NVIDIA QUADRO VIRTUAL DATA CENTER WORKSTATION APPLICATION SIZING GUIDE FOR ESRI ARCGIS PRO

APPLICATION GUIDE

Ver 1.0

EXECUTIVE SUMMARY

This document provides insights into how to deploy NVIDIA® Quadro® Virtual Data Center Workstation (Quadro vDWS) software for Esri ArcGIS Pro users. Recommendations are based on actual customer deployments and benchmarking data and cover three common questions:

Which NVIDIA GPU should I use for my business needs?

How do I select the right profile(s) for the types of users I will have?

How many users can be supported (user density) per server?

Since user behavior varies and is a critical factor in determining the best GPU and profile size, recommendations are made for three user types along with two levels of quality of service (QoS) for each user type: Dedicated Performance and Typical Customer Deployment. User types are segmented as either light, medium or heavy based on type of workflow and the size of the model/data they are working with. Users with more advanced graphics requirements and using larger data sets are categorized as heavy users, for example. Light and medium users require less graphics and typically work with smaller model sizes. Recommendations for each of those users within each level of service along with the server configuration are shown below.

The vGPU profiles listed in the Dedicated Performance and Typical Customer Deployment tables on the next page were created by first understanding the graphic performance of a Quadro workstation GPU (for example, Quadro P2000). The benchmark scores of the physical workstation card were then aligned with the scores outputted for the virtual GPU. It is important to note; the Dedicated Performance table is based upon the Equal Share scheduler and does not oversubscribe the GPU compute engine, resulting in the same GPU performance at all times. Similar to vCPU to physical core oversubscription, many virtual GPUs can utilize the same physical GPU compute engine. The GPU compute engine can be oversubscribed by selecting the Best Effort GPU scheduler policy which best utilizes the GPU during idle and not fully utilized times. For many customer deployments, it is not typical that 12 users will be executing rendering requests simultaneously or even to the degree which were replicated in dedicated performance testing, therefore selecting the best effort scheduler often results in a 2X to 3X oversubscription of the GPU compute engine which results in two to three times the number of users. The degree to which higher scalability is achieved is dependent on the typical day to day activities of your users, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc. It is recommended that you test and validate the appropriate GPU scheduling policy to meet the needs of your users.

DEDICATED PERFORMANCE *

12 Users per Server	6 Users per Server	3 Users per Server
T4-8Q 8vCPU 8GB RAM	T4-16Q 10vCPU 16GB RAM	P40-24Q or RTX6000-24Q 12vCPU 16-32GB RAM
Light User	**Medium User**	**Heavy User**

**Please refer to Appendix B for additional information.*

TYPICAL CUSTOMER DEPLOYMENT

16-24 Users per Server	10-12 Users per Server	6-9 Users per Server
T4-4Q 6vCPU 6-10GB RAM	T4-8Q 8vCPU 10-16GB RAM	P40-8Q / P40-12Q or RTX6000-8Q/RTX6000-12Q 12vCPU+ >32GB RAM
Light User	**Medium User**	**Heavy User**

REFERENCE SERVER LAB BUILDS

6x NVIDIA T4 GPUs	3x Tesla P40 GPUs	3x Quadro RTX 6000 GPUs
2x Intel Xeon Gold 6154 128-512-768 GB RAM 10GbE Network Flash Based Storage	2x Intel Xeon Gold 6154 512-768+ GB RAM 10GbE Network Flash Based Storage	2x Intel Xeon Gold 6154 512-768+ GB RAM 10GbE Network Flash Based Storage
Light User	**Medium User**	**Heavy User**

Table 1. Esri ArcGIS Pro VDI Deployment Configurations

These recommendations are meant to be a guide. The most successful customer deployments start with a proof of concept and are "tuned" throughout the lifecycle of the deployment. Beginning with a POC enables customers to understand the expectations and behavior of their users and optimize their deployment for the best user density while maintaining required performance levels. Continued maintenance is important because user behavior can change over the course of a project and as the role of an individual changes in the organization. A GIS Analyst that was once a light graphics user might become a heavy graphics user when they change teams or are assigned a different project. Management and monitoring tools enable administrators and IT staff to ensure their deployment is optimized for each user.

ABOUT ESRI ARCGIS PRO

Esri builds the industry leading mapping and spatial analytics software for desktop, software as a service (SaaS) and enterprise applications. ArcGIS Pro is the latest generation of desktop GIS which supports every aspect of the GIS workflow, from data creation through 3D visualization and analysis, cartographic map production, as well as sharing geographic maps, data and analytical models. ArcGIS Pro provides an integrated toolset which users can then utilize to extend the reach of the GIS data throughout their organization using ArcGIS Enterprise. The version of ArcGIS Pro used in this sizing guide focuses on 3D visualization as well as spatial analytical tools which execute compute and deep learning inferencing within ArcGIS Pro 2.2.

Esri works closely with NVIDIA to certify the deployment of ArcGIS Pro in the cloud using VDI with NVIDIA Quadro vDWS software. VDI certification eliminates the need to install ArcGIS Pro on a local client, which can help reduce IT support and maintenance costs and enables greater mobility and collaboration. This virtual workstation deployment option enhances flexibility and further expands the wide variety of platform choices available to Esri customers.

ABOUT NVIDIA QUADRO VIRTUAL DATA CENTER WORKSTATION

NVIDIA virtual GPU (vGPU) software enables the delivery of graphics-rich virtual desktops and workstations accelerated by NVIDIA GPUs. NVIDIA Quadro vDWS software is based on NVIDIA virtual GPU technology and includes the Quadro graphics driver required by professional 3D applications. The Quadro vDWS license enables sharing the same NVIDIA GPU across multiple virtual machines running any application, so every virtualized user has access to an experience that ESRI has intended; only provided with NVIDIA Quadro.

NVIDIA Quadro is the world’s preeminent visual computing platform, trusted by millions of creative and technical professionals to accelerate their workflows. With Quadro vDWS software, you can deliver the most powerful virtual workstation from the data center. This frees your most innovative professionals to work from anywhere and on any device, with access to the familiar tools they trust. Certified with over 140 servers and supported by every major public cloud vendor, Quadro vDWS is the industry standard for virtualized enterprises.

To deploy an NVIDIA vGPU solution for Esri ArcGIS Pro, you will need NVIDIA GPUs and a Quadro vDWS software license for each user.

ESRI ARCGIS PRO BENCHMARK AND METRICS

To determine the optimal configuration of Quadro vDWS for Esri’s ArcGIS Pro, both user performance and scalability were considered. The datasets used within the Esri ArcGIS Pro benchmarks are considered less complex and are smaller size (less than 10GB) and focus on three GPU accelerated workflows which include data visualization and spatial analytics.

1. 3D Multi-Patch Rendering - Source: Esri Philadelphia dataset

The Esri PerfTools Add-in for ArcGIS Pro allows gathering of rendering metrics during benchmarking. Esri provided their “3D Philly” dataset and project file which contained 10 spatial bookmarks. PerfTools scripting ran the application through 10 spatial bookmarks with a 10 second think time. This allows the benchmark to more closely mimic a user pause between interactions within the application. An automated scripting framework enabled the test to be scaled out to multiple virtual desktops. Esri applies a frame rate limiter for ArcGIS Pro. By default, the application limits frames to 60 frames per second. This application sizing guide uses the default render settings for ArcGIS Pro which includes utilizing the DirectX rendering engine.

The following test metrics are outputted from the PerfTools add-in for each virtual machine and then analyzed:

- Draw Time Sum: The total time elapsed for all of the benchmarks to fully draw. Less time would be a better user experience (UX) and more would be a worsening UX.
- Frames Per Second (FPS): Esri stated that 30FPS is what most users perceive as a good UX, 60 is optimal but most users do not see a significant difference.

- FPS Minimum: Esri stated that a drop below 5-10FPS would appear to an end user that the drawing had stopped or “frozen”.
- Standard Deviation: This would represent the number of tests that were outside the average of the others, typically representing a faulty test or that scalability thresholds have been exceeded. Values should be < 4 for 3D workloads.

2. Deep Learning Inferencing - Source: Kolovai, Tonga OpenAerialMap dataset

Deep Learning Inferencing tests execute the Detect Objects Using Deep Learning tool which is available using Image Analyst license. The test dataset is from the OpenAerialMap and is based upon Esri’s Learn ArcGIS Deep Learning lesson. Test metrics were outputted as Total Execution Time within the ArcGIS Pro output window. The tests used TensorFlow which was configured to run on the GPU. For more information on how to configure TensorFlow to run on the GPU, please refer to Appendix C.

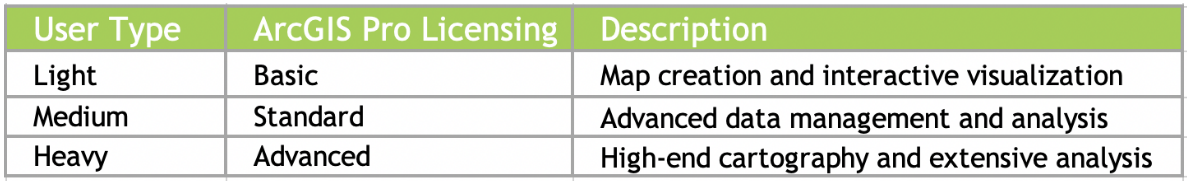
3. Spatial Analysis (CUDA Compute) – Source: California– Shuttle Radar Topography Mission NASA L

Spatial Analysis tests execute Calculate ViewShed2 tool which is available using the Spatial Analyst license. The test dataset was downloaded from the NASA Shuttle Rader Topography Mission website and was geographically exclusion to the State of California. Test metrics were outputted as Total Execution Time within the ArcGIS Pro output window. The tests used CUDA to execute the processing on the GPU. For more information on how to configure 3D Analyst tools to run on the GPU, please refer to Appendix D.

The GPU Profiler was used as a tool for evaluating GPU/CPU utilization rates during each of the three aforementioned benchmarks. ArcGIS Pro demonstrates a great balance between CPU and GPU resources.

ESRI USER TYPES AND ARCGIS PRO LICENSING

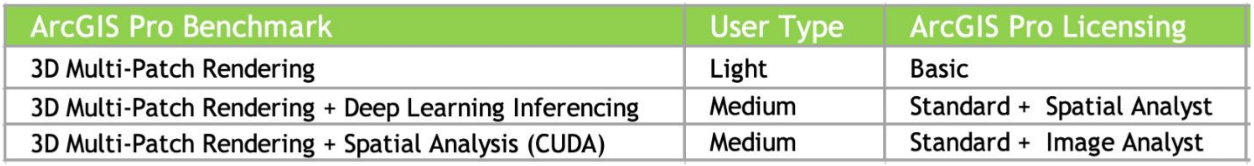
GIS Administrators often segment their end users into user types for each application and bundle similar user types on a host. The following table describes ArcGIS Pro user types used for this sizing guide and aligns user types to ArcGIS Pro licensing and descriptions published by Esri.



User Type	ArcGIS Pro Licensing	Description
Light	Basic	Map creation and interactive visualization
Medium	Standard	Advanced data management and analysis
Heavy	Advanced	High-end cartography and extensive analysis

Table 7. Common user types for ArcGIS Pro and application licensing description

The following table aligns the three Esri ArcGIS Pro benchmarks to user type and ArcGIS Pro licensing.



ArcGIS Pro Benchmark	User Type	ArcGIS Pro Licensing
3D Multi-Patch Rendering	Light	Basic
3D Multi-Patch Rendering + Deep Learning Inferencing	Medium	Standard + Spatial Analyst
3D Multi-Patch Rendering + Spatial Analysis (CUDA)	Medium	Standard + Image Analyst

FINDINGS

To determine the optimal configuration of Quadro vDWS for Esri ArcGIS Pro, both user performance and scalability were considered based on data from industry benchmarks as well as insights from customer best practices.

1. Benchmarking based on the three Esri ArcGIS Pro benchmarks.
2. Documenting customer best practices using Esri ArcGIS Pro with Quadro vDWS

The following tables summarize the recommended configurations based on benchmarking data and customer best practices. These recommendations take into account the performance requirements for different user types as well as optimizing for scale, or user density, on the server to achieve the best total cost of ownership. The performance of the equivalent physical Quadro workstation card was also measured and then analyzed. A 10% threshold was used to align the equivalent physical Quadro workstation card with the reported VDI performance score.

The dedicated performance table illustrates recommendations based upon the fixed share scheduler, which provides the most consistent dedicated performance at all times. However, most customer deployments typically select the best effort GPU scheduler policy to achieve better utilization of the GPU, which usually results in supporting more users per server and better TCO per user. It is important to keep the scheduling policy options in mind when comparing the two tables to one another.

For more on the GPU scheduling options, refer to Deployment Best Practices, Section 5 below.

UNDERSTANDING THE GPU SCHEDULER

DEDICATED PERFORMANCE

User Type	Equivalent Performance Level +/-10%	Users per Server	vCPUs	vGPU Profile	vMemory	CPUs	GPUs	Memory	Storage Type	Networking
Light	**Quadro P2200**	12	8	T4-8Q	8GB	2x Intel Xeon Gold 6154	6x T4	128GB	Flash-Based	10GbE
Medium	**Quadro P4000**	6	10	T4-16Q	10-16 GB	2x Intel Xeon Gold 6154	6x T4	128GB	Flash-Based	10GbE
Heavy	**Quadro P6000**	3	12	P40-24Q	16-32 GB	2x Intel Xeon Gold 6154	3x P40	128GB	Flash-Based	10GbE
	Quadro RTX 6000	3	12	RTX6000-24Q	32-64GB GB	2x Intel Xeon Gold 6154	3x RTX 6000	128GB	Flash-Based	10GbE

TYPICAL CUSTOMER DEPLOYMENT

User Type	Equivalent Performance Level+/-10%	Users per Server	vCPUs	vGPU Profile	vMemory	CPUs	GPUs	Memory	Storage Type	Networking
Light	**Quadro P2200**	16 – 24	6	T4-4Q	8-10 GB	2x Intel Xeon Gold 6154	6x T4	384GB	Flash-Based	10GbE
Medium	**Quadro P4000**	10 – 12	8	T4-8Q	10-16 GB	2x Intel Xeon Gold 6154	6x T4	384-512GB	Flash-Based	10GbE
Heavy	**Quadro P6000**	6 - 9	8-12	P40-8Q P40-12Q	16-32 GB	2x Intel Xeon Gold 6154	3x P40	384-1TB	Flash-Based	10GbE
	Quadro RTX 6000	6 - 9	8-12	RTX6000-8Q RTX6000-12Q	32-64 GB	2x Intel Xeon Gold 6154	3x RTX 6000	384-1TB	Flash-Based	10GbE

Table 2. Server configurations of ArcGIS Pro in a VDI environment.

NVIDIA T4 WITH QUADRO vDWS FOR LIGHT TO MEDIUM USERS

Quadro vDWS combined with NVIDIA T4 is recommended for virtualizing Esri ArcGIS Pro. The T4 GPU performance is in line with commonly used Quadro GPUs, such as the Quadro P4000, used in physical workstations for ArcGIS Pro. The NVIDIA T4 taps into industry leading NVIDIA Turing architecture which sets a new bar for power efficiency and performance. The T4 features both multi-precision Tensor and RT Cores which equates to ground breaking performance for vGPU customers. When compared with the P4, the T4 offers double the frame buffer which enables professional users to work with even larger model sizes and provides about a 25% performance improvement versus the previous generation P4. The T4 adds support for new features like VP9 and H.265 which can be used for video play-back.

The T4 GPU is a single width, half height form factor and requires less power than other GPUs, allowing it to be powered via the standard PCIe bus. This results in a high-density solution accommodating up to six T4 GPUs per 2 rack unit server. Esri ArcGIS Pro benchmark results show that six T4 GPUs in a server configured with two Intel Xeon Gold 6154 CPUs is a well-balanced configuration for ArcGIS Pro.

Based on benchmark results, there are enough CPU resources available to host six T4 GPUs in a single 2 rack unit (RU), 2-socket server running Esri ArcGIS Pro on 12 virtual machines.

QUADRO RTX 6000 WITH QUADRO vDWS FOR HEAVY USERS

Quadro vDWS combined with Quadro RTX™ 6000 is the recommended GPU for heavy users that requires additional performance over a T4. These GIS users work with high-end cartography and execute extensive analysis (compute) workloads which involves inferencing on large datasets. The RTX 6000 is based upon the same NVIDIA Turing architecture as the T4 but it offers 1.8 times the amount of NVIDIA CUDA and Tensor cores. The RTX 6000 is a dual slot card, which allows for up to three GPUs to be installed in many 2RU, 2-socket servers.

TESLA P40 WITH QUADRO vDWS FOR HEAVY USERS

Quadro vDWS combined with NVIDIA Tesla® P40 is an alternative choice for users that require more graphics and compute acceleration than the T4 due to the additional performance and frame buffer of this GPU. It is also sufficient for heavy GIS users but is based upon the NVIDIA Pascal architecture, a predecessor to NVIDIA Turing architecture. Like the RTX6000, the P40 GPU can be recommended for heavy users that require the additional performance of a Tesla P40 over a T4, but the P40 does not offer the enhanced Turing generation benefits. The Tesla P40 is a dual slot card, which allows for up to three GPUs to be installed many 2RU, 2-socket servers.

SERVER RECOMMENDATION: DUAL SOCKET, 2U RACK SERVER

A 2RU, 2-socket server configured with two Intel Xeon Gold 6154 processors is recommended. With a high-frequency 3.0 GHz combined with 18-cores, this CPU is well-suited for optimal performance for each end user while supporting the highest user scale, making it a cost-effective solution for Esri ArcGIS Pro.

SUFFICIENT SYSTEM MEMORY FOR EACH INDIVIDUAL USER

While Esri ArcGIS Pro performs optimally with 8 GB of system memory for each virtual machine, customers typically assign 10 - 16 GB of system memory to medium users for optimal performance. System memory requirements don't change with the transition to virtual workstations powered by Quadro vDWS, therefore the same amount of system memory that is used in a physical workstation should be assigned to the Quadro vDWS accelerated virtual machine.

FLASH BASED STORAGE FOR BEST PERFORMANCE

The use of flash-based storage, such as solid-state drives (SSDs) are recommended for optimal performance. Flash-based storage is the common choice with ArcGIS Pro users on physical workstations and similar performance can be achieved in similarly configured virtual environments.

A typical configuration for non-persistent virtual machines is to use the direct attached storage (DAS) on the server in a RAID 5 or RAID 10 configuration. For persistent virtual machines, a high performing all-flash storage solution is the preferred option.

TYPICAL NETWORKING CONFIGURATION FOR QUADRO vDWS

There is no typical network configuration for in a Quadro vDWS powered virtual environment since this varies based on multiple factors including choice of hypervisor, persistent versus non-persistent virtual machines, and choice of storage solution. Most customers are using 10 GbE networking for optimal performance.

OPTIMIZING FOR DEDICATED QUALITY OF SERVICE

For comparative purposes, we also considered the requirements for a configuration optimized for performance only. This configuration option does not take into account the need to also optimize for scale, or user density. Additionally, this configuration option is based solely on performance results using the aforementioned rendering and compute benchmarks.

As with the recommended best practice, which is optimized for both performance and user density, NVIDIA T4 is recommended for both light and medium users. For heavy users, the Tesla P40 or RTX 6000 is recommended. We also recommend that a larger profile size be used, a T4-8Q for light users, T4-16Q for medium users and P40-24Q or RTX6000-24Q for heavy users. As a result, fewer users can be supported on each server. If only performance is important, it is recommended that the fixed share scheduler is utilized.

This configuration for “performance-only” is based on running ArcGIS Pro across all virtual machines since it shows the impact of a peak workload on all resources of the server, including CPU, memory, GPU, and network, to best architect the solution. The dedicated performance data in this application sizing guide shows benchmarks running at scale.

Tests are simultaneously executed on all virtual machines with minimal pauses or idle time. This workflow is not typical in a true production environment but provides a methodology for assessing dedicated performance during these worst-case scenarios.

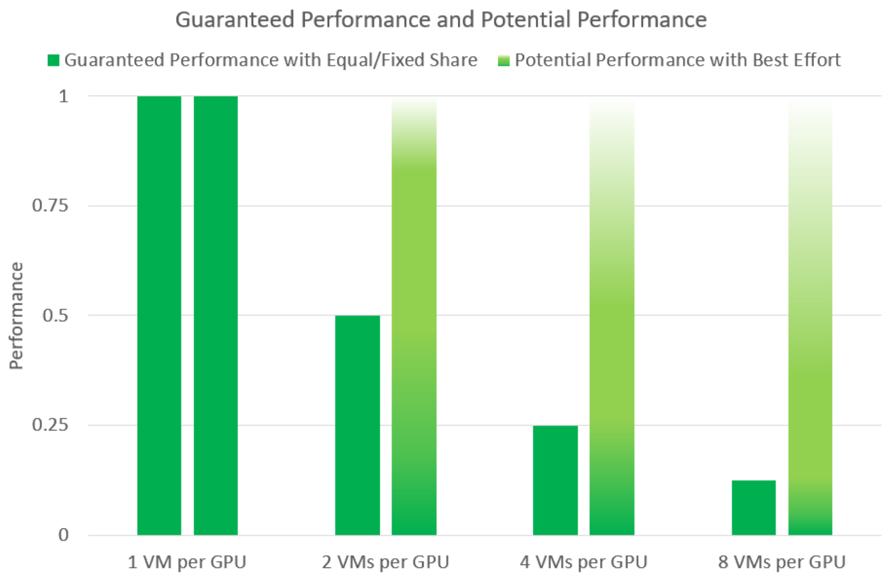


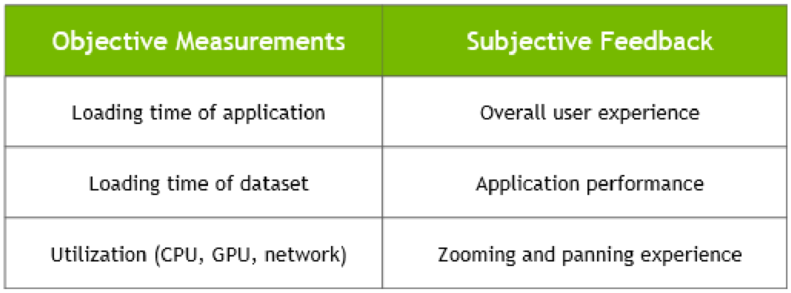
Table 5. Comparison of VMs per GPU performance utilization based on Dedicated Performance vs. Best Effort Configurations

DEPLOYMENT BEST PRACTICES

1. RUN A PROOF OF CONCEPT

The most successful deployments are those that balance user density (scalability) with performance. This is achieved when Quadro vDWS-powered virtual machines are used in production while objective measurements and subjective feedback from end users is gathered.

We highly recommend a proof of concept (POC) is run prior to doing a full deployment to provide a better understanding of how your users work and how many GPU resources they really need, analyzing the utilization of all resources, both physical and virtual. Consistently analyzing resource utilization and gathering subjective feedback allows for optimizing the configuration to meet the performance requirements of end users while optimizing the configuration for best scale.



Objective Measurements	Subjective Feedback
Loading time of application	Overall user experience
Loading time of dataset	Application performance
Utilization (CPU, GPU, network)	Zooming and panning experience

Table 6. Example metrics for a successful POC

2. LEVERAGE MANAGEMENT AND MONITORING TOOLS

Quadro vDWS software on NVIDIA GPUs provides extensive monitoring features enabling IT to better understand usage of the various engines of an NVIDIA GPU. The utilization of the compute engine, the frame buffer, the encoder, and decoder can all be monitored and logged through a command line interface called the NVIDIA System Management Interface (nvidia-smi), accessed on the hypervisor or within the virtual machine. In addition, NVIDIA vGPU metrics are integrated with Windows Performance Monitor (PerfMon) and through management packs like VMware vRealize Operations.

To identify bottlenecks of individual end users or of the physical GPU serving multiple end users, execute the following nvidia-smi commands on the hypervisor.

Virtual Machine Frame Buffer Utilization:

```
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Total" -e "Used" -e "Free"
```

Virtual Machine GPU, Encoder and Decoder Utilization:

```
nvidia-smi vgpu -q -l 5 | grep -e "VM ID" -e "VM Name" -e "Utilization" -e "Gpu" -e "Encoder" -e "Decoder"
```

Physical GPU, Encoder and Decoder Utilization:

```
nvidia-smi -q -d UTILIZATION -l 5 | grep -v -e "Duration" -e "Number" -e "Max" -e "Min" -e "Avg" -e "Memory" -e "ENC" -e "DEC" -e "Samples"
```

3. UNDERSTAND YOUR USERS

Another benefit of performing a POC prior to deployment is that it enables more accurate categorization of user behavior and GPU requirements for each virtual workstation. Customers often segment their end users into user types for each application and bundle similar user types on a host. Light users can be supported on a smaller GPU and smaller profile size while heavy users require more GPU resources, a large profile size and, may be best supported on a larger GPU.

4. USE OF STANDARD BENCHMARKS

Benchmarks, like the three which were used for Esri ArcGIS Pro, can be used to help size a deployment but they have some limitations. Benchmarks simulate peak workloads, when there is the highest demand for GPU resources across all virtual machines. The benchmark doesn’t account for the times when the system isn’t fully utilized, or which hypervisors, and the best effort scheduling policy to leverage to achieve higher user densities with consistent performance.

The graphic below demonstrates how workflows processed by end users are typically interactive, which means there are multiple short idle breaks when users require less performance and resources from the hypervisor and NVIDIA vGPU.

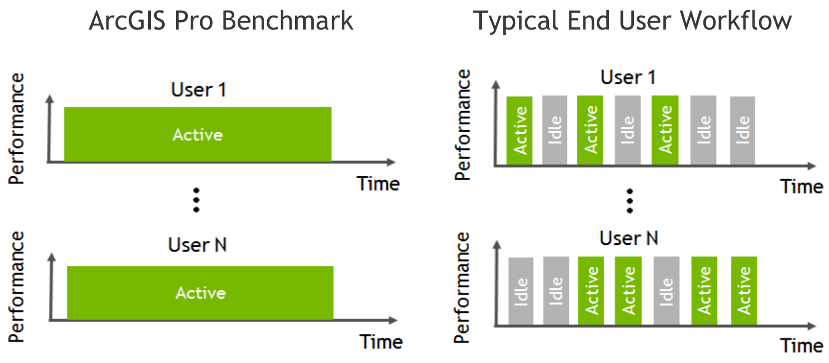


Table 8. Comparison of the Esri ArcGIS Pro benchmarks utilization versus a typical end user workflow

NVIDIA used a custom-designed benchmarking engine to conduct vGPU testing at scale. This benchmarking engine automates the testing process from provisioning virtual machines, establishing remote connections, executing the benchmark, and analyzing the results across all virtual machines. Dedicated performance scores mentioned in this application guide are based on benchmark data which was run in parallel on all virtual machines with scores averaged across three runs.

5. UNDERSTANDING THE GPU SCHEDULER

NVIDIA Quadro vDWS provides three GPU scheduling options to accommodate a variety of QoS requirements of customers.

1) **Fixed share scheduling** guarantees the same dedicated quality of service at all times.
2) **Best effort scheduling**[1] provides consistent performance at a higher scale and therefore reduces the TCO per user.
3) **Equal share scheduling** provides equal GPU resources to each running VM. As vGPUs are added or removed, the share of GPU processing cycles allocated changes accordingly, resulting in performance to increase when utilization is low, and decrease when utilization is high.

Organizations typically select the best effort GPU scheduler policy for their deployment to achieve better utilization of the GPU, which usually results in supporting more users per server with a lower quality of service (QoS) and better TCO per user.

The below example demonstrates the different numbers of users per server that can be reached by applying different QoS thresholds via GPU Scheduling policies. Choosing the Fixed Share Scheduler always guarantees a particular QoS. In this example, two users on a T4 will always experience performance similar to a workstation with Quadro P2200 GPU. Using the Best Effort Scheduler, which is the most commonly chosen GPU scheduling option for enterprises and does not provide the same level of QoS, could allow more users to experience a Quadro P2200 level performance but user performance will vary depending on load from other users on the same T4 at any given time. A single user on a T4 will experience performance similar to a Quadro P4000 but as density increases to 3-4 users per GPU, the performance can be similar to a workstation with a Quadro P620 card. The below example assumes sufficient frame buffer at all scales to demonstrate options on how GPU scheduling policies can impact scale.

[1] Available since 2013 when NVIDIA virtual GPU technology was first introduced

	Dedicated Performance (Fixed Share scheduler)	Typical Customer Configuration (Best Effort Scheduler)
Users/Server Host (6 x NVIDIA T4)	12 (2 users per GPU with the performance of P2000 at all times)	16-24 (3 - 4 users per GPU with the performance of P620-P4000)

Table 9. T4 user density with Fixed Share Scheduler versus Best Effort Scheduler

The **fixed share scheduling** policies guarantee equal GPU performance across all vGPUs sharing the same physical GPU. Dedicated quality of service simplifies a POC since it allows the use of common benchmarks used to measure physical workstation performance such as SPECviewperf, to compare the performance with current physical or virtual workstations.

The **best effort scheduler** leverages a round-robin scheduling algorithm which shares GPU resources based on actual demand which results in optimal utilization of resources. This results in consistent performance with optimized user density. The best effort scheduling policy best utilizes the GPU during idle and not fully utilized times, allowing for optimized density and a good QoS.

The table below shows that when using the best effort GPU scheduling policy, performance for an individual user that shares a GPU with other users can be as good as having a dedicated GPU, if the other end users aren’t executing GPU intensive tasks in parallel.

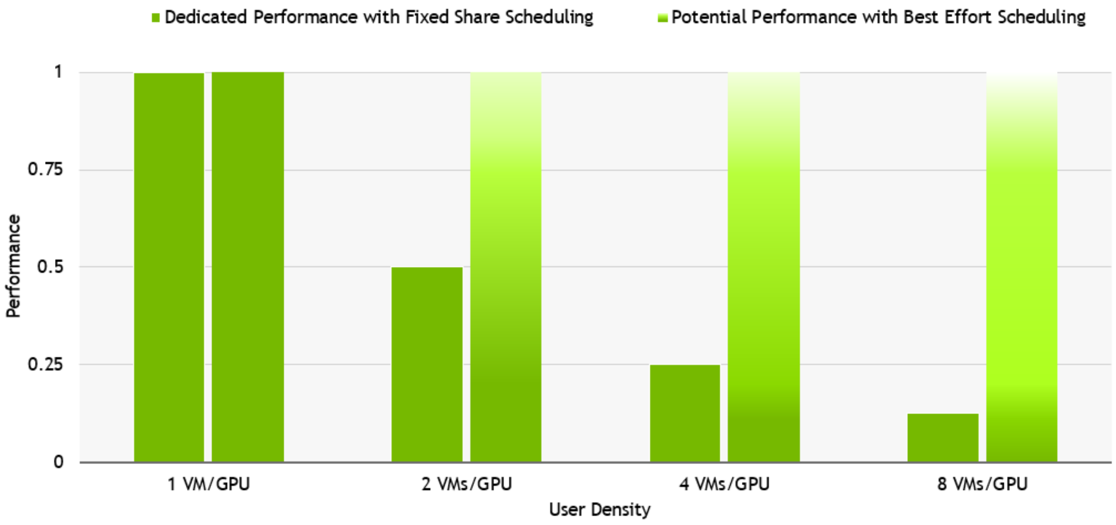


Table 10. Dedicated performance and potential performance comparison.

For details on the NVIDIA test environment used for this report, refer to the Appendix.

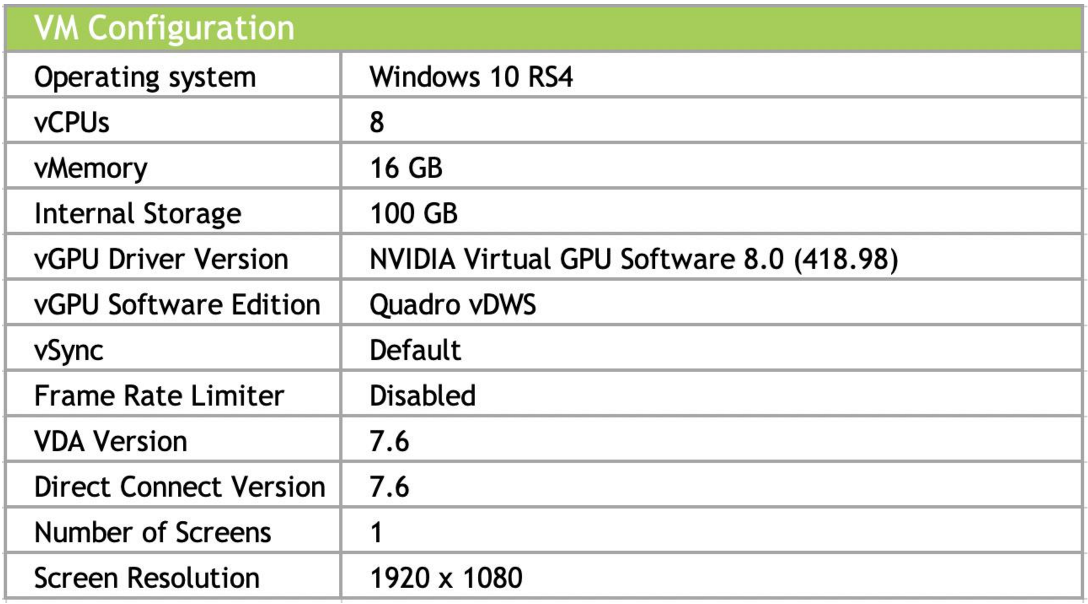
SUMMARY

When sizing a Quadro vDWS deployment for Esri ArcGIS Pro, NVIDIA recommends conducting a POC and fully analyzing resource utilization using objective measurements and subjective feedback. The best effort scheduler option is recommended for enterprise deployments, and user density will be dependent on the hardware configuration and user types.

To see how you can virtualize Esri ArcGIS Pro using Quadro vDWS software, try it for free. Or learn more about Quadro vDWS software.

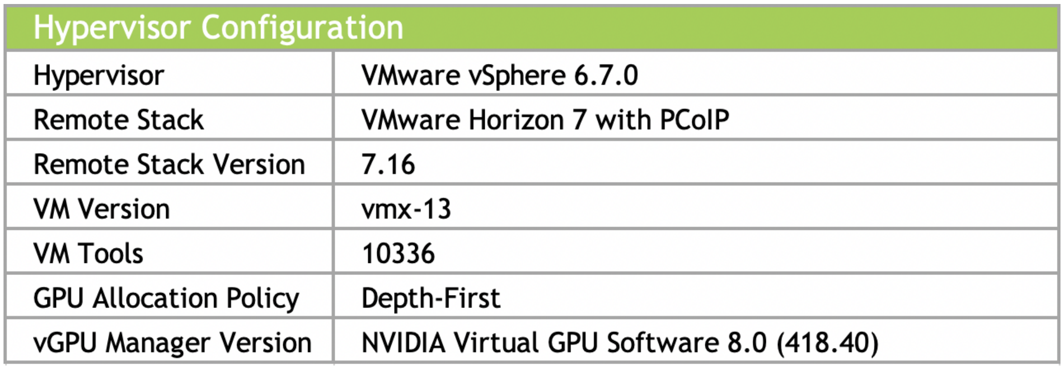
APPENDIX A

NVIDIA TEST ENVIRONMENT



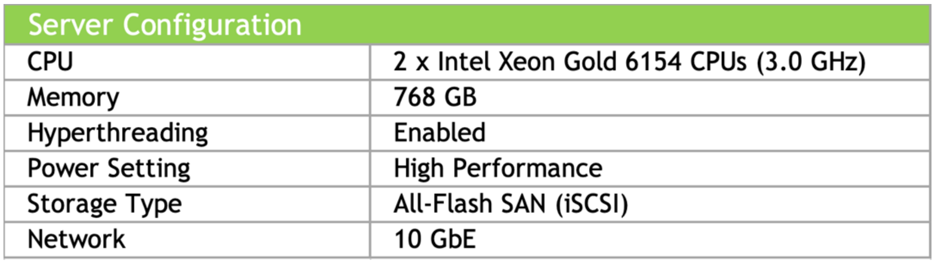
VM Configuration	
Operating system	Windows 10 RS4
vCPUs	8
vMemory	16 GB
Internal Storage	100 GB
vGPU Driver Version	NVIDIA Virtual GPU Software 8.0 (418.98)
vGPU Software Edition	Quadro vDWS
vSync	Default
Frame Rate Limiter	Disabled
VDA Version	7.6
Direct Connect Version	7.6
Number of Screens	1
Screen Resolution	1920 x 1080

Table 11. Virtual Machine (VM) configuration details



Hypervisor Configuration	
Hypervisor	VMware vSphere 6.7.0
Remote Stack	VMware Horizon 7 with PCoIP
Remote Stack Version	7.16
VM Version	vmx-13
VM Tools	10336
GPU Allocation Policy	Depth-First
vGPU Manager Version	NVIDIA Virtual GPU Software 8.0 (418.40)

Table 12. Hypervisor configuration details



Server Configuration	
CPU	2 x Intel Xeon Gold 6154 CPUs (3.0 GHz)
Memory	768 GB
Hyperthreading	Enabled
Power Setting	High Performance
Storage Type	All-Flash SAN (iSCSI)
Network	10 GbE

Table 13. Server configuration details

APPENDIX B

Dedicated Performance – Users Per Server

The recommendations within the Dedicated Performance table, ensures that the Host GPU and CPU can deliver the most optimal performance to ArcGIS Pro.

The recommended users per server within the Dedicated Performance table is limited to 12 light users per server since the Host Utilization reached 100% during the beginning of 24 VM test execution. The following graph indicates Host CPU Utilization at scale of 24 VM's. This spike in Host CPU illustrates that the host was CPU bound and in result, rendering times within VM's increased greater than 25%.

Host CPU Utilization – 24 VMs on ESXi Host

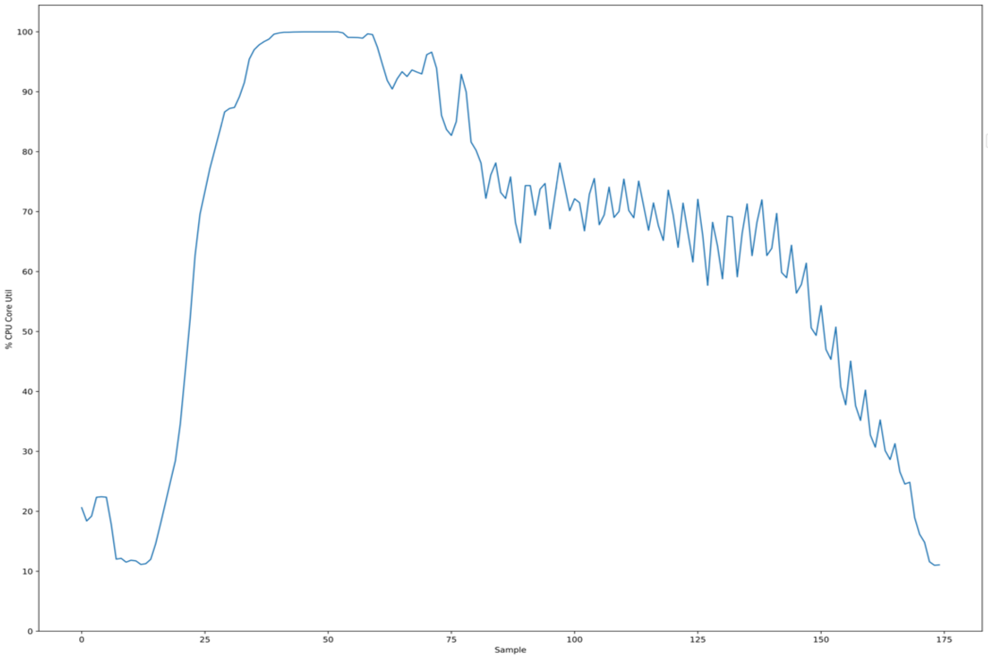


Table 1. Host CPU Utilization – 24 VM's running on ESXi host.

Host GPU was not a bottleneck during the 24 VM test. The following graphs illustrates that all six T4 were heavily utilized during test execution and were optimally utilized:

Host GPU Utilization – 24 VMs on ESXi Host

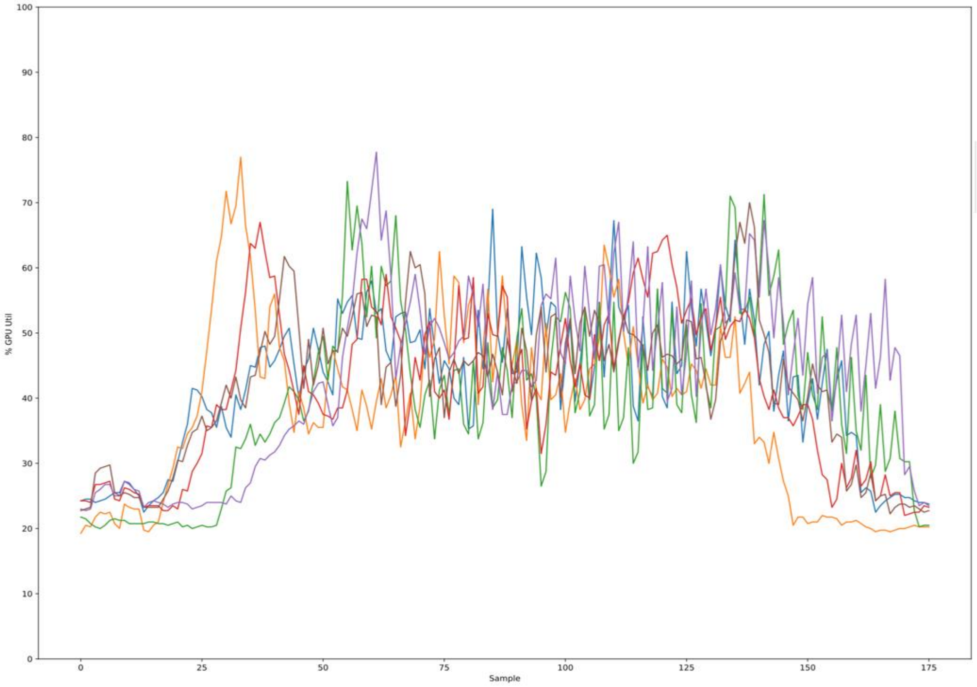


Table 2. Host GPU Utilization – 24 VM's running on ESXi host.

The following graph provides a baseline for the benchmark dataset by showing the GPU and vRAM utilization on a single T4-8Q VM. It is important to keep in mind, that the dataset used within the 3D multi-patch rendering benchmark is considered less complex and is smaller size (less than 1.5GB on disk) than a typical production level dataset. It is highly recommended to baseline your own production datasets using the GPU Profiler tool in order to understand the frame buffer requirements of your own datasets. In doing so, you will be able to determine if your own production datasets will fit within the 8GB frame buffer for dedicated performance.

VM vGPU and vRAM Utilization Rates

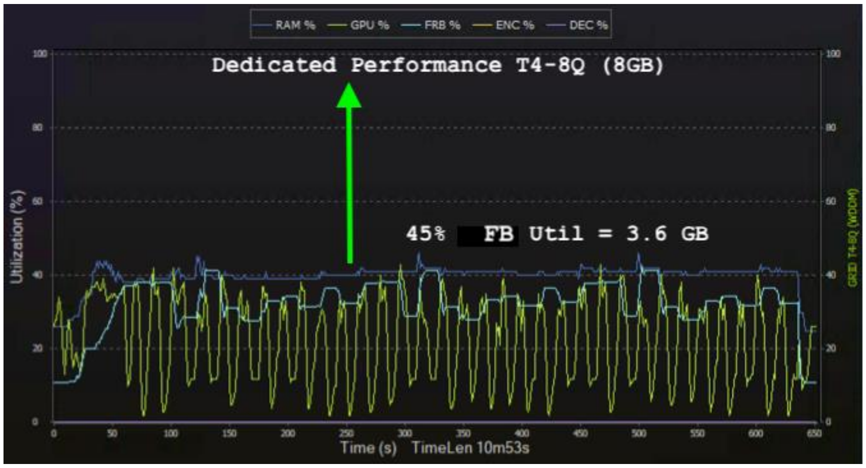


Table 3. VM vGPU and vRAM Utilization baseline.

Using the information from the two aforementioned graphs (Host CPU and VM GPU utilization rates), the Dedicated Performance recommendations, ensures that Host CPU is not a bottleneck and the VM has plenty of framebuffer and GPU compute to deliver the most optimal performance to ArcGIS Pro.

Throughout this guide, it is noted that there are some limitations to benchmarks, like the three which were used for Esri ArcGIS Pro. The benchmark doesn't account for the times when the system isn't fully utilized, or which hypervisors, and the best effort scheduling policy to leverage to achieve higher user densities with consistent performance. Therefore, the typical customer deployment table illustrates customers have achieved 16-24 users per server using the T4-4Q profile. We highly recommend a proof of concept (POC) is ran prior to doing a full deployment to provide a better understanding of how your users work and how many GPU resources they really need, analyzing the utilization of all resources, both physical and virtual. Consistently analyzing resource utilization and gathering subjective feedback allows for optimizing the configuration to meet the performance requirements of end users while optimizing the configuration for best scale.

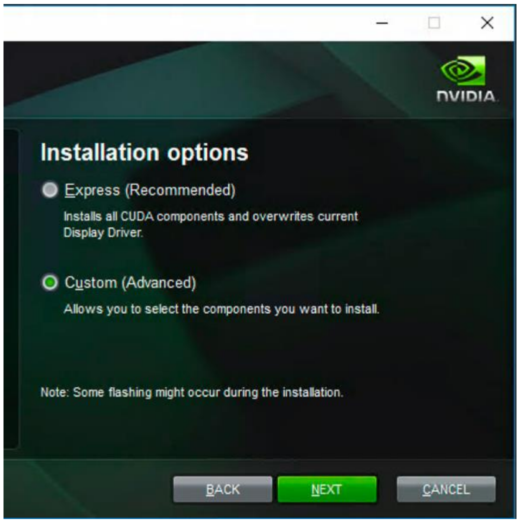
APPENDIX C

Configuring TensorFlow to execute on GPU.

Esri ArcGIS Pro can use TensorFlow to execute Deep Learning. To run TensorFlow on a GPU, first you will need to install Nvidia GPU drivers, CUDA toolkit and CuDNN SDK.

> *NOTE: If TensorFlow was previously added within the ArcGIS Python Package Manager, you will need to first uninstall. Once TensorFlow (TensorFlow-base and Tensorboard) have been un-installed from the Python Package Manager, you can verify the uninstall by typing `import tensorflow as tf` within the python command window. If no module named tensorflow was found, then the uninstall was successful.*

1. Download CUDA Toolkit version 10.1
2. Select custom install



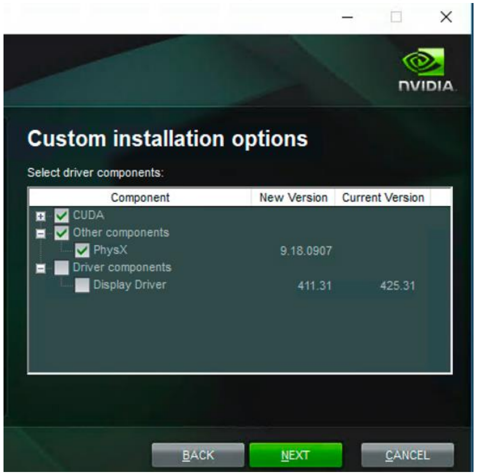
Installation options

○ Express (Recommended)
Installs all CUDA components and overwrites current Display Driver.

● Custom (Advanced)
Allows you to select the components you want to install.

Note: Some flashing might occur during the installation.

3. Unselect Driver Components, **Display Driver**



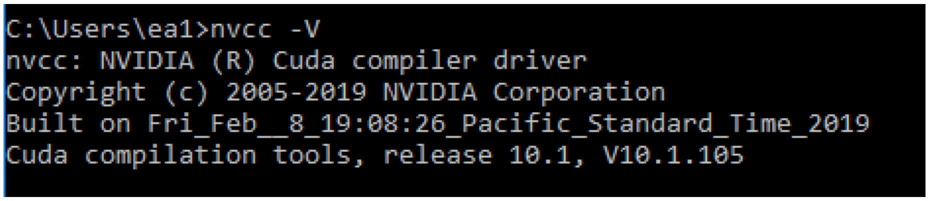
Custom installation options

Select driver components:

Component	New Version	Current Version
☑ CUDA		
☑ Other components		
☑ PhysX	9.18.0907	
□ Driver components		
□ Display Driver	411.31	425.31

NOTE: Unselecting the Display Driver will keep the virtualization host/guest driver in sync which is essential for a successful deployment of vGPU.

4. Open Python command prompt to verify the install and typing in `nvcc -V`



```
C:\Users\ea1>nvcc -V
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2019 NVIDIA Corporation
Built on Fri_Feb__8_19:08:26_Pacific_Standard_Time_2019
Cuda compilation tools, release 10.1, V10.1.105
```

5. Download cUDNN v7.5.1 for CUDA 10.1

   NOTE: This is not a traditional installer file. The download link will contain a zip file with several folders, each containing the CuDNN files (1-Dll, 1-header and 1-library).

   a. Locate your CUDA installation (should be something like C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1).

      NOTE: the directories within the CUDA installation are also the directory within the zip file. There is a bin, and include, a lib.

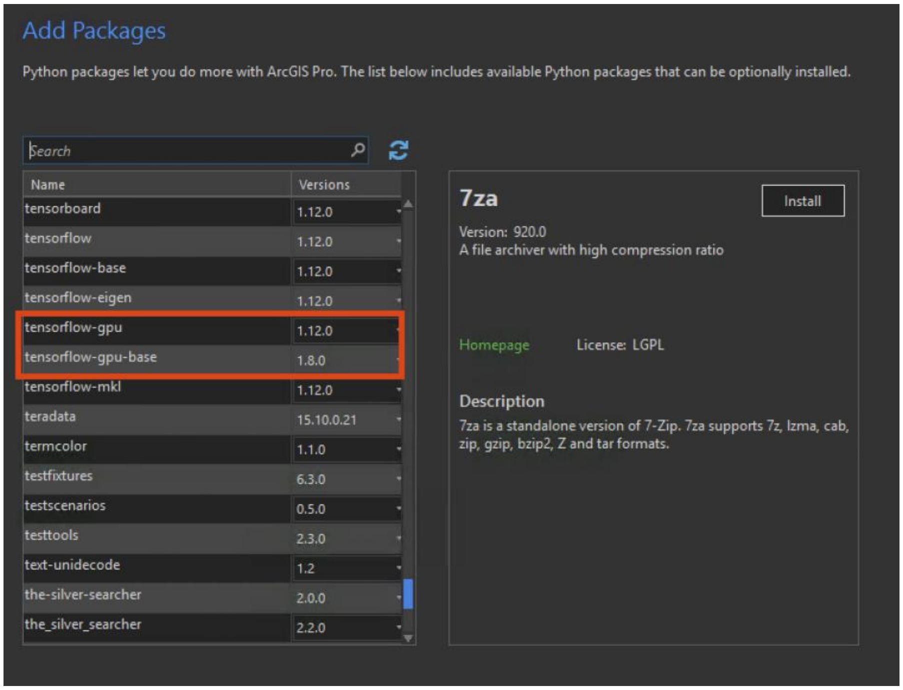
   b. Copy the files from the zip to the relevant directory. For example, cudnn64_7.dll into C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1\bin. Do the same for the other files.

   c. Add the CUDA, cuDNN installation directories to the %PATH% environmental variable.

```
SET PATH=C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1\bin;%PATH%
SET PATH=C:\Program Files\NVIDIA GPU Computing
Toolkit\CUDA\v10.1\extras\CUPTI\libx64;%PATH%
SET PATH=C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1\include;%PATH%
SET PATH=C:\tools\cuda\bin;%PATH%
```

6. Within ArcGIS Pro, to create an environment, in the Python backstage, click Manage Environments button, click an environment and click the environment's Clone button.

7. Add Tensorflow-gpu and Tensorflow-gpu-base Packages within the Python Package Manager in ArcGIS Pro.



Add Packages

Python packages let you do more with ArcGIS Pro. The list below includes available Python packages that can be optionally installed.

Name	Versions
tensorboard	1.12.0
tensorflow	1.12.0
tensorflow-base	1.12.0
tensorflow-eigen	1.12.0
tensorflow-gpu	1.12.0
tensorflow-gpu-base	1.8.0
tensorflow-mkl	1.12.0
teradata	15.10.0.21
termcolor	1.1.0
testfixtures	6.3.0
testscenarios	0.5.0
testtools	2.3.0
text-unidecode	1.2
the-silver-searcher	2.0.0
the_silver_searcher	2.2.0

7za

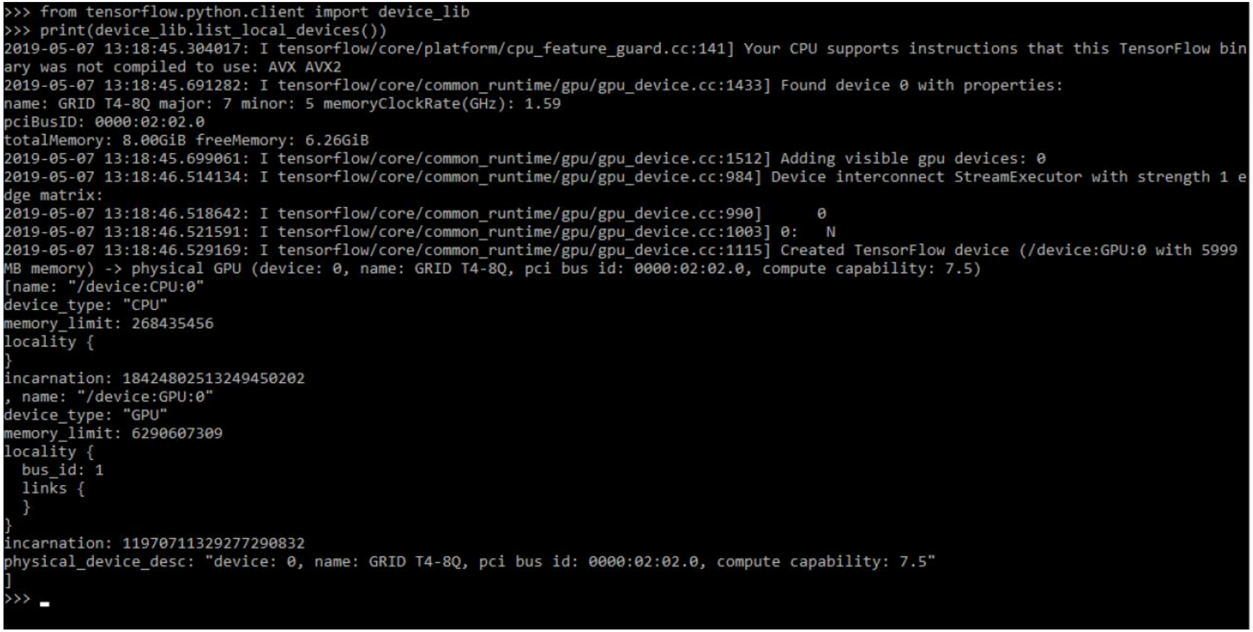
Version: 920.0
A file archiver with high compression ratio

Homepage License: LGPL

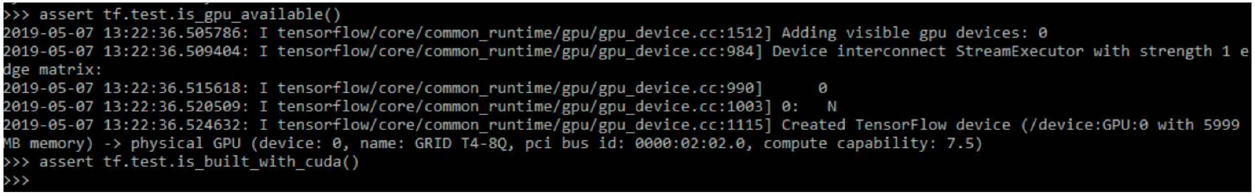
Description

7za is a standalone version of 7-Zip. 7za supports 7z, lzma, cab, zip, gzip, bzip2, Z and tar formats.

8. Open Python command prompt and verify Tensorflow-gpu install.



```
>>> from tensorflow.python.client import device_lib
>>> print(device_lib.list_local_devices())
2019-05-07 13:18:45.304017: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow bin
ary was not compiled to use: AVX AVX2
2019-05-07 13:18:45.691282: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1433] Found device 0 with properties:
name: GRID T4-8Q major: 7 minor: 5 memoryClockRate(GHz): 1.59
pciBusID: 0000:02:02.0
totalMemory: 8.00GiB freeMemory: 6.26GiB
2019-05-07 13:18:45.699061: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1512] Adding visible gpu devices: 0
2019-05-07 13:18:46.514134: I tensorflow/core/common_runtime/gpu/gpu_device.cc:984] Device interconnect StreamExecutor with strength 1 e
dge matrix:
2019-05-07 13:18:46.518642: I tensorflow/core/common_runtime/gpu/gpu_device.cc:990]      0
2019-05-07 13:18:46.521591: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1003] 0:   N
2019-05-07 13:18:46.529169: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1115] Created TensorFlow device (/device:GPU:0 with 5999
MB memory) -> physical GPU (device: 0, name: GRID T4-8Q, pci bus id: 0000:02:02.0, compute capability: 7.5)
[name: "/device:CPU:0"
device_type: "CPU"
memory_limit: 268435456
locality {
}
incarnation: 18424802513249450202
, name: "/device:GPU:0"
device_type: "GPU"
memory_limit: 6290607309
locality {
  bus_id: 1
  links {
  }
}
incarnation: 11970711329277290832
physical_device_desc: "device: 0, name: GRID T4-8Q, pci bus id: 0000:02:02.0, compute capability: 7.5"
]
>>> _
```



```
>>> assert tf.test.is_gpu_available()
2019-05-07 13:22:36.505786: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1512] Adding visible gpu devices: 0
2019-05-07 13:22:36.509404: I tensorflow/core/common_runtime/gpu/gpu_device.cc:984] Device interconnect StreamExecutor with strength 1 e
dge matrix:
2019-05-07 13:22:36.515618: I tensorflow/core/common_runtime/gpu/gpu_device.cc:990]      0
2019-05-07 13:22:36.520509: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1003] 0:   N
2019-05-07 13:22:36.524632: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1115] Created TensorFlow device (/device:GPU:0 with 5999
MB memory) -> physical GPU (device: 0, name: GRID T4-8Q, pci bus id: 0000:02:02.0, compute capability: 7.5)
>>> assert tf.test.is_built_with_cuda()
>>>
```

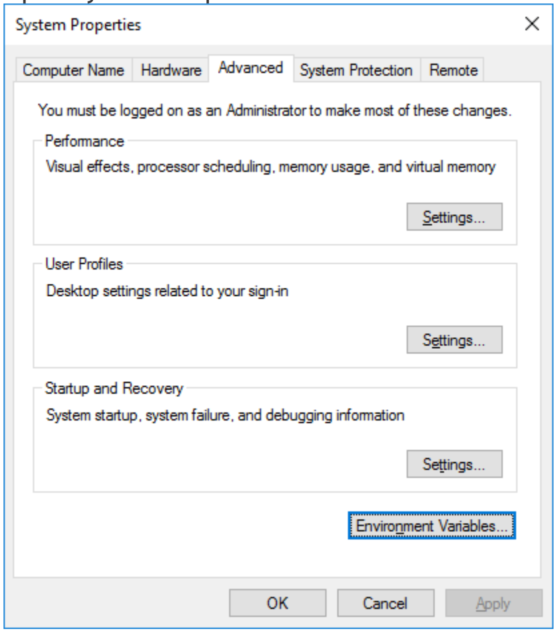
TensorFlow has now been successfully configured to run on the GPU when executing Esri Deep Learning Tools within ArcGIS Pro.

APPENDIX D

Configuring 3D Analyst tools to execute on GPU.

Esri ArcGIS Pro can use CUDA to execute Spatial Analysis tools on the GPU. For more information regarding which tools currently support GPU processing, please refer to the Esri ArcGIS Pro documentation.

1. Open System Properties and select Environmental Variables.



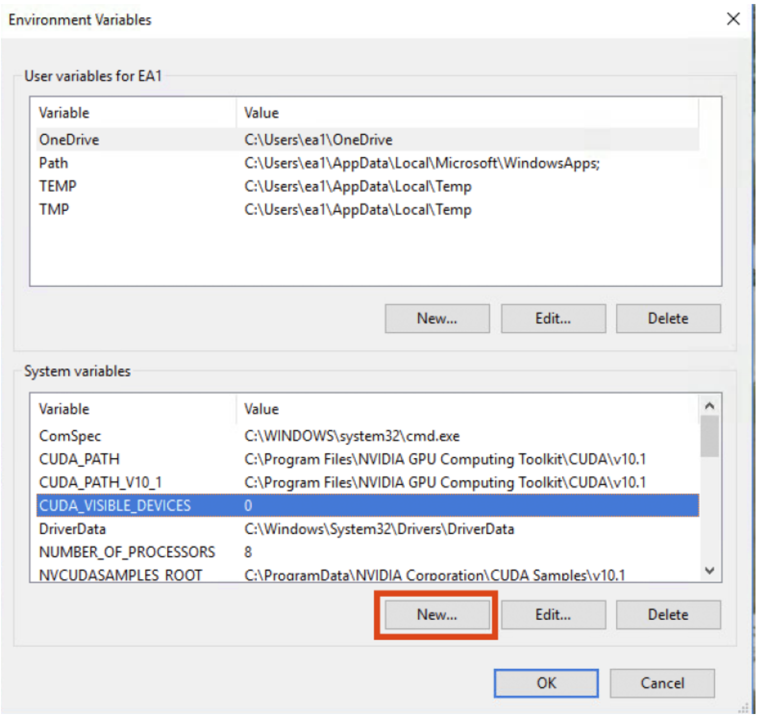
You must be logged on as an Administrator to make most of these changes.

Visual effects, processor scheduling, memory usage, and virtual memory

Desktop settings related to your sign-in

System startup, system failure, and debugging information

2. Create a New System Variable called CUDA_VISIBLE_DEVICES



User variables for EA1

Variable	Value
OneDrive	C:\Users\ea1\OneDrive
Path	C:\Users\ea1\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\ea1\AppData\Local\Temp
TMP	C:\Users\ea1\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
CUDA_PATH	C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1
CUDA_PATH_V10_1	C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v10.1
CUDA_VISIBLE_DEVICES	0
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
NVCUDASAMPLES_ROOT	C:\ProgramData\NVIDIA Corporation\CUDA Samples\v10.1

3. Edit the System variable to use the GPU, value of 0 will use the GPU.

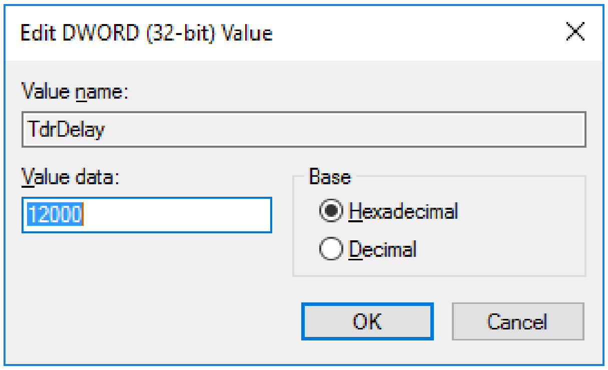
   NOTE: Modifying the value to -1 will use the CPU to execute.

4. Optional: TDRDelay, a timeout value, may prematurely timeout GPU CUDA processing if the following timeout is not set in the following REGKEY:

   HKEY_LOCAL_MACHINE\System\CurrentControlSet\Control\GraphicsDrivers

   Type: DWORD (32-bit)
   Value Name: TdrDelay
   Value Data: 12000

   For more information regarding TDRDelay.

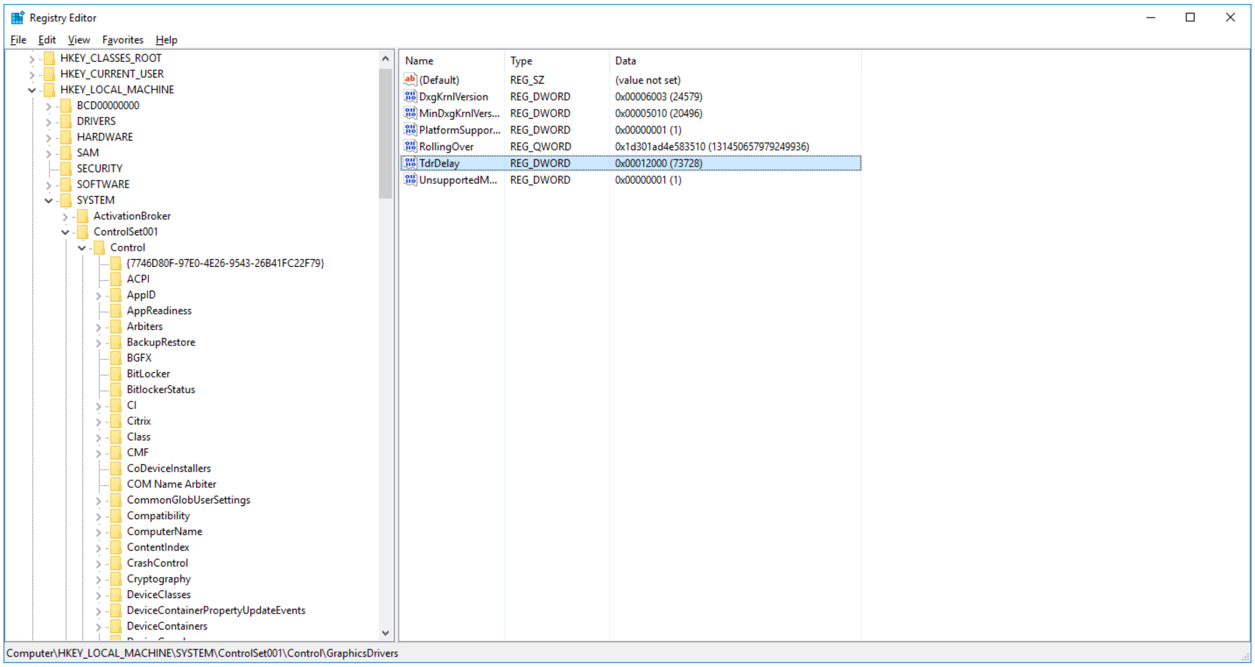


Edit DWORD (32-bit) Value

Value name: TdrDelay

Value data: 12000

Base
● Hexadecimal
○ Decimal



Name	Type	Data
(Default)	REG_SZ	(value not set)
DxgKrnlVersion	REG_DWORD	0x00006003 (24579)
MinDxgKrnlVers...	REG_DWORD	0x00005010 (20496)
PlatformSuppor...	REG_DWORD	0x00000001 (1)
RollingOver	REG_QWORD	0x1d301ad4e583510 (131450657979249936)
TdrDelay	REG_DWORD	0x00012000 (73728)
UnsupportedM...	REG_DWORD	0x00000001 (1)

Computer\HKEY_LOCAL_MACHINE\SYSTEM\ControlSet001\Control\GraphicsDrivers