Large Language Model Inference Performance with NVIDIA AI Enterprise in vSphere

Performance Study – October 2, 2023

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com

Table of Contents

Executive Summary

According to a NASA news release, the exoplanet K2-18b is the Goldilocks Zone that could support bodies of water—and life. It is 120 light-years away. Luckily, the Goldilocks Zone for running your large language models (LLMs) is not 120 light-years away.

Now you can run your LLMs in the Goldilocks Zone of VMware with NVIDIA AI Enterprise, get near bare metal performance for large language models, and have virtualization benefits from VMware vSphere with NVIDIA AI Enterprise.

This paper presents performance results for the VMware vSphere® virtualization platform with NVIDIA H100-80GB vGPUs using the large language model GPT-J with 6 billion parameters. We use the MLPerf Inference 3.1 benchmark for our performance testing. The results fall into the "Goldilocks Zone," which is the area of outstanding performance and increased security and manageability that the power of virtualization makes possible.

Our tests show that the GPT-J with 6 billion parameters inference performance tested with NVIDIA vGPUs in vSphere is 97.9% to 99.8% of the performance on the bare metal system, measured as queries served per second (qps) in the Server scenario and samples per second in the Offline scenario.

For a more comprehensive list of training and inference benchmark results, refer to:

- VMware vSphere 8 Performance Is in the "Goldilocks Zone" for AI/ML Training and Inference Workloads
- No Virtualization Tax for MLPerf Inference v3.0 Using NVIDIA Hopper and Ampere vGPUs and NVIDIA AI Software with vSphere 8.0.1

Introduction

This technical paper highlights the inference performance of LLM GPT-J with 6 billion parameters with NVIDIA AI Enterprise. Our previous performance numbers are published in the MLPerf Inference benchmark results for v0.7, v1.1, and v3.0.

Note: Our performance results for MLPerf v3.1 Inference were not verified by the MLCommons Association[1]. VMware is the only virtualization platform that has submitted virtualized MLPerf benchmark results. VMware, Dell, and NVIDIA achieved performance close to or higher than the corresponding bare metal configuration with the following setup:

- Dell PowerEdge R750xa rack server
- 2x virtualized NVIDIA H100 Tensor Core GPU PCIe cards with 80GB each of GPU memory

Only 32 of the total 128 logical CPU cores were needed for inference in both configurations, leaving the other 96 logical CPU cores in the data center accessible for other work.

VMware vSphere provides easy management and fast workload processing using NVIDIA vGPUs, device groups connected by flexible NVLinks, and VMware vSphere virtualization technologies to leverage AI/ML infrastructure for training, inferencing, and graphics. Virtualization lowers the total cost of ownership of an AI/ML infrastructure by requiring less hardware.

Our tests show results that are in the "Goldilocks Zone," where you get the best of both worlds: bare metal performance and ease of data center management with hardware cost savings.

[1] Our MLPerf™ v3.1 Inference Closed GPTJ-99 and GPTJ-99.9 results were not verified by the MLCommons Association. The MLPerf™ name and logo are trademarks of the MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use is strictly prohibited. See www.mlcommons.org for more information.

Hardware/Software

Table 1 shows the hardware configurations used to run the LLM workloads on the bare metal and virtualized systems. The most salient difference in the configurations is that the virtual setup used a virtualized NVIDIA H100 Tensor Core GPU, denoted by GRID H100-80c vGPU. Note that the H100-80c vGPU is a full profile for time-sliced mode. Both the systems had the same 2x NVIDIA H100-PCIE-80GB physical GPUs. The benchmarks were optimized with NVIDIA TensorRT-LLM. TensorRT-LLM consists of the TensorRT deep learning compiler and includes optimized kernels, pre- and post-processing steps, and multi-GPU/multi-node communication primitives for groundbreaking performance on NVIDIA GPUs.

	Bare Metal	Virtual Configuration
System	Dell PowerEdge R750xa	Dell PowerEdge R750xa
Processors	2x Intel Xeon Platinum 8358	2x Intel Xeon Platinum 8358
Logical Processor	128	32 allocated to the VM for inferencing (96 available for other VMs/workloads)
GPU	2x NVIDIA H100-PCIE-80GB	2x NVIDIA GRID H100-80c vGPU (full profile)
Memory	256GB	128GB (for inferencing VM)
Storage	3.2TB NVMe SSD	3.2TB NVMe SSD
OS	Ubuntu 22.04	Ubuntu 22.04 VM in vSphere 8.0.1
NVIDIA AI Enterprise VIB for ESXi	N/A	vGPU GRID Driver 535.54.06
CUDA	12.2	12.2 CUDA and Linux vGPU Driver 535.54.03
TensorRT	TensorRT-LLM	TensorRT-LLM
Special VM Settings	N/A	`pciPassthru0.cfg.enable_uvm = "1"`

Table 1. Bare metal vs. virtual server configurations for virtualized NVIDIA H100 Tensor Core GPU

GPTJ-6B LLM Inference Performance in VMware vSphere with NVIDIA AI Enterprise

VMware used GPT-J 6B from the MLPerf Inference v3.1 suite to test the data center performance shown in table 2 below. We focused on Offline and Server scenarios. The Offline scenario processes queries in a batch where all the input data is immediately available. Latency is not a critical metric in this scenario. In the Server scenario, the query arrival is random. Each query has an arrival rate determined by the Poisson distribution parameter. Each query has only one sample.

Area	Task	Model	Dataset	QSL Size	Quality	Scenarios	Server Latency Constraint
Language	Language processing	BERT-large	GPT-J 6B	13368	99.9% or 99% of the original FP32 ROUGE 1 - 42.9865 ROUGE 2 - 20.1235 ROUGE L - 29.9881	Server, Offline	N/A

Table 2. GPT-J 6B Inference benchmark used in our performance study

The quality is measured using Recall-Oriented Understudy of Gisting Evaluation (ROUGE). ROUGE is a way to compare the quality of a system-generated summary to one or more reference summaries. It is often used in natural language processing and text summarization tasks. It looks at how much the system-generated summary and the reference summaries match up and overlap.

The following three metrics were used.

- **ROUGE-1** refers to the overlap of *unigrams* (each word) between the system and reference summaries.
- **ROUGE-2** refers to the overlap of *bigrams* (pairs of consecutive words) between the system and reference summaries.
- **ROUGE-L** refers to the Longest Common Subsequence (LCS)[3] based statistics. This takes into account sentence-level structure similarity naturally and automatically identifies the longest co-occurring ones in sequence n-grams.[2]

[2] An n-gram is a contiguous sequence of n words or characters in a text or speech sample. It includes unigrams, bigrams, and longer sequences. N-grams are used in natural language processing and text analysis to identify patterns, relationships, and context.

Table 3 shows the accuracy achieved in our experiments. The GPT-99 and GPTJ-99.9 indicate normal and high-accuracy configurations.

	Required Minimum	Offline Accuracy	Server Accuracy	Offline Accuracy	Server Accuracy
		GPTJ-99	GPTJ-99	GPTJ-99.9	GPTJ-99.9
ROUGE-1	42.9865	43.076	43.0664	43.0756	43.0299
ROUGE-2	20.1235	20.195	20.1473	20.195	20.1399
ROUGE-L	29.9881	30.0363	30.0092	30.0363	30.001

Table 3. GPT-J 6B Inference benchmark accuracy achieved

Figures 2 and 3 compare the throughput samples per second for the Offline scenario and queries processed per second in the Server scenario of LLM GPT-J 6B benchmark using vSphere 8.0.1 with NVIDIA vGPU H100-80c against the bare metal H100 GPU configuration. The bare metal baseline is set to 1.000, and the virtualized result is presented as a relative percentage of the baseline. vSphere with NVIDIA vGPUs deliver near bare metal performance ranging from 98% to 100% for the Offline and Server scenarios when using the GPTJ 6B Inference benchmark, as shown in figures 2 and 3.

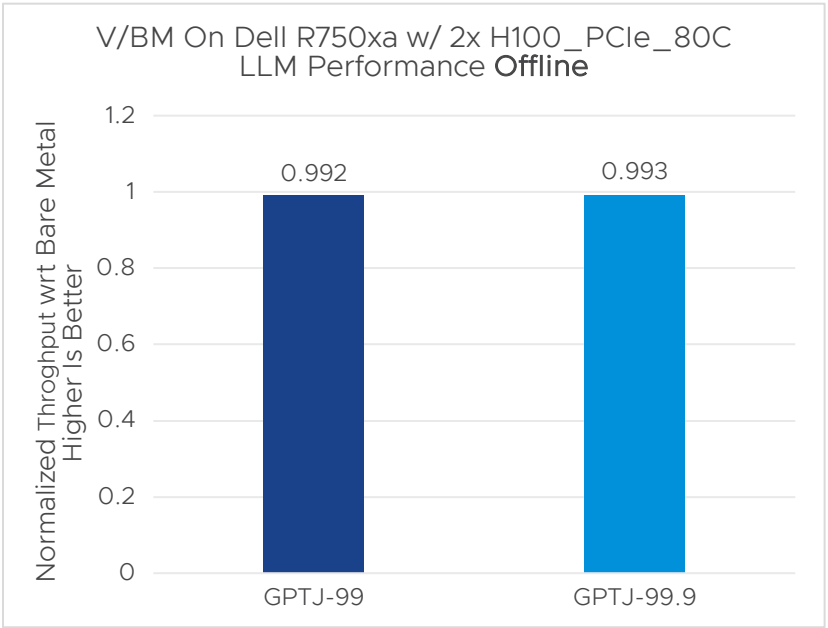

Figure 2. Normalized throughput for the **Offline** scenario (samples per second): vGPU 2x NVIDIA H100-80c vs bare metal 2x NVIDIA H100-80GB

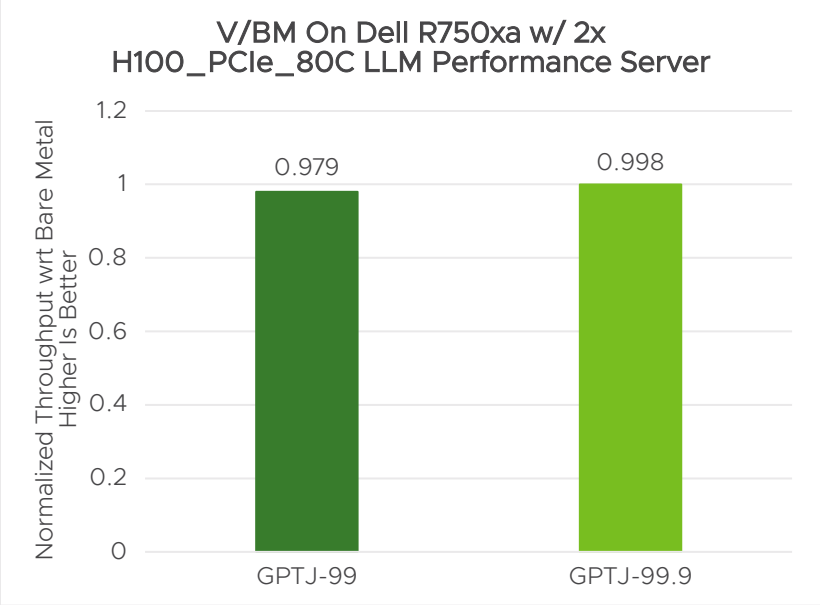

Figure 3. Normalized throughput for **Server** scenario (qps): vGPU 2x NVIDIA H100-80c vs bare metal 2x NVIDIA H100-80GB

Tables 4 and 5 show normalized throughput numbers in queries per second for MLPerf Inference benchmarks on 2x NVIDIA H100 for the Offline and Server scenarios.

Benchmark	vGPU/BM
GPTJ-99 Offline	0.992
GPTJ-99.9 Offline	0.993

Table 4. vGPU 2x H100-80c vs bare metal 2x H100 normalized throughput (samples per second) for **Offline** scenario

Benchmark	vGPU/BM
GPTJ-99 Server	0.979
GPTJ-99.9 Server	0.998

Table 5. vGPU 2x H100-80c vs bare metal 2x H100 normalized throughput (queries per second) for **Server** scenario

Takeaways

- NVIDIA AI Enterprise in VMware vSphere is in the Goldilocks Zone for AI/ML workloads, including LLMs. NVIDIA AI Enterprise in VMware vSphere has near bare metal performance and the virtualization benefits of VMware vSphere.
- NVIDIA AI Enterprise in VMware vSphere delivers from 97.9% to 99.8% of the bare metal performance measured as throughput for LLM GPT-J with 6 billion parameters.
- VMware achieved inference performance with only 32 logical CPU cores out of 128 available CPU cores, thus leaving 96 logical CPU cores for other jobs in the data center.
- For a comprehensive list of training and Inference benchmark results, refer to:
  - VMware vSphere 8 Performance Is in the "Goldilocks Zone" for AI/MLTraining and Inference Workloads
  - No Virtualization Tax for MLPerf Inference v3.0 Using NVIDIA Hopper and Ampere vGPUs and NVIDIA AI Software with vSphere 8.0.1

References

- Webb Discovers Methane, Carbon Dioxide in Atmosphere of K2-18
  https://www.nasa.gov/goddard/2023/webb-discovers-methane-carbon-dioxide-in-atmosphere-of-k2-18b
- MLCommons
  https://mlcommons.org/en/
- ROUGE (metric)
  https://en.wikipedia.org/wiki/ROUGE_(metric)
- NVIDIA TensorRT-LLM Supercharges Large Language Model Inference on NVIDIA H100 GPUs
  https://developer.nvidia.com/blog/nvidia-tensorrt-llm-supercharges-large-language-model-inference-on-nvidia-h100-gpus/
- VMware vSphere 8 Performance Is in the "Goldilocks Zone" for AI/MLTraining and Inference Workloads
  https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/vmware-ml-training-and-inference-perf.pdf
- No Virtualization Tax for MLPerf Inference v3.0 Using NVIDIA Hopper and Ampere vGPUs and NVIDIA AI Software with vSphere 8.0.1
  https://blogs.vmware.com/performance/2023/04/no-virtualization-tax-for-mlperf-inference-v3-0-using-nvidia-hopper-and-ampere-vgpus-and-nvidia-ai-software-with-vsphere-8-0-1.html
- ML Commons June 27, 2023 Training v3.0 results
  https://mlcommons.org/en/training-normal-30/
- vSphere 8 Expands Machine Learning Support: Device Groups for NVIDIA GPUs and NICs
  https://core.vmware.com/blog/vsphere-8-expands-machine-learning-support-device-groups-nvidia-gpus-and-nics
- What is NVLink? - NVIDIA Blog
  https://blogs.nvidia.com/blog/2023/03/06/what-is-nvidia-nvlink/
- MLCommons April 05, 2023 – Inference: Datacenter v3.0 Results
  https://mlcommons.org/en/inference-datacenter-30/
- MLCommons September 22, 2021 – Inference: Datacenter v1.1 Results
  https://mlcommons.org/en/inference-datacenter-11/
- NVIDIA Ampere Architecture
  https://www.nvidia.com/en-us/data-center/ampere-architecture/
- NVIDIA Hopper Architecture In-Depth
  https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth

- NVIDIA Ampere Architecture In-Depth
  https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth
- NVIDIA Docs Hub - NVIDIA AI Enterprise Latest Release (v3.2)
  https://docs.nvidia.com/ai-enterprise/latest/user-guide/index.html#supported-gpus-grid-vgpu
- MIG or vGPU Mode for NVIDIA Ampere GPU: Which One Should I Use? (Part 1 of 3)
  https://blogs.vmware.com/performance/2021/09/mig-or-vgpu-part1.html
- Introduction to MLPerf Inference v1.1 with Dell EMC Servers
  https://infohub.delltechnologies.com/p/introduction-to-mlperf-tm-inference-v1-1-with-dell-emc-servers
- MLPerf Inference Virtualization in VMware vSphere Using NVIDIA vGPUs
  https://blogs.vmware.com/performance/2020/12/mlperf-inference-virtualization-in-vmware-vsphere-using-nvidia-vgpus.html
- NVIDIA T4
  https://www.nvidia.com/en-us/data-center/tesla-t4/
- NVIDIA Triton
  https://developer.nvidia.com/triton-inference-server
- NVIDIA TensorRT
  https://developer.nvidia.com/tensorrt
- NVIDIA Turing GPU Architecture
  https://images.nvidia.com/aem-dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf
- V. J. Reddi et al., "MLPerf Inference Benchmark," 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 2020, pp. 446-459, doi: 10.1109/ISCA45697.2020.00045

Authors

Uday Kurkure works on accelerators for machine learning at VMware. He has a very broad skill set ranging from writing compilers, designing ASICS/FPGAs for computer graphics, and working on reconfigurable computing to virtualizing systems with VMware products. His current interests are machine learning (ML) and high performance computing (HPC). VMware awarded him the most prolific inventor award twice: he has 25 granted patents and 14 pending patent applications. He has published 13 research papers. His educational background includes an MS degree in Computer Science from Stanford and a B Tech in Electronics and Telecommunications from the Indian Institute of Technology. Before VMware, he worked for Synopsys, Adobe Systems, Transmeta, and MIPS Computers.

Lan Vu has worked at VMware for eight years focusing on GPU virtualization with vSphere and machine learning. Lan is interested in developing solutions that bring customers high performance and low cost too their IT infrastructure so they can configure their systems to best utilize cloud resources. Lan holds a PhD in Computer Science from the University of Colorado Denver and has 19 issued and pending patents. She is a winner of VMware's Most Prolific Inventor award.

Hari Sivaraman is a staff performance engineer at VMware. Hari has extensive experience in designing and running experiments to measure and predict the performance of computer systems, and he has built analytical and simulation models to estimate performance and answer "what-if" questions. He uses machine learning to solve problems in cloud management systems. Hari has written and collaborated on several papers and blog articles about AI/ML topics regarding vSphere performance. Hari has 25 patents and has been awarded VMware's Most Prolific Inventor award twice.

Acknowledgements

VMware thanks Liz Raymond and Yunfan Han of Dell; and Charlie Huang, Manvendar Rawat, and Jason Kyungho Lee of NVIDIA for providing the hardware and software for VMware's MLPerf Inference submission. The authors acknowledge Julie Brodeur of VMware for technical writing. The authors acknowledge Juan Garcia-Rovetta and Tony Lin of VMware for their management support.