WHITEPAPER

QUANTIFYING THE IMPACT OF VIRTUAL GPUs

See how NVIDIA benchmarks UX in VMware virtualized environments

JUNE 2019

NVIDIA

CONTENTS

Executive Summary 3

Introduction 4

VDI deployments are challenged by the increasing graphics requirements of everyday applications 4
Design a virtual desktop environment that users love 5

Methodology: Factors for Measuring UX 6

Modeling knowledge worker behavior at scale 6
Automating and running the workload at scale 8
Test environment 9

Test Results: Measuring the End User Experience 10

Metric 1: End User Latency 10
What is it?
How GPUs improve end user latency by 3X

Metric 2: Consistency of End User Latency 11
What is it?
How GPUs improve consistency of end user latency by 2X

Metric 3: Remoted Frames 12
What is it?
How GPUs increase remoted frames by 1.5X

Metric 4: Image Quality 13
What is it?
How GPUs ensure uncompromised image quality

Metric 5: Server Utilization 14
What is it?
How GPUs reduce server CPU load by 19%

Meeting Windows 10 Requirements Today and Tomorrow 15

Conclusion 17

EXECUTIVE SUMMARY

Employees will not accept user experience (UX) compromise. In fact, according to Gartner, *"infrastructure and operations leaders focused on mobile, endpoint and wearable computing strategies (should) focus on performance levels and user experience, both of which are critical to successful Desktop as a service (DaaS) adoption, to improve the odds of success with your DaaS initiative."* [1] Existing benchmarking tools provide limited insights on the quality of the user experience since they measure the response time on the virtual desktop and don't take into account the responsiveness at the end user's access device.

To address the need for better insights of the actual end user experience, NVIDIA developed a benchmarking tool based on a methodology that measures key aspects of the user experience, including end user latency, consistency of user experience, remoted frames, image quality, and server utilization.

The results of our testing showed a GPU-enabled VDI environment significantly improves the user experience as well as the server density when compared to a CPU-only VDI environment:

- 3X improved end user latency enabling snappier response times
- 2X improved consistency of user experience removing "peaks and valleys"
- 1.5X more remoted frames for better fluidity
- Uncompromised image quality
- 33% more users supported on the server with GPU, increasing density and lowering costs

The testing also shows that NVIDIA® Tesla® M10 GPU on NVIDIA® GRID more than meets the needs of today's demanding digital workplace. It provides enough GPU headroom, encoder bandwidth, and frame buffer to support Windows 10 requirements as well as today's modern productivity applications.

This white paper details how NVIDIA measures the quality of the end-user experience and quantifies the benefits of NVIDIA GRID on VMware virtualized environments.

[1] Gartner. Market Guide for Desktop as a Service. 28 June 2018.

INTRODUCTION

VDI DEPLOYMENTS ARE CHALLENGED BY THE INCREASING GRAPHICS REQUIREMENTS OF EVERYDAY APPLICATIONS

In recent years everyday business applications like Microsoft Office, Google Chrome, Skype, and PDF readers have evolved to offer graphics-rich features, providing users with better interactivity and improved overall performance. To further enhance their daily experience, users often have multiple monitors and are upgrading to 4K resolution monitors, which have become more mainstream as they have become more affordable. Between 2013-2018 the market for 4K display resolution grew at a CAGR of 23.1%.[2] Due to these trends, users are now consuming more CPU resources and active memory than ever before.

In tandem with these upgrades, the graphics requirements of operating systems have also increased. Windows 10 is a good case in point. It has the highest graphics requirement of any operating system to date, with a 32% increase in CPU requirements compared to Windows 7.[3] Offloading graphics workloads currently performed by CPUs and moving them to virtual GPUs through NVIDIA GRID frees up CPU resources on server hardware and enables a local desktop-like experience to remote users.

[2] Mordor Intelligence. March 2018. "4K Display Resolution Market - Segmented by Product (Monitor, Smart TV, Smart Phone), Vertical (Media & Entertainment, Retail, Consumer Electronics), and Region - Growth, Trends and Forecasts (2018 - 2023)"

[3] Lakeside Software, Inc. "Elevating User Experience Through GPU Acceleration: A Windows 7 Analysis." Lakeside Software White Paper. 2017

DESIGN A VIRTUAL DESKTOP ENVIRONMENT THAT USERS LOVE

NVIDIA's performance engineering teams have developed a methodology and set of benchmarking tools that simulates, at scale, the end user workflow and measures the below metrics. Throughout the paper, we will refer to this as the NVIDIA benchmarking tool.

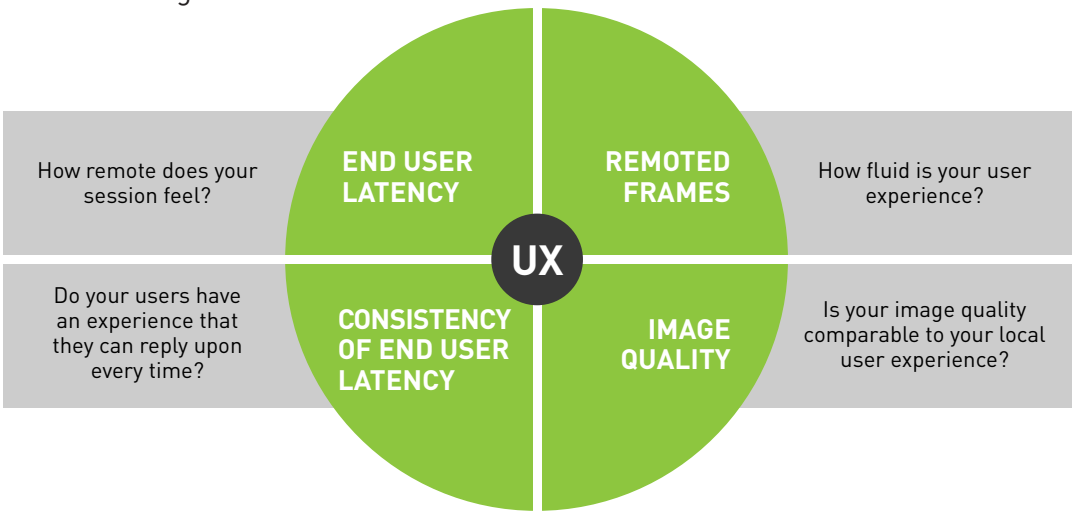


Figure 1. Quantifying User Experience and Scale with NVIDIA Expertise

NVIDIA measures the quality of the user experience across four specific metrics (Figure 1):

- End-User Latency - Measures how remote the session feels or how interactive the session is (amount of lag)
- Consistency - Measures how much the user experience varies during the test run
- Remoted Frames - Measures the number of frames that are sent to the end user
- Image Quality - Measures how much the image was impacted and manipulated by the remote protocol

The paper outlines how the test environment is set up to simulate the day-to-day tasks that knowledge workers perform on standard applications. Then, it explains how end user experience is measured, as well as the trade-offs between user experience and server density. This data will help you design a smarter VDI infrastructure with the best possible user experience while optimizing user density.

METHODOLOGY: FACTORS FOR MEASURING UX

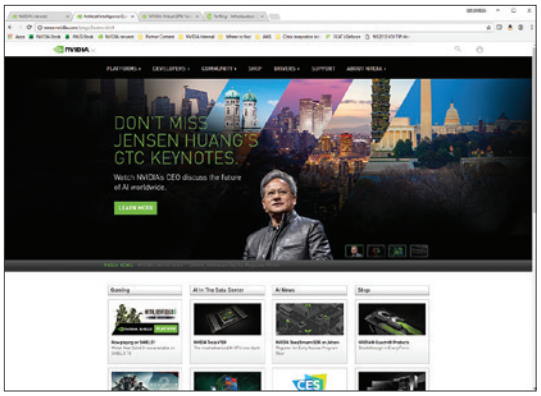
Typical VDI deployments have two conflicting goals: Achieving the best possible user experience and maximizing user density on server hardware. Problems arise as density is scaled up, however, because after a certain point it negatively impacts user experience.

NVIDIA's benchmarking tool measures these trade-offs by modeling how knowledge workers use applications and what happens to performance when workloads are run at scale.

MODELING KNOWLEDGE WORKER BEHAVIOR AT SCALE

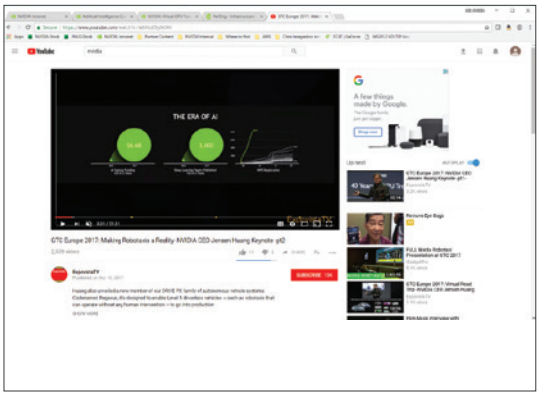
To measure the quality of the knowledge worker experience on VDI, NVIDIA built a workload that simulates user behavior for a set of applications that are a good representation of knowledge workers' most widely used applications (Figure 2). The sample set includes Microsoft Word 2016, Microsoft Excel 2016, Microsoft PowerPoint 2016, Google Chrome for web browsing and video streaming, and Microsoft Edge for browsing PDF documents.

Web browsing with modern web browsers



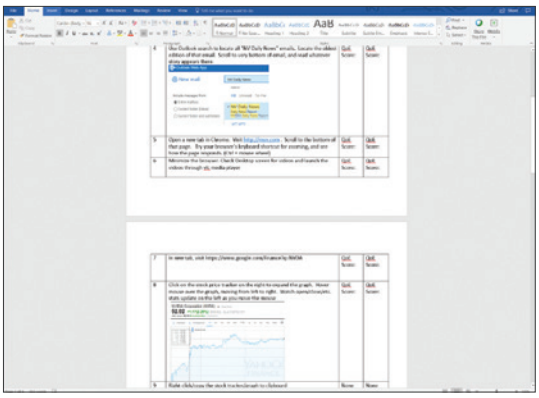
Google Chrome

Viewing videos (web or local)



Google Chrome, Windows Media Player

Viewing/Editing content



Microsoft Word 2016, Excel 2016, PowerPoint 2016, Edge

Figure 2. NVIDIA's Definition of a Knowledge Worker. The images are screenshots taken of the worklfows tested which include web browsing, video viewing, and viewing/editing content.

The simulated workload also takes into account how these applications are used and is built with human speed input and scrolling. For example, all of these applications typically go through several stages when they're being used (Figure 3). The end-user will:

1. Open the application to either create new content or load pre-existing content
2. Modify the content
3. Review the changes by scrolling and/or zooming in and out of the content
4. Save the content
5. Copy content on clipboard and use it in a different application (optional)
6. Close the application

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)	...

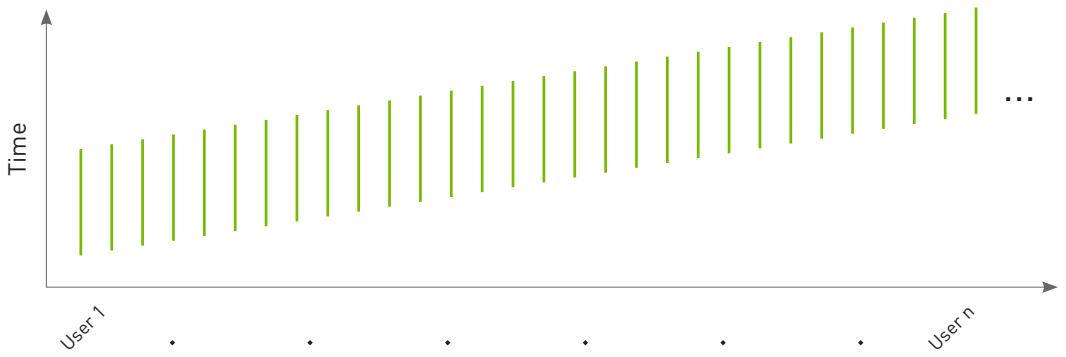


Figure 3. Characteristics of NVIDIA's Benchmarking Tool. The above table shows the workflow of each user. The graph shows cumulative increase in the number of users running workloads through time. Multiple users are tested at a time to simulate scale, with start and end times staggered to be more representative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

AUTOMATING AND RUNNING THE WORKLOAD AT SCALE

The NVIDIA benchmarking tool runs the simulated workflow of a knowledge worker, at scale. This part of the test requires performance monitoring to measure resource utilization. Acting as an execution engine, NVIDIA's benchmarking tool orchestrates the necessary stages that are involved in measuring end user experience for a pre-defined number of VDI instances (Figure 4):

1. Provision a number of VDI instances with predefined settings like vCPU, vRAM, and frame buffer, and provision an equal number of virtual machines that act as virtual thin clients
2. Establish remote connections using the virtual clients to VDI desktops
3. Measure resource utilization stats on the server, as well as the guest OS of the VDI desktop
4. Run a workload that emulates the knowledge worker on all of the VDI instances
5. Collect and analyze performance data and end user experience measurements
6. Generate a report that reflects the trade-off between end user experience and user density (scale)

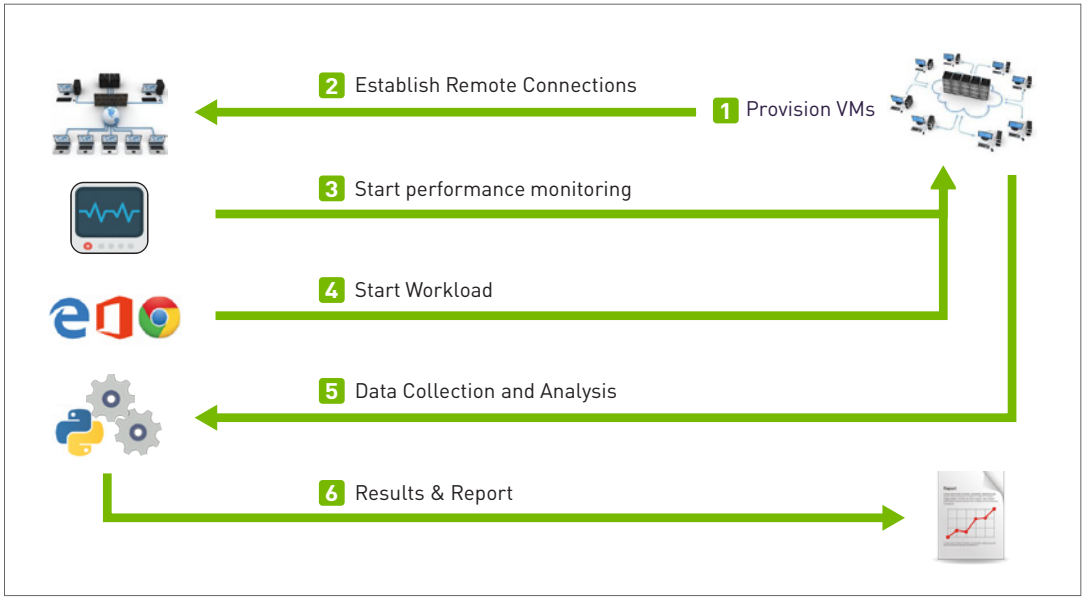


Figure 4. Stages involved in NVIDIA's benchmarking tool in measuring UX

TEST ENVIRONMENT

We quantify the benefits of NVIDIA GRID software (March 2018 release a.k.a. NVIDIA GRID 6.0) running on a Tesla M10. The virtualized environment included VMware Horizon 7.4 and VMware vSphere 6.5. The test modeled 32 VDI users each using a 4K UHD monitor (Table 1). Our results demonstrated significant improvements with the introduction of NVIDIA GRID Virtual PC (GRID vPC) on end user experience, as well as dramatic resource savings.

Table 1: NVIDIA's benchmarking tool runs using the following system configurations:

Host Configuration	VM Configuration	Virtual Client
2 Rack Unit (RU), 2-socket server	vCPU – 4	vCPU - 4
Xeon Gold 6148 CPU @ 2.40GHz	vRAM – 4096 MB/8192	vRAM – 4096 MB
VMware ESXi 6.5 5969303	NIC – 1 (E1000)	NIC – 1 (E1000)
Number of CPUs: 40 (2 x 20)	Hard Disk – 32 GB	Hard Disk – 32 GB
Memory: 512 GB	vGPU – M10-2B	vGPU – M10-2B
Storage: All-Flash SAN (iSCSI)	Virtual Hardware – vmx-11	Virtual Hardware – vmx-11
Hyperthreading, Turbo boost	FRL enabled - No	Single/Dual Screen (4K, 3860 * 2160)
Power Setting: High Performance	VMware Horizon 7.4 (Blast Extreme, 4:2:0)	Guest OS: Windows 7 Ent
GPU: 2 x M10	GRID 6.0 GA	
	Guest OS: Windows 10 Ent Build 1703	
	Number of VMs – 32	

RESULTS OVERVIEW

- 3X improved end user latency enabling snappier response times
- 2X improved consistency of user experience removing "peaks and valleys"
- 1.5X more remoted frames for better fluidity
- Uncompromised image quality
- 33% more users supported on the server with GPU, increasing density and lowering costs

TEST RESULTS: MEASURING THE END USER EXPERIENCE

This paper has demonstrated how NVIDIA's benchmarking tool models knowledge worker behavior, as well as the orchestration stages that are involved in executing this workload at scale. This next section will dig deeper into how the end user experience is measured and the results obtained.

METRIC 1: END USER LATENCY

WHAT IS IT?

One of the most commonly used applications by knowledge workers is Microsoft PowerPoint. Examples of common tasks within the Office Suite that are part of the automated workload include resizing of a shape when building slides or advancing a slide within PowerPoint. The change is then rendered back to the client using the remoting protocol where it is decoded (Figure 5). When a user input—such as a mouse click—is driven from the client virtual machine, the application state changes in the form of transitioning the application to the next slide on the VDI instance. NVIDIA's benchmarking tool measures the time between (T1) when the user provides the input such as a mouse click, and (T2) when the user's client registers the altered state of the application. This method of measuring latency takes into account delays caused by the application processing time, the remoting protocols used, and network latency. This concept is also known as Click-to-Photon. The results from Click-to-Photon will vary depending on the application and how many pixels change between frames. In this testing, majority of pixels between frames change, which is representative of typical knowledge worker actions such as transitioning between slides in PowerPoint, scrolling on a multimedia rich website, or on a PDF with images. Each VDI session collects multiple samples of end user latency, which are aggregated to get a full picture of latency for all users.

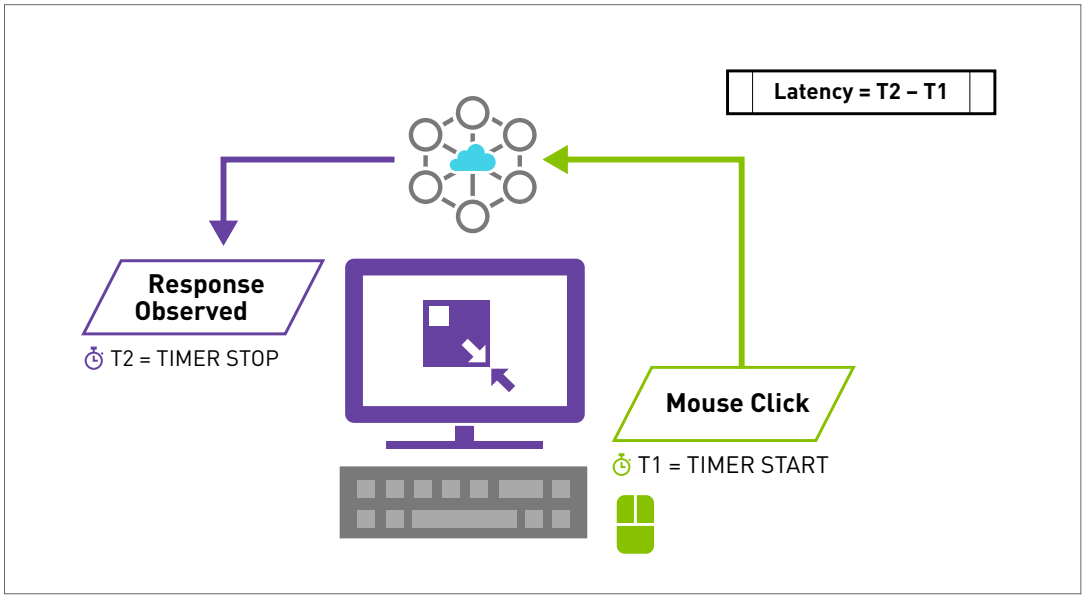


Figure 5. End User Latency (Click-to-Photon)

HOW GPUs IMPROVE END USER LATENCY BY 3X

During 32 VDI sessions, a total of 1,600 end user latency samples were collected. The 90th percentile of the entire sample was used as a good measure of end user latency. Results demonstrated that GPU-accelerated VDI sessions saw 3X better performance in latency (Figure 6). CPU-only VDI users experienced latency at 714 ms, while GPU-accelerated VDI users experienced latency at 233 ms. To put this into perspective, if you could travel at the speed of light, you could go around the earth's equator three times with 400ms of savings from each click.

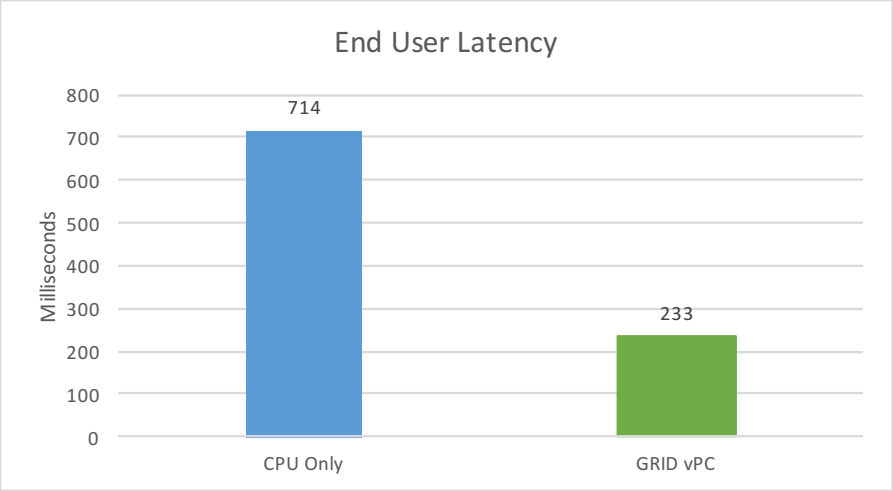


Figure 6: End-user latency of a CPU-only vs. GPU-accelerated VDI environment (lower is better)

METRIC 2: CONSISTENCY OF END USER LATENCY

WHAT IS IT?

Imagine that you are working on PowerPoint and 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 right 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 benchmarking tool measures the variation in end user latency and how frequently it is experienced.

HOW GPUs IMPROVE CONSISTENCY OF END USER LATENCY BY 2X

The standard deviation of the entire sample set of end user latencies is a good measure of "consistency of end user latency." Our data demonstrates that consistency is 2X better on a pool of VDI VMs that are accelerated by NVIDIA GRID vPC versus the pool that is on CPU only. (Figure 7)

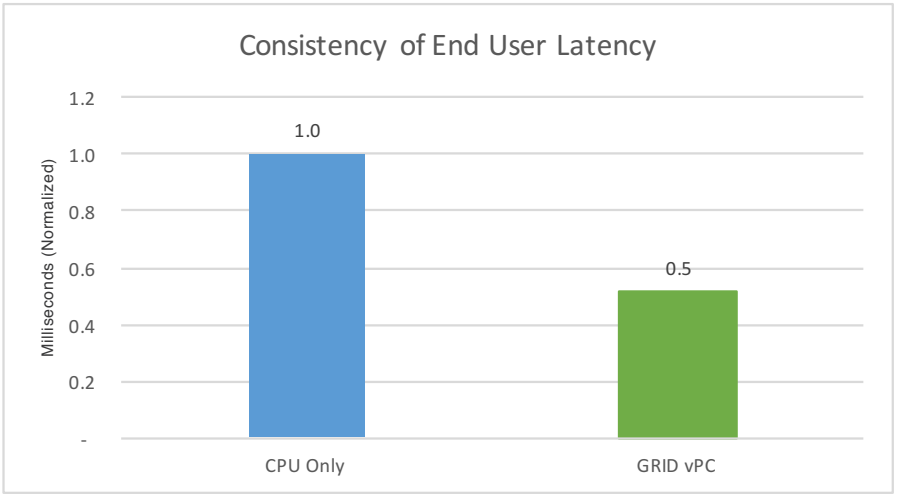


Figure 7. Consistency of end user latency of a CPU only vs GPU-accelerated VDI environment (lower is better)

METRIC 3: REMOTED FRAMES

WHAT IS IT?

Remoted frames is a common measure of user experience. For the entire duration of the workload, NVIDIA's benchmarking tool collects data on the 'frames per second' provided by the remote protocol vendor. 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 in indicator of quality of user experience. The greater this number, the more fluid the user experience.

HOW GPUS INCREASE REMOTED FRAMES BY 1.5X

GPU-accelerated VDI sessions experienced 1.5X more remoted frames (Figure 8), resulting in a more fluid and native-like experience. Fewer number of rendered frames delivers sluggish response times and sub-optimal interactivity.

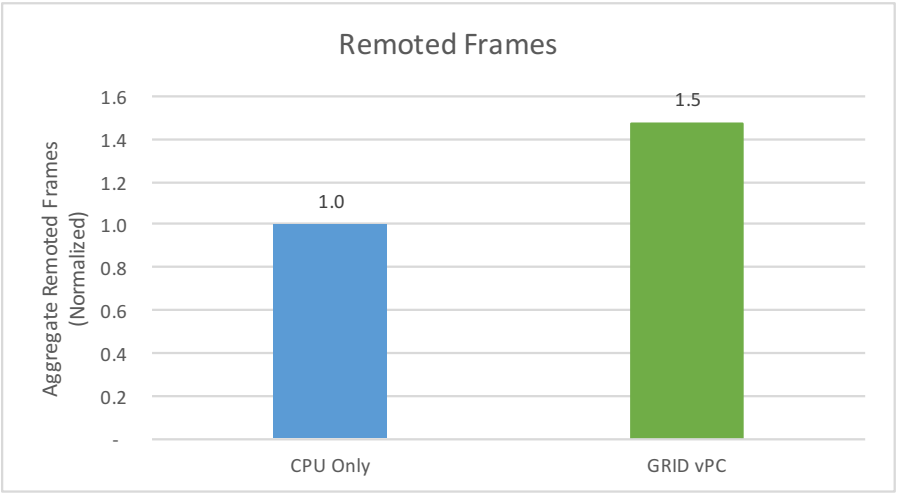


Figure 8. Remoted frames of a CPU-only vs. GPU-accelerated VDI environment (higher is better)

METRIC 4: IMAGE QUALITY

WHAT IS IT?

NVIDIA's 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 on. 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 is obtained. The average SSIM index of all pairs of images is computed to provide the overall remote session quality for the entire population of all users.

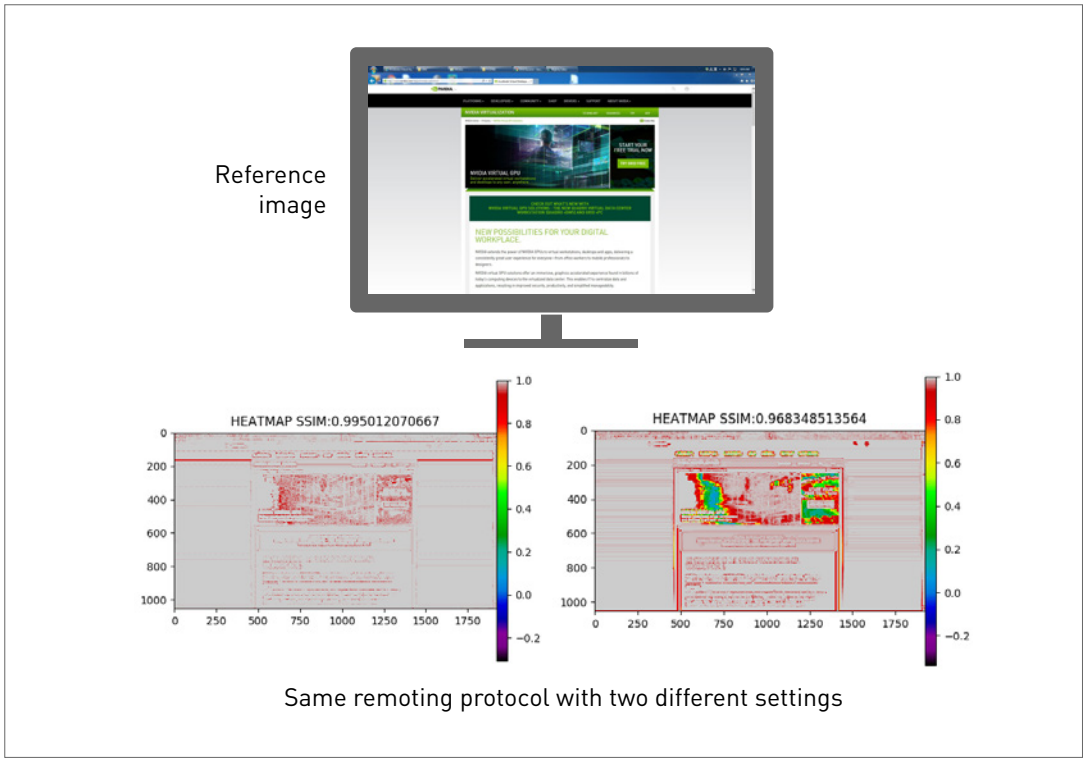


Figure 9. SSIM as a Measure of Image Quality

HOW GPUs ENSURE UNCOMPROMISED IMAGE QUALITY

Image quality is dictated by the remoting protocol and the configuration and the policies set in the VDI environment. Our test demonstrates that GPU-accelerated VDI users will experience uncompromised image quality, as structural similarity (SSIM) of the screen capture are both above 0.99 for both CPU and GPU-accelerated VDI environments (Figure 10).

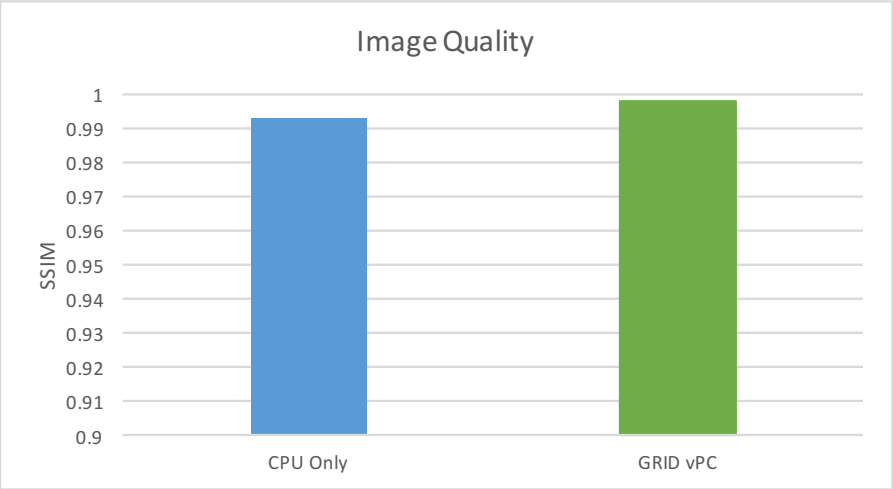


Figure 10. Image quality of a CPU-only vs. GPU-accelerated VDI environment (higher is better)

METRIC 5: SERVER UTILIZATION

WHAT IS IT?

Observing overall server utilization will allow you to assess the trade-offs between end user experience and resource utilization. In order to do this, the tool periodically samples CPU core utilization during a single workload session. To determine the 'steady state' portion of the workload, samples are filtered, leaving out the times when users have all logged on and the workload start ramps up and down. Once steady state has been established, all samples are aggregated to get the total CPU core utilization on the server.

HOW GPUs REDUCE SERVER CPU LOAD BY 19%

Steady state was determined when all VDI sessions were active, and all 32 users were interacting with applications and data. When comparing CPU-only and GPU-accelerated VDI sessions during steady state, a consistent gap in server CPU utilization was observed, which averaged 19% and could go up to 39% (Figure 11). Larger gaps were even observed during some periods, which could be attributed to variations in the workload (video, multimedia, PowerPoint, etc.).

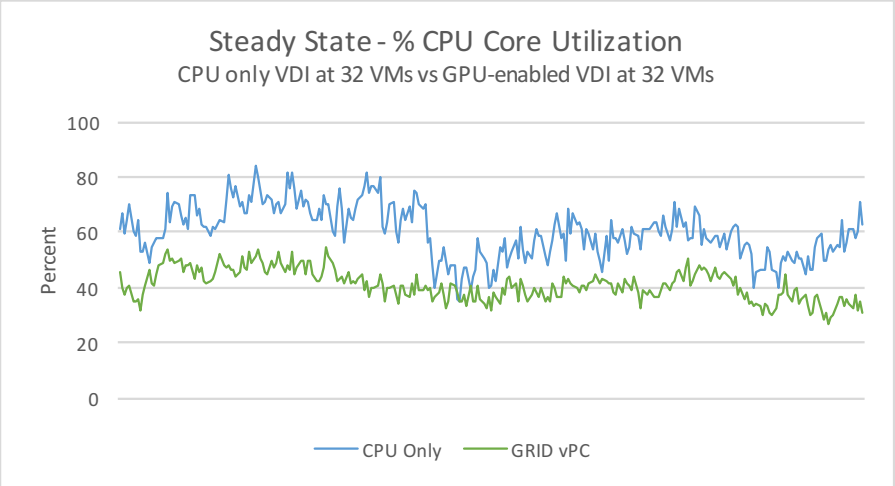


Figure 11. CPU utilization a CPU-only vs. GPU-accelerated VDI environment (lower is better)

When the number of VMs on the CPU-only VDI environment were lowered down to 24 VMs, the CPU utilization dropped to the level of the GPU-enabled VDI environment with 32 VMs (Figure 12). This means you can support up to 33% more users on GPU-enabled VDI environments with a better user experience.

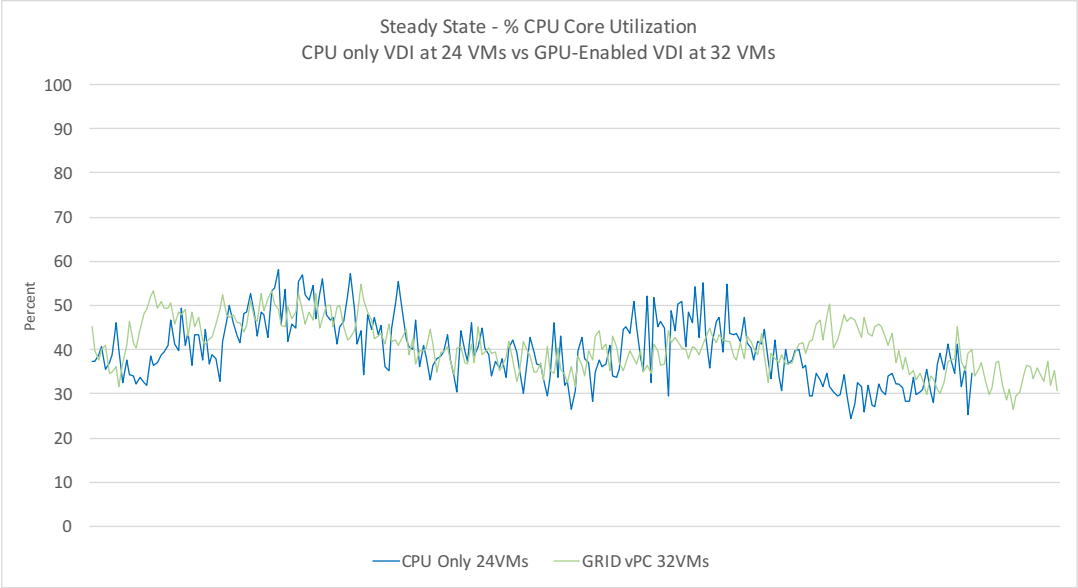


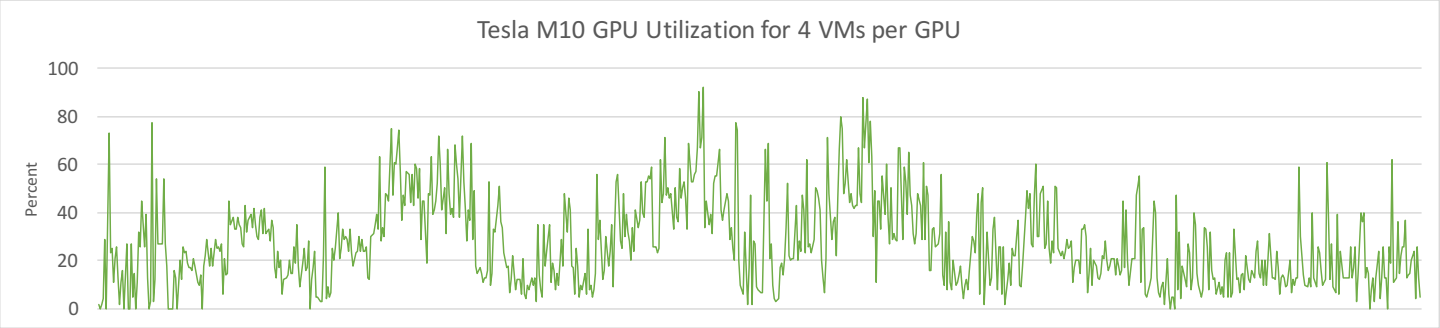
Figure 12. Closing the CPU utilization gap between a CPU only vs GPU-accelerated VDI environment (lower is better)

MEETING WINDOWS 10 REQUIREMENTS TODAY AND TOMORROW

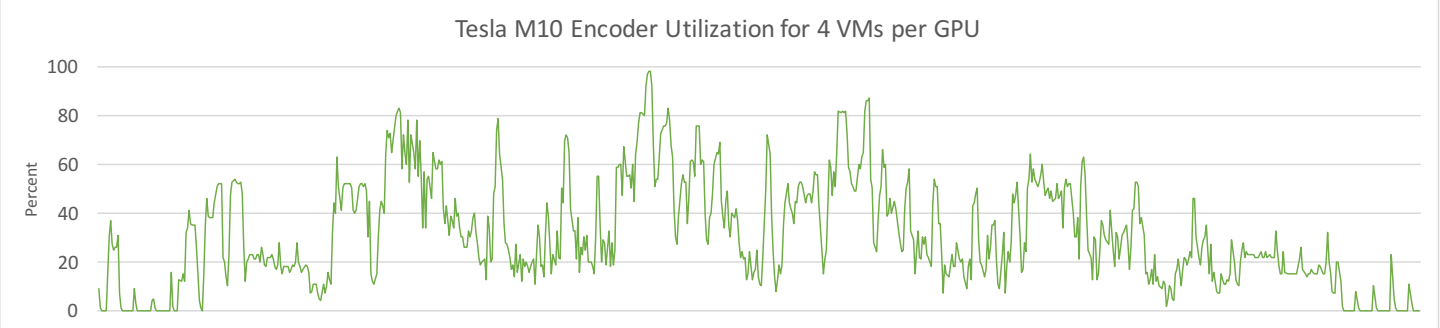
As workloads were running, utilization for GPUs, the encoder, and the frame buffer were analyzed. When examining Tesla M10 GPU utilization for four VMs per GPU, the following observations were made:

- There was plenty of headroom to run today's modern applications.
- Encoder bandwidth was well below 50% most of the time, proving there was enough headroom to meet knowledge workers' application workloads.
- With the Windows operating system and applications both requiring 1GB of video memory, when you are operating this combination in 4K resolution, the 2GB of frame buffer more than meets the requirement, ensuring you can size your users appropriately.

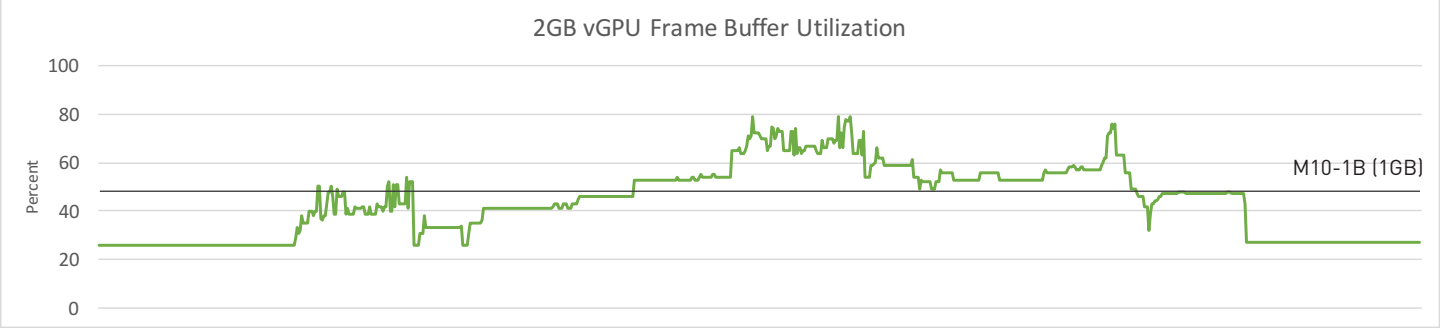
Plenty of GPU Headroom for Today's Modern Apps



Encoder Bandwidth for the Most Demanding Digital Workplace



Today's Operating Systems need 2GB Video Memory



CONCLUSION

Our test results demonstrate that virtualized environments enhanced with NVIDIA GRID for GPU virtualization deliver significant performance improvements. Ultimately, they provide an experience so streamlined and consistent that it allows users to be more productive than they would be on a traditional CPU-only VDI environment. These test results are particularly relevant to today's digital workplaces. With the prevalence of graphics-intensive applications and deployment of Windows 10 across the enterprise, adding graphics acceleration to VDI-powered by NVIDIA GPU virtualization technology is critical to preserving the user experience. Moreover, adding NVIDIA GRID to VDI deployments increases user density on each server which means more users can be supported with better experience.

To learn more about measuring user experience in your own environments, contact your NVIDIA Account Executive.


NVIDIA