

Train your models faster by learning how to profile and apply system-level optimizations

TMLS 2022 Workshop

November 28, 2022

Akbar Nurlybayev, Xin Li, Yubo Gao



Agenda

- Who we are
- What we do
- Why bother
- Hands-on workshop
- Recap

Who we are



Gennady Pekhimenko CEO

UofT CS and ECE
Professor, PhD (CMU)
Vector Faculty Member

Major Awards

ISCA Hall of Fame
MLPerf Research Co-Chair
Google Scholar Research Award
Amazon AWS ML Award
Facebook Faculty Research Award
CIFAR AI Chair
Published 50+ top-tier papers,
4 patents

<https://dblp.org/search?q=mlsys>



a service of SCHLOSS DAGSTUHL
Leibniz Center for Informatics

[home](#) | [browse](#) | [search](#) | [about](#)



mlsys

[+] Search dblp
[-] powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home

Dagstuhl

[~] Venue search results

Likely matches

- Conference on Machine Learning and Systems (MLSys)

[~] Publication search results

found 175 matches

2022

- Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, Brennan Saeta, Parker Schuh, Ryan Sepassi, Laurent El Shafey, Chandramohan A. Thekkath, Yonghui Wu:
Pathways: Asynchronous Distributed Dataflow for ML. MLSys 2022
- Jie Zhao, Xiong Gao, Ruijie Xia, Zhaochuang Zhang, Deshi Chen, Lei Chen, Renwei Zhang, Zhen Geng, Bin Cheng, Xuefeng Jin:
Apollo: Automatic Partition-based Operator Fusion through Layer by Layer Optimization. MLSys 2022
- Runsheng Guo, Victor Guo, Antonio Kim, Josh Hildred, Khuzaima Daudjee:
Hydrozoa: Dynamic Hybrid-Parallel DNN Training on Serverless Containers. MLSys 2022

[~] Refine list



refine by author

Gennady Pekhimenko (8)
Matei Zaharia (7)
Carole-Jean Wu (6)
Ion Stoica (6)
Vijay Janapa Reddi (5)
Matthew Mattina (5)
Sudip Roy (5)
Paul N. Whatmough (4)
Anshumali Shrivastava (4)
Dimitris S. Papaliopoulos (4)
926 more options

Who we are



**Gennady
Pekhimenko**
CEO

UofT CS and ECE
Professor, PhD (CMU)
Vector Faculty Member

Major Awards

ISCA Hall of Fame
MLPerf Research Co-Chair
Google Scholar Research Award
Amazon AWS ML Award
Facebook Faculty Research Award
CIFAR AI Chair
Published 50+ top-tier papers,
4 patents

<https://dblp.org/search?q=mlsys>

[-] Refine list



refine by author

- | | |
|--------------------------------|--|
| Gennady Pekhimenko (8) | - Co-founder and Chief Technologist at Databricks |
| Matei Zaharia (7) | - Creator of Apache Spark |
| Carole-Jean Wu (6) | - Professor at Stanford University |
| Ion Stoica (6) | - Co-founder and Executive Chairman at Anyscale, creators of Ray framework |
| Vijay Janapa Reddi (5) | - Co-founder and Executive Chairman at Databricks |
| Matthew Mattina (5) | - Professor at UC Berkeley |
| Sudip Roy (5) | |
| Paul N. Whatmough (4) | |
| Anshumali Shrivastava (4) | |
| Dimitris S. Papailiopoulos (4) | |
| 926 more options | |

**We Make
Machine Learning
Affordable**



What we do: system-level optimizations



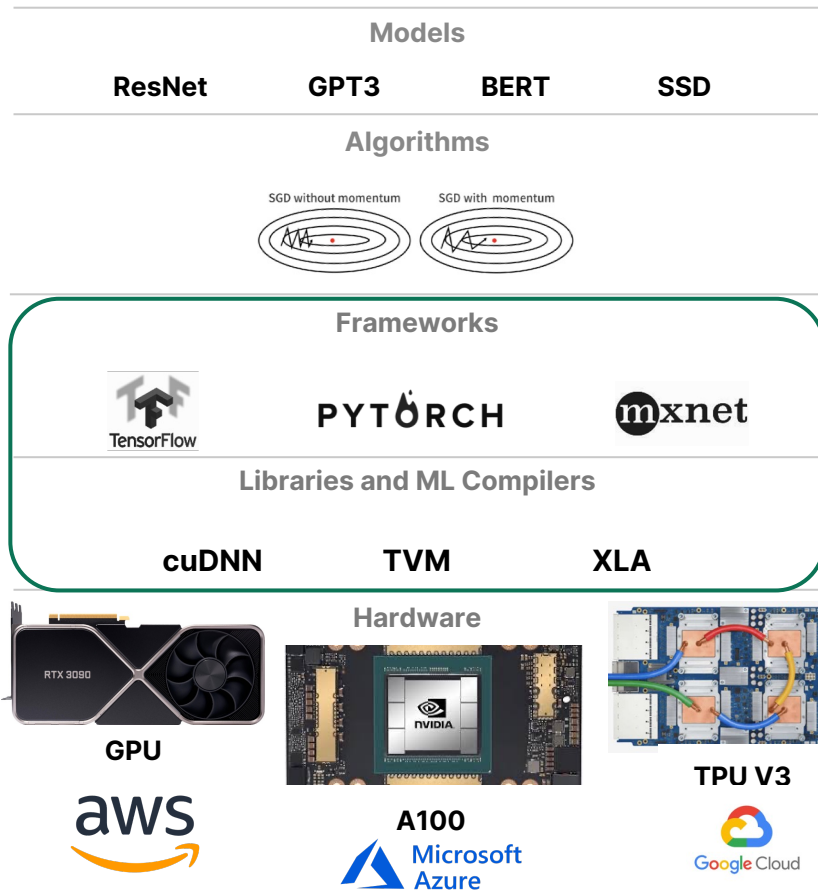
ML Tools



System-Level Optimizations



Hardware-Specific Optimizations



Training compute (FLOPs) of milestone Machine Learning systems over time

[illegible]

<https://arxiv.org/abs/2202.05924>

Hands-on workshop

<https://centml.ai/events/tmls2022>



Recap

- Applying system-level optimizations can **significantly improve** your ML workloads without changing your model.
- System-level optimizations **require expertise**.

Recap

ML Tools

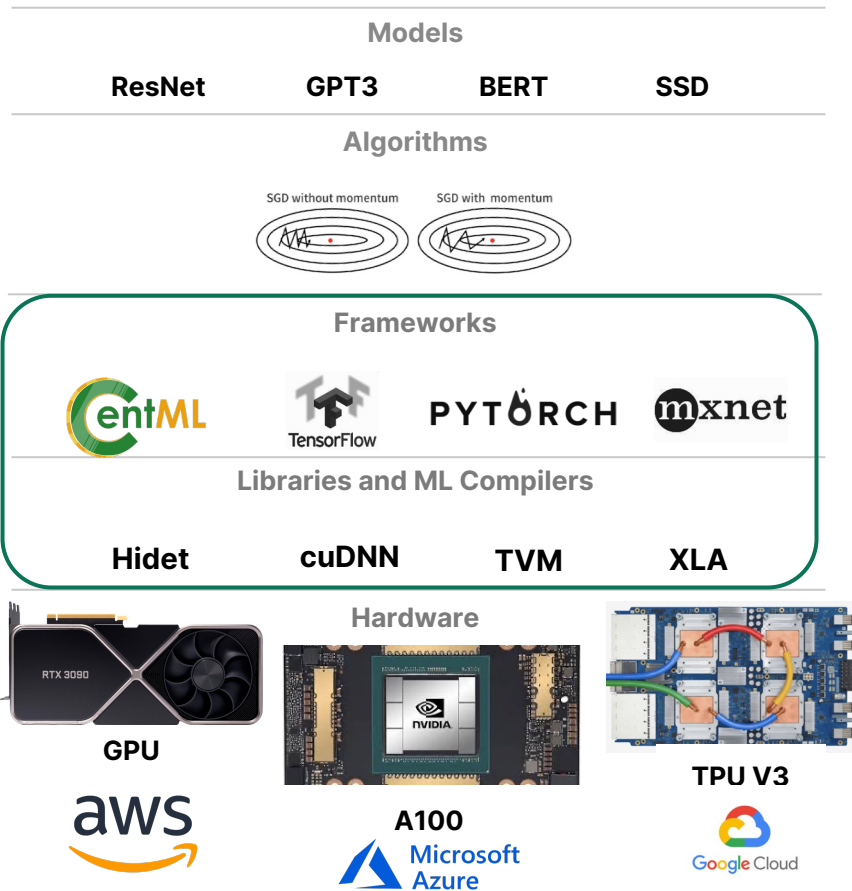
- Computational performance profiler
- GPU execution time predictor
- Cost estimator

System-Level Optimizations

- Fused optimizer
- torch.jit
- CUDAGraph
- Horizontal Fusion

Hardware-Specific Optimizations

- Mixed precision training



We Make Machine Learning Affordable

<https://centml.ai>

Xin Li
xin@centml.ai

Yubo Gao
ybgao@centml.ai

Akbar Nurlybayev
akbar@centml.ai
@atkabai



Are we a good fit?

- ☐ Do you engage in ML Training?
- ☐ What is the relative model size that you train?
- ☐ Where do you currently train?
- ☐ Do you use GPU?
- ☐ Do you perform hyperparameter search?
- ☐ How often do you train/re-train?
- ☐ What is the order of magnitude of your training and/or inference costs? 10s of thousand, 100s of thousands and etc.
- ☐ Do you currently use scheduler?
- ☐ Do you experience challenges with GPU allocations?