



NVIDIA Virtual PC Best Practices

Solution Brief

Document History

TB-10862-001_v01

Version	Date	Authors	Description of Change
01	February 16, 2022	UD, JL	Initial release

Table of Contents

- Executive Summary..... 1**
- Graphics Recommendations 2**
 - Recommended GPU: NVIDIA A16..... 2
 - vGPU Profiles for NVIDIA vPC Deployments 3
 - vGPU Configuration on the NVIDIA A16 GPU 4
 - Recommended ECC Memory Settings 4
 - Support for Multiple and High-Resolution Monitors..... 5
- Server, VM, and Display Protocol Recommendations..... 6**
 - Server Recommendations..... 6
 - VM Configuration Recommendations 6
 - Golden Image Recommendations 7
 - Display Protocol Recommendations..... 8
- Windows 10 Settings for Optimum Graphics Performance 9**
 - Hardware Accelerated GPU Scheduling..... 9
 - GPU Preferences for Software Applications 9
- Monitoring..... 11**
- Conclusion 12**

List of Tables

Table 1	NVIDIA vPC User Profiles	3
Table 2	Windows 10 GPU Preferences.....	10

Executive Summary

Having a reliable virtual desktop infrastructure (VDI) is essential to any user's virtualized experience. For NVIDIA Virtual PC (vPC), this means reliably built virtual machines that provide users a consistent experience with flexible desktops to perform their daily work.

The NVIDIA vGPU solution is the industry's most advanced technology for virtualizing GPU hardware acceleration. NVIDIA vPC improves virtual desktops for every user, with proven performance built on NVIDIA GPUs for exceptional productivity, security, and IT manageability. NVIDIA virtual GPU software divides NVIDIA GPU resources, enabling the GPU to be shared among multiple virtual machines running any application.

This best-practices guide provides considerations for sizing your VDI desktop (such as vCPU, and memory sizes), properly sizing a virtual GPU (vGPU) profile, and enabling hardware-based acceleration to provide a rich user experience. The goal of this guide is to answer frequently asked questions such as:

- ▶ What is the recommended NVIDIA GPU for my NVIDIA vPC VDI?
- ▶ What considerations should I make when evaluating NVIDIA virtual GPU (vGPU) profiles for my users' workloads?
- ▶ What VM configurations should I consider when designing and deploying NVIDIA vPC?

Understanding users' workloads and their related frame buffer requirements can help you determine the overall density requirements of your VDI desktop environment. Digital worker or knowledge worker use cases will vary per user depending on many factors, including:

- ▶ Number and type of applications
- ▶ File sizes
- ▶ Number of monitors and their resolutions

Graphics Recommendations

To provide a rich user experience for VDI desktop users, select a GPU that is purpose built for high-density, graphics-rich VDI and a vGPU profile that is optimized for VDI desktop users' workloads. Also ensure that the GPU's error-correcting code (ECC) memory setting is compatible with NVIDIA vPC and that the amount of frame buffer in your selected vGPU profiles is sufficient for the number and resolutions of the monitors that you plan to use with each vGPU.

Recommended GPU: NVIDIA A16

Density-optimized GPUs are typically recommended for knowledge worker VDI desktops to run office productivity applications, streaming video, and Windows 10. They are designed to maximize the number of VDI users supported on a server.

The NVIDIA A16 GPU is purpose-built for high-density, graphics-rich VDI. Leveraging the NVIDIA Ampere architecture, the NVIDIA A16 provides double the user density of the previous generation of NVIDIA GPUs while ensuring the best possible user experience.



Note: Although NVIDIA vPC is supported on other [data center GPUs](#), it is essential to consider the cost per card and total return on investment (ROI) when selecting a GPU for use with NVIDIA vPC. For additional information, refer to the chapter about NVIDIA GPUs recommended for virtualization in the [Virtual GPU Positioning Guide](#).

The following features of the [NVIDIA A16](#) GPU offer an exceptional VDI user experience:

- ▶ 64 GB of memory (on four GPUs, each with 16GB of memory)
 - A16-1B (1 GB of frame buffer)
 - A16-2B (2 GB of frame buffer)
- ▶ Superior quality video through [support for the latest codecs](#):
 - H.265 encoding and decoding
 - VP9
 - AVI
- ▶ Higher resolution monitors for streaming video & multi-media
 - Four encoders (7th generation) – [NVENC](#)
 - Eight decoders (5th generation) – [NVDEC](#)

To deploy NVIDIA Ampere GPUs, ensure that your environment meets the following requirements:

- ▶ [SR-IOV](#) is supported.
- ▶ The following software is running in your environment:
 - NVIDIA virtual GPU software version 13 or higher
 - VMware vSphere ESXi 7.0 Update 2 (7.0.2) or higher

Support for NVIDIA vGPU software requires the Enterprise Plus Edition of VMware vSphere Hypervisor (ESXi). For details, refer to [VMware vSphere Edition Comparison \(PDF\)](#).

vGPU Profiles for NVIDIA vPC Deployments

vGPU profiles determine the amount of frame buffer that can be allocated to a vGPU. Understanding user workloads and their related frame buffer requirements can help you choose the most suitable vGPU profile, which determines overall density of your environment.

The following table describes the vGPU profiles for NVIDIA vPC and the types of users and applications.

Table 1 NVIDIA vPC User Profiles

	1B Profile	2B Profile
Workload Type	Knowledge Workers	Power Workers
Definition	Runs a variety of applications, web browsing, using email, creating complex documents, presentations, and spreadsheets. Highly mobile, including road warriors.	Users with specific requirements such as multiple, high-resolution monitors to handle larger model files and higher resolution media.
Applications	Video conferencing apps, Windows 10, office productivity apps, design apps, streaming video, multi-media using the latest web standards like WebGL, and industry-specific apps like EMR applications.	Workloads running multiple vGPU accelerated applications and technologies simultaneously, such as Windows 10 and office 365 productivity applications, streaming 4K video and other multimedia; and industry-specific applications that use graphics-intensive browsers that support: WebGL, communications platforms, secure access capabilities, and live real-time streaming data.
Examples	Accountants, sales managers, marketing analysts, lawyers, 2D software developers	Financial analysts, traders, medical specialists
Monitor Resolution	Single HD (1920×1080) Single QHD (2560×1440)	Single 4K (4096×2160) Single 5K (5120×2880)

It is important to note that workloads will vary for each user depending on many factors, including the number of applications, the types of applications, file sizes, and the number of monitors and their

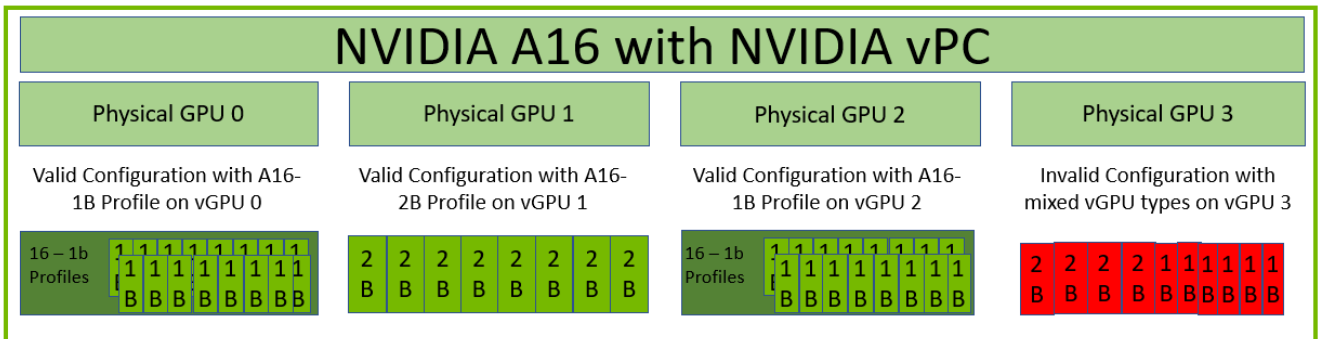
resolution. Frame buffer should be continuously monitored because user behavior can change throughout a project, and individual roles change within the organization. A user that was once a light user of graphics might become a heavy user of graphics after a change of teams or assignment to a different project. Management and monitoring tools enable administrators and IT staff to ensure that their deployment is optimized for each user.

A commonly used monitoring tool that helps organizations track their needs is the [GPU Profiler](#). The profiler quickly captures real-time resource utilization while workloads are being executed within the virtual machine. Refer to the [NVIDIA vPC Sizing Guide](#) for additional information.

vGPU Configuration on the NVIDIA A16 GPU

Optimized for high user density with NVIDIA vPC, the NVIDIA A16 GPU card contains 4 physical GPUs, each having 16 GB of frame buffer. Each physical GPU can host one of two vPC-specific vGPU profiles that carry a predetermined amount of GPU frame buffer. The profile types can vary between different physical GPUs on the same card but must be the same on each physical GPU. The following graphic illustrates this point and highlights the following configurations:

- ▶ A valid configuration with A16-1B vGPUs on GPU 0 and 2
- ▶ A valid configuration with A16-2B vGPUs on GPU 1
- ▶ An **invalid** configuration with mixed vGPU types on GPU 3



Note: The maximum number of vGPUs that can be created simultaneously on a physical GPU is defined by the vGPU type and the GPU frame buffer size. Refer to the NVIDIA vGPU documentation for [maximizing density](#) on GPUs.

Recommended ECC Memory Settings

Many NVIDIA GPUs that support vGPU software support error-correcting code (ECC) memory. ECC memory improves data integrity by detecting and correcting the most common memory data corruption. However, ECC is not supported on B-series vGPU profiles. Although B-series vGPUs can be created on physical GPUs on which ECC is enabled by default, keeping ECC enabled when it is not supported will incur some costs. This incurred cost reduces the amount of usable frame buffer on the

GPU, and it is recommended to disable ECC. For more information on disabling ECC memory refer to the [NVIDIA Virtual GPU Software Documentation](#).

Support for Multiple and High-Resolution Monitors

An increasing need for multiple and high-resolution monitors has impacted vGPU frame buffer requirements, exposing a need for hardware-accelerated encoding and decoding. Each additional monitor requires an average of 15% more frame buffer. Refer to the [NVIDIA vPC Sizing Guide](#) for additional information on the impact of multiple monitors with higher screen resolutions.

Server, VM, and Display Protocol Recommendations

It is essential to evaluate user experience properly as you design your VDI infrastructure and consider the effects of broad-scale deployment. For more information about quantifying user experience, refer to chapter 4 of the [NVIDIA vPC Sizing Guide](#) for deployment best practices.

Server Recommendations

Before deploying NVIDIA vPC, it is essential to confirm that your server is an [NVIDIA vGPU certified OEM server](#). It is recommended to cross-check this list with the [VMware Compatibility Matrix](#) to ensure compatibility for your deployment.

The following server specifications are recommended for hosting NVIDIA vPC:

- ▶ 3.0 GHz CPU
- ▶ High-speed RAM
- ▶ Fast networking
- ▶ High-speed storage

VM Configuration Recommendations

The following VM hardware specifications are recommended for NVIDIA vPC:

- ▶ **vCPU:** 4 vCPU
- ▶ **Memory:** 4 GB to 6 GB
- ▶ **vGPU profile:** 1B or 2B
- ▶ **Network adapter:** on VMware vSphere, VMXNET3

For optimum VDI performance and ease of maintenance, follow these recommendations:

- ▶ Use a flash storage array for the OS HDD with thin provisioning when possible.
- ▶ Use a separate vCenter Server instance for the VDI infrastructure.
- ▶ Use VMware Instant Clone Technology when deploying large instances.
- ▶ Remove all unused virtual hardware devices, such as CD or DVD drives, SATA controllers, COM ports, and floppy disk drives.

- ▶ Use separate VLANs for the VDI environment.

Several best-practice designs for configuring a secure and robust network for VDI environments are available in the [Horizon Architecture Planning](#) guide.

Golden Image Recommendations

Creating a single VM that serves as the golden image or base template VM for your VDI desktops provides consistency in your deployment.

Follow these recommendations when creating your golden image VM:

- ▶ Install the following components:
 - Hypervisor management tools, such as VMware Tools
 - A Windows operating system that is compatible with VMware Horizon

To ensure that your chosen Windows OS is supported by vGPU software, refer to the [Supported Products](#) page in the NVIDIA Virtual GPU Software Documentation.
 - Standard Microsoft runtimes, features, and software applications
 - Microsoft updates
- ▶ Generalize Windows with the [Sysprep](#) or other tools.
- ▶ Configure Windows for optimum graphics performance as explained in "Windows 10 Settings for Optimum Graphics Performance," on page 9.
- ▶ (Optional) Install the NVIDIA vGPU graphics driver for Windows.



Note: NVIDIA vGPU drivers are backwards compatible. Cross-check compatibility in the [NVIDIA Virtual GPU Software Release Documentation](#) for more information.

- ▶ Create a template of the image.

Some enterprises have the flexibility to deploy Windows 10 optimization tools for their VDI environments Golden Image VM. Here are two commonly used optimization tools:

- ▶ VMware VDI environments – [VMware OS Optimization Tool](#)
- ▶ Citrix VDI environments – [Citrix Optimizer](#)



Note: Ensure that you review and test any tool before use in a production environment.

Display Protocol Recommendations

NVIDIA GPUs contain a hardware-based encoder and decoder, which provide acceleration for several popular codecs. Complete encoding (which can be computationally complex) is offloaded from the CPU to the GPU by using NVENC. Hardware-based decoders (referred to as NVDEC) provide faster real-time decoding for video playback applications. When NVIDIA hardware-based encoders and decoders are used, the graphics engine and the CPU are free for other operations, and there is no impact on graphics performance or quality, resulting in performance that is faster than real-time video processing. For more information, refer to [NVIDIA video codec](#) documentation.

The following display protocols leverage NVIDIA's NVENC and NVDEC hardware-based accelerators for VDI deployments:

- ▶ For Horizon VDI deployments, NVIDIA vPC best practice is to use the [VMware Blast 3D Protocol \(H.264/H.265\)](#). The VMware Blast 3D display protocol supports virtual shared pass-through graphics acceleration technology or vGPU software. This technology allows the VM to pass all graphics commands directly to the NVIDIA GPU, bypassing the hypervisor. The VDI desktop takes full advantage of the NVIDIA vGPU features, accelerating application responsiveness, enhancing the user experience, and delivering consistent performance for your NVIDIA vPC VDI desktops.
- ▶ For Citrix deployments, refer to the [Remoting Protocols For GPU-Accelerated, VDI environments](#) whitepaper, which provides an overview of different VDI policy settings and their impact on the user experience. This white paper illustrates how the recommended setting YUV444 is the optimal remoting protocol for image quality and leveraging bandwidth consumption when using an NVIDIA vPC vGPU in your Citrix VDI environment.

Windows 10 Settings for Optimum Graphics Performance

With NVIDIA vPC, Windows 10 delivers a great user experience by leveraging NVIDIA vGPU hardware acceleration. Hardware acceleration (or offloading) allows the GPU to do what it does best, graphics processing. Moving rendering processes from the CPU to the GPU results in a richer user experience, with better graphics performance than emulated software acceleration. When vGPU resources are assigned to a VM, Windows 10 is considered fully accelerated.

Hardware Accelerated GPU Scheduling

Windows has enhanced the GPU scheduling for hardware acceleration within the Windows Display Driver Model (WDDM). The May 2020 update for Windows 10 introduced a new GPU scheduler as a user opt-in, but it is off by default. This new scheduler reduces the overhead of GPU scheduling. For NVIDIA vPC, it is recommended to enable the **Hardware-Accelerated GPU Scheduling** feature, which allows the NVIDIA vGPU to manage its video memory instead of the operating system. Refer to [Hardware Accelerated GPU Scheduling](#) on the Microsoft DirectX Developer Blog for additional information on advanced graphic settings and Windows 10 hardware acceleration configuration.

GPU Preferences for Software Applications

User workloads are the most important factor in determining which vGPU profile will deliver optimal performance. A typical Windows 10 active workload includes the Microsoft Office 365 suite of applications. Most of these applications have hardware acceleration turned on by default (through application settings), but additional OS configuration settings can further optimize these applications to improve the user experience and graphics performance.

In Windows 10, you can customize the graphics performance preferences of any application. You can force graphic-intensive applications to use the dedicated NVIDIA vGPU, improving the applications' performance. To change the graphics processor preference for an application, go to the graphics settings in the System display section and set the **Graphics Performance Preference** to high performance for a quality user experience.

Table 2 Windows 10 GPU Preferences

GPU Preference	Description
System default	Allow Windows to decide the best GPU for your application.
Power saving	Run the application on the most power-saving GPU available.
High performance	Run the application on the highest-performing GPU available.

Monitoring

Typical VDI deployments have two goals: achieving the best possible user experience and maximizing user density on the hypervisor host server. Problems can arise as density is scaled up because it can negatively impact user experience after a certain point.

It is recommended to regularly monitor hypervisor host server and VM metrics to ensure optimal performance and resource utilization of your VDI environment. These measurements can be used to assess the trade-offs between your end-user experience and resource utilization. Capturing performance metrics and logging the metrics for later analysis can show trends and patterns over time, which helps optimize your vPC VDI configuration to meet your end users' performance requirements and helps determine scalability.

NVIDIA vPC provides extensive monitoring features to better understand the usage of the various engines of the NVIDIA GPU. The utilization of the compute engine, the frame buffer, the encoder, and the decoder can all be monitored and logged on the hypervisor host server and the VM. It is important to continually monitor resource usage to ensure system health, stability, and scalability, as your deployment needs may change over time.

When monitoring resource usage by the hypervisor host server and guest VMs, pay close attention to the following metrics:

- ▶ **GPU engine usage on the hypervisor host server.** To ensure adequate server resources and prevent the server from doing more than it has capacity for, GPU engine usage rates should not exceed 90%.

These metrics are essential for understanding the number of vPC VM's that can be hosted on the server. If metrics are high, conduct additional testing by analyzing vGPU and vCPU metrics on individual VM's and reduce the number of VM's per server.



Note: In general, 80% CPU usage is a reasonable ceiling on the Host Server, and 90% usage should be a warning that the CPUs are approaching an overloaded condition.

- ▶ **Frame buffer usage in VMs.** Frame buffer usage should not exceed 90% for a short time nor average over 70% on the 1 GB (1B) profile. If high frame buffer usage is observed, the vPC virtual machine should be assigned a 2 GB (2B) profile.

Refer to the [NVIDIA vPC Sizing Guide](#) for additional information about and tools for monitoring hypervisor host server and VM resource usage.

Conclusion

The NVIDIA vGPU solution is the industry's most advanced technology for virtualizing GPU hardware acceleration. This best-practices guide provides considerations for optimally configuring your NVIDIA vPC deployment and choosing the correct virtual GPU profile.

Density-optimized GPUs are typically recommended for knowledge worker VDI desktops to run office productivity applications, streaming video, and Windows 10. They are designed to maximize the number of VDI users supported in a server. The NVIDIA A16 GPU is density optimized for vPC and can increase user density compared to the previous generation.

Windows 10 features, such as hardware acceleration, allow the GPU to do what it does best, namely, graphics processing. Customizing the graphics performance preferences of applications within the Windows 10 operating system can further enhance the user experience.

Workloads may vary for each user and depend on many factors, including the number of applications, the types of applications, file sizes, number of monitors, and their resolution. It is recommended that hypervisor host server and VM metrics be continuously monitored over a product lifecycle to ensure optimal user experience and performance while meeting organizational needs.

Notice

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

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

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

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

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

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

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

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

Trademarks

NVIDIA and the NVIDIA logo 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.

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.

OpenCL

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

Copyright

© 2022 NVIDIA Corporation. All rights reserved.

