



NVIDIA Virtual GPU Positioning

Selecting the Right GPU for Your Virtualized Workload

Technical Brief

Table of Contents

- Intent of this Technical Brief..... 1**
- Executive Summary..... 2**
- Introduction..... 3**
 - Selecting the Right Virtual GPU Software..... 3
- NVIDIA GPUs Recommended for Virtualization..... 7**
 - Selecting the Right GPU 9
 - Professional Graphics 11
 - AI Deep Learning Training..... 12
 - AI Deep Learning Inference 14
 - Knowledge Workers..... 17
- NVIDIA AI Enterprise vs. Bare Metal 19**
- Impact of GPU Sharing 20**
- NVIDIA AI Enterprise Scale Out..... 22**
- Conclusion 23**
- Resources Links..... 24**

Intent of this Technical Brief

The flexibility of the NVIDIA vGPU solution sometimes leads to the question, “How do I select the right software license and GPU combination to best meet the needs of my workloads?”

In this technical brief, you will find guidance on how to select the best virtual GPU software license and graphics processing unit (GPU) combination, based on your workload. This guidance is based on variables such as performance and performance per dollar¹. Other factors that should be considered include things like which NVIDIA vGPU certified [OEM server](#) you’ve selected, which NVIDIA GPUs are supported in that platform, as well as any power and cooling constraints.



Note:

¹Performance per dollar assumes estimated GPU street price plus NVIDIA virtual GPU software license cost with 3- or 4-year subscription divided by the number of users.

Executive Summary

This technical brief provides general guidance based on performance and price for virtualized workloads using NVIDIA virtual GPU software. It is recommended that you test your unique workloads to determine the best NVIDIA virtual GPU solution to meet your needs. However, for those who want to better understand best practices for accelerating workloads in a virtualized infrastructure, this technical brief serves as a great starting point.

Table 1 summarizes the recommended GPU for running a specific virtualized workload, based only on performance. For this testing, we selected a representative benchmark for each workload, described in Table 5. For the specific benchmarks run with NVIDIA virtual GPU software, NVIDIA® A40 GPUs provided the best performance for professional graphics and rendering workloads, while the A100 provided the best performance for artificial intelligence (AI) including deep learning training and deep learning inferencing.

In many cases, raw performance is not the only factor considered when selecting the right virtual GPU solution for your workload. Cost is often also considered. Table 2 summarizes the recommended GPU if only performance per dollar is considered. If the infrastructure will support only a knowledge worker VDI workload, the A16 GPU provides the best performance per dollar, while also providing the best user density. The A40 GPU provides the best performance per dollar for professional graphics applications. It is important to note, for AI training workloads, time-to-solution is extremely important, and for that reason, costs outside of just infrastructure should be considered. For example, highly paid data scientists and analysts can achieve results orders of magnitude faster when using the unprecedented acceleration offered by the A100. As such, A100 would be recommended for this workload when considering these other cost factors.

Table 1. Best Performance GPU per Workload

Workload	Professional Graphics	AI Deep Learning Training	AI Deep Learning Inference	Knowledge Workers
GPU	A40	A100	A100	A40 and A16 perform the same ¹

Table 2. Best Performance per Dollar GPU per Workload

Workload	Professional Graphics	AI Deep Learning Training	AI Deep Learning Inference	Knowledge Workers
GPU	A40	A100	A100	A16

¹ NVIDIA A100 and A30 are not supported for graphics workloads.

Introduction

The [NVIDIA virtual GPU \(vGPU\)](#) solution provides a flexible way to accelerate virtualized workloads – from AI to VDI. The solution includes NVIDIA virtual GPU software and NVIDIA data center GPUs. There are several unique NVIDIA virtual GPU software licenses, each priced and designed to address a specific use case.

For VDI workloads, leverage the following software licenses:

- ▶ [NVIDIA Virtual PC \(vPC\)/Virtual Applications \(vApps\)](#) – accelerates office productivity applications, streaming video, Windows, RDSH, multiple and high-resolution monitors and 2D electric design automation (EDA).
- ▶ [NVIDIA RTX Virtual Workstation \(RTX vWS\)](#) – accelerates professional design and visualization applications including Autodesk Revit, Maya, Dassault Systèmes CATIA, Solidworks, Esri ArcGIS Pro, Petrel, and more.

For AI, data science and high-performance computing workloads, leverage the following software licenses:

- ▶ [NVIDIA AI Enterprise](#) – an end-to-end cloud-native suite of AI and data analytics software, including NVIDIA vGPU software, that is optimized, certified, and supported by NVIDIA to run on VMware vSphere with [NVIDIA-Certified Systems](#).
- ▶ [NVIDIA Virtual Compute Server \(vCS\)](#) – accelerates artificial intelligence (AI), deep learning (DL), data science and high-performance computing (HPC) workloads in a virtualized environment with Red Hat Virtualization or other KVM-based hypervisors.

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 a dependency on purchasing new GPU hardware. It also provides the flexibility for IT to architect the optimal solution to meet the specific needs of users in their environment.

Selecting the Right Virtual GPU Software

Select your NVIDIA virtual GPU software license based on the workload(s) your users are running. Table 3 shows the feature differences between the NVIDIA vGPU software license options. NVIDIA vPC software is selected for knowledge worker VDI to run office productivity applications. NVIDIA RTX vWS is selected to virtualize professional visualization applications which benefit from the RTX Enterprise platform drivers and ISV certifications, support for NVIDIA® CUDA® and OpenCL, higher resolution displays, and larger profile sizes. For server virtualization to run compute workloads such as AI, data science and HPC, with VMware vSphere, the NVIDIA AI Enterprise license, which includes a driver that has been tested to run these compute workloads, would be selected. Or leverage the NVIDIA Virtual Compute Server (vCS) software license for compute workloads with KVM hypervisors.

Table 3. NVIDIA Virtual GPU Software Features

Configuration and Deployment	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise or vCS
Windows OS Support	✓	✓	
Linux OS Support	✓	✓	✓
NVIDIA Graphics Driver	✓	✓	
NVIDIA RTX Enterprise Driver	✓		
NVIDIA Compute Driver			✓
Multi-vGPU/NVLink	✓		✓
GPUDirect Support			✓
GPU Operator			✓
ECC Reporting and Handling	✓		✓
Page Retirement	✓		✓
Display	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise or vCS
Maximum Hardware Rendered Display	Four 5K, Two 8K	Four QHD, Two 4K, One 5K	One 4K
Maximum Resolution	7680x4302	5120x2880	4096x2160
Maximum Pixel Count	66,355,200	17,694,720	8,847,360
Advanced Professional Features	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise or vCS
ISV Certifications	✓		
NVIDIA CUDA/OpenCL	✓		✓
Graphics Features and APIs	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise or vCS
NVENC	✓	✓	✓

OpenGL Extensions (WebGL)	✓	✓	
Insitu Graphics/GL Support			✓
RTX Platform Optimizations	✓		
DirectX	✓	✓	
Vulkan Support	✓		✓
Profiles	NVIDIA RTX vWS	NVIDIA vPC	NVIDIA AI Enterprise or vCS
Max Frame Buffer Supported	48GB	2GB	80GB
Available Profiles	0Q, 1Q, 2Q, 3Q, 4Q, 6Q, 8Q, 12Q, 16Q, 24Q, 32Q, 48Q	0B, 1B, 2B	4C, 5C, 6C, 8C, 10C, 12C, 16C, 20C, 24C, 40C, 80C

Refer to the respective sizing guides for [NVIDIA Virtual PC \(vPC\)](#), [NVIDIA RTX Virtual Workstation \(vWS\)](#) and [NVIDIA AI Enterprise](#) for additional details on how to best configure an accelerated virtualized infrastructure.

NVIDIA GPUs Recommended for Virtualization

Table 4 shows the [NVIDIA GPUs recommended for virtualization](#) workloads. The GPUs in this table are tested and supported with NVIDIA virtual GPU software. Refer to the NVIDIA virtual GPU [product documentation](#) for the full support matrix details.

Table 4. NVIDIA GPUs Recommended for Virtualization

	A100	A30	A40	A16
GPUs/Board (Architecture)	1 (Ampere)	1 (Ampere)	1 (Ampere)	4 (Ampere)
RTX Technology	--	--	✓	✓
Memory Size	40/80GB HBM2	24GB HBM2	48GB GDDR6	64GB GDDR6 (16GB per GPU)
vGPU Profiles (GB)	4, 5, 8, 10, 16, 20, 40, 80	4, 6, 8, 12, 24	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	1, 2, 4, 8, 16
MIG Support	Up to 7	Up to 4	No	No
NVLink Support	Yes	Yes	Yes	No
Form Factor	SXM4 and PCIe 4.0 Dual Slot	PCIe 4.0 Dual Slot	PCIe 4.0 Dual Slot	PCIe 4.0 Dual Slot
Power (W)	400/250	165	300	250
Thermal	passive	passive	passive	passive
Optimized For	performance	performance	performance	density
Target Workloads	Highest Performance AI and HPC	Mainstream AI and HPC	High-end Virtual Workstations or mixed virtual workstations and compute (AI, data science)	Knowledge Worker Virtual Desktops, Entry Virtual Workstations

The NVIDIA GPUs recommended for virtualization are divided into two categories:

- ▶ Performance Optimized GPUs are typically recommended for high-end virtual workstations running professional visualization applications, or for running compute-intensive workloads such as artificial intelligence, deep learning, or data science workloads.
- ▶ Density Optimized GPUs are typically recommended for knowledge worker virtual desktop infrastructure (VDI) to run office productivity applications, streaming video and Windows. They are designed to maximize the number of VDI users supported in a server.

Built on the innovative [NVIDIA RTX platform](#), the NVIDIA A40 GPU is uniquely positioned to power the most demanding professional visualization workloads. NVIDIA A40 is an integral part of the NVIDIA EGX Platform for Professional Visualization, which can run various accelerated graphics workloads including powerful virtual workstations. The NVIDIA A40 provides the best performance for graphics workloads and features larger (48GB) memory, but with NVIDIA® [NVLink](#)® can support up to 96GB to power virtual workstations that support very large animations, files, or models.

The NVIDIA A100 is the most advanced data center GPU ever built to accelerate AI, high performance computing, and data science. Customers who train or use neural networks, use computationally intensive applications, or run simulations requiring double precision accuracy (FP64 performance) should be using the A100, which provides the best time-to-solution. A100 is available in two form factors, PCIe and SXM module. The SXM module is available with servers that support NVIDIA® [NVLink](#)®, provide the best performance and strong scaling for hyperscale and HPC data centers running applications that scale to multiple GPUs, such as deep learning.

Selecting the Right GPU

While many organizations seek the highest performing GPU or the GPU that provides the best performance per dollar, there are other factors like performance per watt or form-factor that can be taken into consideration.

Workloads have been executed on an industry standard dual socket server with VMware vSphere 7 U2 and NVIDIA vGPU 13.0 using vGPU 1:1 profile unless otherwise stated. 1:1 vGPU profiles correspond to the full GPU allocated to a single virtual machine. This was chosen as the impact of scaling does not differ between GPUs². See “Impact of GPU Sharing” section for more details.

Note that the comparisons should be used as general guidance when choosing GPUs based on performance or performance per dollar. All recommendations are based on the workloads listed in Table 5 which could differ from the applications being used in production.

² Assumes that enough frame buffer is available on all vGPUs across all GPUs.

Table 5. Description of Benchmarks Used

Workload	Description	vGPU Software Edition
Professional Graphics	<p>SPECviewperf 2020 (4K)</p> <p>The SPECviewperf 2020 is a standard benchmark for measuring graphics performance based on professional applications. The benchmark measures the 3D graphics performance of systems running under the OpenGL and Direct X application programming interfaces.</p>	NVIDIA RTX vWS
AI Deep Learning Training	<p>BERT Large Fine Tune Training, Batch Size: 32* number of GPUs; Precision: Mixed; Data=Real; Sequence Length: 384; cuDNN Version: 8.1.0.77; NCCL Version: 2.8.4; Baseline: DL 21.02; Installation Source: NGC</p> <p>BERT is one of today's most widely used natural language processing models.</p>	NVIDIA AI Enterprise
AI Deep Learning Inference	<p>BERT Large Inference; Batch Size: 128 INT8; Seq-Len: 128</p> <p>BERT is one of today's most widely used natural language processing models.</p>	NVIDIA AI Enterprise
Knowledge Worker	<p>NVIDIA nVector Digital Worker Workload</p> <p>NVIDIA's nVector benchmarking tool that simulates the end user workflow and measures key aspects of the user experience, including end-user latency, framerate, image quality and resource utilization.</p>	NVIDIA vPC

Professional Graphics

The NVIDIA A40 is based on the NVIDIA Ampere™ architecture, which features second-generation RT Cores to deliver massive speedups for workloads like photorealistic rendering of movie content, architectural design evaluations, and virtual prototyping of product designs. RT cores also speed up the rendering of ray-traced motion blur for faster results with greater visual accuracy and can simultaneously run ray tracing with either shading or denoising capabilities. The significantly higher power budget of the NVIDIA A40 enables it to provide the highest graphics performance.

Figure 1. RTX vWS SPECviewperf2020 Performance

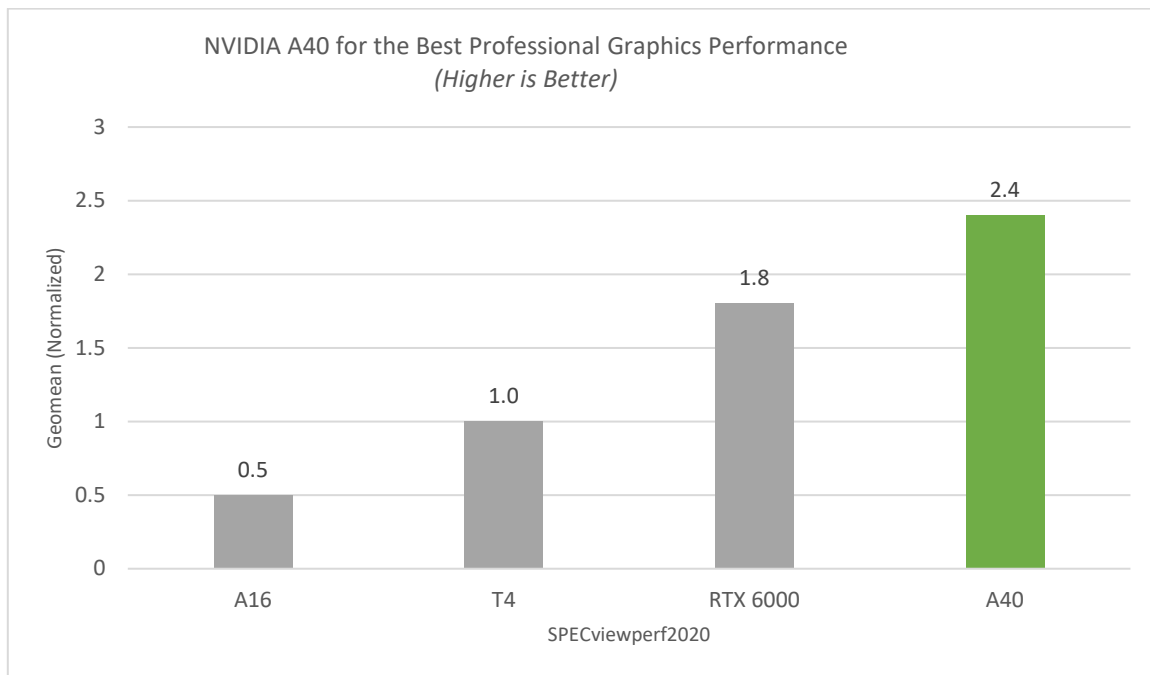


Figure 1 Test Configuration:	
Server CPU	Intel Xeon Gold 6154 (18C, 3.0GHz)
vGPU License	RTX vWS software
Hypervisor	VMware ESXi 7 U2
vGPU Driver Version (Host/Guest)	471.68
VM OS	Windows 10
VM vCPU	8 vCPU
VM vMemory	16 GB

Figure 2. RTX vWS SPECviewperf2020 Performance per Dollar

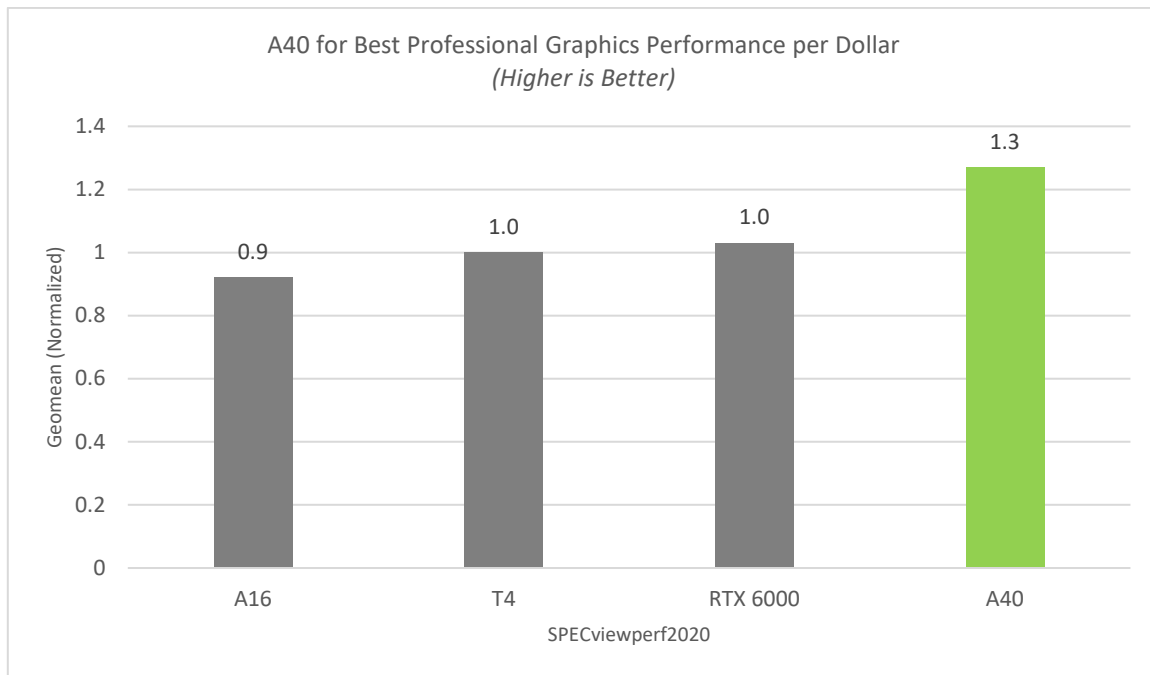


Figure 2 assumes estimated GPU street price plus NVIDIA RTX vWS software cost with 4-year subscription with one user per GPU.

AI Deep Learning Training

A100, based on the NVIDIA Ampere architecture, is designed to bring AI to every industry. The A100 is built to accelerate AI, and it is no surprise that it provides the highest performance for deep learning training workloads. It is important to note, for deep learning training workloads, time-to-solution is extremely important. For example, the cost of having highly paid data scientists wait for results could outweigh the benefits of a slightly lower cost solution, so A100 would be recommended when considering these other cost factors.

Figure 3. NVIDIA AI Enterprise Deep Learning Training Performance

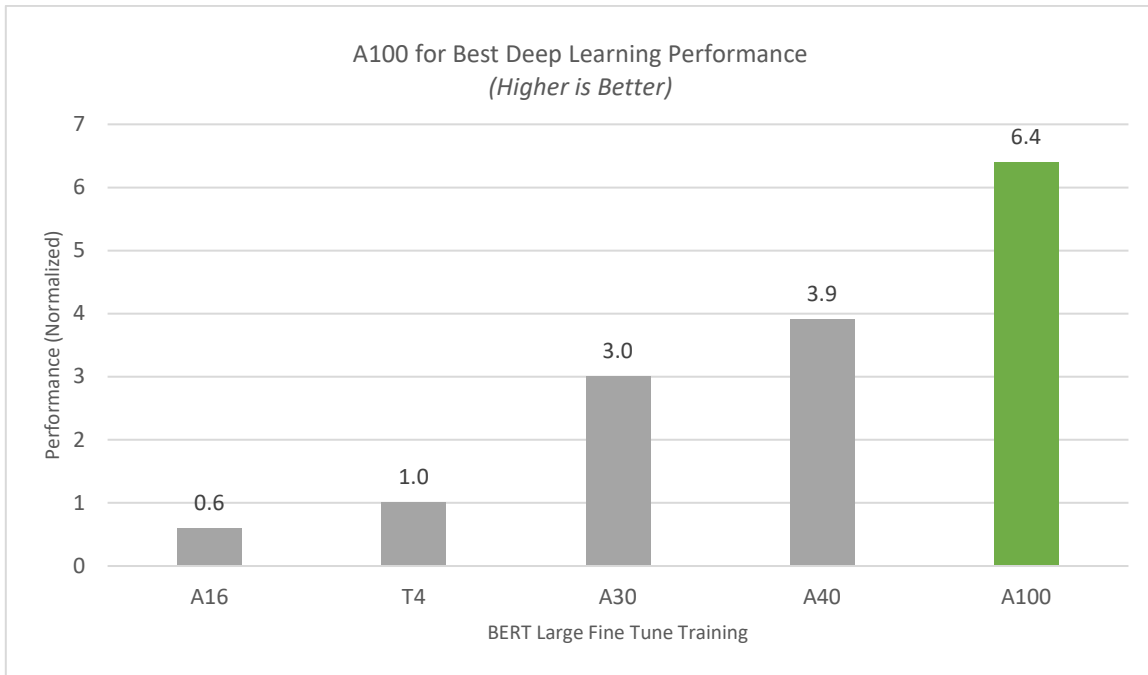


Figure 3 Test Configuration:	
Workload	BERT Large Fine Tune Training
Batch Size	32* number of GPUs
Precision	Mixed
Server CPU	AMD EPYC 7742
Hypervisor	ESXi 7 U2
vGPU Manager Version	vGPU 13.0

Figure 4. NVIDIA AI Enterprise Deep Learning Training Performance per Dollar

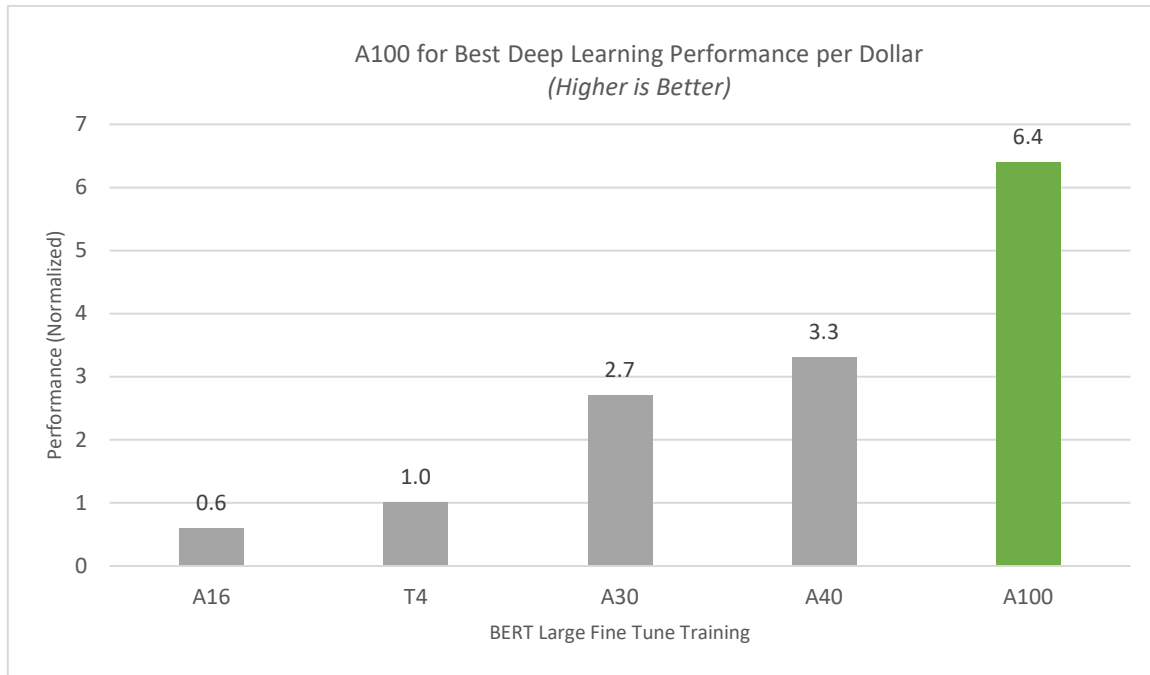


Figure 4 assumes estimated GPU street price plus NVIDIA AI Enterprise software cost with 3-year subscription.

AI Deep Learning Inference

For deep learning inference workloads, cost is often an important consideration. Therefore, the NVIDIA A30 is typically the preferred solution. Environments that prioritize performance as the most important consideration would select the A100. Note that both the A100 and the A30 GPUs support NVIDIA Multi-Instance GPU (MIG), which partitions the single A100 or A30 GPU into smaller, independent GPU instances which run simultaneously, each with its own memory, cache and streaming multiprocessors. MIG is not supported on other GPUs such as A16, A40 and T4.

Figure 5. NVIDIA AI Enterprise Deep Learning Inference Performance

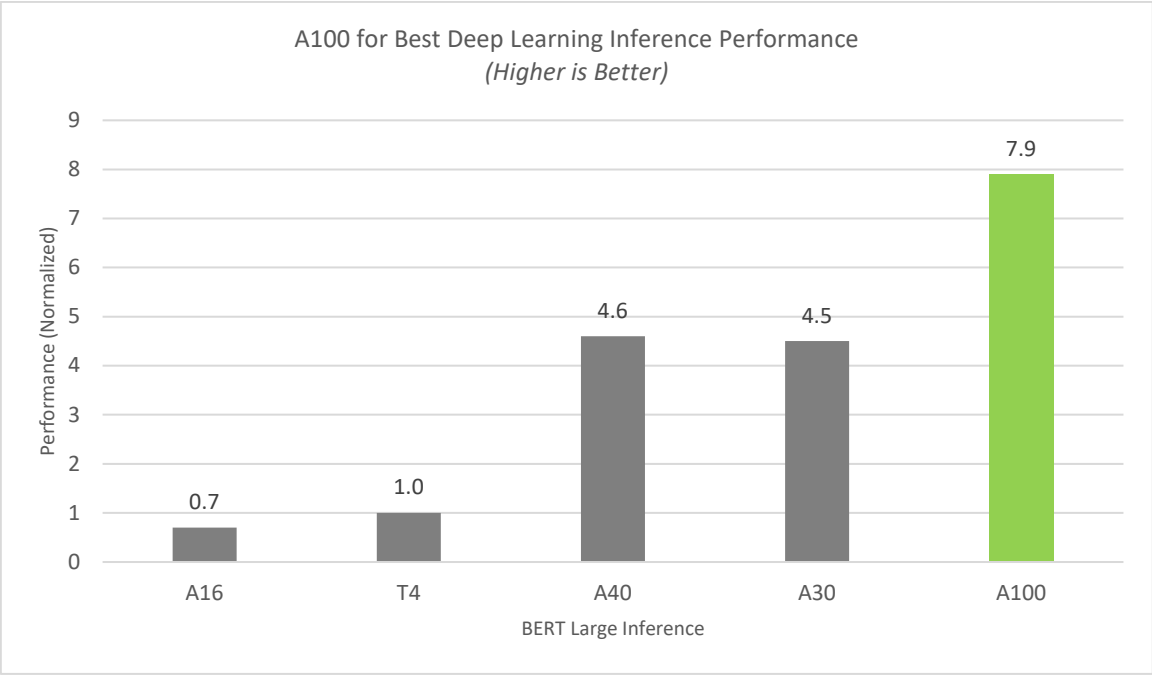


Figure 5 Test Configuration:	
Workload	BERT Large Inference
Batch Size	128
Integer Data Type	INT 8
Sequence Length	SEQ-Len 128
Precision	Mixed
Server CPU	AMD EPYC 7742
Hypervisor	ESXi 7
vGPU Manager Version	vGPU 13.0

Figure 6. NVIDIA AI Enterprise Deep Learning Inference Performance per Dollar

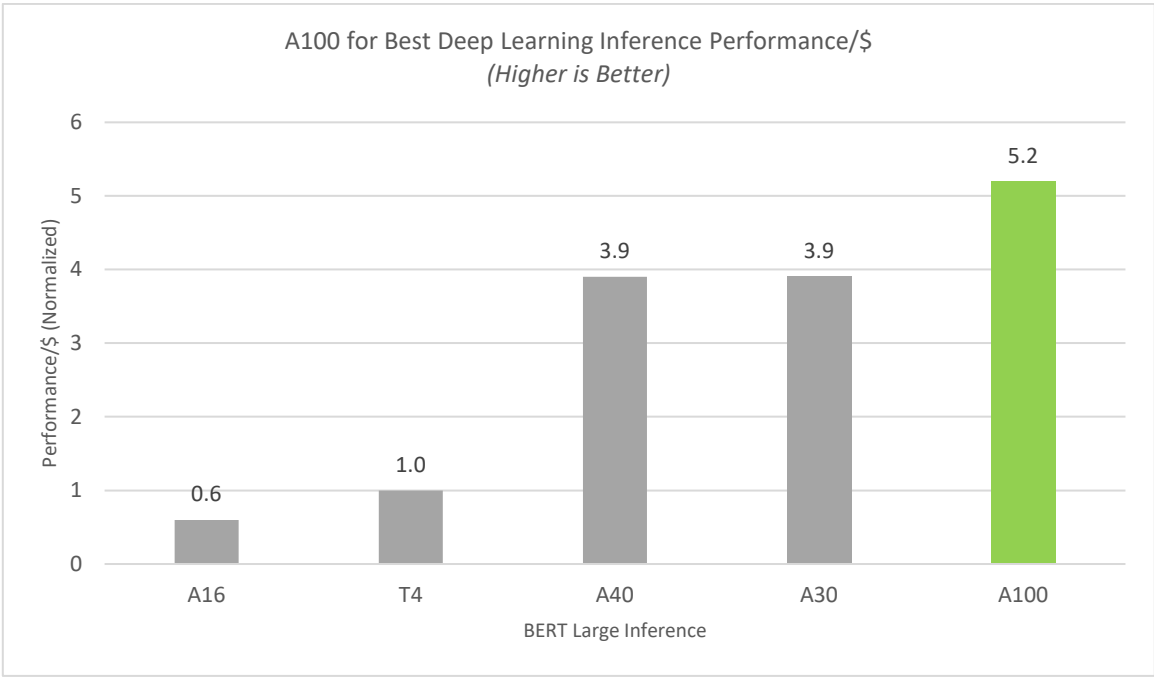


Figure 6 assumes estimated GPU street price plus NVIDIA AI Enterprise software cost with 3-year subscription.

Knowledge Workers

As more knowledge worker users are added on a server, the server runs out of CPU resources. Adding an NVIDIA GPU for this workload offloads constraints on the CPU resulting in improved user experience and performance for end users. The [NVIDIA nVector](#) knowledge worker VDI workload was used to test user experience and performance with NVIDIA GPUs combined with NVIDIA vPC software. NVIDIA M10, T4, A16 and A40 achieve similar performance for this workload.

For knowledge worker VDI workloads where users are accessing office productivity applications, web browsers and streaming video, achieving the highest user density per server and the best performance per dollar are important factors. The NVIDIA A16, with its unique quad-GPU on a board design is ideal for providing a high number of users per GPU and the most cost-effective performance for this use case.

Customers are realizing the benefits of increased resource utilization by leveraging common virtualized GPU accelerated server resources to run virtual desktops and workstations but leveraging these same resources to run compute when users are logged off. Customers who want to be able to run compute workloads on the same infrastructure that they run VDI, might leverage an A40 to do so. Learn more about [Using NVIDIA Virtual GPUs to Power Mixed Workloads](#) in our whitepaper.

Despite having 48GB of frame buffer, the A40 supports a maximum of only 32 users due to reaching the context switching limit per GPU. Refer to Table 6 to see how many VDI users can be supported for each GPU (with 1GB profile size).

Table 6. Maximum Number of Supported NVIDIA vPC Knowledge Workers (with 1GB Profile Size)

GPU	M10	T4	A40	A16
Max. Users per GPU board	32	16	32	64
Max. Users per server ³	96	96	96	192

Table 6 assumes that each user needs a 1GB profile. However, it is best to conduct a POC to determine the appropriate profile sizes for the users in the environment to provide the best user experience.

Figure 7. NVIDIA vPC VDI Cost per User

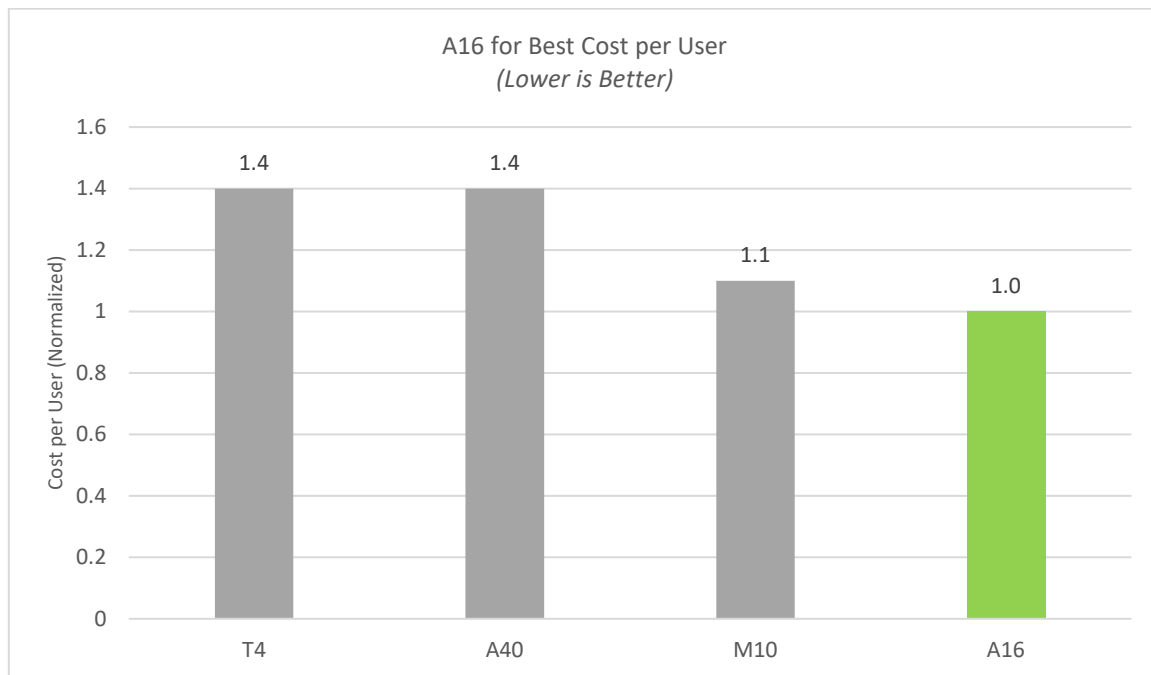


Figure 7 assumes estimated GPU street price plus NVIDIA vPC software cost with 4-year subscription divided by the number of users.

³ Assumes maximum of 3× M10, A40 or A16 boards per 2U server or up to 6× T4 boards per 2U server. See the specifications for your preferred OEM server to determine the maximum number of boards supported.

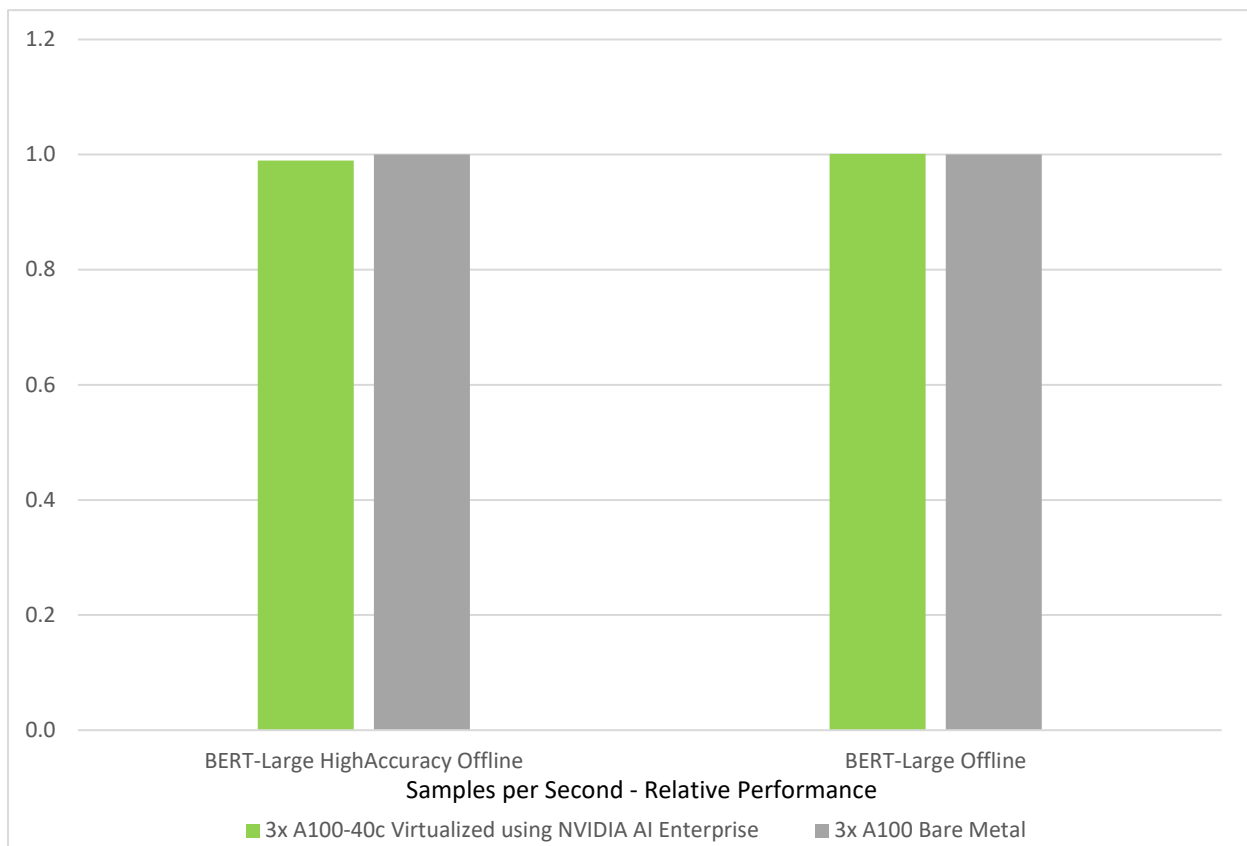
NVIDIA AI Enterprise vs. Bare Metal

Organizations choose to virtualize servers and applications for various reasons (manageability, flexibility, and security to name a few), but traditionally came with performance sacrifice. However, the performance difference of using NVIDIA AI Enterprise is negligible and will depend on the workload, as well as various other configuration variables. The following example illustrates near bare-metal performance with NVIDIA AI Enterprise in comparison to a bare metal server running the MLPerf Inference: Datacenter v1.1 Natural Language Processing (BERT-Large) benchmark in a 1:1 configuration using the NVIDIA A100 Tensor Core GPU.

Figure 9 represents per-accelerator performance derived from the best MLPerf results for respective submissions using the reported accelerator count in Data Center Offline. The server is a Dell EMC PowerEdge R7525, with 3 A100-PCIE-40GB GPUs (configured with 3 GRID A100-40C profiles), TensorRT, AMD EPYC 7502, NVIDIA A100-PCIE-40GB, TensorRT 8.0.2, CUDA 11.3

The MLPerf name and logo are trademarks. Refer to www.mlcommons.org for more information.

Figure 9. Inference Benchmark



Impact of GPU Sharing

Improving overall utilization through sharing a GPU across multiple virtual machines with NVIDIA vGPU software is implemented by scheduling the time which each virtual machine can use the GPU. NVIDIA vGPU software provides multiple GPU scheduling options to accommodate a variety of Quality of Service (QoS) levels for sharing the GPU. View the NVIDIA vGPU product documentation for more information about GPU scheduling options.

In general, the performance per virtual machine when sharing a GPU with n virtual machines will be $1/n$ of the total performance of the GPU. Therefore, two virtual machines sharing a GPU will result in approximately 50 percent of the overall performance per virtual machine and four virtual machines will result in approximately 25 percent of the overall performance per virtual machine.

Figure 10 is an illustration of multiple virtual machines with an overall throughput increase of 16%.

Figure 10. Virtual GPU Sharing

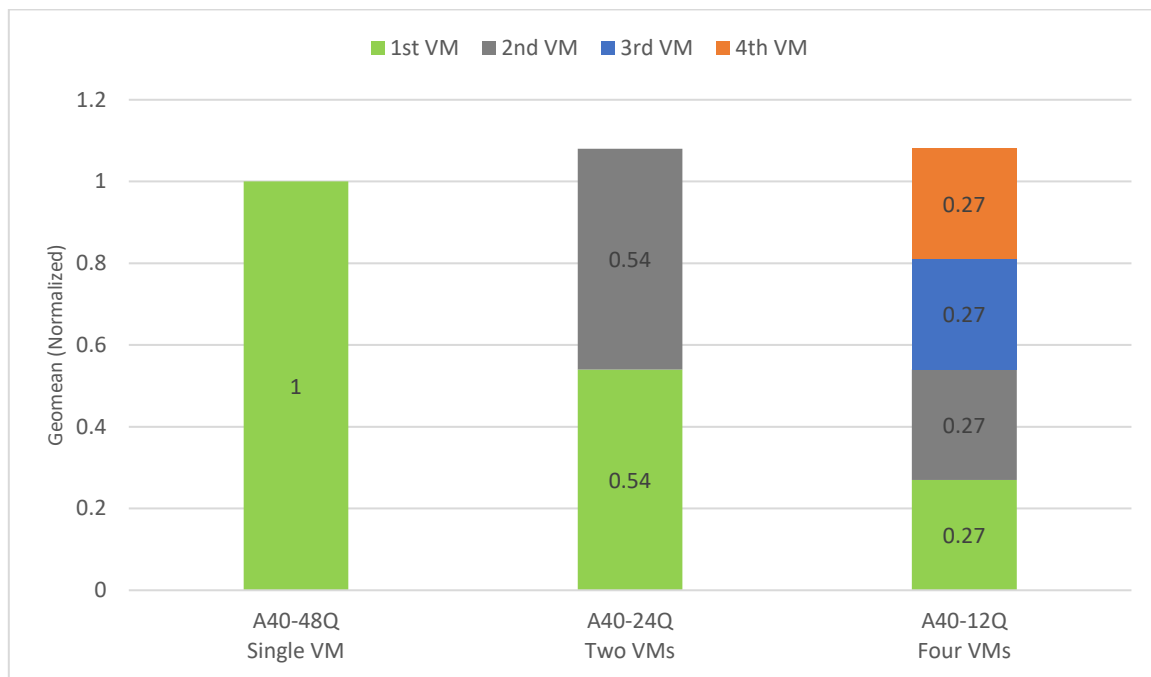
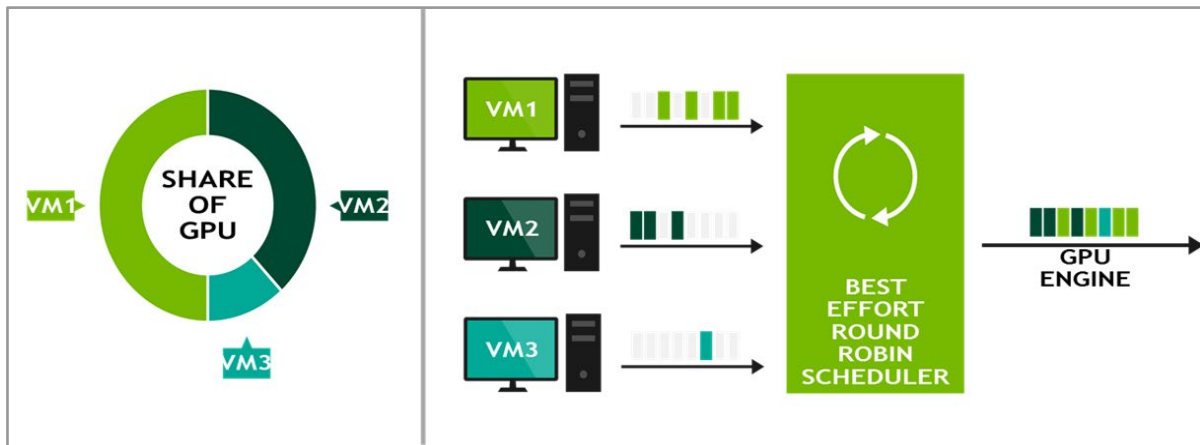


Figure 10 Test Configuration:	
Workload	Specviewperf 2020
Server CPU	Intel Xeon Gold (18C, 3.0GHz)
GPU Spec	RTX vWS with A40 with Equal Share scheduler
Hypervisor	VMware ESXi 7 U2
vGPU Driver Version (host, guest)	471.68
VM Spec	Windows 10, 8 vCPU, 16GB memory.

However, when workloads across virtual machines are not executed at the same time, or aren't always GPU bound, the performance can exceed the expected performance. The default GPU scheduling policy, "Best Effort," will be selected for this to happen as it leverages unused GPU time of other virtual machines. See Figure 11 for a simplified view of how the "Best Effort" GPU scheduler works.

Figure 11. Best Effort GPU Scheduler

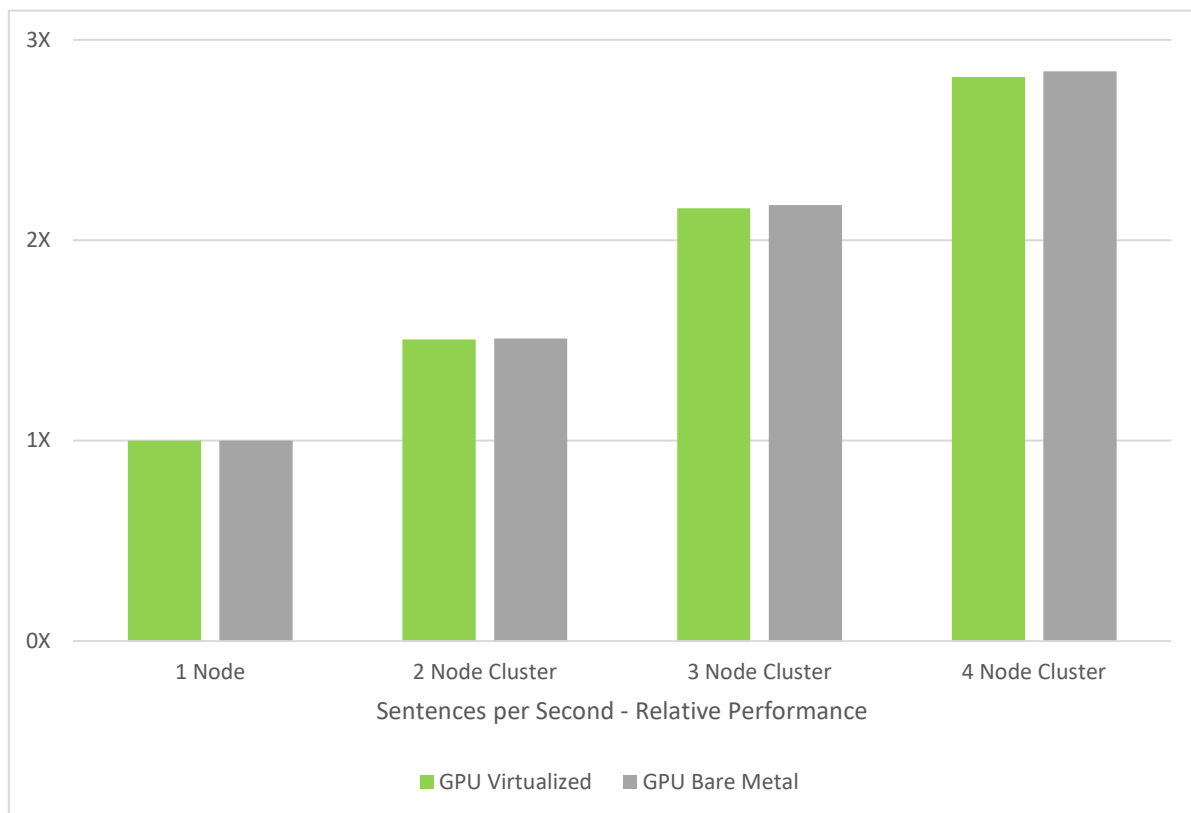


NVIDIA AI Enterprise Scale Out

The scaling factor of virtual machines with NVIDIA AI Enterprise distributed deep learning training is similar to the scaling factor for non-virtualized configurations. NVIDIA AI Enterprise is optimized on VMware vSphere and supports features such as GPUDirect RDMA, ATS (Address Translation Services), and RoCE (RDMA over Converged Ethernet) to provide multinode scale out performance that is nearly indistinguishable from bare metal.

Figure 12 represents a BERT-Large training workload to train a natural language processing model using NVIDIA A100 Tensor Core GPUs. With 1 GPU per node, the workload is scaled to 4 nodes with bare metal performance. The system is an Intel Xeon Gold (6240R @ 2.4GHz), Ubuntu 18.04, NVIDIA Mellanox ConnectX6 Dx, RoCE enabled, ATS enabled, TensorFlow BERT Large Training using Horovod, FP16, BS:30, Seq Len: 384, 1 NVIDIA A100 GPU per node, guest driver 460.32.04, GPU virtualized with VMware vSphere 7.0u2 and an NVIDIA vGPU 12.0 (40C profile).

Figure 12. NVIDIA AI Enterprise Distributed Deep Learning Training Performance



Conclusion

The NVIDIA vGPU software solution offers unmatched flexibility and performance when paired with the latest generation GPUs based on the NVIDIA 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 NVIDIA A40 GPU is uniquely positioned to power the most demanding graphics and rendering workloads for dynamic virtual workstations while also offering the most cost-effective performance for professional graphics applications. If the infrastructure will support knowledge worker VDI workloads, the A16 GPU provides the most cost-effective performance, while also providing the best user density. And finally, for AI workloads including deep learning training and deep learning inferencing, the NVIDIA A100 is the most advanced data center GPU ever built to accelerate AI, high-performance computing, and data science with unprecedented acceleration, while the NVIDIA A30 provides the most cost-effective performance for inferencing workloads.

While this technical brief provides general guidance on how to select the right NVIDIA GPU for your workload, actual results may vary depending on the specific application being virtualized.

The most successful deployments are those that balance virtual machine density (scalability) with required performance. This is achieved when a proof of concept (POC) with production workloads is conducted while analyzing the utilization of all resources of a system and gathering subjective feedback from all stakeholders. Consistently analyzing resource utilization and gathering subjective feedback allows for optimizing the configuration to meet the performance requirements while optimizing the configuration for best scale.

Resources Links

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](#)

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 products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

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 NON-INFRINGEMENT, 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.

OpenCL

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

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NVIDIA RTX, NVIDIA Ampere, NVIDIA Turing, NVIDIA Volta, 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.

Copyright

© 2021 NVIDIA Corporation. All rights reserved.