: High-Performance Code Generation for Dynamic Tensor Programs”
- ◇ NSERC USRA Award, **$6,000 total**. *2021–2021*
  “Machine Learning Compilers”, USRA: Benjamin Chislett
- ◇ DCS Award **$6,000 total**. *2021–2021*
  “Methodology for Developing and Evaluating Machine Learning Chips”, DCS: Chenhao Jiang
- ◇ Mitacs, Accelerate, **$30,000 total**. *2021–2022*
  “Adversarial Robustness of Deep Learning Algorithms on Next-Gen AI Accelerators”
- ◇ Canada Foundation for Innovation (CFI). *2021–2024*
  John Evans Leaders Fund program, Co-PI.
  Total: **CAD$276,000**, My share: **CAD$138,000**
  “Computer systems support for machine learning and artificial intelligence ”
- ◇ Amazon, AWS Machine Learning Research Award. *2020–2021*
  USD$40,000 cash and USD$80,000 in AWS cloud credits
  “Efficient DNN Training at Scale: from Algorithms to Hardware”
- ◇ Facebook, Faculty Research Award, **USD$49,500**. *2020–2021*
  AI Systems HW/SW Co-Design
  “Efficient DNN Training at Scale: from Algorithms to Hardware”
- ◇ Facebook, Facebook/University of Toronto unrestricted gift, Co-PI. *2020–2021*
  Total: **USD$150,000**, My share: **USD$50,000**.
  “Efficient ML Everywhere: From the Edge to the Data Center, From SW to HW”
- ◇ Mitacs, Accelerate, **$30,000 total**. *2020–2021*
  “Explore efficiently automated parallel hyperparameter search for optimizing machine learning models over large scale cloud cluster”

- NSERC UTEA, **$4,875 total**. *2020–2020*
  “Fair and Efficient Scheduling of Machine Learning Workloads in High Performance Computer Clusters”, UTEA: Yu Bo Gao

- ESROP - U of T, **$3,000 total**. *2020–2020*
  “Efficient Streaming Engines for Time Series Data”, ESROP: Kimberly Hau

- NSERC CRD, **$273,000 total**. *2020–2023*
  “Efficient Distributed DNN Training and Inference”

- NSERC CRD, **$180,000 total**. *2020–2023*
  “Efficient Compiler-Driven Pointer Compression”

- CIFAR, AI Chair, **$1,000,000**. *2019–2024*
  “Systems for Machine Learning”

- NSERC, Strategic Networks Grant, Co-PI. *2019–2021*
  Total: **$1,000,000 per year**, My share: **$35,000 per year**
  “COHESA Network”

- Mitacs, Accelerate, **$30,000 total**. *2019–2020*
  “Next Generation AI Accelerator Algorithm Hardware Co-Optimization”

- Huawei, Research Grant, **$428,400 total**. *2019–2022*
  “Efficient Data Compression/Deduplication for Persistent Memory and DRAM”

- IBM Canada, CAS Program, Award #1112, **$90,000 total**. *2019–2022*
  “Efficient Compiler-Driven Pointer Compression”

- Huawei, Research Grant, **$289,300 total**. *2019–2022*
  “Compiler Infrastructure for Optimizing DNN Workloads”

- NSERC Discovery Grant (Increase) **$12,500 total**. *2018–2023*
  “Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications”

- NSERC CRD, **$204,000 total**. *2019–2022*
  “Efficient Memory Footprint Reduction for Java Performance”

- Huawei, Research Grant, **$199,546 total**. *2018–2021*
  “Efficient Distributed DNN Training”

- NSERC UTEA, **$4,875 total**. *2018–2018*
  “Parallelism and Hardware Heterogeneity Support in Modern Compilers”, UTEA: Qiongsi Wu

- Connaught New Researcher Award, Connaught Fund, **$10,000 total**. *2018–2019*
  “Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications”

- NSERC Discovery Accelerator Supplement Grant, **$120,000 total**. *2018–2021*
  “Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications”

- NSERC Discovery Grant NSERC (RGPIN-2018-06514), **$140,000 total**. *2018–2023*
  “Exploiting Hardware Heterogeneity for Efficient Execution of Emerging Applications”

- IBM Canada, CAS Program, Award #1063, **$102,000 total**. *2018–2021*
  “Efficient Memory Footprint Reduction for Java Performance”

- Huawei, HiRP Open Program, **$87,044 total**. *2017–2019*
  “Hardware/Software Optimization and Compiler Support for Heterogeneous Systems”

◇ Canada Foundation for Innovation (CFI), **$240K total**. *2017–2020*
John Evans Leaders Fund program, CFI (Award #36585)
"Heterogeneous Systems Laboratory"

Section 2: Scholarly and Professional Work

Career Publication Statistics	
Total Number of Citations*	4315
H-index*	30
I10-index*	44
Journals Articles/Book Chapters	10
Peer-reviewed Workshop Papers	4
Peer-reviewed Conference Papers	50
Patents	3

∗ based on Google Scholar information on Nov 30, 2022

PEER-REVIEWED CONFERENCE PUBLICATIONS

50. Anand Jayarajan, Yudi Sun, Wei Zhao, Gennady Pekhimenko.
TiLT: A Time-Centric Approach for Stream Query Optimization and Parallelization.
ASPLOS'23 (to appear).

49. Yaoyao Ding, Cody Hao Yu, Bojian Zheng, Yizhi Liu, Yida Wang, Gennady Pekhimenko.
Task-Mapping-Oriented Programming Paradigm for DNN Tensor Programs.
ASPLOS'23 (to appear).

48. Muralidhar Andoorveedu, Zhanda Zhu, Bojian Zheng, Gennady Pekhimenko.
Tempo: Memory Footprint Optimization for Transformer-Based Model Training.
NeurIPS'22 (to appear).

47. Han Jie Qiu, Sihang Liu, Xinyang Song, Samira Khan, Gennady Pekhimenko.
Pavise: Integrating Fault Tolerance Support for Persistent Memory Applications.
Parallel Architectures and Compilation Techniques (**PACT'22**). October 2022

46. Xiaodan Serina Tan, Pavel Golikov, Nandita Vijaykumar, Gennady Pekhimenko.
GPUPool: A Holistic Approach to Fine-Grained GPU Sharing in the Cloud.
Parallel Architectures and Compilation Techniques (**PACT'22**). October 2022

45. Bojian Zheng, Ziheng Jiang, Cody Hao Yu, Haichen Shen, Josh Fromm, Yizhi Liu, Yida Wang, Luis Ceze, Tianqi Chen, Gennady Pekhimenko.
DietCode: Automatic Optimization for Dynamic Tensor Programs.
Machine Learning and Systems Conference (**MLSys'22**). September 2022

44. Hongyu Zhu, Ruofan Wu, Yijia Diao, Shanbin Ke, Haoyu Li, Chen Zhang, Jilong Xue, Lingxiao Ma, Yuqing Xia, Wei Tsui, Fan Yang, Mao Yang, Lidong Zhou, Asaf Cidon, Gennady Pekhimenko.
Roller: Fast and Efficient Tensor Compilation for Deep Learning.
USENIX Symposium on Operating Systems Design and Implementation (**OSDI'22**). July 2022

43. Sana Tonekaboni, Gabriela Morgenshtern, Azadeh Assadi, Aslesha Pokhrel, Xi Huang, Anand Jayarajan, Robert Greer, Gennady Pekhimenko, Melissa McCradden, Fanny Chevalier, Mjaye Mazwi, Anna Goldenberg.
How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at the ICU.
Conference on Health, Inference, and learning (**CHIL'22**). April 2022.

42. Ao Li, Bojian Zheng, Gennady Pekhimenko, Fan Long.
Automatic Horizontal Fusion for GPU Kernels.

International Symposium on Code Generation and Optimization (**CGO'22**). April 2022.

41. Max Ryabinin, Eduard Gorbunov, Vsevolod Plokhotnyuk, Gennady Pekhimenko.
Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices. International Conference on Neural Information Processing Systems (**NeurIPS'21**). December 2021.

40. Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitry Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas Wolf, Gennady Pekhimenko.
Distributed Deep Learning In Open Collaborations. International Conference on Neural Information Processing Systems (**NeurIPS'21**). December 2021.

39. Omar Mohamed Awad, Mostafa Mahmoud, Isak Edo, Ali Hadi Zadeh, Ciaran Bannon, Anand Jayarajan, Gennady Pekhimenko, Andreas Moshovos.
FPRaker: A Processing Element For Accelerating Neural Network Training. International Symposium on Microarchitecture (**MICRO'21**). October 2021.

38. Geoffrey Yu, Pavel Golikov, YuBo Gao, Gennady Pekhimenko.
Habitat: Prediction-guided Hardware Selection for Deep Neural Network Training. USENIX Annual Technical Conference (**ATC'21**). July 2021.

37. Ziqi Wang, Michael A. Kozuch, Todd C. Mowry, Vivek Seshadri, Chulhwan Choo, Gennady Pekhimenko.
NVOverlay: Efficient, Scalable and Flexible Full-System Checkpointing on NVM with Low Write Amplification. ACM/IEEE International Symposium on Computer Architecture (**ISCA'21**). June 2021.

36. Shang Wang, Peiming Yang, Yuxuang Zheng, Xin Li, Gennady Pekhimenko.
Horizontally Fused Training Array: An Effective Hardware Utilization Squeezer for Training Novel Deep Learning Models. Machine Learning and Systems Conference (**MLSys'21**). April 2021.

35. James Gleeson, Srivatsan Krishnan, Moshe Gabel, Vijay Janapa Reddi, Eyal de Lara, Gennady Pekhimenko.
RL-Scope: Cross-stack Profiling for Deep Reinforcement Learning Workloads. Machine Learning and Systems Conference (**MLSys'21**). April 2021.

34. Yaoyao Ding, Ligeng Zhu, Zhihao Jia, Gennady Pekhimenko, Song Han.
IOS: An Inter-Operator Scheduler for CNN Acceleration. Machine Learning and Systems Conference (**MLSys'21**). April 2021.

33. Isak Edo Vivancos, Sayeh Sharify, Milos Nikolic, Ciaran Bannon, Mostafa Mahmoud, Alberto Delmás Lascorz, Gennady Pekhimenko, Andreas Moshovos.
Boveda: Building an On-Chip Deep Learning Memory Hierarchy Brick by Brick. Machine Learning and Systems Conference (**MLSys'21**). April 2021.

32. Anand Jayarajan, Kimberly Hau, Andrew Goodwin, Gennady Pekhimenko.
LifeStream: A High-performance Stream Processing Engine for Periodic Streams. International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS'21**).

31. Mostafa Mahmoud, Isak Edo Vivancos, Ali Hadi Zadeh, Omar Mohamed Awad, Jorge Albericio, Gennady Pekhimenko, Andreas Moshovos.
TensorDash: Exploiting Sparsity to Accelerate Deep Neural Network Training. International Symposium on Microarchitecture (**MICRO'20**). October 2020.

30. Geoffrey Yu, Tovi Grossman, Gennady Pekhimenko.
Skyline: Interactive In-editor Computational Performance Profiling for Deep Neural Network Training. ACM Symposium on User Interface Software and Technology (**UIST'20**). October 2020.

29. Hongyu Zhu, Amar Phanishayee, Gennady Pekhimenko.
    Daydream: Accurately Estimating the Efficacy of Performance Optimizations for DNN Training. USENIX Annual Technical Conference (**ATC’20**). July 2020.

28. Bojian Zheng, Nandita Vijaykumar, Gennady Pekhimenko.
    Echo: Compiler-based GPU Memory Footprint Reduction for LSTM RNN Training. ACM/IEEE International Symposium on Computer Architecture (**ISCA’20**). June 2020.

27. Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, Brian Anderson, Maximilien Breughe, Mark Charlebois, William Chou, Ramesh Chukka, Cody Coleman, Sam Davis, Pan Deng, Greg Diamos, Jared Duke, Dave Fick, J. Scott Gardner, Itay Hubara, Sachin Idgunji, Thomas B. Jablin, Jeff Jiao, Tom St. John, Pankaj Kanwar, David Leei, Jeffery Liao , Anton Lokhmotov, Francisco Massa, Peng Meng, Paulius Micikevicius, Colin Osborne, Gennady Pekhimenko, Arun Tejus, Raghunath Rajan, Dilip Sequeira, Ashish Sirasao, Fei Suni, Hanlin Tang, Michael Thomson, Frank Wei, Ephrem Wu, Lingjie Xu, Koichi Yamada, Bing Yu, George Yuan, Aaron Zhong, Peizhao Zhang, Yuchen Zhou.
    MLPerf Inference Benchmark. ACM/IEEE International Symposium on Computer Architecture (**ISCA’20**). June 2020.
    IEEE MICRO Top Picks Award
    HiPEAC Paper Award

26. Shang Wang, Yifan Bai, Gennady Pekhimenko.
    Scaling Back-propagation by Parallel Scan Algorithm. Machine Learning and Systems Conference (**MLSys’20**). March 2020.

25. Peter Mattson, Christine Cheng, Gregory Diamos, Cody Coleman, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St John, Carole-Jean Wu, Lingjie Xu, Cliff Young, Matei Zaharia.
    MLPerf Training Benchmark. Machine Learning and Systems Conference (**MLSys’20**). March 2020.

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.
    IEEE MICRO Top Picks Honorable Mention

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. Machine Learning and Systems Conference (**MLSys’19**). April 2019.

21. Hongyu Zhu, Mohamed Akrout, 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 Technical 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.

JOURNALS & BOOK CHAPTERS

10. Anirudh Mohan Kaushik, Gennady Pekhimenko, Hiren Patel. *Gretch: A Hardware Prefetcher for Graph Analytics*. ACM Transactions on Architecture and Code Optimization (**TACO'21**). 2021.

9. Amir Yazdanbakhsh, Gennady Pekhimenko, Hadi Esmaeilzadeh, Onur Mutlu, Todd C. Mowry.
Towards Breaking the Memory Bandwidth Wall Using Approximate Value Prediction.. **Approximate Circuits**. 2019.

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 Transac-

tions on Architecture and Code Optimization (**TACO'16**). 2016.

5. Donghyuk Lee, Saugate 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.

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

6. Amar Phanishayee, Gennady Pekhimenko, Animesh Jain.
Efficient data encoding for deep neural network training. Patent No. 20190347549. November 2019.

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.

Section 3: Teaching and Supervision

TEACHING EXPERIENCE

- **Instructor** at the University of Toronto, *Fall 2020*
CSC B58H: Computer Organization, Undergraduate
- **Instructor** at the University of Toronto, *Winter 2023, 2021, 2020, 2019, 2018*
CSC D70H: Compiler Optimization, Undergraduate
- **Instructor** at the University of Toronto, *Fall 2022, 2021, 2020, 2019, 2018, 2017*
CSC 2224H: Parallel Computer Architecture and Programming, Graduate
- **Teaching Assistant** at Carnegie Mellon University, *Spring 2012*
Optimizing Compilers, Graduate
- **Teaching Assistant** at Carnegie Mellon University, *Fall 2011*
Introduction to Computer Systems, Undergraduate
- **Teaching Assistant** at the University of Toronto, *Fall 2007, Spring 2008*
Operating Systems, Undergraduate
- **Teaching Assistant** at the University of Toronto, *Spring 2007*
Computer Programming, Undergraduate
- **Teaching Assistant** at the University of Toronto, *Fall 2006*
Software Engineering, Undergraduate

SUPERVISION EXPERIENCE

Career Student Statistics			
	In Progress	**To Begin**	**Graduated**
Doctoral Supervisor	10	0	2
Master Supervisor	7	0	17
Doctoral Committee Member	5	1	8
Postdoctoral Fellows	0	1	0
Undergraduate Research Assistants	3	0	17

STUDENTS AND POSTDOCS SUPERVISED

Current:

- Christina Giannoula, Postdoc Fellow.

- Bojian Zheng, PhD Student. Machine Learning Compilers.
- Anand Jayarajan, PhD Student. Efficient stream processing engine.
- Shang (Sam) Wang, PhD student. Horizontal fusion for efficient DNN training.
- Jiacheng Yang, PhD student. DNN Training at the edge.
- Yaoyao Ding, PhD. student. Efficient tensor compilers.
- Pavel Golikov, PhD. student. Stream processing at the edge.
- Qidong Su, PhD student.
- Daniel Snider, PhD student.
- Yu Bo Gao, PhD student.
- Renbo Tu, PhD student.
- Kevin Song, MASc. student.
- Jasper Zhu, MSc. student.
- Peiming Yang, MASc. student.
- Xin Li, MASc. student.
- Chenhao Jiang, MASc. student.

- Zhanda Zhu, MASc. student.
- John Zhou, MScAC. student

- Wei Zhao, BSc. student.
- Yudi Sun, BASc. student.
- Baorun Mu, BSc. student.

Graduated:

- James Gleeson, PhD (co-advised with Eyal de Lara, 2022). Optimizing reinforcement learning training. First position: Samsung research.
- Ellina Zhang, BASc. student (2022).
- Pavel Golikov, MSc. (2022). FlexIoT: Flexible IoT Application Development With Stream Processing Engines Fist position: PhD student at UofT.
- Daniel Snider, MSc. (2022). Hotline Profiler: A Multi-Scale Timeline for Visualizing Time-Use in DNN Training Fist position: PhD student at UofT.
- Yaoyao Ding, MASc. (2022). IOS: Inter-Operator Scheduler for CNN Acceleration Fist position: PhD student at UofT.
- Alexandra Tsvetkova, MSc. (2022). Data compression for Java applications. Fist position: Google.
- Muralidhar Andoorveedu, BASc. (2022). Memory Footprint Reduction for Transformer-based models. First position: Amazon AWS.
- Qingyang Qie, BSc. (2022). Machine Learning Compilers. First position: Amazon AWS.
- Yaoyao Ding, MASc. (2022). IOS: Inter-Operator Scheduler for CNN Acceleration. First position: PhD student at UofT.
- Hongyu Zhu, PhD. (2022). Benchmarking, Profiling and White-Box Performance Modeling for DNN Training. First position: ByteDance.
- Jiacheng Yang, MASc. (2021). Enabling Privacy-Preserving Model Personalization via On-Device Incremental Training . First position: PhD student at UofT.
- Alexander Cann, MScAc. (2021). Modeling Application Performance under Multi-Instance (Multi-Stream) Execution Scenarios. First position: AMD.
- Qiongsi Wu, MSc. (2021). Compiler support for multi-threading with OpenMP. First position: IBM.
- Hanjie Qiu, MSc. (2021). Pavise: Integrating Fault Tolerance Support for Persistent Memory Applications. First position: AMD.
- Xiaodan (Serina) Tan, MASc. (2021). GPUPool: A Holistic Approach to Fine-Grained GPU Sharing in the Cloud. First position: Amazon AWS.
- Jiahuang (Jacob) Lin, MScAC. (2021). Speech recognition using DeepSpeech2 model.
- Yvonne Yang BASc. (2021). Intern with James Gleeson. RL Optimizations.
- Benjiamin Chislett, BSc. (2021). Machine Learning Compilers.
- Maryam Gohargani, BSc. (2021). Intern with Qiongsi Wu.
- Kimberly Hau, BASc. (2021). Intern with Anand Jayarajan. Streaming Engines.
- Cong Wei, BSc. (2021). Intern with Hongy Zhu. ML Benchmarks.
- Geoffrey Yu, MSc. (2020). Habitat: Prediction-guided Hardware Selection for Deep Neural Network Training. First position: PhD student at MIT EECS.
- Shang (Sam) Wang, MSc. (2020). Back-propagation by Parallel Scan Algorithm. First position: Nvidia.
- Yingying Fu, MScAC (2020). Next Generation AI Accelerator Algorithm Hardware Co-Optimization. First position: Untether AI, Toronto, ON.
- Izaak Niksan, BASc. (2020). Memory profiler for DNN training.
- Pavel Klishin, MSc. (2019). DNN training acceleration with FPGAs. First position: Industry, Moscow, Russia
- Andrew Pelegris, MSc. (2019) . Binarized DNNs acceleration. First position: Stealth-mode startup.

- Mohamed Akrout, MScAC (2019). Reinforcement learning profiling and training. First position: Research Scientist at Triage, Toronto, ON.
- Ming (Michael) Yang, BASc. (2019). New simulator infrastucture for GPUs. First position: Engineer at Cerebras, Bay Area, CA.
- Yifan Bai, BSc. (2019). Jacobian-based approach for DNN training. First position: Graduate student at UC Berkeley, CA.
- Kuei-Fang (Albert) Hsueh, BASc. (2019). Machine translation using Transformer model for inference. First position: Graduate student at UofT, Toronto, ON
- Akshay Nair, BSc. (2018). Simulation infrastructure for GPUs. First position: Software Engineer at Google, Mountain View.

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, Undegraduate 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.

STUDENT AWARD HIGHLIGHTS

- Geoffrey Yu (NSERC CGS-D, NSERC CGS-M, Snap Research Scholarship, Vector Institute Scholarship in Artificial Intelligence, Queen Elizabeth II Graduate Scholarship)
- Hanjie Qiu (OGS, Vector Institute Scholarship in Artificial Intelligence)
- Qiongsi Wu (NSERC CGS-M, Vector Institute Scholarship in Artificial Intelligence)
- Serina Tan (Vector Institute Scholarship in Artificial Intelligence)
- Bojian Zheng (Doctoral Completion Award, Third place in MICRO 2018 ACM SRC)
- James Gleeson (Bell Scholarship (twice))
- Qidong Su (Wolfand Scholarship)
- Jasper Zhu (OGS, Vector Institute Scholarship in Artificial Intelligence, NSERC-CGS-M)
- Kevin Song (Queen Elizabeth II Graduate Scholarship, OGS)
- Zhanda Zhu (Kai Yin Shen Graduate Scholarship)
- Chenhao Jiang (Vector Institute Scholarship in Artificial Intelligence)
- Yubo Gao (Wolfond Fellowship)
- Xin Li (OGS)
- Shang Wang (OGS)
- Yaoyao Ding (Amazon Post-Internship Fellowship)

Section 4: Internal and External Service

INTERNAL DEPARTMENT COMMITTEES

- **Faculty Hiring Committee** *2022–2023*

- **Faculty Hiring Committee, Systems and Data Systems** *2021–2022*
  Outcome: Hired one systems faculty and one security/systems faculty.
- **Graduate Admission Committee, Co-Chair** *2020–2021*
- **Graduate Admission Committee, Co-Chair** *2019–2020*
- **Faculty Hiring Committee, Systems and Computer Architecture** *2018–2019*
  Outcome: Hired one systems faculty (for UTSC position).

Internal Faculty Committees

- **IT Strategic Committee, UTSC** *2020–2022*

External Organization Services

- **ML Chip and Compilers Symposium, co-located with MLSys, Co-Chair** *2022*
- **MLCommons/MLPerf Research Co-Chair** *2019–2022*
- **MLBench Workshop/Tutorial Organizer** *2019–2022*

External Conference Services

- **General Chair**, MICRO 2023 *2022–2023*
- **PC Member**, MLSys 2023 *2022–2023*
- **Heavy PC Member**, EuroSys 2023 *2022–2023*
- **PC Member**, MLSys 2022 *2021–2022*
- **PC Member**, ASPLOS 2022 *2021–2022*
- **PC Member**, MICRO 2021 *2021–2021*
- **Co-Chair**, Artifact Evaluation Committee at MICRO 2021 *2021–2021*
- **PC Member**, OSDI 2021 *2020–2021*
- **ERC (External Review Committee) Member**, ISCA 2021 *2020–2021*
- **Chair**, Artifact Evaluation Committee at ASPLOS 2021 *2020–2021*
- **PC Member**, MLSys 2021 *2020–2021*
- **PC Member**, HPCA 2021 *2020–2021*
- **PC Member**, MICRO 2020 *2020*
- **ERC (External Review Committee) Member**, ISCA 2020 *2019–2020*
- **PC Member**, MICRO TopPicks 2020 *2019–2020*
- **PC Member**, MLSys 2020 *2019–2020*
- **PC Member**, EuroSys 2020 *2019–2020*
- **Co-Chair**, Artifact Evaluation at MLSys 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**, MLSys 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*

PROFESSIONAL MEMBERSHIPS

- **IEEE Computer Society** *2014–present*
- **Association of Computing Machinery (ACM)** *2012–present*
- **ACM SIGARCH** *2012–present*

Section 5: Additional Professional Activities

INVITED TALKS

74. *Efficient DNN Training at Scale*
Google, Mountain View *September 2022*
73. *Keynote Talk on Efficient DNN Training*
ExSAIS workshop @IPDPS, online *June 2022*
72. *Efficient DNN Training at Scale*
Cruise, online *May 2022*
71. *Memory Footprint Reduction Techniques for DNN Training: An Overview*
Higher School of Economics (HSU), online lecture *March 2022*
70. *Moshpit SGD: Communication-Efficient Decentralized Training on Heterogeneous Unreliable Devices*
Vector - NeurIPS Highlights, online *February 2022*
69. *Efficient DNN Training at Scale*
Huawei STW Workshop, online *October 2021*
68. *Efficient DNN Training at Scale*
Facebook PyTorch Seminar Series, online *October 2021*
67. *Efficient DNN Training at Scale*
SAFARI @ETH Zurich, online *August 2021*
66. *LifeStream: A High-Performance Stream Processing Engine for Periodic Streams*
Microsoft Research, online *May 2021*
65. *Apple ODML Workshop Keynote Invited Talk*
Apple ODML Workshop, online *April 2021*
64. *ASPLOS MLBench Workshop*
ASPLOS MLBench'21, online *April 2021*
63. *MLSys MLBench Tutorial*
MLSys MLBench'21, online *April 2021*
62. *HPCA MLBench Tutorial*
HPCA MLBench'21, online *February 2021*
61. *Efficient DNN Training at Scale: from Algorithms to Hardware*
Facebook, Facebook Faculty Summit, online *October 2020*
60. *Keynote talk on Efficient DNN Training at Scale*
Vector Institute NLP Symposium, online *September 2020*
59. *ISPASS Tutorial on ML Benchmarking*
ISPASS ML Performance'20, online *August 2020*
58. *VCEW Invited Talk: ML Benchmarking*
VCEW'20, online *June 2020*
57. *ISCA Mini-Panel: Accelerators*
ISCA'20, online *June 2020*
56. *Efficient DNN Training at Scale: from Algorithms to Hardware*
Microsoft, Microsoft Research Seminar, online *May 2020*

55. *Efficient DNN Training at Scale: from Algorithms to Hardware*
Facebook, SysML Seminar, online *May 2020*

54. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Fields Institute, Toronto, ON *January 2020*

53. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Uber ATG, Toronto, ON *November 2019*

52. *Holistic Approach to DNN Training Efficiency: Analysis and Optimizations*
Yandex, Moscow, Russia *August 2019*

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*

Selected Media and Press

6. DietCode, MLSys'22 paper published in the Amazon Science blog: Link
   Amazon Science *Sep 2022*
5. Assistant Professor Gennady Pekhimenko inducted into International Symposium on Computer Architecture Hall of Fame
   UofT News *Sep 2021*
4. MLCommons Launches and Unites 50+ Global Technology and Academic Leaders in AI and Machine Learning to Accelerate Innovation in ML
   BusinessWire *Dec 2020*
3. Assistant Professor Gennady Pekhimenko receives Amazon and Facebook research awards to train deep neural networks faster
   UofT News *Oct 2020*
2. NeurIPS 2019, Amii, Mila, and Vector Researchers Discuss AI in Canada
   Medium.com *Dec 2019*
1. How to evaluate machine learning? U of T research supports latest benchmark initiative
   UofT News *May 2018*