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

UNIVERSITY OF TORONTO

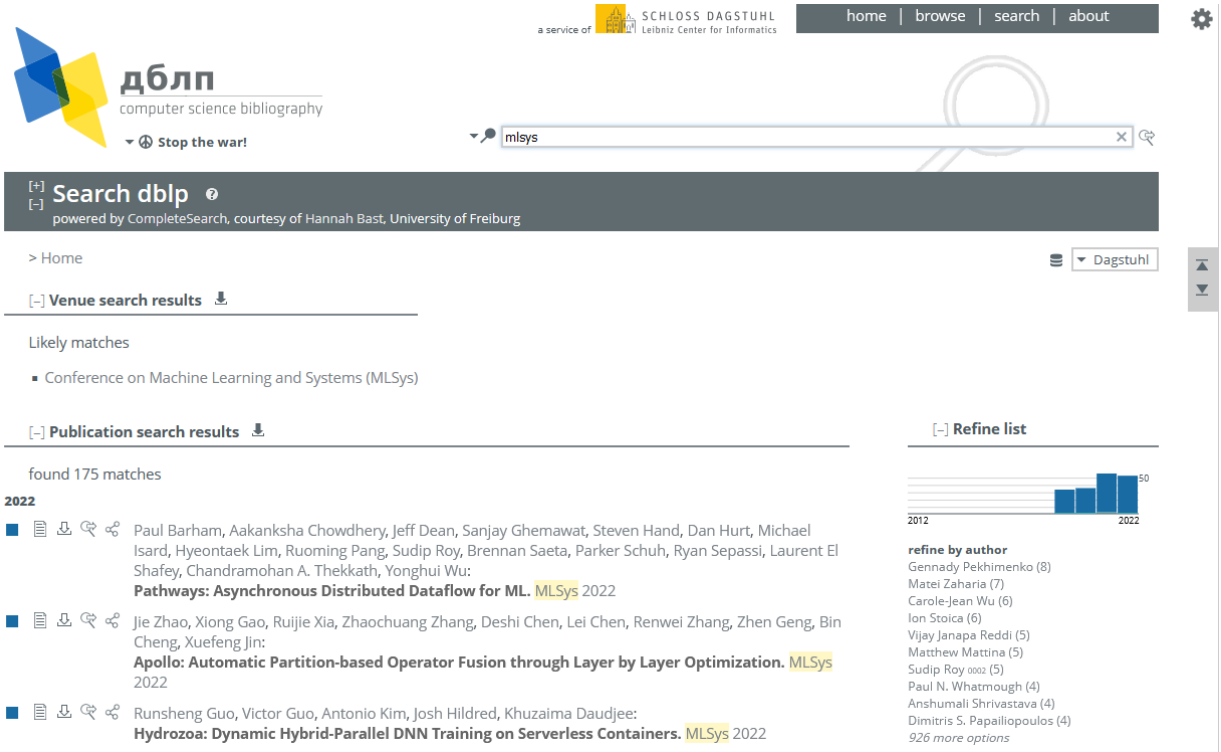
NVIDIA. IBM intel. aws Microsoft Research

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



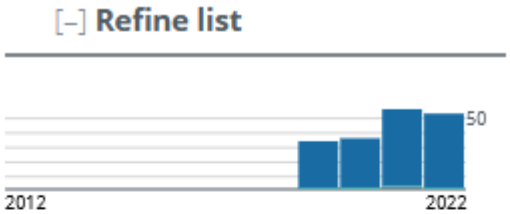
Who we are

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



- Co-founder and Chief Technologist at Databricks
- Creator of Apache Spark
- Professor at Stanford University
- Co-founder and Executive Chairman at Anyscale, creators of Ray framework
- Co-founder and Executive Chairman at Databricks
- Professor at UC Berkeley

We Make Machine Learning Affordable

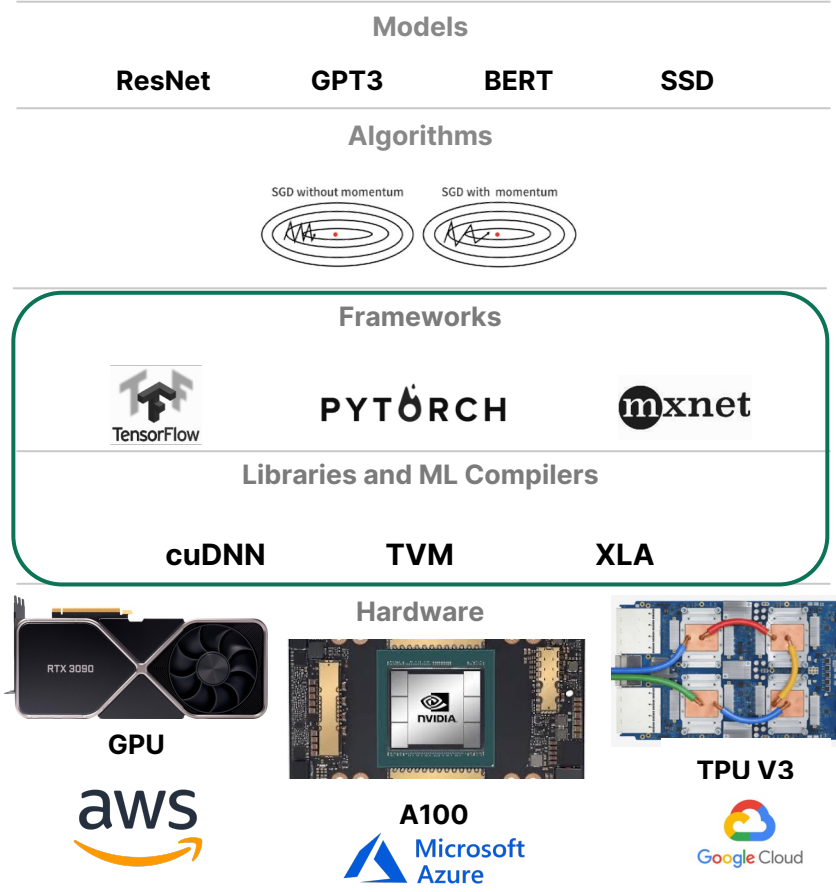
What we do: system-level optimizations

CentML expertise

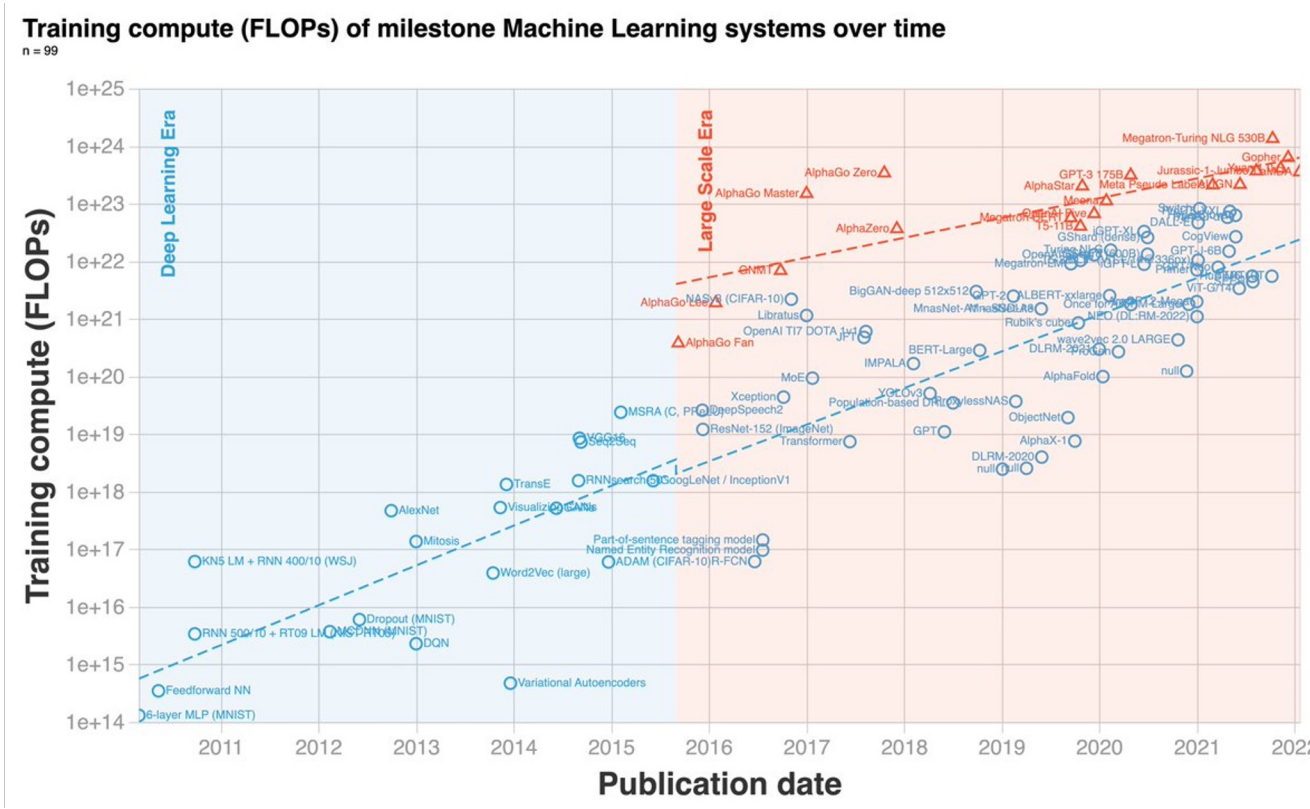
ML Tools

System-Level Optimizations

Hardware-Specific Optimizations



Larger models ⇒ longer training time



Source: Compute Trends Across Three Eras of Machine Learning, 2022

https://arxiv.org/abs/2202.05924


Hands-on workshop
https://centml.ai/events/tmls2022
CentML

Recap

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

- System-level optimizations **require expertise**.

Recap

ML Tools

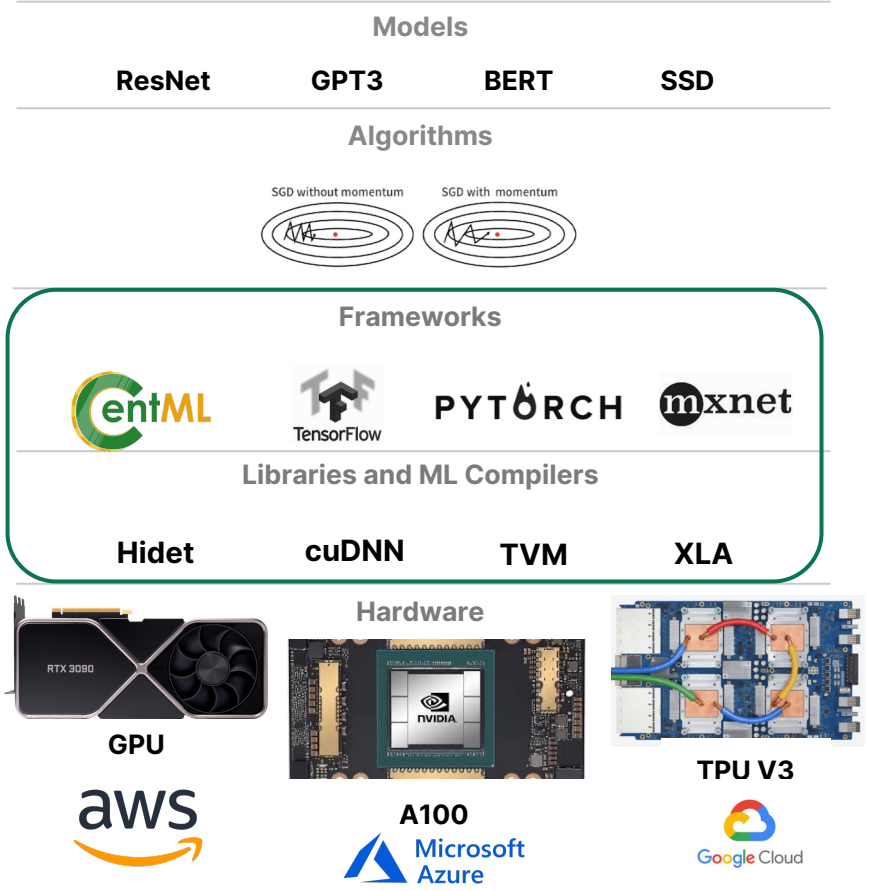
- Computational performance profiler
- GPU execution time predictor
- Cost estimator

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?