NVIDIA Virtual PC (vPC) Sizing Guide

Application Sizing Guide

| May 2021

Document History

nv-quadro-vpc-profile-sizing-guide-v3-072021

Version	Date	Authors	Description of Change
01	May 5, 2021	AS, AH	Updated document with RTX 6000, RTX 8000, and V100D
02	May 19, 2021	AS, AH	vGPU Channel Count
03	July	AH	Updated document with A16 and A10

Stakeholder Approvals

Stakeholder Name	Role	Date	Approver Comments & Version Approved
Judy Lee	PMM	7/15	Approved
Manvender Rawat	PM		Approved
Anne Bird	Enterprise Support	8/8/2022	Approved
Adi Padhye	Engineering		
Shailesh Deshmukh	SA		
Simon Schaber	SA	06/27/2022	Nothing to add
Justin Hodgson	Sales		

Table of Contents

Chapter 1. Executive Summary

This document provides insights into leveraging NVIDIA Virtual PC (vPC) for knowledge workers. It gives recommendations based on NVIDIA's nVector knowledge worker benchmarking and covers common questions such as:

- Which NVIDIA® GPU should I use for my business needs?
- How do I select the right NVIDIA virtual GPU (vGPU) profile(s) for the types of users I will have?
- What advantages of running NVIDIA vPC versus traditional CPU-only virtual desktop infrastructure (VDI)?

Knowledge worker workloads will vary per user depending on many factors, including applications, the types of applications, file sizes, and the number of monitors and their resolution. This document used NVIDIA's nVector as the testing framework for executing a typical knowledge worker workload that simulates application workflow and is a tool for capturing real-world metrics. Since the number of monitors and their resolution directly impact sizing, our testing to support this document explored various screen resolutions and the number of monitors. Tests were executed on CPU-only VM's as well as VM's with NVIDIA vGPU.

It is recommended that you test your unique workloads to determine the best NVIDIA virtual GPU solution to meet your needs. The most successful customer deployments start with a proof of concept (POC) 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 essential because user behavior can change throughout a project, and an individual's role changes within said organization. For example, a user 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 optimize their deployment for each user.

1.1 About NVIDIA nVector Benchmark

NVIDIA's performance engineering team developed a methodology and benchmarking tool which simulates, at scale, a knowledge worker workflow. This workflow is a good representation of commonly used software applications:

- Microsoft Word 2016
- Microsoft Excel 2016
- Microsoft PowerPoint 2016
- Google Chrome (32-bit) web browser and video streaming
- PDF document viewing

These applications will perform various functions throughout the test, replicating an end user's task. Microsoft Word, Excel, and PowerPoint create new content, modify existing content, and move content between applications. Tasks within these applications include scrolling, zooming, menu navigation, and PDF creation. Google Chrome streams live videos and visits interactive websites. Microsoft Edge acts as a PDF viewer.

When running the nVector benchmark at scale, nVector randomizes Knowledge Worker (KW) workloads across multiple virtual machines.

Simulating Many Users, Many Behaviors

User #1	User #2	User #3	User #4	...
Google Chrome (Video)	MS Word 2016	Windows Media Player	Google Chrome (Web)	...
Windows Media Player	Microsoft Edge (PDF)	MS Word 2016	Google Chrome (Video)	...
MS Word 2016	MS Excel 2016	Microsoft Edge (PDF)	Windows Media Player	...
Microsoft Edge (PDF)	Google Chrome (Web)	MS Excel 2016	MS Word 2016	...
MS Excel 2016	Google Chrome (Video)	Google Chrome (Web)	Microsoft Edge (PDF)	...

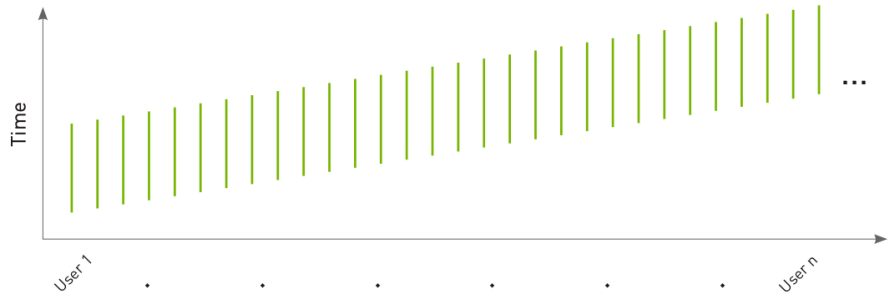


Figure 3. Characteristics of NVIDIA's Benchmarking Tool. The above table shows the workflow of each user. The graph shows cumulative increase in the number of users running workloads through time. Multiple users are tested at a time to simulate scale, with start and end times staggered to be more representative of real VDI environments.

1.2 What is NVIDIA vPC?

NVIDIA virtual PC (vPC) software enables the delivery of graphics-rich virtual desktops accelerated by NVIDIA GPUs. NVIDIA vPC allows sharing the same GPU across multiple virtual machines, delivering a native-PC experience to knowledge workers while improving user density. Because tasks typically done on the CPU are offloaded to the GPU, the user has a much better experience, and more users can be supported.

Virtual GPU profiles determine the amount of frame buffer allocated to your virtual machine. The vGPU profiles supported on NVIDIA GPUs with NVIDIA vPC software are 1B (with 1024 MB of frame buffer) and 2B (with 2048 MB of frame buffer). The 1B profile typically supports single and dual HD

configurations and is used when considering density. The 2B profile can utilize multi-monitor setups and is typically for higher resolution configurations. Using NVIDIA's nVector testing framework, we conducted extensive testing on various configurations with both profiles to give IT admins an idea of what to expect when they scale within their own environments. Because users work in applications with varying levels of utilization, performing a POC with your workload against the testing done within this document is recommended.

NVIDIA Feature List

Configuration and Deployment	vPC
Desktop Virtualization	✓
Remote Desktop Session Host (RDSH) App Hosting	✓
RDSH Desktop Hosting	✓
Windows OS Support	✓
Linux OS Support	✓[4]
GPU Pass-Through Support[5]	
Bare Metal Support[6]	
NVIDIA Graphics Driver	✓
Guaranteed Quality-of-Service Scheduling[7]	✓

Display	vPC
Maximum Hardware Rendered Display	Four HD, Two 4K[4], One 5K[13]
Maximum Resolution	5120 x 2880[13]

[4] Support starts with the NVIDIA virtual GPU software March 2018 release (version 6.0).

[13] 5K resolution support starts with the NVIDIA virtual GPU software December 2019 release (10.0).

NVIDIA vPC delivers an engaging user experience for the digital workplace. Employees can be most productive using modern applications and work the way they want, from anywhere. Delivering up to 50% better performance[1] over CPU-only VDI, NVIDIA vPC combined with A16 enables IT to cost-effectively scale virtualization to every employee with performance that rivals a physical PC.

1.3 Recommended NVIDIA GPUs for NVIDIA vPC

[1] Performance measured using NVIDIA nVector benchmark running knowledge worker workloads (Excel, Word, PowerPoint, Chrome, Media Player, PDF) running on dual 1920x1080 resolution displays with NVIDIA vPC (vGPU 13.0) and NVIDIA A16-1B measuring frames per second.

Density-optimized GPUs are typically recommended for virtual desktop users running office productivity applications, streaming video, and Windows 10. They are designed to maximize the number of VDI users supported in a server.

	NVIDIA A16	NVIDIA A10
# Boards [Architecture]	4 [Ampere]	1 (Ampere)
RT Cores	40 [4 x 10 per GPU]	72
Memory Size	64 GB GDDR6 [4 x 16GB per GPU]	24 GB GDDR6
Form Factor	PCIe 4.0 Dual Slot FHFL	PCIe 4.0 1-slot FHFL
Power	250W	150W
Thermal	Passive	Passive
Optimized for	Density	Performance
1B Users per Board	64	24

The NVIDIA A16 and NVIDIA A10 are based on the NVIDIA Ampere™ architecture. The NVIDIA A16 GPU offers the best user density option for NVIDIA vPC customers as well as the lowest cost per user. The A16 is a 64 GB (4x GPUs with 16 GB per card) dual-slot FHFL card that draws up to 250 W and is passively cooled. The NVIDIA A10 is a 24 GB Single slot FHFL card that draws 150 W maximum and is passively cooled.

While the A16 provides the best value for knowledge worker deployments, selecting the A10 brings the added benefits of enabling compute for AI in addition to accelerated graphics and video. This enables IT to maximize data center resources by running virtual workstations, deep learning inferencing, rendering, and other graphics and compute intensive workloads -- all leveraging the same data center infrastructure. This ability to run mixed workloads can increase user productivity by enabling your datacenter to run day and night. For example, IT can distribute resources by day for knowledge workers, and at night run rendering and compute workloads. This approach maximizes utilization and reduces costs in the data center. For additional information regarding A16 and A10 technology enhancements, please check out the A16 and A10 product features.

Note: For the ability to run mixed workloads with the A16 and A10, please note the appropriate software licenses within the NVIDIA Virtual GPU Software Packaging, Pricing, and Licensing Guide.

GPUs that leverage ECC memory are enabled by default. When enabled, ECC has an overhead cost due to using extra VRAM to store the ECC bits themselves. This will result in a lower frame buffer on the vGPU compared to the physical GPU. It is essential to resize your environment when switching from Maxwell, Pascal, and Turing GPUs; to newer GPUs like the A16. Additional information can be found here.

The maximum number of vGPUs created simultaneously on a physical GPU is defined by the amount of frame buffer per VM and how many VMs can share that physical GPU. For example, an NVIDIA A10 GPU with 24GB of GPU Memory can support up to 24 1B profiles (24 GB total with 1GB per VM). You cannot oversubscribe GPU memory, and it must be shared equally for each physical GPU.

Note: While the NVIDIA A40 has 48 GB of GPU memory, the maximum vGPUs per A40 GPU is limited to 32 for optimal performance.

The NVIDIA A16 is recommended by NVIDIA for vPC as this card is optimized for density. The complete list of NVIDIA GPUs that support vPC can be found here.

Chapter 2. Testing Methodology

2.1 Single VM Testing

The first phase of testing explored the impact of higher resolution and multi-monitor scenarios where the following tests were executed:

Resolution	Monitors
High Definition (HD) 1920x1080	1
	2
Quad High Definition (QHD) 2560x1440	2
	3
4K 4096x2160	1
	2
5K 5120x2880	1

Note: This table reflects the configurations that were tested, NOT our recommendations.

Tests were executed on a single VM using 1B and 2B vGPU profiles to determine the optimal vGPU profile based on the nVector KW workload.

Important: The nVector Knowledge worker workload is designed to simulate peak usage scenarios using the typical productivity apps where all the concurrent users are actively using the system resources simultaneously. These results are meant to give administrators an outline with which to plan POC deployments. Workloads within your environment might be less resource intensive than the nVector knowledge workload.

2.1.1 Test Environment

Single VM testing leveraged two physical servers, the first hosting the target vPC VMs and the second hosting the virtual clients. Both server hosts used VMware vSphere ESXi 7.0.2 and NVIDIA Virtual GPU Software. The target VM acts as a standard vPC VDI that an end-user would connect to, and the virtual client serves as an example of an endpoint that the end-user would use to connect to the target VM.

Host Configuration	VM Configuration	Virtual Client
2U NVIDIA-Certified Server	vCPU: 4	vCPU: 2
Intel® Xeon® Gold 6248 @ 3.00 GHz	vRAM: 6144 MB	vRAM: 4096 MB
VMware ESXi, 7.0.2, 17630552	NIC: 1 (vmxnet3)	NIC: 1 (vmxnet3)
Number of CPUs: 40 (2 x 20)	Hard disk: 48 GB	Hard disk: 48 GB
Memory: 766 GB	Virtual Hardware: vmx-13	Virtual Hardware: vmx-13
Storage: Local Flash	VMware Horizon 8.12	VMware Horizon 8.12
Power Setting: High Performance	Blast Extreme (4:4:4)	Blast Extreme (4:4:4)
GPU: A16	vGPU Software: 13 (Windows Driver 471.68)	vGPU Software: 13 (Windows Driver 471.68)
Scheduling Policy: 0x00 (Best Effort)	Guest OS: Windows 10 Pro 20H2	Guest OS: Windows 10 Pro 20H2

2.1.2 Test Metrics - Framebuffer Usage

Frame buffer utilization is based upon many factors, including application load, monitor configuration, and screen resolution. Since our test focuses on the impact of higher resolutions and multi-monitor scenarios, frame buffer utilization is a critical test metric.

2.1.2.1 GPU Profiler

GPU Profiler is a commonly used tool that can quickly capture resource utilization while a workload is being executed on a virtual machine. This tool is typically used during a POC to help size the virtual environment to ensure acceptable user performance. GPU Profiler was run on a single VM with various vGPU profiles while the nVector knowledge worker workload ran. The following metrics were captured:

- Framebuffer %
- vCPU %
- RAM %
- Video Encode

- Video Decode

A good rule of thumb to follow is that frame buffer utilization should not exceed **90%** for a short time or an average of over **70%** on the 1GB (1B) profile. If high utilization is noted, the vPC VM should be assigned a 2GB (2B) profile. These results are reflective of the work profile mentioned in section 1.1. Due to users leveraging different applications with varying degrees of utilization, we recommend performing a POC within your internal environment.

2.2 Scalability Testing

Typical VDI deployments have two conflicting goals: Achieving the best possible user experience and maximizing user density on server hardware. Problems can arise as density is scaled up because it can negatively impact user experience after a certain point. Scalability testing used nVector to execute tests at scale on 64 and 128 VMs while leveraging dual HD (1920x1080) monitors. Capacity planning for the server is often dependent upon server resource utilization metrics and user experience. This testing phase examined both, and the following sections summarize their importance and how to analyze these metrics.

2.2.1 Server Utilization Metrics

Observing overall server utilization will allow you to assess the trade-offs between end-user experience and resource utilization. To do this, monitoring tools periodically sample CPU core and GPU utilization during a single workload session. To determine the ‘steady state’ portion of the

workload, samples are filtered, leaving out when users have all logged on, and the workload ramps up and down. Once a steady state has been established, all samples are aggregated to get the total CPU core utilization on the server.

The utilization of the GPU compute engine, the frame buffer, the encoder, and the decoder can all be monitored and logged through NVIDIA System Management Interface (nvidia-smi), a command-line interface tool. In addition, NVIDIA vGPU metrics are integrated through management packs like VMware vRealize Operations. For our testing purposes, nVector automated the capture of the following server metrics. Since it is highly recommended that you test your unique workloads during a POC, nvidia-smi commands can run on the hypervisor, which will allow you to monitor GPU utilization of the physical GPU. Please refer to Section 4: Deployment Best Practices for further syntax information.

2.2.2 User Experience Metrics

NVIDIA's nVector benchmarking tool has built-in mechanisms to measure user experience. This next section will dig deeper into how the end-user experience is measured and how results are obtained.

2.2.2.1 Latency Metrics

Latency defines the response or feel of the end-user when working with applications in the VDI. Increased latency can provide a poor experience, including mouse cursor delay, text display issues when typing, and audio/video sync issues. The lower latency, the better! Imagine that you are working on a PowerPoint, adding a shape, and resizing it. On the first attempt, this process is instantaneous. However, the second attempt is delayed by several seconds or is sluggish. With such inconsistency, the user tends to overshoot or have trouble getting the mouse in the correct position. This lack of a consistent experience can be very frustrating. Often, it results in the user experiencing high error rates as they click too fast or too slow, trying to pace themselves with an unpredictable response time. NVIDIA's nVector benchmarking tool measures the variation in end-user latency and how frequently it is experienced.

2.2.2.2 Remoted Frames Metrics

Frame rate metrics are captured on the endpoint and provides an excellent metric on the possible end-user experience. The average frame rate is captured and calculated across the simulated workload. A lower frame rate can cause slow response during screen refresh and stuttering during scrolling or zooming. The higher the frame rate, the better!

Remoted frames are a standard measure of user experience. NVIDIA's nVector benchmarking tool collects data on the 'frames per second' provided by the remote protocol vendor for the entire workload duration. The tool then tallies the data for all VDI sessions to get the total number of frames remoted for all users. Hypervisor vendors likewise measure total remoted frames as an indicator of the quality of user experience. The greater this number, the more fluid the user experience.

2.2.2.3 Image Quality

Image quality is determined by the remoting protocol, configuration of the VDI environment, and endpoint capability. For this sizing guide, the protocol used is VMware Blast Extreme with High Color

Accuracy (HCA) YUV444, a protocol specific to vPC use cases. HCA no longer removes any chroma information from images and provides a much better image quality.

NVIDIA's nVector benchmarking tool uses a lightweight agent on the VDI desktop and the client to measure image quality. These agents take multiple screen captures on the VDI desktop and on the thin client to compare later. The structural similarity (SSIM) of the screen capture taken on the client is computed by comparing it to the one taken on the VDI desktop. When the two images are similar, the heatmap will reflect more colors above the spectrum shown on its right with an SSIM value closer to 1.0 (Figure 9). As the images become less similar, the heatmap will reflect more colors down the spectrum with a value less than 1.0. More than a hundred pairs of images across an entire set of user sessions are obtained. The average SSIM index of all pairs of images is computed to provide the overall remote session quality for all users. The threshold SSIM value is 0.98; scores above 0.98 indicate good image quality, with 1.0 perfect.

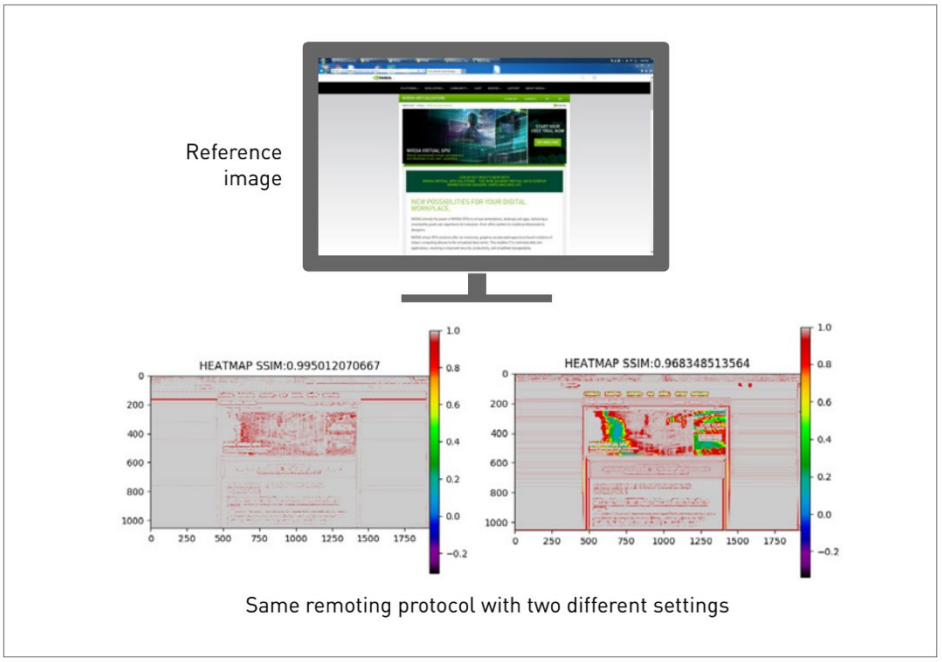


Figure 9. SSIM as a Measure of Image Quality

Chapter 3. Test Findings

3.1 Single VM Multi-Monitor Resolution Test Results

The following table summarizes the results of single VM testing configurations, where we explored the impact on frame buffer (FB) for higher resolution and multi-monitor scenarios based upon the nVector KW workload. As monitor resolutions continue to increase, more pixels are being delivered to the screen. As a result, the frame buffer usage in a virtual environment increases. While HD (1920x1080) is currently the most common resolution, an increasing number of devices are being released with higher resolution screens.

nVector Knowledge Worker Workload Test Results*		
Resolution	Monitors	A16 vGPU Profile
High Definition 1920x1080	1	1B
	2	2B
Quad High Definition 2560x1440	2	2B
	3	2B
4K 4096x2160	1	2B
	2	2B
5K 5120x2880	1	2B
*Based upon benchmark testing, test your workloads to ensure FB sizing is appropriate for your users.		

Knowledge worker workloads will vary per user depending on many factors, including use of multiple applications, the types of applications utilized, file sizes, in addition to the number of monitors and their resolution. Additional monitor and resolution support, including mixed displays, can be found

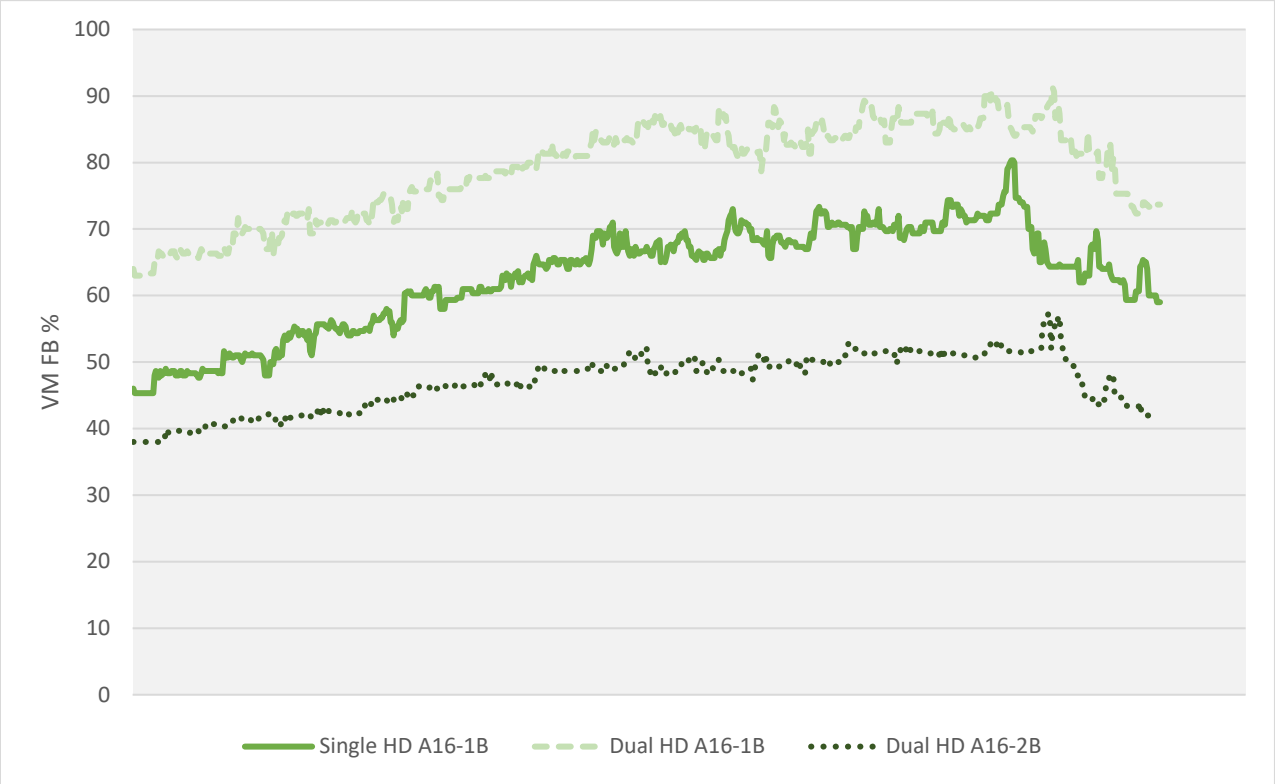
here. It is highly recommended that you test your workloads during a POC since mileage may vary. Our nVector test results should be used for guidance purposes only.

> **!** Important: The nVector Knowledge worker workload is designed to simulate heavy usage scenarios using the typical productivity Apps where all the concurrent users are actively using the system resources simultaneously. These results are meant to give administrators an outline with which to plan POC deployments. Workloads within your environment might be less resource intensive than the nVector knowledge workload.

The results of Single VM frame buffer analysis results are used for sizing purposes since the maximum number of vGPUs that can be created (and then assigned to a VM) is defined by the amount of GPU memory per VM. The following sections describe the frame buffer usage captured on the VM for the nVector KW workload.

3.1.1 High Definition (1920x1080) Displays

When the number of monitors is increased, more pixels are being delivered to the screen. Our nVector KW workload reported an average 15% increase in FB Usage when monitors were increased from 1 to 2. The following graphs demonstrate the framebuffer utilization on single and dual HD monitors while a nVector KW workload was executed:



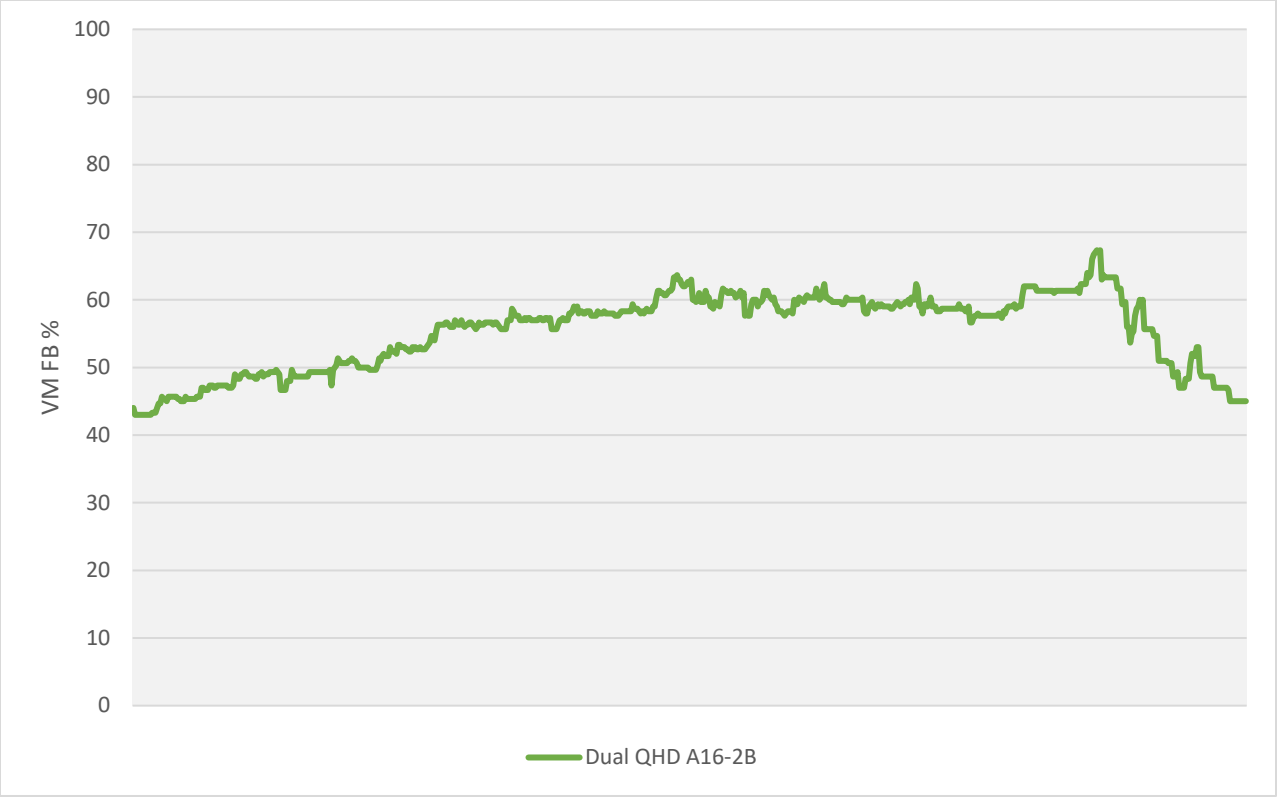
3.1.2 Quad High Definition (2560x1440) Displays

Quad High Definition (2560x1440) resolution tests were executed using 2, 3, and 4 monitors. Quad high definition (QHD) has almost double as many pixels as HD; therefore, FB requirements for QHD

monitors are greater than HD. Overall, our nVector test results for the KW workload illustrate that the 2B profile was sufficient for 2 QHD monitors. When monitors were increased from 2 to 3, there was a 15% increase in FB usage. The 2B profile provided an adequate amount of FB for 2 and 3 QHD monitors. The following sections portray our test findings with nVector.

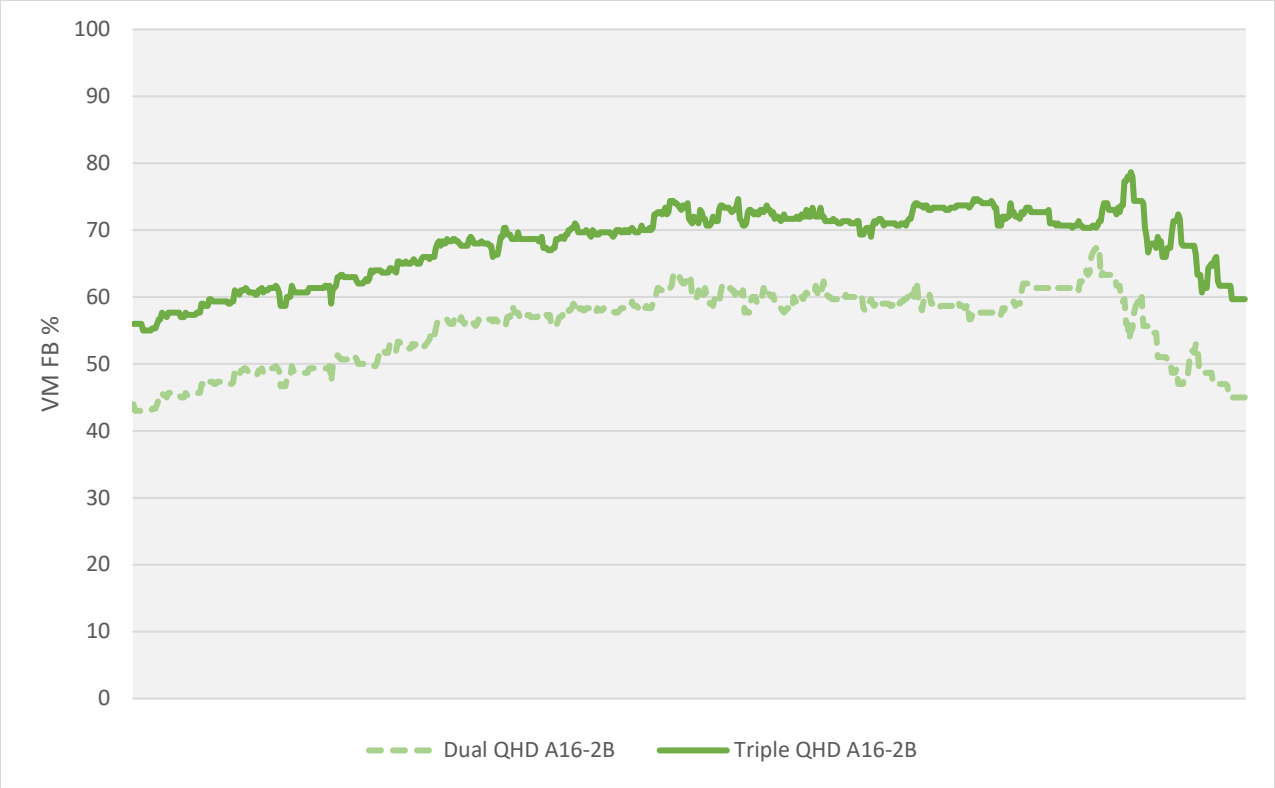
3.1.2.1 Dual QHD Monitor Test Results

The following graph illustrates the frame buffer usage captured while executing the nVector KW workload on dual QHD monitors.



3.1.2.2 Triple QHD Monitor Test Results

The following graph illustrates the impact on frame buffer when monitors are increased from 2 to 3 QHD monitors in separate nVector KW workload tests.



3.1.3 4K (4096x2160) Displays

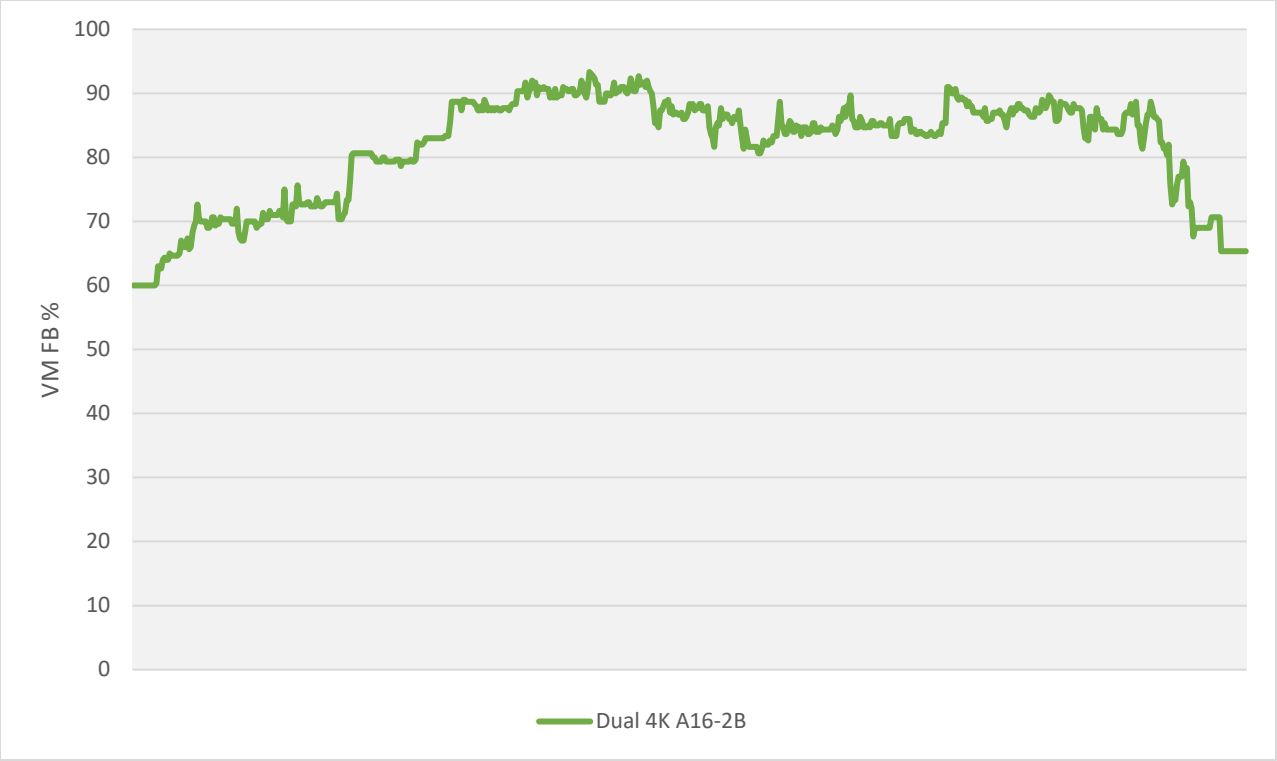
Tests were executed using a single 4K monitor as well as a dual 4K monitor. The nVector KW workload test results illustrated that a 2B vGPU profile for a single 4K monitor was sufficient; however, it certainly utilized the FB. Based upon this information, when the number of monitors was increased to two 4K monitors, admins will need to best determine whether a 2B profile can support Dual 4K displays at desired performance levels.

3.1.3.1 Single 4K Monitor Test Results

The following graph illustrates the FB usage of the nVector KW workload using the 2B profile for a single 4K monitor:

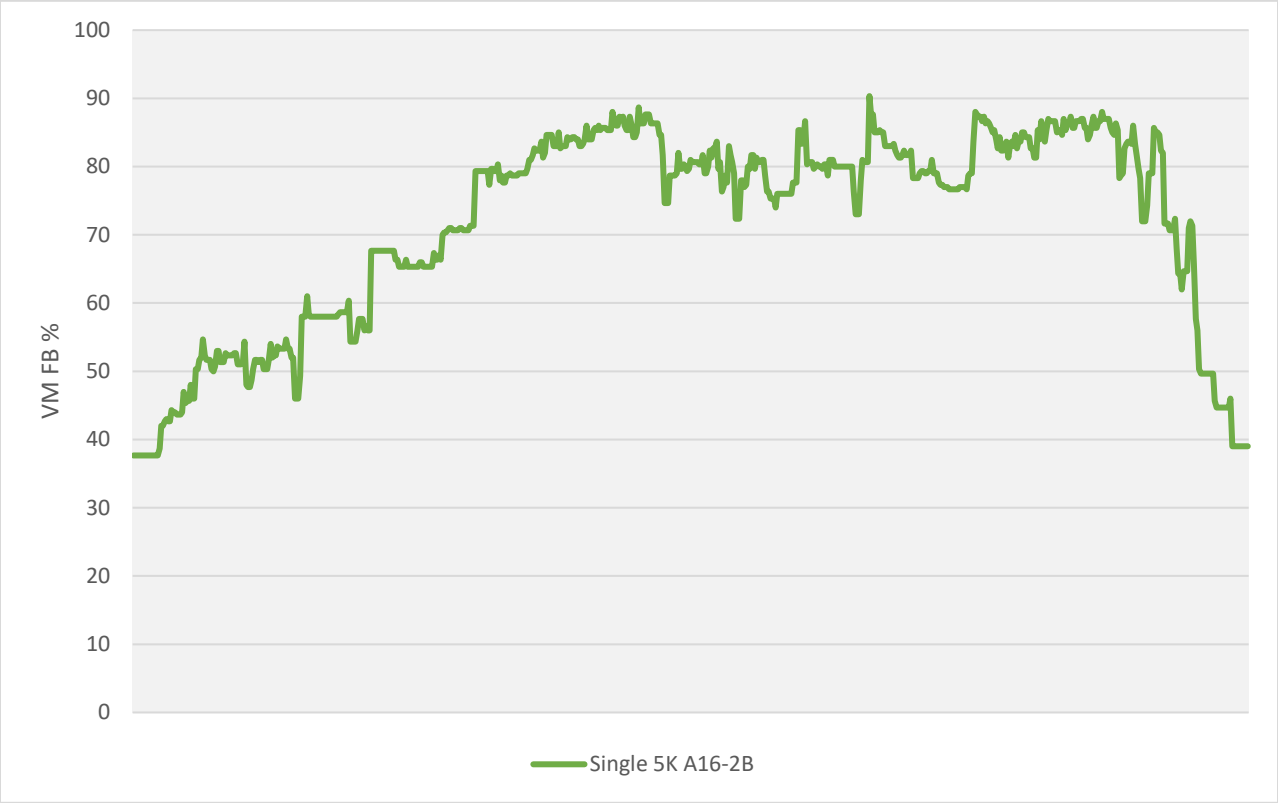
3.1.3.2 Dual 4K Monitor Test Results

The following graph illustrates the FB usage of the nVector KW workload using the 2B profile for a Dual 4K monitor configuration:



3.1.4 5K (5120x2880) Display:

NVIDIA vPC supports only a single 5K monitor, and for our nVector KW workload, the 2B vGPU profile was chosen based upon the known FB requirements of 5K resolution. 5K displays have a resolution of about seven times the pixels than high-definition displays (1920x1080). The following graph illustrates FB Usage for the nVector KW workload:



3.1.5 Single VM Multi-Monitor Resolution Summary

To recap, it is important to remember when sizing your environment, that a good rule of thumb to follow is that frame buffer utilization should not exceed **90%** for a short time or an average of over **70%** on the 1GB (1B) profile. If high utilization is noted, the vPC VM should be assigned a 2GB (2B) profile. In addition, when deciding whether to increase monitors for your configuration, our nVector KW workload reported an average 15% increase in FB Usage when monitors were increased from 1 to 2.

For a single HD or Dual HD configuration, the A16-1B will be adequate for most knowledge workers. However, if an A16-1B profile is not meeting organizational needs for a Dual HD configuration, switching to the A16-2B profile will be adequate.

Quad high definition (QHD) has almost double as many pixels as HD; therefore, FB requirements for QHD monitors are greater than HD. For QHD configurations our nVector test results show that the 2B profile was sufficient for 2 and 3 QHD monitors.

For a single 4K configuration, the 2B profile was sufficient however, it certainly utilized the FB. Based upon this information, when the number of monitors was increased to two 4K monitors, admins will need to best determine whether a 2B profile can support Dual 4K displays at desired performance levels given user application utilization.

With a single 5K monitor, an A16-2B profile can be utilized, however it should be noted again that admins will need to conduct internal testing to see if a single 5K monitor can support their respective workloads.

Important: The nVector Knowledge worker workload is designed to simulate heavy usage scenarios using the typical productivity apps where all the concurrent users are actively using the system resources simultaneously. These results are meant to give administrators an outline with which to plan POC deployments. Workloads within your environment might be less resource intensive than the nVector knowledge workload.

3.2 Multi-Monitor Resolution Scalability Test Results

Running a single VM on a large environment does not allow you to capture the usage of a production environment. Since full HD is currently the most common resolution, our scalability testing results within this document focus on dual HD (1920x1080) monitors. It is highly recommended that you test your workloads during a POC since mileage may vary. Our nVector test results should be used for guidance purposes only.

For our server utilization testing, the scale was configured for 64 VMs on the ESXi host with an NVIDIA T4, contrasting CPU core utilization versus GPU utilization to show CPU offload with an NVIDIA GPU.

To demonstrate user experience, tests were conducted with identical CPUs and server configurations for a 128 VM test with two NVIDIA A16's contrasted against a 64 VM test with four NVIDIA T4's. Within nVector's user experience metrics, the 128 VM test with two A16's has similar levels of performance metrics as the 64 VM test with four T4's. With an identical CPU, the NVIDIA A16 has effectively doubled our density while providing a consistent user experience in line with the previous generation NVIDIA T4.

The following table summarizes the multi-monitor high-resolution test environment as well as how many NVIDIA GPUs were used for each scalability test:

# of VMs at Scale	# GPUs Cards	VGPU Profile	Monitor Resolution	# of Monitors
64	4	T4-1B	1920x1080	2
128	2	A16-1B	1920x1080	2

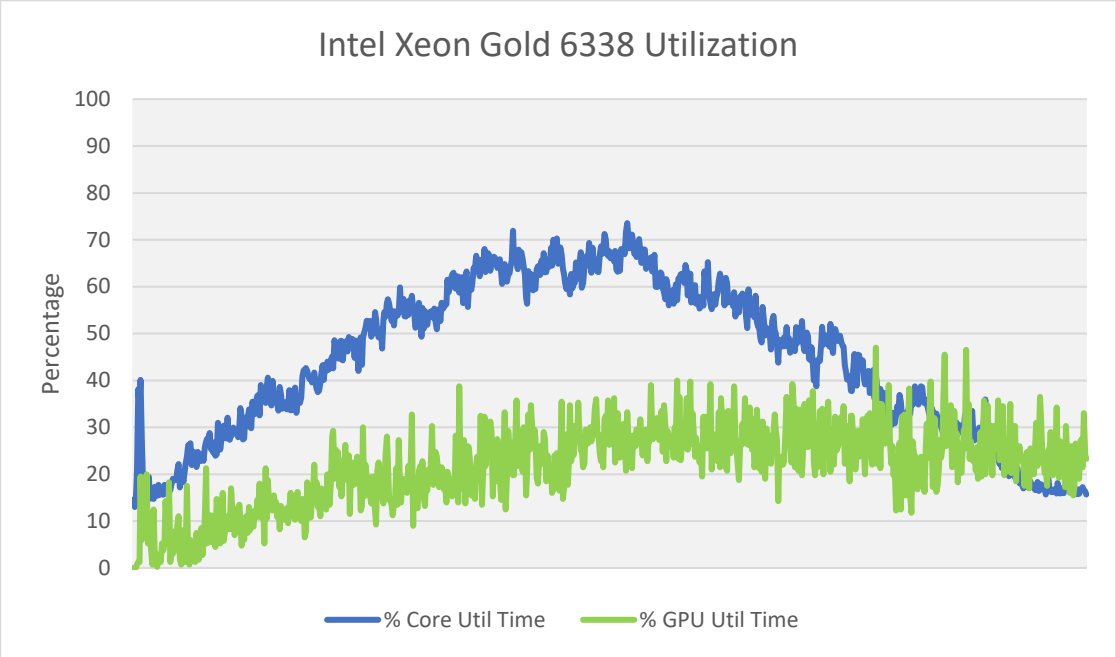
During this process, the benchmark was used to execute various nVector KW workflows across multiple VMs with start and end times staggered across the environment.

3.2.1 Server Utilization Metrics

Choosing the correct CPU for virtualization and proper configuration can directly affect scalability even when a virtual GPU is present. Processor resources are often hyper-threaded and

overprovisioned to a certain degree. In terms of CPU specs, you should evaluate the number of cores and clock speed. The following paragraphs describe our test findings when using 64 VMs:

The following graph illustrates CPU Core utilization using an Intel Xeon Gold 6338 3.2 GHz Turbo (Skylake) for 64 VMs. This server configuration has 64 cores with hyperthreading enabled.



This graph of the 64 VM T4 test shows that as the GPU utilization ramps up during the nVector KW workload test, CPU Core utilization is offloaded and drops accordingly.

3.2.2 nVector User Experience Metrics

To further assess the trade-offs between end-user experience and resource utilization, we used nVector's built-in mechanisms to measure user experience. The following sections describe our findings for the nVector KW workload for a 128 A16 vPC VM test contrasted with a 64 T4 vPC VM test using identical server specs and CPU's.

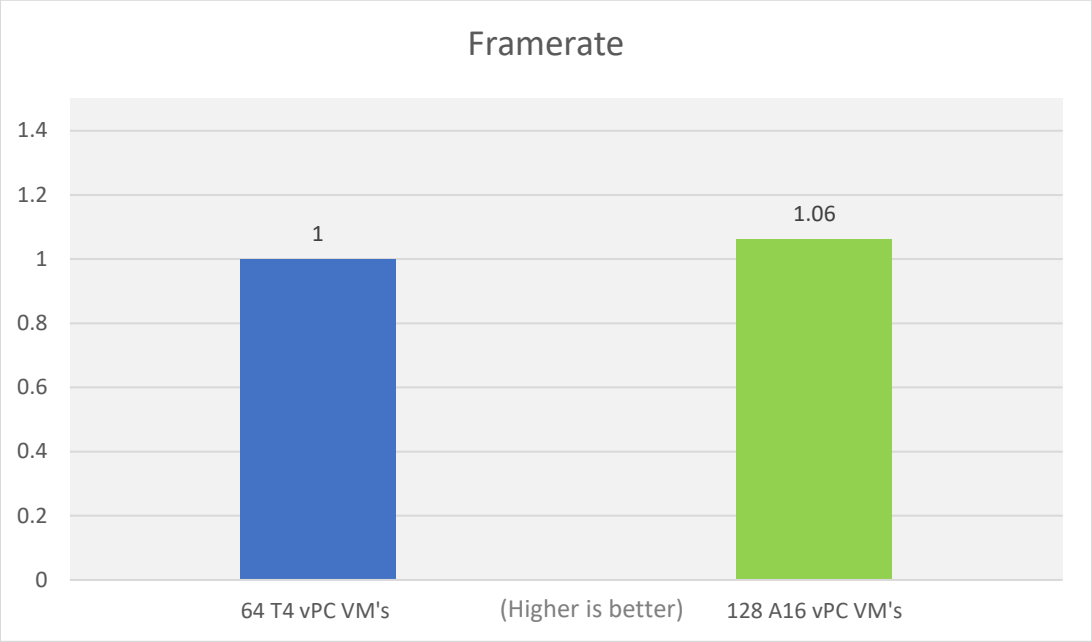
Our results show a consistent user experience and performance across both the NVIDIA A16 and NVIDIA T4, however the NVIDIA A16 has doubled our density in the server to 128 VM's.

3.2.2.1 Frame Rate

The nVector benchmark tool captures frame rate, which provides an excellent metric in determining the end-user experience. Providing a consistent and high frame rate can lead to a smoother experience for the user, while an inconsistent frame rate will create a less than acceptable experience.

The following graph illustrates the frame rate differences for Dual High Definition 1920x1080 monitors while running the nVector KW workload. The average frame rates were nearly identical

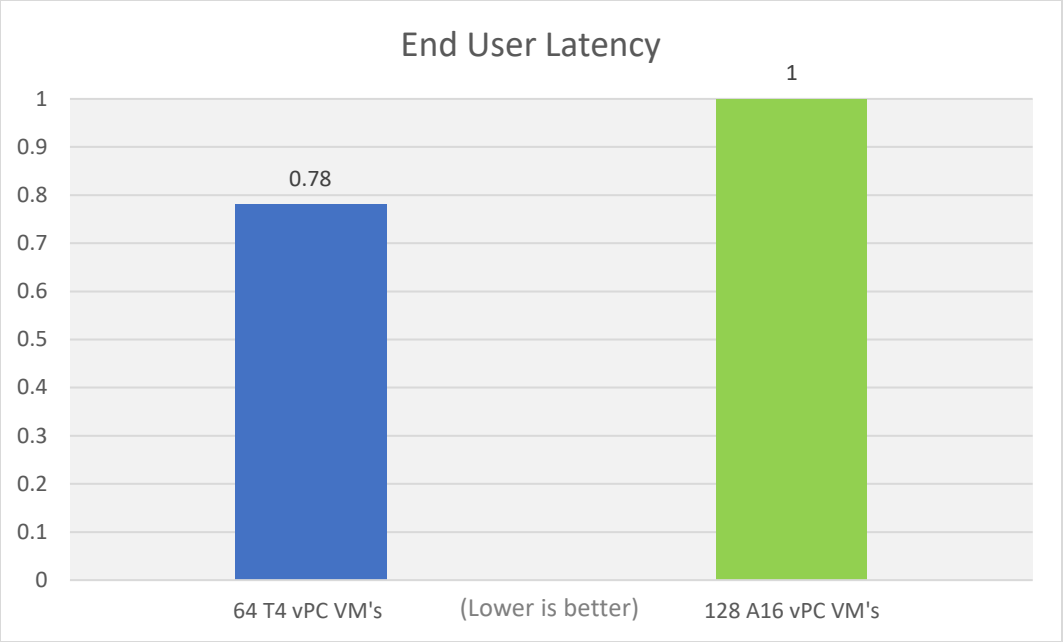
between the NVIDIA A16 and NVIDIA T4, with a slight edge going to the A16.



3.2.2.2 Latency Metrics

Another critical metric captured by the nVector benchmark tool is latency or, in this case, end-user latency. Latency can affect mouse speed, characters showing up on the screen behind what is typed, and poor video playback.

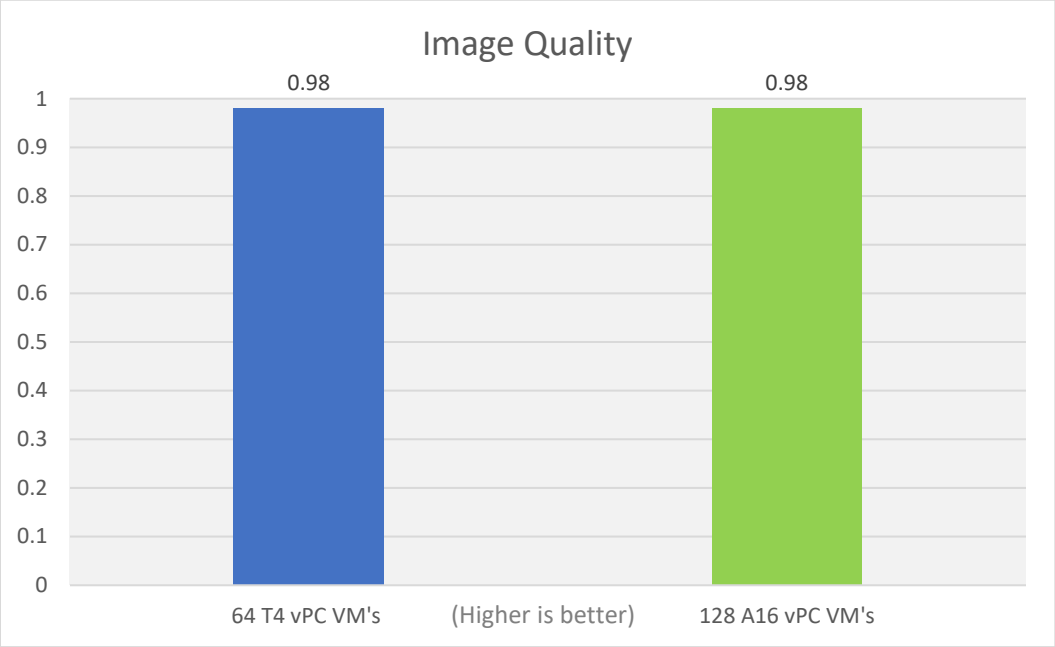
The following graph illustrates end-user latency for Dual High Definition 1920x1080 monitors while running the nVector KW workload. User latency had a 22 percent increase with the A16 when compared with the T4. However, the NVIDIA A16 was able to double our density within the server. Striking a balance between performance and density is best determined by admins when conducting an internal POC to appropriately address the needs of the organization.



3.2.2.3 Image Quality

The nVector benchmark tool calculates image quality. It is determined by the remoting protocol, the configuration, and policies set in the VDI environment (please refer to Appendix A regarding the configuration used within our testing). Poor image quality, under .90, can cause issues with text display, line sharpness, and other graphical issues.

Our nVector testing illustrates that GPU-accelerated VM's with vPC deliver uncompromised image quality as SSIM of the screen capture using Dual High Definition 1920x1080 monitors. Both the NVIDIA A16 and NVIDIA T4 tests for vPC VM's reported higher than the 0.90 thresholds at an exceptional .98 SSIM.



3.2.3 Multi-Monitor Resolution Scalability Summary

To recap our multi-monitor resolution scalability tests, we used identical CPUs and server configurations for a 128 VM test with two NVIDIA A16's contrasted against a 64 VM test with four NVIDIA T4's. User experience metrics such as framerate, end-user latency, and image quality were noted to have similar performance levels in the NVIDIA A16 when compared to the NVIDIA T4. However, with an identical CPU, the NVIDIA A16 effectively doubled our density while providing a consistent user experience in line with the previous generation NVIDIA T4.

During the 64 VM T4 test to assess CPU offload by the GPU, GPU utilization was effectively shown to ramp up during the nVector workload in tandem with the CPU Core utilization being offloaded and dropping accordingly.

Admins should expect to reliably scale an NVIDIA Certified Server with 2x NVIDIA A16's to 128 VM's while using a Dual HD configuration for optimal density. In addition, admins should expect similar performance with 4x NVIDIA T4's in combination with an NVIDIA Certified Server in a Dual HD configuration scaling to 64 VMs.

Chapter 4. Deployment Best Practices

4.1 Run a Proof of Concept

The most successful deployments strike a balance between user density (scalability) and quality user experience. This is achieved when vPC virtual machines are used in production while objective measurements for adequate density sizing and subjective feedback from end-user's experience are gathered.

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

4.2 Leverage Management and Monitoring Tools

NVIDIA vPC on NVIDIA GPUs provides extensive monitoring features, enabling IT to use the various engines of an NVIDIA GPU. The utilization of the compute engine, the frame buffer, the encoder, and the decoder can all be monitored and logged through a command-line interface tool 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 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"
```

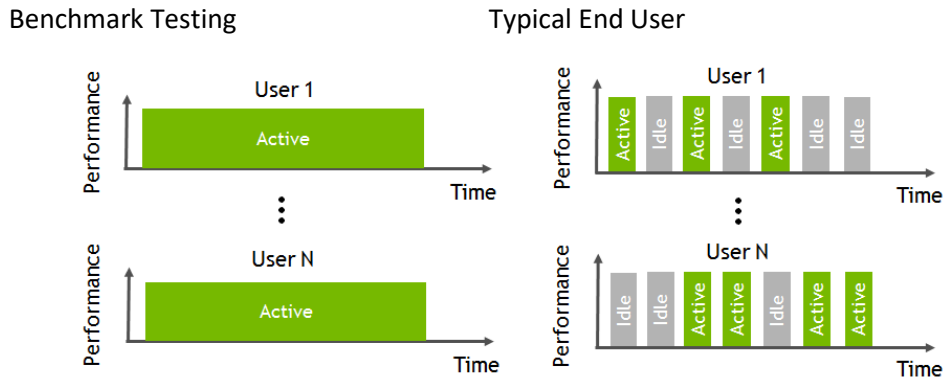
4.3 Understand Your Users

Another benefit of performing a POC before deployment is that it enables a 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, large profile size, and, may be best kept on an upgraded vGPU license like NVIDIA RTX Virtual Workstation (RTX vWS).

4.4 Use Benchmark Testing

Benchmarks like nVector can be used to help size a deployment, but they have some limitations. The nVector benchmarks simulate peak workloads with the highest demand for GPU resources across all virtual machines. The benchmark does not account for the times when the system is not fully utilized. Hypervisors and the best effort scheduling policy can be leveraged 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. The degree to which higher scalability is achieved depends on your users' typical day-to-day activities, such as the number of meetings and the length of lunch or breaks, multi-tasking, etc. It is recommended to test and validate your internal workloads to meet the needs of your users.



NVIDIA used the nVector benchmarking engine to conduct vGPU testing at scale. This benchmarking engine automates the testing process from provisioning virtual machines, establishing remote connections, executing KW workflow, and analyzing the results across all virtual machines. Test

results shown in this application guide are based on the nVector KW benchmarks run in parallel on all virtual machines with metrics averaged.

> **!** CAUTION: When done well, benchmarking can improve an organization's processes and overall performance. However, there are a plethora of pitfalls to the use of benchmark testing, especially social benchmarking, as the results of an organization's internal POC may yield different results than the social benchmark. This is due to a variety of reasons such as configuration differences, unreliable or incomplete data, lack of proper framework for standardized testing, etc. Conducting an internal POC with benchmark testing is valuable, but it can provide an incomplete picture when not paired with organizational strategy and goals.

4.5 Understanding the GPU Scheduler

NVIDIA vPC provides three GPU scheduling options to accommodate a variety of QoS requirements of customers. Additional information regarding GPU scheduling can be found here.

- **Fixed share scheduling** guarantees the same dedicated quality of service at all times.
- **Best effort scheduling** provides consistent performance at a larger scale and therefore reduces the TCO per user.
- **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.

4.6 Understanding GPU Channels

A physical GPU has a fixed number of channels. The number of channels allocated to each vGPU is proportional to the maximum number of vGPUs allowed on the physical GPU. Issues occur when the channels allocated to a vGPU are exhausted, and the guest VM to which the vGPU is assigned fails to allocate a channel to the vGPU.

To remove channel errors, use a vGPU type with more frame buffer, which reduces the maximum number of vGPUs allowed on the physical GPU. As a result, the number of channels allocated to each vGPU is increased.

Pascal, Volta, and Turing cards have 4096 channels per physical card, and the Ampere cards have a total of 2048. The number of channels per profile can be found with this formula:

Divide the channel count by the maximum number of vGPUs per GPU and subtract eight channels required for overhead.

For example, the maximum number of 1B vGPU profiles on an NVIDIA T4 is 16. The 4096 channels are divided by 16, which provides 256 channels per VM before overhead. Next, 8 channels are subtracted from the 256 channels, which offers 248 channels per VM.

(4096/16)-8 = 248 channels per VM

Example channel allocation per GPU Profile:

GPU profile	Channels per VM
A16-1B	120 channels per VM (16 x 4 card = 64 max)
A16-2B	248 channels per VM
A10-1B	77 channels per VM (24 max)
A10-2B	162 channels to each VM
A40-1B	56 channels to each VM (32 max)
A40-2B	77 channels to each VM
T4-1B	248 channels to each VM (16 VM)
T4-2B	504 channels to each VM

Channel utilization is related to single and multi-threaded applications running in a vGPU VM in addition to OS and boot overhead. Therefore, different apps will have different channel utilization.

When the channels allocated to a vGPU are exhausted, and the guest VM fails to allocate a channel, the following errors are reported on the hypervisor host or in an NVIDIA bug report:

```
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): Guest attempted to
allocate channel above its max channel limit 0xfb
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): VGPU message 6
failed, result code: 0x1a
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
0xc1d004a1, 0xff0e0000, 0xff0400fb, 0xc36f,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1,
0xff1fe314, 0xff1fe038, 0x100b6f000, 0x1000,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
0x80000000, 0xff0e0200, 0x0, 0x0, (Not logged),
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1, 0x0
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): , 0x0
```

Chapter 5. Summary

When sizing an NVIDIA vPC deployment for Knowledge Workers, NVIDIA recommends conducting a POC and thoroughly 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 Digital Knowledge Worker workloads using NVIDIA vPC software, try it for free. Learn more about NVIDIA vPC here.

Appendix A. Framebuffer Utilization Master List

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A10-1B Profile - 1GB of Frame Buffer						
Knowledge	A10	1920x1080	1	84.33% (863 MB)	68.24% (698 MB)	53% (542 MB)
Knowledge	A10	1920x1080	2	94% (962 MB)	83.46% (854 MB)	77% (788 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A10-2B Profile - 2GB of Frame Buffer						
Knowledge	A10	(2560x1440) QHD	2	72% (1475 MB)	61.64% (1262 MB)	48% (983 MB)
Knowledge	A10	(2560x1440) QHD	3	84% (1720 MB)	75.91% (1555 MB)	62% (1270 MB)
Knowledge	A10	4096x2160 (4K)	1	85% (1740 MB)	63.46% (1299 MB)	39% (798 MB)
Knowledge	A10	5120x2880 (5K)	1	91.67% (1877 MB)	68.95% (1412 MB)	41% (839 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A16 -1B Profile - 1GB of Frame Buffer						
Knowledge	A16	1920x1080	1	80% (819 MB)	62.5% (640 MB)	45% (461 MB)
Knowledge	A16	1920x1080	2	91% (932 MB)	78.15% (800 MB)	63% (645 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A16 -2B Profile - 2GB of Frame Buffer						
Knowledge	A16	1920x1080	2	57% (1167 MB)	46.98% (962 MB)	38% (778 MB)
Knowledge	A16	(2560x1440) QHD	2	67% (1372 MB)	55.42% (1135 MB)	43% (881 MB)
Knowledge	A16	(2560x1440) QHD	3	79% (1618 MB)	67.71% (1387 MB)	57% (1167 MB)
Knowledge	A16	(4096x2160) 4K	1	81% (1659 MB)	64.99% (1331 MB)	39% (799 MB)
Knowledge	A16	(4096x2160) 4K	2	93% (1905 MB)	81.62% (1672 MB)	60% (1229 MB)
Knowledge	A16	(5120x2880) 5K	1	90% (1843 MB)	71.71% (1469 MB)	37% (758 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A40-1B Profile - 1GB of Frame Buffer						
Knowledge	A40	1920x1080	1	80.33% (822 MB)	63.66% (651 MB)	50% (512 MB)
Knowledge	A40	1920x1080	2	94% (962 MB)	80.19% (821 MB)	74% (757 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC A40-2B Profile - 2GB of Frame Buffer						
Knowledge	A40	1920x1080	2	72% (1475 MB)	61.01% (1249 MB)	49% (1003 MB)
Knowledge	A40	(2560x1440) QHD	2	74.67% (1529 MB)	64.83% (1328 MB)	52.33% (1072 MB)
Knowledge	A40	(2560x1440) QHD	3	88.33% (1802 MB)	80.03% (1639 MB)	68% (1392.64 MB)
Knowledge	A40	4096x2160 (4K)	1	86.67% (1775 MB)	65.54% (1342 MB)	43% (880 MB)
Knowledge	A40	(4096x2160) 4K	2	95% (1946 MB)	85.66% (1761 MB)	71% (1454 MB)
Knowledge	A40	5120x2880 (5K)	1	92.33% (1890 MB)	71.14% (1456 MB)	45% (921 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC M10-1B Profile - 1GB of Frame Buffer						
Knowledge	M10	1920x1080	1	68% (696 MB)	50.62% (518 MB)	28% (287 MB)
Knowledge	M10	1920x1080	2	69.67% (714 MB)	52.86% (541 MB)	40% (410 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC T4-1B Profile - 1GB of Frame Buffer						
Knowledge	T4	1920x1080	1	74% (758 MB)	61.37% (628 MB)	47% (481 MB)
Knowledge	T4	1920x1080	2	90% (922 MB)	76.07% (779 MB)	65% (666 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC T4-2B Profile - 2GB of Frame Buffer						
Knowledge	T4	2560x1440 (QHD)	2	68% (1,393 MB)	53.93% (1,104 MB)	42% (860 MB)
Knowledge	T4	2560x1440 (QHD)	3	88% (1,802 MB)	77.33% (1,583 MB)	53% (1,085 MB)
Knowledge	T4	4096x2160 (4K)	1	82.33% (1,686 MB)	59.96% (1,228 MB)	37% (756 MB)
Knowledge	T4	4096x2160 (4K)	2	95% (1,946 MB)	77.08% (1,579 MB)	58% (1,188 MB)
Knowledge	T4	5120x2880 (5K)	1	96% (1,966 MB)	71.52% (1,465 MB)	39% (799 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC RTX6000p-1B Profile - 1GB of Frame Buffer						
Knowledge	RTX6000p	1920x1080	1	89.33% (915 MB)	71.2% (729 MB)	56.39% (577 MB)
Knowledge	RTX6000p	1920x1080	2	95.67% (980 MB)	85.87% (879 MB)	75% (768 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC RTX6000p-2B Profile - 2GB of Frame Buffer						
Knowledge	RTX6000p	2560x1440 (QHD)	2	73.33% (1,501 MB)	61.87% (1,267 MB)	51% (1,044 MB)
Knowledge	RTX6000p	2560x1440 (QHD)	3	85.67% (1,755 MB)	75.2% (1,540 MB)	64% (1,311 MB)
Knowledge	RTX6000p	4096x2160 (4K)	1	85% (1, 741 MB)	67.55% (1,383 MB)	45% (922 MB)
Knowledge	RTX6000p	4096x2160 (4K)	2	95.33% (1,952 MB)	81.08% (1,661 MB)	66% (1,352 MB)
Knowledge	RTX6000p	5120x2880 (5K)	1	92.33% (1,891 MB)	72.93% (1,494 MB)	48% (983 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC RTX8000p-1B Profile - 1GB of Frame Buffer						
Knowledge	RTX8000p	1920x1080	1	89.33% (915 MB)	71.74% (735 MB)	59% (604 MB)
Knowledge	RTX8000p	1920x1080	2	98% (1,004 MB)	86.8% (889 MB)	81.24% (832 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC RTX8000p-2B Profile - 2GB of Frame Buffer						
Knowledge	RTX8000p	2560x1440 (QHD)	2	73.67% (1,509 MB)	62.23% (1,274 MB)	50% (1024 MB)
Knowledge	RTX8000p	2560x1440 (QHD)	3	77.67% (1,591 MB)	63.56% (1,302 MB)	50.03% (1025 MB)
Knowledge	RTX8000p	4096x2160 (4K)	1	89.67% (1,837 MB)	67.54% (1,383 MB)	45.77% (937 MB)
Knowledge	RTX8000p	4096x2160 (4K)	2	95.67% (1,959 MB)	81.7% (1,673 MB	66% (1352 MB)
Knowledge	RTX8000p	5120x2880 (5K)	1	92.67% (1,898 MB)	71.96% (1,474 MB)	47% (962 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC V100D-1B Profile - 1GB of Frame Buffer						
Knowledge	V100D	1920x1080	1	86.67% (888 MB)	67.59% (692 MB)	51.14% (524 MB)
Knowledge	V100D	1920x1080	2	94.67% (969 MB)	83% (850 MB)	73.04% (747 MB)

Workload	GPU	Resolution	Monitors	Peak FB Usage	Average FB Usage	Idle FB Usage
vPC V100D-2B Profile - 2GB of Frame Buffer						
Knowledge	V100D	2560x1440 (QHD)	2	71.33% (1,461 MB)	59.01% (1,209 MB)	45.33% (928 MB)
Knowledge	V100D	2560x1440 (QHD)	3	87.33% (1,789 MB)	72.62% (1,487 MB)	58.67% (1201 MB)
Knowledge	V100D	4096x2160 (4K)	1	83.33% (1,699 MB)	65.72% (1,346 MB)	41% (839 MB)
Knowledge	V100D	4096x2160 (4K)	2	95.67% (1,959 MB)	81.67% (1,673 MB)	63% (1290 MB)
Knowledge	V100D	5120x2880 (5K)	1	90.33% (1,850 MB)	70.33% (1,440 MB)	41.67% (853 MB)

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 NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, NVIDIA OptiX, NVIDIA RTX, NVIDIA Turing, Quadro, Quadro RTX, and TensorRT 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

© 2020 NVIDIA Corporation. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051

http://www.nvidia.com