ML Commons

MLCommons® Community Meeting 4Q22

December 8, 2022

This community meeting is being recorded and will be shared

Schedule

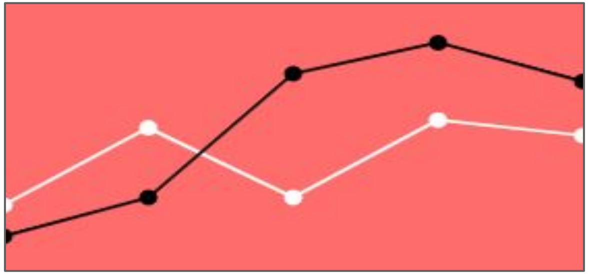
9:00 AM	Breakfast
9:30 AM	Welcome
10:00 AM	MLC Update
10:30 AM	Break
11:00 AM	Algorithms
11:10 AM	Medical
11:20 AM	Best Practices
11:30 AM	DataPerf
11:40 AM	Dynabench
11:50 AM	Science
12:00 PM	Lunch

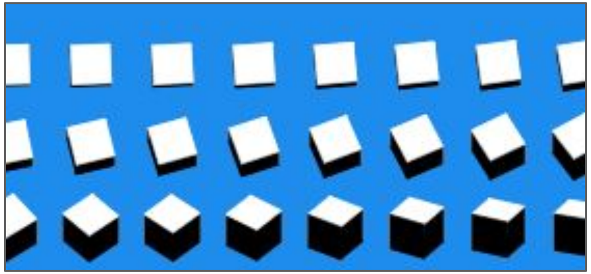
1:00 PM	Tiny
1:10 PM	Mobile
1:20 PM	Datasets
1:30 PM	Inference
1:40 PM	Training
1:50 PM	HPC
2:00 PM	Break
2:30 PM	Storage
2:40 PM	Power
2:50 PM	Benchmark Infra
3:00 PM	Research
3:10 PM	Closing Discussion

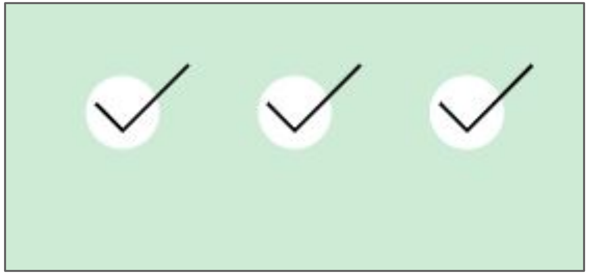
Welcome

MLCommons Update

4Q22 MLCommons Hero Awards







Matthew Frank:

MLPerf Training reference quality

Sanghyun Son: MLPerf Mobile Super-resolution model, dataset, and tons of work

New folks

BoD (Graphcore): Phil Brown

BoD (Alibaba): Weiming Zhao

PM: David Tafur

PM: Ashish Thaker

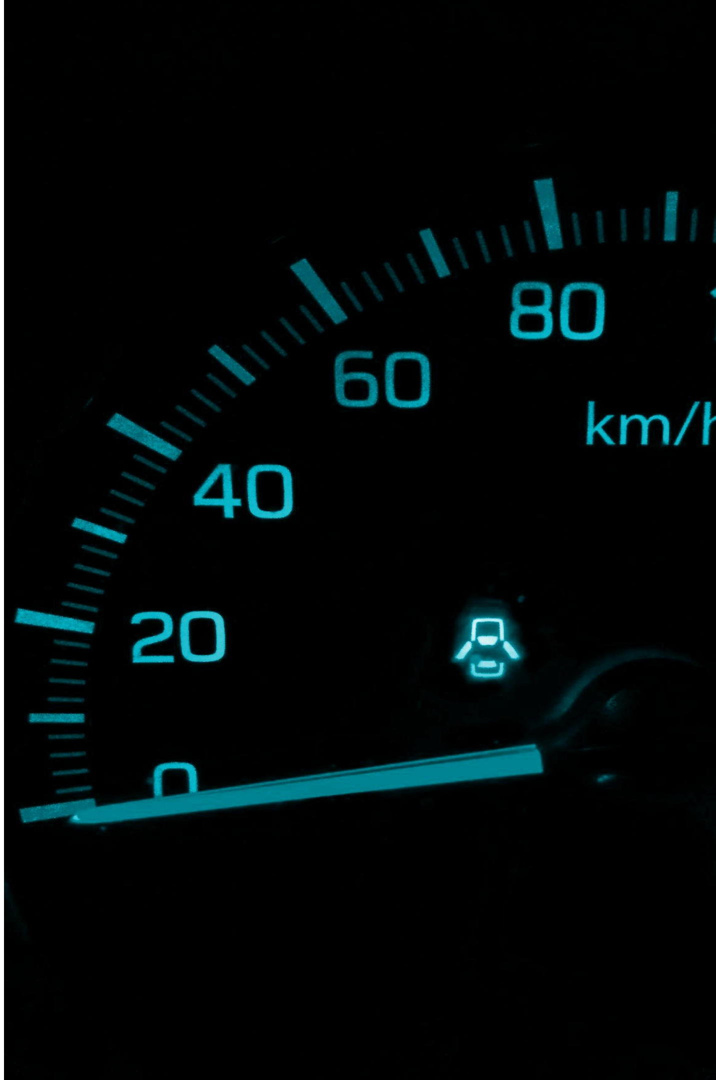
Admin: Anita Mikes

Full intros next meeting. :-)

Benchmarking 2022

- Results
  - Collectively 11,929+ results
  - (including 4000+ power results!)

- Benchmark Pipeline
  - RetinaNet, MOSAIC benchmarks launched
  - LLM, DLRM v2.0 nearing completion
  - Automotive in development
  - Power measurement for training effort started



Data & Best Practices 2022

- People's Speech v1.1, DollarStreet datasets released
- DataPerf alpha benchmarks: metrics for dat
- Dataset ecosystem discussions underway

- Much better submission infrastructure
- Dynabench platform moved to MLCommons
    - enabled DataPerf alpha
- MLCube enabled MedPerf for FeTs challenge

Research 2022

- MedPerf platform supported MACCAI FeTS challenge
- Algorithmic benchmarks nearing completion
- Scientific benchmarks nearing completion

Looking ahead to 2023

Introducing MLCommons roadmap [must be on community@ list to access, you can join here]
Quarter by quarter plans for each working group

Themes

- Benchmark launches:
  - LLM, DLRMv2, Auto, algorithms, DataPerf, MedPerf, science…
- Benchmark quality and presentation
- Public data ecosystem
- Organization capability,
  - Especially marketing and project management

Public datasets fuel ML

Enable researchers to communicate

Enable entrepreneurs, and often large company projects, to get started

Support benchmarks that drive progress

Support regulation (soon)

Bottom line: drive market growth

State of public datasets is limiting progress

Mostly aging academic efforts scraping web and cheaply labeling it

Legal, ethical, and safety issues
Don't address emerging research challenges
Don't provide high quality tests for solutions
Don't address practical data lifecycle
Miss opportunities for societal good
No consistent format or license

Need a better ecosystem!

Data + access + metrics + tools + community

Wishlist for 2023

1. Increased results utility via consumer input
2. Quality, well-documented reference code
3. Lighter-weight benchmarking, esp. inference

4. Dataperf, algorithms, and science submissions
5. MedPerf disease area funding/contacts
6. Dataset ecosystem funding/contacts

7. More brilliant, active academics
8. Marketing Director
9. Improved community meeting format

It would not be possible without our members

Founding Members

Members

Academics from educational institutions including:

Harvard University
Polytechnique Montreal
Peng Cheng Laboratory
Stanford University
University of California, Berkeley
University of Toronto
University of Tübingen
University of Virginia
University of York, United Kingdom
Yonsei University
York University, Canada


Thank you
Happy Holidays

Break

Schedule

9:00 AM	Breakfast
9:30 AM	Welcome
10:00 AM	MLC Update
10:30 AM	Break
11:00 AM	Algorithms
11:10 AM	Medical
11:20 AM	Best Practices
11:30 AM	DataPerf
11:40 AM	Dynabench
11:50 AM	Science
12:00 PM	Lunch

1:00 PM	Tiny
1:10 PM	Mobile
1:20 PM	Datasets
1:30 PM	Inference
1:40 PM	Training
1:50 PM	HPC
2:00 PM	Break
2:30 PM	Storage
2:40 PM	Power
2:50 PM	Benchmark Infra
3:00 PM	Research
3:10 PM	Closing Discussion

Algorithms

Mission

Create a set of rigorous and relevant benchmarks to measure neural network training speedups due to algorithmic improvements.

Model Track

- Tests models
  Model, initialization, loss, data augmentation
- Needs to perform well on one task
  Target evaluation metric and a set of compatible datasets
- All use the same (standard) optimizer and tuning procedure
  Avoid models that are extremely sensitive to optimizer details

Generally useful models

Training Algorithm Track

- Tests training algorithms
  Parameter update, training data selection, hyperparameter space
- Needs to perform well on **multiple** tasks
  Scored based on performance on a set of provided workloads
- No workload-specific adaptation is allowed
  Avoid optimizers that don't work on unseen workloads.

Generally useful training algorithms

Working Group Deliverables

1. **Rules**: We will produce a set of rules for rigorous benchmarking of algorithmic improvements.
2. **Harness**: We will produce a testing harness for these rules that is executable on commonly available clouds.
3. **Baseline** training algorithm/model implementations: We will produce a baseline training algorithm and model implementation for each benchmark, which can also serve as submission skeletons.
4. **Call for participation** & **initial submission round**: Once rules and harness/references are developed we will call for participation by the research/industry community.

Additional submission rounds on a regular schedule main · algorithmic-efficiency / RULES.md · Go to file · ···

fsschneider Removed private link to workload brainstorming · Latest commit 61851c4 2 days ago · History

1 contributor

381 lines (263 sloc) | 37.7 KB · Raw · Blame

MLCommons™ Algorithmic Efficiency Benchmark Rules

Version: 0.0.2 *(Updated 21 September 2021)*

> TL;DR New training algorithms and models can make neural net training faster. We need a rigorous training time benchmark that measures time to result given a fixed hardware configuration and stimulates algorithmic progress. We propose a Training Algorithm Track and a Model Track in order to help disentangle optimizer improvements and model architecture improvements. This two-track structure lets us enforce a requirement that new optimizers work well on multiple models and that new models aren't highly specific to particular training hacks.

- Introduction
- Training Algorithm Track
  - Submissions
    - Specification
    - Evaluation during training
    - Valid submissions
  - Tuning
    - External Tuning Ruleset
    - Self-Tuning Ruleset
  - Workloads
    - Public workloads
    - Held-out workloads
  - Scoring
    - Competition hardware
    - Defining target performance
    - Summary score using performance profiles
- Model Track

Introduction

We need a more scientifically sound methodology for evaluating training speedups due to new algorithms, including both new optimizers and new model architectures. Cutting edge machine learning (ML) models are exceeding the compute budgets of many researchers, and ML compute is becoming a larger and larger cost in industry. To reduce the compute and potentially environmental cost of ML research and practice, we need rigorous benchmarking of efficiency. Such benchmarks will guide us in selecting the best directions to evolve existing

Working Group Deliverables

1. **Rules**: We will produce a set of rules for rigorous benchmarking of algorithmic improvements.
2. **Harness**: We will produce a testing harness for these rules that is executable on commonly available clouds.
3. **Baseline** training algorithm/model implementations: We will produce a baseline training algorithm and model implementation for each benchmark, which can also serve as submission skeletons.
4. **Call for participation** & **initial submission round**: Once rules and harness/references are developed we will call for participation by the research/industry community.

Additional submission rounds on a regular schedule

Algorithms Working Group

Group

https://mlcommons.org/en/groups/research-algorithms/

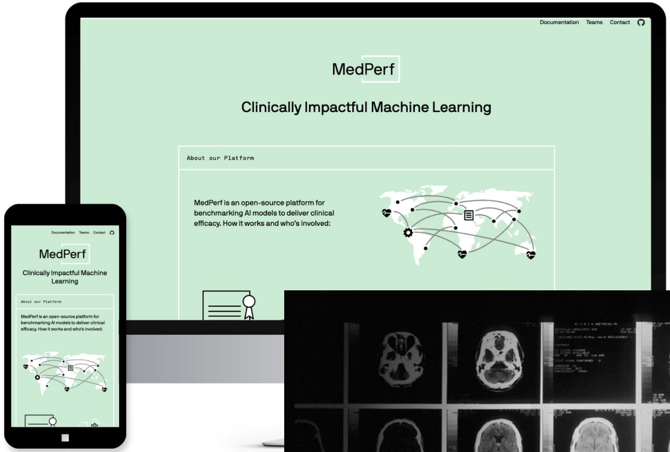
Meetings

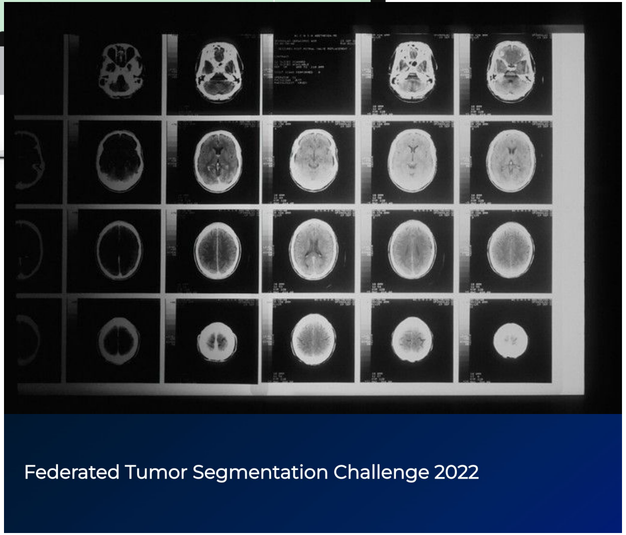
Weekly on Thursday from 11:30 AM - 12:30 PM (Pacific Time)

Medical

2022 Recap

- MedPerf Release
  - Federated Data Registry
  - Model Registry
  - Server
  - Client
  - Website
- MedPerf used in FeTS Challenge
  - First and largest federated brain tumor segmentation challenge (30+ hospitals, 500+ results, thanks Spyros Bakas!)
- Paper rebuttal at Nature Machine Intelligence
- Welcomed GanDLF to the Medical WG as the open-sourced deep learning framework for medical imaging





2023 Roadmap

Focus on **3 Pillars**

- **<u>Strategy</u>**
    - 2 benchmarks:
    - Continuous Federated Evaluation of brain tumor segmentation
    - Impactful clinical problem in oncology
    - Partnerships
        - Sage
        - ACR
        - Tata Medical Center
- **<u>Core technology</u>**
    - Web client for MedPerf
    - Model Protection for MedPerf
    - Multimodal data support for GaNDLF
- **<u>Integration with frameworks</u>**
    - Synapse
    - HuggingFace
    - MONAI

New additions

Alexander (Alex) Getka is a senior software engineer at the University of Pennsylvania and current lead developer of GaNDLF. He is interested in using machine learning for making science easier and for creating interactive experiences like games. In his spare time, he wonders how he managed to get here!

MedPerf Technical Updates

- Harness the FeTS experience:
  - Improved benchmark management utilities
    - Model prioritization
    - Multiple experiments in one command
  - Better user experience
    - Configuration profiles
    - Cached experiments
    - Partial Submissions
- Better documentation
  - Extensive explanation of the medperf ecosystem
  - Code refactoring and best practices
- Version releases coming up soon

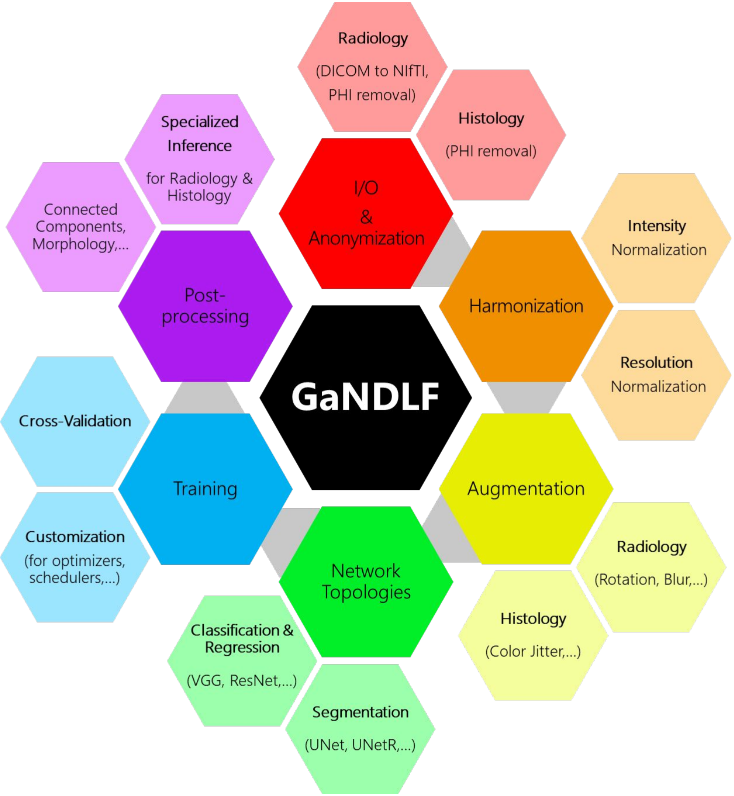
XNAT Support for Data Input

XNAT is an open source imaging informatics platform.

MedPerf will be able to:

- Expect input datasets to be hosted on XNAT
    - Automation for clinicians
- *(Maybe) Expect MLCubes to read data from the XNAT server*

The Generally Nuanced Deep Learning Framework



- **Low-code / no-code** deep learning for medical imaging
- Support for **multiple:**
  - Imaging domains (radiology, histology)
  - Tasks (segmentation, regression, classification)
- **Extensible**
- Common data preparation utilities (e.g. Anonymization)
- Model **Training** utilities:
  - Augmentation
  - cross-validation
- Built-in CPU-**optimization** (OpenVINO)

Thanks for your efforts and contribution!

Everyone at the Medical Working group for making all of this possible.

BIG THANKS!

Special thanks to David, Relja and Debo for their support

Bi-weekly meetings!

Same time: 10am PST

Best Practices

MLCube™

December 2022
MLCommons Community Meeting

Diane Feddema, Sergey Serebryakov &
The Best Practices Working Group

MedPerf Update

- Enablement of read-only setting for input parameters on mlcube CLI
- Added Singularity Support
- Enablement of GPU settings
- MLCube is being tested with Gandlf
- Continuous Support for the MedPerf project from the Best Practices WG

DynaBench Update

- Vision Challenge support
- Speech Challenge support
- Export of final results in json format
- Weekly meetings a continuous support for the project

Documentation updates 1/2

- Explaining MLCube concepts such as MLCube platform, task, parameter, effective configuration, workspace etc.
- Explaining the concept of system settings and describing the structure of system settings file.
- Adding automatically generated documentation for MLCube CLI (with `mkdocs-click` plugin).
- Adding custom epilogues to MLCube `click`-based commands to provide usage example (not available yet in web version)

Documentation updates 2/2

```
Usage: mlcube configure [OPTIONS]

  Configure MLCube. Some MLCube projects need to be configured first. For instance, docker-based MLCubes distributed via
  GitHub with source code most likely will provide a Dockerfile to build a docker image. In this case, the process of
  building a docker image before MLCube runner can run it, is called a configuration phase. In general, users do not need to
  run this command manually - MLCube runners should be able to figure out when they need to run it, and will run it as part
  of mlcube run command.

Options:
  --mlcube PATH       Path to an MLCube project. It can be a directory path, or a path to an MLCube configuration file. When it is
                      a directory path, MLCube runtime assumes this directory is the MLCube root directory containing mlcube.yaml
                      file. When it is a file path, this file is assumed to be the MLCube configuration file (mlcube.yaml), and a
                      parent directory of this file is considered to be the MLCube root directory. Default value is current
                      directory.

  --platform NAME     Platform name to run MLCube on (a platform is a configured instance of an MLCube runner). Multiple platforms
                      are supported, including docker (Docker and Podman), singularity (Singularity). Other runners are in
                      experimental stage: gcp (Google Cloud Platform), k8s (Kubernetes), kubeflow (KubeFlow), ssh (SSH runner).
                      Default is docker. Platforms are defined and configured in MLCube system settings file.

  -P, -p PARAMS       MLCube configuration parameter is a key-value pair. Must start with -P or '-p'. The dot (.) is used to refer
                      to nested parameters, for instance, -Pdocker.build_strategy=always. These parameters have the highest
                      priority and override any other parameters in system settings and MLCube configuration.

  -h, --help          Show help message and exit.

EXAMPLES:
  - Configure MNIST MLCube project:

      $ git clone https://github.com/mlcommons/mlcube_examples
      $ cd ./mlcube_examples
      $ mlcube configure --mlcube=mnist --platform=docker

MLCube online documentation: https://mlcommons.github.io/mlcube
```

Join Best Practices WG!

https://mlcommons.org/en/get-involved/#getting-started

- Working Group Meetings
  - General updates/discussions every Friday 9:00 - 10:00 PST
  - Mid-week technical sync every Tuesday 10:00 - 10:15 PST
- Topics of interest
  - How to share ML projects.
  - Reproducibility and portability of ML projects .
  - Docker/singularity best practices in HPC/enterprise data centers.
  - Challenges associated with ML tests and benchmarks in enterprise environments

DataPerf

Working Group Chairs: **Newsha Ardalani (FAIR)** and Praveen Paritosh (Google)

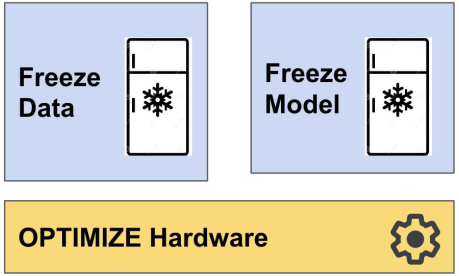
Mark Mazumder[1] Colby Banbury[1] Xiaozhe Yao[2] Bojan Karlaš[2] William Gaviria Rojas[3] Sudnya Diamos[3] Greg Diamos[5] Lynn He[6] Douwe Kiela[4] David Jurado[7] David Kanter[7] Rafael Mosquera[7] Juan Torres[7] Newsha Ardalani[8] Praveen Paritosh[9] Lora Aroyo[9] Bilge Acun[8] Sabri Eyuboglu[10] Amirata Ghorbani[10] Tariq Kane[3] Christine R. Kirkpatrick[11] Tzu-Sheng Kuo[12] Jonas Mueller[13] Tristan Thrush[4] Joaquin Vanschoren[14] Margaret Warren[15] Adina Williams[8] Serena Yeung[10] Ce Zhang[2] James Zou[10] Carole-Jean Wu[8] Cody Coleman[3] Andrew Ng[7] Peter Mattson[9] and Vijay Janapa Reddi[1]

[1]Harvard University [2]ETH Zurich [3]Coactive.AI [4]Hugging Face [5]Landing.AI [6]DeepLearning.AI [7]ML Commons [8]Meta [9]Google [10]Stanford University [11]San Diego Supercomputer Center, UC San Diego [12]Carnegie Mellon University [13]Cleanlab [14]TU Eindhoven [15]Institute for Human and Machine Cognition

MLPerf

Benchmarks for

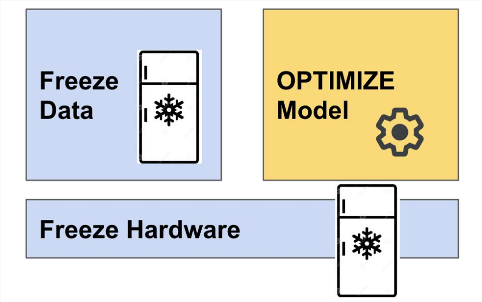
Hardware



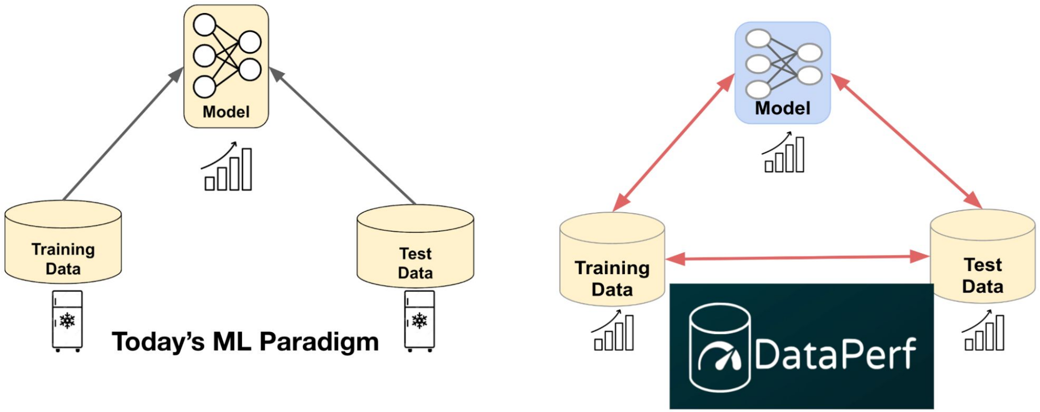
kaggle

Benchmarks for

ML

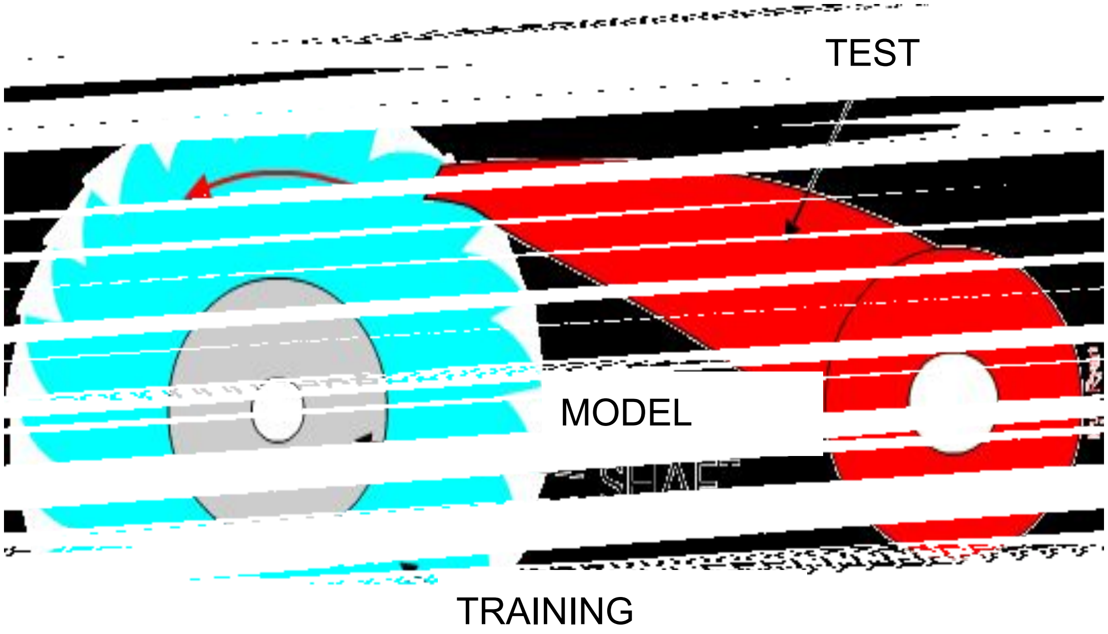


Benchmarking Data and Data-Centric Algorithms



What is the purpose of the DataPerf WG?

- Enable data-centric AI by open rigorous data benchmarks
- Create a virtuous cycle of progress in models, training and test



DataPerf Challenges

Training Set Selection - Vision

Challenge: Design a data selection strategy that chooses the best training set from a large candidate pool of training images.

Task: Image classification
Dataset: Custom subset of the Open Images Dataset

Training Set Selection - Speech

Challenge: Design a data selection strategy which chooses the best training set from a candidate pool of spoken words.

Task: Keyword spotting
Dataset: The Multilingual Spoken Words Corpus

Training Set Debugging - Vision

Challenge: Design a data cleaning strategy that chooses samples to relabel from a noisy training set.

Task: Image classification
Dataset: Custom subset of the Open Images Dataset with noisy labels

Dataset Valuation

Challenge: Design a data acquisition strategy that chooses data subsets from multiple data-sellers

Task: Sentiment Analysis
Dataset: Custom subset of the Amazon review, the YELP review, the IMDB review, and the sentiment 140 dataset

Challenge 1: Vision | Training Data Selection

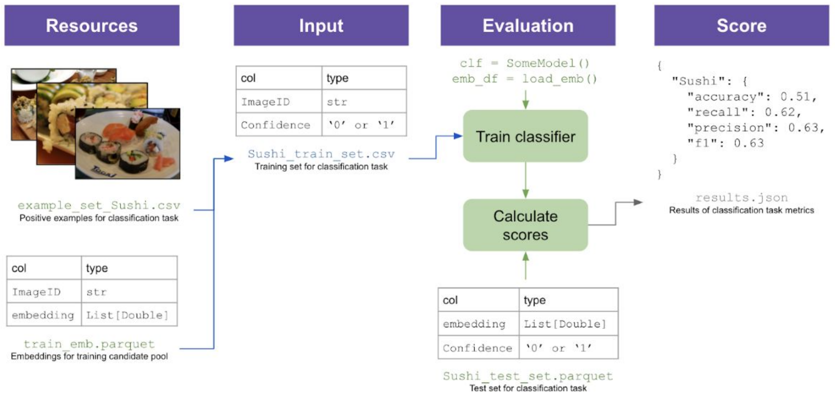
By William Gaviria Rojas and Cody Coleman (Coactive AI)

Challenge: Design a data selection strategy that chooses the best training set from a large candidate pool of training images.

Evaluation: Submissions will be scored using mean average precision across a set of image classification tasks.

Link: https://dynabench.org/tasks/vision-dataperf

Status: Finished Beta testing and gathered all feedbacks, in the process of implementing the improvements



Benchmark: Training data selection

Task: Image classification

Dataset: Custom subset of the Open Images Dataset

Challenge 3: Speech | Training Data Selection

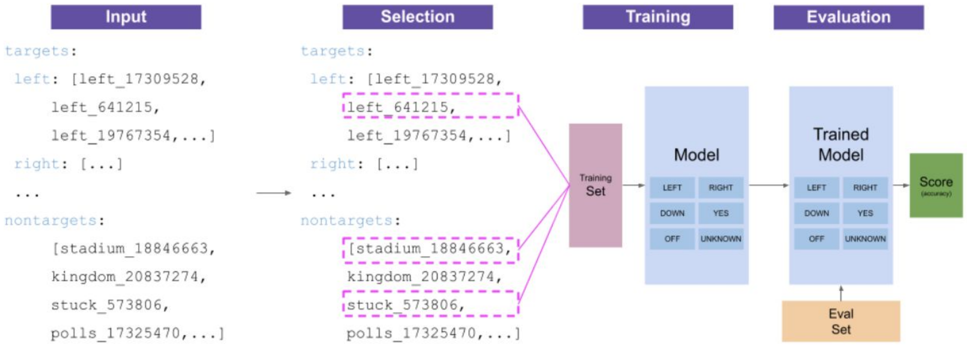
By Colby Banbury, Mark Mazumder and Vijay Janapa Reddi (Harvard)

Challenge: Design a data selection strategy which chooses the best training set from a candidate pool of spoken words.

Evaluation: Submissions will be scored using classification accuracy across a limited set of keywords.

Link: https://dynabench.org/tasks/speech-selection

Status: still running its beta, no feedback, no dynabench submission



Benchmark: Training data selection
Task: Keyword spotting
Dataset: The Multilingual Spoken Words Corpus

Challenge 2: Vision | Training Data Debugging

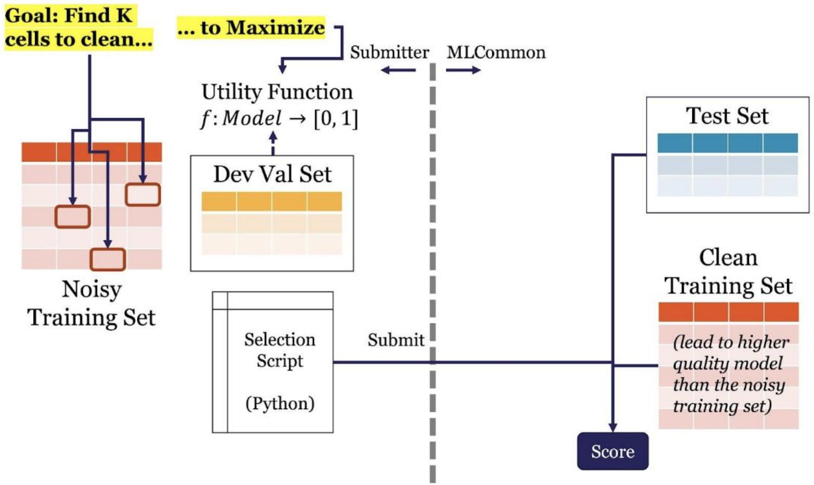
By Xiaozhe Yao and Ce Zhang (ETH Zürich)

Challenge: Design a data cleaning strategy that chooses samples to relabel from a noisy training set.

Evaluation: Submissions will be scored using mean average precision across a set of image classification tasks.

Link: https://dynabench.org/tasks/vision-debugging

Status: Still finishing beta testing, no feedback yet, some successful submission on dynabench



Benchmark: Training data label cleaning
Task: Image classification
Dataset: Custom subset of the Open Images Dataset with noisy labels

Challenge 4: Acquisition | Training Data Creation

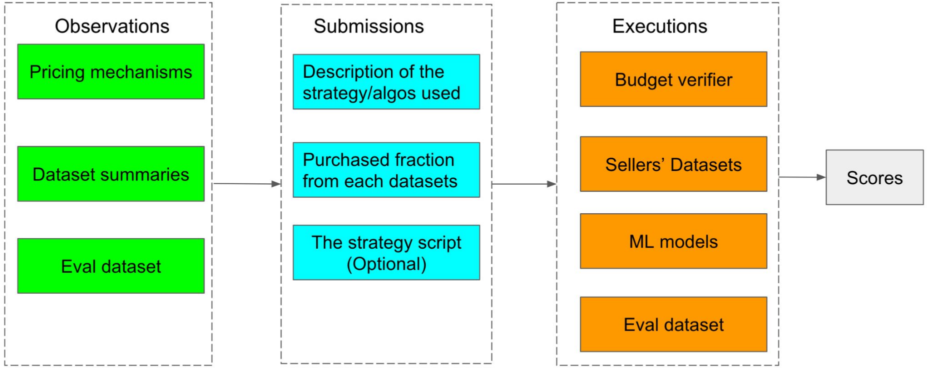
By Lingjiao Chen (Stanford), Bilge Acun, Newsha Ardalani (Meta), Sudnya Diamos (Co-active AI)

Challenge: Design a data acquisition strategy that chooses data subsets from multiple data-sellers

Evaluation: Submissions will be scored using mean average precision across a set of sentiment analysis dataset.

Link: https://github.com/facebookresearch/Data_Acquisition_for_ML_Benchmark

Status: Still in dynabench integration, should be ready for beta testing soon



Benchmark: Training data creation

Task: Sentiment Analysis

Dataset: Custom subset of the Amazon review, the YELP review, the IMDB review, and the sentiment 140 dataset

2022 Challenges Launch Status

Alpha	DynaBench Integration	Beta	Final touch	Public Launch	Results
-	- Data Acquisition	-Data Selection/Speech -Data Debugging	-Data Selection/Vision	Jan 2023	July 2023 @ ICML

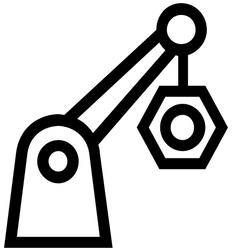
Presented DataPerf at
Neurips 2022
ML for System workshop

What’s Next in 2023?

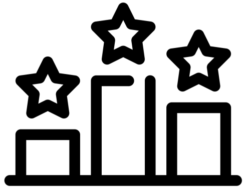
- Public launch of the 4 challenges from 2022
- Results sharing at ICML 2023
- Adaptation of high-impact datasets for ongoing challenges
- Expand DataPerf challenges to include unsupervised learning
- Academic outreach

DataPerf: Call for Action

Join the Working Group and help us design and develop DataPerf.



Sign up to participate in leaderboards and challenges in 2023.

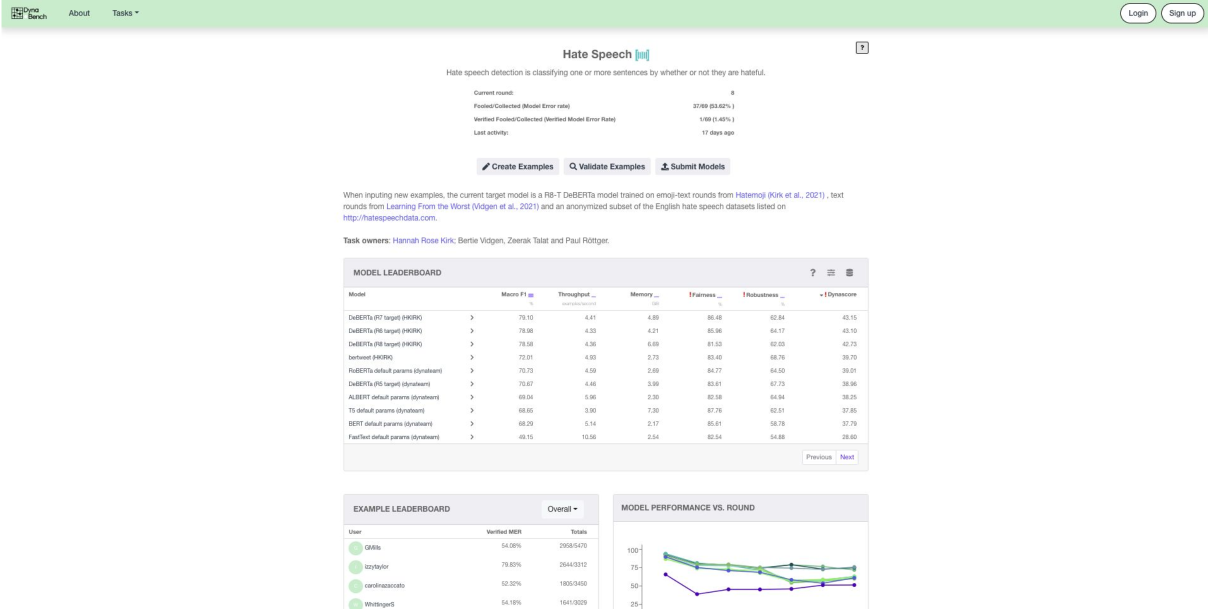


MLCommons™ Dynabench

Juan Ciro
December 08, 2022

Re branding

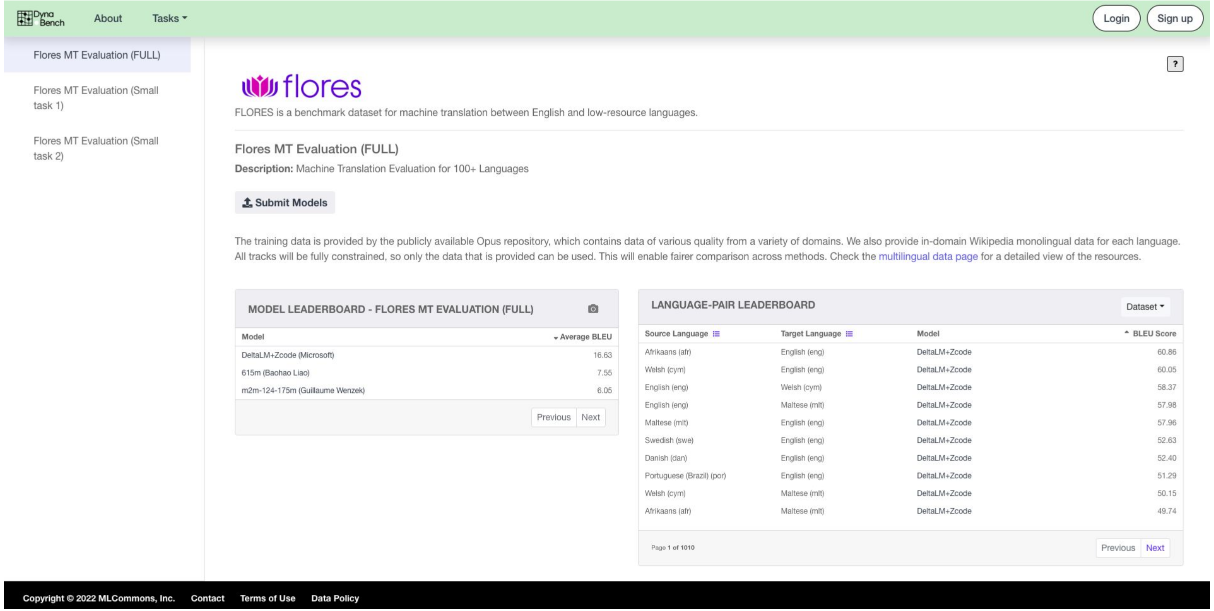
What are the new changes?



- Incorporated MLCommons style
- Improved security
- Improved UX for task owners
- New features

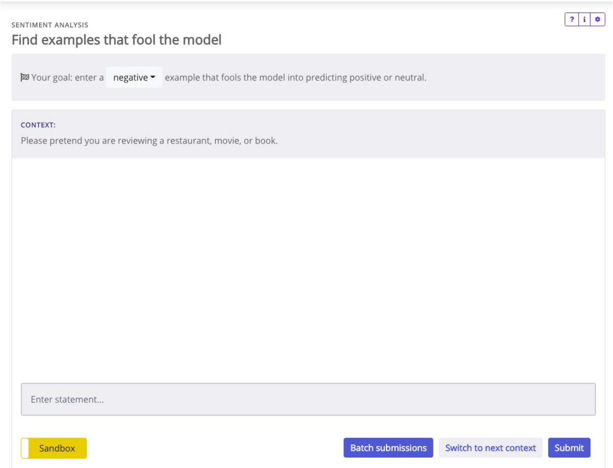
New way of structuring tasks

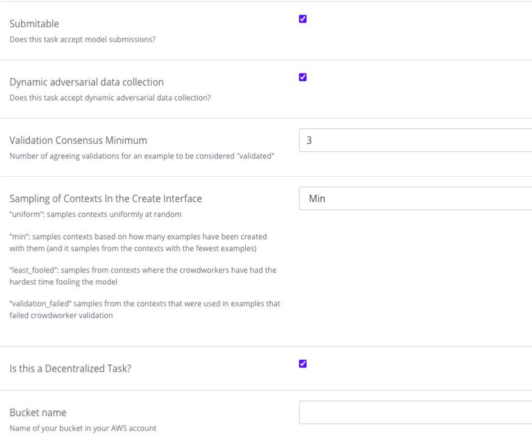
- Organize tasks by community
- Initiatives powered by Dynabench (e.g. Dataperf)



New features in the platform

- Batch adversarial sampling
- New task configurations
- New “Models in the loop” implementation
- New decentralized pipeline/tutorial





Dataperf powered by Dynabench

The Dynabench engineering team has been working on integrating the Dataperf challenges as tasks on the platform.

Here's the current status for each of them:

	Challenge Design	MLCube Integration	Dynabench Integration	Beta Testing
Vision	Done	Done	Done	Done
Speech	Done	Done	Done	In progress
Debugging	Done	Done	Done	In progress
Acquisition	Done	Pending	In progress	Pending

What's next

1. Meta RfPs are nearly done with legal agreements and funding should all be disbursed early 2023, so 5 more tasks will begin shortly

2. Counter speech coming before EoY

3. More dataperf challenges

4. More tutorials and documentation

5. Increased PR strategy seeking to grow the open source community

6. General application improvements (frontend, backend, deployment, etc.)

ML
Commons

Any ideas? Contact us!

MLCommons™
Dynabench

Science

Jeyan Thiyagalingam
Tony Hey
Geoffrey Fox

ML Commons

Science MLCommons Working Group

Science Working Group contribution to MLCommons Community Meeting November 30, 2022

Farzana Yasmin Ahmad,
Gregg Barrett,
Wahid Bhimji,
Wesley Brewer,
Cade Brown,
Bala Desinghu,
Murali Emani,
Steve Farrell
Geoffrey Fox,
Grigori Fursin,
Tony Hey,
Shantenu Jha,
David Kanter,
Christine Kirkpatrick,

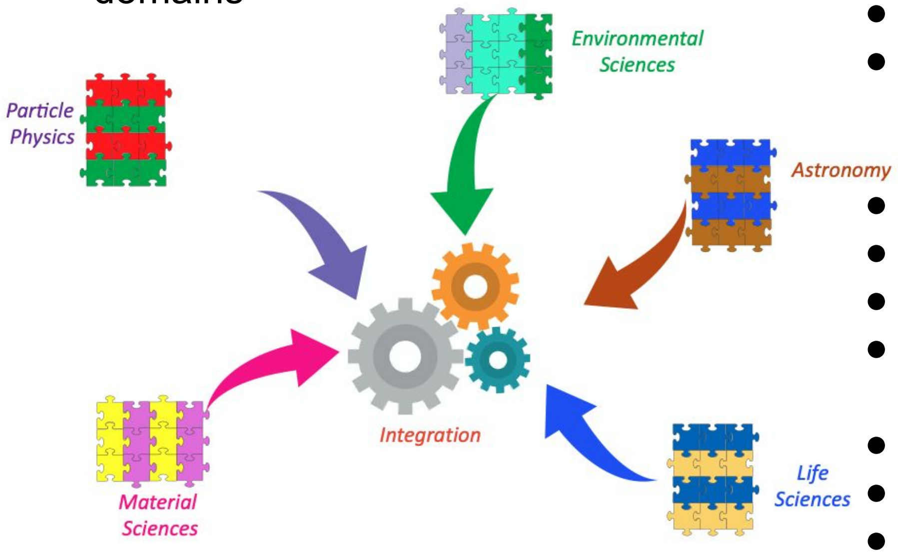
Piotr Luszczek,
Lauren Moos,
Piotr Hai Ah Nam,
Juri Papay,
Amit Ruhela,
Mallikarjun Shankar,
Dingwen Tao,
Jeyan Thiyagalingam
Aristeidis Tsaris,
Gregor von Laszewski,
Feiyi Wang,
Junqi Yin,

Tom Gibbs,
Rushil Anirudh,
Hyojin Kim,
Sergey Samsonau,

attended September 2021 – December 2022 meetings; Purple co-chairs

Science Research MLCommons working group

- Science like industry involves edge and data-center issues, end-to-end systems, inference, and training, There are some similarities in the datasets and analytics as both industry and science involve image data but also differences; science data associated with simulations and particle physics experiments are quite different from most industry exemplars
- When fully contributed, the benchmark suite will cover (at least) the following domains: **material sciences, environmental sciences, life sciences, fusion, particle physics, astronomy, earthquake and earth sciences**, with more than one representative problem from each of these domains



- https://mlcommons.org/en/groups/research-science/
- One aim is to provide a mechanism for assessing the capability of different ML models in addressing different scientific problem
- i.e. **one benchmark measure is Scientific Discovery**
- Cover rich range of problem classes
- “End-to-end” is one class
- Provide common environment to store and run benchmarks (Software)
- **4 Initial Benchmarks (2 from DOE labs, 1 UK, 1 UVA**)
- Surrogates Included (1 from LLNL next round)
- Lead use of FAIR metadata for MLCommons

Science-based Metrics

- Metrics will include those measuring **performance on science discovery,** e.g., could be one or more of:
  - Accuracy achieved
  - Time to solution (to meet a specific accuracy target)
  - Top-1 or Top-5 score
  - Chance your home will suffer a big earthquake …..
- Goal of our benchmarks is to **stimulate development of new methods relevant for scientific outcomes**. We aim to:
  - Offer well-defined “science data” sets
  - Provide a reference implementation - to help others overcome any format/interpretation/usage hurdles
  - Specify target benchmark metrics (to outperform)
  - Require a description of the improved method or code used by respondents
- *The science data should have enough richness to allow experimentation with innovative approaches.*
- Also allow **traditional system performance benchmarks**

Benchmark	Science	Task	Owner Institute	Specific Benchmark Issues
CloudMask	Climate	Segmentation	RAL	Classify cloud pixels in images
STEMDL	Material	Classification	ORNL	Classifying the space groups of materials from their electron diffraction patterns
CANDLE-UNO	Medicine	Classification	ANL	Cancer prediction at cellular, molecular and population levels.
TEvolOp Forecasting	Earthquake	Regression	Virginia	Predict Earthquake Activity from recorded event data
ICF or Inertial Confinement Fusion	Plasma Physics	Simulation surrogate	LLNL	There are other possible LLNL benchmarks from collection of 10

Benchmark contains Datasets, Science Goals, Reference Implementations; hosted at SDSC or RAL
Specification of 4 Benchmarks https://github.com/mlcommons/science

Benchmark Status

- First 4 benchmarks have been loaded into MLCommons GitHub with datasets, reference implementations and goals https://github.com/mlcommons/science
- All these benchmarks have Apache 2.0 licenses
- All have been run on multiple platforms
- Focus is Open Division with Science Discovery Metric
- The benchmarks are ready although continuing work on text suggesting approaches that are promising for improving Science Discovery Metric
- Science WG Policy and Submissions Document complete
  - Testing logging with trial open-division submissions
- Announced at SC22 Meeting Dallas November 16,2022
- Expect first submissions in next 2 months

Rolling Submissions Policy

- Submissions can be made to the MLCommons Science GitHub at any time for any benchmark that has been released.
- Submissions will be automatically checked and then reviewed by the working group which has a review committee for each benchmark.
- Depending on the number of scientific innovations in the submission, the review time will vary.
- The submitters will get an automatic acknowledgment on submission and a customized response from the committee within a week of the submittal date.
- This second response will indicate the estimated time for a committee review to be completed.
- On completion of the committee review, all submissions that are considered in scope will be posted on the working group GitHub which includes a scientific discovery "leaderboard" for each benchmark.
    - Updates will be summarized quarterly
- The innovations will be described and can include aspects other than final accuracy
    - e.g. the submission might need a smaller dataset to achieve an interesting accuracy. It is expected that benchmarks will be posted for at least a year so as to gather a rich set of input.

Science WG Benchmark Status/Futures

- ML (currently deep learning) will transform most scientific fields
  - e.g. new book “AI for Science” has 40 chapters
- ML approaches can be grouped so methods are cross-science domain
  - Time Series e.g. Earthquake, Tokamak electron density, Weather, Stocks, Particle motion
  - Simulation surrogates
  - Map properties to capabilities (chemistry to drug, materials to manufacturing value)
  - Image analysis (note can have time series of images so categories mixed)
  - GAN scenario generation
  - Publication/science text
  - Control systems (networks, accelerators, tokamaks)
- Could collect 1-4 examples in each group; explore Foundation/giant models
  - Recent meetings discuss new benchmarks (Surrogates, CFD microstructure, Autotuning, Particle Physics Inference, Radiology)
- Join Working group https://mlcommons.org/en/groups/research-science/ at https://mlcommons.org/en/get-involved/
- See minutes at https://docs.google.com/document/d/167m7FK6-Ud4M5gXta5cIc1hKqaRHkk2B1GyKasdeQLc/edit?usp=sharing

Lunch Break Welcome back at 1 PM Pacific time

Schedule

9:00 AM	Breakfast
9:30 AM	Welcome
10:00 AM	MLC Update
10:30 AM	Break
11:00 AM	Algorithms
11:10 AM	Medical
11:20 AM	Best Practices
11:30 AM	DataPerf
11:40 AM	Dynabench
11:50 AM	Science
12:00 PM	Lunch

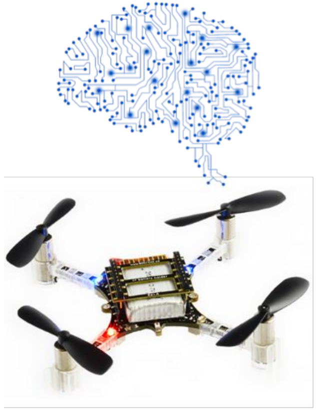
1:00 PM	Tiny
1:10 PM	Mobile
1:20 PM	Datasets
1:30 PM	Inference
1:40 PM	Training
1:50 PM	HPC
2:00 PM	Break
2:30 PM	Storage
2:40 PM	Power
2:50 PM	Benchmark Infra
3:00 PM	Research
3:10 PM	Closing Discussion

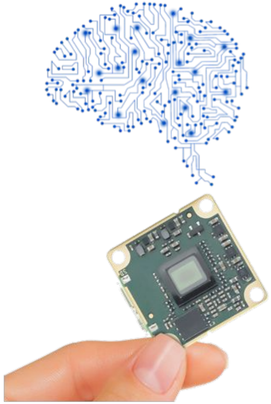
Tiny

MLPerf Tiny

What we are

- A benchmark suite for ultra-low-power ML systems (TinyML)
- On-device real-time batch-of-one inference.
- Measure energy/inference and latency on 4 different models





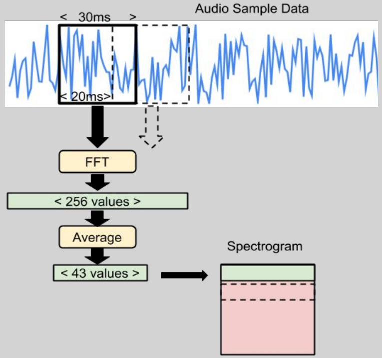
Typical Systems

- MCUs, some accelerators
- 10s-100s MHz
- ≦ MB Flash, SRAM
- ~mW Power

- Lightweight Models (<1M Param)

MLPerf Tiny Reference Models - Single Stream only

Keyword Spotting

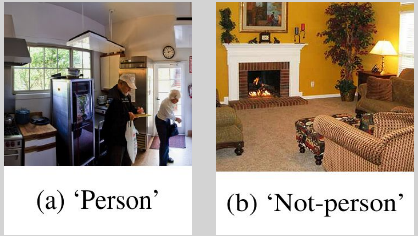


Google Speech Commands

DS-CNN
(52 Kpar)

Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

Visual Wake Words

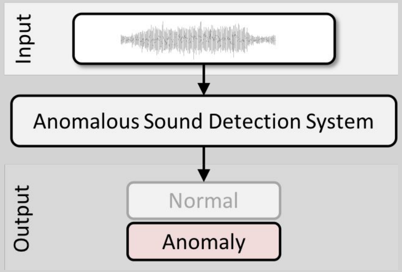


Visual Wake Words Dataset

MobileNetV1 .25
(325 Kpar)

Chowdhery, Aakanksha, et al. "Visual wake words dataset." *arXiv preprint arXiv:1906.05721* (2019).

Anomaly Detection

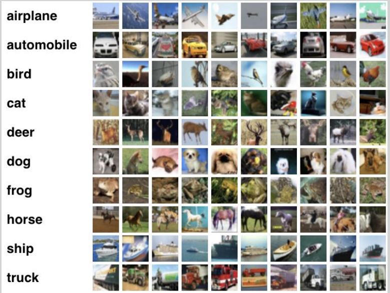


DCASE2020-Task2 / ToyADMOS

FC-AutoEncoder
(270 Kpar)

Yuma Koizumi, Shoichiro Saito, Noboru Harada, Hisashi Uematsu and Keisuke Imoto, "ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection," in Proc of WASPAA, 2019.

Image Classification



CIFAR10

ResNet8
(96 Kpar)

Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

MLPerf Tiny Reference Platform

ST NUCLEO-L4R5ZI

- Open platform
- Widely known, available, and affordable
- ARM Cortex-M4 with FPU
- 2 MB Flash
- 640KB SRAM

\+

- Open source
- Portable reference implementation

\+

- Open source
- Full toolchain with everything needed
- Continuously improving quantization support

Latest Round

- Published November 9
- Submitting Organizations: 8
  - Including 3 new submitters
- Systems Submitted: 17
  - 11 w/ Energy
- Good variety of hardware represented: Arm, RISC, FPGA, Custom Accelerators

What's Next for Tiny WG

- **MLPerf Tiny v1.1** (3rd round of submissions)
    - Submissions deadline May 19, 2023.
    - Public Release June 21
        - Shortly before TinyML Summit Europe
- **New Benchmark** in progress
    - Streaming workload – typical of many tiny applications (e.g. wakeword)
    - Sustained inference on a continuous time-series
    - Exercise rapid duty-cycling for energy efficiency
    - Add RNN to the benchmark suite

Interested in Submitting?

- Join the working group! We meet Mondays at 9am PT

https://groups.google.com/a/mlcommons.org/g/tiny

- Contact the WG chairs:

Jeremy Holleman (jeremy@syntiant.com)

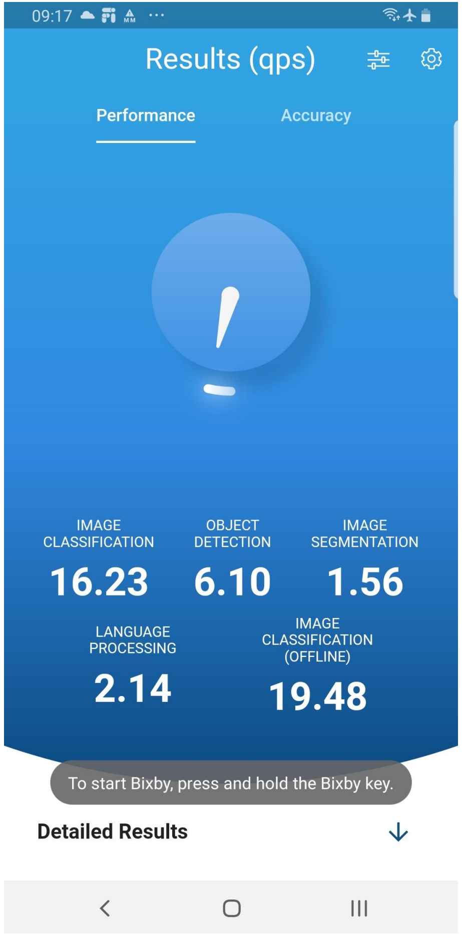
Csaba Kiraly (csaba@mlcommons.org)

Mobile

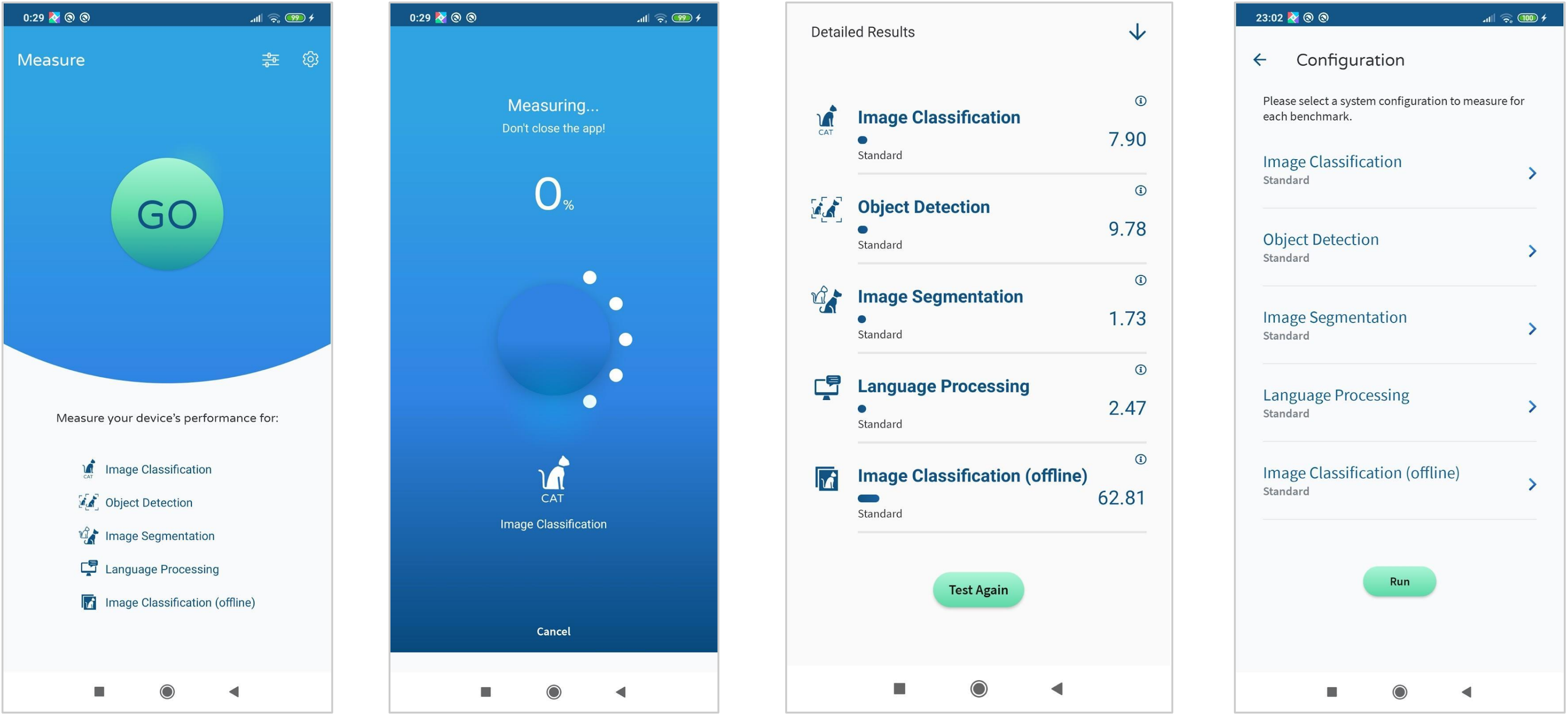
Mobile Group

WG Purpose:

- Develop a performance-accuracy benchmark suite for consumer mobile devices (phones & laptop) with different AI stacks
- Goal: Allowing general public to examine the AI performance of their devices
  - Providing a MLPerf benchmark app



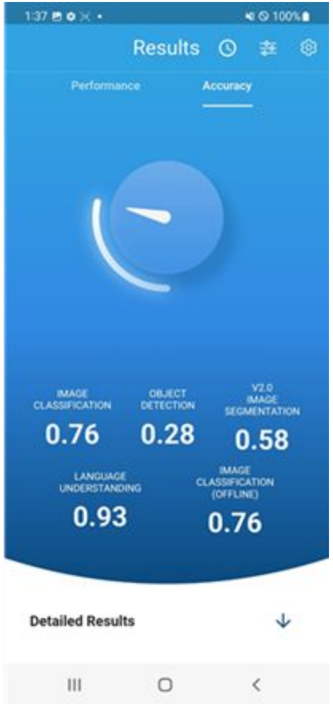
MLPerf Mobile Benchmark Application

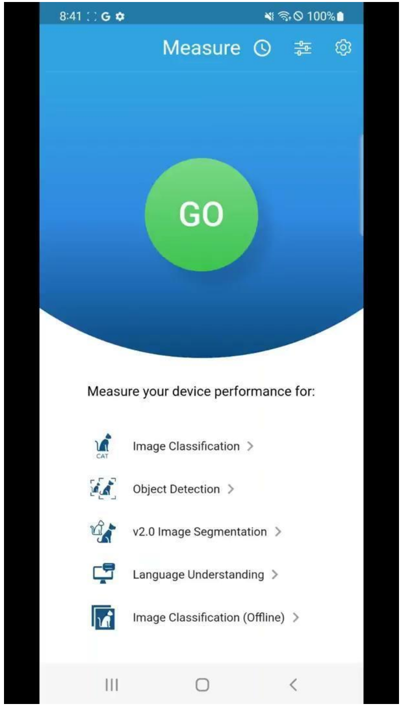


MLPerf App

- Supports Android, IOS, and Windows Operating Systems

  Running latest v2.1 Flutter App on Android Phone:





Updates From Last Community Meeting

- V2.1 MLPerf App Release
  - Support Qualcomm Snapdragon 8+ Gen1 & Samsung Exynos 2300

Upcoming Activities

- New Super Resolution model for v3.0
  - Thanks to Seoul National University's help on this
- Windows Flutter support for v3.0
- Long term: score collection & default runtime

Default Runtime

- Allowing devices that the members did not submit to still be able to run MLPerf
  - Very important to meet the press demand
  - Important for making the app on Play store
- Current proposal: use TFlite Delegate
  - For low tier devices without NPU, use CPU/GPU delegate instead
  - Will need consensus from all members
- Target Q4’2023

Score Collection

- Allowing benchmark app user to submit results to MLC website
  - The results are also visible from the benchmark app
- Target Groups
  - For the users who want to submit MLPerf benchmark result to MLC website
  - For academics, press, OEMs, and SoC vendors who want to refer to the results from different devices
- Benefits
  - More traffic for the MLC website
  - Great source to see how devices perform over time
  - Less burden to the submitters
    - Collect much more devices results
    - Allowing off-cycle submission

What’s ahead

Feb 2022 v2.0

New features and support expansion

- New segmentation model

Aug 2022 v2.1

Cross Platform Enablement

- Porting Android implementation to flutter
- Core ML support

Mar 2023 v3.0

New features and Cross platform support

- New SR model
- Windows Flutter support

Aug 2023 v3.1

Increase adoption

- Internal launch of the score collection website
- Having MLPerf app to use a default runtime on all mobile devices

Interested in Participating?

- Join the working group! We meet Wednesday at 3 PM PT

- Contact the WG chairs:
  - William Chou
    - William@mlcommons.org
  - Manasa Kankanala
    - manasa@mlcommons.org

Datasets

Rafael Mosquera
December 08, 2022

People's Speech Update

1. Faster download: the dataset is now hosted on Hugging Face to improve download speed.

2. Higher quality: the threshold for character error rate has decreased from 20% to 10% thereby improving overall quality of the clean dataset

3. Better file structure

4. Better documentation: we included a tutorial on how to train an ASR model using NeMo and GCP! It can be found here: https://github.com/mlcommons/peoples-speech/blob/main/model-training/tutorial.md

5. To download People's Speech go to: https://mlcommons.org/en/peoples-speech/

The Dollar Street Dataset

1. New public dataset, containing 38,479 images of household items from around the world

2. Every image contains geographical information (country, region), the topic it belongs to (home, TV, shower, bed, etc.), and a proxy for household income derived from the self-reported consumption as well as the income levels

3. Data was manually collected and verified

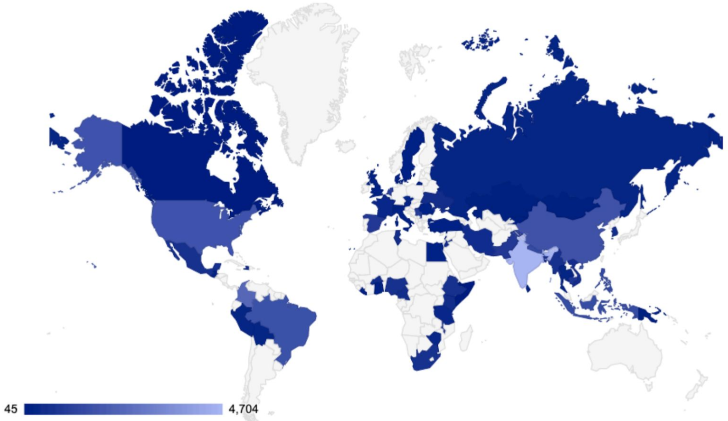
4. Includes homes with no internet access

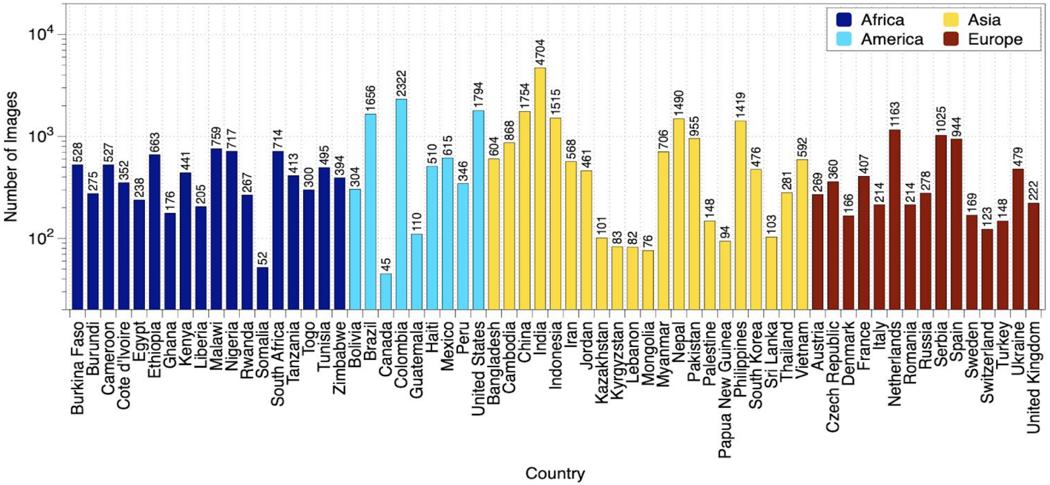
5. All data is licensed under CC-BY 4.0

6. Currently hosted on Kaggle at https://www.kaggle.com/datasets/mlcommons/the-dollar-street-dataset

William A Gaviria Rojas, Sudnya Diamos*, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, Cody Coleman*

Geographic representation





Number of images per country

The Dollar Street dataset contains collected from homes in 63 countries in four regions of the world: Africa, America, Asia and Europe.

Join Datasets!

https://mlcommons.org/en/groups/datasets/

Google group link: Datasets Google Group

- Join group to be
  - 1. invited to the weekly meetings (Thur 11-12am PT)
  - 2. Receive emails from the email list

- Interested in helping? Contact one of the WG chairs, Peter Mattson or Greg Diamos

Inference

Inference v2.1 Submission

- Inference benchmarks for Data Center and Edge.
- **Submitters**: Alibaba, ASUSTeK, Azure, Biren, Dell, Fujitsu, GIGABYTE, H3C, HPE, Inspur, Intel, Krai, Lenovo, Moffett, Nettrix, NeuralMagic, NVIDIA, OctoML, Qualcomm, SAPEON, Supermicro
- **Results**:
  - Over 5,300 performance results, 1.4x more compared to v2.0
  - >2,400 power measurement results, 1.1x more compared to v2.0
- **New**:
  - Reinanet Object detection benchmark
  - Inference-over-network scenario
- Inference results are available at https://mlcommons.org/en/inference-datacenter-21/ and https://mlcommons.org/en/inference-edge-21/

Issues and lessons learned from v2.1 submission

- Multiple sources of information for new models *[refer to github rules as final source, only look at github main branch]*

- No mechanism to process last minute emergency changes in a systematic way. *[Special WG session day before deadline to review emergency changes]*

- Late and incomplete submissions create uncertainty and hard feelings. *[tighten the rules and more education]*

- Very few network submissions. *[effort underway to get feedback from sub-WG members to fix this issue in next submission cycle]*

ML Inference Network Division

- A new Network division was introduced in v2.1
  - Number of submissions was low, format not as expected
- Corrective plan introduced for v3.0
  - How to make it easier for submissions?
    - Provide a simple reference QDL implementation (An working LON Demo). Show rules implemented
  - Clarify submission rules
    - What can run in LON node, what must run in remote SUT
    - “SUT over the Network” and “remote attached DLA” submissions
    - Separate QDL class from SUT class to avoid confusion
    - Clarify and add required accompanying submission material
  - Support for Multi-Node Network submission

Looking ahead in 2023

Next submission (v3.0) - March 3, 2023

Highlights

- Large Language Model
  - Will be based on Training WG LLM benchmark
  - Plan to intercept in v3.1
- DLRM v2.0
  - Plan to update the benchmark in v3.1
- Automotive benchmark in v3.1
- Fix the issues observed in v2.1

Taskforce on education and reproducibility: progress report

https://bit.ly/modular-mlperf Weekly conf-calls: every Thursday 11am PST Grigori Fursin and Arjun Suresh (OctoML)

Mission:

- Make it easier to run and reproduce MLPerf inference benchmarks across any software and hardware stack.
- Automate optimization, design space exploration and comparison of ML Systems.

Progress since October 2022:

- Released stable MLCommons CK2 v1.1.3 to decompose MLPerf into portable and reusable scripts with a unified API.
- Implemented **modular CK2 workflow and container** to automatically plug different ML engines, models and datasets to MLPerf.
- **Successfully validated this CK2-MLPerf workflow at the Student Cluster Competition at SuperComputing'22**: 10 teams managed to prepare end-to-end MLPerf inference submission in less than 30 minutes using this tutorial!
- Added GitHub actions with the CK2 workflow to the MLPerf inference repo to automatically test various reference implementations.
- Prototyped CK2 automation for power measurements and TinyMLPerf submissions (v1.0).
- Prototyped universal & modular CK2 wrapper for loadgen to measure performance of any model with any hardware.

The next steps:

- Extend modular CK2 workflows and prepare tutorials to reproduce all past MLPerf inference submissions (Nvidia, Qualcomm, Intel, ...) across any supported platform - collaboration is welcome!
- Support other MLCommons workgroups including TinyML, Mobile, MedPerf, Research and Best Practices.
- **Collaborate with anyone interested to automate and optimize their MLPerf inference v3.0 submissions.**

Interested in joining Inference WG ?

Meets every Tuesday at 8:30 (PST).
https://mlcommons.org/en/get-involved/#getting-started

Google group link: https://groups.google.com/a/mlcommons.org/g/inference

Training

v2.1 Results!

- 200 results from 18 submitters, from small workstations to large scale data centers with thousands of processors
  - Azure, Azure-HazyResearch, ASUSTeK, Baidu, Dell, Fujitsu, GIGABYTE, HPE, Inspur, Intel, Intel-HabanaLabs, Krai, Lenovo, MosaicML, NVIDIA, Samsung, Supermicro, *xFusion (new submitter)*

- New Preview HW: Intel Sapphire Rapids and NVIDIA H100

- New SW-only submissions from multiple submitters

- No new benchmarks for Training this round. All benchmarks compatible with v2.0.

- Coincided with HPC and Tiny Publication

v2.1 Lessons learned

- What went well
  - Submission infrastructure / process
  - RCP Process
  - Spreadsheet automation
  - In general, this was a smooth review period - many cancelled meetings

- What didn't go well
  - Nothing major identified in the post mortem

v3.0 - Spring 2023

- Submission May 19, Publication June 23
- New benchmarks
  - LLM - GPT3-175B with C4 English Dataset
  - DLRM - DCNv2 with synthetically multihot Criteo dataset
- Deprecating MiniGo early
- Reference cleanup, for reals
- Power/Energy (aggressive)

v3.1 - Fall 2023

- Submission October 13, Publication November 10

- New benchmarks under consideration
  - Auto - PointPainting with latest Waymo dataset
  - Stable Diffusion or Graph Neural Networks - actively debating

- Continued reference clean up

- Power/Energy (more likely)

Join Training!

https://mlcommons.org/en/groups/training/

Google group link: training@mlcommons.org

- Join group to be
  - 1. invited to the weekly meetings (Thur 830-10am PT)
  - 2. Receive emails from the email list

HPC

Get involved

- Join the HPC group: https://mlcommons.org/en/groups/training-hpc/
- Meetings: Mondays, weekly alternating between 8-9AM PT and 3-4PM PT.
- Reach out to Murali Emani (memani@anl.gov), Steve Farrell (sfarrell@lbl.gov)

HPC Working Group

We do ML Benchmarking on supercomputer systems

- E.g. Top500 machines, national labs, vendors, etc.
- ML for large scale scientific simulations and analytics

We publish the MLPerf HPC benchmark suite

- Modeled after MLPerf Training, with some mods
- Two annual submission rounds so far
- Just done with v2.0 submissions and results announced at SC22

Other recent activities

- Presented in Birds of a Feather session at ISC'22
- Presented (by Steve) at ISC'22 workshop on Measuring Effective Performance of HPC Systems

Top500 November 2022

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
8	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
9	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
10	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT	4,981,760	61.44	100.68	18,482

MLPerf HPC rules

Two measurement types

- Time-to-train: traditional strong-scaling metric
  - Same as MLPerf Training.
- Throughput: new weak-scaling metric added in v1.0
  - Train many models concurrently to convergence
  - Measure and report time-to-train-all, and throughput (models/min)
  - Allows us to overcome scaling limitations of training applications and measure capabilities of systems of *any size*

Other deviations from MLPerf Training

- Data staging from parallel file system to node-local / accelerated storage must be included in measured time

MLPerf HPC v2.0 schedule

Feb 14, 2022

Finalize process for new benchmarks

Feb 21, 2022

New benchmark proposals

May 23, 2022

Benchmark freeze

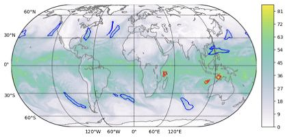
Sep 16, 2022

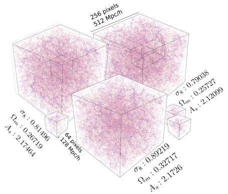
Submission deadline

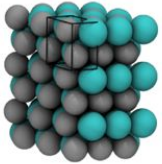
Nov, 2022

Publication, SC22

MLPerf HPC v2.0 details







Benchmark	Task	Dataset	Reference Model	Quality Target
DeepCAM	Climate segmentation	CAM5+TECA simulation, image size (768x1152x16), 8.8 TB	DeepCAM 2D CNN based on DeepLabv3+	0.82 IOU
CosmoFlow	Cosmology parameter prediction	CosmoFlow N-body simulation, 3D cubes of size 128^3, 10.2 TB	CosmoFlow 3D CNN	0.124 MAE
OpenCatalyst	Quantum molecular modeling	Open Catalyst 2020 (OC20) S2EF, 300GB	DimeNet++ GraphNN	0.036 Forces MAE

MLPerf HPC v2.0 Results

- **MLCommons press release:** https://mlcommons.org/en/news/mlperf-training-4q2022/
- **Link to results:** https://mlcommons.org/en/training-hpc-20/

- **5 Submitters:** Dell (***new***), Fujitsu+RIKEN, HelmholtzAI, NVIDIA, TACC
- **16 performance results:**
  - 9 closed division strong-scaling measurements
  - 5 closed division weak-scaling (throughput) measurements
  - 2 open division measurements
- Strong-scaling submission scale up to 2,048 GPUs
- Weak-scaling submission scale up to 4,096 GPUs (Selene), and 82,944 CPUs (Fugaku)

MLPerf HPC v2.0 Highlights

- **Performance improvements show software sophistication**
  - CosmoFlow - best strong scaling result 2x faster
    - best throughput 3x higher
  - DeepCAM - best strong scaling result 6% faster
    - best throughput 21% higher
  - OpenCatalyst - best strong scaling result 5x faster
    - *first throughput measurements!*

MLPerf HPC v3.0 schedule

Feb 27, 2022 (tentative)
New benchmark proposals

Apr 10, 2022 (tentative)
New benchmark decisions

June 12, 2022 (tentative)
Benchmark freeze

Oct 06, 2023
Submission deadline

Nov, 2023
Publication, SC23

MLPerf HPC v3.0 plans

New benchmark(s)

- Starting to consider proposals ~now
- Previously considered candidate benchmarks:
  - Drug design with transformers (almost added in v2.0)
  - Protein folding (AlphaFold2, OpenFold)

Power

- Discussions resume Dec 12
- Need to accommodate variations in HPC power measurement capabilities at sites
- Likely to propose something similar to Green500 methodology

Ideas to increase participation, add value, decrease cost

How can we create opportunities for participants to *learn* how to measure and optimize ML performance for their HPC systems?

- Tutorials (planning for ISC'23)
- Guest speakers (vendors, tools, methods)
- Cooperative effort, common goal to maximize perf on *all* systems
- All participants publish results in submission, follow-up papers

How do we maximize value of results, e.g. for publications?

- More results at more training scales
- Power measurements for efficiency studies, comparisons
- Publish full optimization story, including baseline perf + stages of optimization

How can we reduce the cost and challenges of participation?

- Cooperative effort, less competitive (submitters helping each other)
- Incentivize vendors to join and help submitters
- Increase options for asynchronous submissions

Thank you

Get involved

- Join the HPC group: https://mlcommons.org/en/groups/training-hpc/
- Meetings: Mondays, weekly alternating between 8-9AM PT and 3-4PM PT.
- Reach out to Murali Emani (memani@anl.gov), Steve Farrell (sfarrell@lbl.gov)

Break

Schedule

9:00 AM	Breakfast
9:30 AM	Welcome
10:00 AM	MLC Update
10:30 AM	Break
11:00 AM	Algorithms
11:10 AM	Medical
11:20 AM	Best Practices
11:30 AM	DataPerf
11:40 AM	Dynabench
11:50 AM	Science
12:00 PM	Lunch

1:00 PM	Tiny
1:10 PM	Mobile
1:20 PM	Datasets
1:30 PM	Inference
1:40 PM	Training
1:50 PM	HPC
2:00 PM	Break
2:30 PM	Storage
2:40 PM	Power
2:50 PM	Benchmark Infra
3:00 PM	Research
3:10 PM	Closing Discussion

Storage

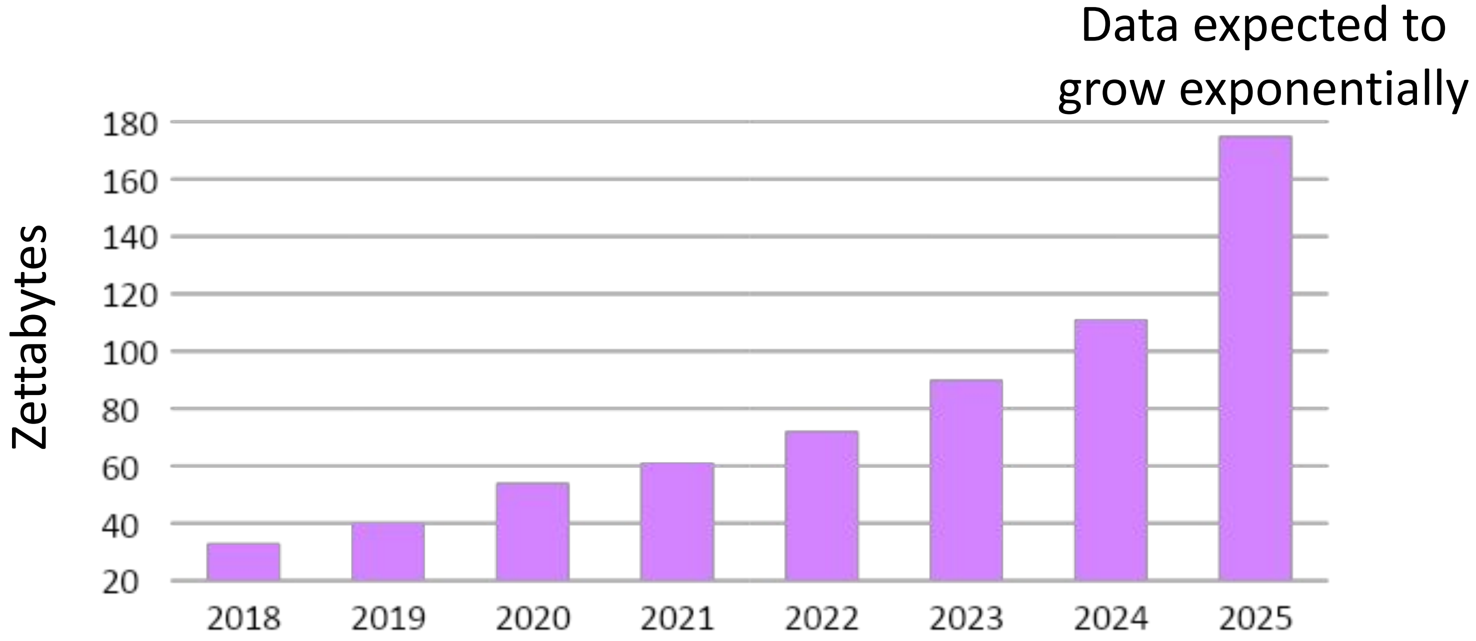
ML
C

ML Commons

MLPerf Storage

Oana Balmau, McGill University
MLPerf Storage Co-chair
December 8th, 2022

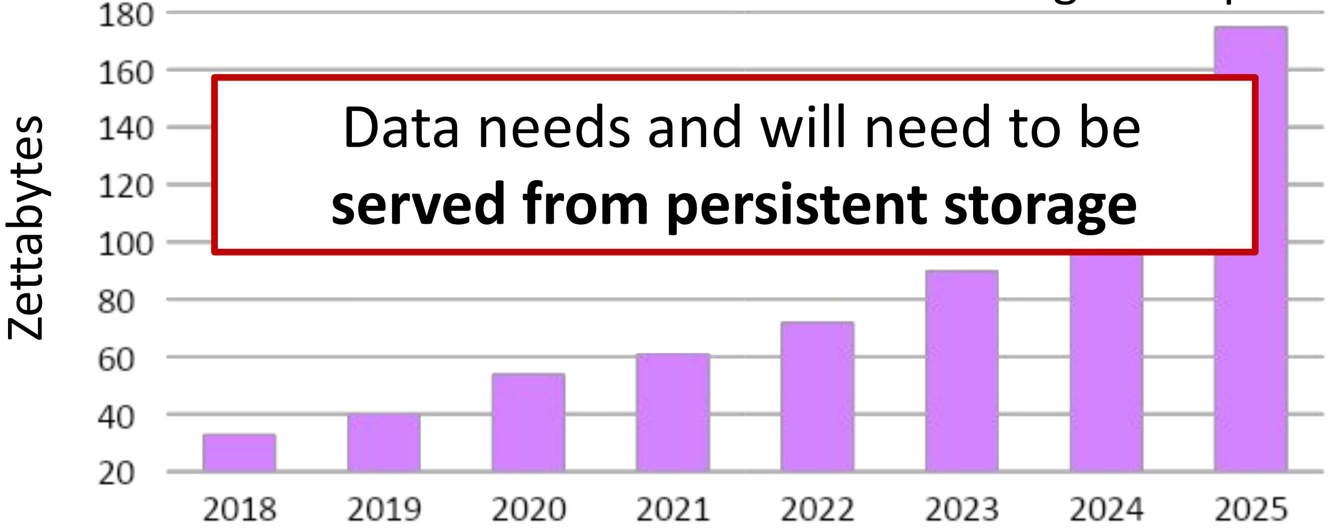
Humanity produces a lot of data



Source: IDC 2022

Humanity produces a lot of data

Data expected to
grow exponentially



Data needs and will need to be
served from persistent storage

Source: IDC 2022

Data is the moving force of ML algorithms

… but in many projects the **storage decision is an afterthought**

MLPerf Storage Goals

- **Understand storage bottlenecks in ML workloads and propose optimizations**

- **Help AI/ML researchers and practitioners make an informed *storage* decision**

Why Create an ML Storage Benchmark

Many existing ML/AI benchmarks

DeepMind Lab

MLPerf

OpenAI

DLBT

PMLDB

DAWNBench

Current ML/AI benchmarks

- Focus on **end-to-end testing**

  - hard to isolate value of each component

- Insist on **training and inference** speed

  - tend to simplify storage

  - ignore pre-processing

- **Expensive accelerators** needed to run

- Require **extensive entry knowledge**

DeepMind Lab

MLPerf

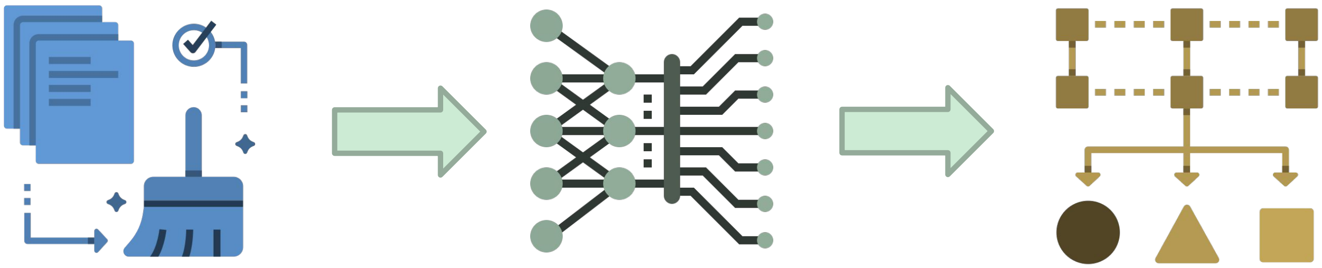
OpenAI

DLBT

PMLDB

DAWNBench

Stages of the ML Pipeline

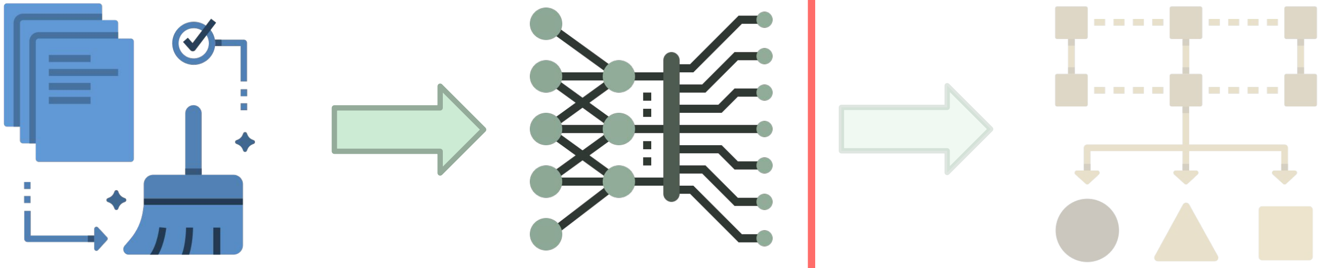


Data cleaning & pre-processing **Training** **Inference**

MLPerf Storage focus in ML pipeline

I/O intensive [1,2] – Our focus

*As much as **50% of the Watts** can go into storage and data cleaning [2]*



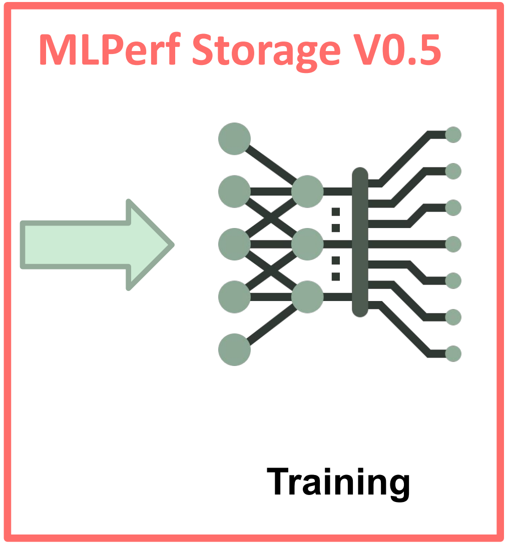
Data cleaning & pre-processing

Training

Inference

[1] Murray et al. ***tf.data: A Machine Learning Data Processing Framework***, *VLDB 21.*
[2] Zhao et al. ***Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training*** *ISCA 22.*

MLPerf Storage V0.5



- ✅ Focus on **storage impact in ML/AI**
- ✅ Realistic **storage** settings in **training phase**
- ✅ **No accelerator required** to run
- ✅ **Minimal AI/ML knowledge** required

MLPerf Storage V0.5 – workloads

Workload	Image segmentation	Natural language processing	Recommender Systems
Model	Unet3D	BERT	DLRM
Seed data	KiTS19 Set of images	Wikipedia 2020 Text	Criteo Terabyte Click logs
Framework	Pytorch	Tensorflow	Pytorch
I/O behavior	Random access inside many small files	Sequential access of small subset of files, streamed.	Random access inside one large file

https://github.com/mlcommons/storage

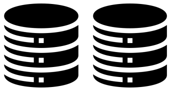
Preview package coming soon!

- **Single node**
- Many **simulated accelerators.**
- **Synthetic datasets** generated from real dataset seed.
- **Local storage**

Data pipeline in ML workload

Storage resources

Disk



Cleaned
dataset

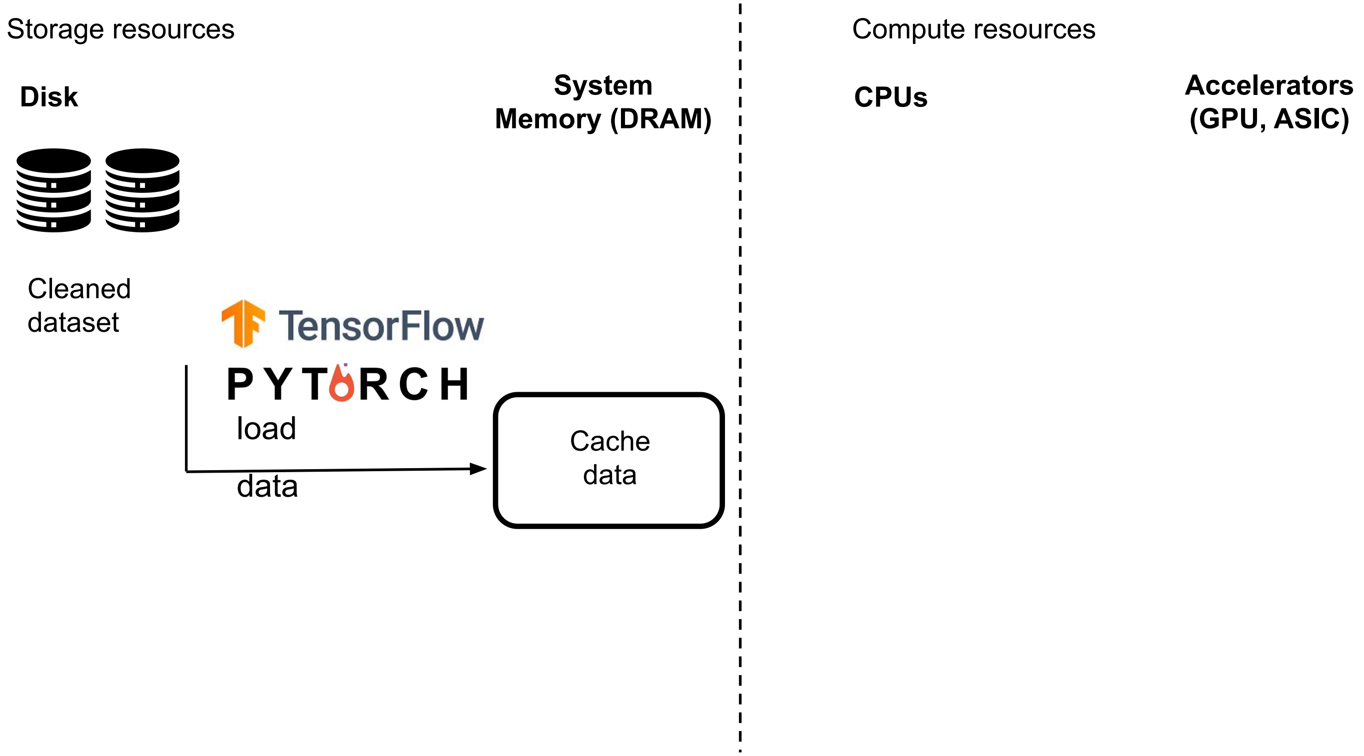
**System
Memory (DRAM)**

Compute resources

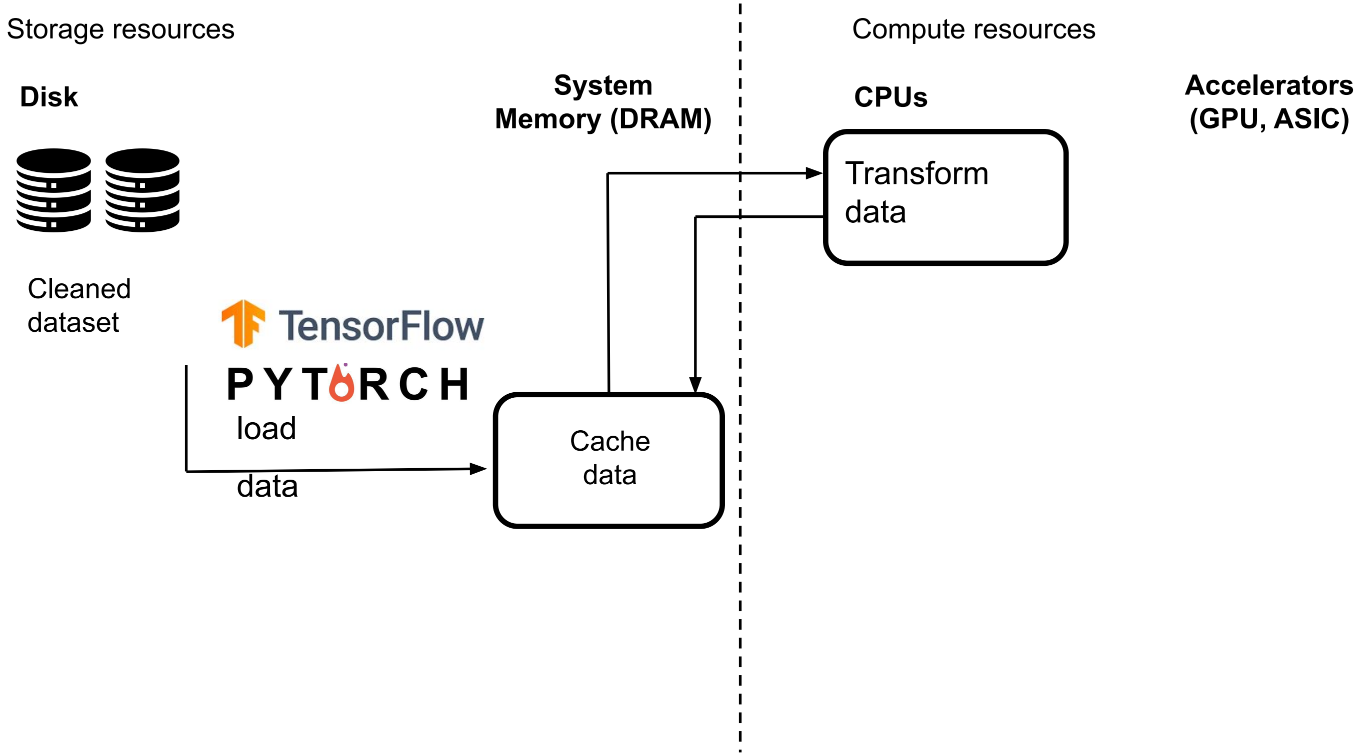
CPUs

**Accelerators
(GPU, ASIC)**

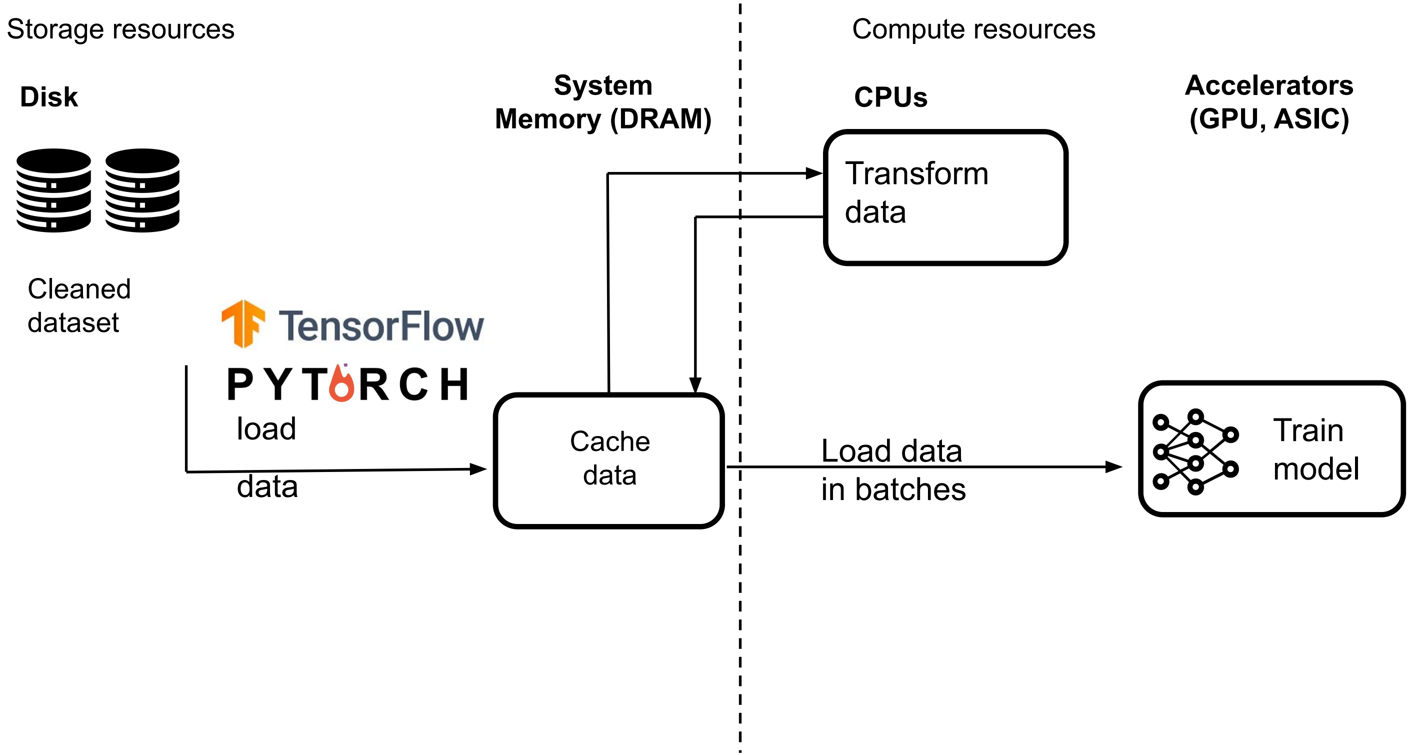
Data pipeline in ML workload



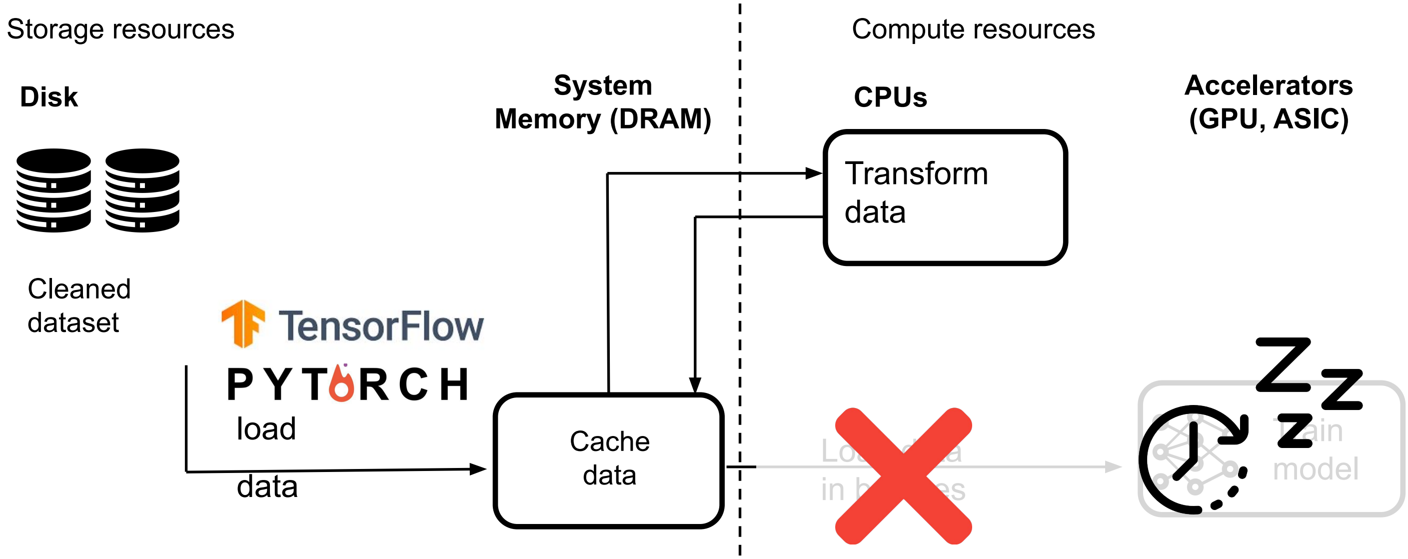
Data pipeline in ML workload



Data pipeline in ML workload



Data pipeline in MLPerf Storage benchmark

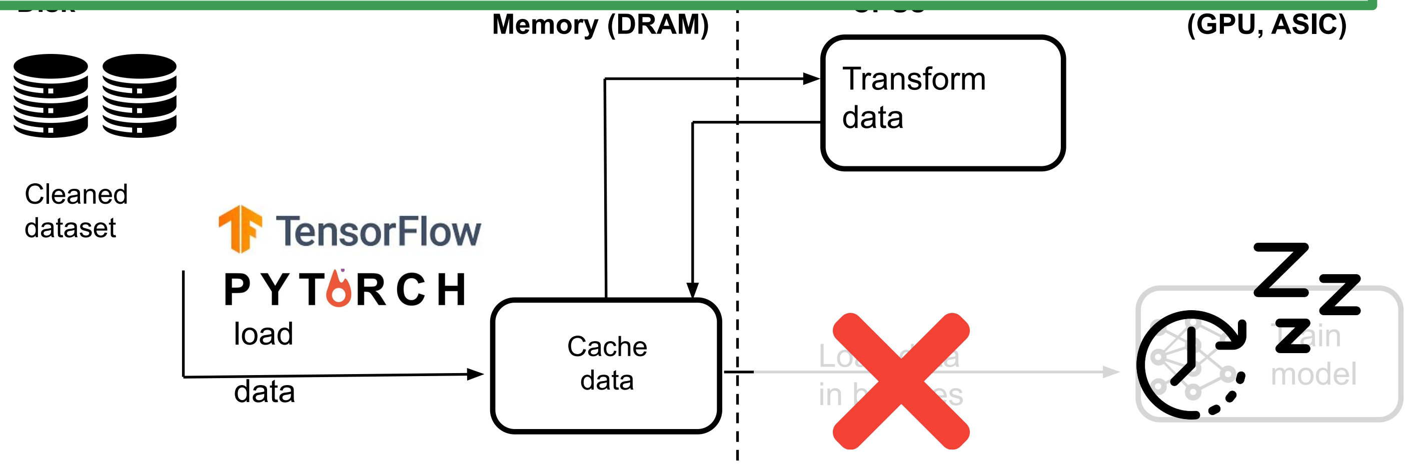


Benchmark is built as an extension of DLIO [1]

[1] H. Devarajan, H. Zheng, et al. DLIO: A Data-Centric Benchmark for Scientific Deep Learning Applications, CCGrid '21.

Data pipeline in MLPerf Storage benchmark

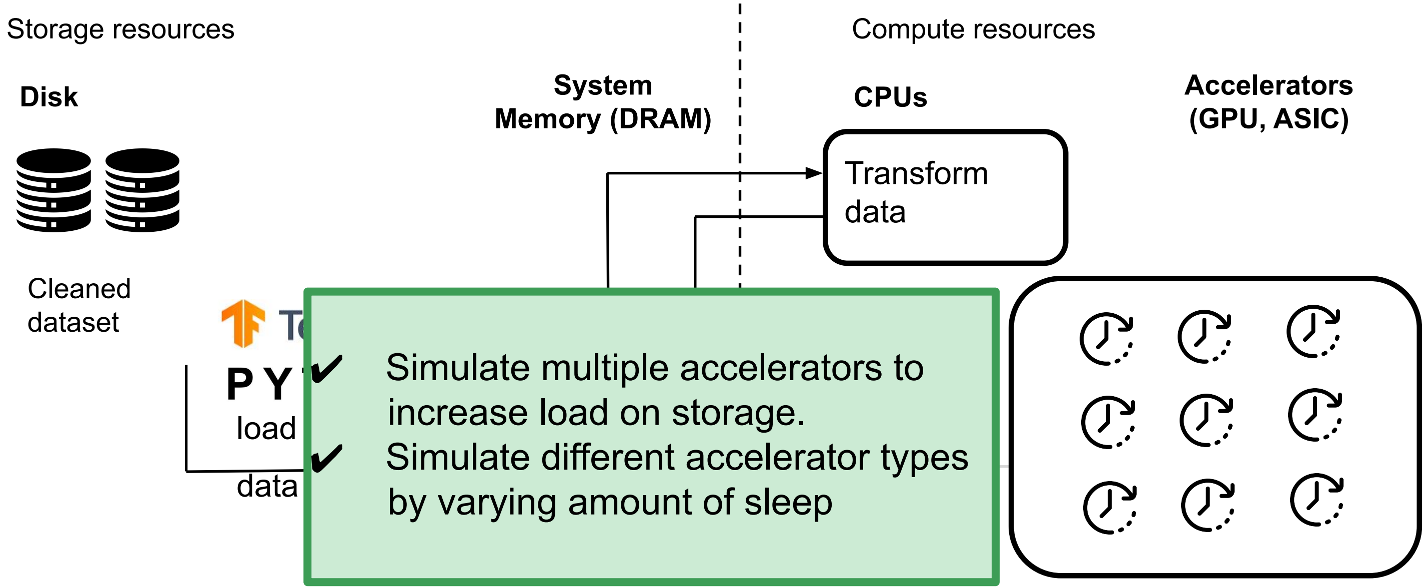
✔ Realistic storage settings: nothing changes in data pipeline, apart from training computation.



Benchmark is built as an extension of DLIO [1]

[1] H. Devarajan, H. Zheng, et al. DLIO: A Data-Centric Benchmark for Scientific Deep Learning Applications, CCGrid '21.

Data pipeline in MLPerf Storage benchmark

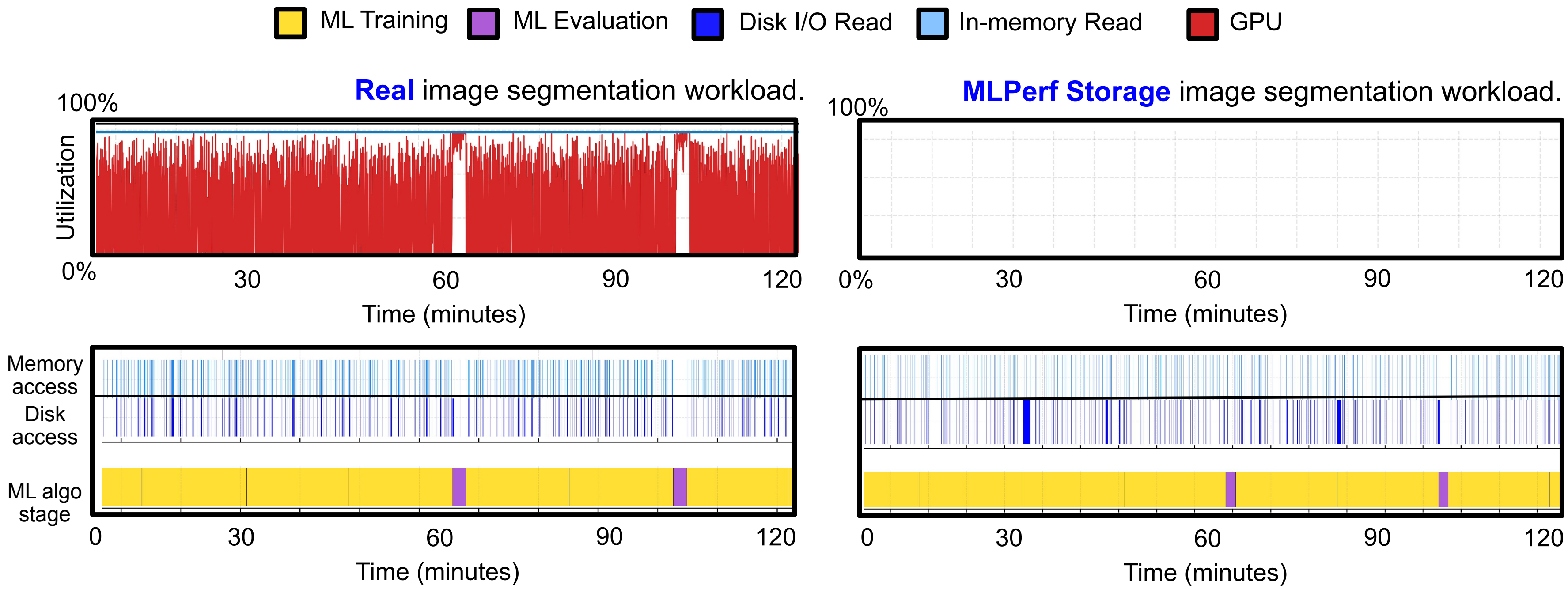


- ✔ Simulate multiple accelerators to increase load on storage.
- ✔ Simulate different accelerator types by varying amount of sleep

Benchmark is built as an extension of DLIO [1]

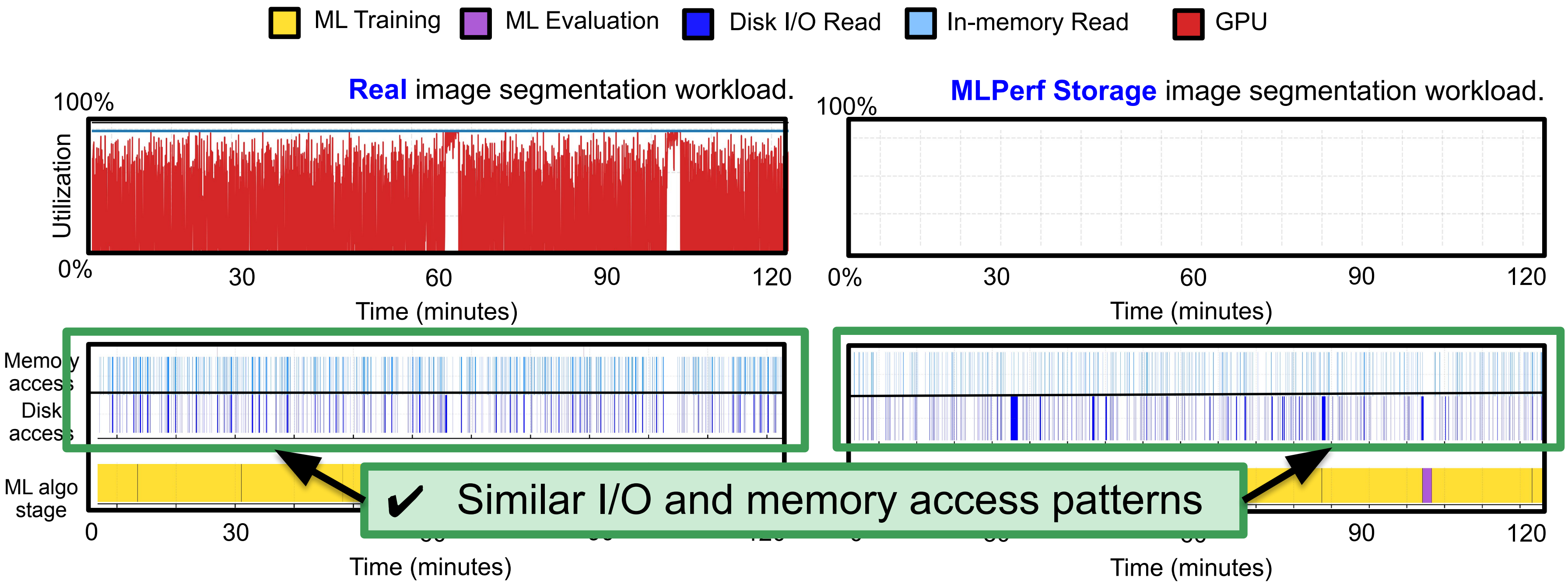
[1] H. Devarajan, H. Zheng, et al. DLIO: A Data-Centric Benchmark for Scientific Deep Learning Applications, CCGrid '21.

Simulating training time does not impact I/O patterns



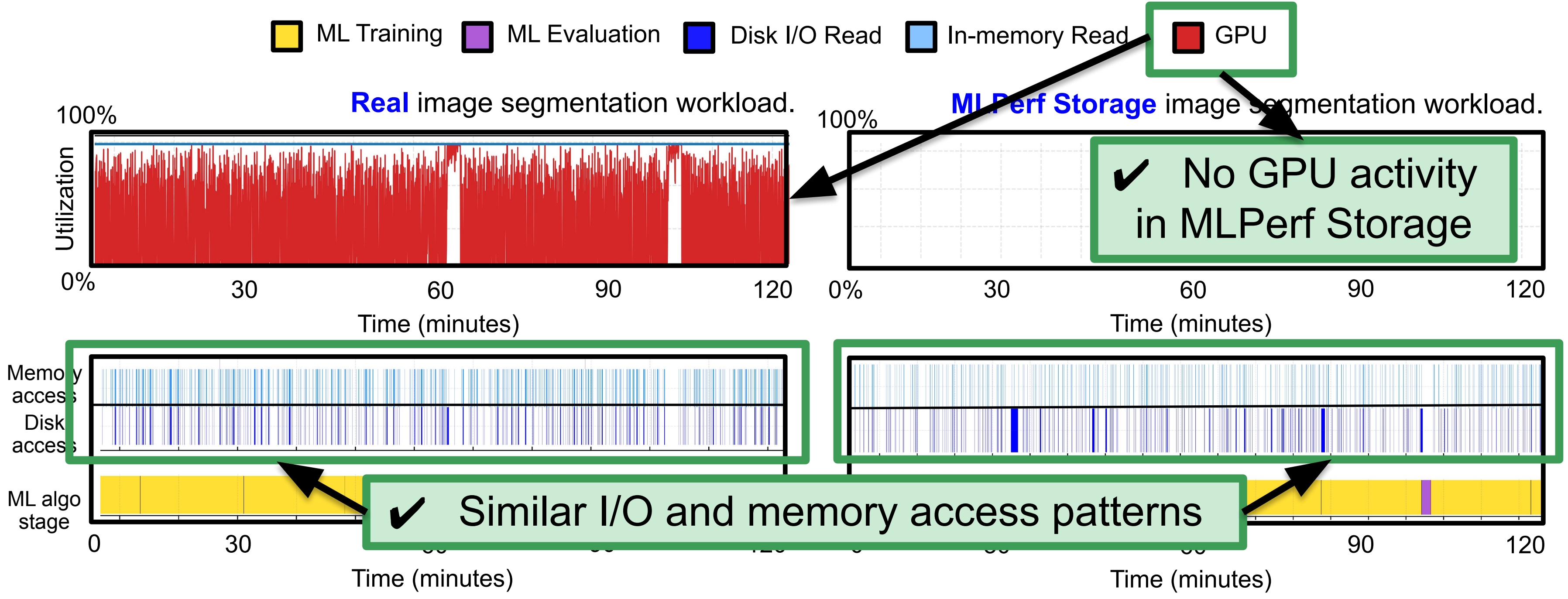
Experiment setup: DGX-1 with 8xV100 GPUs, 512GB DRAM. Dataset : KiTS19, Dataset size:Memory size ratio 2:1

Simulating training time does not impact I/O patterns



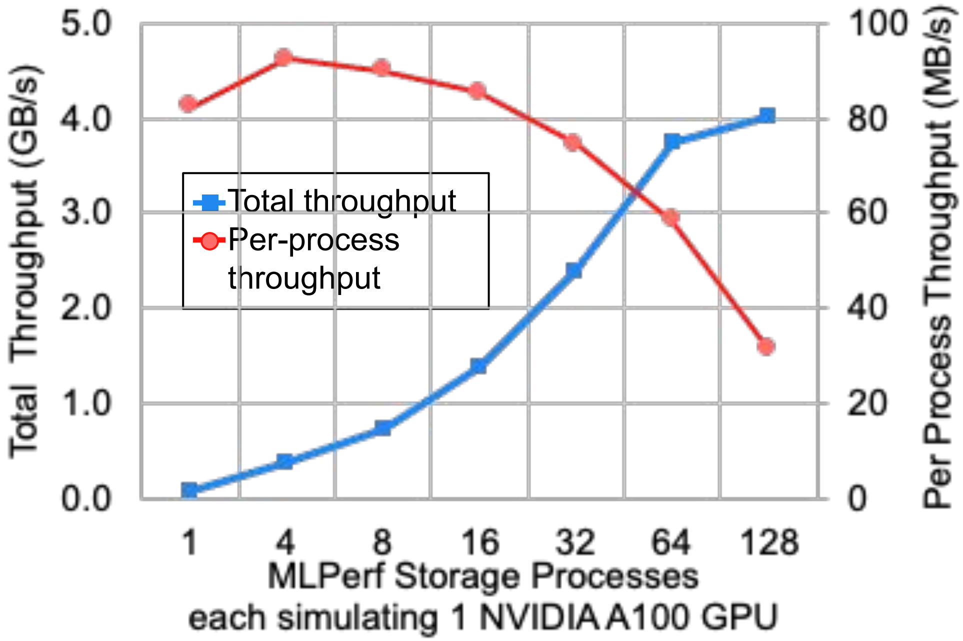
Experiment setup: DGX-1 with 8xV100 GPUs, 512GB DRAM. Dataset : KiTS19, Dataset size:Memory size ratio 2:1

Simulating training time does not impact I/O patterns



Experiment setup: DGX-1 with 8xV100 GPUs, 512GB DRAM. Dataset : KiTS19, Dataset size:Memory size ratio 2:1

Single node, single SSD Scalability: Micron Data Center



- Load is added to the storage system by adding simulated processes
- SSD reaches peak throughput with load from 128 simulated GPUs.

Image segmentation workload. Seed dataset : KiTS19, dataset size: 500GB.

Next Steps

- Release preview package: https://github.com/mlcommons/storage

- Open **MLPerf Storage** for first round of submissions: expected 2023.

- Scale simulated compute to **multi-node training.**

- Paper submission target **VLDB 23.**

Key Takeaways – MLPerf Storage

MLPerf Storage is a new benchmark

- ✅ Realistic **storage** settings
- ✅ **No accelerators required** to run
- ✅ Follow MLPerf Storage repository for updates:

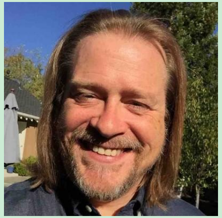
  https://github.com/mlcommons/storage

Get involved
mlcommons.org/en/get-involved/

We appreciate your feedback

Share your thoughts
Email oana.balmau@cs.mcgill.ca

Thanks to all working group co-chairs!



Curtis Anderson
Panasas

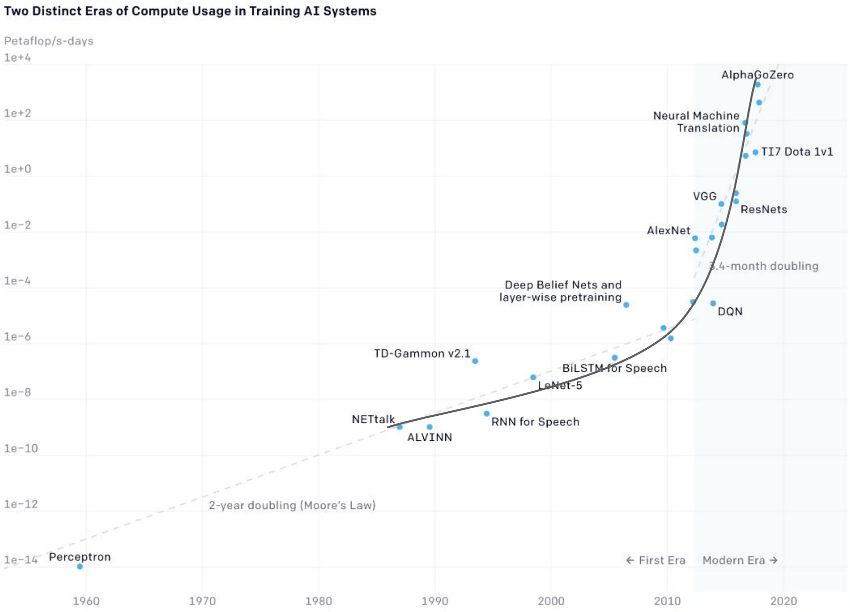
Huihuo Zheng
Argonne National Labs

Johnu George
Nutanix

ML
C

Power

Measure Energy Consumption in MLPerf Benchmark categories



Source: OpenAI

One algorithm that lets a robot manipulate a Rubik's Cube used as much energy as 3 nuclear plants produce in an hour. PHOTOGRAPH: GETTY IMAGES

Initiative to measure the energy cost of Algo+SW+HW in AI/ML applications

Power Working Group - Objective and Goals

- MLPerf Power: A Best Practices Working Group
  - Objective : Enable measuring energy consumed of submissions across Benchmark categories

- What :
  - **Measured** metric: ***System Power*** (and Energy consumed)
  - Submissions enabled for Inference (Datacenter, Edge) and TinyML
  - Key OKR for 2023 : Enable Power in MLPerf HPC and MLPerf Training

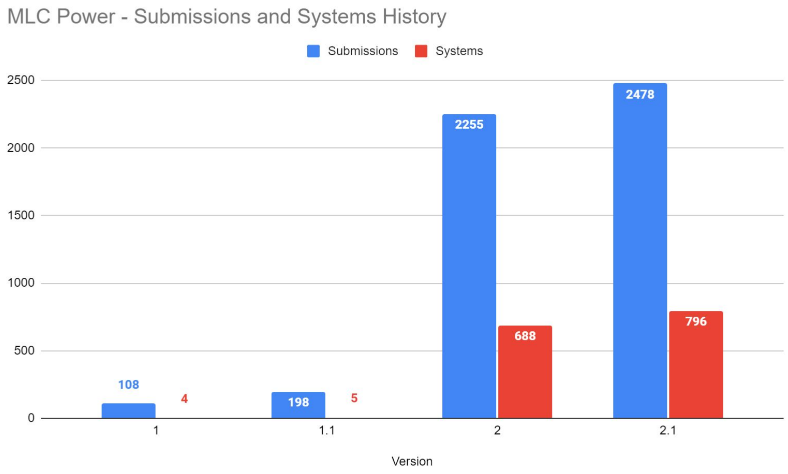
- Key Goals (Next several months) :
  - Improve the current measurement methodology for existing Inference Power (v3.0)
  - Enable adoption in distributed systems - Inference (Network category), HPC , Training
  - Increase/Encourage Inference submissions to Power
    - We have had 1.2K+ power related submissions in 2021 and over 4.6K+ in 2022 . Still, < 50% of total submissions

Inference 2.1 Power submissions

- **Submitters**: Dell, Fujitsu, H3C, Krai, NVIDIA, Qualcomm, HPE and Lenovo

- **Results**:
  - 2478 submissions on 796 unique systems
  - 10% more power results compared to v2.0

- Good trajectory, but need more adoption



Energy Efficiency (geomean) across submissions has increased by 37-70%* in 1 year

*data center, v1.0 -> v2.0

Power measurement for Distributed Systems

- MLPerf Inference (Network category)
  - Well documented methodology worked out
  - Proof of Concept in HW done with multiple analyzers
  - Targeting v3.1 MLPerf-I submissions

- MLPerf HPC
  - Working through the strawperson with MLPerf HPC-WG
  - Measure power in performance phase as in Green500
  - Support the same levels as in Green500 - but need consistency
  - Target next MLPerf HPC submission

- MLPerf Training
  - Working with MLPerf Training co-chairs on the proposal
  - Feedback from submitters will help guide the methodology
  - Phase the measurements starting from single node/small scale initially

MLPerf Power WG meetings

- Attendees must join “power” alias
- Call for all MLPerf community members to actively participate to expand to different verticals and take part in feature testing/development efforts
- Currently MLPerf Power WG meets weekly
  - Tuesday’s 3PM Pacific Time
    - Moved to accommodate attendees from Asia
- Please reach out if you have any questions
  - sidgunji@mlcommons.org
  - tejus@mlcommons.org

Benchmark Infra

MLPerf Submission Infra

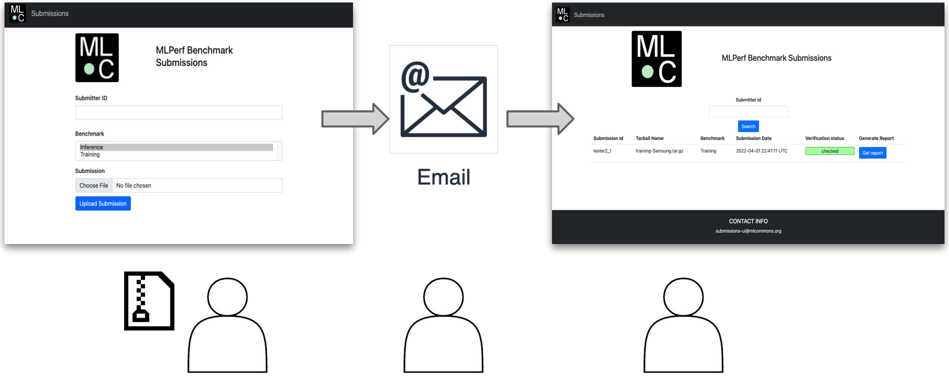
submissions-ui.mlcommons.org

Goals

- Web tool for automating benchmark submission process (Submission UI)
- Automate validations, enforce compliance
- Submitters get instant compliance feedback
- Submissions kept private before deadline

Timeline

- Aug 2022: Inference v2.1
- Sep 2022: HPC v2.0
- Oct 2022: Training v2.1
- Mar 2023: Inference v3.0
- May 2023: Training v3.0
- Oct 2023: HPC v3.0



Infra: Latest Updates

Training v2.1

- Further automation for submission repo.
  - Github actions for checking changes in the results.
  - Generate results table automatically for the review committee.

update_summary.yml

1	division	availability	submitter	system	host_processor_model_name	host_processors_count
2	closed	Available on-premise	ASUSTeK	ESC8000A-E11-8xA100-PCIE-80GB-NVBridge	AMD EPYC 7763	2
3	closed	Available on-premise	ASUSTeK	ESCN4A-E11	AMD EPYC 7773X	1
4	closed	available	Intel-HabanaLabs	HLS-Gaudi2-PT	Intel(R) Xeon(R) Platinum 8380	2
5	closed	available	Intel-HabanaLabs	HLS-Gaudi2-TF	Intel(R) Xeon(R) Platinum 8380	2
6	closed	available	Lenovo	Lenovo ThinkSystem SR670 V2 Server with 4x 40GB SXM4 A100	Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz	2
7	closed	available	Lenovo	Lenovo ThinkSystem SR670 V2 Server with 8x 80GB PCIe A100	Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz	2
8	closed	available	Supermicro	AS-4124GS-TNR	AMD EPYC 7713 64-Core Processor	2
9	closed	available	xFusion	G5500V6x8xA30	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz	2
10	closed	available	xFusion	G5500V6x8xA30	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz	2
11	closed	available	xFusion	G5500V6x8xA30	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz	2
12	closed	available	xFusion	G5500V6x8xA30	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz	2
13	closed	cloud	Azure	NC96ads_A100_v4	AMD EPYC 7V13 64-Core Processor	2
14	closed	cloud	Azure	NC96ads_A100_v4	AMD EPYC 7V13 64-Core Processor	2
15	closed	cloud	Azure	NC96ads_A100_v4	AMD EPYC 7V13 64-Core Processor	2
16	closed	cloud	Azure	ND96amsr_A100_v4_n1	AMD EPYC 7V12 64-Core Processor	2
17	closed	cloud	Azure	ND96amsr_A100_v4_n1	AMD EPYC 7V12 64-Core Processor	2
18	closed	cloud	Azure	ND96amsr_A100_v4_n1	AMD EPYC 7V12 64-Core Processor	2
19	closed	cloud	Azure	ND96amsr_A100_v4_n1	AMD EPYC 7V12 64-Core Processor	2

summary.csv

compare_summary.yml

1	division	availability	submitter	system	host_processor_model_name	host_processors_count	accelerator_model_name	accelerators_count
2	closed	cloud	Azure-HazyResearch	ND96amsr_A100_v4_n16	AMD EPYC 7V12 64-Core Processor	32	NVIDIA A100-SXM4-80GB	128
3	closed	cloud	Azure-HazyResearch	ND96amsr_A100_v4_n8	AMD EPYC 7V12 64-Core Processor	16	NVIDIA A100-SXM4-80GB	64

added_rows.csv

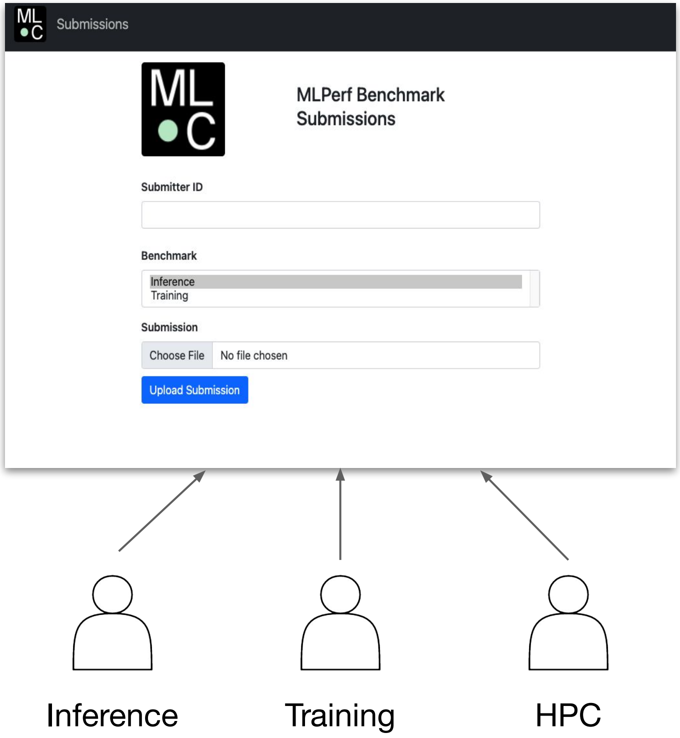
1	division	availability	submitter	system	host_processor_model_name	host_processors_count	accelerator_model_name	accelerators_count
2	closed	cloud	Azure-HazyResearch	ND96amsr_A100_v4_n16	AMD EPYC 7V12 64-Core Processor	32	NVIDIA A100-SXM4-80GB	128
3	closed	cloud	Azure-HazyResearch	ND96amsr_A100_v4_n8	AMD EPYC 7V12 64-Core Processor	16	NVIDIA A100-SXM4-80GB	64

removed_rows.csv

Infra: Latest Updates

Growing Benchmark Support

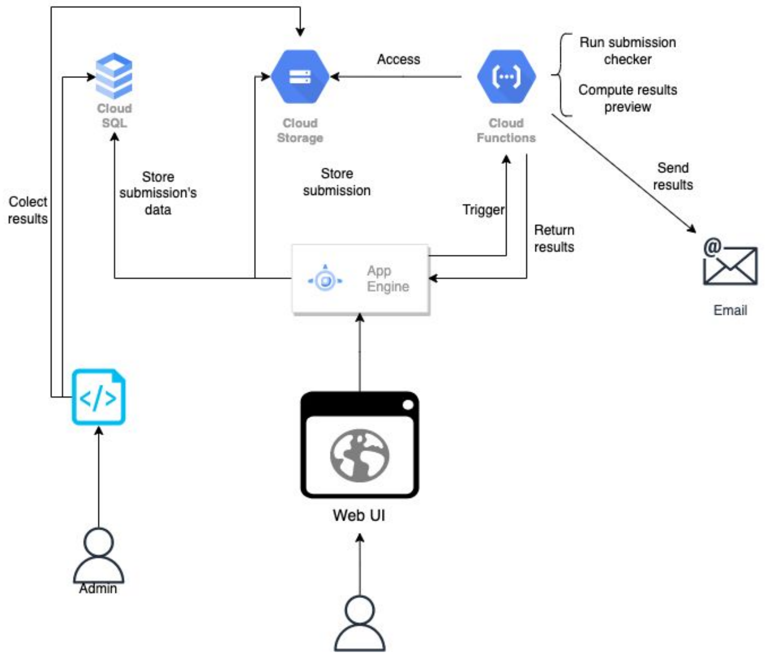
- Submission UI now supports Inference, Training and HPC submissions
- Working in progress with more WGs



Submission UI: currently in progress

- **Goal:** Improve scalability, security and network management.
- **Solution:** Migrate the UI to App Engine (GCP) and adapt the infrastructure around this service.
- App engine manages the application and satisfies the goal.
- Invisible to submitters (end users).

New Infrastructure



Next Steps, Goals and Challenges

Submission Infra

- Potential improvements in auth and verifications (CLA verification, Trademark Agreement verification, EULA verification…).

- Submission UI support for more WG.
    - Standardize requirements

Group Info

Mail list: benchmark-infra@mlcommons.org

Works: https://github.com/mlcommons/logging

Meetings: Tuesdays 10:30-11:00 Pacific Time

Research

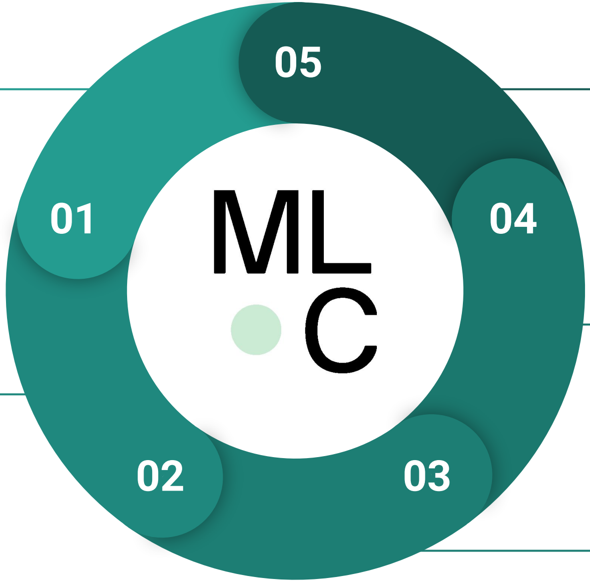
MLCR Activities

Research / Publications

Document the technical challenges that are being addressed and document the solutions we are pioneering for ML systems.

Workshops / Tutorials

Help share the wealth of activities that MLPerf is doing with the academic communities.



Seminar Series

Host invited seminar series that is broadly open to the ML and systems community.

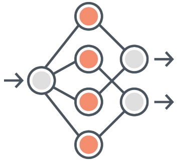
Rising Stars

Identify junior researchers to help promote their visibility within the community.

Prize Awards

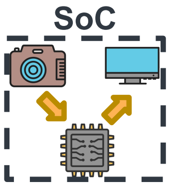
Sponsor research in academic institutions to help build relationships with MLC and look for more forward looking ideas

Real-time and Multi-model ML Inference



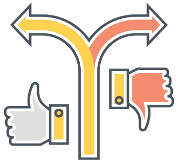
Concurrent and Cascaded Models

Real-time Processing



SoC-level Pipeline

Multi-Modality



User-input-driven

Context-driven

arXiv:2211.08675v1 [cs.LG] 16 Nov 2022

XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse

Hyoukjun Kwon [1] Krishnakumar Nair [1] Jamin Seo [2*] Jason Yik [3*] Debabrata Mohapatra [1] Dongyuan Zhan [1] Jinook Song [1] Peter Capak [1] Peizhao Zhang [1] Peter Vajda [1] Colby Banbury [3] Mark Mazumder [3] Liangzhen Lai [1] Ashish Sirasao [1] Tushar Krishna [2] Harshit Khaitan [1] Vikas Chandra [1] Vijay Janapa Reddi [3]

Abstract

Real-time multi-model multi-task (MMMT) workloads, a new form of deep learning inference workloads, are emerging for applications areas like extended reality (XR) to support metaverse use cases. These workloads combine user interactivity with computationally complex machine learning (ML) activities. Compared to standard ML applications, these ML workloads present unique difficulties and constraints. Real-time MMMT workloads impose heterogeneity and concurrency requirements on future ML systems and devices, necessitating the development of new capabilities. This paper begins with a discussion of the various characteristics of these real-time MMMT ML workloads and presents an ontology for evaluating the performance of future ML hardware for XR systems. Next, we present XRBench, a collection of MMMT ML tasks, models, and usage scenarios that execute these models in three representative ways: cascaded, concurrent, and cascaded-concurrency for XR use cases. Finally, we emphasize the need for new metrics that capture the requirements properly. We hope that our work will stimulate research and lead to the development of a new generation of ML systems for XR use cases.

1 Introduction

Applications based on machine learning (ML) are becoming prevalent. The number of ML models that must be supported on the edge, mobile devices, and data centers is growing. The success of ML across tasks in vision and speech recognition is furthering the development of increasingly sophisticated use cases. For instance, the *metaverse* (Meta, 2022c) combines multiple unit use cases (e.g., image classification, segmentation, speech recognition, etc.) to create more sophisticated use cases (e.g., real-time interactivity via virtual reality). Such sophisticated use cases demand more functionality, for which application engineers are increasingly relying on composability; rather than developing different large models for different use cases, they are mixing multiple smaller, specialized ML models to compose task functionality (Barham et al., 2022).

In this paper, we focus on this new class of ML workloads referred to as multi-model multi-task (MMMT) ML workloads, specifically in the context of extended reality (XR) for metaverse use cases. A real-time MMMT application for extended reality is illustrated by Figure 1. The figure depicts how several MMMT models can be cascaded and operated concurrently, sometimes dynamically subject to certain conditions, to provide complex application-level functionality. The center section of the figure demonstrates that processing throughput requirements can vary depending on the usage scenario. The right side of the figure shows how there can be a variety of interleaved execution patterns for each of the concurrent jobs. MMMT workloads exhibit model heterogeneity, expanded computation scheduling spaces (Kwon et al., 2021), and usage-dependent real-time constraints, which makes them challenging to support compared to today's single-model single-task (SMST) workloads.

We identify three key issues that arise with MMMT workloads that present interesting system level design challenges. The first is *scenario-driven behavior*. All ML pipelines operate at a set frames per second (FPS) processing rate that is determined by a particular use case (e.g., virtual reality gaming, augmented reality social interaction, and outdoor activity recognition). A scenario may sometimes even demand zero FPS (i.e. deactivating a model) for models not required for the scenario. This fluctuating FPS is due to the context-based behavior that drives system resource utilization, which presents a hurdle when designing the underlying DNN accelerator—the heterogeneous workload makes it difficult to employ traditional DNN specialization.

Second, MMMT workloads exhibit *complex dependencies*. XR use cases display substantial data dependency (e.g., eye segmentation to tracking) and control dependency (e.g.,

*Equal contribution [1]Meta, Menlo Park, California, USA [2]ECE, Georgia Institute of Technology, Atlanta, Georgia, USA [3]SEAS, Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Hyoukjun Kwon <hyoukjun.kwon.p@gmail.com>.

https://arxiv.org/pdf/2211.08675.pdf

MLCommons Rising Star Awards Program

- Recognize outstanding MLSys PhD students
- Recognize outstanding junior faculty doing research in MLSys, ML Data…
- Promote activities and alignment with conferences
- Create and nurture long-term relationships
- Requesting support from MLC member companies to invest in your future!
- Please contact vj@eecs.harvard.edu

Closing discussion

ML
C