Unleashing the power of AI: MLPerf Benchmarking Outcomes Show a Clear Lead for Lenovo ThinkEdge Servers

Article

In the domain of Artificial Intelligence (AI) and machine learning (ML), the assessment of computational performance across diverse hardware settings is facilitated by benchmarks such as MLPerf. These benchmarks are instrumental in evaluating the efficacy of machine learning models under various operational conditions.

This technical discourse delves into the outcomes of the latest MLPerf benchmarks, particularly examining the performance across three distinct server configurations:

- Lenovo ThinkEdge SE455 V3 with 2x NVIDIA L40 GPUs
- Lenovo ThinkEdge SE450 with 2x NVIDIA L40 GPUs
- Lenovo ThinkEdge SE360 V2 with 1x NVIDIA L4 GPU

The implications of these results are pivotal for optimizing machine learning operations in both commercial and research settings and the latest benchmark results show a clear Lenovo leadership in performance handling ML tasks.

Server Configurations and Benchmarks

The MLPerf benchmarks are structured to test a range of server configurations across different operational modes:

- ThinkEdge SE455 V3 with 2x NVIDIA L40 GPUs: Tested in both Server and Offline modes, this configuration is designed for high-throughput and latency-sensitive applications.
- ThinkEdge SE450 with 2x NVIDIA L40 GPUs: Evaluated across Single-Stream, Multi-Stream, and Offline modes, this setup addresses diverse requirements from real-time processing to batch analytics.
- ThinkEdge SE360 V2 with 1x NVIDIA L4 Tensor Core GPU: This configuration, while intended for more mainstream AI applications, provides insights into performance efficiencies in constrained environments.

Analysis of Benchmark Results

An analysis of the recent MLPerf benchmark results shows some interesting results.

Lenovo ThinkEdge SE455 V3 with 2x NVIDIA L40 GPUs

ResNet50 and RetinaNet: These models demonstrated high performance in both Server and Offline scenarios. The throughput of ResNet50 in Server mode (59,982.60) suggests its suitability for real-time image classification systems in industrial quality control or surveillance applications. Similarly, RetinaNet's performance (949.03 in Server mode) aligns with use cases in real-time object detection for autonomous vehicles or advanced security systems.

3D-Unet (99.0% and 99.9%): The exclusive Offline completion time of six minutes, indicates this model's potential in non-real-time medical imaging tasks, such as tumor segmentation in volumetric scans, where batch processing is feasible.

RNN-T and BERT 99.9%: These models excel in both Server and Offline modes, ideal for real-time and batch natural language processing tasks. RNN-T's higher score in Offline mode (18,850.6) could be leveraged in automated translation services during large-scale events, while BERT's consistency (1,699.3 in Server mode) makes it suitable for context-aware customer support bots.

Lenovo ThinkEdge SE450 with 2x NVIDIA L40 GPUs

ResNet50: This model's performance across Single-Stream, Multi-Stream, and Offline modes underlines its adaptability. Its potential applications include real-time processing in wearable health monitors and batch processing in large-scale video analysis for content moderation.

3D-Unet (99.0% and 99.9%): Notable for its application in single-patient medical imaging, enabling rapid diagnostic assessments.

RNN-T and BERT 99.0%: Appropriate for deployment in edge computing devices where latency is critical, such as in-field language translation devices.

Lenovo ThinkEdge SE360 V2 with 1x NVIDIA L4

ResNet50 and RetinaNet: Despite the lower hardware capabilities, these models perform adequately, suggesting their utility in mobile or embedded systems where power efficiency and space constraints are paramount.

Lenovo Leads the Majority of the Tests

The benchmark results indicate a significant performance lead in 21 out of 27 tested categories, highlighting the strengths of the Lenovo ThinkEdge SE455 V3 and ThinkEdge SE450 configurations in handling advanced ML tasks efficiently. These results underscore the potential for tailored hardware selection based on specific application needs, enhancing both performance and cost-efficiency.

Conclusion

The insights from the latest MLPerf benchmarks are critical for stakeholders in the machine learning ecosystem, from system architects to application developers. They provide a quantitative foundation for hardware selection and optimization, crucial for deploying scalable and efficient ML systems. Future developments in hardware and software are anticipated to further influence these benchmarks, continuing the cycle of innovation and evaluation in the field of machine learning.

Professionals in the field are encouraged to consider these results in their future hardware procurement and system design strategies. For further discussion or consultation on leveraging these insights in specific use cases, engage with our expert team at aidiscover@lenovo.com.

For more information

For more information, see the following resources:

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

Engage the Lenovo AI Center of Excellence:
https://lenovo-ai-discover.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/training-normal-20/
- Latest news about MLCommons: https://mlcommons.org/news-blog

Author

Carlos Huescas is the Worldwide Product Manager for NVIDIA software at Lenovo. He specializes in High Performance Computing and AI solutions. He has more than 15 years of experience as an IT architect and in product management positions across several high-tech companies.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- Edge Servers
- MLPerf Benchmark

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, LP1969, was created or updated on June 6, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1969
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1969.

Trademarks