NEWS RELEASE

MLCommons Releases New MLPerf Results that Highlight Growing Importance of Generative AI and Storage

9/11/2023

Latest benchmarks include LLM in inference and the first results for storage benchmark

SAN FRANCISCO--(BUSINESS WIRE)-- Today, MLCommons, the leading open AI engineering consortium, announced new results from two MLPerf™ benchmark suites: MLPerf Inference v3.1, which delivers industry standard Machine Learning (ML) system performance benchmarking in an architecture-neutral, representative, and reproducible manner, and MLPerf Storage v0.5. This publication marks the first ever release of results from the MLPerf Storage benchmark, which measures the performance of storage systems in the context of ML training workloads.

MLPerf Inference v3.1 Introduces New LLM and Recommendation Benchmarks

MLPerf Inference v3.1 includes record participation, with over 13,500 performance results and up to 40% performance gains, from 26 different submitters, and over 2000 power results. Submitters include: ASUSTeK, Azure, cTuning, Connect Tech, Dell, Fujitsu, Giga Computing, Google, H3C, HPE, IEI, Intel, Intel-Habana-Labs, Krai, Lenovo, Moffett, Neural Magic, NVIDIA, Nutanix, Oracle, Qualcomm, Quanta Cloud Technology, SiMA, Supermicro, TTA, and xFusion. In particular, MLCommons® would like to congratulate first time MLPerf Inference submitters Connect Tech, Nutanix, Oracle, and TTA.

“Submitting to MLPerf is not trivial. It’s a significant accomplishment, as this is not a simple point and click benchmark. It requires real engineering work and is a testament to our submitters' commitment to AI, to their customers, and to ML,” said David Kanter, Executive Director of MLCommons.

The MLCommons MLPerf Inference benchmark suite measures how fast systems can run models in a variety of deployment scenarios. ML inference is behind everything from the latest generative AI chatbots to safety features in vehicles, such as automatic lane-keeping and speech-to-text interfaces. Improving performance and power efficiency is key to deploying more capable AI systems that benefit society.

MLPerf Inference v3.1 introduces two new benchmarks to the suite. The first is a large language model (LLM) using the GPT-J reference model to summarize CNN news articles, which garnered results from 15 different submitters, reflecting the rapid adoption of generative AI. The second is an updated recommender, modified to be more representative of industry practices, using the DLRM-DCNv2 reference model and a much larger datasets, with 9 submissions. These new tests help advance AI by ensuring that industry-standard benchmarks represent the latest trends in AI adoption to help guide customers, vendors, and researchers.

“The submissions for MLPerf inference v3.1 are indicative of a wide range of accelerators being developed to serve ML workloads. The current benchmark suite has broad coverage among ML domains and the most recent addition of GPT-J is a welcome addition to the generative AI space. The results should be very helpful to users when selecting the best accelerators for their respective domains,” said Mitchelle Rasquinha, MLPerf Inference Working Group co-chair.

MLPerf Inference benchmarks primarily focus on datacenter and edge systems. The v3.1 submissions showcase several different processors and accelerators across use cases in computer vision, recommender systems, and language processing. There are both open and closed submissions related to performance, power, and networking categories. Closed submissions use the same reference model to ensure a level playing field across systems, while participants in the open division are permitted to submit a variety of models.

First Results for New MLPerf Storage Benchmark

The MLPerf Storage Benchmark Suite is the first open-source AI/ML benchmark suite that measures the performance of storage for ML training workloads. The benchmark was created through a collaboration spanning more than a dozen leading industry and academic organizations and includes a variety of storage setups including: parallel file systems, local storage, and software defined storage. The MLPerf Storage Benchmark will be an effective tool for purchasing, configuring, and optimizing storage for machine learning applications, as well as for designing next-generation systems and technologies.

Training neural networks is both a compute and data-intensive workload that demands high-performance storage to sustain good overall system performance and availability. For many customers developing the next generation of ML models, it is a challenge to find the right balance between storage and compute resources while making sure that both are efficiently utilized. MLPerf Storage helps overcome this problem by accurately modeling the I/O

patterns posed by ML workloads, providing the flexibility to mix and match different storage systems with different accelerator types.

“Our first benchmark has over 28 performance results from five companies which is a great start given this is the first submission round,” explains Oana Balmau, Storage Working Group co-chair. “We’d like to congratulate MLPerf Storage submitters: Argonne National Laboratory (ANL), DDN, Micron, Nutanix, WEKA for their outstanding results and accomplishments.”

The MLPerf Storage benchmark suite is built on the codebase of **DLIO**, a benchmark designed for I/O measurement in high performance computing, adapted to meet current storage needs.

View the Results

To view the results for MLPerf Inference v3.1 and MLPerf Storage v0.5, and to find additional information about the benchmarks please visit:

https://mlcommons.org/en/storage-results-05/
https://mlcommons.org/en/inference-edge-31/
https://mlcommons.org/en/inference-datacenter-31/

About MLCommons

MLCommons is an open engineering consortium with a mission to make machine learning better for everyone through benchmarks and data. The foundation for MLCommons began with the MLPerf benchmark in 2018, which rapidly scaled as a set of industry metrics to measure machine learning performance and promote transparency of machine learning techniques. In collaboration with its 50+ members - global technology providers, academics, and researchers, MLCommons is focused on collaborative engineering work that builds tools for the entire machine learning industry through benchmarks and metrics, public datasets, and best practices.

For additional information on MLCommons and details on becoming a Member or Affiliate, please visit **MLCommmons** or contact **participation@mlcommons.org**.

Press Contact:

Kelly Berschauer
kelly@mlcommons.org

Source: MLCommons