Impact of NVIDIA Virtual GPU on Video Conference Tools

Elevating User Experience with NVIDIA GRID Virtual PC (GRID vPC)

Technical Brief

TB-10169-001_v01 | October 2020

Table of Contents

Intent of this Technical Brief

NVIDIA virtual GPU (NVIDIA GRID vGPU™) technology is enabling organizations to seamlessly shift to remote work. With this shift, video conferencing tools are becoming increasingly prevalent.

In this technical brief, you will find the impact of NVIDIA GRID® virtual PC (GRID vPC) on video conferencing tools. As video and audio are offloaded to the graphics processing unit (GPU), you will be able to see the difference in user experience when executing common video conferencing features like video activation and screenshare.

Executive Summary

As more organizations shift to remote work, video conferencing applications have become critical tool for employees to work seamlessly with remote Teams. To be effective, it is imperative for users to have a great user experience when using video conferencing tools. Video and audio quality are key to productive real-time collaboration. Any delays in video or audio can lead to miscommunication and negatively impact overall productivity.

This technical brief provides an overview of the impact of NVIDIA GRID virtual PC (GRID vPC) on Microsoft Teams, Zoom, and Cisco Webex compared to CPU only virtual desktop infrastructure (VDI). Tests were executed to determine if GRID vPC offers CPU offload when users are accessing video conferencing tools, like the results seen while testing the NVIDIA nVector knowledge worker workflow. As published within the existing *NVIDIA GRID vPC Sizing Guide,* NVIDIA GRID vPC offers improved server density as well as user experience satisfaction when comparing to a CPU-only virtualized environment. When introducing NVIDIA vGPU technology into the environment, virtual systems will no longer rely solely on CPU for graphics processing. Therefore, systems without a GPU have higher overall CPU usage due to software applications ability to execute certain functions on the GPU, offloading CPU.

Overall, our video conferencing test results showed that by having a vGPU present within the virtual machine (VM), there was significant amount of vCPU offload which frees vCPU resources within the VM to run other tasks. The amount of CPU offload is dependent on how well the web conferencing application can take advantage of the GPU as well as network and web camera conditions. The CPU offload within VM's ultimately results in decreased CPU resource usage on Host Server due to the nature of virtualization using shared server resources. Therefore, NVIDIA GRID vPC offers improved server density while boosting user experience.

Introduction

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

- NVIDIA GRID Virtual PC/Virtual Applications (NVIDIA GRID) – accelerates office productivity applications, streaming video, Windows 10, RDSH, multiple and high-resolution monitors and 2D electric design automation (EDA).
- NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS)– accelerates professional design and visualization applications including Autodesk Revit, Maya, Dassault Systèmes CATIA, Solidworks, Esri ArcGIS Pro, Petrel, and more.
- NVIDIA Virtual Compute Server (vCS)– accelerates artificial intelligence (AI), deep learning (DL), data science and high-performance computing (HPC) workloads run in a virtualized environment.

Decoupling the GPU hardware and virtual GPU software options enables customers to benefit from innovative features delivered in the software at a regular cadence, without a dependency on purchasing new GPU hardware. It also provides the flexibility for IT to architect the optimal solution to meet the specific needs of users in their environment.

NVIDIA GRID vPC

This technical brief will focus on NVIDIA GRID vPC. NVIDIA GRID vPC software enables the delivery of graphics-rich virtual desktops accelerated by NVIDIA graphics processing units (GPUs). NVIDIA GRID vPC enables 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.

Testing Methodology

Testing was executed in 2 phases to determine the impact of NVIDIA GRID vPC. These phases represent different workflows which a typical knowledge worker performs when using video conferencing tools. Table 1 describes the 2 phases of testing as well as the VM configuration tested.

Table 1. Test Workflow

Phase	Workflow	vCPU	vGPU Profile	Memory
1	Web Camera	2	T4-1B	4 GB
1	Web Camera	2	N/A (CPU-Only)	4 GB
2	Screenshare	2	T4-1B	4 GB
2	Screenshare	2	N/A (CPU-Only)	4 GB

- Phase 1 - Web Camera
  - Incremental activation of 3 cameras
  - Incremental deactivation of 3 cameras
- Phase 2 - Screenshare
  - Viewing screenshare
  - No web camera activation

There was a total of 3 users executing tests during all phases of testing. The first test user acted as the video conferencing host, whereas the second and third test users were attendees and accessed video conferencing tools within a VM that was configured for GRID vPC or CPU-only VDI.

The following video conferencing tools were tested during each phase of testing:

- MS Teams
- Zoom
- Cisco Webex

Testing was not automated, instead test users followed a very defined script to ensure performance logging was synchronized for CPU only and GRID vPC VMs. Testing focused on single VMs and were not executed to scale (i.e. 64 users) since enterprise users typically not work exclusively with these tools but have multiple software applications active while interacting with video conferencing tools. Testing methodology primarily focused on the value of CPU offload which drives the value of GRID vPC - improved server density as well as user experience satisfaction.

Test Setup

GPU Profiler is commonly used tool which 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 NVIDIA nVector knowledge worker workload was running. The following metrics were captured:

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

Figure 1 shows a screen capture of the GPU Profiler.

Figure 1. GPU Profiler Screen Capture

The following list summarizes the VM configuration for GRID vPC as well as CPU-only VDI.

- vGPU Profile: 1B
- vCPU: 2
- vRAM: 4096 MB
- NIC: 1 (vmxnet3)
- Hard disk: 40 GB
- Virtual hardware: vmx-13
- VMware Horizon: 7.9
- vGPU Driver: NVIDIA GRID 10.2 (Windows Driver 442.06)
- Guest OS: Windows 10 Enterprise 1909
- Server: Intel Xeon Gold 6148 CPU @ 2.4 GHz

Variables Impacting Performance

There are several variables which can impact performance as well as the amount of resource usage within the VM's. These include:

- Network latency
- Web camera model, driver, and quality
- CPU speed
- Monitor setup – quantity and resolution

Additionally, how users configure their software such as layout of video conferencing tools and the amount of user movement/activity can impact performance.

We tried to standardize across the test to limit the impact of these variables. The greatest factor to consider is network latency and web camera quality, if these are of poor quality, the impact which NVIDIA GRID vPC offers to increase software performance will be less. Logitech 1080p web cameras used for testing GRID vPC as well as CPU-only VDI. VM screen resolution was also standardized at 2560 × 1440 for both test cases.

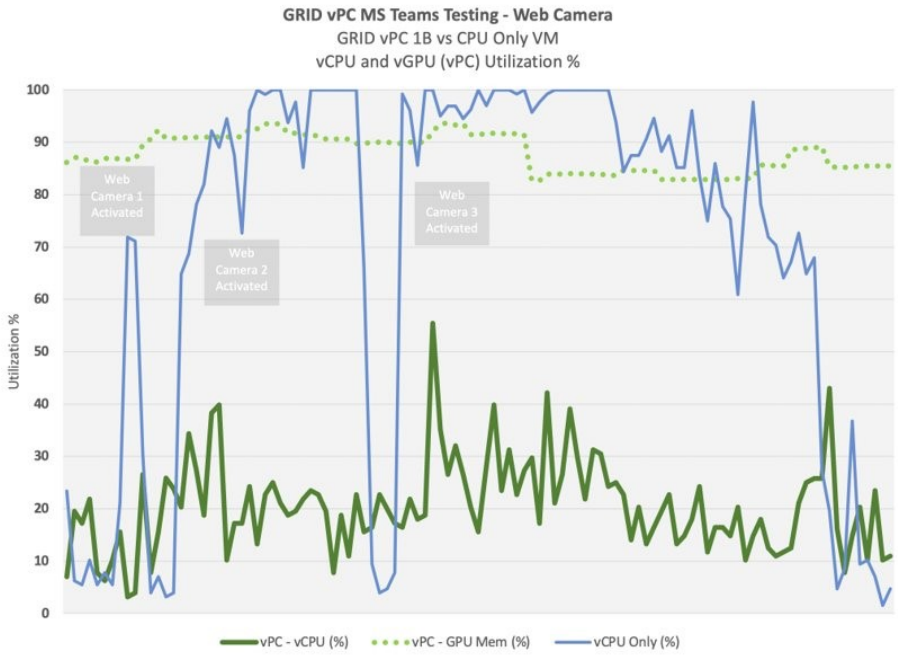
Video conferencing optimization/management packs offered by different hypervisors can also have an impact on the amount of resource offload within the VDI. Video and audio are offloaded to thin client therefore the VM CPU utilization is equivalent with CPU only and GRID vPC VM. But there is added complexity in administering and managing endpoints with optimization packs and they are not suitable for every on-premise or cloud NVIDIA vGPU solution.

Impact of NVIDIA GRID vPC

Overall, Microsoft Teams, Zoom, and Cisco Webex exhibited a certain level of vCPU offload. Since each web conferencing application is uniquely engineered, the amount of CPU offload is dependent on how well the software itself can take advantage of the GPU.

The user experience suffered on the CPU-only VM due to vCPU bottlenecks when accessing Microsoft Teams and Zoom video conferencing tools. The CPU-only VM (2vCPU) reached 100% utilization therefore the video quality was poor, resulting in pixel loss and stutter. Audio quality was also poor with audio delay and breakage. In order to improve the user experience within the CPU-only VM, the vCPU configuration should be increased (i.e. from 2 to 4vCPU) in order to more closely match the quality of user experience offered by a GRID vPC VM. With this in mind, the increased amount of server density value prop really shines with value of GRID vPC since the vCPU count of 2 was adequate for GRID vPC. Figure 2 illustrates the CPU resource usage within the CPU only VM vs GRID vPC as well the utilization of the 1 GB GPU memory frame buffer.

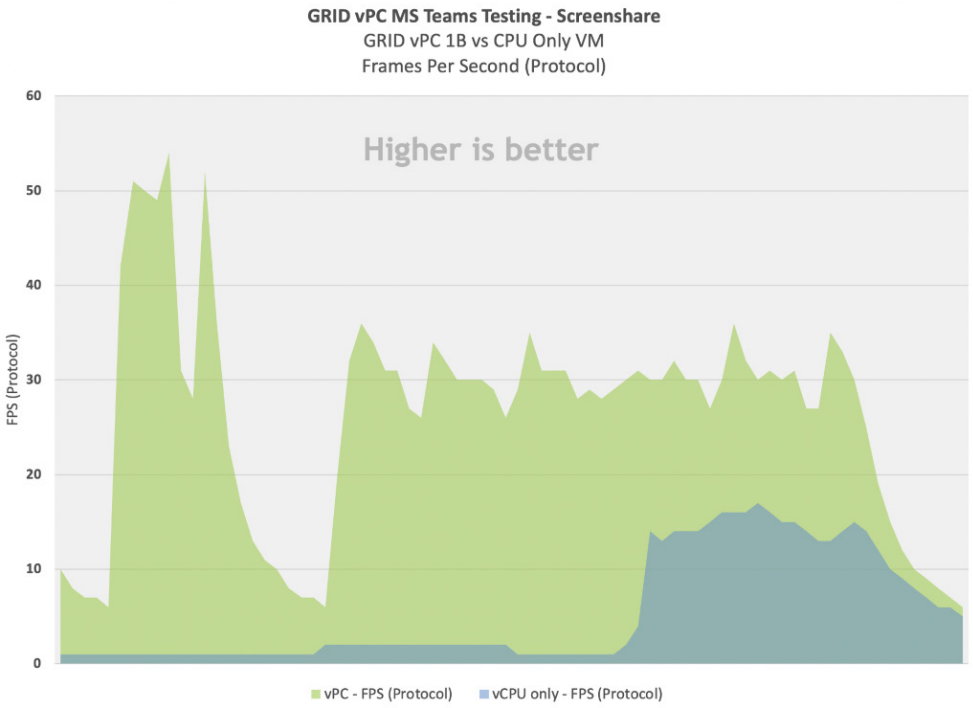
Figure 2. MS Teams Testing – Web Camera



Note: MS Teams results tested on a server with Intel Xeon Gold 61454 CPU @3.0 GHz, NVIDIA GRID vPC software, VMware ESXi 6.7 Update 3, VMware Horizon 7, host/guest driver 442.06, VM config, WIn10 (1909), 2vCPU, 4GB Memory, T4-1B, VMware Blast Protocol. Test workflow incrementally turned on 3 web cameras, ending in a total of 3 web camera feeds (one/attendee), prior to concluding tests web cameras were turned off in MS Teams.

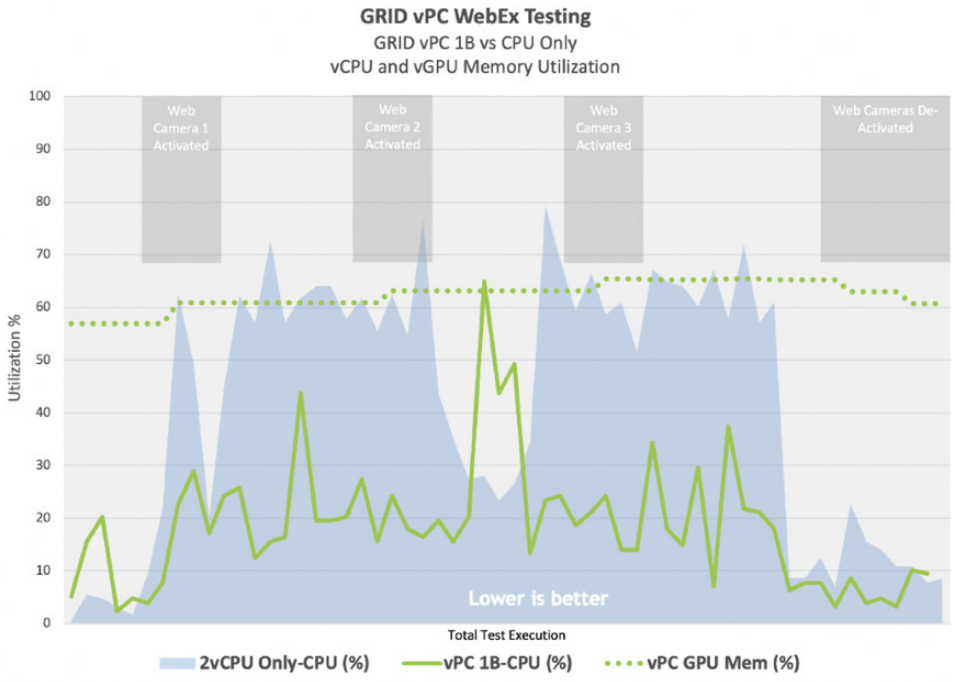
The remoted frames per second (FPS) were also much higher for the GRID vPC VM, providing a better user experience while using less vCPU resources. Figure 3highlights the difference in FPS between GPU-accelerated VDI and CPU-only VDI.

Figure 3. MS Teams Testing – Screenshare



Although WebEx software did not exercise the 2vCPU as much as Teams and Zoom, the amount of CPU offload was still significant. CPU spikes occurred when web cameras were activated but spikes also occurred when there was movement detected in web camera by web conference attendees. Figure 4 illustrates the amount of CPU offload which GRID vPC offered within our web camera workflow testing.

Figure 4. Cisco WebEx Testing – Web Camera



Conclusion

As 2020 progresses, remote work is moving from a short-term response to a long term and potentially permanent solution. For teams to be productive, video conferencing tools, including Microsoft Teams, Zoom, and Cisco Webex, have become necessary for real-time collaboration and communication. It is not enough for organizations to provide video conferencing tools; they also must ensure a suitable user experience with optimal video and audio quality. Video breakage and audio delays can lead to miscommunication, loss of information, lower productivity, and overall frustration. GPU-accelerated VDI with NVIDIA GRID vPC delivers consistent user experience when using video collaboration tools.

With CPU offload in mind, GRID vPC offers increased server density since virtualized server resources are shared across multiple virtual desktops. This offload of CPU also frees up vCPU resources within the VM to run other tasks. With CPU-only VM, more vCPU resources would need to be allocated to match the quality of user experience offered by a GRID vPC VM, increasing the amount of hardware to purchase, deploy, and manage.

Although optimization packs can improve the user experience with video collaboration tools in VDI, there is added complexity in administering and managing endpoints when using optimization packs. By using GPU-accelerated VDI for video conferencing applications, NVIDIA GRID vPC, not only is the user experience improved, but the added benefit of centralized data and applications simplifies IT administration and management, saving time, costs, and IT resources.

Resources Links

NVIDIA GRID Resources:

NVIDIA GRID vPC Sizing Guide

Quantifying the Impact of NVIDIA Virtual GPUs

NVIDIA GRID Solution Overview

NVIDIA GRID webpage

Other Resources:

Try NVIDIA vGPU for free

NVIDIA vGPU 90 day evaluation

NVIDIA Virtual GPU Software Documentation

NVIDIA vGPU Certified Servers

Notice

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

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

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

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

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

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

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

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

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF 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, NVIDIA GRID, NVIDIA GRID vGPU, and Quadro 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

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com