

PERFORMANCE STUDY
June 2024

VMware vSphere 8 Performance Is in the “Goldilocks Zone” for AI/ML Training and Inference Workloads

Performance Study
June 11, 2024

Contents

Executive summary	3
Introduction	3
VMware and NVIDIA AI Enterprise.....	4
Hardware and software.....	4
AI/ML training performance in vSphere with NVIDIA vGPU.....	5
Flexible AI/ML infrastructure using device groups in vSphere with NVIDIA GPUs and NVLink	7
MLPerf Inference performance in vSphere with NVIDIA vGPU.....	8
MLPerf Inference performance results for bare metal and virtual configurations.....	9
Takeaways	14
References	15
Authors.....	17
Acknowledgments	17

Executive summary

In this paper, we give AI/ML training workload performance test results for the VMware vSphere virtualization platform using multiple NVIDIA A100-80GB GPUs with NVIDIA NVLink; the results fall into the “Goldilocks Zone,” which refers to the area of good performance with virtualization benefits.

Our results show that several virtualized MLPerf Training¹ v3.0 benchmarks’ training times perform within 1.06% to 1.08% of the same workloads run on a comparable bare metal system. Note that lower is better.

In addition, we show the MLPerf Inference v3.0 test results for the vSphere virtualization platform with NVIDIA H100 and A100 Tensor Core GPUs. Our tests show that when NVIDIA vGPUs are used in vSphere, the workload performance measured as queries served per second (qps) is 94% to 105% of the performance on the bare metal system. Note that higher is better.

Introduction

VMware vSphere provides easy management and fast workload processing using NVIDIA vGPUs, device groups connected by flexible NVLinks, and other vSphere virtualization technologies to leverage AI/ML infrastructure for training, inference, 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 cost savings.

This technical paper highlights the power of NVIDIA GPUs connected with NVLink in vSphere 8.0.1 used for AI/ML training and inference workloads. Our performance numbers are published in the MLPerf Inference benchmark results for v0.7, v1.1, and v3.0. We’re the only virtualization platform to do so.

Note: Our performance results for MLPerf Training v3.0 were not verified by the MLCommons Association.

VMware, Dell, and NVIDIA achieved performance close to or higher than the corresponding bare metal configuration with the following setups:

- Dell PowerEdge XE8545 server with 4x virtualized NVIDIA SXM A100-80GB GPUs
- Dell PowerEdge R750xa with 2x virtualized NVIDIA H100-PCIE-80GB GPUs

Only 16 of the total 128 logical CPU cores were needed for inference in both configurations, leaving the other 112 logical CPU cores in the data center accessible for other work.

Achieving the best performance for the VMs during training requires 88 logical CPU cores out of 128. The remaining 40 logical CPU cores in the data center can be used for other activities.

¹ Unverified MLPerf® Training v3.0 closed BERT-large and RNNT. Results were not verified by MLCommons. The MLPerf® name and logo are registered and unregistered 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.

VMware and NVIDIA AI Enterprise

The partnership between VMware and NVIDIA brings virtualized GPUs to vSphere with NVIDIA AI Enterprise. This lets you not only achieve quicker processing times for virtualized machine learning and artificial intelligence workloads—it also lets you leverage the many benefits of vSphere, such as cloning, vMotion, distributed resource scheduling, and suspending and resuming VMs.

Hardware and software

Table 1 shows the hardware configurations used to run the AI/ML workloads on bare metal and virtualized systems. The most salient difference in the configurations is that the virtual setup used a virtualized NVIDIA H100 GPU, denoted by GRID H100-80c vGPU. Note that the H100-80c vGPU profile is for time-sliced mode. Both systems had the same 2x H100-PCIE-80GB physical GPUs. The benchmarks were optimized with NVIDIA TensorRT.

Table 1. Bare metal vs. virtual server configurations for virtualized H100

	Bare metal	Virtual
System	Dell PowerEdge R750xa	Dell PowerEdge R750xa
Processors	2x Intel Xeon Platinum 8358	2x Intel Xeon Platinum 8358
Logical processors	128	<ul style="list-style-type: none"> • 16 allocated to the VM for inference (112 available for other VMs/workloads) • 88 allocated to the VM for training (40 available for other VMs/workloads)
GPUs	2x NVIDIA H100-PCIE-80GB	2x NVIDIA GRID H100-80c vGPU
Memory	256GB	128GB (for inference VM)
Storage	3.0TB NVMe SSD	3.0TB NVMe SSD
OS	Ubuntu 20.04	Ubuntu 20.04 VM in vSphere 8.0.1
NVIDIA AIE VIB for ESXi	–	vGPU GRID Driver 525.85.07
CUDA	12	12
TensorRT	8.6.0	8.6.0
MLPerf Inference	v3.0	v3.0

Table 2 describes the hardware configurations used for bare metal and virtual runs for the virtualized NVIDIA A100. The most salient difference in the configurations is that the virtual configuration used a virtualized A100 GPU, denoted by GRID A100-80c vGPU. Note that the A100-80c vGPU profile is for time-sliced mode. Both systems had the same 4x A100-SXM-80GB physical GPUs. The benchmarks were optimized with NVIDIA TensorRT.

Table 2. Bare metal vs. virtual server configurations for virtualized A100

	Bare metal	Virtual
System	Dell PowerEdge XE8545	Dell PowerEdge XE8545
Processors	2x AMD EPYC 7543	2x AMD EPYC 7543
Logical processors	128	<ul style="list-style-type: none"> • 16 allocated to the VM for inference (112 available for other VMs/workloads) • 88 allocated to the VM for training (40 available for other VMs/workloads)
GPUs	4x NVIDIA A100-SXM-80GB	4x NVIDIA GRID A100-80c vGPU
Memory	1TB	128GB (for inference VM) 900GB (for training VM)
Storage	3TB NVME SSD	3TB NVME SSD
OS	Ubuntu 20.04	Ubuntu 20.04 VM in vSphere 8.0.1
NVIDIA AIE VIB for ESXi	-	vGPU GRID Driver 525.85.07
CUDA	12	12
TensorRT	8.6.0	8.6.0
MLPerf Training	v3.0 BERT-large, RNNT	v3.0 BERT-large, RNNT
MLPerf Inference	v3.0	v3.0

AI/ML training performance in vSphere with NVIDIA vGPU

ML/AI workloads are becoming pervasive in today’s data centers and cover many domains. To show the flexibility of vSphere virtualization in disparate environments, we chose different types of workloads. For training, we chose natural language processing, represented by BERT, and speech, represented by RNNT, as shown in table 3.

Table 3. The MLPerf Training workloads used in our performance study

Area	Benchmark	Dataset	Quality target	Reference implementation model
Language	Speech recognition	LibriSpeech	0.058 Word Error Rate	RNNT
Language	NLP	Wikipedia 2020/01/01	0.72 Mask-LM accuracy	BERT-large (340 Million parameters)

Figure 1 compares the training times of MLPerf Training v3.0 benchmarks using vSphere 8.0.1 with NVIDIA vGPU 4x A100-80c against the bare metal 4x A100-80GB GPU configuration. The bare metal baseline is set to 1.00, and the virtualized result is presented as a relative percentage of the baseline. (Note that lower is better.)

vSphere with NVIDIA vGPUs delivers near bare metal performance with a difference of just 6-8% for training when using RNNT and BERT.

Figure 1. Normalized training times: vGPU 4x A100-80c vs bare metal 4x A100-80GB

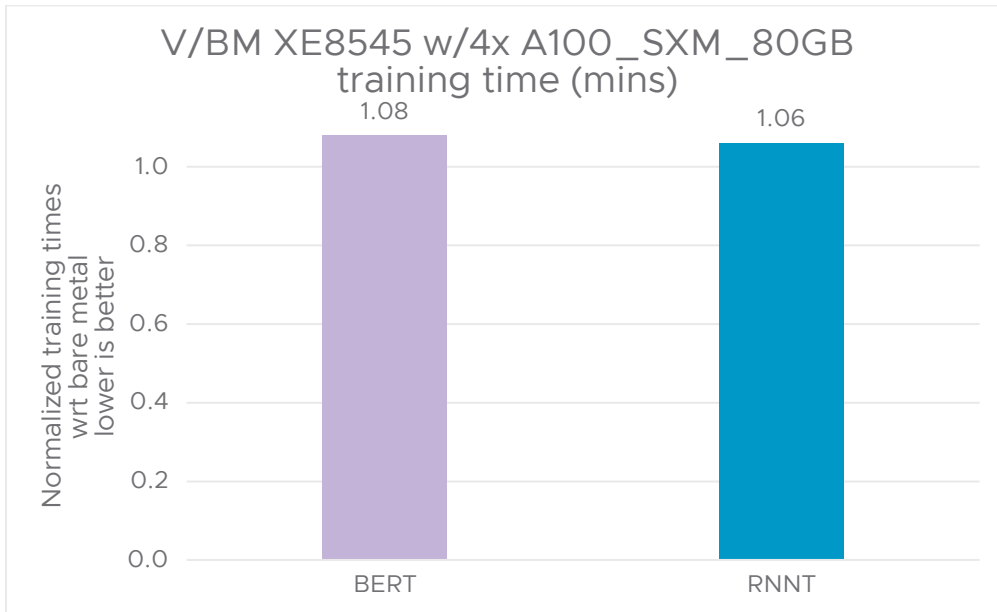


Table 4 shows training times in minutes for the MLPerf Training v3.0 benchmarks.

Table 4. vGPU4x A100-80C vs. Bare Metal 4x A100-80GB training times (mins)

Benchmark	Bare metal 4x A100 training times (mins)	vGPU 4x A100-80c training times (mins)	vGPU/BM
BERT-large	32.792	35.28	1.08
RNNT	55.086	58.447	1.06

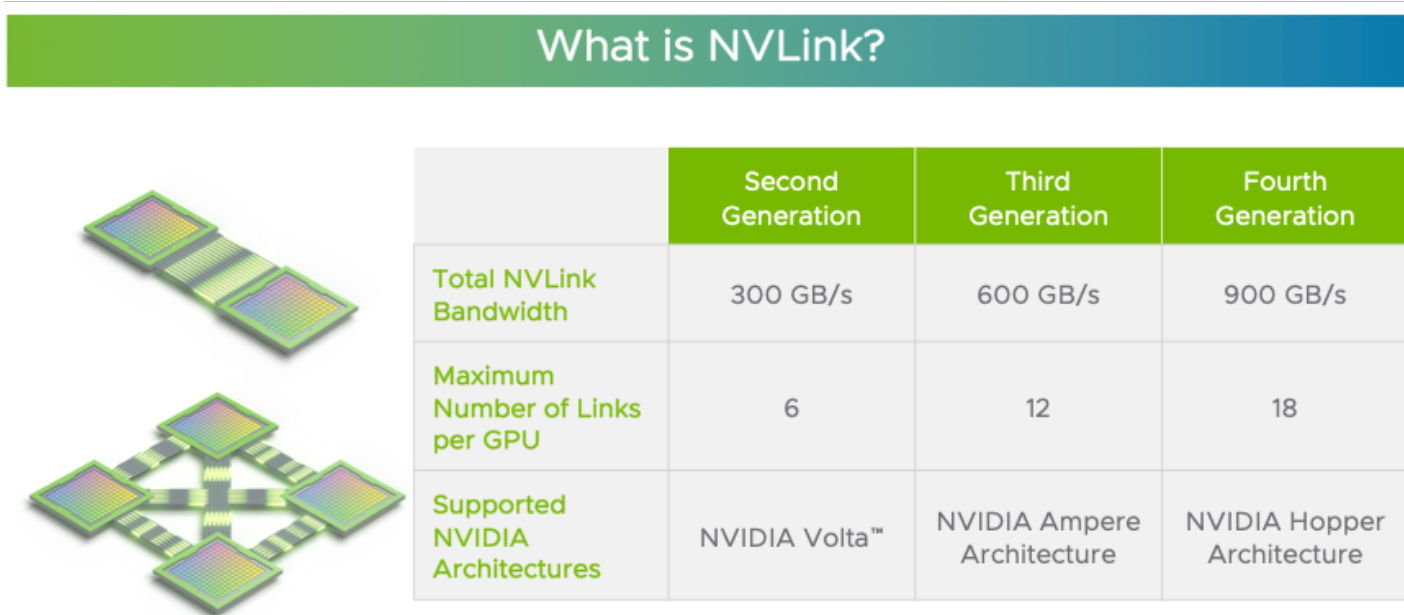
Dell obtained the bare metal results, which are published in the MLPerf Training v3.0 closed division with the submitter id of 3.0-2050.²

² The virtualized results by VMware were not verified by the MLCommons Association. The MLPerf® name and logo are registered and unregistered 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.

Flexible AI/ML infrastructure using device groups in vSphere with NVIDIA GPUs and NVLink

NVLink is a high-speed connection for GPUs and CPUs formed by a robust software protocol, typically riding on multiple pairs of wires printed on a computer board. A device group in vSphere 8 can comprise 2 or more GPUs connected by NVLinks, allowing more efficient communication between them. Figure 2 shows device groups consisting of GPUs connected by NVLinks. It also shows a table listing different generations of NVLinks and throughput improvements. NVLinks are extremely critical for ML/AI performance. They allow multiple GPUs to share the high bandwidth memory (HBM) associated with each of the GPUs on a single host. Thus, it allows larger machine learning models to fit into HBM.

Figure 2. Device groups and generations of NVLink (source: www.nvidia.com/en-us/data-center/nvlink/)



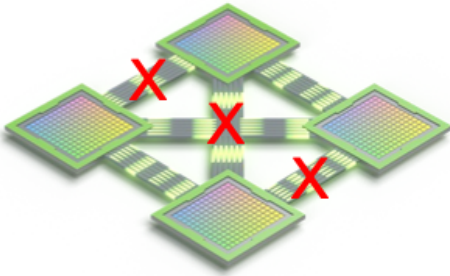
NVLink can be programmatically enabled and disabled to create new device groups. When a machine learning model is trained, it is typically trained with FP32 precision. Before deploying this model for inference, it is often quantized to use lower precisions like FP16, BF16, or Int8. The conversion to lower precision reduces the model size and its computation requirements. Now the model can be deployed on a smaller number of GPUs for inference.

Figure 3 shows disabling NVLinks to go from a 4-GPU device group to a 2-GPU device group. These two groups of GPUs are isolated from each other and can be allocated to two different VMs belonging to different tenants. They could be used for deploying two ML models for inference. This is what VMware did when leveraging the same hardware for ML training and inference.

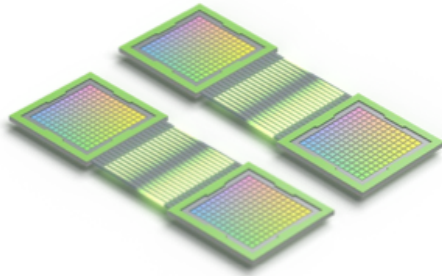
Figure 3. Enabling and disabling NVLinks to create device groups in vSphere


What are Device Groups in vSphere?

1 device group NVIDIA: 4@grid_a100x-40c%NVLink 2x Device Groups of NVIDIA: 2@grid_a100x-40c%NVLink



NVLinks between device groups are disabled providing isolation for the device groups.



vmware 

11

MLPerf Inference performance in vSphere with NVIDIA vGPU

VMware used the MLPerf Inference v3.0 suite to test the data center apps shown in table 4 below. MLPerf [published the official results](#) for these benchmarks. For inference, we chose natural language processing, represented by BERT; object detection, represented by Retinanet; medical imaging, represented by 3D UNET; and speech, represented by RNNT.

Table 4. The MLPerf Inference benchmarks used in our performance study

Area	Task	Model	Dataset	QSL Size	Quality	Scenarios	Server latency constraint
Vision	Object detection	Retinanet	OpenImages (800x800)	64	99% of FP32 (0.20 mAP)	Server, Offline	100ms
Vision	Medical image segmentation	3D UNET	KITS 2019 (602x512x512)	16	99% of FP32 and 99.9% of FP32 (0.86330 mean DICE score)	Offline	N/A
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)	Server, Offline	1000ms
Language	Language processing	BERT-large	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 and 99.9% of FP32 (f1_score=90.874%)	Server, Offline	130ms

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, and, in this case, the latency for serving a query is a critical metric.

MLPerf Inference performance results for bare metal and virtual configurations

Figures 4 and 5 compare the throughput (queries processed per second) of MLPerf Inference benchmark workloads using vSphere 8.0.1 with the 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.

Figures 5 and 6 show that virtualized performance is very near that of bare metal, with results ranging from only a 3-5% difference between the two systems.

Figure 4. Normalized throughput for OFFLINE scenario (qps): vGPU 2x H100-80c vs bare metal 2x H100

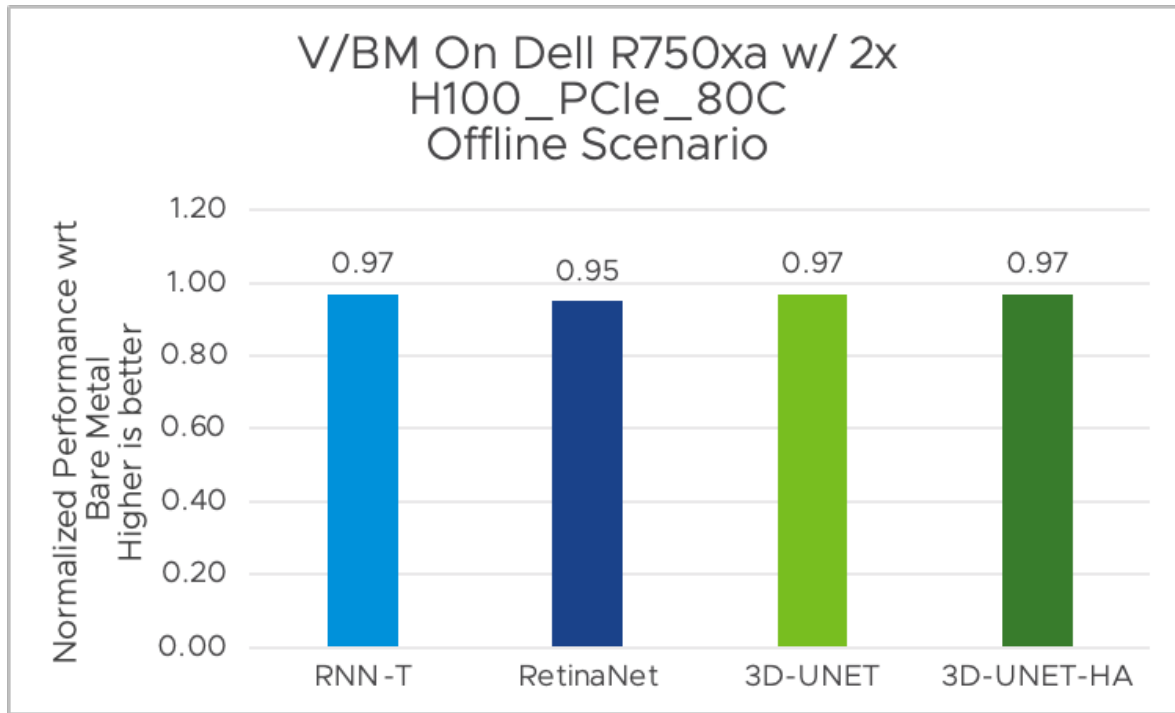
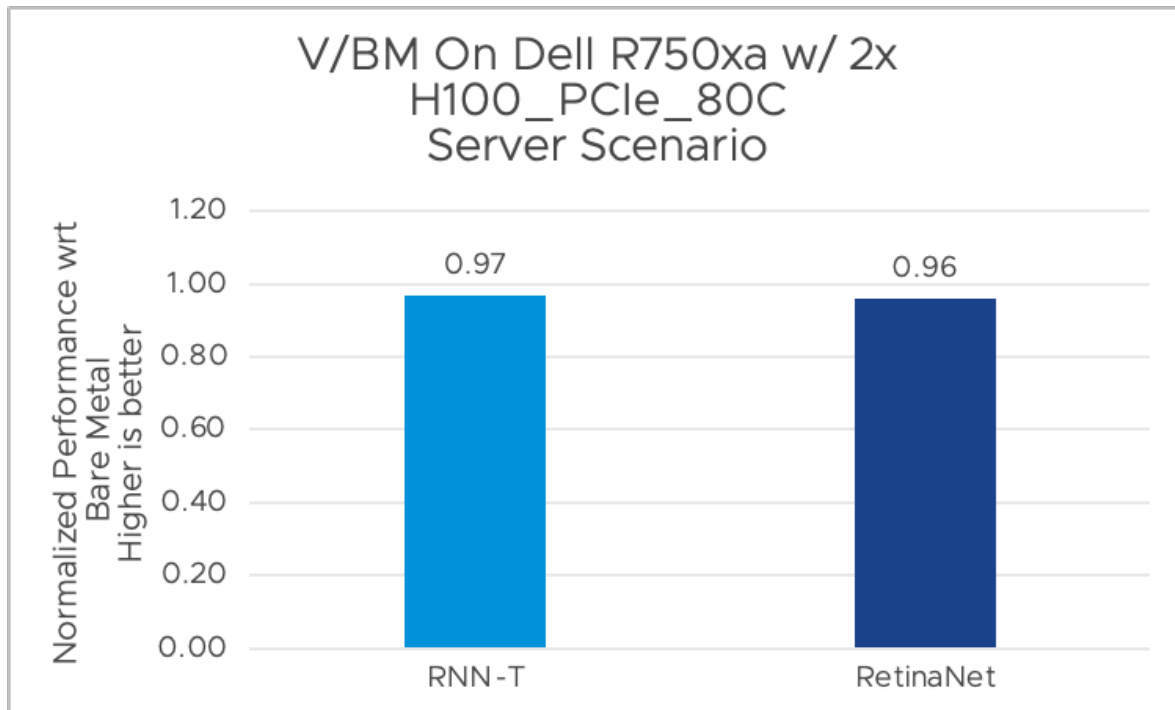


Figure 5. Normalized throughput for SERVER scenario (qps): vGPU 2x H100-80c vs bare metal 2x H100



VMware vSphere 8 Performance Is in the “Goldilocks Zone” for AI/ML Training and Inference

Tables 5 and 6 show throughput numbers in queries per second for MLPerf Inference benchmarks on 2x H100 for the Offline and Server scenarios.

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

Benchmark	Bare metal 2x H100	vGPU 2x H100-80c	vGPU/BM
RNNT Offline	33741.00	32771.40	0.97
Retinanet Offline	1892.09	1800.60	0.95
3d-UNET-99 Offline	9.05	8.76	0.97
3d-UNET-99.9 HA Offline	9.05	8.76	0.97

Table 6. vGPU 2x H100-80c vs bare metal 2x H100 throughput (queries per second) for SERVER scenario

Benchmark	Bare metal 2x H100	vGPU 2x H100-80c	vGPU/BM
RNN-T Server	32004.00	31131.20	0.97
RetinaNet Server	1852.24	1772.81	0.96

The [above results are published by MLCommons](#) in the closed division with the submitter id of 3.0-0017.

Figures 6 and 7 compare throughput (queries processed per second) for MLPerf Inference benchmarks using vSphere 8.0.1 with the NVIDIA vGPU A100-80c against the bare metal A100 GPU configuration. The bare metal baseline is set to 1.000, and the virtualized result is presented as a relative percentage of the baseline.

Our results show that vSphere with NVIDIA vGPUs delivers near bare metal and above bare metal performance ranging from 94.4% to 105% for Offline and Server scenarios when using the MLPerf Inference benchmarks.

Figure 6. Normalized throughput for OFFLINE scenario (qps): vGPU 4x A100-80c vs bare metal 4x A100-80GB

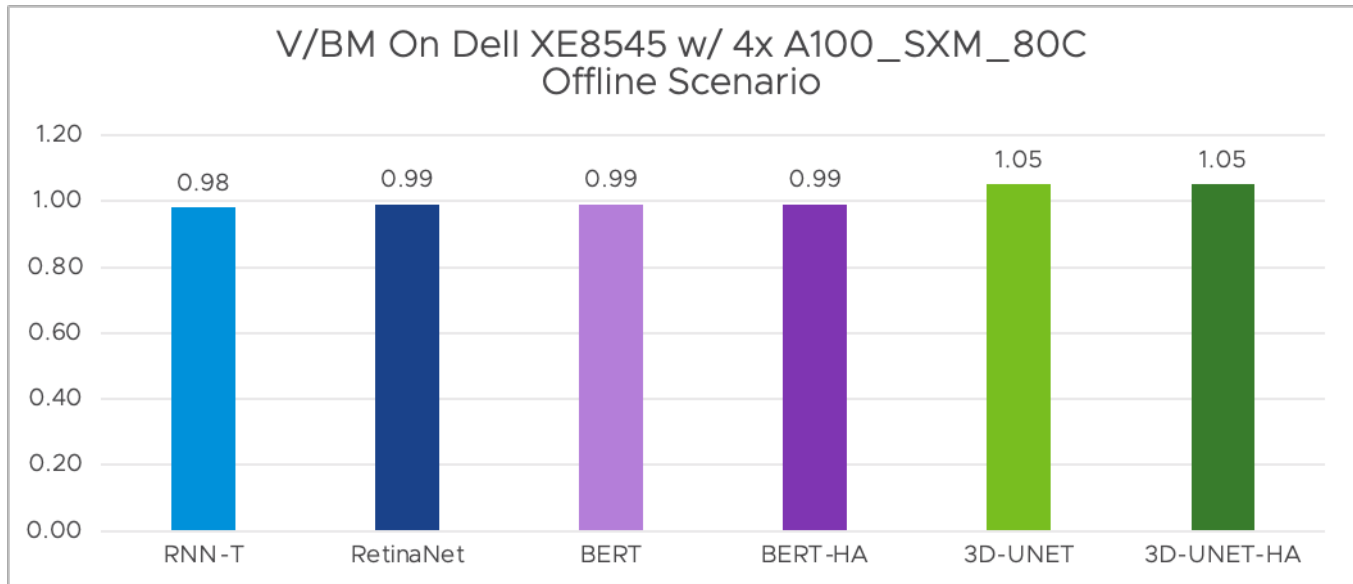
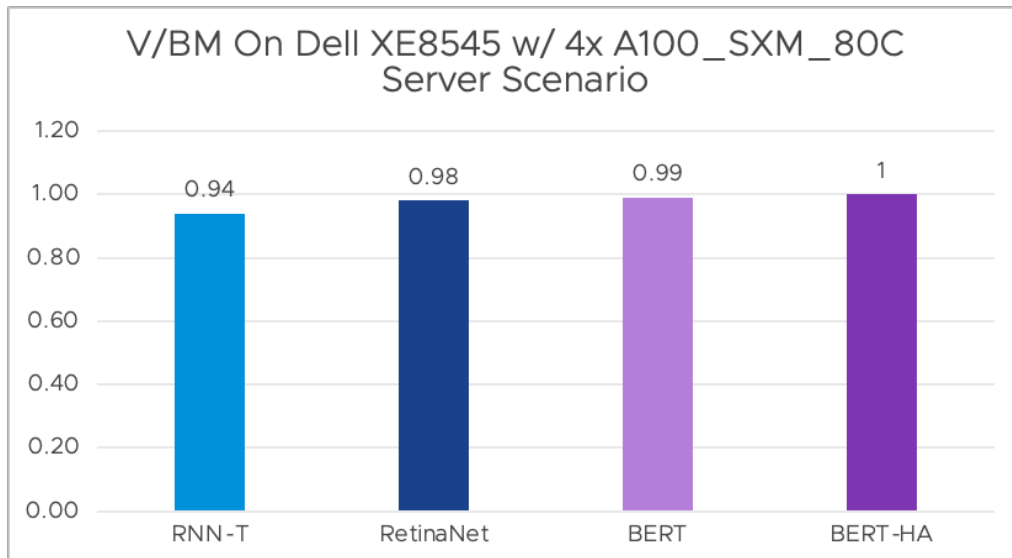


Figure 7. Normalized throughput for SERVER scenario (qps): vGPU 4x A100-80c vs bare metal 4x A100



VMware vSphere 8 Performance Is in the “Goldilocks Zone” for AI/ML Training and Inference

Tables 7 and 8 show throughput numbers for MLPerf Inference benchmarks.

Table 7. vGPU4x A100-80C vs bare metal 4x A100-80GB throughput (queries/second) for OFFLINE scenario

Benchmark	Bare Metal 4x A100	vGPU 4x A100-80c	vGPU/BM
RNN-T Offline	57084.00	56174.00	0.98
RetinaNet Offline	2910.78	2876.56	0.99
BERT Offline	15090.00	14923.10	0.99
BERT HighA Offline	7880.00	7767.84	0.99
3d-UNET-99 Offline	14.44	15.10	1.05
3d-UNET-99.9 HA Offline	14.44	15.10	1.05

Table 8. vGPU4x A100-80C vs bare metal 4x A100-80GB throughput (queries/second) for the SERVER scenario

Benchmark	Bare Metal 4x A100	vGPU 4x A100-80c	vGPU/BM
RNN-T Server	54000.40	51001.80	0.94
RetinaNet Server	2848.84	2798.93	0.98
Bert Server	13597.00	13497.90	0.99
Bert HighA Server	7004.00	7004.02	1.00

The above inference results are [published by MLCommons](#) in the closed division with the submitter id 3.0-0018.

Takeaways

- VMware vSphere with NVIDIA vGPU and AI technology performs within the “Goldilocks Zone”—the area of performance for good virtualization of AI/ML workloads.
- vSphere with NVIDIA AI Enterprise using NVIDIA vGPUs and NVIDIA AI software delivers from 106% to 108% of the bare metal performance measured as training times for MLPerf Training v3.0 benchmarks.
- vSphere achieved training performance with only 88 logical CPU cores out of 128 available CPU cores, thus leaving 40 logical CPU cores for other jobs in the data center.
- VMware used NVIDIA NVLinks and flexible device groups to leverage the same hardware setup for ML training and ML inference.
- vSphere with NVIDIA AI Enterprise using NVIDIA vGPUs and NVIDIA AI software delivers from 94% to 105% of the bare metal performance measured as queries served per second for MLPerf Inference v3.0 benchmarks.
- vSphere achieved inference performance with only 16 logical CPU cores out of 128 available CPU cores, thus leaving 112 logical CPU cores for other jobs in the data center.

vSphere combines the power of NVIDIA vGPUs and NVIDIA AI software with the data center management benefits of virtualization.

References

- MLCommons
<https://mlcommons.org/>
- MLPerf Training Benchmark Suite Results (filter on June 27, 2023 for v3.0)
<https://mlcommons.org/benchmarks/training/>
- VMware vSphere
<https://www.vmware.com/products/vsphere.html>
- 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?
<https://blogs.nvidia.com/blog/2023/03/06/what-is-nvidia-nvlink/>
- NVLink and NVLink Switch
<https://www.nvidia.com/en-us/data-center/nvlink/>
- MLPerf Inference: Datacenter Benchmark Suite Results (filter on April 05, 2023 for v3.0)
<https://mlcommons.org/benchmarks/inference-datacenter/>
- MLPerf Inference: Datacenter Benchmark Suite Results (filter on September 22, 2021 for v1.1)
<https://mlcommons.org/benchmarks/inference-datacenter/>
- 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: Virtual GPU Types for Supported GPUs
<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-vgpu.html>

- NVIDIA AI Enterprise
<https://www.nvidia.com/en-us/data-center/products/ai-enterprise/>
- NVIDIA A100 Tensor Core GPU
<https://www.nvidia.com/en-us/data-center/a100/>
- NVIDIA H100 Tensor Core GPU
<https://www.nvidia.com/en-us/data-center/h100/>
- NVIDIA Triton Inference Server
<https://developer.nvidia.com/nvidia-triton-inference-server>
- NVIDIA TensorRT
<https://developer.nvidia.com/tensorrt>
- 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 to 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.

Acknowledgments

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 thank Julie Brodeur of VMware for technical writing. The authors acknowledge Juan Garcia-Rovetta and Tony Lin of VMware for their management support.

