The Vision Behind MLPerf

A Community-driven ML Benchmark Suite for Software Frameworks and Hardware Accelerators in Cloud and Edge Computing

---

Prof. Vijay Janapa Reddi

MLPerf Inference Chair

ARM Research Summit
September 17th, 2019

HARVARD
School of Engineering and Applied Sciences

MLPerf is the Work of Many

Founding leads: Peter Bailis (Stanford), Greg Diamos (Baidu), Peter Mattson (Google), David Patterson (UC Berkeley / Google), Gu-Yeon Wei (Harvard), Matei Zaharia (Stanford)

Training chairs: Victor Bittorf (Google), Paulius Micikevicius (NVIDIA), Andy Hock (Cerebras)

Inference chairs: Christine Cheng (Intel), David Kanter (RWI), Vijay Janapa Reddi (Harvard), Carole-Jean Wu (Facebook), Guenther Schmuelling (Microsoft), Hanlin Tang (Intel), Bing Yu (MediaTek)

Many others see **mlperf.org/about**

MLPerf

The New York Times

Big Bets on A.I. Open a New Frontier for Chip Start-Ups, Too

ML hardware is projected to be a ~$60B industry in 2025.

(Tractica.com $66.3B, Marketsandmarkets.com: $59.2B)

than $100 million from investors. Venture capitalists invested more than $1.5 billion in chip start-ups last year, nearly doubling the investments made two years ago, according to the research firm CB Insights.

MLPerf

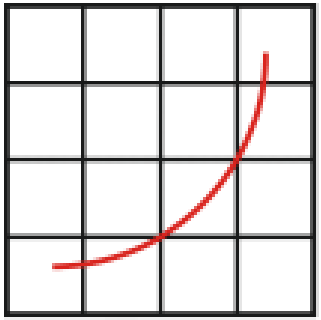
What task?
What model?
What dataset?
What batch size?
What quantization?
What software libraries?
…

MLPerf

Agenda

- ✅ *Why ML needs a benchmark suite?*
- **Are there lessons we can borrow?**
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?

Yes! Look to successful history in benchmarks.



spec®

TPC™

MLPerf

SPEC Impact

- Settled **arguments in the marketplace** (grow the pie)
- Resolved internal **engineering debates** (better investments)
- Became a **standard** from research to practice in the industry
- Needed to **revise regularly** to maintain usefulness:

  SPEC89, SPEC92, SPEC95, SPEC2000, SPEC2006, SPEC2017

SPEC fueled the Golden Age of microprocessor design.

MLPerf

Can we start a new Golden Age for ML Systems?

MLPerf

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- **What is MLPerf?**
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?

What is MLPerf?

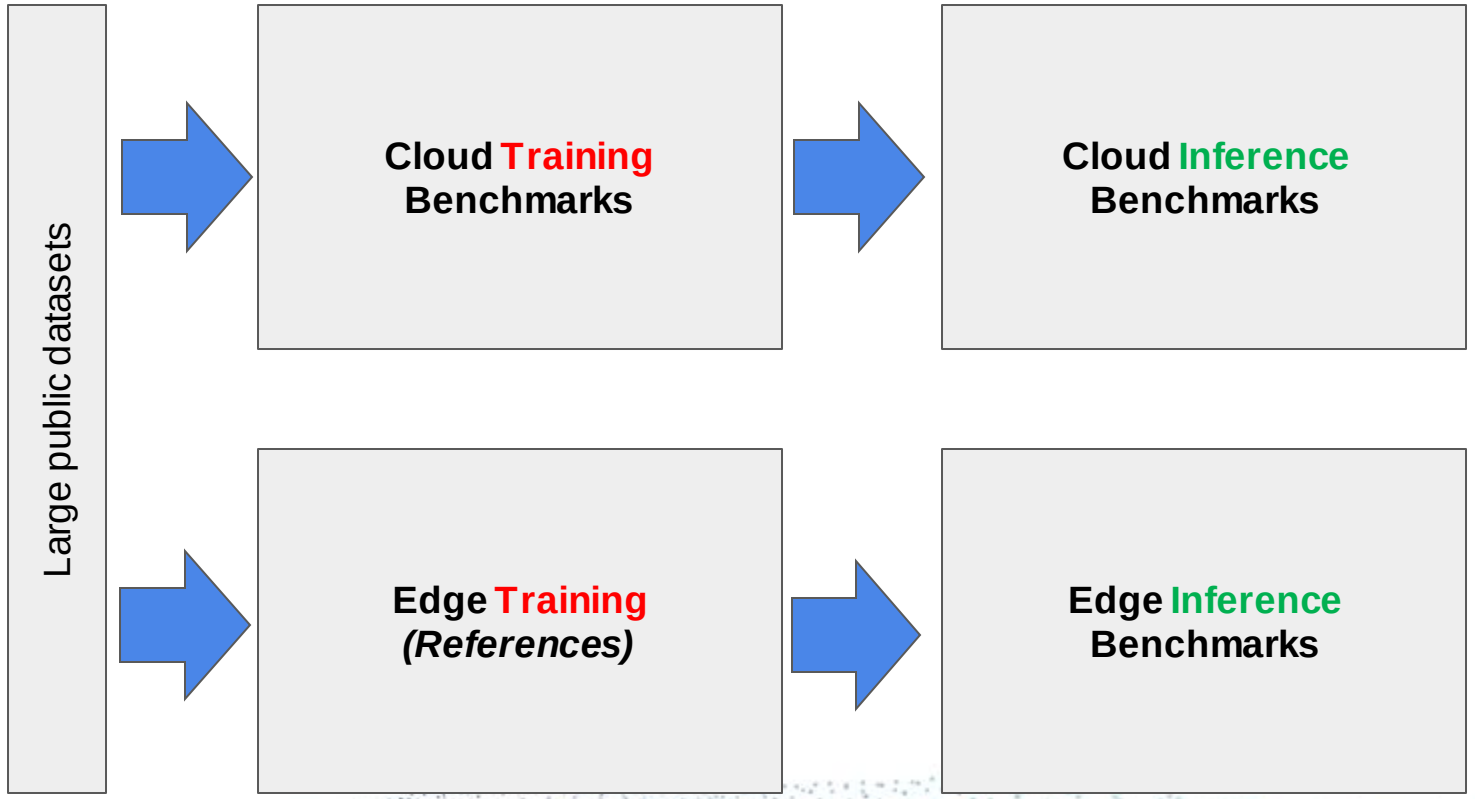
A machine learning performance benchmark suite with broad industry and academic support.

Contributions by researchers from

Harvard University, Stanford University, University of Arkansas, Littlerock, University of California, Berkeley, University of Illinois, Urbana Champaign, University of Minnesota, University of Texas, Austin, University of Toronto

Alibaba, AMD, Arm, Baidu, Cadence, Cerebras, Cisco, Cray, Dividiti, DDN Storage, Enflame Tech, Esperanto, Facebook, Google, Groq, Habana, Iluvatar, Inspur, Intel, In-Q-Tel, MediaTek, Mentor Graphics, Microsoft, Myrtle, Mythic, NetApp, NVIDIA, One Convergence, Qualcomm, Rpa2ai, Sambanova, Samsung S.LSI, Sigopt, Synopsys, Tencent, Tensyr, Teradyne, VMware, Wave Computing, WekaIO, Xilinx

MLPerf
(Sept. 2019)



MLPerf

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - **How does MLPerf curate a benchmark?**
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?

The ML **Landscape is Large**

Area	Vision	Language	Audio	Commerce	Action / RL	Other
Problem	Image Classification Object Detection / Segmentation Face ID HealthCare (Radiology) Video Detection Self-Driving	Translation Language Model Word Embedding	Speech Recognition Text-to-Speech Question Answering Keyword Spotting Language Modeling Chatbots Speaker ID Graph embeddings Content ID	Rating Recommendations Sentiment Analysis Next-action Healthcare (EHR) Fraud detection Anomaly detection Time series prediction Large scale regression	Games Go Robotics Health Care Bioinformatics	GANs 3D point clouds Word embeddings
Models	ResNet-50 TF Object Detection Detectron	BERT Transformer OpenNMT	Deep Speech 2 SQuAD Explorer	Neural Collaborative Filtering CNNs	DQN PPO A2C	
Datasets	ImageNet COCO	WMT English-German	LibriSpeech SQuAD LM-Benchmark	MovieLens-20M Amazon IMDB	Atari Go Chess Grasping	
Metrics	COCO mAP Prediction accuracy	BLEU	WER Perplexity	Prediction accuracy	Prediction accuracy Win/Loss	

MLPerf

ML Benchmark Design Choices

Big Questions	Training	Inference
1. Benchmark **definition**	What is the definition of a benchmark task?	
2. Benchmark **selection**	Which benchmark task to select?	
3. **Metric** definition	What is the measure of performance in ML systems?	
4. Implementation **equivalence**	How do submitters run on different hardware/software systems?	
5. **Issues** specific to training or inference	Which hyperparameters can submitters tune?	Quantization, calibration, and/or retraining?
	Reduce result variance?	
6. **Results**	Do we normalize and/or summarize results?	

Decision Making Guidelines

Model Range	Example	Principle
Maturity: Lowest common denominator, most widely used, or most advanced?	**Image recognition**: AlexNet, ResNet, or EfficientNet?	**Cutting edge**, not bleeding edge
Variety: What broad kind of deep neural network to choose?	**Translation**: GNMT with RNN vs. Transformer with Attention	Try and **ensure coverage** at a whole suite level
Complexity: Less or more weights?	**Object detection**: SSD vs. Mask R-CNN? Resolution?	**Survey end-users** and anticipate market demand
Practicality	**Feasibility**: Is there a public dataset?	**Good for now** > perfect.

MLPerf **Training** Benchmarks 0.5v, 0.6v

Area	Benchmark	Model	Dataset
Vision	Image classification	ResNet-50 v1.5	ImageNet
	Object detection	Mask R-CNN	COCO
		Mask-R-CNN	COCO
Language/Audio	Translation	Transformer	WMT Eng-Germ
		NMT	WMT Eng-Germ
Commerce	Recommendation	NCF	MovieLens-20M
Action	Reinforcement Learning	Mini-go	Go

MLPerf **Inference** Benchmarks 0.5v

Area	Benchmark	Model	Dataset
Vision	Image Classification	MobileNet v1 ResNet50	ImageNet (224x224) ImageNet (224x224)
	Object Detection	SSD-MobileNet v1 SSD-ResNet34	MS-COCO (300x300) MS-COCO (1200x1200)
Language	Translation	Google NMT	WMT Eng-Germ

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
    - **What is the "science" behind the curation?**
    - Where are we heading now?
- What comes next for MLPerf?

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
    - **What is the "science" behind the curation?**
      - **Training**
      - Inference
      - Results
    - Where are we heading now?
- What comes next for MLPerf?

Challenges in Benchmarking ML Systems

Training

- **Hyperparameters**
- **Metric**
- Scale
- Power
- Cost
- Variance
- On-premise vs. cloud
- ...

Inference

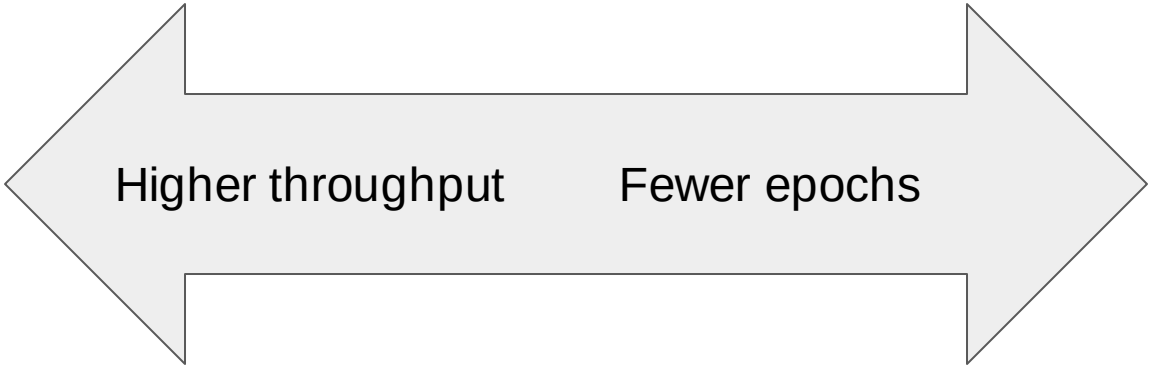
- **Scenarios**
- **Quantizations**
- Pruning
- Scores
- **Power**
- ...

Training **Metric**: Throughput vs. Time-to-Train

Throughput (samples / sec)
Easy / cheap to measure

Can increase throughput at cost of total time to train!

Time-to-train (end-to-end)
Time to solution!



MLPerf

Training **Hyperparameter Tuning**

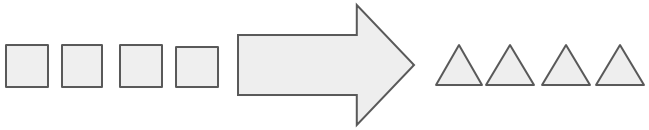
- Different **system sizes** ⇒ different **batch sizes** ⇒ different **hyperparameters**
- But, some working hyperparameters are better than others
- Finding good hyperparameters is expensive and not the point of the benchmark, **we do not want a hyperparameter tuning competition**!

- Solution v0.5, v0.6:
    - Hyperparameter "borrowing" during review process

Agenda

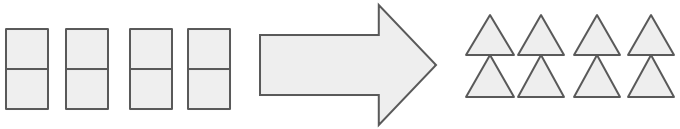
- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
    - **What is the "science" behind the curation?**
      - Training
      - **Inference**
      - Results
    - Where are we heading now?
- What comes next for MLPerf?

Inference **Scenarios & Metrics** to Measure



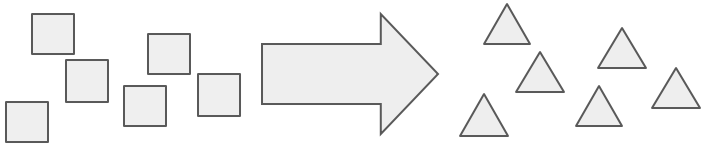
Single stream
(e.g. cell phone augmented vision)

Latency



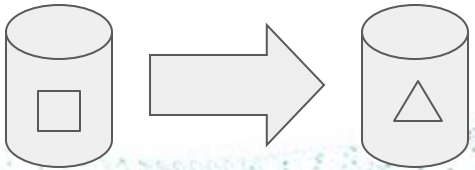
Multiple stream
(e.g. multiple camera driving assistance)

Number streams
subject to latency bound



Server
(e.g. translation app)

QPS
subject to latency bound



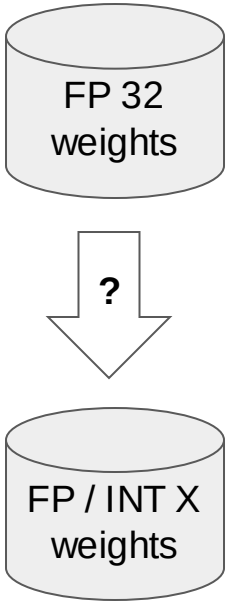
Offline
(e.g. photo sorting app)

Throughput

MLPerf

Inference **Quantization** and **Retraining**

- Quantization is key to efficient inference, but do not want a quantization contest – some can do it better than others
- Can the submitters quantize?
  - **Yes**, but must be principled: describe reproducible method
- Can the submitters calibrate?
  - **Yes**, but must use a fixed set of calibration data
- Can the submitters retrain?
  - **No**, not a retraining contest. But, provide retrained 8 bit.



MLPerf

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
    - **What is the "science" behind the curation?**
      - Training
      - Inference
      - **Results**
    - Where are we heading now?
- What comes next for MLPerf?

Results **Normalization** and/or Scale

- Do you present only the results?

System
Foo
Bar

Results lack scale information.

If so, an inefficient larger system can look better than an efficient smaller system.

Need supplemental normalization and/or scaling information

MLPerf provides some scale information

Current: Number of chips
Planned: Power

Results **Summarization** (or Not)?

- Should we have a single MLPerf score that summarizes all results?

System	ResNet	GNMT
Foo	3m	4m
Bar	1m	6m

- Summarized results **Pros**
  - Easy to communicate
    Do it consistently
- Summarized results **Cons**
  - Oversimplifies
  - Some vendors only submit subsets
  - Users care about different subsets

MLPerf doesn't summarize.
We recommend weighted geometric mean.

Agenda

- ✅ *Why ML needs a benchmark suite?*
- ✅ *Are there lessons we can borrow?*
- ✅ *What is MLPerf?*
  - ✅ *How does MLPerf curate a benchmark?*
  - ✅ *What is the "science" behind the curation?*
  - ○ **Where are we heading now?**
- What comes next for MLPerf?

NVIDIA

USA - United States

PLATFORMS · DEVELOPERS · COMMUNITY · SHOP DRIVERS · SUPPORT ABOUT NVIDIA ·

HOME DEEP LEARNING DRIVING GAMING PRO GRAPHICS AUTONOMOUS MACHINES HEALTHCARE AI PODCAST

NVIDIA Sets Six Records in AI Performance

Tensor Core GPUs achieve best performance across every MLPerf benchmark submitted;

Google Cloud

Blog Latest Stories Product News Topics

AI & MACHINE LEARNING

MLPerf benchmark establishes that Google Cloud offers the most accessible

intel AI

Technology Industries Library Research Careers Programs Search

MLPERF RESULTS VALIDATE CPUS FOR DEEP LEARNING TRAINING

Inference Submissions Due Soon (Oct 11th)

Closed division submissions

- Requires using the specified model
- Enables apples-to-apples comparison
- Simplifies work for HW groups

Open division submissions

- Open division allows using any model
- Encourages innovation
- Ensures Closed division does not stagnate

MLPerf

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
  - *What is the "science" behind the curation?*
  - *Where are we heading now?*
- **What comes next for MLPerf?**

ML Commons

We are creating a non-profit called MLCommons to "**accelerate ML innovation and increase its positive impact on society.**"

Datasets x Benchmarks x Best Practices

The People's Speech

100k hours of human speech boosting the tail of the 2030 earth population

Strawman: Brutally honest feedback is welcome

(seeking authors, add yourself here) Greg Diamos

Objective. MLCommons aims to boost speech recognition accuracy for the entire earth's population by creating a new speech dataset with 100,000 hours of transcribed speech focusing on improving word error rates for multiple languages, accents, and variation. Smart speakers and personal assistants are expected make advanced speech recognition technology available to the entire earth's population by the mid 2020s. Existing systems are tuned for the most common speakers of the most common languages such as English and Mandarin. However, the real distribution of human speech is very diverse. There are approximately one thousand spoken languages with more than one million speakers each, many regional accents, and significant variation among speakers and environments. We will focus on boosting accuracy for the tail of this distribution by sampling more speech from it. We will also include speech drawn from emerging applications of speech recognition in healthcare and hard speech technology problems including far field and diarized speech. The dataset will be licensed to permit unrestricted academic and commercial reuse.

Historical Inspiration. We are inspired in part by the success of the DARPA grand challenge model which has led to rapid improvement in speech recognition accuracy by focusing the community on clear metrics. We plan to develop new test sets that allow us to measure word error rates of speech recognition systems from more languages, accents, new applications, and with more variation. We are inspired by efforts to create community speech recognition resources such as the Linguistic Data Consortium, Mozilla Common Voice, LibriSpeech, DIHARD, and DeepSpeech. We plan to leverage internet platforms, crowdsourcing, and efficient labeling technology to collect and label a very large collection of speech. Emerging generative speech and language modeling technology such as WaveNet, TacoTron, BERT, GPT, and DeepVoice have enabled algorithms to generate realistic speech and language. We will use these technologies to augment our dataset. We will align our work with the MLPerf speech benchmark to enable the creation of high performance computing systems that can train such a speech recognition system efficiently.

Agile Dataset Development. A small core team will rapidly iterate on the creation of test sets drawn from real speech used by industry and academic partners as well as new training data sources. Accuracy of our best models trained on new data sources will be evaluated on these test sets. We will quickly release an initial low quality dataset with 100,000 hours to enable building efficient training systems. Subsequent releases will be staged over a two year duration as new data achieves significant improvements in test set accuracy.

Data Platforms. We will partner with internet platforms to reach the earth's population. The rise of personal computing and the internet has enabled the emergence of content platforms that mediate access to content created by diverse communities of users. Platforms such as YouTube, Vimeo, StoryCorps, and many others include diverse speech data resources, some of which are released with permissive licenses like CC-BY. We will also leverage speech synthesis, crowdsourcing, and assisted labeling technology to assemble our dataset.

Speech Recognition System. We will evaluate the impact of new data on a state of the art deep learning based speech recognition system starting from a recurrent neural network transducer architecture that uses convolutional and recurrent neural networks to process mel spectrogram frames that is trained using a transducer loss function and evaluated using a separately trained language model with a beam search decoder. During this project we will develop more accurate models ourselves and in cooperation with competitions such as the MLPerf Open Division, DIHARD, and Kaggle.

High Performance Training System. Training a speech recognition system on 100,000 hours of speech will create a very large computational problem. We anticipate that the dataset will require 10TB of storage and speech recognition models with billions of parameters. This is a very large computational problem, estimated to take at least ten years to train with an unoptimized V100 GPU implementation. We will develop a highly optimized training system capable of training a very accurate speech recognition system in less than one week.

Call For Participation. We are seeking contributors to this project who can donate near full time. See this document for information about how to apply.

MLPerf

More at **MLPerf.org,** or contact **info@mlperf.org**

MLPerf