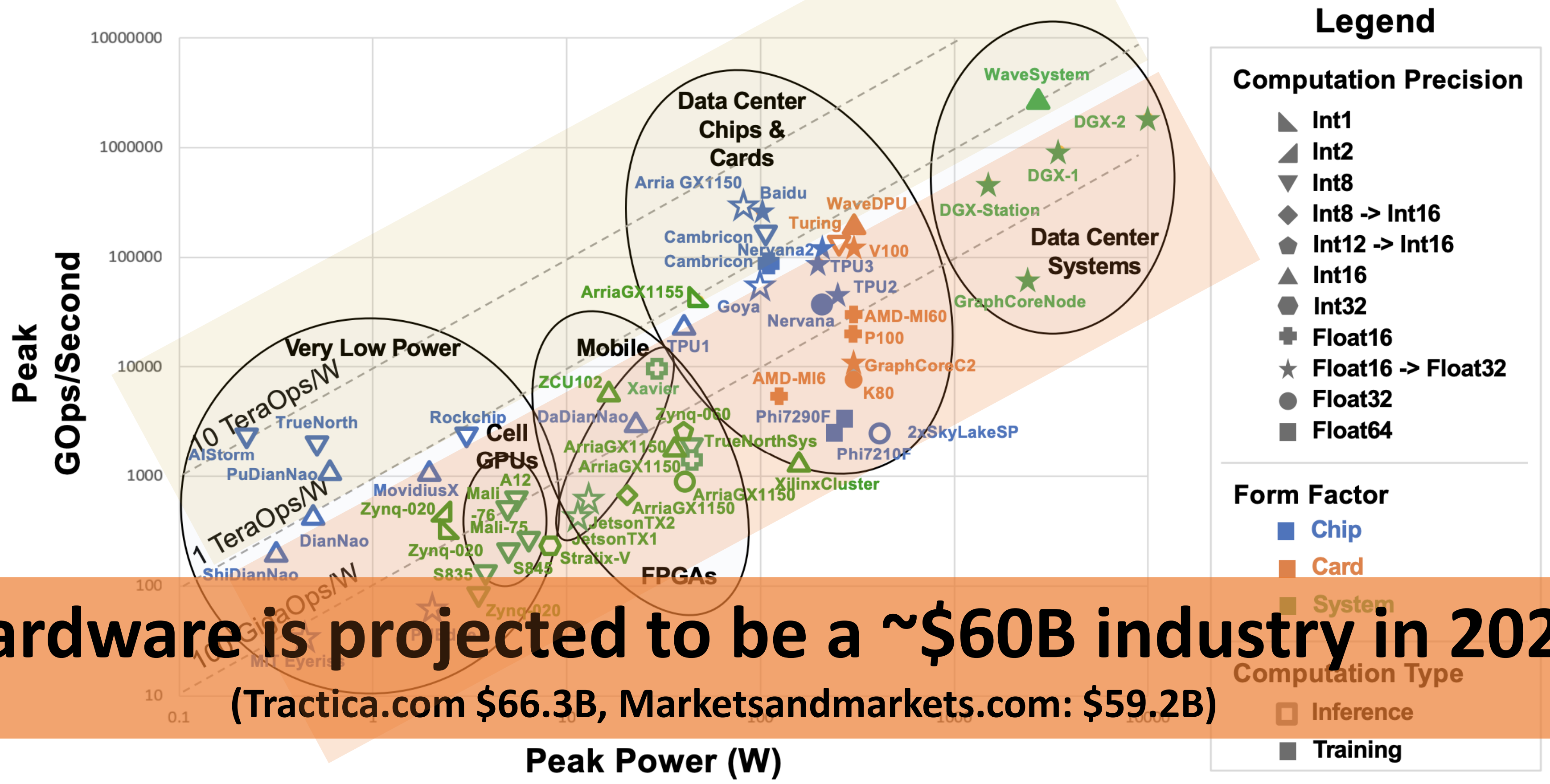
MLPerf: An Industry Standard Benchmark Suite for Machine Learning

Carole-Jean Wu
carolejeanwu@fb.com
Facebook, ASU
MLPerf Inference Chair and
Recommendation Benchmark Advisory Board Chair
(Work by Many People in the MLPerf Community)

2020 embedded VISION summit

MLPerf

Deep Learning is Fueling the HW Renaissance



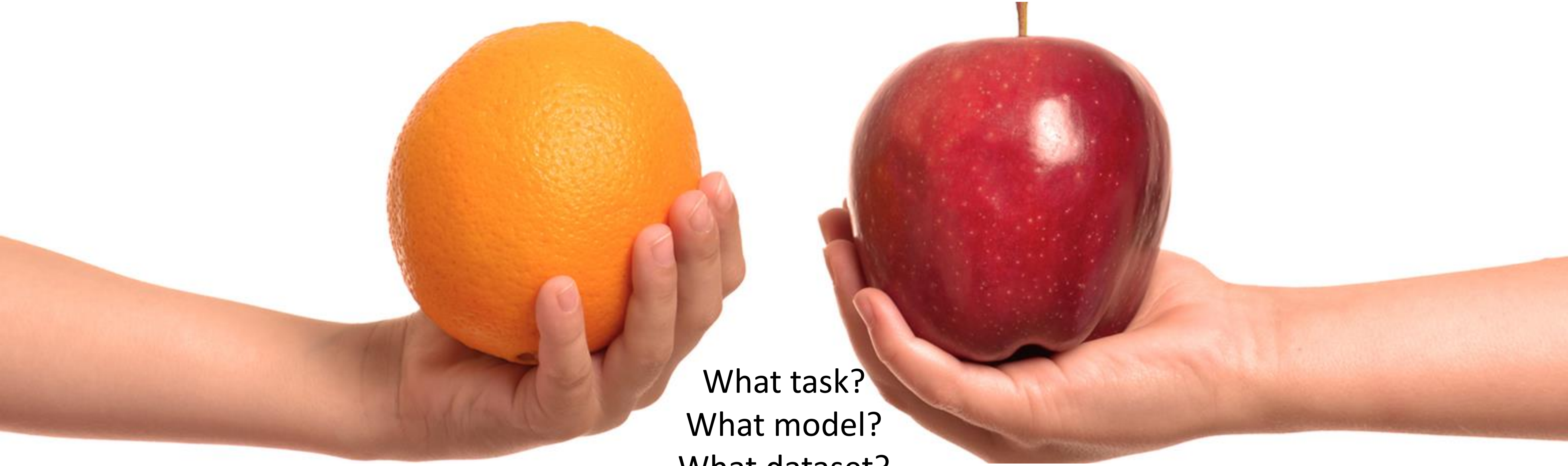
ML hardware is projected to be a ~$60B industry in 2025.

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

Survey and Benchmarking of AI Accelerators. Reuther et al. MIT Lincoln Lab Supercomputing Center. Arxiv-2019

MLPerf

How Do We Compare AI ~~Hardware?~~ *Systems*



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

We Need a Benchmark for ML

"What get measured, gets improved."

— Peter Drucker

Benchmarking aligns research with development,
engineering with marketing,
and competitors across the industry in
pursuit of a clear objective

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?
- What comes next for MLPerf?
- How can we contribute to MLPerf?

Are There Lessons We Can Borrow?

MLPerf

Successful History in Benchmarks

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

⇒ **Fueled the golden age of microprocessor design**

How Do 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?
    - Training
    - Inference
- What comes next for MLPerf?
- How can we contribute to MLPerf?

MLPerf is an ML Performance Benchmarking Effort with Wide Industry and Academic Support

2020 embedded VISION summit

AI Labs.tw
Alibaba
AMD
Andes Technology
Aon Devices
Arm
Baidu
Cadence
Calypso AI
Centaur Technology
Cerebras
Ceva
Cisco
Code Reef
Cray
CTuning Foundation
Dell
Dividiti
DDN Storage
Edgify
Enflame Tech
Esperanto
Facebook
FuriosaAI
Google
Groq
Habana
Hewlett Packard Enterprise
Hop Labs
Horizon Robotics
Iluvatar
Inspur
Inzone AI
Intel
In-Q-Tel
Lenovo
MediaTek
Mentor Graphics
Microsoft
Myrtle
Mythic
NetApp
NVIDIA
One Convergence
PathPartner Technology
Qualcomm
Rpa2ai
Sambanova
Samsung S.LSI
Sigopt
SiMa AI
Skymizer
Supermicro
Synopsys
Tencent
Tensyr
Teradyne
Transpire Ventures
VMware
Volley
Wave Computing
Wiwynn
WekaIO
Xilinx

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

MLPerf

ML Benchmark Design Overview

<table>
<tr><th>Big Questions</th><th>Training</th><th>Inference</th></tr>
<tr><td>1. Benchmark definition</td><td colspan="2">What is a benchmark task?</td></tr>
<tr><td>2. Benchmark selection</td><td colspan="2">Which benchmark tasks?</td></tr>
<tr><td>3. Metric definition</td><td colspan="2">What is performance?</td></tr>
<tr><td>4. Implementation equivalence</td><td colspan="2">How do submitters run on very different hardware/software systems?</td></tr>
<tr><td rowspan="2">5. Issues specific to training or inference</td><td>Which hyperparameters can submitters tune?</td><td rowspan="2">Quantization, calibration, and/or retraining?</td></tr>
<tr><td>Reduce result variance?</td></tr>
<tr><td>6. Presentation</td><td colspan="2">Do we normalize and/or summarize results?</td></tr>
</table>

Training Benchmark Definition

Dataset

Train a model

Target Quality[1]

75.9%

Do we specify the model?

1. Target quality set by experts in area, raised as SOTA improves

MLPerf Training Benchmark Suite (v0.7)

Benchmark	Data set	Model	Quality Threshold
Image classification	ImageNet (Deng et al., 2009)	ResNet-50 v1.5 (MLPerf, 2019b)	75.9% Top-1 accuracy
Object detection (lightweight)	COCO 2017 (Lin et al., 2014)	SSD-ResNet-34 (Liu et al., 2016)	23.0% mAP
Instance segmentation and object detection (heavyweight)	COCO 2017 (Lin et al., 2014)	Mask R-CNN (He et al., 2017a)	37.7 Box min AP, 33.9 Mask min AP
NLP	Wikipedia 01/01/2020	BERT	0.712 Mask-LM accuracy
Translation (nonrecurrent)	WMT17 EN-DE (WMT, 2017)	Transformer (Vaswani et al., 2017)	25.0 BLEU
Recommendation	Criteo Terabyte CTR	DLRM (Naumov et al., 2019)	0.8025 AUC
Reinforcement learning	Go (19 x 19 Board)	MiniGo (MLPerf, 2019a)	50% win rate

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 SOTA
- Time includes preprocessing and validation over of N runs
- MLPerf provides the reference implementations that achieve quality target

Agenda

- *Why ML needs a benchmark suite?*
    - *Are there lessons we can borrow?*
- **What is MLPerf?**
    - *How does MLPerf curate a benchmark?*
    - *Training*
    - **Inference**
- What comes next for MLPerf?
- How can we contribute to MLPerf?

Inference Benchmark Definition

Input

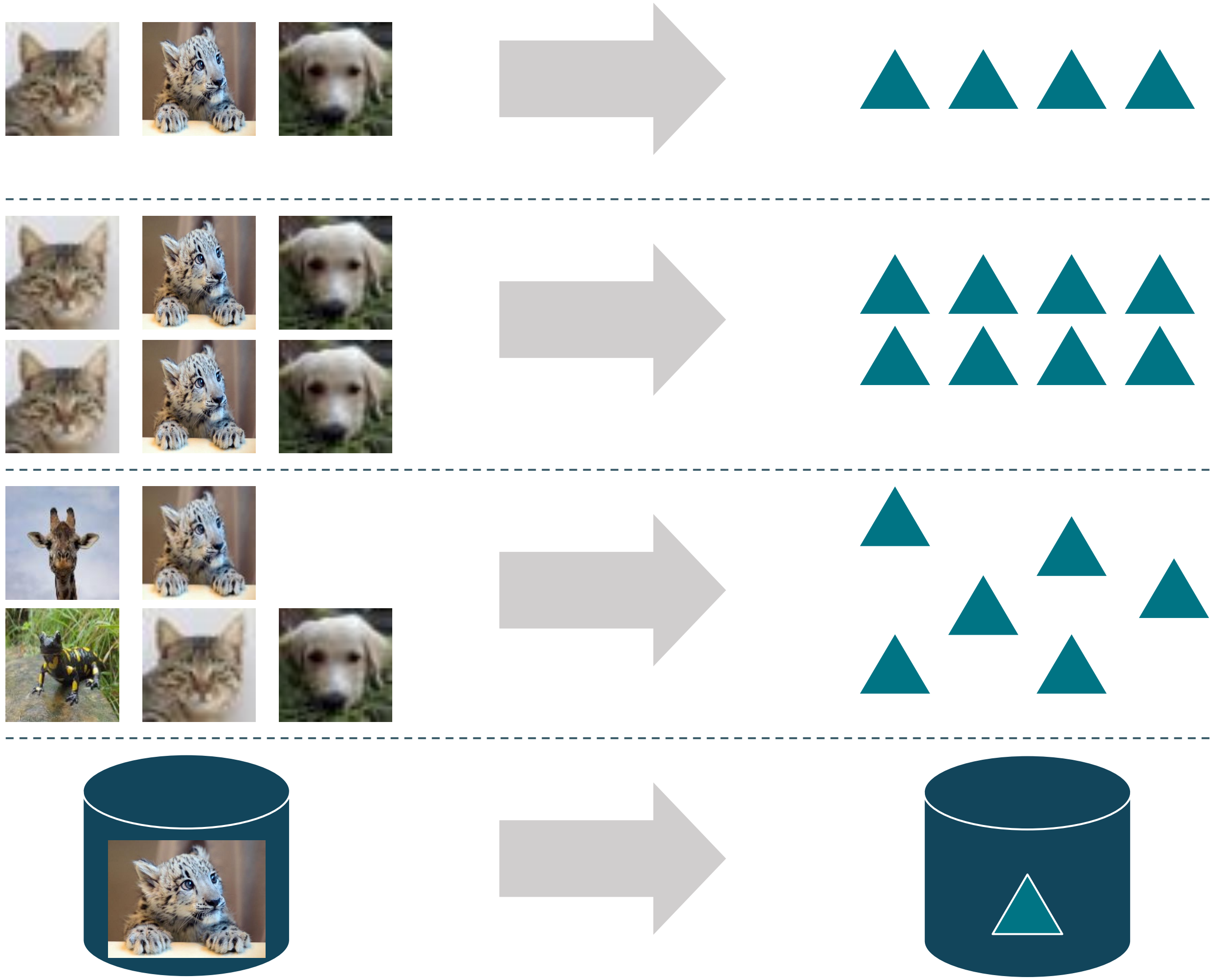
Result
(with required quality, e.g. 75.1%)

Trained ResNet

"Leopard"

Do you specify the model? Closed Division does, Open Division does not.

But How is Inference Really Used?



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

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

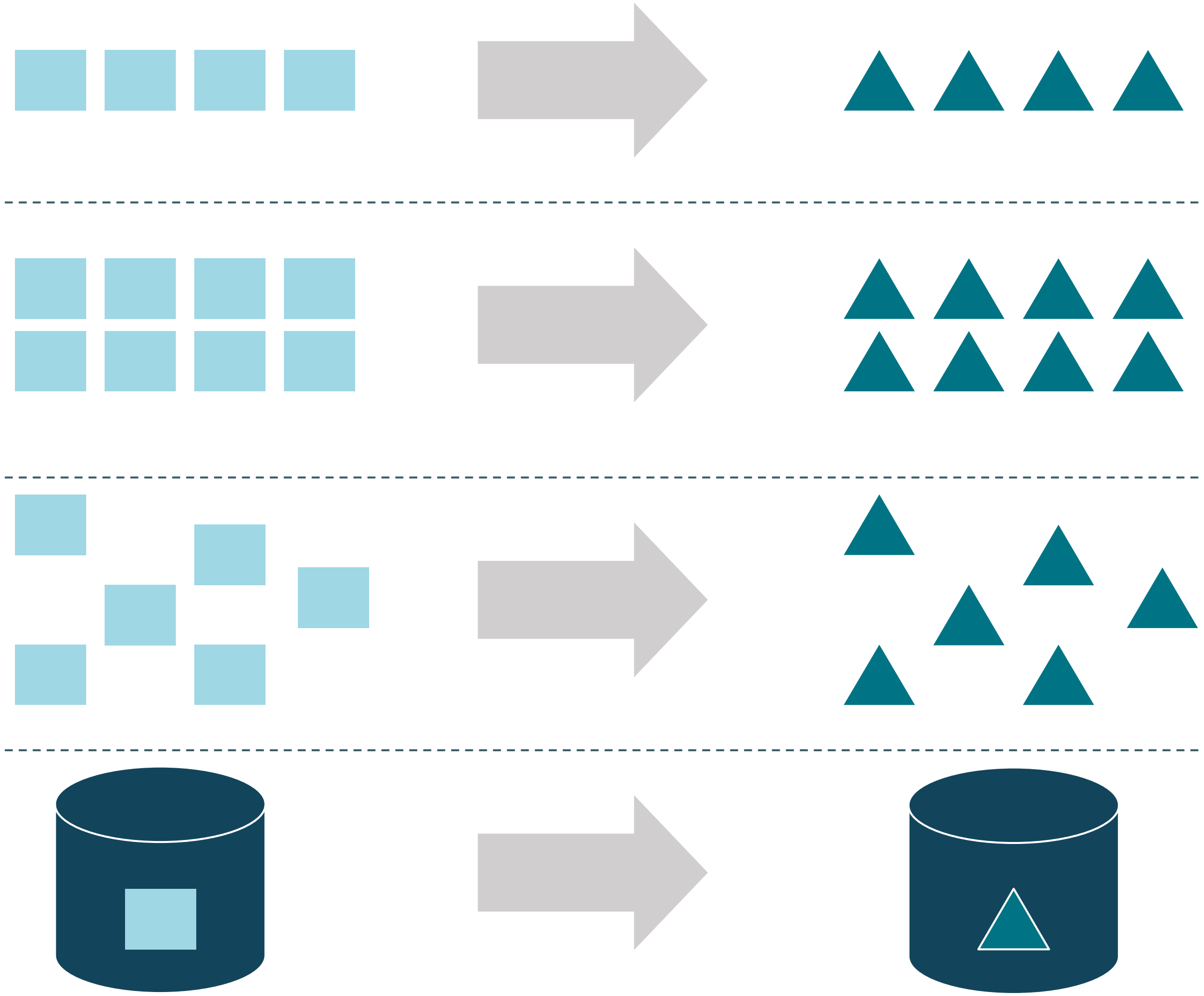
Server
(e.g. translation app)

Offline
(e.g. photo sorting app)

MLPerf Inference Benchmark Suite (v0.5)

Area	Task	Model	Dataset	Quality	Server latency constraint	Multi-Stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	99% of FP32 (76.46%)	15 ms	50 ms
Vision	Image classification	MobileNets-v1 224	ImageNet (224x224)	98% of FP32 (71.68%)	10 ms	50 ms
Vision	Object detection	SSD-ResNet34	COCO (1200x1200)	99% of FP32 (0.20 mAP)	100 ms	66 ms
Vision	Object detection	SSD-MobileNets-v1	COCO (300x300)	99% of FP32 (0.22 mAP)	10 ms	50 ms
Language	Machine translation	GNMT	WMT16	99% of FP32 (23.9 BLEU)	250 ms	100 ms

Inference Metric: One Metric for Each Scenario



Scenario	Metric
Single stream (e.g. cell phone augmented vision)	**Latency**
Multiple stream (e.g. multiple camera driving assistance)	**# of streams** subject to latency bound
Server (e.g. translation app)	**QPS** subject to latency bound
Offline (e.g. photo sorting app)	**Throughput**

Inference Benchmark Suite v0.5

Area	Task	Model	Dataset	Quality	Server latency constraint	Multi-Stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	99% of FP32 (76.46%)	15 ms	50 ms
Vision	Image classification	MobileNets-v1 224	ImageNet (224x224)	98% of FP32 (71.68%)	10 ms	50 ms
Vision	Object detection	SSD-ResNet34	COCO (1200x1200)	99% of FP32 (0.20 mAP)	100 ms	66 ms
Vision	Object detection	SSD-MobileNets-v1	COCO (300x300)	99% of FP32 (0.22 mAP)	10 ms	50 ms
Language	Machine translation	GNMT	WMT16	99% of FP32 (23.9 BLEU)	250 ms	100 ms

MLPerf Inference Benchmark Suite v0.7

	Neural Network
Vision	ResNet-50 v1.5
	SSD ResNet-34
	SSD MobileNet v1 (edge)
	3D UNET
Speech	RNN-T
Language	BERT Large
Commerce	DLRM (datacenter)

MOBILE	Neural Network
Vision	MobileNet EdgeTPU
	SSD-MobileNet v2
	DeepLabv3
Language	Mobile-BERT

MLPerf

MLPerf Training and Inference Call for Submissions

Closed division submissions

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

Open division submissions

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

Timeline for Benchmark Result Submission

Week -12	Week -6	Week 0	Week 4
Benchmark Freeze	Code & Rule Freeze	Submission Deadline	Result Publication

- Discuss details on rules and models and implementations
- Develop the SW stack

- Clarify rules
- Fine-tune SW for HW
- Pass compliance check
- Sign CLA
- Release SW

- Review submission results
- Prepare result release

ATTEND WEEKLY SUBMITTER WORKING GROUP MEETINGS

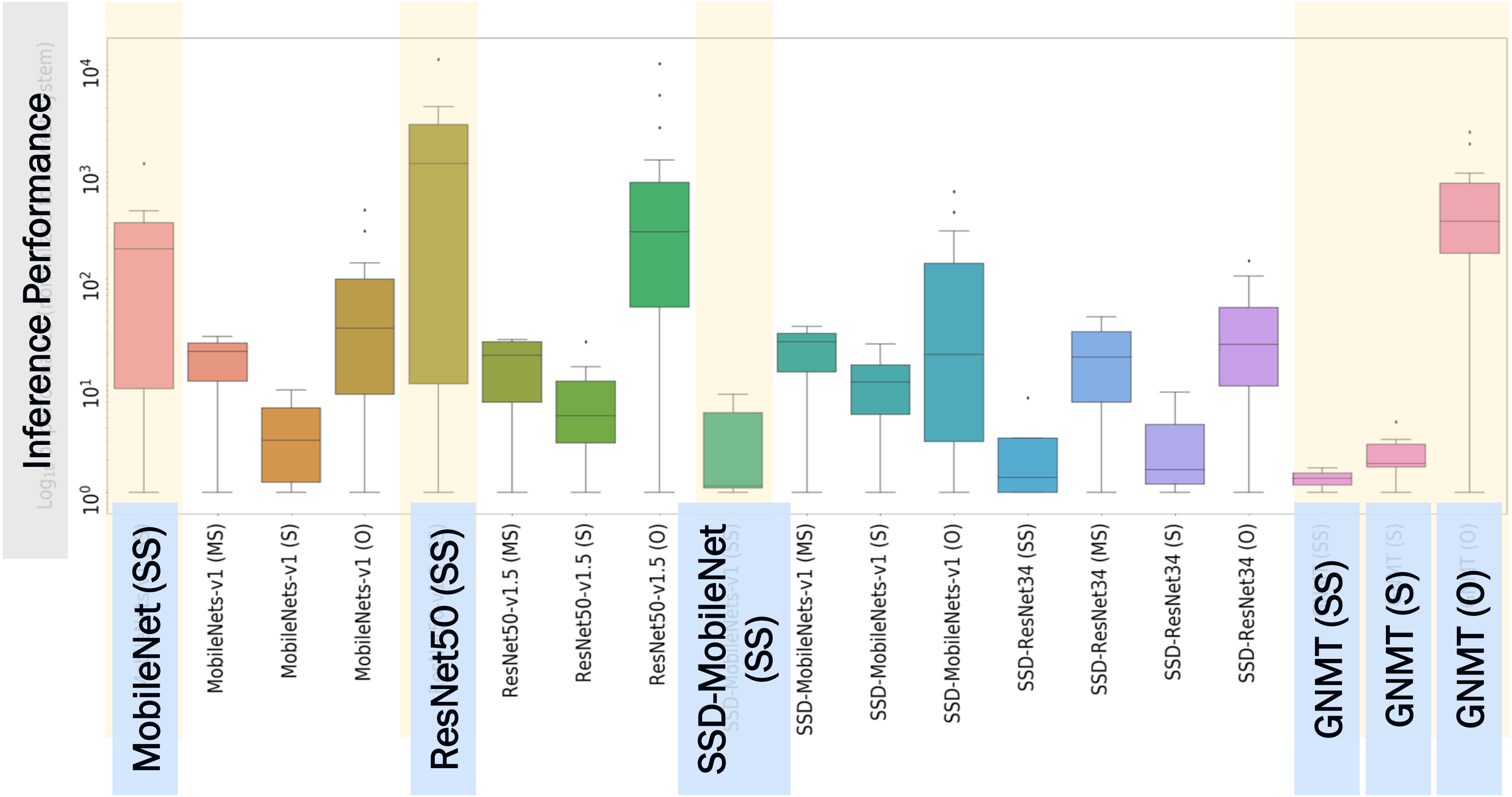
MLPerf Inference Results

https://mlperf.org/inference-results/

Submissions · Benchmarks · Scenarios · Systems · Inference Scores

ID	Submitter	System	MobileNet-v1 Stream	MobileNet-v1 MultiS	MobileNet-v1 Server	MobileNet-v1 Offline	ResNet-50 v1.5 Stream	ResNet-50 v1.5 MultiS	ResNet-50 v1.5 Server
CATEGORY: Available									
Inf-0.5-1	Alibaba Cloud	Alibaba Cloud T4				17,473.60			
Inf-0.5-2	Dell EMC	Dell EMC R740			67,124.18	71,214.50			20,742.8
Inf-0.5-3	Dell EMC	Dell EMC R740xd with 2nd generation Intel® Xeon® Scalable Processor							
Inf-0.5-4	Dell EMC	Dell EMC R740xd with 2nd generation Intel® Xeon® Scalable Processor							
Inf-0.5-5	dividiti	Raspberry Pi 4 (rpi4)	394.34				1,916.65		
Inf-0.5-6	dividiti	Raspberry Pi 4 (rpi4)	103.60				448.31		
Inf-0.5-7	dividiti	Linaro HiKey960 (hikey960)	121.11				518.07		
Inf-0.5-8	dividiti	Linaro HiKey960 (hikey960)	50.77				203.99		
Inf-0.5-9	dividiti	Linaro HiKey960 (hikey960)	143.07				494.90		
Inf-0.5-10	dividiti	Huawei Mate 10 Pro (mate10pro)	74.20				354.13		
Inf-0.5-11	dividiti	Huawei Mate 10 Pro (mate10pro)	111.60				494.92		
Inf-0.5-12	dividiti	Firefly-RK3399 (firefly)	120.56				695.11		
Inf-0.5-13	dividiti	Firefly-RK3399 (firefly)	106.49				447.90		
Inf-0.5-14	dividiti	Firefly-RK3399 (firefly)	80.12				391.02		
Inf-0.5-15	Google	Cloud TPU v3							16,014.2
Inf-0.5-16	Google	2x Cloud TPU v3							
Inf-0.5-17	Google	4x Cloud TPU v3							
Inf-0.5-18	Google	8x Cloud TPU v3							
Inf-0.5-19	Google	16x Cloud TPU v3							
Inf-0.5-20	Google	32x Cloud TPU v3							
Inf-0.5-21	Habana Labs	HL-102-Goya PCI-board					0.24	700.00	
Inf-0.5-22	Intel	Intel® Xeon® Platinum 9200 processors							
Inf-0.5-23	Intel	Intel® Xeon® Platinum 9200 processors	0.49		27,244.81	29,203.30	1.37		4,850.6

MLPerf Inference v0.5 Presented Almost 600 Results Across a Wide Range of Platform Scales



Normalized performance on log10 scale for models and scenarios. Results are normalized to the slowest system and show up to a 10,000X range in performance.

Agenda

- *Why ML needs a benchmark suite?*
- *Are there lessons we can borrow?*
- *What is MLPerf?*
  - *How does MLPerf curate a benchmark?*
  - *Training*
  - *Inference*
- **What comes next for MLPerf?**
- How can we contribute to MLPerf?

MLPerf

MLPerf Focuses for 2020

Evolve the benchmark suites fairly

Improve efficiency information

Reduce result sparsity

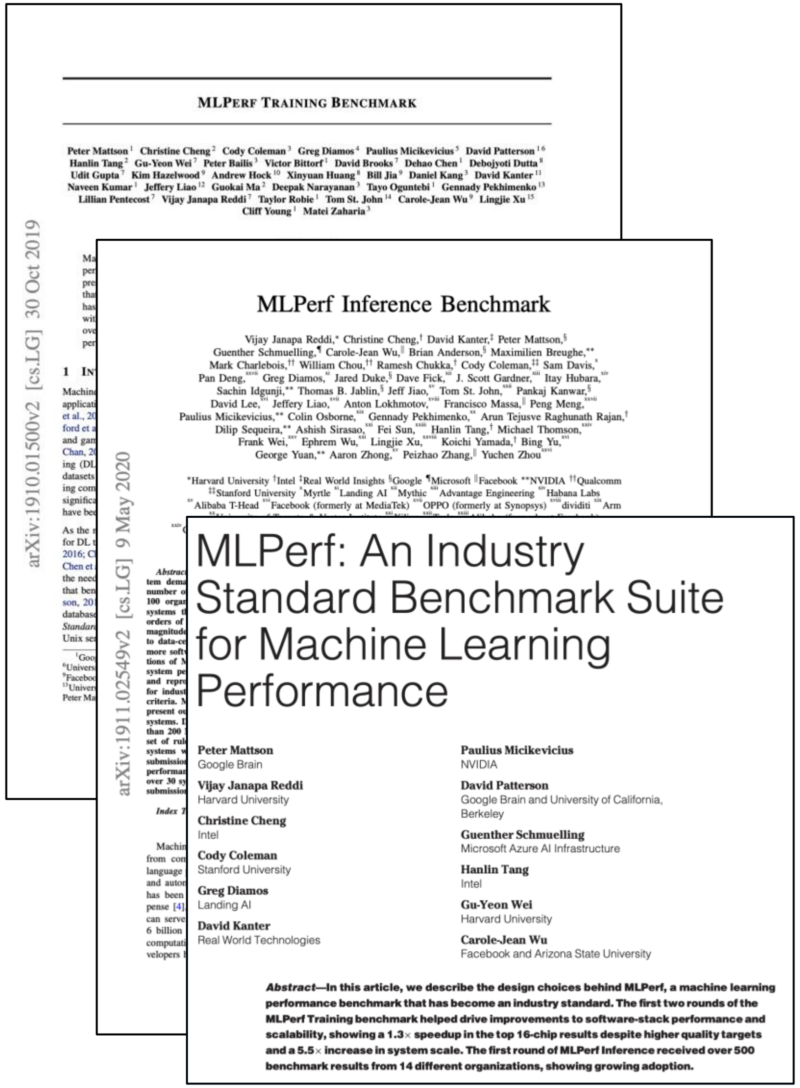
Reduce benchmarking cost

Inference (Datacenter/Edge)

Inference (Mobile)

Conclusion

- **Benchmarking ML Systems is hard** due to the fragmented ecosystem
- MLPerf is a **community-driven ML benchmark** for the HW/SW industry
- The benchmark suite helps level the playing field, **enabling ML system comparison**
  - Defines Tasks, Scenarios, Datasets, Methods
  - Establish clear set of metrics and divisions
  - Allows for hardware/software flexibility



MLPerf is the Work of Many

Aaron Zhong
Abid Muslim
Andrew Hock
Ankur Ankur
Anton Lokhmotov
Arun Rajan
Ashish Sirasao
Atsushi Ike
Bill Jia
Bing Yu
Brian Anderson
Carole-Jean Wu
Christine Cheng
Cliff Young
Cody Coleman
Colin Osborne
Daniel Kang
Dave Fick
David Brooks
David Kanter
David Patterson
Debajyoti Pal
Debo Dutta
Deepak Narayanan
Dehao Chen
Dilip Sequeira
Ephrem Wu
Fei Sun
Francisco Massa
Frank Wei
Gennady Pekhimenko
George Yuan
Greg Diamos
Gu-Yeon Wei
Guenther Schmuelling
Guokai Ma
Hanlin Tang
Itay Hubara
J. Scott Gardner
Jacob Balma
Jared Duke
Jeff Jiao
Jeffery Liao
Jonah Alben
Jonathan Cohen
Kim Hazelwood
Koichi Yamada
Lillian Pentecost
Lingjie Xu
Mark Charlebois
Masafumi Yamazaki
Matei Zaharia
Maximilien Breughe
Michael Thomson
Naveen Kumar
Pan Deng
Pankaj Kanwar
Paulius Micikevicious
Peizhao Zhang
Peng Meng
Peter Bailis
Peter Baldwin
Peter Mattson
Ramesh Chukka
Sachin Idgunji
Sam Davis
Sarah Bird
Sergey Serebryakov
Steve Farrell
Taylor Robie
Tayo Oguntebi
Thomas B. Jablin
Tom St. John
Tsuguchika Tabaru
Udit Gupta
Victor Bittorf
Vijay Janapa Reddi
William Chou
Xinyuan Huang
Yuchen Zhou

MLPerf

MLPerf Needs Your Help!

- Join the discussion community at **mlperf.org**
- Help us by joining a working group
  - Cloud training and inference
  - Mobile Inference, HPC
  - On-premises scale, submitters, special topics
  - Help us design submission criteria, to include the data you want
- Propose new benchmarks and data sets
- Address challenging “special topic” issues
- Submit your benchmark results!

More at MLPerf.org

info@mlperf.org

Latest Inference Results
October 19, 2020

MLPerf: An Industry Standard Benchmark Suite for Machine Learning

Carole-Jean Wu
carolejeanwu@fb.com

MLPerf Inference Chair
Recommendation Benchmark Advisory Board Chair