MLPerf

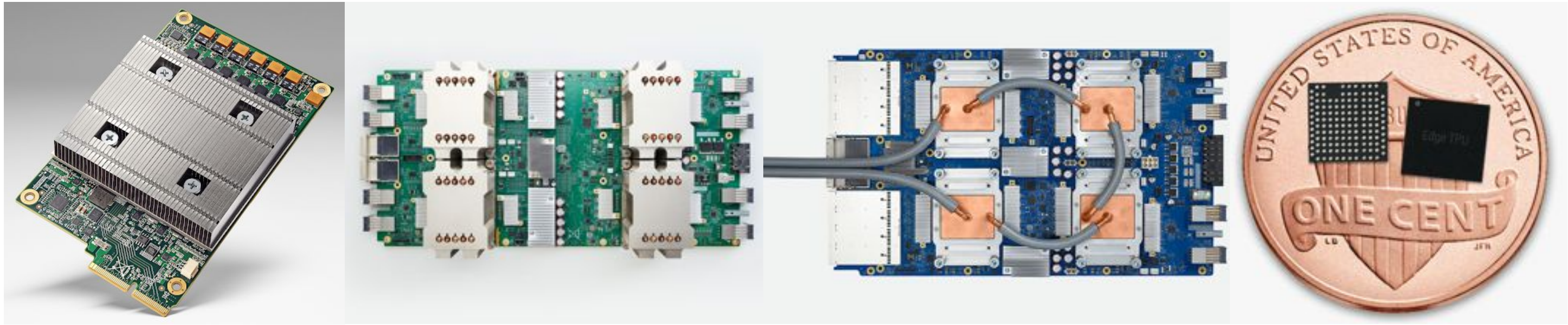
Fair and Useful Benchmarks for Machine Learning

Cliff Young, Google AI (but representing the work of many)
Stanford CS217
November 13, 2018

Deep Learning has Reinvigorated Hardware

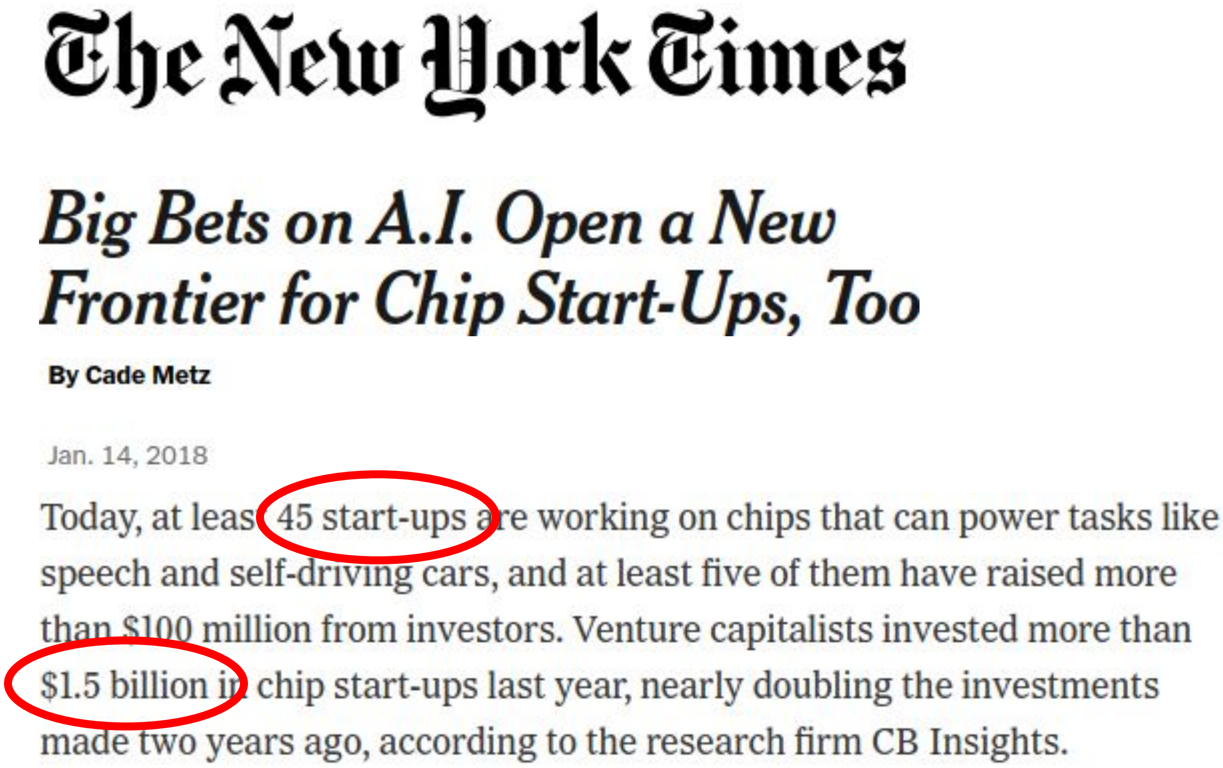
GPUs ⇒ AlexNet, Speech.

TPUs ⇒ Many Google applications: AlphaGo and Translate, WaveNet speech.



Startups ⇒ both training and inference, many different approaches.

I’m looking forward to test-driving new systems.



The New York Times

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

By Cade Metz

Jan. 14, 2018

Today, at least 45 start-ups are working on chips that can power tasks like speech and self-driving cars, and at least five of them have raised more 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.

Q: How do We Measure ML Performance Today?

A: Surprisingly badly.

- Example: single-benchmark measurement of throughput, with synthetic training data, ignoring accuracy.
- Poor reproducibility, comparability.
- “ResNet-50” is not a precise specification, but it’s what everyone reports.

Q: How could we do better?

A1: Look to successful history in benchmark suites: SPEC and TPC.

A2: Draw on experiences of those who have done ML benchmarking.

Idea: Could we build a “SPEC for Machine Learning”?

SPEC impact

- Settled arguments in marketplace (grow the pie).
- Resolved internal engineering debates (better investments).
- Cooperative ⇒ nonprofit Corporation with 22 members.
- Universities join at modest cost.
- Became standard in marketplace, papers, and textbooks.
- Needed to revise suite regularly to maintain usefulness: SPEC89, SPEC92, SPEC95, SPEC2000, SPEC2006, SPEC2017.

Coincides with (caused?) the Golden Age of microprocessors.

Can we start a new Golden Age for ML Systems?

Agenda

Why Machine Learning Needs Benchmarks

MLPerf Principles (stealth-mode phase)

Launching MLPerf Training (exponential growth phase)

Work-in-Progress: MLPerf Inference (new product line)

Agenda

~~Why Machine Learning Needs Benchmarks~~

MLPerf Principles (stealth-mode phase)

Launching MLPerf Training (exponential growth phase)

Work-in-Progress: MLPerf Inference (new product line phase)

First Half of 2018: MLPerf Research Phase

Gathered researchers from Baidu (DeepBench), Google (TF benchmarks), Harvard (Fathom), and Stanford (DAWNBench).

Combined the best parts from all of our experiences.

Planned to cover both training and inference; initial focus on **training**.

MLPerf Goals

- Accelerate progress in ML via fair and useful measurement.
    - Moving target ⇒ launch and iterate rapidly
- Serve both the commercial and research communities.
- Enable fair comparison of competing systems yet encourage innovation to improve the state-of-the-art of ML.
- Enforce replicability to ensure reliable results.
- Use representative workloads, reflecting production use-cases.
- Keep benchmarking effort affordable (so all can play).

Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Correct inevitable mistakes in the formulation
  - Scale problems to match faster hardware
- At least initially, revise annually? MLPerf18, MLPerf19, …
- Like SPEC, have quarterly deadlines and then publish results for that quarter via searchable database

Open/Closed Divisions

- **Closed** division requires using the specified model
  - Limits overfitting
  - Enables apples-to-apples comparison
  - Simplifies work for HW groups

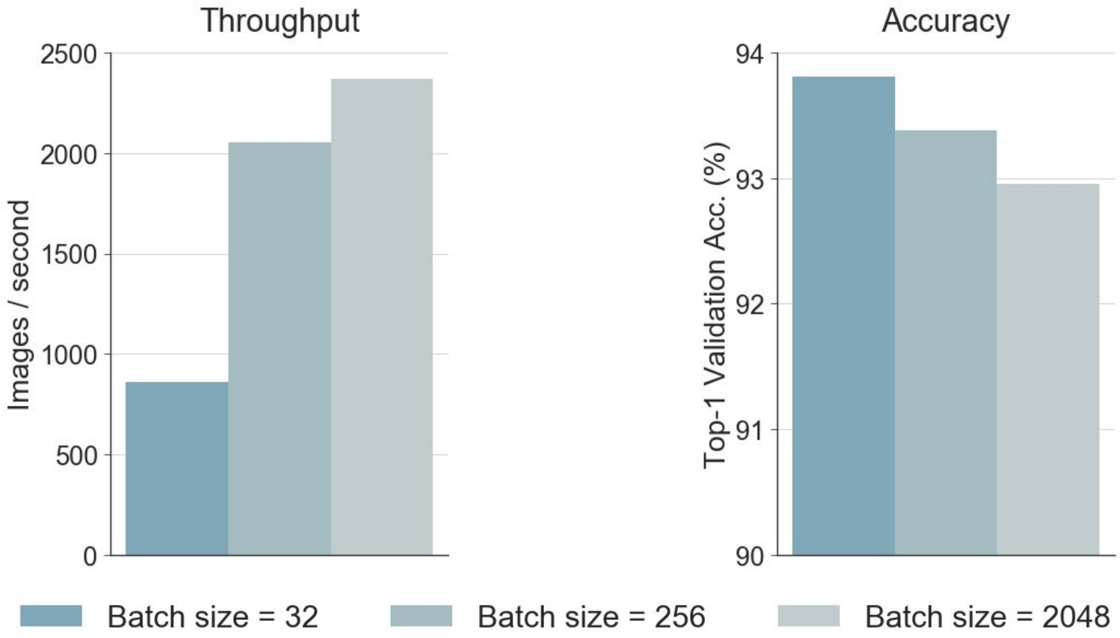
- **Open** division allows using any model
  - Encourages innovation
  - Ensures Closed division does not stagnate

Metrics Should Capture Performance and Quality

- ***Performance***: how fast is a model for training, inference?
- ***Quality***: how good are a model's predictions?

Important for benchmark to capture **both** performance and quality

Performance and Quality **aren’t always correlated**



End-to-end training of a ResNet56 CIFAR10 model on a Nvidia P100 machine with 512 GB of memory and 28 CPU cores, using TensorFlow 1.2 compiled from source with CUDA 8.0 and CuDNN 5.1.

MLPerf metric: **Training time** to reach quality target

- Quality target is *specific for each benchmark* and *close to state-of-the-art*
  - Updated w/ each release to keep up with the state-of-the-art

- Time includes preprocessing, validation over median of 5 runs

- Available: reference implementations that achieve quality target

MLPerf

A Benchmark for Machine Learning from an Academic/Industry Cooperative

Researchers from:
Baidu, Google, Harvard, Stanford, and UC Berkeley

Agenda

~~Why Machine Learning Needs Benchmarks~~

~~MLPerf Principles (stealth-mode phase)~~

Launching MLPerf Training (exponential growth phase)

Work-in-Progress: MLPerf Inference (new product line)

2H2018: MLPerf Community Phase

May	First general meeting.
June	Added benchmarks (volunteers!).
July	Chartered working groups: on-prem, Cloud, submitters, special topics
August	WGs report solid progress; inference WG chartered.
September	More WG progress.
October	First v0.5 submissions, with review period.
November 9	First submissions!

Practical Compromises for Useful Comparisons

- Restricted hyperparameter tuning in Closed: avoid resource competition.
- Scale metric for Cloud: find the 2x wins, avoid price micro-optimization.
- Scale for on-premises = power; methodology under way.
- Variance issues in measuring the Time-to-accuracy metric.
- Benchmark choices adapted based on community feedback
  +Mask R-CNN, +Transformer, -Sentiment.

Design Debate: Summary Scores?

Pro: If you don’t, people will anyway (and badly)

We know how to do the statistics correctly.

Con: Not every user or supplier cares about all benchmarks

This is an era of specialization.

Avoid over-optimization/focus on particular benchmarks.

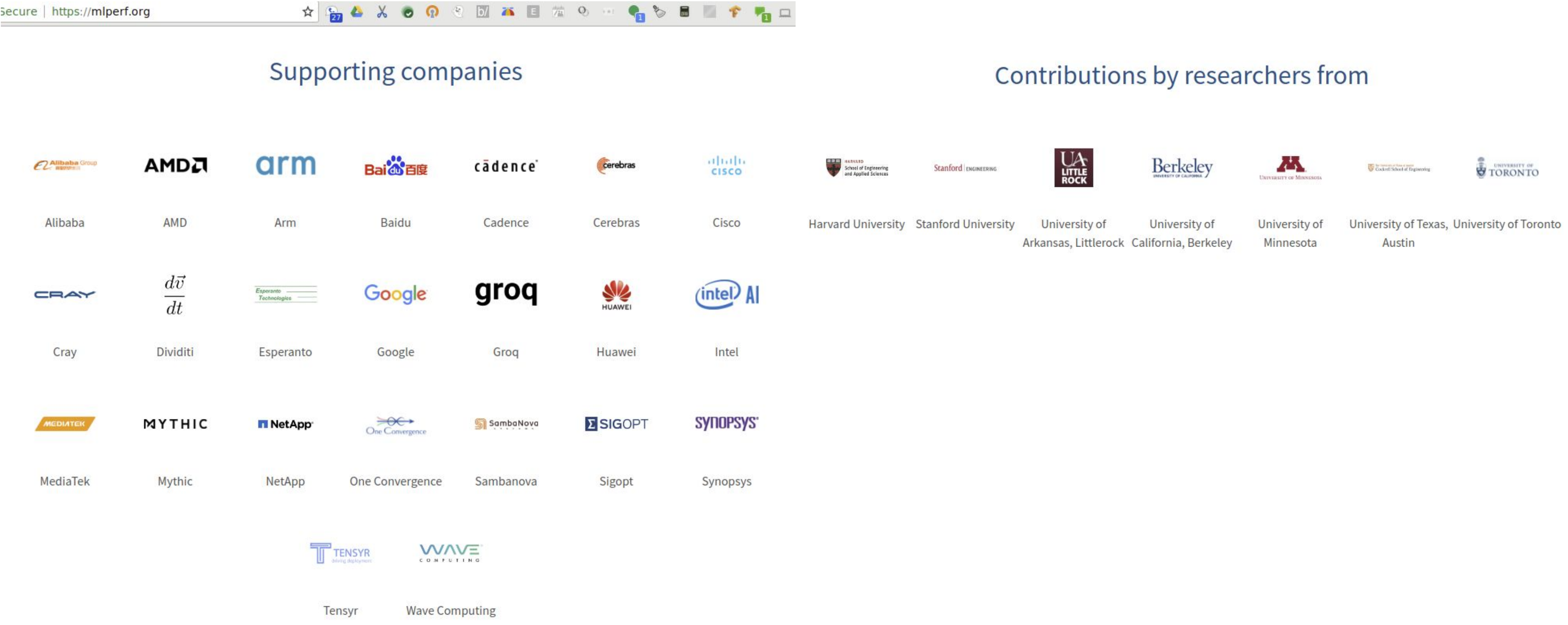
Working compromise: no summary scores,

but a required methodology if you want to combine scores.

MLPerf Training Benchmarks

Task	Model	Dataset
Image Classification	ResNet-50	ImageNet
Object Detection	Mask-RCNN SSD	MS-COCO 2017
Translation	Google NMT Transformer	WMT16 WMT17
Recommendation	Neural Collaborative Filtering	MovieLens ml-20m
Reinforcement Learning	Minigo	NA
Speech Recognition	DeepSpeech2*	Librispeech

Supporting Organizations / Researchers / Folks



Secure | https://mlperf.org

Supporting companies

Alibaba

AMD

Arm

Baidu

Cadence

Cerebras

Cisco

Cray

Dividiti

Esperanto

Google

Groq

Huawei

Intel

MediaTek

Mythic

NetApp

One Convergence

Sambanova

Sigopt

Synopsys

Tensyr

Wave Computing

Contributions by researchers from

Harvard University

Stanford University

University of Arkansas, Littlerock

University of California, Berkeley

University of Minnesota

University of Texas, Austin

University of Toronto

450+ member discussion group

What are we working on?

- First version: **reference** code, in two frameworks, of each benchmark.
- Resolving or controlling the **variance** issues.
- Launching an **inference** suite (deferred from first release).
- Getting to **governance**, and an umbrella organization.

Agenda

~~Why Machine Learning Needs Benchmarks~~

~~MLPerf Principles (stealth-mode phase)~~

~~Launching MLPerf Training (exponential growth phase)~~

Work-in-Progress: MLPerf Inference (new product line)

Inference is Under Construction

WG launched just 10 weeks ago.

Open issues:

- How many divisions?
- Where do you draw the system/measurement boundary?
- What metric(s) to measure?
- How do we report results out?

Inference: What's the Right Metric?

Accuracy.

This seems required, else the results are irrelevant.

Like the training division, relax slightly from SOTA research accuracy.

Latency (TPU paper highlights importance for web serving).

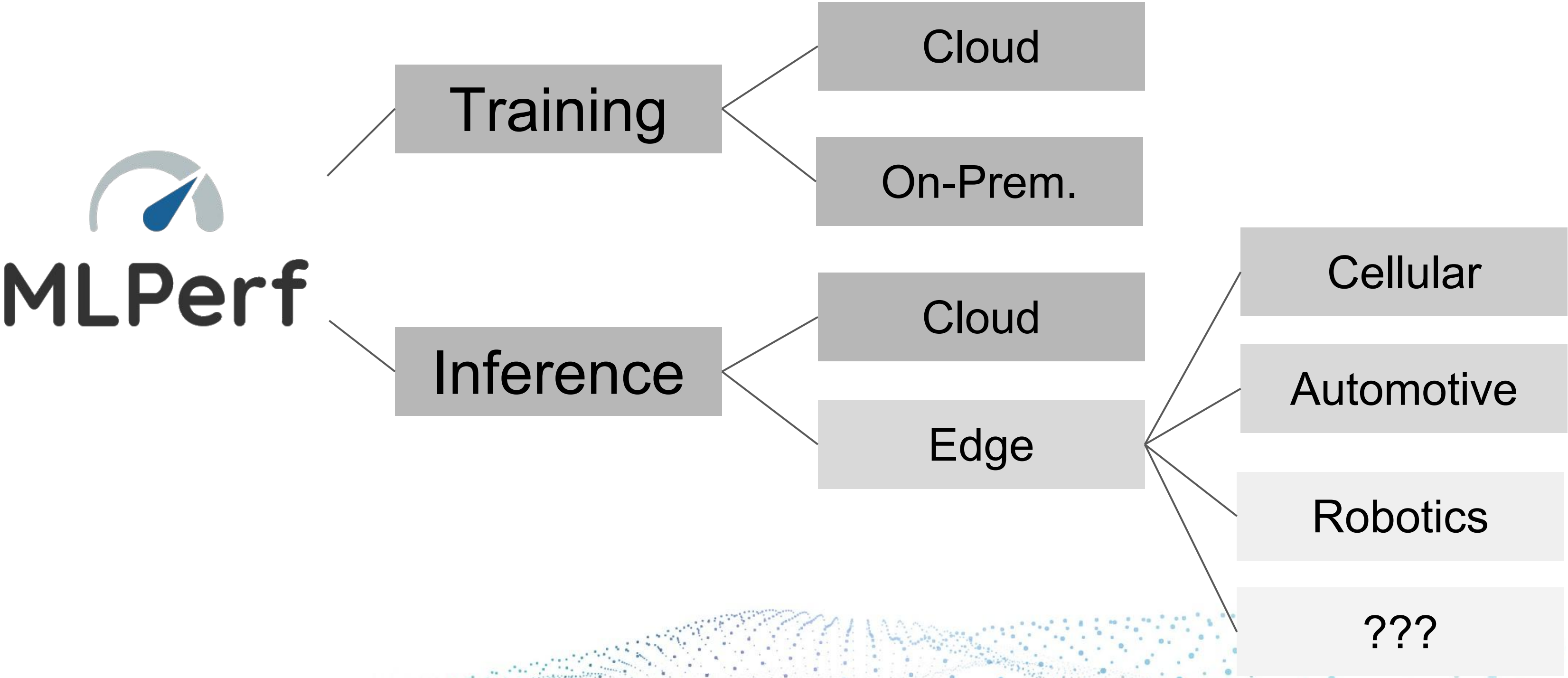
Batch/offline processing might not care about latency.

Do we need many latency buckets?

Throughput.

But with an accuracy and a latency constraint.

How many MLPerf Divisions?



MLPerf

Training

Cloud

On-Prem.

Inference

Cloud

Edge

Cellular

Automotive

Robotics

???

Recap

Machine Learning Needs Benchmarks!

Goals: agility, both research and development, replicability, affordability.

MLPerf Training: v0.5 deadline is **October 31**.

MLPerf Inference is **under construction**.

For rapid iteration to work, we need good input!

MLPerf needs your help!

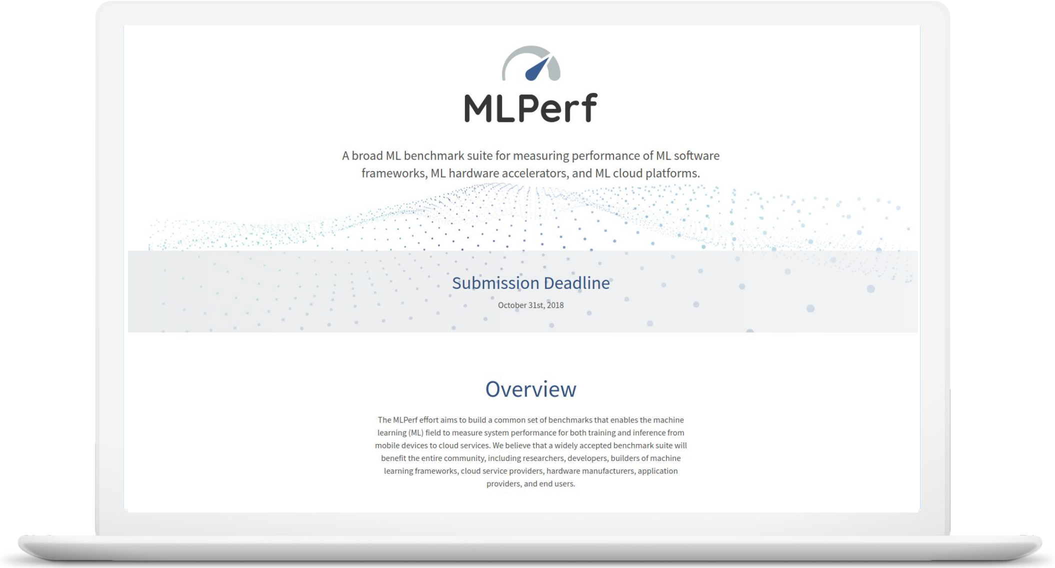
- Join the discussion community at MLPerf.org.

- Help us by joining a working group:
  Cloud scale, on-premises scale, submitters, special topics, inference.
  Help us design submission criteria, to include the data you want!

- Propose new benchmarks and data sets.

- Submit your benchmark results!

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



MLPerf

A broad ML benchmark suite for measuring performance of ML software frameworks, ML hardware accelerators, and ML cloud platforms.

Submission Deadline

October 31st, 2018

Overview

The MLPerf effort aims to build a common set of benchmarks that enables the machine learning (ML) field to measure system performance for both training and inference from mobile devices to cloud services. We believe that a widely accepted benchmark suite will benefit the entire community, including researchers, developers, builders of machine learning frameworks, cloud service providers, hardware manufacturers, application providers, and end users.

v0.5 Submission Deadline: October 31!