Application Deployment Guide

Esri ArcGIS Pro with NVIDIA GRID vGPU on VMware Horizon

esri | NVIDIA. | vmware

THE QUESTION

HOW MANY USERS CAN I GET ON A SERVER?

This is a typical conversation we have with customers considering NVIDIA GRID vGPU:

How many users can I get on a server?

NVIDIA: What is their primary application?

Esri ArcGIS Pro 1.0.

NVIDIA: Are they primarily 3D or 2D data users?

3D mostly.

NVIDIA: Would you describe them as light, medium, or heavy users?

Medium to heavy.

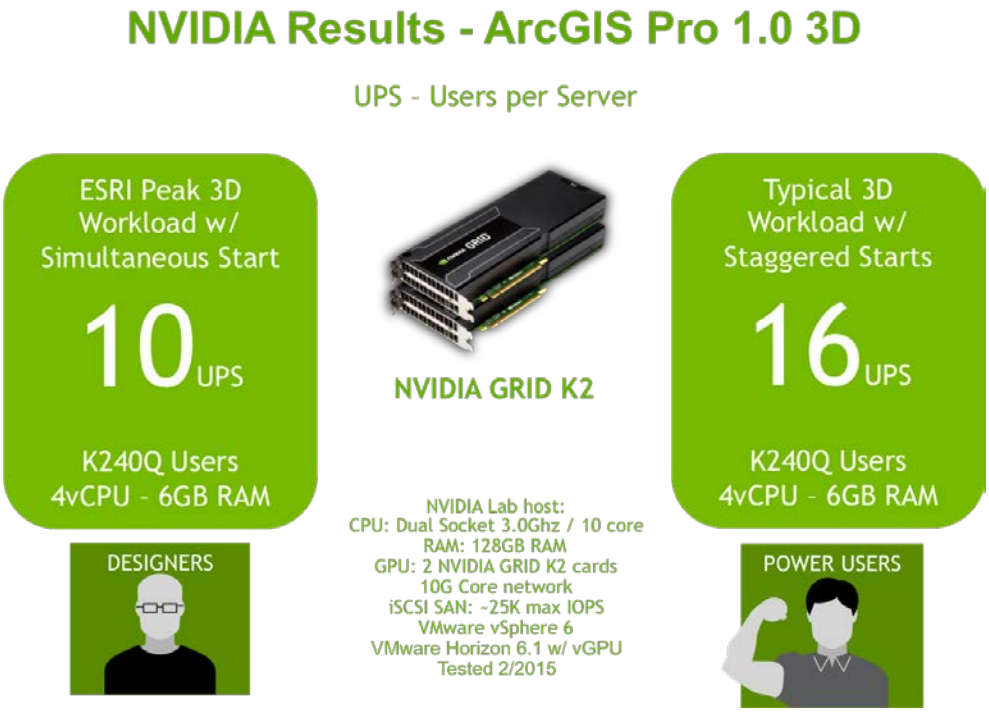
NVIDIA: Power users to designers then.

I need performance AND scalability numbers that I can use to justify the project.

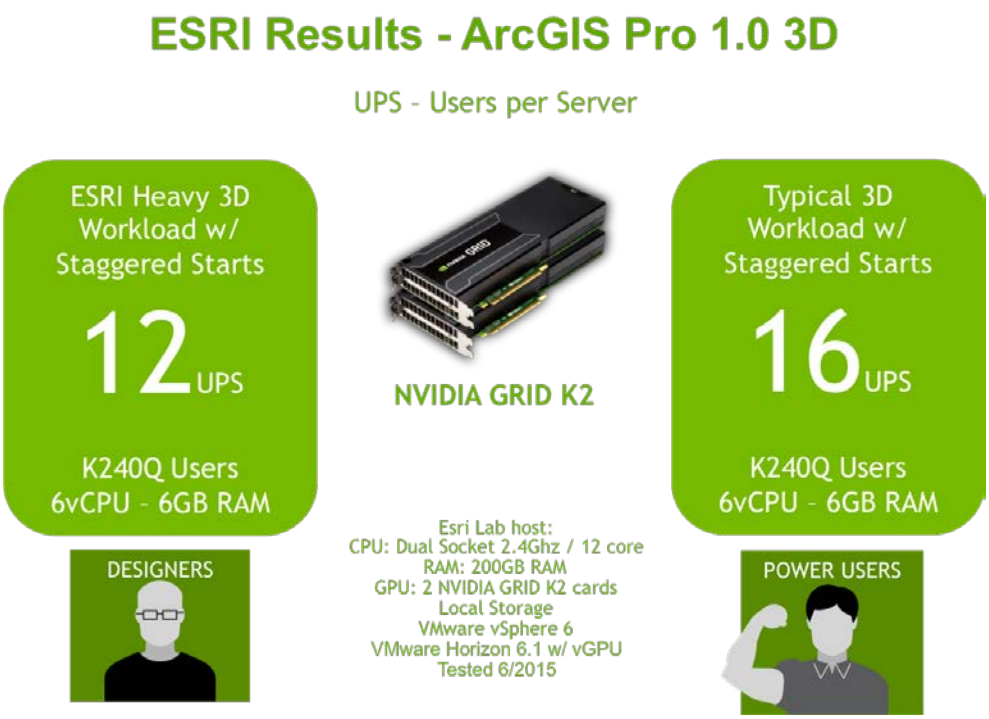
THE ANSWER – USERS PER SERVER

UPS – USERS PER SERVER

Based on NVIDIA Performance Engineering Lab findings, NVIDIA GRID provides the following performance and scalability metrics for Esri ArcGIS 3D Pro 1.0. These metrics are based on tests with the lab equipment shown in the graphic below, using the Esri API based "heavy 3D" benchmark and in working with Esri to determine acceptable performance. Of course, your usage will depend on your models, but this is guidance to help guide your implementation.



Esri's performance engineering team also performed tests using slightly different synthetic human behavior and a different CPU clock and core to give a broader perspective of results. Notice their lab equipment in the following graphic; also note the increase in vCPU count/VDI:



The details regarding these results are detailed later in this application guide.

ABOUT THE APPLICATION: ESRI ARCGIS PRO 1.0

ArcGIS Pro 1.0 is the premier Geographic Information Systems (GIS) application for mapping, visualizing, editing, and analyzing spatial data. Esri recommends a GPU for best end user experience, but as ArcGIS Pro 1.0 also generates heavy CPU load, this also needs to be considered in architecting your vGPU solution. The size of your map data, the concurrency of your users, and the level of interaction with 3D data all need to be considered when defining your user groups.

User Classification Matrix

Esri classifies its users as follows in Table-01. We then correlate these to our own NVIDIA user classifications as a reference:

Given they are the most graphics intensive users, we focused our tests on the designer and power user groups.

Table-01

User Classification Matrix						
NVIDIA User Classifications	**Knowledge Workers**		**Power User**		**Designer**	
ESRI User Classifications	Light 2D	Medium 2D	Heavy 2D	Light 3D	Medium 3D	Heavy 3D

HOW TO DETERMINE USERS PER SERVER

This section contains an overview of the NVIDIA GRID Performance Engineering Lab, recommended virtual desktop builds, the testing methodology used, and the metrics and results that support the findings in this deployment guide.

The Performance Engineering Lab

The NVIDIA GRID Performance Engineering Team's mandate is to measure and validate the performance and scalability delivered via the GRID platform, GRID vGPU software running on NVIDIA GRID GPUs, on all enterprise virtualization platforms. The goal of this team is to provide proven testing that gives customers the ability to create a successful deployment.

Leveraging its lab of enterprise virtualization technology, the Performance Engineering Team has the capacity to run a wide variety of tests ranging from standard benchmarks to reproducing customer scenarios across a wide range of hardware.

None of this is possible without working with ISVs, OEMs, vendors, partners, and their user communities to determine the best methods of benchmarking in ways that are both accurate and reproducible. As a result, the Performance Engineering Team works closely with its counterparts in the enterprise virtualization community.

The NVIDIA Performance Engineering Lab holds a wide variety of different OEM servers, with varying CPU specifications, storage options, client devices, and network configurations. We work closely with OEMs and other third party vendors to develop accurate and reproducible benchmarks that ultimately will assist our mutual customers to build and test their own successful deployments.

TYPICAL ESRI ARCGIS PRO 1.0 3D VIRTUAL DESKTOPS

Esri delivers a recommended hardware specification to help choose a physical workstation. These recommendations provide a good starting point to start architecting your virtual desktops. Based on our Esri PerfTools testing results, along with feedback from early customers, this is our recommended virtual system requirement. Your own tests with your own models will determine if these recommendations meet your specific needs.

VMware Recommended ArcGIS Pro Virtual System Requirements

Working with VMware, Esri, and our shared customers, the NVIDIA GRID Performance Engineering Team recommends in Table-02 the following system requirements for deploying Esri ArcGIS Pro 1.0 in a virtual environment:

Table-02

VMware: Recommended Level Configuration			
VMware Software	VMware vSphere 6 or later w/ VMware Horizon 6.1 or later		
Virtual Machine Operating System	Microsoft® Windows® 7 SP1 64-bit: Enterprise, Ultimate, or Professional Microsoft® Windows® 8.1 64-bit: Enterprise, Pro, or Windows 8.1		
Host Server Recommendation	Minimum (Light 3D)	Recommended (Medium 3D)	Optimal (Heavy 3D)
CPU (Haswell, Intel® Xeon E5 v3, or greater recommended)	2.6 GHz+ Intel® Xeon E5 v2 or greater 2.3 GHz+ Intel® Xeon E5 v3 or greater	3.0 GHz+ Intel® Xeon E5 v2 or greater 2.3 GHz+ Intel® Xeon E5 v3 or greater	3.0 GHz+ Intel® Xeon E5 v2 or greater 2.3 GHz+ Intel® Xeon E5 v3 or greater
Memory	128 GB	160 GB	192-256 GB
Networking	1 Gb minimum 10 Gb recommended	10 Gb	10 Gb
Storage	~250+ IOPS Per User	~500+ IOPS Per User	~500+ IOPS Per User
GPU	NVIDIA GRID K2 or later	NVIDIA GRID K2 or later	NVIDIA GRID K2 or later

Virtual Machine Settings	Minimum (Light 3D)	Recommended (Medium 3D)	Optimal (Heavy 3D)
Memory	4 GB RAM	8 GB RAM	16 GB RAM or greater
vCPUs	4 vCPUs	4 vCPUs	4 vCPUs
Disk Space	50 GB free disk space	50 GB free disk space	50 GB free disk space
Graphics Adapter	NVIDIA GRID K220Q (512 MB) or later	NVIDIA GRID K240Q (1 GB) or later	NVIDIA GRID K260Q (2 GB) or later

For the test, the NVIDIA GRID Performance Engineering Team keys on recommended specifications when feasible. The goal is to test both performance and scalability; to maintain the flexibility and manageability advantages of virtualization without sacrificing the performance end users expect from NVIDIA powered graphics.

UX – THE VDI USER EXPERIENCE

Defining user experience (UX) requires careful examination of user and application interaction. This can be obvious, like the rendering time for an image to appear or smoothly panning across that image. It can also be less obvious, like the ability to smoothly scroll down a page or the "snappy" reaction for a menu to appear after a right click. While elements such as these can be measured, the user's perception is much harder to measure.

Users also add variables like "think time", the time they spend looking at their display before interacting again with the application. This time offers an advantage to the underlying resources, such as CPU, as it allows tasks to finish and processes to complete. It is even more beneficial in a shared resource environment such as VDI, where one user "thinking" frees up resources for another user who chose that moment to interact with their application. Now factor in other time away from the application (meetings, lunch, etc.) and one could expect to get even more benefits from shared resources. These benefits equate to more resources for the user's session and typically a more responsive application, thus a better-perceived experience by the end user.

Using a known data set, "Philly 3D", the Esri 3D test cycles through 11 pre-defined bookmarks. Testing started with a single VM benchmark test, which reported a total test execution time of 30 seconds, averaging 3.3 seconds per bookmark. The scalability

threshold was determined by examining the total execution time as well as the average bookmark time. This average bookmark time is indicative of time the user is waiting for a single bookmark map display to complete. The initial 10 VM test reported a 4.4 second average bookmark time, about 1 second more than the single VM test. After several rounds of tests, with individual desktops being viewed to confirm the experience, it was determined that overall results 45 seconds or greater were less than acceptable for end user experience.

In the tests that ESRI performed, the addition of staggered starts and additional think time made for longer total test times. As a result, a different scoring method, one based on actual user interaction was employed. See below for additional details.

Esri Benchmark Metrics

Esri provides a PerfTools add-in that allows gathering of UX metrics during benchmarking. Esri, as the ISV, knows their product best and defines great UX as the combination of the following metrics:

- Draw Time Sum: The total time elapsed for all of the benchmarks to fully draw. This was defined by Esri to be acceptable up to 45 seconds. Less time would be a better 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 <2 for 2D and <4 for 3D workloads.

After initial testing, it was clear that Draw Time Sum is the logical key metric for UX. If Draw Time Sum is inside its acceptability threshold, then all other three metrics were also inside their respective metrics. This gave us a single value to track, and then validate the rest of the results were within acceptable ranges as well.

Real Life Experience Versus Benchmarking

The goal is to find the most accurate possible proxy for testing; however, this is still not the same as real users doing real work with their data. The NVIDIA GRID Performance Engineering Labs is committed to working with customers to find more and better models, and field confirmation of findings.

The Importance of Eyes On!

It's important to view the tests to be sure that the experience is enjoyable to users. That said, it's also important to keep perspective, especially if you are not a regular user of applications such as ArcGIS Pro 1.0. While a data center admin deploying an ArcGIS Pro 1.0 VDI workload might view a testing desktop and think the experience is slow or sluggish, a daily user who works in it daily might find it normal. An actual 3D designer user using the app in a virtual desktop is the ultimate test of success. As an example, we discovered that the tests did not include panning, or "Navigation", an action in the application that users may typically leverage. This activity increases both CPU and GPU utilization and as a result may negatively impact user experience if either approaches their limits. To ensure this, in their tests Esri leveraged individual users on each end device to witness the tests and judge usability. The table below shows the scoring methodology they used:

Table-03 (The Usability rating)

User Experience Ranking	User Experience Examples
1 to 5	Unacceptably jerky; Poor to annoying user experience
6	Jerky, but usable
7	Moderately smooth, moderate jerkiness
8	Smooth, minor jerkiness
9 to 10	Very smooth; little to no difference between a non-VDI solution. No apparent jerkiness, lag/tearing, or texturing delays.

TESTING METHODOLOGY

To ensure that test results are repeatable we have deliberately chosen a peak workload and executed simultaneous tests, meaning all testing virtual desktops are executing the same activities at the same time. A "Peak Workload" should be an unlikely demonstration of real user interaction, the result showing the number of users per host when the highest possible load generated by the application in question is put on the shared resources.

In the case of ArcGIS Pro 1.0, the NVIDIA Performance Engineering team focused on Esri's graphics-rendering pipeline using DirectX 11 to determine the impact of GPU on performance and scalability. OpenGL testing will be covered in future guides.

These tests did not focus on analytics, as this has a greater impact on networking (assuming remote data) and CPU. With NVIDIA GRID vGPU allowing the virtual desktops to be moved into the data center, Esri theoretically expects analytic performance to improve. With data proximity the entire UX should improve, since

analytics are performed by the application and with VDI the application is in theory closer to the data with both located in the same data center. The NVIDIA GRID Performance Engineering Team intends to test analytics operations in the future. Given ArcGIS Pro 1.0 is asynchronous you can render while running analytics in the background. The impact of this will need to be tested.

The following information details the test methodology used with the Esri PerfTools add-in:

- Sample workload: Esri provided their "Philly 3D" workload for us to test with. This test is described by Esri as a representative data set for a "heavy 3D" user.
- Scripting: Esri provided application scripting designed to run the application through several bookmarks, performing functions, and capturing the length of time to complete the tasks.
- Think Time: The Esri script allowed for "think time" adjustment, allowing us to create synthetic human behavior. We started with the default of 2 seconds, then adjusted it to 5 seconds, then 10 seconds. This imitates time a user pauses between interactions with the application.
- Scalability: Automation scripting enables the team to run tests on specific quantities of virtual desktops as required. In general, we run 1, then 8, then 16, to get a baseline of results and accompanying logs (CPU, GPU, RAM, networking, storage IOPS, etc.), then narrow down the optimal number of desktops based on UX.

RESULTS OVERVIEW

- **vCPU:**

  Many fundamental tasks associated with spatial analysis are CPU intensive, thus ArcGIS Pro 1.0 benefits from well-configured vCPUs. Four vCPU performed better than six or eight because the app spawns more threads, negatively impacting performance. When architecting your solution, keep in mind that one vCPU is used by the OS, one by VMware ESXi, and two by the application. Esri, using a slower CPU but with more cores, found that 4 vCPUs was resource constrained and caused VDI to crash; as a result they moved to 6 vCPUs and the issue was resolved. Esri also tested 8, but found little difference with the additional 2 vCPUs.

  - ***Result: NVIDIA testing found 4 vCPUs performed the best based on the benchmarks ESRI provided. ESRI, using a slower CPU but with more cores, found 6 vCPUs to perform best.***

- **vRAM:**

  Esri ArcGIS Pro 1.0 is typically not RAM intensive so based on recommended specifications we started with 6GB/virtual desktop. During testing with 4GB of

vRAM CPU congestion occurred. We determined this was caused by the application requiring 5+GB of vRAM to load the "Philly 3D" dataset. Tests with 8GB of vRAM per virtual desktop did not produce improved performance supporting Esri's statement that the application is not RAM intensive.

- ***Result: For this workload ≥6GB of vRAM is required. You should base the amount of vRAM on the needs of your actual workloads.***

- **vGPU**

  Esri states that a GPU is required. Tests used servers hosting pairings of K2 cards. A test with the full GPU and all 4GB of frame buffer, the K280Q profile, offered high performance but lacked scalability as it limits tests to 4 virtual desktops in a two K2 host. Next the team tested with the K260Q profile and its 2GB of frame buffer - up to its maximum of 8 users (4 per K2 card) results were under the acceptability metrics provided by ESRI. This meant there were resources remaining for more users if frame buffer was lowered to 1GB via the K240Q profile. The K220Q profile and its 512MB of frame buffer was tested, but this was too little frame buffer and caused CPU swapping impacting performance and scalability, and further proved GPU is necessary.

  - ***Result: For Esri ArcGIS Pro 1.0 performance AND scalability, running a map with the characteristics of Philly3D, the K240Q profile was best.***

- **Storage**

  The NVIDIA GRID Performance Engineering Lab used a Pure Storage iSCSI attached SAN over 10G non-trunked networking. The Esri tests never exceed ~25,000 IOPS, and thus never taxed the flash based cache Pure Storage SAN.

  - ***Result: Clearly, local spindle bound storage would have been a bottleneck and impacted performance - the fast cache SAN handled the IOPS load.***

- **Networking**

  The NVIDIA GRID Performance Engineering Lab used 10G core, and 1GB distribution, networking. At no time was networking a bottleneck and results were unremarkable.

  - ***Result: It is very clear that moving users to the data, versus users pulling data over the wire to themselves, increases productivity.***

NVIDIA PERFORMANCE ENGINEERING TEST RESULTS

The following are the full results of our testing. The baseline was the 45 second draw time sum - anything greater than that value represented a worsening UX while less

would be a better UX. Looking for both performance and scalability, we tested looking for the greatest number of virtual desktops, and therefore the greatest scalability, while still within performance expectations and the threshold of 45 seconds. It's important to note that your users, your data, and your hardware will impact these results and you may decide a different level of performance or scalability is required to meet your individual business needs.

Tests were also run to look for potential NUMA issues that can negatively impact performance. This is where the physical GPU and its PCI-e channels are tied to one physical CPU, while the virtual desktop is running on the other physical CPU, so communication with the physical GPU has to move over the QPI between the two physical CPUs. This creates a bottleneck and can cause performance issues. However, in our testing, the application is sufficiently CPU bound that NUMA affinity made little difference.

The results in the table below show the decrease in performance as we increased vCPU counts, and then the increase in scalability with synthetic human behavior (think time):

Table-04

VM Config	VM count	Draw Time (min:sec)	FPS	Min FPS	Standard Deviation
K240Q 8vcpu 6GB vRAM	1	00:35.0	59.95	13.86	NA
	8	01:06.0	53.34	8.09	21.99
	16	02:03.0	47.79	3.8	4.74
K240Q 6vcpu 6GB vRAM	1	00:34.4	60.03	ND	NA
	8	00:49.0	51.48	7.3	3.6
	9	00:56.4	50.77	6.8	3.48
	16	01:39.6	46.60	4.52	6.27
K240Q 4vcpu 6GB vRAM (Best Results)	1	00:30.3	61.84	17.75	NA
	8	00:40.6	49.76	11.45	1.45
	9	00:41.6	46.87	10.21	3.2
	10	00:45.0	45.76	9.46	2.62
	12	00:51.2	40.31	7.54	6.3
	16	01:09.6	37.52	5.3	2.7
K240Q 4vcpu 6GB vRAM Think Time Increased	1	00:30.2	60.02	24.81	NA
	8	00:34.0	57.05	14.15	0.79
	12	00:38.1	52.87	12.00	1.2
	16	00:46.0	49.55	9.85	2.8

GREEN – DENOTES BEST RESULTS

NA – NOT APPLICABLE

ND – NOT DETERMINED

As the table shows, the Draw Time Sum hits 45 seconds at 10 users with the K240Q profile, 4 vCPUs, and 6GB of vRAM. Then by increasing "think time" scalability increased to 16 before we once again hit (in this case slightly exceeded – 1sec) the acceptability threshold of 45 seconds. Note that changing other variables, such as vCPU count, resulted in lower scalability and performance.

ESRI PERFORMANCE ENGINEERING TEST RESULTS

The following are the results of Esri's testing. Esri performed tests on a host with dual Ivy Bridge 2.4GHz CPUs, with 12 cores each. With the testing the NVIDIA Performance Team has done it was clear that lower clock speeds would cause a decrease in scalability with a peak (simultaneous start, limited think time) test. Esri views the Philly 3D benchmark as exceeding typical user models and causes more impact to the system than would be commonly seen in the field. As a result, they used two synthetic human behaviors to more accurately achieve what would be experienced by typical users: 10 second staggered starts, and 10 second think times between test segments.

Usability Scoring

To judge the usability Esri had users viewing each of the test end devices, one per VDI session. These users were asked to score 1-10, 10 being the best usability, 8 was deemed the threshold with anything lower being suboptimal. Keep in mind that the 4 vCPU tests suffered resource contention with CPU and saw crashes that would have relieved the remaining virtual desktops and given them artificially better scores. They are included below to show that 6 vCPUs did support the necessary usability.

Table-05 (The Usability Scores)

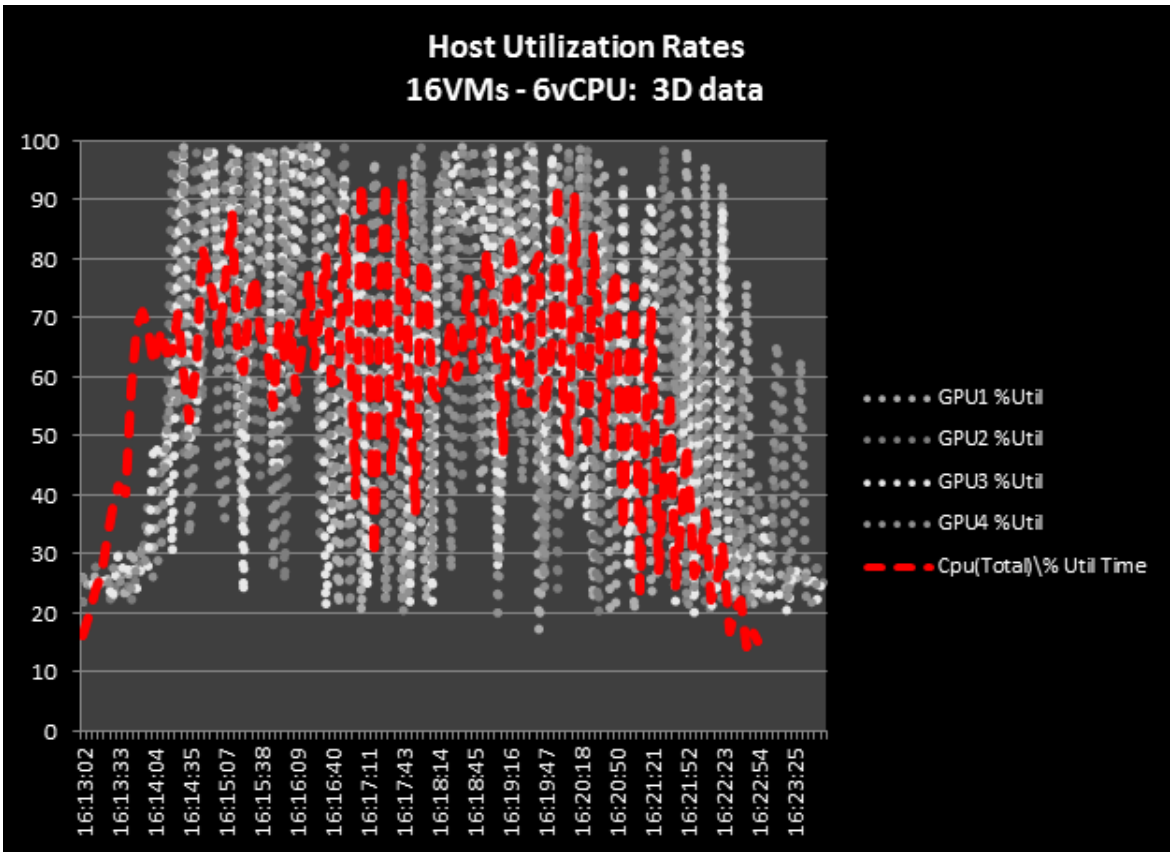
	16 users	**12 users**	**8 users**	**4 users**	**1 user**
4vCPU*	8	8	8	8.25	8.5
6vCPU	7.75	8	8	8.5	8.75

*Crashes occur, as a result remaining guests benefited.

GPU and CPU Utilization

Esri has worked hard to improve GPU us in ArcGIS Pro 1.0 and provided the following graph as a means to show that use. The red line is the general CPU utilization across the tests for the 16 virtual desktops while the white lines are GPU. As you can see graphics

traditionally have bursts in usage where users will receive as much of the GPU as required. In doing so the CPU is not asked to do that work, this offloading allows the CPU to focus on its tasks and in general the sharing of resources allows for more scalability.



Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. 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 publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification 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. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

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 these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do 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 specification. NVIDIA does not accept any 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 specification, 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 specification. 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 specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

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. 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 NVIDIA terms and conditions of sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

ROVI Compliance Statement

NVIDIA Products that support Rovi Corporation's Revision 7.1.L1 Anti-Copy Process (ACP) encoding technology can only be sold or distributed to buyers with a valid and existing authorization from ROVI to purchase and incorporate the device into buyer's products.

This device is protected by U.S. patent numbers 6,516,132; 5,583,936; 6,836,549; 7,050,698; and 7,492,896 and other intellectual property rights. The use of ROVI Corporation's copy protection technology in the device must be authorized by ROVI Corporation and is intended for home and other limited pay-per-view uses only, unless otherwise authorized in writing by ROVI Corporation. Reverse engineering or disassembly is prohibited.

OpenCL

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

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA GRID, and NVIDIA GRID vGPU™ 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