



GPU Positioning for Virtualized Compute and Graphics Workloads

Selecting the Right GPU for Your Virtualized Workload

Technical Brief

Table of Contents

Executive Summary	1
Selecting the Right NVIDIA GPU for Virtualization.....	2
NVIDIA GPUs for Virtualization	2
NVIDIA H100 Tensor Core GPU	4
NVIDIA L40	4
NVIDIA L4	5
NVIDIA A100 Tensor Core GPU	5
NVIDIA A30 Tensor Core GPU	5
NVIDIA A40	5
NVIDIA A16	6
NVIDIA A10	6
NVIDIA A2	6
GPU Performance Benchmark Tests	7
Knowledge Worker VDI	7
Professional Graphics	9
AI Deep Learning Training	11
AI Deep Learning Inference	13
Selecting the Right NVIDIA GPU Virtualization Software	17
Summary of Product Features	17
Optimal Workloads	18
Product Details	18
NVIDIA Virtual Applications	18
NVIDIA Virtual PC	19
NVIDIA RTX Virtual Workstation	19
NVIDIA AI Enterprise	19
Impact of GPU Sharing	20
Performance Allocation for a Shared GPU	20
Scheduling Options for GPU Sharing	20
Best Effort GPU Scheduler	20
Effect of GPU Sharing on Overall Throughput	21
NVIDIA AI Enterprise Performance and Scaling	23
NVIDIA AI Enterprise Single-Node Performance	23
NVIDIA AI Enterprise Multinode Scaling Performance	25
Conclusion	27
Additional Resources	28

List of Figures

Figure 1	NVIDIA vPC VDI Cost per User	9
Figure 2	RTX vWS SPECviewperf2020 Performance	10
Figure 3	RTX vWS SPECviewperf2020 Performance per Dollar	10
Figure 4	NVIDIA AI Enterprise Deep Learning Training Performance.....	12
Figure 5	NVIDIA AI Enterprise Deep Learning Training Performance per Dollar	12
Figure 6	NVIDIA AI Enterprise Deep Learning Inference Performance.....	14
Figure 7	NVIDIA AI Enterprise Deep Learning Inference Performance per Dollar	15
Figure 8	Best Effort GPU Scheduler.....	21
Figure 9	Effect of GPU Sharing on Overall Throughput.....	21
Figure 10	Inference Benchmarks for Measuring NVIDIA AI Enterprise Performance	24
Figure 11	Training Benchmarks for Measuring NVIDIA AI Enterprise Multinode Scaling Performance.....	25

List of Tables

Table 1	Optimal NVIDIA vGPU Solutions for Virtualized Workloads.....	1
Table 2	NVIDIA GPUs for Virtualization	3
Table 3	GPU Performance Benchmark Tests and Results.....	7
Table 4	Best GPUs for Knowledge Worker VDI Workloads	8
Table 5	Maximum Number of Supported NVIDIA vPC Knowledge Workers (with 1 GB Profile Size)	8
Table 6	Best GPUs for Professional Graphics Workloads.....	9
Table 7	Server Configuration for Benchmarking Professional Graphics Workloads.....	11
Table 8	Best GPUs for AI Deep Learning Training Workloads.....	11
Table 9	Server Configuration for Benchmarking AI Deep Learning Training Workloads	13
Table 10	Best GPUs for AI Deep Learning Inference Workloads.....	14
Table 11	Server Configuration for Benchmarking AI Deep Learning Inference Workloads	15
Table 12	NVIDIA GPU Virtualization Software Features	17
Table 13	Optimal Workloads for NVIDIA GPU Virtualization Software Products	18
Table 14	Server Configuration for Measuring the Effect of GPU Sharing on Overall Throughput.....	22
Table 15	Server Configuration for Measuring NVIDIA AI Enterprise Performance.....	24
Table 16	Server Configuration for Measuring NVIDIA AI Enterprise Multinode Scaling Performance.....	26

Executive Summary

The [NVIDIA virtual GPU](#) (vGPU) solution provides a flexible way to accelerate virtualized workloads – from Artificial Intelligence (AI) to Virtual Desktop Infrastructure (VDI). This solution includes NVIDIA graphics processing units (GPUs) for virtualization and NVIDIA software for virtualizing these GPUs.

Decoupling the GPU hardware and virtual GPU software options enables customers to benefit from innovative features delivered in the software at a regular cadence, without the need to purchase new GPU hardware. It also provides the flexibility for IT departments to architect the optimal solution to meet the specific needs of users in their environment.

The flexibility of the NVIDIA vGPU solution sometimes leads to the question, “How do I select the right combination of NVIDIA GPUs and virtualization software that best meets the requirements of my workloads?” In this technical brief, you will find guidance to help you answer that question.

This guidance is based on factors such as raw performance, performance per dollar¹, and overall cost effectiveness. It serves as a great starting point for understanding best practices for accelerating workloads in a virtualized infrastructure. However, to determine the NVIDIA virtual GPU solution that best meets your needs, you must test the solution with your own workloads.

You should also consider other factors, such as which NVIDIA vGPU certified [OEM server](#) to select, which NVIDIA GPUs are supported by that server, and any power and cooling constraints.

Table 1 summarizes the NVIDIA vGPU solutions for virtualized workloads.

Table 1 Optimal NVIDIA vGPU Solutions for Virtualized Workloads

Workload	GPU Virtualization Software	Best Raw Performance GPU	Most Cost-Effective GPU
Knowledge worker VDI	NVIDIA vPC	NVIDIA T4, A2, A10, A16, A40, L4, and L40 achieve similar performance ²	NVIDIA A16
Professional graphics	NVIDIA RTX vWS	NVIDIA L40	NVIDIA L4
AI deep learning inference	NVIDIA AI Enterprise	NVIDIA H100	NVIDIA H100
AI deep learning training	NVIDIA AI Enterprise	NVIDIA H100	NVIDIA H100

¹ Performance per dollar is calculated by adding the estimated GPU street prices to the cost of a 4-year or 5-year subscription to NVIDIA virtual GPU software and dividing the total cost by the number of users.

² NVIDIA H100, A100 and NVIDIA A30 do not support graphics workloads.

Selecting the Right NVIDIA GPU for Virtualization

The GPU that best meets the requirements of your workloads depends on the importance to you of factors such as raw performance, time-to-solution, performance per dollar, performance per watt, form factor, and any power and cooling constraints.

NVIDIA GPUs for Virtualization

Table 2 summarizes the features of the [NVIDIA GPUs for virtualization](#) workloads based on the NVIDIA Ampere GPU architecture.

The NVIDIA H100 Tensor Core GPU, NVIDIA A100 Tensor Core GPU and NVIDIA A30 Tensor Core GPU support the NVIDIA [Multi-Instance GPU](#) (MIG) feature. The MIG feature partitions a single GPU into smaller, independent GPU instances which run simultaneously, each with its own memory, cache, and streaming multiprocessors. The MIG feature is not supported on other GPUs such as the NVIDIA A2, NVIDIA A10, NVIDIA A16, NVIDIA A40, NVIDIA L4, NVIDIA L40, and NVIDIA T4.

GPUs for graphics workloads based on the NVIDIA Lovelace, Hopper, and Ampere GPU architecture feature second and third-generation RT Cores. RT Cores are accelerator units that are dedicated to performing ray tracing operations with extraordinary efficiency.

The GPUs in Table 2 are tested and supported with NVIDIA software for virtualizing GPUs, namely: NVIDIA AI Enterprise and NVIDIA virtual GPU software. For the full product support matrixes for the NVIDIA software for virtualizing GPUs, refer to the following documentation:

- > [NVIDIA AI Enterprise Product Support Matrix](#)
- > [Virtual GPU Software Supported Products](#)

Table 2 NVIDIA GPUs for Virtualization

	H100	A100	A30	L4	L40	A40	A10	A16	A2
GPUs/Board (Architecture)	1 (Hopper)	1 (Ampere)	1 (Ampere)	1 (Lovelace)	1 (Lovelace)	1 (Ampere)	1 (Ampere)	4 (Ampere)	1 (Ampere)
RTX Technology	--	--	--	✓	✓	✓	✓	✓	✓
Memory Size and Type	80GB HBM3	40/80GB HBM2	24GB HBM2	24GB GDDR6	48GB GDDR6	48GB GDDR6	24GB GDDR6	64GB GDDR6 (16GB per GPU)	16GB GDDR6
vGPU Profile Sizes (GB)	4, 5, 8, 10, 16, 20, 40, 80	4, 5, 8, 10, 16, 20, 40, 80	4, 6, 8, 12, 24	1, 2, 3, 4, 6, 8, 12, 24	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 3, 4, 6, 8, 12, 24	1, 2, 4, 8, 16	1, 2, 4, 8, 16
MIG Support	Up to 7	Up to 7	Up to 4	No	No	No	No	No	No
NVLink Support	Yes	Yes	Yes	No	No	Yes	No	No	No
Form Factor	> PCIe 5.0 > Dual slot FHFL	> SXM4 > PCIe 4.0 > Dual slot FHFL	> PCIe 4.0 > Dual slot FHFL	> PCIe 4.0 > Single slot HHHH	> PCIe 4.0 > Dual slot FHFL	> PCIe 4.0 > Dual slot FHFL	> PCIe 4.0 > Single slot FHFL	> PCIe 4.0 > Dual slot FHFL	> PCIe 4.0 > Single slot HHHH
Power (W)	350	400/300	165	72	300	300	150	250	60
Cooling	Passive	Passive	Passive	Passive	Passive	Passive	Passive	Passive	Passive
Optimized For ³	Performance	Performance	Performance	Performance	Performance	Performance	Performance	Density	Density
Target Workloads	AI training and inference, HPC, data analytics	AI training and inference, HPC, data analytics	AI inference	VDI, mid-level to high-end virtual workstations	High-end virtual workstations/mixed virtual workstations and compute (AI, data science)	High-end virtual workstations/mixed virtual workstations and compute (AI, data science)	Entry-level to mid-level virtual workstations	Knowledge worker virtual desktops	AI inference, VDI, and virtual workstations

NVIDIA H100 Tensor Core GPU

The NVIDIA H100 Tensor Core GPU enables an order-of-magnitude leap for **large-scale AI and HPC** with unprecedented performance, scalability, and security for every data center. H100 is designed to provide the highest and most cost-effective performance for deep learning inference and training workloads and includes the NVIDIA AI Enterprise software suite to streamline AI development and deployment. With NVIDIA fourth generation NVLINK, H100 accelerates exascale scale workloads with a dedicated Transformer Engine for trillion parameter language models. For small jobs, H100 can be partitioned down to right-sized Multi-Instance GPU (MIG) partitions. With Hopper Confidential Computing, this scalable compute power can secure sensitive applications on shared data center infrastructure. The inclusion of the NVIDIA AI Enterprise software suite reduces time to development and simplifies deployment of AI workloads and makes H100 the most powerful end-to-end AI and HPC data center platform.

NVIDIA L40

The NVIDIA® L40, based on the NVIDIA Ada Lovelace GPU architecture, delivers unprecedented visual computing performance for the data center and provides revolutionary neural graphics, compute, and AI capabilities to accelerate the most demanding visual computing workloads. The L40 features 142 third-generation RT Cores that enhance real-time ray tracing capabilities and 568 fourth-generation Tensor Cores with support for the FP8 data format. These new features are combined with the latest generation CUDA Cores and 48GB of graphics memory to accelerate visual computing workloads from high-performance virtual workstation instances to large-scale digital twins in NVIDIA Omniverse. With up to twice the performance of the previous generation at the same power, the NVIDIA L40 is uniquely suited to provide the visual computing power and performance required by the modern data center. When combined with NVIDIA RTX™ Virtual Workstation (vWS) software, the NVIDIA L40 delivers powerful virtual workstations from the data center or cloud to any device. Millions of creative and technical professionals can access the most demanding applications from anywhere with awe-inspiring performance that rivals physical workstations—all while meeting the need for greater security.

³ **Performance-optimized GPUs** are designed to maximize raw performance for a specific class of virtualized workload. They are typically recommended for the following classes of virtualized workload:

- > **High-end virtual workstations** running professional visualization applications.
- > **Compute-intensive workloads** such as artificial intelligence, deep learning, or data science workloads.

Density-optimized GPUs are designed to maximize the number of VDI users supported in a server. They are typically recommended for knowledge worker virtual desktop infrastructure (VDI) to run office productivity applications, streaming video, and the Windows OS.

NVIDIA L4

The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for **video, AI, virtual workstations, and graphics applications** in the enterprise, in the cloud, and at the edge. And with NVIDIA's AI platform and full-stack approach, L4 is optimized for video and inference at scale for a broad range of AI applications to deliver the best in personalized experiences. As the most efficient NVIDIA accelerator for mainstream use, servers equipped with L4 power up to 120X higher AI video performance over CPU solutions and 2.5X more generative AI performance, as well as over 4X more graphics performance than the previous GPU generation. L4's versatility and energy-efficient, single-slot, low-profile form factor makes it ideal for edge, cloud, and enterprise deployments.

NVIDIA A100 Tensor Core GPU

The NVIDIA® A100 Tensor Core GPU is designed to bring AI to every industry, accelerating **compute-intensive workloads** such as artificial intelligence, deep learning, or data science workloads.

Where time-to-solution is extremely important, factors other than the performance per dollar of the virtual GPU infrastructure can have a significant effect on the cost-effectiveness of a virtual GPU solution. By using the acceleration offered by the NVIDIA A100, data scientists and analysts can achieve results orders of magnitude faster than with a less expensive but less capable virtual GPU infrastructure.

The NVIDIA A100 is available in the PCIe and SXM module form factors. The SXM module form factor is available with servers that support NVIDIA® [NVLink](#)®. With these servers, the NVIDIA A100 provides high performance and strong scaling for hyperscale and HPC data centers running applications that scale to multiple GPUs, such as deep learning applications.

NVIDIA A30 Tensor Core GPU

The NVIDIA A30 Tensor Core GPU is designed to provide cost-effective performance for **deep learning inference workloads**. For deep learning inference workloads, the most important consideration is typically the combination of performance per dollar and the flexibility that is a result of support for the Multi-Instance GPU (MIG) feature.

NVIDIA A40

Built on the RTX platform, the NVIDIA A40 GPU is uniquely positioned to power high-end virtual workstations running professional visualization applications, accelerating the **most demanding graphics workloads**. The second-generation RT Cores of the NVIDIA A40 enable it to deliver massive speedups for workloads such as photorealistic rendering

of movie content, architectural design evaluations, and virtual prototyping of product designs.

The NVIDIA A40 features 48 GB of frame buffer, but with the NVIDIA® [NVLink®](#) GPU interconnect, it can support up to 96 GB of frame buffer to power virtual workstations that support very large animations, files, or models. Although the NVIDIA A40 has 48 GB of frame buffer, the context switching limit per GPU limits the maximum number of users supported to 32. Refer to Table 6 to see how many VDI users can be supported by each GPU when each user has a vGPU profile with 1 GB of frame buffer.

The NVIDIA A40 is also suitable for running VDI workloads and compute workloads on the same infrastructure. Resource utilization can be increased by using common virtualized GPU accelerated server resources to run virtual desktops and workstations while users are logged on, and compute workloads after the users have logged off. Learn more from the NVIDIA whitepaper about [Using NVIDIA Virtual GPUs to Power Mixed Workloads](#).

NVIDIA A16

The NVIDIA A16 is designed to provide the most cost-effective graphics performance for **knowledge worker VDI workloads**. For these workloads, where users are accessing office productivity applications, web browsers, and streaming video, the most important consideration is achieving the best performance per dollar and the highest user density per server. With four GPUs on each board, the NVIDIA A16 is ideal for providing the best performance per dollar and a high number of users per GPU for these workloads.

NVIDIA A10

The NVIDIA A10 is designed to provide cost-effective graphics performance for accelerating and optimizing the performance of **mixed workloads**. When combined with NVIDIA RTX vWS software, it accelerates graphics and video processing with AI on mainstream enterprise servers. Its second-generation RT Cores make the NVIDIA A10 ideal for mainstream professional visualization applications running on high-performance mid-range virtual workstations.

NVIDIA A2

The NVIDIA A2 is designed to provide cost-effective compute performance for **deep learning inference workloads** and cost-effective graphics performance for **knowledge worker VDI workloads** with vPC. It is a versatile, low-power, small-footprint, entry-level GPU that delivers low cost per user.

GPU Performance Benchmark Tests

The GPU performance benchmark tests measure GPU performance for virtualized workloads that use NVIDIA GPU virtualization software. To measure the performance of a GPU running a specific virtualized workload, a representative benchmark test for the workload is run on the GPU.

In many cases, cost rather than raw performance is the principal factor in selecting the right virtual GPU solution for a specific workload. For this reason, the GPU performance benchmark tests measure both raw performance **and** performance per dollar.

Unless otherwise stated, the tests are run with vGPU profiles that are allocated all the physical GPU's frame buffer. This vGPU profile size was chosen because the impact of scaling does not vary between different GPUs⁴. Refer to "NVIDIA AI Enterprise Multinode Scaling Performance" for details.

Table 3 summarizes the results of the benchmark tests to determine which GPUs provide the best raw performance and the best performance per dollar for specific workloads.



Note: When choosing GPUs based on raw performance or performance per dollar, use these results **for general guidance only**. All results are based on the workloads listed in Table 3, which could differ from the applications being used in production.

Table 3 GPU Performance Benchmark Tests and Results

Workload	Benchmark	Best Raw Performance GPU	Most Cost-Effective GPU
Knowledge worker VDI	NVIDIA nVector Digital Worker Workload	Performance for all GPUs is the same ⁵	A16
Professional graphics	SPECviewperf 2020 (3840x2160)	L40	L4
AI deep learning inference	BERT Large Fine Tune Training	H100	H100
AI deep learning training	BERT Large Inference	H100	H100

Knowledge Worker VDI

GPU performance for knowledge worker VDI workloads was measured by using the **NVIDIA nVector Digital Worker Workload** benchmark test. NVIDIA nVector Digital Worker Workload is a benchmarking tool that simulates end users' workflows and measures key aspects of the user experience, including end-user latency, framerate, image quality, and resource utilization.

⁴ Assumes that enough frame buffer is available on all vGPUs across all GPUs.

⁵ NVIDIA H100, A100 and NVIDIA A30 do not support graphics workloads.

Test Results

The GPUs that provide the best raw performance and cost effectiveness for knowledge worker VDI workloads are listed in Table 4. For knowledge worker VDI workloads, the principal factor in determining cost effectiveness is the combination of performance per dollar and user density.

As more knowledge worker users are added to a server, the server consumes more CPU resources. Adding an NVIDIA GPU for this workload conserves CPU resources by offloading graphics rendering tasks to the GPU. As a result, user experience and performance are improved for end users.

Table 4 Best GPUs for Knowledge Worker VDI Workloads

Best Raw Performance	Most Cost Effective
NVIDIA T4, A2, A10, A16, A40, L4, and L40 achieve similar performance for this workload	NVIDIA A16

Table 5 Maximum Number of Supported NVIDIA vPC Knowledge Workers (with 1 GB Profile Size)

GPU	Maximum Users per GPU Board	Maximum Boards per Server ⁶	Maximum Users per Server
A16	64	3	192
L4	24	6	144
A10	24	6	144
L40	32	3	96
A40	32	3	96
A2	16	6	96
T4	16	6	96

Table 5 assumes that each user requires a vGPU profile with 1GB of frame buffer. However, to determine the profile sizes that provide the best user experience for the users in your environment, you must conduct a proof of concept (POC).

⁶ The maximum number of boards per server assumes a 2U server. Refer to the specifications for your preferred OEM server to determine the maximum number of boards supported.

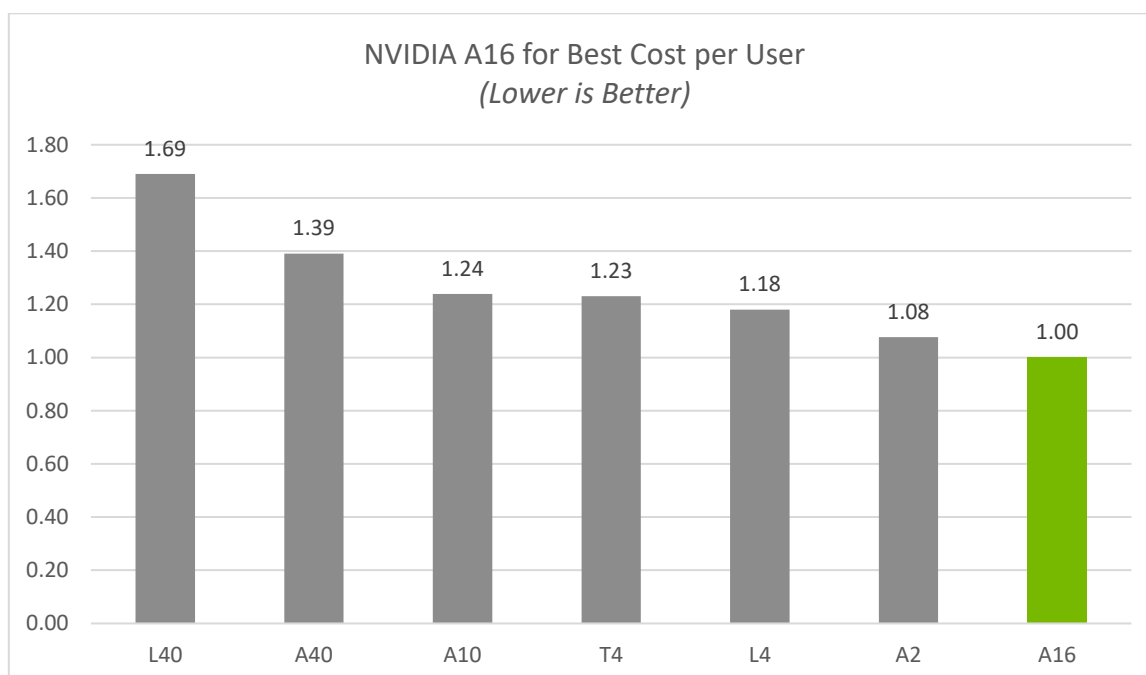
Figure 1 **NVIDIA vPC VDI Cost per User**

Figure 1 assumes an estimated GPU street price plus the cost of NVIDIA vPC software with a four-year subscription divided by number of users.

Professional Graphics

GPU performance for professional graphics workloads was measured by using the **SPECviewperf 2020 (3840x2160)** benchmark test. SPECviewperf 2020 is a standard benchmark for measuring the graphics performance of professional applications. It measures the 3D graphics performance of systems running under the OpenGL and Direct X application programming interfaces.

Test Results

The GPUs that provide the best raw performance and cost effectiveness for professional graphics workloads are listed in Table 6. For professional graphics workloads, the principal factor in determining cost effectiveness is performance per dollar.

Table 6 **Best GPUs for Professional Graphics Workloads**

Best Raw Performance	Most Cost Effective
NVIDIA L40	NVIDIA L4

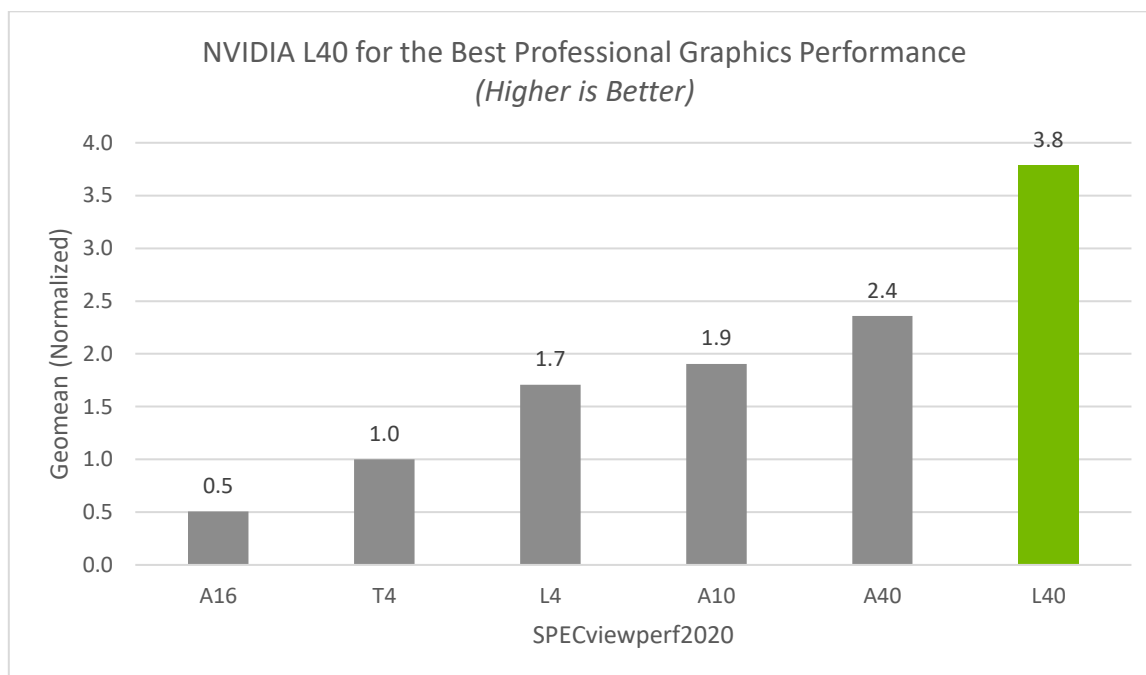
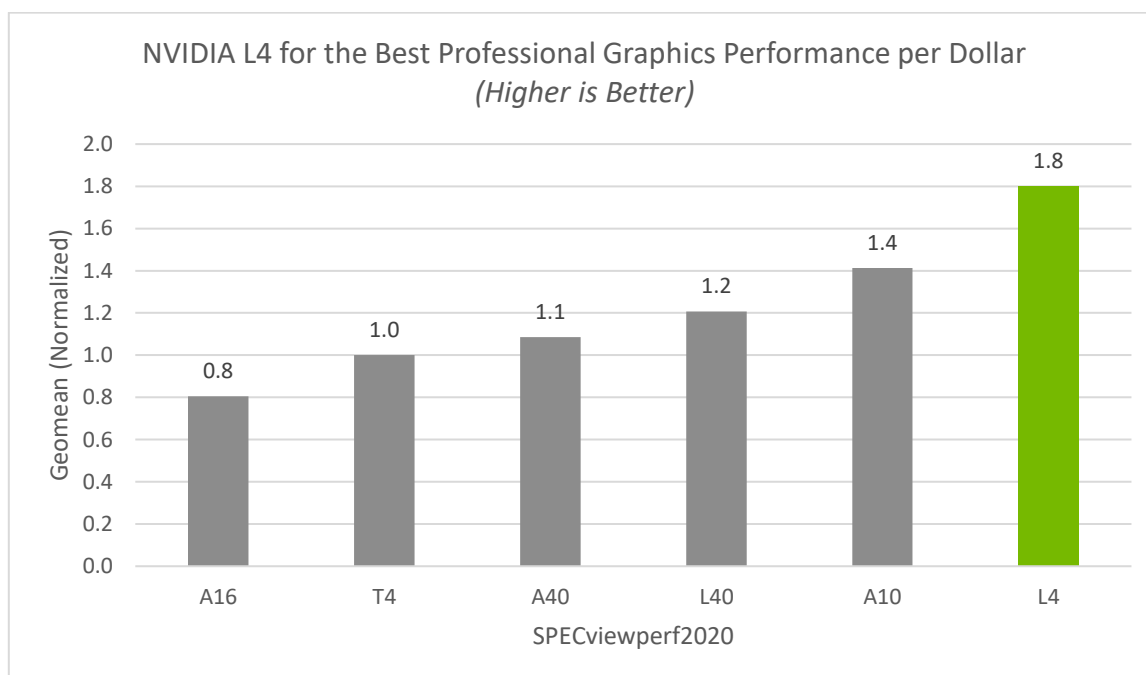
Figure 2 RTX vWS SPECviewperf2020 Performance**Figure 3 RTX vWS SPECviewperf2020 Performance per Dollar**

Figure 3 assumes an estimated GPU street price plus the cost of NVIDIA RTX vWS software with a four-year subscription.

Benchmark Server Configuration

The server configuration for benchmarking professional graphics workloads is listed in Table 7.

Table 7 **Server Configuration for Benchmarking Professional Graphics Workloads**

Property	Value
Server CPU	Intel Xeon Gold 6154
Hypervisor software	VMware ESXi 7.0 U2
VM vCPUs	8 vCPU
VM vMemory	16 GB
VM guest OS	Windows 11
GPU virtualization software	NVIDIA RTX vWS
Virtual GPU Manager driver version	525.76
Guest driver version	525.89
vGPU profile	L4-24Q, L40-48Q, A10-24Q, A16-16Q, A40-48Q, T4-16Q

AI Deep Learning Training

GPU performance for AI deep learning training workloads was measured by using the **BERT Large Fine Tune Training** benchmark test. BERT is one of the most widely used natural language processing models today.

Test Results

The GPUs that provide the best raw performance and cost effectiveness for AI deep learning training workloads are listed in Table 8. For AI deep learning training workloads, the principal factor in determining cost effectiveness is time-to-solution.

Table 8 **Best GPUs for AI Deep Learning Training Workloads**

Best Raw Performance	Most Cost Effective
NVIDIA H100 PCIe	NVIDIA H100 PCIe

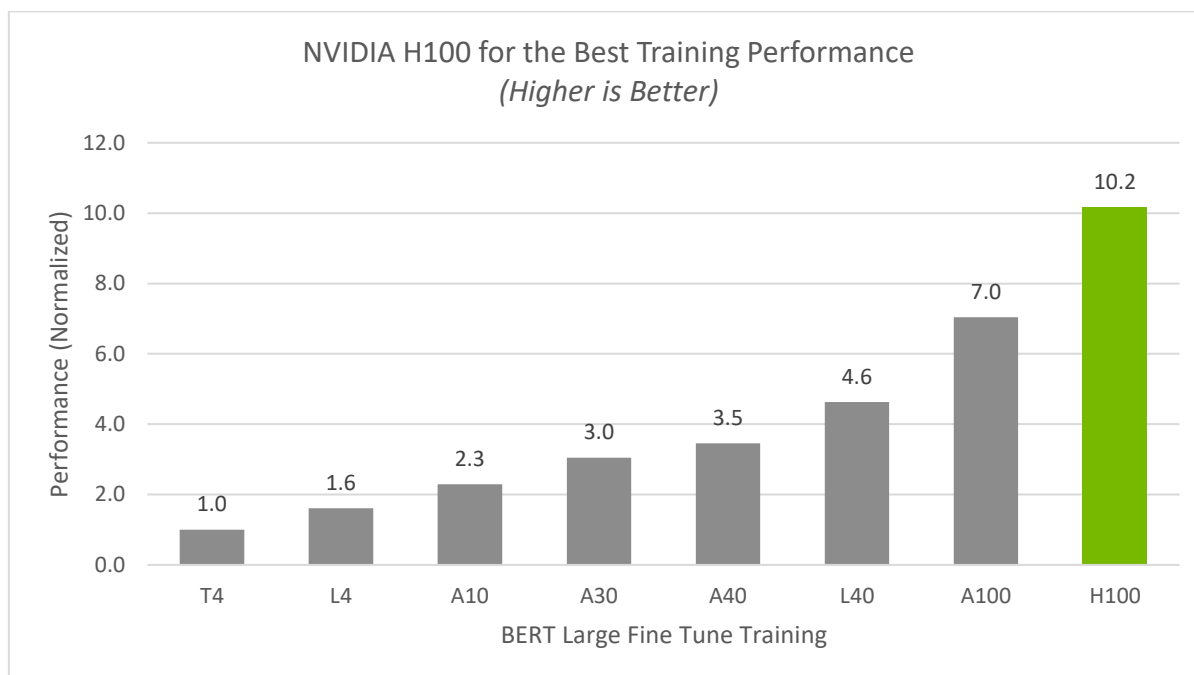
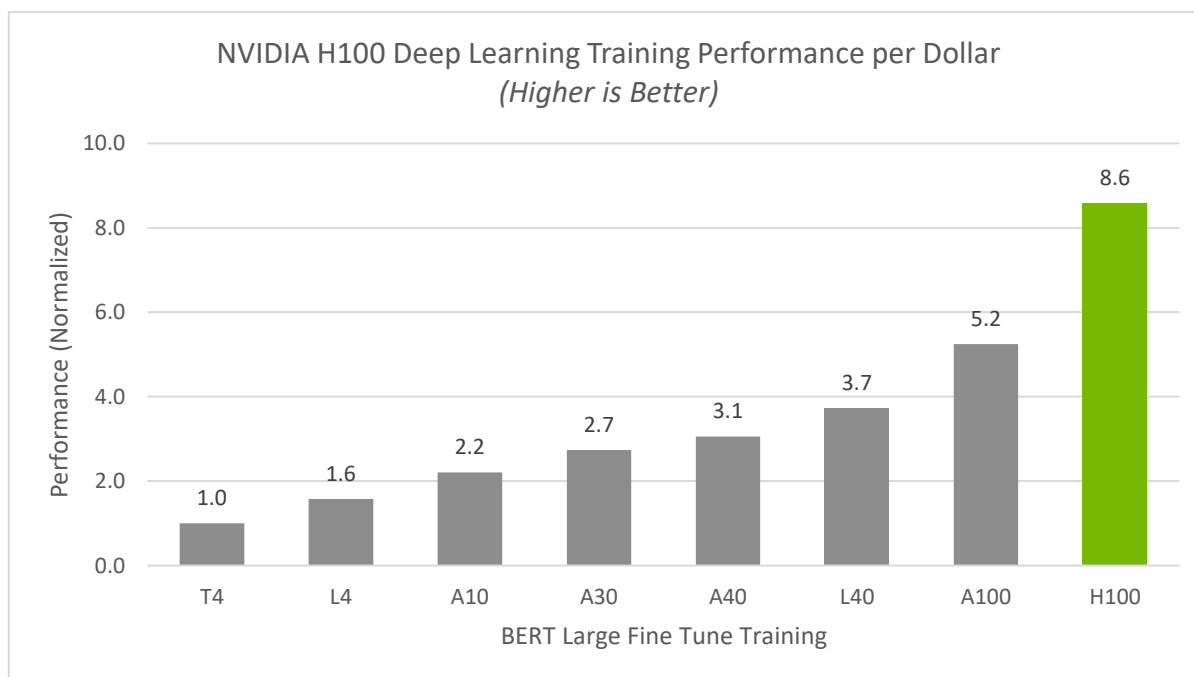
Figure 4 NVIDIA AI Enterprise Deep Learning Training Performance**Figure 5** NVIDIA AI Enterprise Deep Learning Training Performance per Dollar

Figure 5 assumes an estimated GPU street price plus the cost of NVIDIA AI Enterprise software with a five-year subscription.

Benchmark Server Configuration

The server configuration for benchmarking AI deep learning training workloads is listed in Table 9.

Table 9 Server Configuration for Benchmarking AI Deep Learning Training Workloads

Property	Value
Server CPU	AMD EPYC 7763, Intel Xeon Gold 6354
Hypervisor software	VMware ESXi 8.0
VM virtual CPUs	128: A10, A30, A40, A100, L40; 72: H100; 64: T4, L4
VM virtual memory	448: A10, A30, A40, A100, L40; 384: H100; 256: T4, L4
VM guest OS	Ubuntu 20.04.1
GPU virtualization software	NVIDIA AI Enterprise
Virtual GPU Manager driver version	525.52
Guest driver version	525.52
vGPU profile	L4-24C, L40-48C, H100-80C, A10-24C, A30-24C, A40-48C, A100DX-7-80C, T4-16C
Batch size	32× the number of GPUs
Precision	Mixed
Data	Real
Sequence length	384
cuDNN version	8.6.0.152
NCCL version	2.15.1
Baseline	DL 22.09
Installation source	NGC

AI Deep Learning Inference

GPU performance for AI deep learning inference workloads was measured by using the **BERT Large Inference** benchmark test. BERT is one of the most widely used natural language processing models today.

Test Results

The GPUs that provide the best raw performance and cost effectiveness for AI deep learning inference workloads are listed in Table 10. For AI deep learning inference workloads, the principal factor in determining cost effectiveness is the combination of performance per dollar and the flexibility that is a result of support for the Multi-Instance GPU (MIG) feature.

Table 10 Best GPUs for AI Deep Learning Inference Workloads

Best Raw Performance	Most Cost Effective
NVIDIA H100 PCIe	NVIDIA H100 PCIe

Figure 6 NVIDIA AI Enterprise Deep Learning Inference Performance

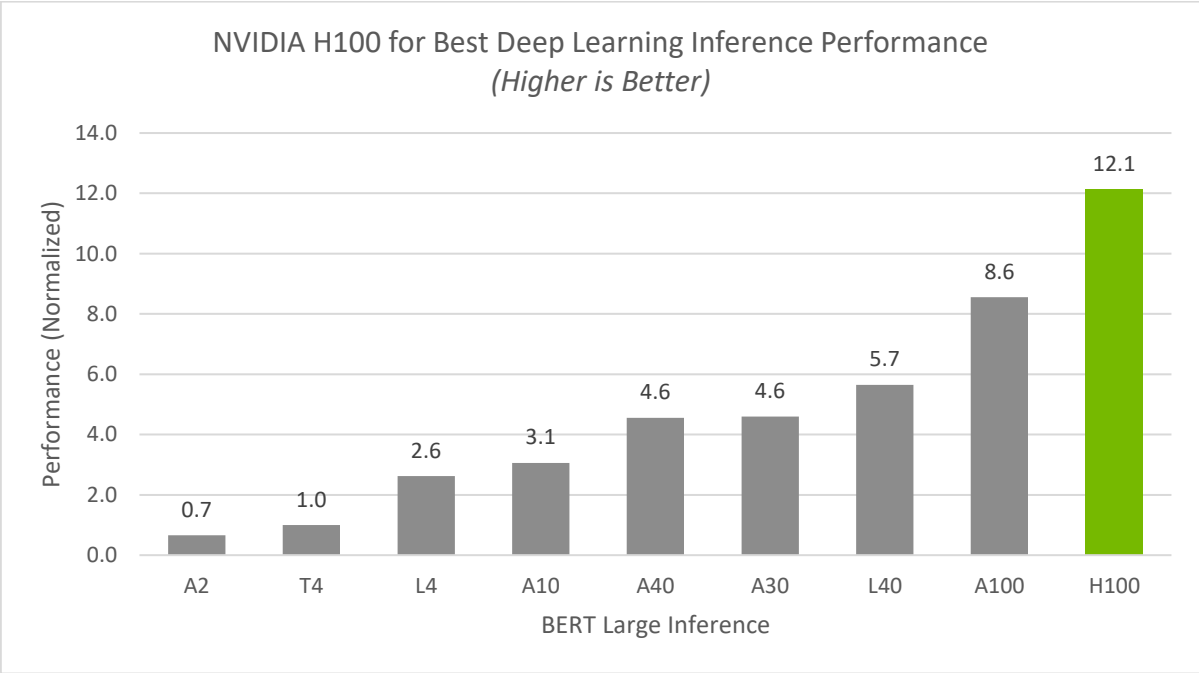


Figure 7 **NVIDIA AI Enterprise Deep Learning Inference Performance per Dollar**

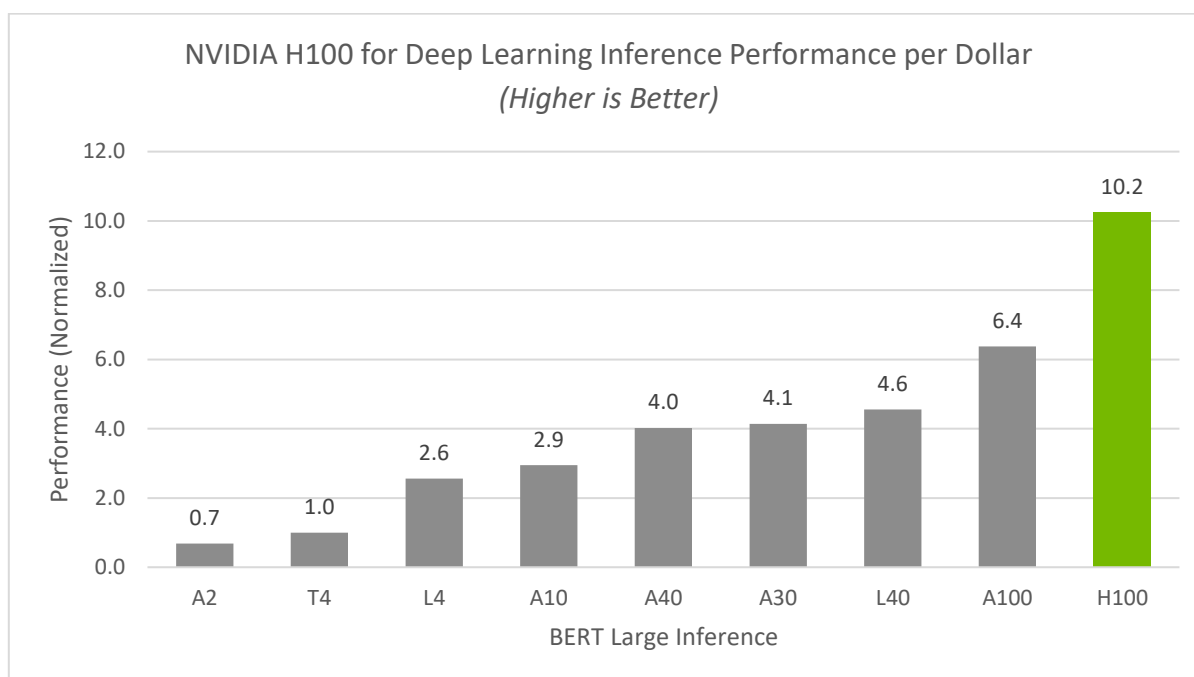


Figure 7 assumes an estimated GPU street price plus the cost of NVIDIA AI Enterprise software with a five-year subscription.

Benchmark Server Configuration

The server configuration for benchmarking AI deep learning inference workloads is listed in Table 11.

Table 11 **Server Configuration for Benchmarking AI Deep Learning Inference Workloads**

Property	Value
Server CPU	AMD EPYC 7763, Intel Xeon Gold 6354
Hypervisor software	VMware ESXi 8.0
VM virtual CPUs	128: A10, A30, A40, A100, L40; 72: H100; 64: T4, L4
VM virtual memory	448: A10, A30, A40, A100, L40; 384: H100; 256: T4, L4
VM guest OS	Ubuntu 20.04.1
GPU virtualization software	NVIDIA AI Enterprise
Virtual GPU Manager driver version	525.52
Guest driver version	525.52
vGPU profile	L4-24C, L40-48C, H100-80C, A10-24C, A30-24C, A40-48C, A100DX-7-80C, T4-16C

Property	Value
Batch size	128
Integer data type	INT 8
Sequence length	128
Precision	Mixed

Selecting the Right NVIDIA GPU Virtualization Software

NVIDIA GPU virtualization software products are optimized for different classes of workload. Therefore, you should select the right NVIDIA GPU virtualization software product on the basis of the workloads that your users are running.

Summary of Product Features

Table 12 summarizes the differences in features between the various NVIDIA GPU virtualization software products.

Table 12 NVIDIA GPU Virtualization Software Features

Configuration and Deployment	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise
Windows OS support	✓	✓	
Linux OS support	✓	✓	✓
NVIDIA graphics driver	✓	✓	
NVIDIA RTX enterprise driver	✓		
NVIDIA compute driver			✓
Multiple vGPUs per VM/NVLink	✓		✓
ECC reporting and handling	✓		✓
Page retirement	✓		✓
Display	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise
Maximum hardware rendered display	Four 5K Two 8K	Four QHD Two 4K One 5K	One 4K
Maximum resolution	7680×4302	5120×2880	4096×2160
Maximum pixel count	66,355,200	17,694,720	8,847,360
Advanced Professional Features	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise
ISV certifications	✓		
NVIDIA CUDA Toolkit/OpenCL support	✓		✓

Graphics Features and APIs	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise
NVENC	✓	✓	✓
OpenGL extensions (WebGL)	✓	✓	
In-situ graphics/GL support			✓
RTX platform optimizations	✓		
DirectX	✓	✓	
Vulkan support	✓		✓
Profiles	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise
Maximum supported frame buffer	48 GB	2 GB	80 GB
Available Profiles	0Q, 1Q, 2Q, 3Q, 4Q, 6Q, 8Q, 12Q, 16Q, 24Q, 32Q, 48Q	0B, 1B, 2B	4C, 5C, 6C, 8C, 10C, 16C, 20C, 40C, 80C

Optimal Workloads

Table 13 shows the different classes of workload for which NVIDIA GPU virtualization software products are optimized.

Table 13 Optimal Workloads for NVIDIA GPU Virtualization Software Products

Workload	GPU Virtualization Software
Knowledge worker VDI	NVIDIA Virtual PC (vPC)
Professional graphics	NVIDIA RTX Virtual Workstation (vWS)
AI deep learning inference	NVIDIA AI Enterprise
AI deep learning training	NVIDIA AI Enterprise

Product Details

Each NVIDIA GPU virtualization software product is designed for a specific class of workload.

NVIDIA Virtual Applications

[Virtual Applications \(vApps\)](#) software is designed for **app streaming** and **Remote Desktop Sharing Host (RDSH) workloads**.

NVIDIA Virtual PC

[NVIDIA Virtual PC \(vPC\)](#) software is designed for **knowledge worker VDI workloads** to accelerate the following software and peripheral devices:

- > Office productivity applications
- > Streaming video
- > The Windows OS
- > Multiple monitors
- > High-resolution monitors
- > 2D electric design automation (EDA)

NVIDIA RTX Virtual Workstation

[NVIDIA RTX Virtual Workstation \(RTX vWS\)](#) software is designed for **professional graphics workloads** that benefit from the following NVIDIA RTX vWS features:

- > RTX Enterprise platform drivers and ISV certifications
- > Support for NVIDIA® CUDA® Toolkit and OpenCL
- > Higher resolution displays
- > vGPU profiles with larger amounts of frame buffer

NVIDIA RTX vWS accelerates professional design and visualization applications such as:

- > Autodesk Revit
- > Dassault Systèmes CATIA
- > Esri ArcGIS Pro
- > Maya
- > Petrel
- > Solidworks

NVIDIA AI Enterprise

[NVIDIA AI Enterprise](#) is designed for **compute-intensive workloads**, such as artificial intelligence (AI), deep learning, data science, and high-performance computing (HPC) workloads. It is a secure, end-to-end, cloud-native suite of AI software that includes an extensive library of full-stack software such as:

- > NVIDIA AI workflows
- > Frameworks
- > Pretrained models
- > Infrastructure optimization

NVIDIA AI Enterprise accelerates the data science pipeline and streamlines the development and deployment of production AI including generative AI, computer vision,

speech AI and more. Available in the cloud, the data center and at the edge, NVIDIA AI Enterprise enables organizations to develop once and run anywhere. Global enterprise support and regular security reviews ensure business continuity and AI projects stay on track.

Impact of GPU Sharing

Using NVIDIA vGPU software to share a GPU among multiple virtual machines improves overall utilization of the GPU. NVIDIA vGPU software shares a GPU by scheduling the time that each virtual machine can use the GPU.

Performance Allocation for a Shared GPU

In general, when N virtual machines share a GPU, each a virtual machine is allocated $1/N$ of the total performance of the GPU. For example:

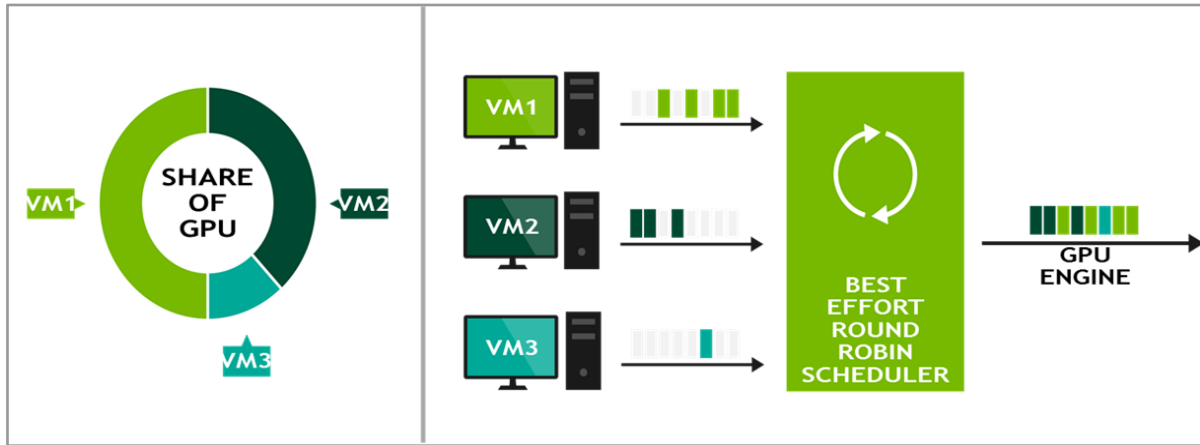
- > When **two** virtual machines share a GPU, each virtual machine is allocated approximately **50 percent** of the total performance of the GPU.
- > When **four** virtual machines share a GPU, each virtual machine is allocated approximately **25 percent** of the total performance of the GPU.

Scheduling Options for GPU Sharing

To accommodate a variety of Quality of Service (QoS) levels for sharing a GPU, NVIDIA vGPU software provides multiple GPU scheduling options. For more information about these GPU scheduling options, refer to [Changing Scheduling Behavior for Time-Sliced vGPUs](#) in *NVIDIA Virtual GPU Software User Guide*.

Best Effort GPU Scheduler

The default best effort GPU scheduling policy takes advantage GPU performance that other virtual machines have not used. When workloads across virtual machines are not executed at the same time, or are not always GPU bound, the best effort GPU scheduling policy enables a virtual machine's share of the GPU's performance to exceed its expected share. Figure 8 is a simplified illustration of how the best effort GPU scheduler works.

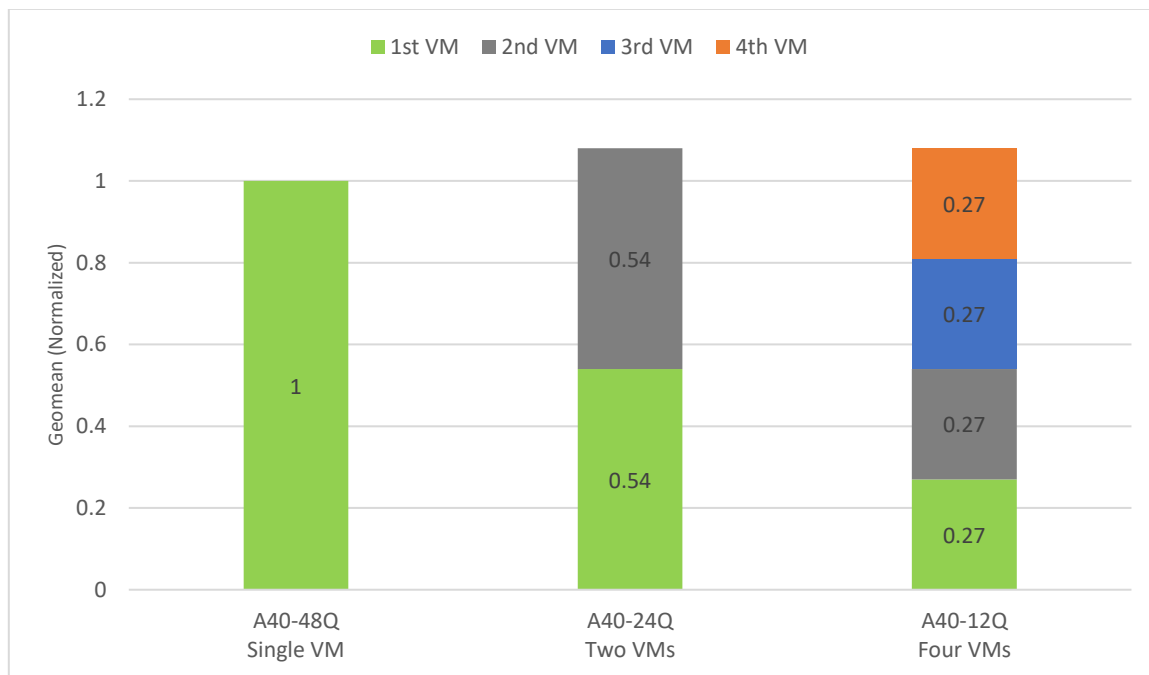
Figure 8 Best Effort GPU Scheduler

Effect of GPU Sharing on Overall Throughput

To measure the effect of GPU sharing on overall throughput, the SPECviewperf 2020 benchmark test was run against a GPU that was allocated to a single VM and then shared among two and four VMs.

Figure 9 shows the results of this test:

- > With **two** virtual machines, throughput is increased by **8%**.
- > With **four** virtual machines, throughput is increased by **16%**.

Figure 9 Effect of GPU Sharing on Overall Throughput

The server configuration for measuring the effect of GPU sharing on overall throughput is listed in Table 14.

Table 14 Server Configuration for Measuring the Effect of GPU Sharing on Overall Throughput

Property	Value
Server CPU	Intel Xeon Gold 6154 (18 cores, 3.0 GHz)
Server GPU	NVIDIA A40
Hypervisor software	VMware ESXi 7.0 U2
VM vCPUs	8 vCPU
VM vMemory	16 GB
VM guest OS	Windows 10
GPU virtualization software	NVIDIA RTX vWS
Virtual GPU Manager driver version	470.63
Guest driver version	471.68
vGPU profiles	A40-48Q, A40-24Q, A40-12Q
Workload	Specviewperf 2020

NVIDIA AI Enterprise Performance and Scaling

The benefits of virtualizing servers and applications, such as manageability, flexibility, and security, have traditionally come at the cost of lower single-node and multimode scaling performance. However, the difference in single-node performance between a virtualized environment with NVIDIA AI Enterprise and a bare-metal server is negligible and depends on the workload and other configuration variables. The multimode scaling performance of virtual machines with NVIDIA AI Enterprise distributed deep learning training equals the performance from bare-metal servers.



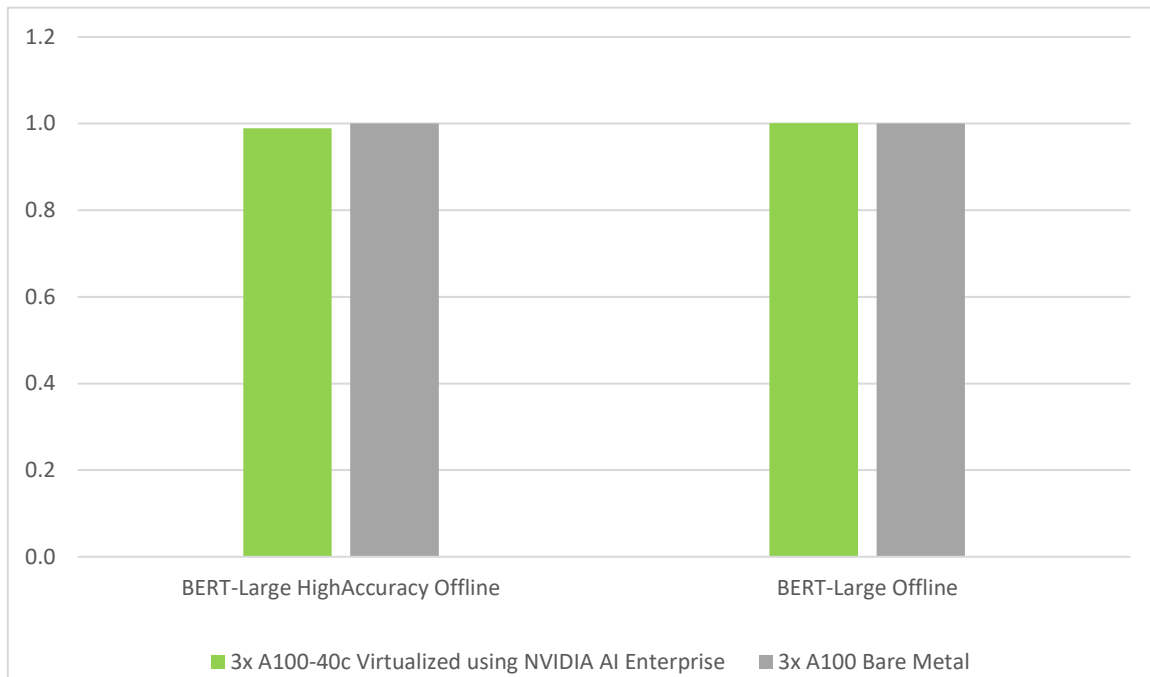
Note: NVIDIA AI Enterprise is also supported for bare metal environments.

NVIDIA AI Enterprise Single-Node Performance

To measure the single-node performance of a virtualized server in comparison to a bare metal server, the same benchmark test was run on both servers. NVIDIA AI Enterprise was used to virtualize the GPU in the virtualized server. The MLPerf™ Inference: Datacenter v1.1 Natural Language Processing (BERT-Large) benchmark test was used for these measurements.

Figure 10 shows the per-accelerator performance derived from the best results for submissions to both servers using the reported accelerator count in [MLPerf Inference v1.1 Results MLCommons](#) and [v1.1 Results MLCommons](#). These results demonstrate near bare-metal performance with NVIDIA AI Enterprise.

Figure 10 Inference Benchmarks for Measuring NVIDIA AI Enterprise Performance



The server configuration for measuring the performance of NVIDIA AI Enterprise is listed in Table 15.

Table 15 Server Configuration for Measuring NVIDIA AI Enterprise Performance

Property	Value
Server form factor	2U
Server sockets	2
Server CPU	AMD EPYC 7502
Server GPUs	3×NVIDIA A100 PCIe 40GB
Hypervisor software	VMware ESXi 7.0 U2
VM guest OS	Ubuntu 20.04 LTS
GPU virtualization software	NVIDIA AI Enterprise software
Virtual GPU Manager driver version	470.63
Guest driver version	470.63.01
vGPU profile	A100-40C
TensorRT version	8.0.2
NVIDIA CUDA Toolkit version	11.3
Workload	MLPerf Inference: Datacenter v1.1 Natural Language Processing (BERT-Large)

NVIDIA AI Enterprise Multinode Scaling Performance

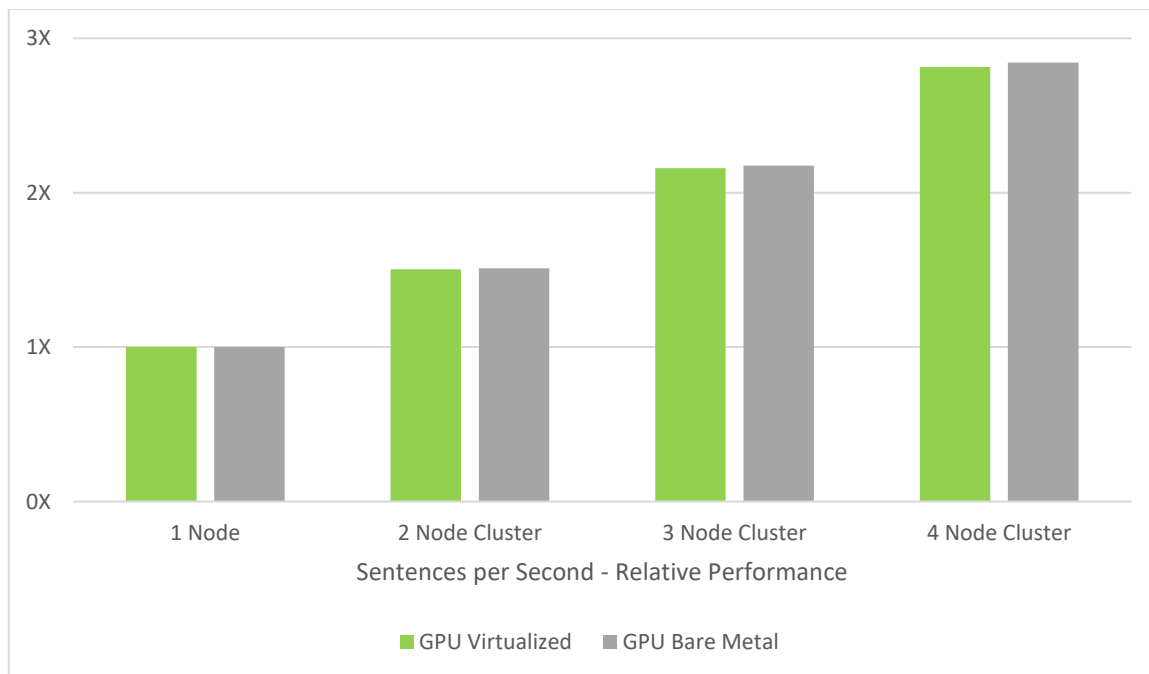
To provide multinode scaling performance that equals the performance from bare-metal servers, NVIDIA AI Enterprise supports features such as:

- > GPUDirect® remote direct memory access (RDMA) technology
- > Address Translation Services (ATS)
- > RDMA over Converged Ethernet (RoCE)

To measure the multinode scaling performance of a virtualized server in comparison to a bare metal server, the same benchmark test was run on a single server of each type and then on a two-node, three-node, and four-node cluster of each type of server. NVIDIA AI Enterprise was used to virtualize the GPU in each virtualized server used in the tests. The BERT Large Training benchmark test was used to train a natural language processing model for these measurements.

Figure 11 shows the results of these tests. They demonstrate for clusters of up to four nodes bare-metal levels of multinode scaling performance with NVIDIA AI Enterprise.

Figure 11 Training Benchmarks for Measuring NVIDIA AI Enterprise Multinode Scaling Performance



The server configuration for measuring the multinode scaling performance of NVIDIA AI Enterprise is listed in Table 16.

Table 16 **Server Configuration for Measuring NVIDIA AI Enterprise Multinode Scaling Performance**

Property	Value
Server CPU	Intel Xeon Gold (6240R @ 2.4GHz)
Server GPU	1×NVIDIA A100 PCIe 40GB
Network interface card	Mellanox Connect-X® 6 SmartNIC with RoCE and ATS enabled
Hypervisor software	VMware vSphere ESXi 7.0 U2
VM guest OS	Ubuntu 18.04 LTS
GPU virtualization software	NVIDIA virtual GPU (vGPU) software
Virtual GPU Manager driver version	460.32.04
Guest driver version	460.32.03
vGPU profile	A100-40C
Deep learning training framework	Horovod distributed deep learning training framework
Workload	BERT Large Training
Batch size	30
Precision	FP16
Sequence length	384

Conclusion

The NVIDIA solutions for virtualized compute and graphics workloads offers unmatched flexibility and performance when paired with GPUs based on the NVIDIA Hopper, Lovelace, and Ampere architecture. The solution is designed to meet the ever-shifting workloads and organizational needs of today's modern enterprises.

For **professional visualization workloads**, the optimal GPU for each class of workload is as follows:

- > The **NVIDIA L40** is uniquely positioned to power the most demanding graphics and rendering workloads for dynamic virtual workstations.
- > The **NVIDIA L4** offers the best performance per dollar for professional graphics workloads.
- > If the infrastructure supports knowledge worker VDI workloads, the **NVIDIA A16** provides the best performance per dollar, while also providing the best user density.

For **AI workloads**, including deep learning training and deep learning inferencing, the optimal GPU for each class of workload is as follows:

- > The **NVIDIA H100** offers the best raw performance **and** cost effectiveness for training and inference workloads. It is the most advanced data center GPU ever built, delivering high performance with unprecedented acceleration.

The NVIDIA H100 GPU supports MIG, optimizing GPU utilization and providing flexibility with dynamic reconfiguration of MIG instances.

NVIDIA GPU virtualization software products are optimized for different classes of workload. For details on how to best configure an accelerated virtualized infrastructure, refer to the sizing guidelines for these GPU virtualization software products:

- > [NVIDIA vPC Windows 10 Profile Sizing Guidance](#)
- > [NVIDIA RTX Virtual Workstation Application Sizing Guide](#)
- > [NVIDIA AI Enterprise Sizing Guide](#)

Although this technical brief provides general guidance on how to select the right NVIDIA GPU and virtualization software for your workloads, actual results may vary depending on the specific workloads that are being virtualized. To balance virtual machine density (scalability) with required performance, conduct a proof of concept (POC) with your production workloads. To allow the configuration to be optimized to meet the requirements for performance and scale, analyze the utilization of all resources of the system and gather subjective feedback from all stakeholders.

Additional Resources

NVIDIA vPC Resources

- > [NVIDIA vPC Windows 10 Profile Sizing Guidance](#)
- > [Quantifying the Impact of NVIDIA Virtual GPUs](#)
- > [NVIDIA vPC Solution Overview](#)
- > [NVIDIA vPC webpage](#)

NVIDIA RTX Virtual Workstation Resources

- > [NVIDIA RTX Virtual Workstation Application Sizing Guide](#)
- > [NVIDIA RTX vWS Solution Overview](#)
- > [NVIDIA RTX vWS webpage](#)

NVIDIA AI Enterprise Software Suite Resources

- > [NVIDIA AI Enterprise webpage](#)
- > [NVIDIA AI Enterprise Solution Overview](#)
- > [NVIDIA AI Enterprise Sizing Guide](#)

Other Resources

- > [Try NVIDIA vGPU for free](#)
- > [Using NVIDIA Virtual GPUs to Power Mixed Workloads](#)
- > [NVIDIA Virtual GPU Software Documentation](#)
- > [NVIDIA vGPU Certified Servers](#)
- > [NVIDIA LaunchPad: The End-to-End AI Platform](#)

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA CUDA, NVIDIA RTX, NVIDIA Turing, NVIDIA Volta, GPUDirect, NVLink, Quadro RTX, and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

MLPerf

The MLPerf name and logo are trademarks of MLCommons Association ("MLCommons") in the United States and other countries.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

© 2023 NVIDIA Corporation. All rights reserved.

TB-09867-001_v03