Marching on with MLPerf, Breaking Some Records While Attempting New Ones

Article

Lenovo AI continues to perform

MLPerf™ is a noble endeavor; without it, Artificial Intelligence (AI) performance would be lost in translation between expectation and reality. AI is often misunderstood, incorporating diverse workflows and architectures. Due to its complexity, it can be difficult for organizations to make informed decisions as to how systems should be configured and how best to invest in improving AI performance. MLPerf was born to address this need and Lenovo is an active participant in these efforts.

MLPerf's mission is to "build fair and useful benchmarks" that provide unbiased evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions. This is not about bragging rights as to who builds the fastest servers; it's about helping our customers make informed decisions about how to best configure the infrastructure they buy from us and showing where to invest in areas they will see the biggest returns.

To be transparent and help our customers make better-informed decisions, we publish results quarterly. This is important because it provides our customers with a better understanding of how the latest versions of technical components perform, and it sometimes involves us testing entirely new types of technology. In this case, our latest round of MLPerf Inference v2.1 involved testing new versions of technologies, as well as new classes of technology.

MLPerf Inference 2.1 highlights

From the highest level, here's what we attempted and discovered in this quarter's MLPerf testing:

- We won 3 benchmarks, and tied 1, using the same GPUs and infrastructure as the previous round of testing. This indicates that even existing AI infrastructure can perform better over time, ostensibly with firmware and software updates.
- We won more ResNet benchmarks than any other technology vendor, and while the ResNet family include some of the earliest MLPerf benchmarks, it is widely perceived as the industry standard, so this is a meaningful accomplishment.
- We made our first Qualcomm submission and will likely be carrying that torch forward.
- We made our first submission into the power category, and it's encouraging to see these benchmarks moving beyond sheer performance to incorporate capabilities organizations truly care about, like power efficiency.

Advancements in performance

This was our second run with some of the same servers, indicating the performance improvements we are seeing are due to software and/or driver improvements, proving they are obtainable for organizations that have already deployed these systems. The most image classification wins for the accelerator and number pair:

- ThinkSystem SR670 V2 with 8x NVIDIA A100-PCIe-80GB – ResNet Offline 318,162 images/sec
- ThinkSystem SR670 V2 with 8x NVIDIA A100-PCIe-80GB – RNNT Offline 107,881.00 samples/s
- ThinkSystem SR670 V2 with 4x NVIDIA A100-SXM-80GB – ResNet Server 150,027 images/s (tie with Dell)
- ThinkSystem SR670 V2 with 4x NVIDIA A100-SXM-80GB – ResNet Offline 174,180 images/s

Fundamentally, Lenovo is making continual progress with supporting more models. This momentum includes extending and enhancing our AI infrastructure portfolio so that customers can make more informed decisions from faster insights. Our goal through MLPerf is to bring clarity to infrastructure decisions so customers can focus on the success of their AI deployment overall.

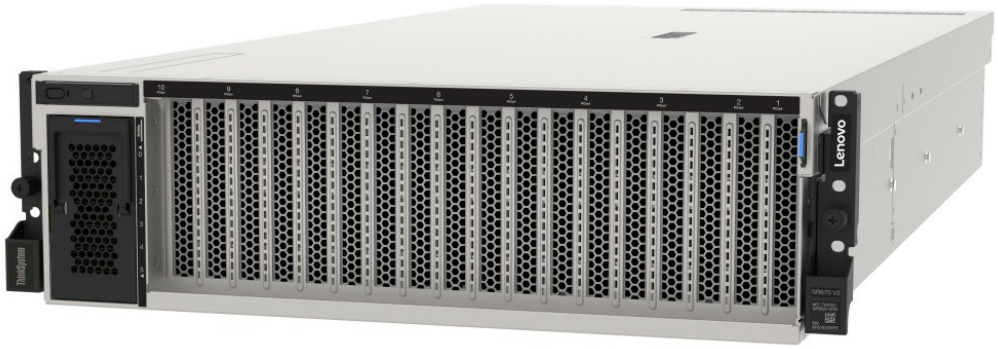

Figure 1. Lenovo ThinkSystem SR670 V2 configured to support eight NVIDIA A100 GPUs

Lenovo collaborates with NVIDIA

Lenovo demonstrated AI performance across various infrastructure configurations, including NVIDIA A16 GPUs, running on Lenovo ThinkSystem platforms We showcased the efficiency and performance of our air-cooled systems, providing both PCIe and HGX deployment options in a standard data center platform that enterprises of all sizes can quickly deploy.

Lenovo collaborates extensively with NVIDIA in the AI realm. Through our Lenovo AI Innovation Centers, we're working with NVIDIA to ensure the success of our mutual customers AI initiatives. This provides customers with access to Lenovo and NVIDIA AI experts to aid with consulting on projects, the proper infrastructure to run a proof of concept, and proof of ROI before deployment. As the AI world continues to evolve, collaborations make coming to market an easier and more effective process.

For more information

For more information, see the following resources:

Explore Lenovo AI solutions:
https://www.lenovo.com/us/en/servers-storage/solutions/analytics-ai/

Engage the Lenovo AI Center of Excellence:
https://lenovoaicodelab.atlassian.net/servicedesk/customer/portal/3

MLCommons®, the open engineering consortium and leading force behind MLPerf, has now released new results for MLPerf benchmark suites:

- Benchmark results: https://mlcommons.org/en/inference-datacenter-21/
- Latest news about MLCommons: https://mlcommons.org/en/news/mlperf-inference-v21/

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- ThinkSystem SR670 V2 Server

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1644, was created or updated on September 9, 2022.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1644
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1644.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkSystem®

Other company, product, or service names may be trademarks or service marks of others.