MLPerf Training Benchmark

Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim Hazelwood, Andrew Hock, Xinyuan Huang, Atsushi Ike, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Tsuguchika Tabaru, Carole-Jean Wu, Lingjie Xu, Masafumi Yamazaki, Cliff Young, and Matei Zaharia

MLSys 2020

Artifacts Available
Artifacts Evaluated Functional
Results Replicated

MLPerf

Why MLPerf?

MLPerf

Why MLPerf?

Machine learning (ML) is changing whole industries such as automotive safety, e-commerce, and medicine.

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

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

Need a standard benchmark to provide the field/industry with clear metrics.

Prior Work

SPEC and TPC, consortium-backed standards but not ML

DeepBench, but only ML primitives

Fathom and TBD, measure throughput for broad ML suite

DAWNBench, measure time-to-train for a few ML tasks

MLPerf = **consortium +**
broad suite +
time-to-train +
novel contributions

https://commons.wikimedia.org/wiki/File:Chocolate02.jpg https://commons.wikimedia.org/wiki/File:Peanut_Butter_Texture.jpg

Goals

What:

Enable fair comparisons

Encourage innovation

Serve commercial and research communities

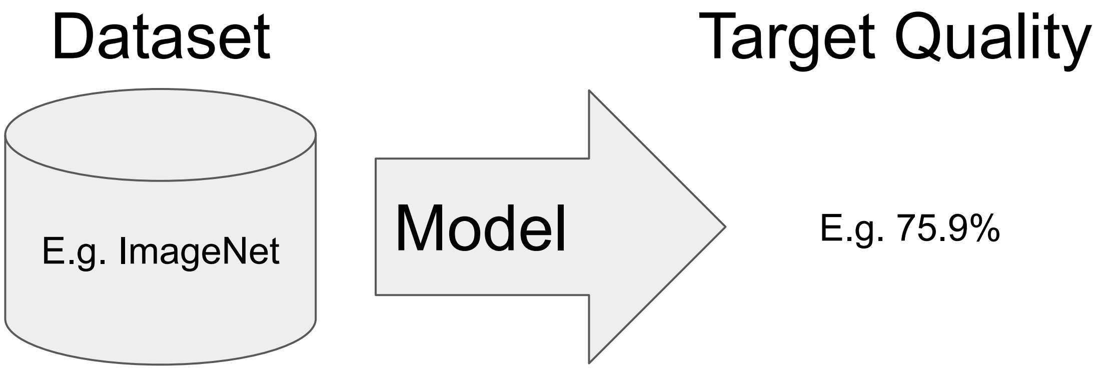
How:

Ensure reliable results

Keep benchmarking easy and affordable

MLPerf Training Benchmark

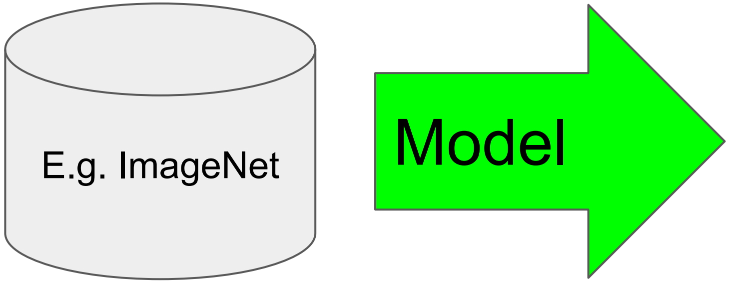
MLPerf Training benchmark definition



Two divisions with different model restrictions

Dataset

Target Quality



E.g. 75.9%

Closed division: specific model e.g. ResNet v1.5 → direct comparisons

Open division: any model → innovation

Benchmark suite

Area	Task	**Dataset**	**Model** (closed)	**Target Quality** (v0.5)
Vision	Image recognition	ImageNet	ResNet	74.9% Top-1
	Object detection	COCO	SSD	21.2 mAP
	Object segmentation	COCO	Mask R CNN	37.7 Box mAp 33.9 Mask minAP
Language	Translation	WMT Eng.-German	NMT	21.8 Sacre Bleu
	Translation	WMT Eng.-German	Transformer	25.0 Bleu
Commerce	Recommendation	Movielens-20M	NCF	0.635 HR @ 10
Research	Go	n/a	Mini go	40.0% move prediction

MLPerf

Metric: time-to-train

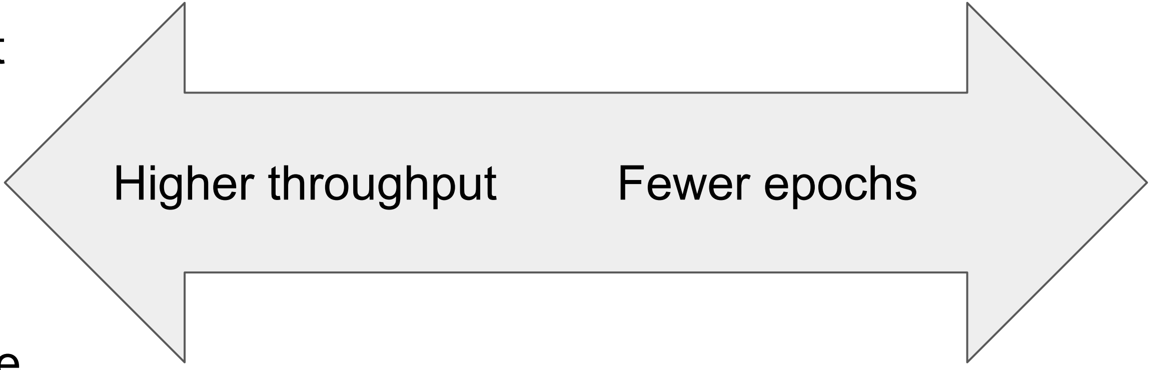
Alternative is throughput

- Easy / cheap to measure

But can increase throughput at cost of total time to train!

Time-to-train (end-to-end)

- Time to solution!
- Computationally expensive
- High variance
- **Least bad choice**



Lower precision
Higher batch size

Higher precision
Lower batch size

Time-to-train excludes

System initialization

Depends on cluster configuration and state

Model initialization

Disproportionate for big systems with small benchmarking datasets

Data reformatting

Mandating format would give advantage to some systems

MLPerf

Challenges and Contributions

ML Training benchmarking challenges

Diverse software stacks and hardware systems	● Can’t use the same executable ● Can’t use the same *code*

ML Training benchmarking challenges

Diverse software stacks and hardware systems	• E.g.: larger systems → larger SGD mini batches → different optimizer hyperparams • Hyperparameter tuning is computationally expensive, can be unfair
Different scales and/or numerics require tuning	

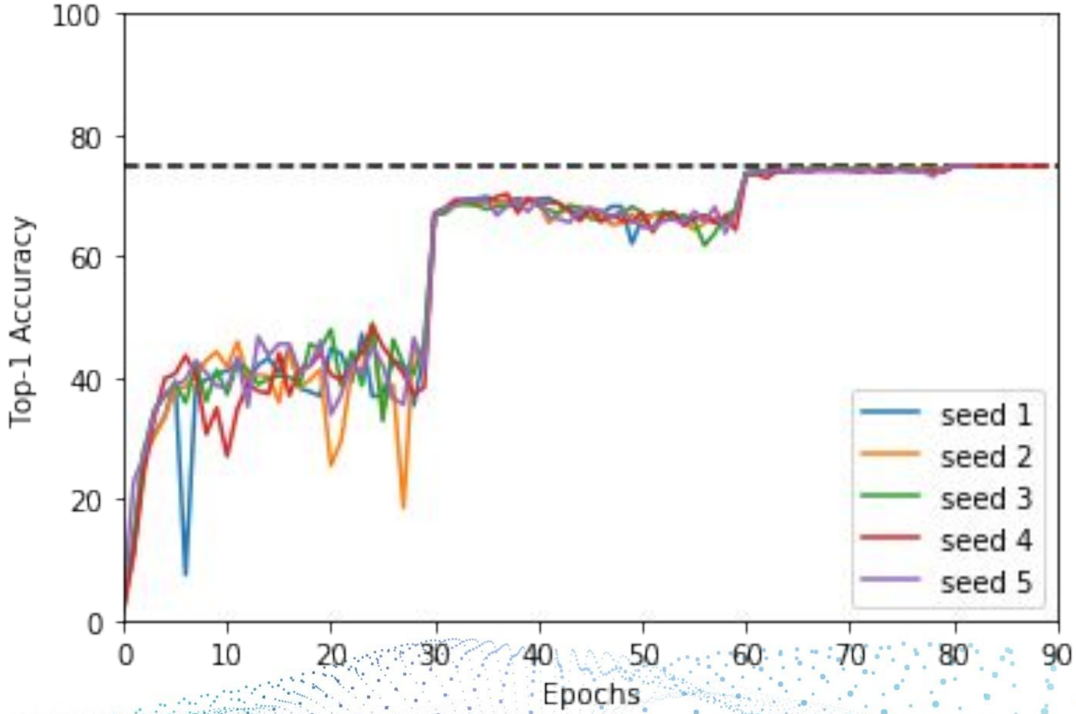
MLPerf

ML Training benchmarking challenges

Diverse software stacks and hardware systems	• Random weight initialization • Non-deterministic floating point effects
Different scales and/or numerics require tuning	
Convergence is stochastic	

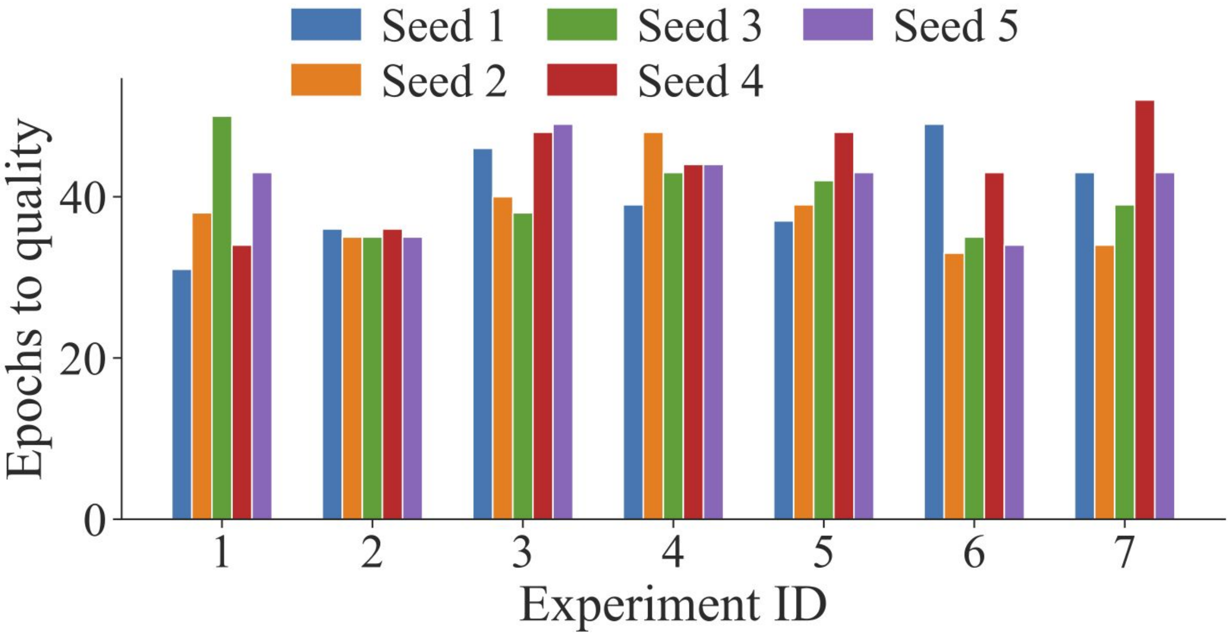
MLPerf

Convergence variance: ResNet



MLPerf

Convergence variance: MiniGo



MLPerf

MLPerf contributions

Diverse software stacks and hardware systems	**Reference implementations** **Rules for reimplementation**
Different scales and/or numerics require tuning	
Convergence is stochastic	

MLPerf

MLPerf contributions

Diverse software stacks and hardware systems	Reference implementations Rules for reimplementation
Different scales and/or numerics require tuning	**Limited tunable hyperparameters; limited values**
Convergence is stochastic	

List of tunable hyperparameters

Benchmark	Tunable hyperparameters
All that use SGD	Mini batch size, Learning-rate schedule parameters
ResNet-50 v1.5	--
SSD-ResNet-34	Maximum samples per training patch
Mask R-CNN	Number of image candidates
GNMT	Learning-rate decay function, Learning rate, Decay start, Decay interval, Warmup function, Warmup steps
Transformer	Optimizer: Adam or Lazy Adam, Learning rate, Warmup steps
NCF	Optimizer: Adam or Lazy Adam, Learning rate, β1, β2
MiniGo	--

MLPerf contributions

Diverse software stacks and hardware systems	Reference implementations Rules for reimplementation
Different scales and/or numerics require tuning	Limited tunable hyperparameters; limited values
Convergence is stochastic	**Require multiple runs** **Drop low and high, average**

LPerf

Submission Process

Pre-submit

Download **reference implementation**, read rules, join submitters working group

↓

Reimplement benchmark for system under test (SUT)

↓

Tune hyperparameters (allowed by list, to allowed values)

↓

Run benchmark required number of times

↓

Submit logs from all runs, code, metadata in Github by deadline

↓

MLPerf

Post-submit

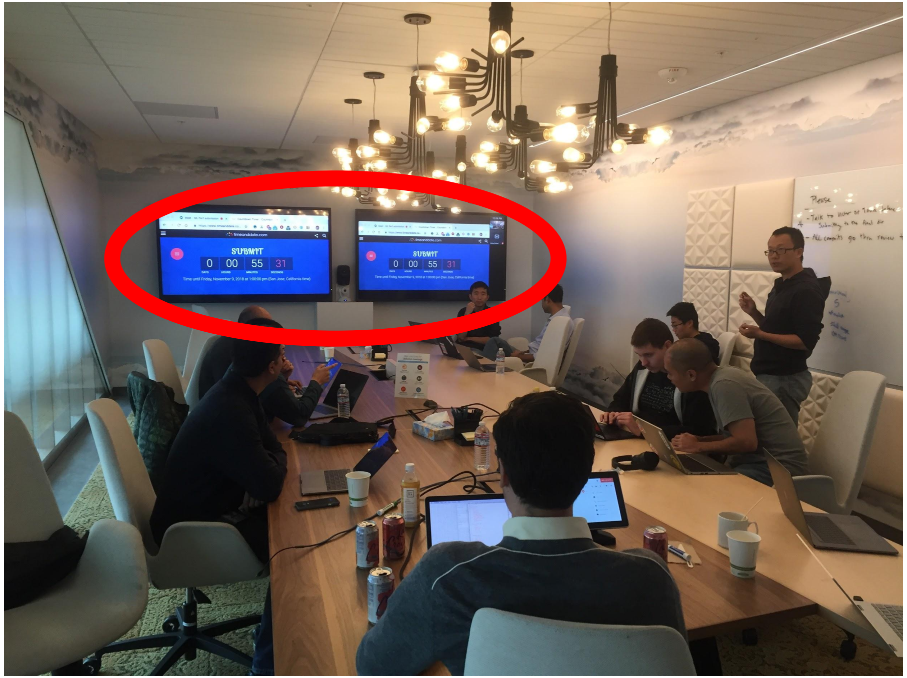
All submitters **peer review** all submissions, raise issues

Borrow hyperparameters from other submissions and resubmit if desired

MLPerf posts all results and makes logs, metadata, and code public under Apache-2

Celebrate!!!

MLPerf


SUBMIT
0 00 55 31
SUBMIT
0 00 55 31

Results and Lessons Learned

Impact of good benchmarks

Benchmarks

- Defined set of problems
- Clear metrics

Competition

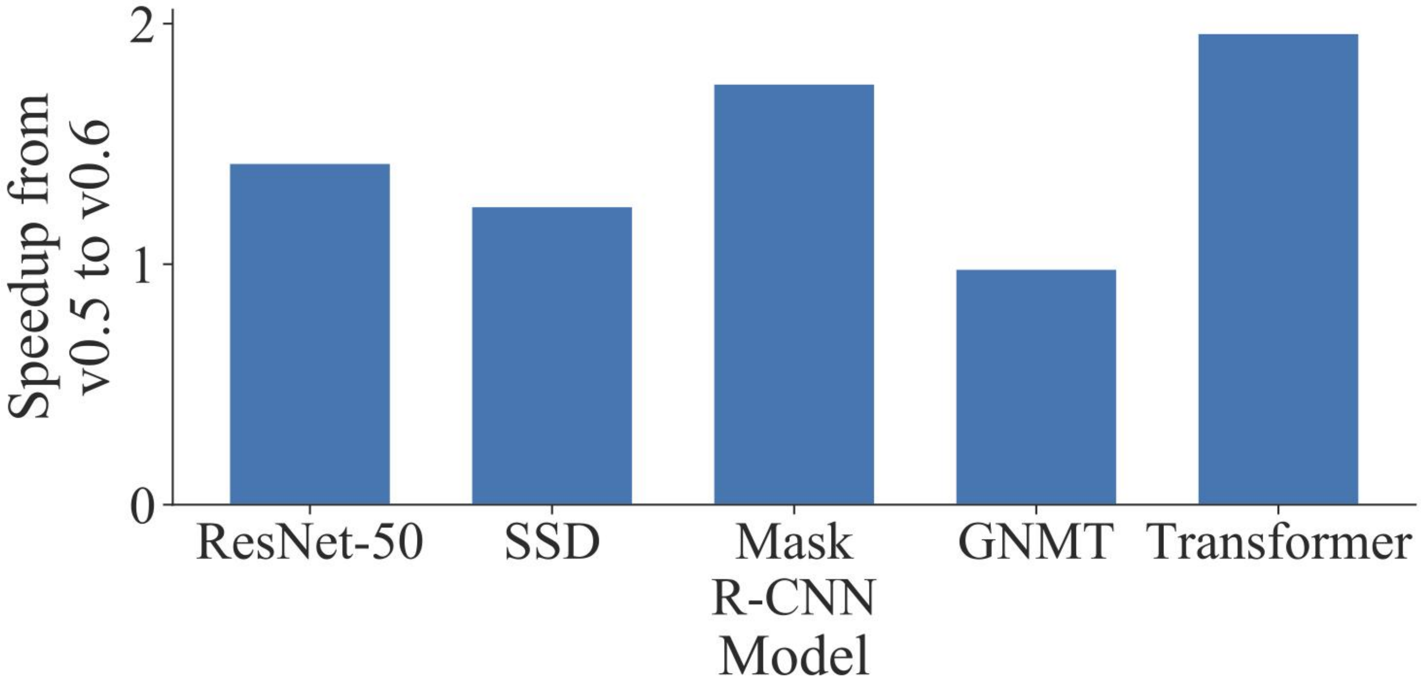
- Competing engineering teams try different approaches
- Results show what works best

Better Software / HW

- Improved understanding of performance
- Faster, more scalable software stacks
- Future hardware designs driven by best-of-breed ideas

MLPerf

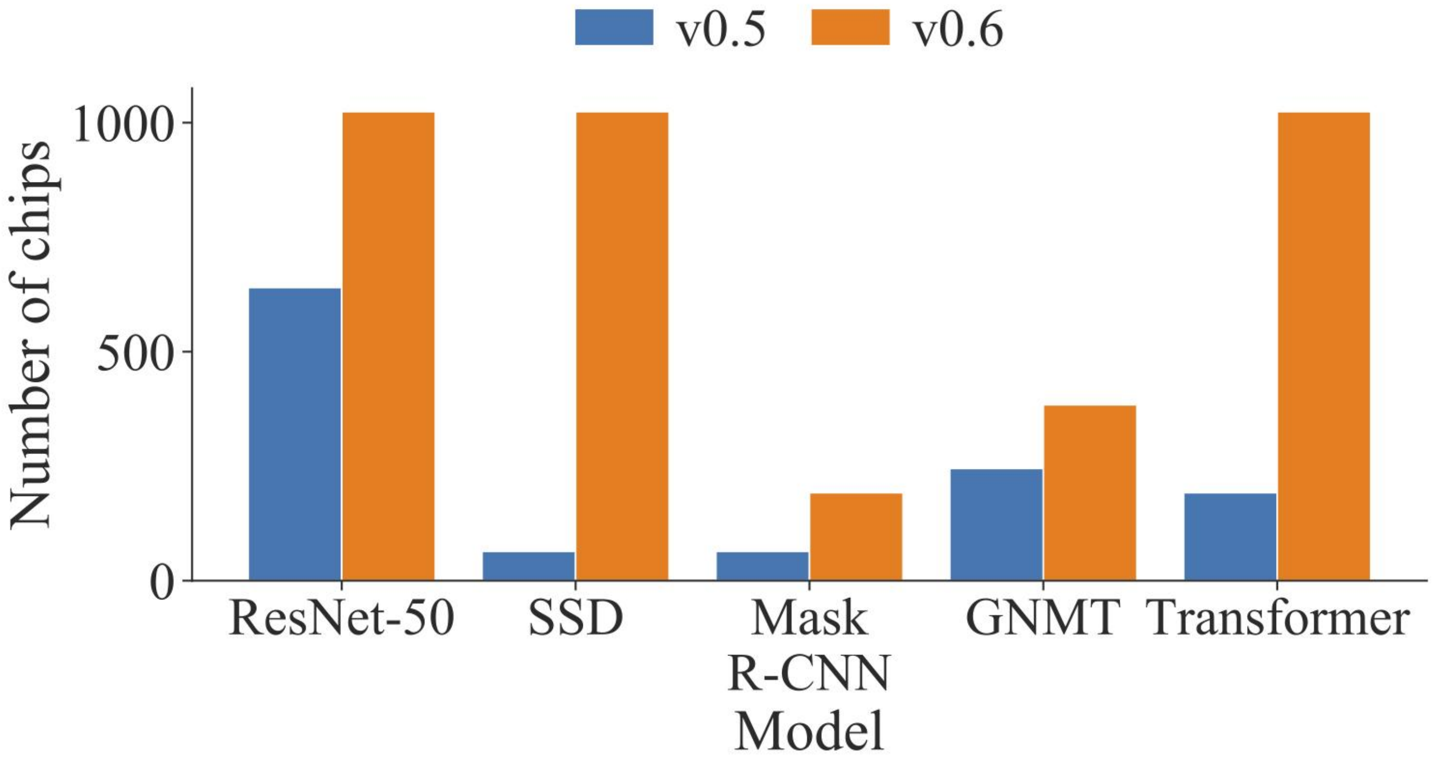
MLPerf Training: 16-chip speedup v0.5 to v0.6*



* Benchmark quality targets, and hence workload, increased in v0.6 for ResNet, SSD, GNMT

MLPerf

MLPerf Training, system scale increase v0.5 to v0.6



MLPerf

Lessons learned

- Benchmarking with reimplementation is possible
- Realistic dataset size is critical to ML performance benchmarking
- Hyperparameters are surprisingly portable at similar scales; borrowing works
- Low ratio of (standard deviation of epochs to train) : (mean epochs to train) is key to acceptable variance
- Variance in time to train increases at high batch sizes
- Frameworks had differences in optimizers that impact convergence

Support and Adoption

MLPerf Support: Companies

AI Labs.tw
Alibaba
AMD
Andes Technology
Aon Devices
Arm
Baidu

Cadence
Calypso AI
Centaur Technology
Cerebras
Ceva
Cirrus
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
Intel
In-Q-Tel

Lenovo
MediaTek
Mentor Graphics
Microsoft
Myrtle
Mythic
NetApp

NVIDIA
One Convergence
Oppo
PathPartner Technology
Pure Storage
Qualcomm
Rpa2ai

Sambanova
Samsung S.LSI
Sigopt
SiMa AI
Skymizer
Supermicro
Synopsys

Tencent
Tensyr
Teradyne
Transpire Ventures
VerifAI
VMind
VMware

Volley
Wave Computing
Wiwynn
WekaIO
Xilinx

MLPerf Support: Researchers

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 Adoption: Press

The Curious Case Of MLPerf Inferencing Benchmark Results
Forbes • Last month

MLPerf Releases First Inference Benchmark Results; Nvidia Touts its Showing
HPCwire • Last month

Centaur Releases In-Depth Analysis from The Linley Group for World's First x86 Processor with AI Coprocessor Technology
StreetInsider.com • 2 days ago

MLPerf Expands Toolset; Launches Inferencing Suite
HPCwire • Jun 24

Is Intel Considering Another AI Acquisition?
EE Times • 6 days ago

Benchmark Scores Reveal Who's Winning the AI Inference Race - EETimes
EE Times • Last month

Google, Nvidia tout advances in AI training with MLPerf benchmark results
ZDNet • Jul 10

MLPerf – Will New Machine Learning Benchmark Help Propel AI Forward?
HPCwire

NVIDIA Turing GPUs and NVIDIA Xavier Achieve Fastest Results on MLPerf Benchmarks Measuring Data Center and Edge AI Inference Performance
EE Journal • Last month

myrtle.ai to Develop a Speech Recognition Benchmark for MLPerf
HPCwire

It Is About Latency
HPCwire • 9 days ago

Reading Between the MLPerf Lines
The Next Platform

NVIDIA Gets Tiny With Jetson Xavier NX
Forbes • Last month

Nvidia Crushes Self to Take AI Benchmark Crown
ExtremeTech

NVIDIA Xavier wins critical AI performance benchmarks
Automotive World • Last month

Why Are Baidu, Google, Harvard And Stanford Collaborating For This ML Benchmark?
Analytics India Magazine • Jul 15

AI Accelerators: TOPS is Not the Whole Story - EETimes
EE Times • 2 days ago

Intel unveils next-gen Movidius VPU, codenamed Keem Bay
ZDNet • Last month

Centaur Unveils an x86 SoC with Integrated AI Coprocessor
CNX Software • Last month

The MLPerf Consortium, with Members like ARM & Google, have introduced Tech Industry's First Standard ML Benchmark Suit
Patently Apple • Jun 26

MLPerf benchmark results showcase Nvidia's top AI training times
ZDNet

Google Cloud and Nvidia Tesla set new AI training records with MLPerf benchmark results
Packt Hub • Jul 15

Who's Winning the AI Inference Race?
Eetasia.com • Last month

AI Gets Inference Benchmarks
EE Times • Jun 24

Intel, GraphCore And Groq: Let The AI Cambrian Explosion Begin
Forbes • Last month

Centaur announces new SoC featuring an 8-core server-class x86 CPU with AVX512 support and an integrated 20 TOPS AI co-processor
Notebookcheck.net • Last month

MLPerf Releases Five Benchmarks
EE Times India • Jun 26

NVIDIA Corp (NVDA) Q3 2019 Earnings Call Transcript
The Motley Fool • Last month

Twitter wants help with deepfakes, and Microsoft Azure will rent out new AI chips for its cloud users, and more
The Register • Last month

Embedded Benchmark Calls for Support
EE Times • Jun 12

Startup Runs AI in Novel SRAM
EE Times • Jul 22

MLPerf Releases v0.6 Training Results
HPCwire • Jul 10

MLPerf To Provide Much Needed Clarity In The Field Of Machine Learning
Forbes • Jun 25

Digging into MLPerf Benchmark Suite to Inform AI Infrastructure Decisions
HPCwire • Apr 9

MLPerf Is Changing the AI Hardware Performance Conversation. Here's how
Data Center Knowledge • Aug 1

GPUs Continue to Dominate the AI Accelerator Market for Now
InformationWeek • Last month

Nvidia tops AI inference benchmarks, also announces 'world's smallest supercomputer' chip for AI tasks
Firstpost • Last month

Why I joined MLPerf
EE Times • Mar 20

MLPerf

Work in Progress / Future Work

Future work

Expand and update benchmark suite

Improve rules: hyperparameter tables, out-of-the box division

More efficiency information: power, cloud cost

New suites:
- Inference (launched in 2019)
- Mobile (launching in 2020)
- HPC (in progress)
- TinyML (in progress)

The next frontier: accuracy?

MLPerf

Shameless Plugs

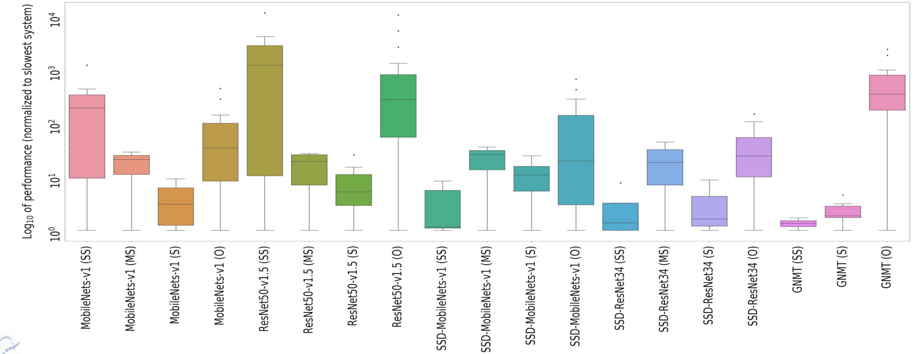
“Benchmarking Machine Learning Workloads” Workshop Tomorrow

Keynote:

MLPerf Inference

Vijay Janapa Reddi, Harvard

9:10 AM



MLPerf

MLPerf Training: Open Division needs you!

Want to Showcase faster models, compilers, pruners, data-set optimizers

Only need to use Dataset and Target to submit

Low overhead, low-risk exposure

Some assembly required

Plan for impact

Think big: conceive of your work as 10% of a larger whole

Great idea + coalition >> great idea alone

- Build different skills sets

Make the world better

MLPerf

Summary

Summary

Introduced MLPerf training

Broad suite of tasks + time-to-train metric + consortium

Solved ML benchmarking challenges: diverse systems, scaling, variance

Results show MLPerf helps drive performance improvements

Achieved broad support/adoption: industry, academia, press

More to do! Join us: **mlperf.org/get-involved** or **info@mlperf.org**.