Enhanced Computational Performance with ThinkSystem Servers

Article

This round of MLPerf intends to analyze and showcase the performance capabilities of two Lenovo ThinkSystem servers, specifically the SR680a V3 with 8-GPU NVIDIA HGX H200 and SR675 V3 models with 8x NVIDIA H100 NVL (PCIe form factor).

Three computationally intensive tasks are employed to evaluate these systems:

- Image generation using Stable Diffusion v2
- LLM fine-tuning using Llama2 70B-LoRA
- Object detection through RetinaNet

Two world records:

- Public ID: 4.1-0039
  - System name: SR675 v3 with 8x H100 NVL
  - Benchmark: Retinanet
- Public ID: 4.1-0040
  - System name: SR680a v3 with 8x H200 SXM5
  - Benchmarks: Llama2 and Stable-Diffusion

Methodology

These benchmarks were conducted at MLCommons with the specified ThinkSystem servers. Key hardware configurations for each system include:

1. ThinkSystem SR680a V3: 8-GPU NVIDIA HGX H200 System
2. ThinkSystem SR675 V3: 8x NVIDIA H100 NVL (PCIe form factor)

The chosen tasks and corresponding runtimes are outlined as follows:

- Stable Diffusion v2 on SR680a V3 (Image Generation): Completed in 29.364 minutes (Public ID: 4.1-0040)
- LLama2 70B-LoRA fine-Tuning on SR680a V3 (LLM Fine-tuning): Completed in 23.062 minutes (Public ID: 4.1-0040)
- RetinaNet on SR675 V3 (Object Detection): Executed in 45.256 minutes (Public ID: 4.1-0039)

Results and Analysis

Based on the benchmark data, the SR680a V3 server consistently demonstrates best-in-class performance in both image generation and LLM fine-tuning tasks, showcasing its robust capabilities for computationally demanding workloads.

- In image generation with Stable Diffusion v2, the SR680a V3 managed to achieve results within a reasonable timeframe of 29.364 minutes. (Public ID: 4.1-0040)
- Similarly impressive was the performance shown during Large Language Model (LLM) fine-tuning with Llama2 70B-LoRA, which completed in just 23.062 minutes. (Public ID: 4.1-0040)

On the other hand, the SR675 V3 server exhibited extraordinary competence in object detection using RetinaNet, marking a notable milestone as the only system with an equivalent GPU configuration to obtain results within this extremely short timeframe (45.256 minutes). (Public ID: 4.1-0039)

Conclusion

This round of benchmarks presents a comprehensive evaluation of the performance capacity of ThinkSystem servers bolstered with NVIDIA accelerated computing, showcasing exceptional results across specified computational tasks. Notable accomplishments include:

- Exceptional image generation capabilities on SR680a V3.
- Efficient LLM fine-tuning on SR680a V3
- Lightning-fast object detection on SR675 V3 using RetinaNet

The experimental tests confirm the powerful performance of these ThinkSystem servers in handling complex computational workloads, thereby validating their capability to satisfy the computational demands of organizations today. The success of these architectures can only be expected as they continue to unlock deeper computing power and accelerate insights generation for future success.

The insights from the latest MLPerf benchmarks are critical for stakeholders in the generative AI and 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 AI/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.

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/benchmarks/inference-datacenter/
  - Latest news about MLCommons: https://mlcommons.org/news-blog

Author

David Ellison is the Chief Data Scientist for Lenovo ISG. Through Lenovo's US and European AI Discover Centers, he leads a team that uses cutting-edge AI techniques to deliver solutions for external customers while internally supporting the overall AI strategy for the World Wide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the US Postal Service. Previous to that, he received a PhD in Biomedical Engineering from Johns Hopkins University. He has numerous publications in top tier journals including two in the Proceedings of the National Academy of the Sciences.

Related product families

Product families related to this document are the following:

- MLPerf Benchmark
- ThinkSystem SR675 V3 Server
- ThinkSystem SR680a V3 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, LP2182, was created or updated on March 17, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2182
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2182.

Trademarks