RETROSPECTIVE: MLPerf

Vijay Janapa Reddi[†] Carole-Jean Wu[⋆] David Kanter[‡] Peter Mattson[§]
[†]Harvard University [⋆]Meta [‡]MLCommons [§]Google

"On behalf of the MLPerf community."

I. The Dawn of MLPerf

The MLPerf benchmark had its inception in the interplay between academia and industry, as a collaborative venture with a singular objective: to accelerate progress in machine learning (ML). The concept of benchmarking is not new to our community; it had been a staple in performance measurement for decades. For instance, the SPEC CPU benchmark has been the bedrock for unleashing the golden age of general-purpose CPUs. But as ML began to dramatically alter the technological landscape in the late 2010s and early 2020s, the need for a ML-specific benchmark became strongly apparent.

The rapidly growing interest in AI and machine learning led to what many dubbed the "Cambrian explosion" of AI hardware startups. This growth, however, also revealed a significant issue. With so many different hardware architectures and machine learning models emerging, it was increasingly difficult to objectively compare their performance. The explosion of AI hardware innovation was, at least in some ways, being hindered by the lack of a universal ML benchmark.

In the years leading up to the creation of MLPerf in 2018, multiple efforts were being made to establish benchmarks for ML systems. For instance, DawnBench, a project from researchers at Stanford University, was an initiative to benchmark the training and inference speed of deep learning models. Harvard's Fathom benchmark aimed to reflect different areas of machine learning that are important to the commercial and research communities and have open datasets. DeepBench from Baidu was focused on developing a benchmark to assess the performance of basic ML operations that are important to deep learning on various hardware platforms. Despite these efforts, there was no industry-wide, agreed-upon standard.

Recognizing this gap, several key stakeholders from academia and industry started MLPerf in 2018. Its mission was simple yet profound: to build fair and useful benchmarks for the burgeoning field of machine learning. It aimed to provide metrics that would enable the scientific community and industry to measure and compare ML systems' performance, thereby fostering innovation and technological advancements.

MLPerf benchmark suite initially encompassed categories like image classification, object detection, speech-to-text, and machine translation, which were further expanded over time to include more tasks for both training and inference. The benchmarking process involved not only the raw speed of training a model but also other aspects like the quality of the model's results, its scalability, and the power consumed. These processes were specialized for ML training and ML inference.

MLPerf is a collaborative effort involving many individuals and organizations. Please see the papers for the full list of authors and acknowledgements.

II. MLPerf Training

Benchmarking the ML training process is important for several reasons. Training helps in the fine-tuning of the model parameters to minimize the discrepancy between predicted and actual output. This 'learning' stage allows models to generalize from given examples and make accurate predictions or decisions about unseen data. Given the resource-intensive nature of ML training, it is important to benchmark the training process to provide a standardized measure of performance.

MLPerf training benchmarks provide valuable insights into the efficiency, scalability, and cost-effectiveness of different ML training hardware and software. It aids in tracking the progress of ML technologies over time and provides an objective comparison of different systems and configurations. Without a training benchmark suite, it would be almost impossible to measure and compare the efficiency of various ML training methods, creating a barrier for future advancements.

When benchmarking ML training, a few critical factors come into play. The time taken to train a model, measured in terms of time-to-accuracy is a key performance indicator, as first indicated by DAWNBench. Other crucial considerations include the computational resources used, energy consumption, and the scalability of the system when subjected to larger datasets and complex tasks. Another important factor is the reproducibility of results. The training benchmark results should be reliable, repeatable, and obtained under fair, unbiased conditions to ensure an accurate comparison.

The advent of MLPerf's ML training benchmark has driven much innovation in the world of ML systems. By providing an industry-standard benchmark for ML training, MLPerf has made significant contributions in streamlining and accelerating progress in ML technologies. One of the grand results of the MLPerf benchmark is the development of more efficient hardware and software for ML training. With MLPerf, companies have been able to objectively test and optimize their products, leading to advancements in GPUs, CPUs, TPUs, and ML frameworks. The results have also driven advancements in ML algorithms. The quest to improve benchmark scores has led to the development of innovative training methods that reduce the time-to-accuracy or increase the scalability of training.

The MLPerf training benchmarks have also provided valuable insights into the interplay between different parts of an ML system—hardware, software, and algorithms. They have helped in highlighting bottlenecks and areas of improvement,

providing a clearer path for future research and development. For instance, training system performance trajectory has increased by up to 30x since the initial release of the benchmark in 2018. MLPerf Training 2.1, which is the seventh iteration of the training-focused benchmark suite, has expanded to encompass several hundred submissions from different submitters.

III. MLPerf Inference

The landscape of ML inference is vastly different from training and it is intricate. Different devices, from datacenter servers to edge devices such as smart home gadgets and mobile phones, to even smaller IoT devices, execute inference tasks. Each has its own distinct performance, power, latency, and throughput constraints. As such, benchmarking ML inference is no small task; it necessitates a comprehensive understanding of the myriad environments where inference is executed.

To address this diversity of tasks and metrics, the MLPerf community aimed at creating a suite of fair, reliable, and practically useful benchmarks for inference software and hardware. MLPerf's ML inference benchmark provides a standardized method to evaluate ML inference performance across this diverse landscape. It focuses on the core aspects of ML inference systems without significantly sacrificing model quality.

Over the years, the MLPerf inference benchmark suite has been instrumental in catalyzing improvements in ML inference hardware and software. It has helped uncover bottlenecks, drive improvements in software frameworks, and inform hardware design, ultimately leading to the creation of more efficient, effective inference systems. The benchmark has been an invaluable resource for the industry to identify which systems offer the best performance per dollar or the highest throughput while adhering to power and latency constraints.

Some grand results have emerged from this process. For example, advances in the optimization of inference processing on GPUs have resulted in significant improvements in latency times for complex tasks, thereby broadening the potential applications of ML models. In addition, we have seen the evolution of specialized ML hardware, new generations of TPUs, GPUs, ASICs, as well as FPGAs, which is evident from the large number of benchmark submitters.

Furthermore, there has been a progressive increase in the efficiency of inference on edge and IoT devices. This has unlocked the potential for advanced ML applications in power and compute-constrained environments. For instance, real-time anomaly detection in industrial IoT is an important use case for MLPerf Tiny. On the other hand, MLPerf Mobile has focused on advanced AI in smartphones, handhelds and laptops.

The MLPerf inference benchmark suite is currently in version 3.0. In this version alone, there are over 6,700 performance results. The general inference trends we observe include a significant number of new hardware systems. The performance in the datacenter has grown by over 30% in some benchmarks since MLPerf Inference v2.1. There is also a growing emphasis on power efficiency, with improvement by over 50%. The new "inference over the network" category has experienced a 3x increase in submissions. Additionally, a wide array of different techniques such as distillation and sparsification, along with new models, are being applied in the submitted results. Thus, the MLPerf inference benchmark suite has, and continues to, shape the ML landscape.

IV. The Future of MLPerf

Originally focused on benchmarking machine learning models, MLPerf is constantly undergoing significant expansion, embracing new application domain and evolving to meet the challenges of an ever-changing technological landscape.

One such pivotal moment in the recent history of MLPerf is recognizing the growing importance of specialized domains, MLPerf is opening its doors to new areas such as autonomous vehicles, medical AI, and AI for science. This expansion is bringing together new and diverse communities, fostering collaboration and driving innovation across various sectors.

Moreover, as AI compute kernels made significant strides in speed and efficiency, the role of storage and network technologies became increasingly vital in achieving high-performance and cost-effective machine learning pipelines. The coordination between data storage, ingestion, and training accelerators became crucial to ensure continued performance scaling during training. Recognizing this critical aspect, MLPerf introduced the Storage benchmark in 2022. This benchmark aims to facilitate fair comparisons among various storage technologies within the context of a realistic training pipeline, addressing an essential component that had previously been overlooked.

Another significant stride that MLPerf is taking is recognizing that models are only as effective as the data they are trained on. With this realization, MLCommons, the organization that supports MLPerf, is expanding the scope to encompass benchmarking data itself. This new dimension of assessment addresses the crucial aspect of data quality, diversity, and legal distribution. By incorporating data benchmarking into its repertoire, practitioners are encouraged to pay closer attention to the complete pipeline of machine learning.

As MLPerf continues to grow, it is of paramount importance that its development and deployment align with principles of responsibility and environmental accountability. To this end, MLPerf has embarked on the standardization of power measurement and submission guidelines. This transition to a more accurate yardstick promises a comprehensive representation of computing infrastructure utilization, which in turn offers more equitable energy efficiency comparisons across ML training and inference systems. In just a span of 2 years, MLPerf has accumulated over 5,000 power submission results from a wide range of experiments conducted across 1,400 systems. Given the rapidly scaling demand for ML, any efficiency gains translate directly into broader use, consequently driving a significantly larger energy and environmental footprint. The hope is that ML – a more efficient alternative – is displacing the environmental footprint of prior solutions.

Benchmarking is an ongoing and continuous process. Therefore, MLPerf has no foreseeable end in its evolution as long as the field of machine learning continues to evolve and expand.