The Vision Behind MLPerf (mlperf.org):

An ML Benchmark Suite for ML Software Frameworks and
ML Hardware Accelerators in ML Cloud and Edge Computing Platforms

2018 International Symposium on Benchmarking, Measuring and Optimizing

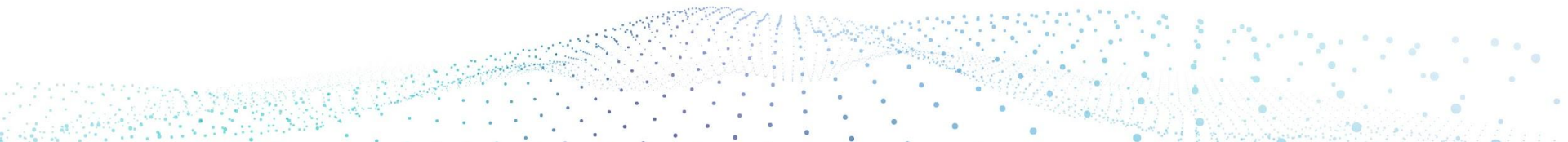December 11th, 2018

Prof. Vijay Janapa Reddi

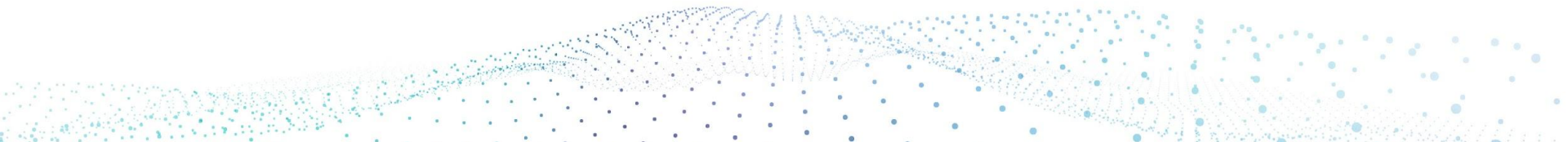HARVARD
UNIVERSITY

THE UNIVERSITY OF
TEXAS
AT AUSTIN

Google

**(representing the viewpoints of many, many, people in MLPerf)**

"[A New Golden Age for Computer Architecture: **Domain-Specific Hardware/Software Co-Design**, Enhanced Security, Open Instruction Sets, and Agile Chip Development](#)"

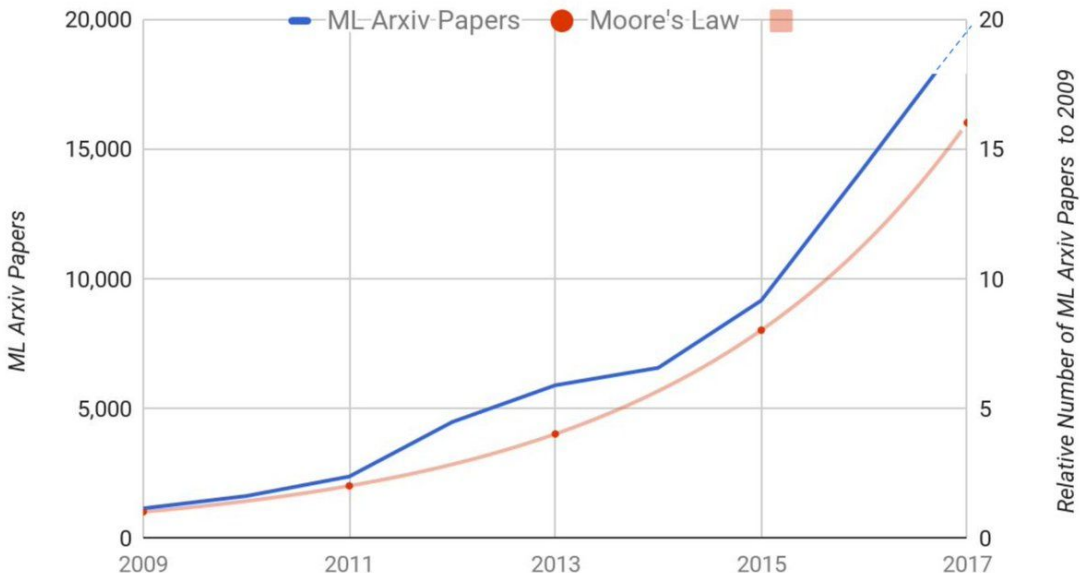John Hennessy and David Patterson

"[A New Golden Age in Computer Architecture: Empowering the **Machine-Learning Revolution**](#)"

Jeff Dean, David Patterson, Cliff Young
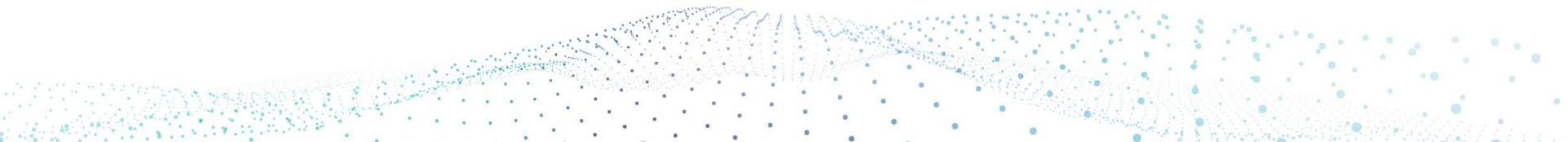
The (Rapid) Rise of ML

- The number of ML papers published Arxiv each year is growing exponentially

- The pace of growth is on par and if not exceeding the rate of Moore's Law scaling

Source: https://blog.openai.com/ai-and-compute/

AI to Compute: 300,000x Increase in Compute

"… **since 2012 the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time (by comparison, Moore's Law had an 18-month doubling period).** Since 2012, this metric has grown by more than 300,000x (an 18-month doubling period would yield only a 12x increase). Improvements in compute have been a key component of AI progress, so as long as this trend continues, it's worth preparing for the implications of systems far outside today's capabilities."

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute
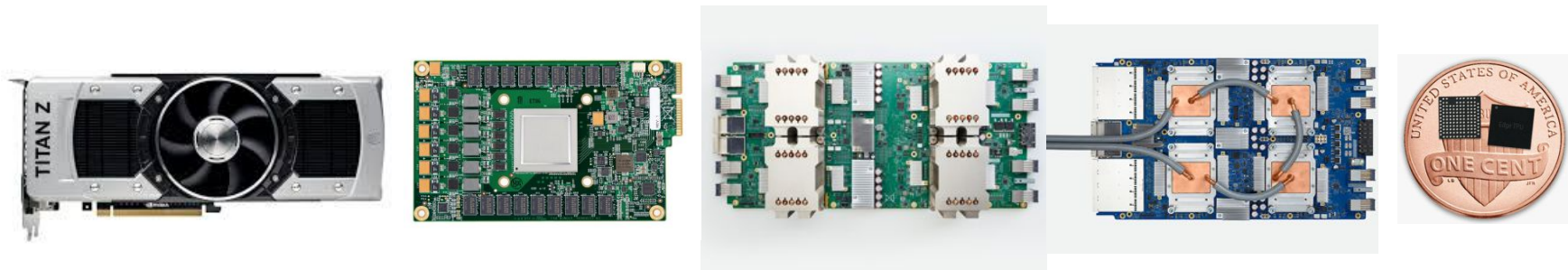

Source: https://blog.openai.com/ai-and-compute/

MLPerf
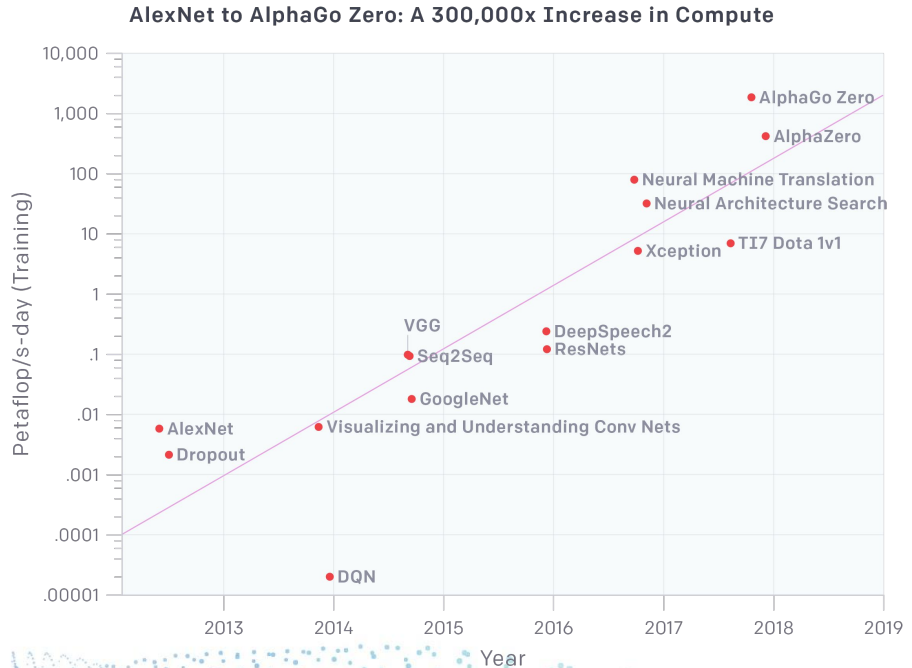
Deep Learning has Reinvigorated Hardware

GPUs ⇒ AlexNet, Speech.

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

→ Rapidly fueling the renaissance of the hardware industry, including startups

MLPerf

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.

8

MLPerf

How do we **compare** the hardware?

MLPerf

MLPerf

How do we compare the hardware, today?

Answer is "**surprisingly badly.**"

- Example: single-benchmark measurement of throughput
  - Synthetic training data
  - Measure performance, **ignoring accuracy**

- Poor reproducibility
  - No means to effectively reproduce the same results
  - Hard to compare numbers across different models, inputs and datasets

- "ResNet-50" is not a precise specification, but it's what everyone reports.

MLPerf

How do we **design** better hardware?

MLPerf

How do we design better hardware? More MACS?!

Latency (ms) vs # MACs (Million)
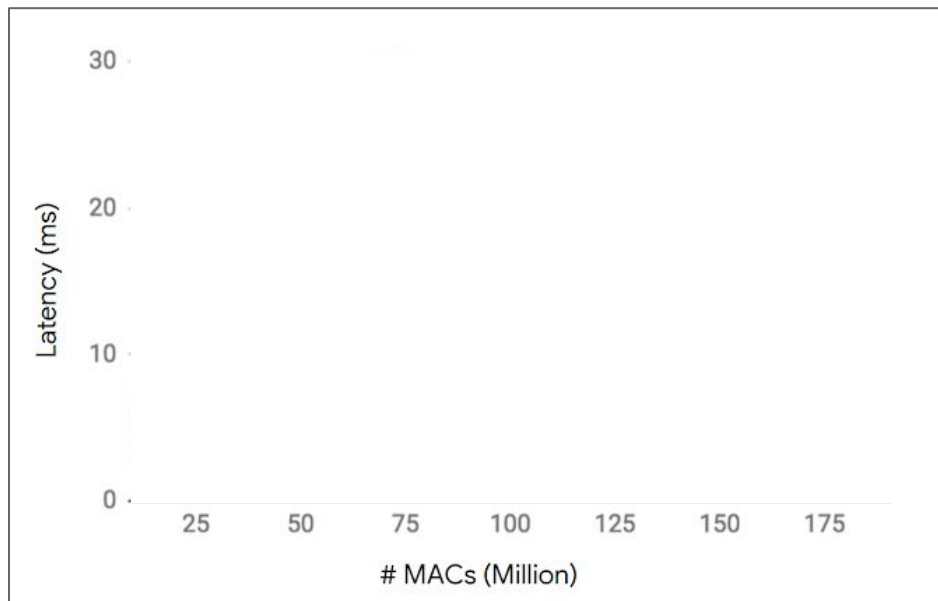
- Model performance cannot be evaluated using raw hardware performance (MACs)

- Model latency varies across different levels of MAC capability

- Latency ultimately impacts or dictates the experience

MLPerf

The **Three Cornerstones** for ML Performance
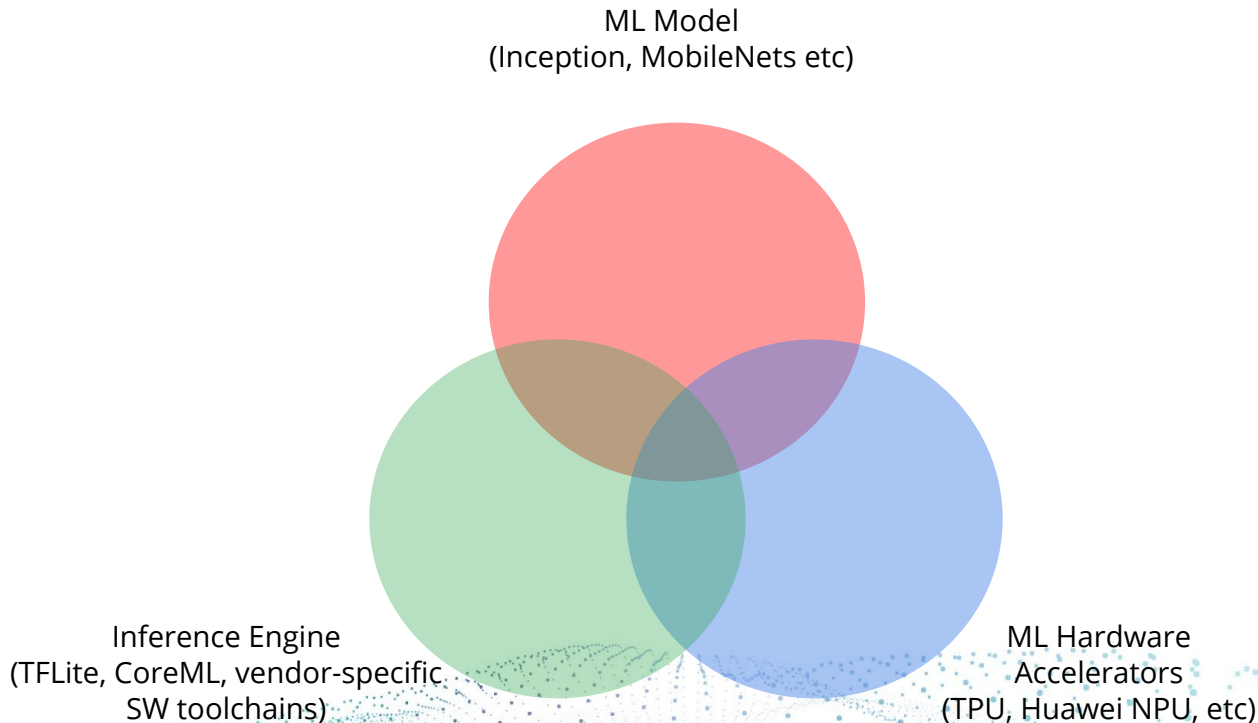
ML Model
(Inception, MobileNets etc)

Inference Engine
(TFLite, CoreML, vendor-specific
SW toolchains)

ML Hardware
Accelerators
(TPU, Huawei NPU, etc)

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?

MLPerf

Are there lessons we can borrow?

MLPerf

Are there lessons we can borrow? Yes!

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

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

MLPerf

SPEC Impact

- Settled **arguments in the marketplace** (grow the pie)

- Resolved internal **engineering debates** (better investments)

- **Cooperative** ⇒ nonprofit Corporation with 22 members

- **Universities** join at modest cost and help drive innovation

- Became **standard** in marketplace, papers, and textbooks

- Needed to **revise regularly** to maintain usefulness:

  SPEC89, SPEC92, SPEC95, SPEC2000, SPEC2006, SPEC2017

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

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?

MLPerf

Growing Number of Supporting Organizations

Alibaba	AMD	Arm	Baidu	Cadence	Cerebras	Cisco
Cray	Dividiti	Enflame Tech	Esperanto	Google	Groq	Huawei
Intel	MediaTek	Mentor Graphics	Mythic	NetApp	NVIDIA	One Convergence
Rpa2ai	Sambanova	Samsung S.LSI	Sigopt	Synopsys	Tensyr	Wave Computing

- 500+ discussion group members

- Researchers from 7 institutions

- 28+ Companies

MLPerf

Supporting Research Institutions

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

MLPerf

MLPerf Goals

- Accelerate progress in ML via **fair and useful measurement**

- Serve both the **commercial and research communities**

- **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 affordable** (so that all can play)

MLPerf

MLPerf Philosophy: **Agile** Benchmark Development
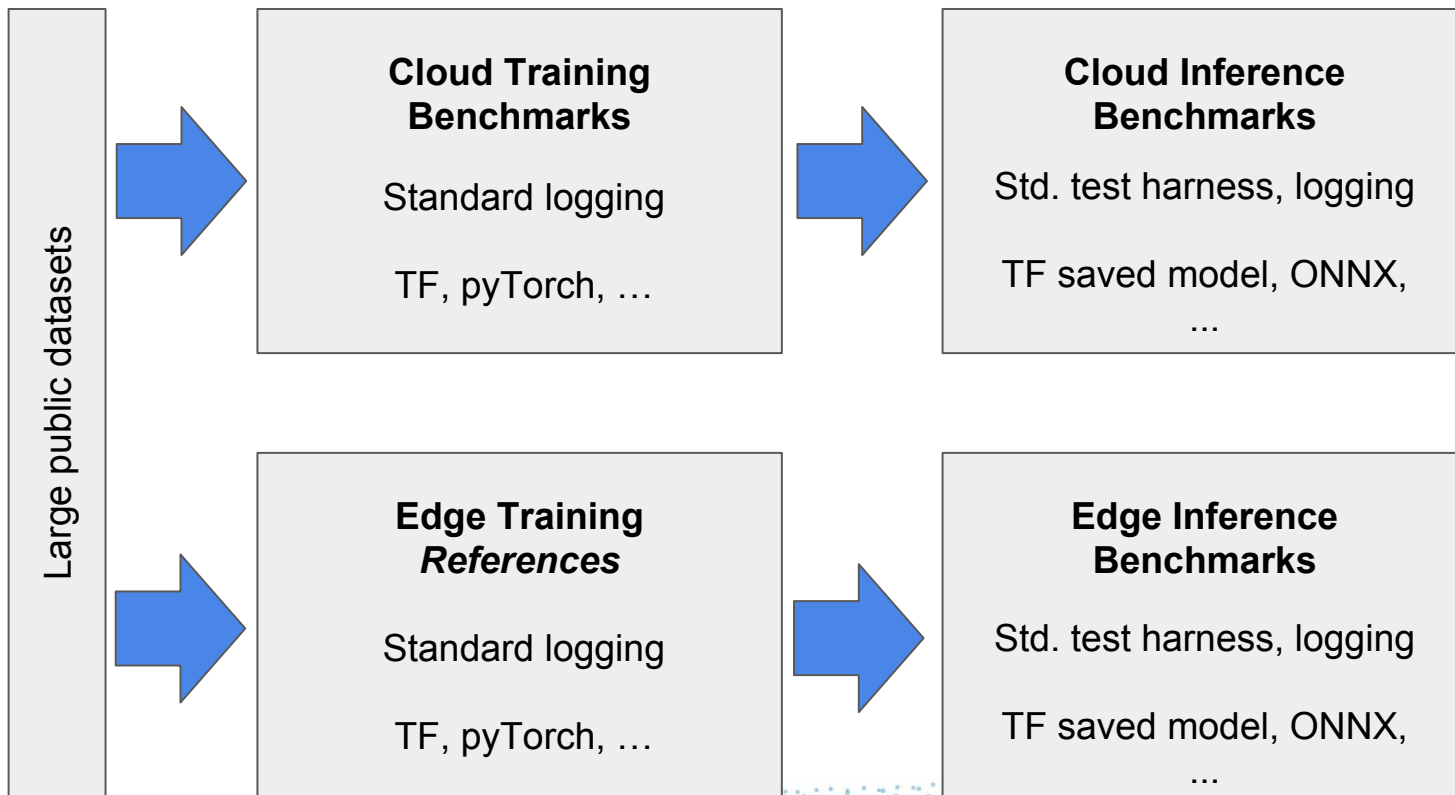

Extremely fast growth...
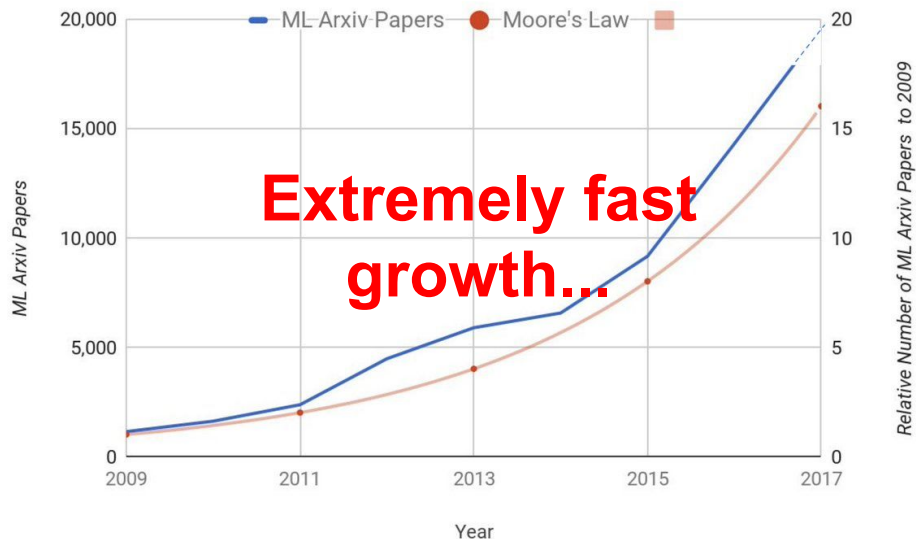
- Rapidly iterate the benchmark suite
  - Remain relevant in the very fast moving machine learning field
  - Correct inevitable mistakes during the fast-paced benchmark formulation
  - Scale problems to match faster hardware, and better systems

- At least initially, revise annually? MLPerf18, MLPerf19, ...

- Like SPEC, have quarterly deadlines and then publish searchable results

MLPerf

MLPerf
(as of Dec 10)

Large public datasets

Cloud Training Benchmarks

Standard logging

TF, pyTorch, …

Cloud Inference Benchmarks

Std. test harness, logging

TF saved model, ONNX, ...

Edge Training *References*

Standard logging

TF, pyTorch, …

Edge Inference Benchmarks

Std. test harness, logging

TF saved model, ONNX, ...

MLPerf

Agile Benchmarking (**Training**) Timeline (in 2018)

May   First general meeting

June   Added benchmarks (volunteers!)

July   Chartered working groups:
                 On-premise, 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   First results submissions!

December   MLPerf results discussion (December 13th)

MLPerf

Agile Benchmarking (**Inference**) Timeline (in 2018)

June   Proposed an inference benchmark suite

July   Gathered the important ML tasks to consider
                  for inclusion

August   WG sanctions the tasks to generate implementations

September   Discussions on the models, datasets and metrics

October   Discussions on power and performance measurement

November   Code development and specification refinement

December   Code drops coming in (as we "speak")

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?

MLPerf

Bootstrapping MLPerf 0.5v **<u>Training</u>** Effort

- Gathered researchers
  - Baidu (DeepBench)
  - Google (TF benchmarks)
  - Harvard (Fathom)
  - Stanford (DAWNBench)

- Combined the best parts from all of our experiences

- Planned to cover both training and inference; initial focus on **<u>training</u>**

MLPerf

MLPerf **Training** Benchmarks 0.5v

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

MLPerf

Toward the Definition of a ML **<u>Inference</u>** Task

- Task description
  - An overview of the ML task

- Dataset
  - A set of inputs and the corresponding ground-truth outputs. The dataset associated with a task also prescribes the input/output data format for the task

- Quality metric
  - A measure of the model's quality/accuracy that is calculated using the ML task's output(s), the ground-truth output(s) from the dataset and a loss function

Task	Task Description	Dataset	Quality metric	Sample Apps
Recognition	Classify an input into one of many categories. Alternatively, generate a high dimensional embedding that can be used for recognition	Imagenet/COCO Input: RGB image of size XX x YY Output: label index	Top-1 error rate	Face authentication, Music recognition

MLPerf

MLPerf Cloud Inference

| ML Tasks	Owner	Framework	Model	Dataset
Image Classification	Guenther	TF and ONNX	Resnet50 1.5v	ImageNet
Object Detection	Itay Hubara ihubara@habana.ai/ christine.cheng@intel.com	PyTorch	(1) VGG16 (2) SSD-MobileNet	MS-COCO
Speech Recognition	Gennady/Anton	PyTorch	DeepSpeech2	Librispeech
Machine Translation	rohit.kalidindi@intel.com	Tensorflow	(1) GNMT http://download.tensorflow.org/models/nmt/10122017/deen_gnmt_model_4_layer.zip (2) transformer	WMT16
Recommendation	adselvar@cisco.com , manasa.kankanala@intel.com	PyTorch	Neural Collaborative Filtering	MovieLens 20M
Text (e.g. Sentiment) Classification	Itay Hubara ihubara@habana.ai	PyTorch	seq2-CNN	IMDB
Language Modeling	gregdiamos@baidu.com	TF	https://github.com/tensorflow/models/tree/master/research/lm_1b	(1) 1 billion words (2) Amazon reviews
Text To Speech	Amit Bleiweiss amit.bleiweiss@intel.com	Caffe2	WaveNet	LJSpeech
Image Segmentation		N/A	MaskRCNN	COCO

MLPerf Edge Inference

| ML Tasks	Owner	Framework	Model	Dataset
Image Classification	(1) Anton (2) Fei and Mejia, Andres <andres.mejia@intel.com>	(1) TF-Lite (2) Caffe2/ONNX	(1) MobileNets-v1.0 224?? (2) ShuffleNet (https://s3.amazonaws.com/download. onnx/models/opset_6/shufflenet.tar.gz)	ImageNet
Object Detection	(1) Yuchen (yuchen.zhou@gm.com) (2) Scott Gardner (MN)/ christine.cheng@intel.com	(1) TF (2) TF-Lite	(1) SSD-ResNet50 (2) SSD-MobileNetsV1	(1) VOC (2) COCO
Speech Recognition	Scott Gardner	TF	DeepSpeech1 (Mozilla)	(1) Librispeech (2) "noisy" validation
Machine Translation	rohit.kalidindi@intel.com	Tensorflow	GNMT http://download.tensorflow.org/models/ nmt/10122017/deen_gnmt_model_4_la yer.zip	WMT16
Text To Speech			WaveNet	
Face Identification	David Lee <david.lee@mediatek.com>	TF-Lite	SphereFace	LFW
Image Segmentation	Carole Wu/Fei Sun <carolejeanwu/feisun@fb.com>	Caffe2/ONNX	MaskRCNN2Go	COCO
Image Enhancement	christine.cheng@intel.com	Tensorflow based on https://github.com/tenso rlayer/srgan	SRGAN (https://github.com/tensorlayer/srgan/re leases/tag/1.2.0)	DIV2K

[MLPerf Cloud/Edge Inference Matrix](#)

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?

MLPerf

"Science"

Metrics   Method

MLPerf

"Science"

Metrics   Method

MLPerf
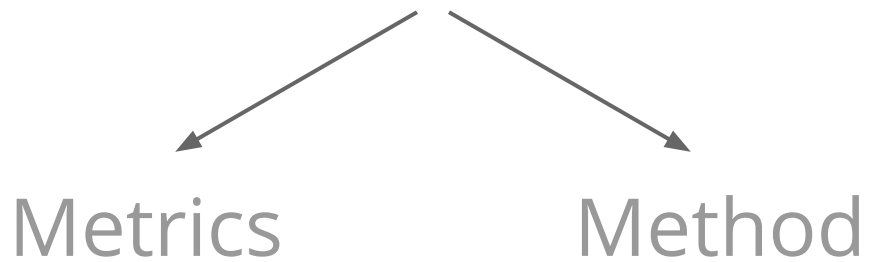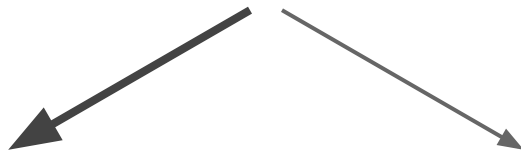
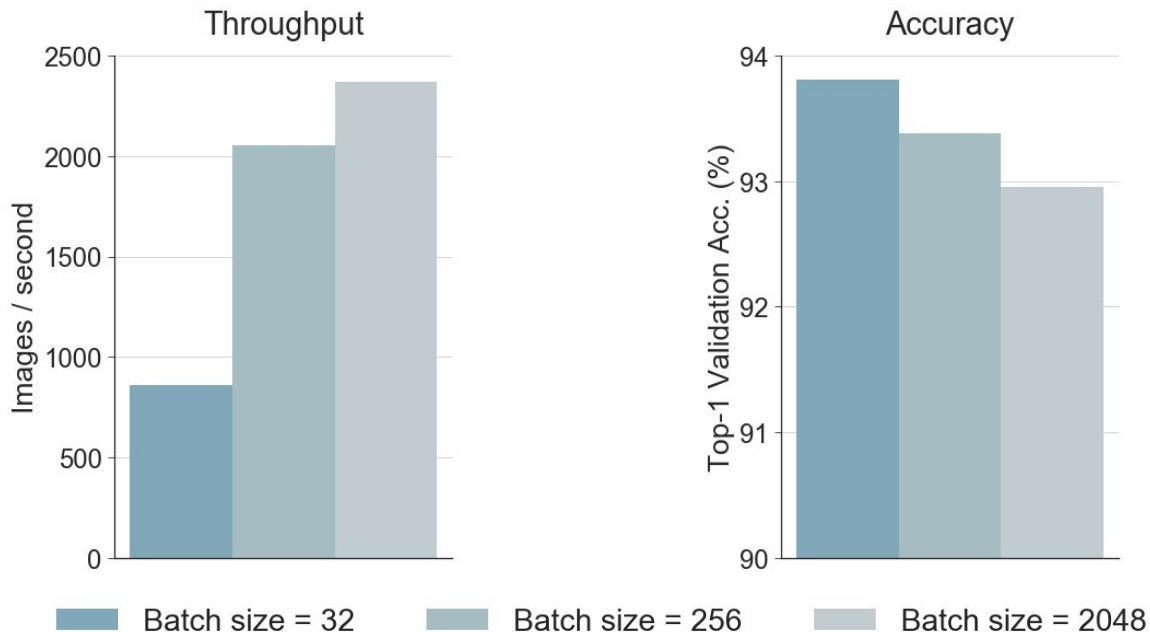# Toward a Unified Metric: Performance *and* Quality

- **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

MLPerf

Performance and Quality Aren't Always Correlated

Training

- End-to-end training of a ResNet56 CIFAR10 model

- Nvidia P100 machine with 512 GB of memory and 28 CPU cores

- TensorFlow 1.2 compiled from source with CUDA 8.0 and CuDNN 5.1

Throughput — Images / second

Accuracy — Top-1 Validation Acc. (%)

Batch size = 32 Batch size = 256 Batch size = 2048

MLPerf

# Training Metric: **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

Performance and Quality Aren't Always Correlated

Inference

- For a given latency target, you can achieve different levels of model quality

- Possible to trade-off model accuracy with complexity

- Model performance (inference/s) is insufficient

https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html

MLPerf
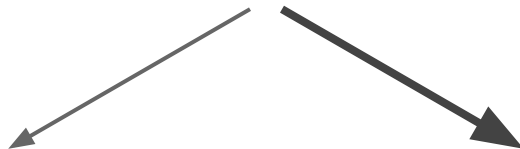
Performance and Quality Aren't Always Correlated

Inference

- For a given latency target, you can achieve different levels of model quality

- Possible to trade-off model accuracy with complexity

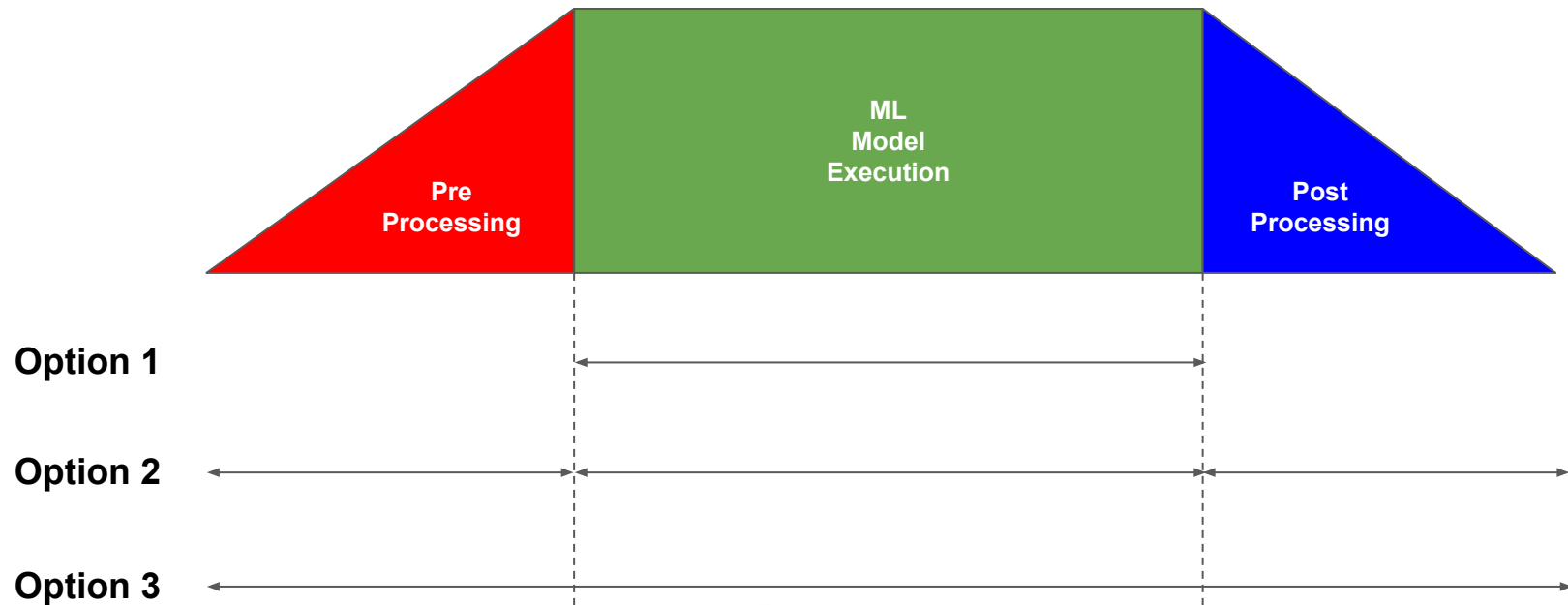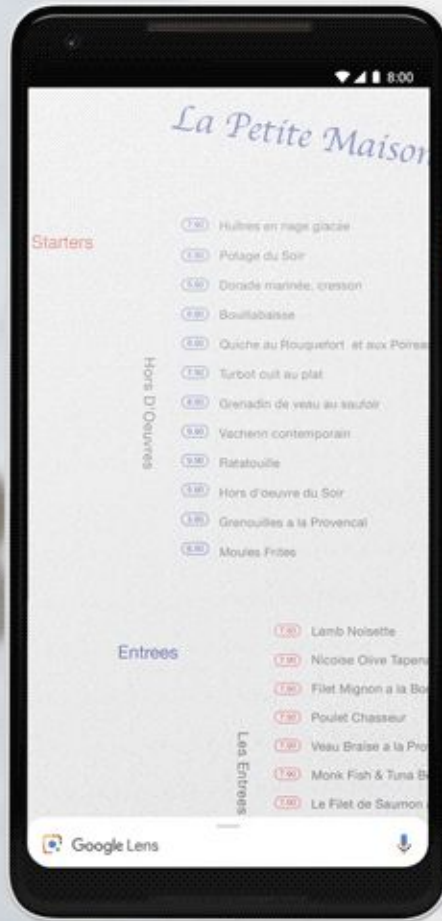- Model performance (inference/s) is insufficient

https://arxiv.org/pdf/1801.04381.pdf

MLPerf

"Science"

Metrics **Method**
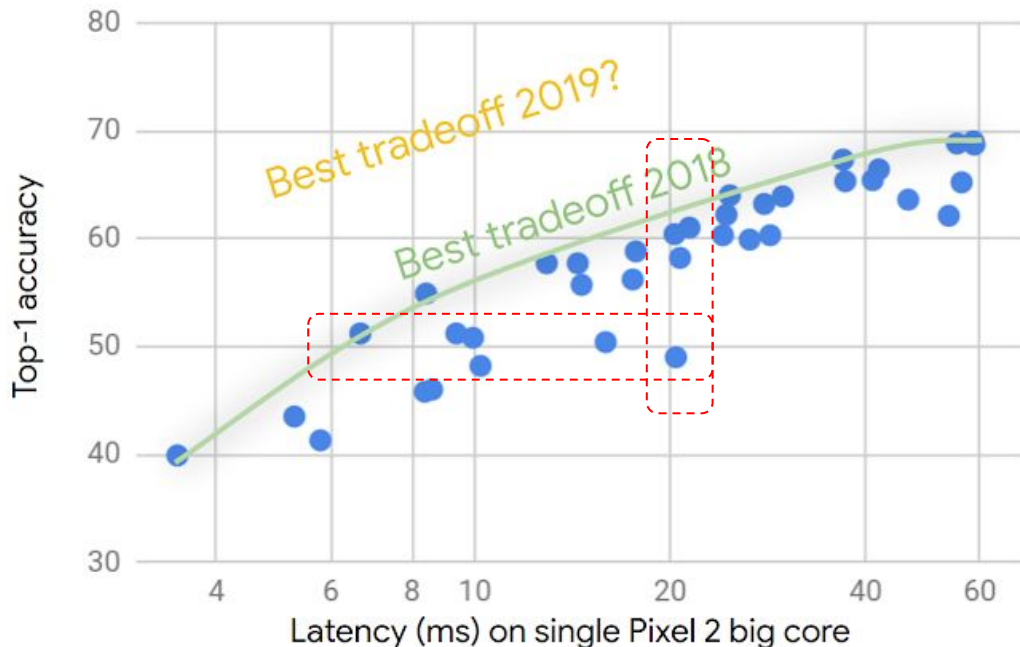
MLPerf
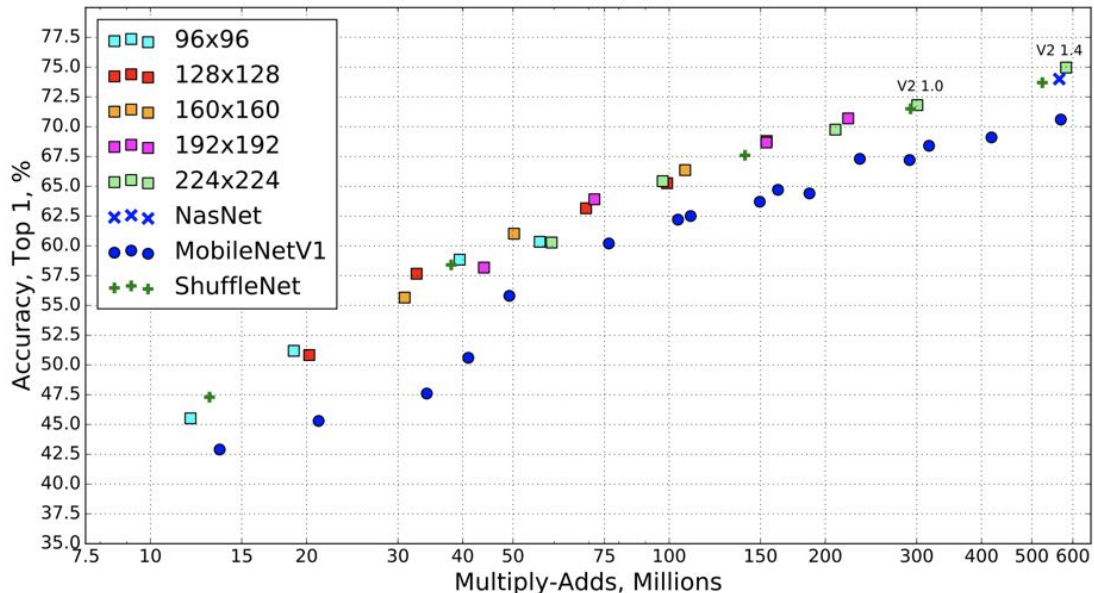
What start/ends do we measure and why?

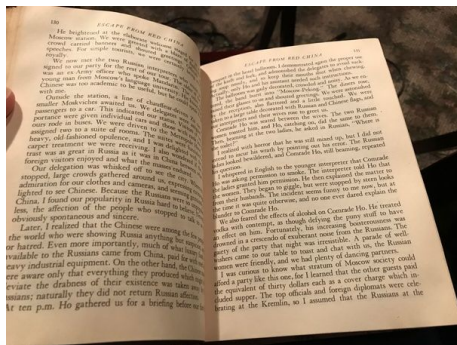MLPerf

On-Device OCR:
A case study

45

[PhotoOCR](#) Normalized Performance (CPU only)

- Sparse

- Dense

Total Inference and Total Non-Inference

Legend: Total Non-Inference, Total Inference

Y-axis: Time normalized to Phone X (sparse)

	Phone X (sparse)	Phone Y (sparse)	Phone X (dense)	Phone Y (dense)
Total Non-Inference	0.17	0.21	1.56 (28%)	1.42 (25%)
Total Inference	0.83	1.36	4.02	4.35
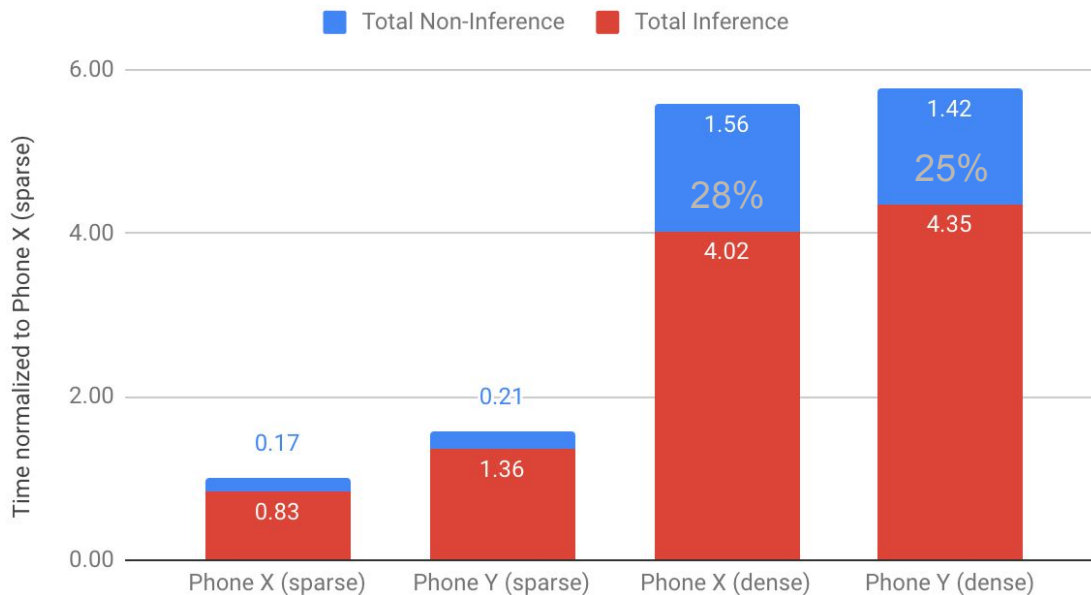
MLPerf

[PhotoOCR](#) Task Breakdown

Detection

Detection (Non-inference time and Inference time)

■ Non-inference time   ■ Inference time

Phone X (sparse)	27% 0.27 / 0.73
Phone Y (sparse)	20% 0.32 / 1.67
Phone X (dense)	49% 0.88 / 0.91
Phone Y (dense)	48% 0.69 / 0.72

Time normalized to Phone X (sparse)

Recognition

Recognition (Non-inference time and Inference time)

■ Non-inference time   ■ Inference time

Phone X (sparse)	0.09 / 0.91
Phone Y (sparse)	0.09 / 0.97
Phone X (dense)	1.30 / 7.20
Phone Y (dense)	1.51 / 8.08

Time normalized to Phone X (sparse)

Do we account for pre- and post-processing times in the inference run test?

MLPerf

MLPerf Challenges

Cloud (Training E.g.)

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

Edge (Inference E.g.)

- **Quantizations**
- Sparsity
- Pruning
- **Scores**
- Variance
- **Power**
- ...

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?

Where are we heading now?

- First version: **reference** code, in two frameworks, of each benchmark.

- Resolving or controlling the **variance** issues.

- Working on the **inference** suite.

- Getting to **governance**, and an umbrella organization.

MLPerf

Reference Implementations → Call for Submissions
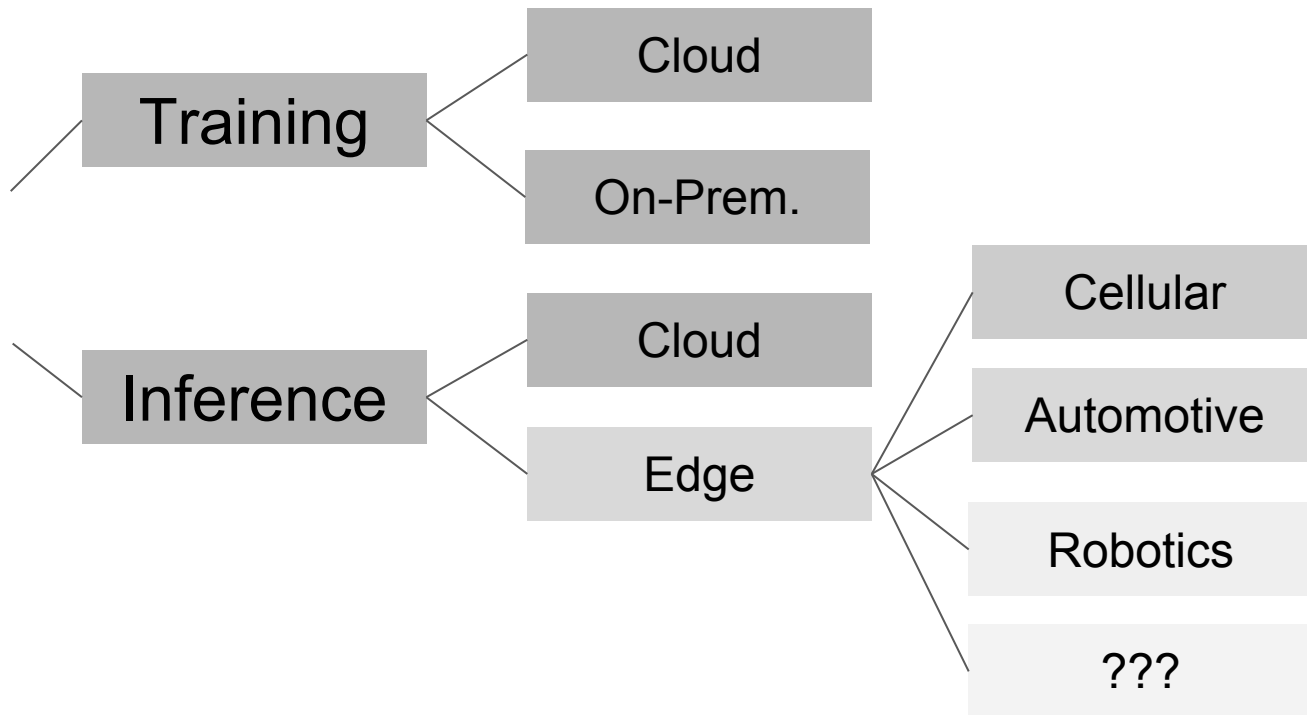
Closed division submissions

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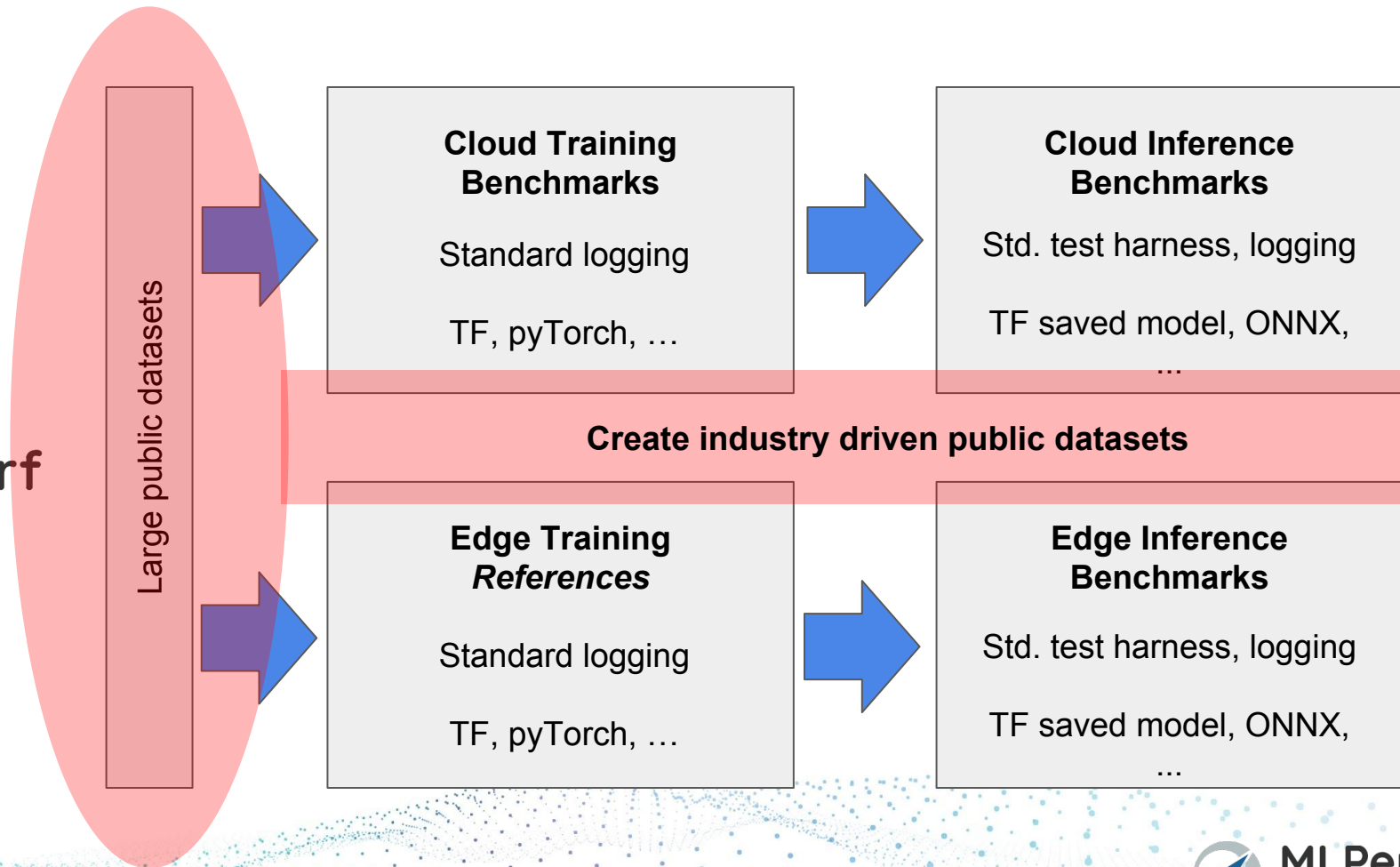- Simplifies work for HW groups

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?**

MLPerf

MLPerf

Large public datasets

Cloud Training Benchmarks

Standard logging

TF, pyTorch, …

Cloud Inference Benchmarks

Std. test harness, logging

TF saved model, ONNX,

...

Create industry driven public datasets

Edge Training *References*

Standard logging

TF, pyTorch, …

Edge Inference Benchmarks

Std. test harness, logging

TF saved model, ONNX,

...

MLPerf

Policy

Large public datasets

Benchmarks and Standardization (MLPerf)

(nothing is set in stone yet, we are looking for ideas)

MLPerf

55

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?*

MLPerf

Concluding thoughts…

MLPerf

Recap of "**The Vision Behind MLPerf**"

- Machine Learning needs benchmarks!

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

- MLPerf Training: v0.5 deadline was in November

- MLPerf Inference is under construction
  - Inference workload suite under development
  - Q1 reference implementations finalized
  - Q2/3 solicit inference result submissions"

(for rapid iteration to work, we need good input!)

MLPerf

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!

MLPerf

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

MLPerf

Acknowledgements

Peter Mattson

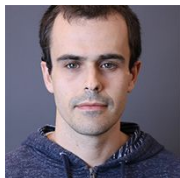Cliff Young

David Patterson

Carole-Jean Wu

Greg Diamos
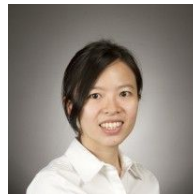
... and countless other working group members!

MLPerf

Thank You

MLPerf