WHITEPAPER

NVIDIA VIRTUAL GPU MANAGEMENT & MONITORING

APRIL 2019

NVIDIA

OVERVIEW

Today's IT departments are facing constantly growing complexity of their virtualized infrastructures and environments. At the same time, they need to find better, faster ways to meet the needs of increasingly demanding users. In North America, the average volume in a virtualized desktop environment is 0.60 tickets per user per month.[1] For a company with 5,000 virtualized employees, for example, that translates to 3,000 tickets per month. With the average cost per ticket for level 1 support at $22 and level 2 support at $62, each ticket filed or even escalated to level 2 that could have been resolved in level 1 is wasted expense.[2]

Oversizing and undersizing are also large cost drivers. Incomplete visibility into the virtualized infrastructure and the inability to properly tune environments can result in misallocation of resources. Oversizing leads to lower server density and waste, while undersizing leads to a poor user experience.

Through better insights into their infrastructure, IT can localize a problem before it starts. This means they can now reduce the number of tickets and escalations, and reduce the time it takes to resolve issues. Furthermore, IT can better understand the requirements of their users and better right-size the allocation of resources—saving operational costs while enabling a better user experience.

But delivering a high-quality user experience is only half the story. Delivering high availability is equally important, especially in today's 24/7 world where hourly costs of downtime for enterprises exceed $300,000 on average. In fact, some industries—such as financial services—stand to lose $1 million or more per year, according to Information Technology Intelligence Consulting (ITIC).[3] Through features like live migration of GPU-accelerated VMs, IT can perform critical services like workload leveling, infrastructure resilience, and server software upgrades without any VM downtime.

---

[1] Jeff Rumburg. The True Cost of Desktop Support: Understanding the Critical Cost Drivers of Desktop Support.

[2] Kinetic Vision. (2013, July 30). The ROI of IT Support Improvements (Think Big).

[3] Information Technology Intelligence Consulting. (2017, May 18). Hourly Downtime Tops $300K for 81% of Firms; 33% of Enterprises Say Downtime Costs >$1M.

CHALLENGES IN THE VIRTUAL WORKSPACE LIFECYLE

Addressing the challenges in todays virtualized environments requires visibility into all phases of the virtual workspace lifecycle—from design and assessment to operations and support. Each phase has its own unique challenges:

- Design and Assessment: IT architects need to be able to right-size their VDI environment for the best end-user experience and ROI.
- Operations: IT administrators need to efficiently deploy VDI with the best performance and minimal downtime, and proactively manage and monitor their environment.
- Support: IT help desk and admins need to provide timely VDI user support and issue resolution.

THE FUTURE OF IT MANAGEMENT: END-TO-END VISIBILITY

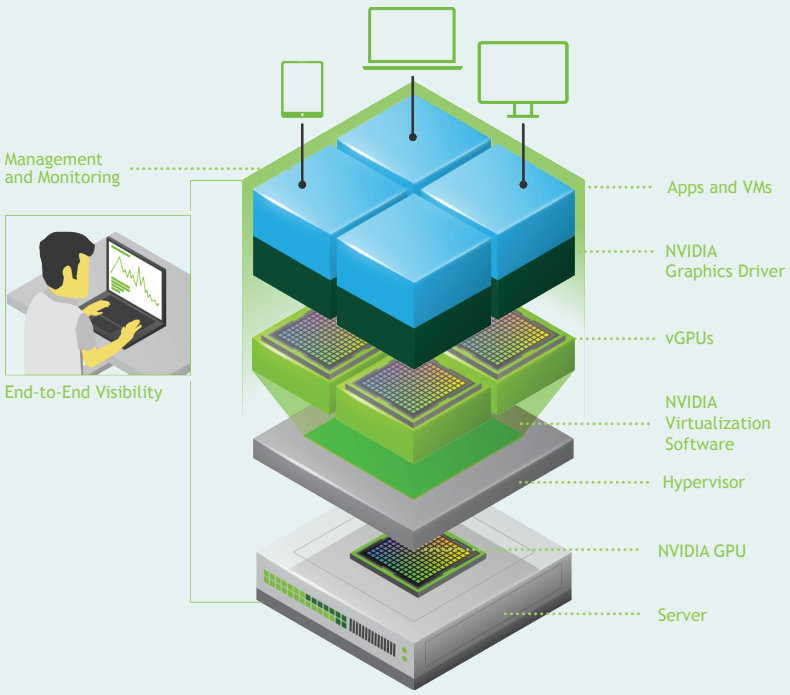
To understand the needs of users, optimize use of resources, and monitor and support with responsiveness and agility, IT needs a comprehensive GPU strategy for their virtual environment.

WHAT IS NVIDIA VIRTUAL GPU?

NVIDIA virtual GPU (vGPU) software enables every virtual machine to get the benefits of a GPU, just like a physical desktop. Because work that was typically performed by the CPU has been offloaded to the GPU, the user has a much better experience and more users can be supported.

NVIDIA vGPU solutions rely on two separate but closely related components to work: (1) the NVIDIA® GPU hardware that goes into the data center servers, and (2) the vGPU software that lives both in the host and the guest.

NVIDIA vGPU is the only GPU virtualization solution that provides end-to-end management and monitoring to deliver real-time insight into GPU performance. It also enables broad partner integrations, so you can use the tools you know and love.



CUSTOMER EXAMPLE

EASY MANAGEMENT OF HIGH-PERFORMANCE VIRTUALIZED DESKTOPS IMPROVE EFFICIENCY AND OPERATIONAL COSTS AT DIGITALGLOBE.

DigitalGlobe is the world's leading provider of high-resolution Earth imagery, data, and analysis, with the most sophisticated commercial satellite constellations in orbit. The company also develops innovative applications and systems that leverage its images. In recent years, an aging virtual environment increasingly impacted the productivity of its developers and office staff who struggled with slow performance and limited mobility. DigitalGlobe leveraged NVIDIA GRID® to lower CPU utilization, improve performance, and enable the staff to collaborate more effectively.

Managing both Linux and Windows VDI environments for many users across the globe was challenging. But with NVIDIA virtual GPU metrics integrated into VMware vRealize Operations (vROps), it took only 30 seconds to deploy and access a dashboard with all the granularity and insights into their virtualized infrastructure.

"We're focused on VDI for all the good things it provides—portability, security, and manageability. With NVIDIA virtual GPU metrics integrated into third party monitoring tools, we get better insight into our VDI environment. We can even see consumption down to the application level: who is using more or less resources, as well as right-size allocation for a better user experience. This ultimately eliminates waste. Thanks to NVIDIA and Nutanix, we have a team of 3.5 people who easily manage 1,500 users daily between the Linux and Windows VDI environments. That's a 500-to-1 management philosophy. You don't get that anywhere. It's a huge cost savings in administrative time."

– Mike Bantz, Engineer and Technical Lead for VDI Environments, DigitalGlobe

NVIDIA VIRTUAL GPU MANAGEMENT SOLUTION

NVIDIA vGPU management features provide end-to-end visibility—from host characteristics to individual vGPU-enabled VMs to a holistic view of all the VMs on a host. It even provides visibility down to individual applications using different GPU components.

The management layer in the guest and host work together to provide visibility into how the GPU is behaving for physical characteristics (such as temperature), as well as how each vGPU enabled VM is consuming the resources. With application-level monitoring, it can also show how a VM is using the GPU resources and inspect each application contributing to it. This brings unprecedented visibility, from accurate sizing to effectively monitoring end user performance and troubleshooting performance issues. The information comes to life either through built-in windows tools, command line, or intuitive dashboards built by our partners.

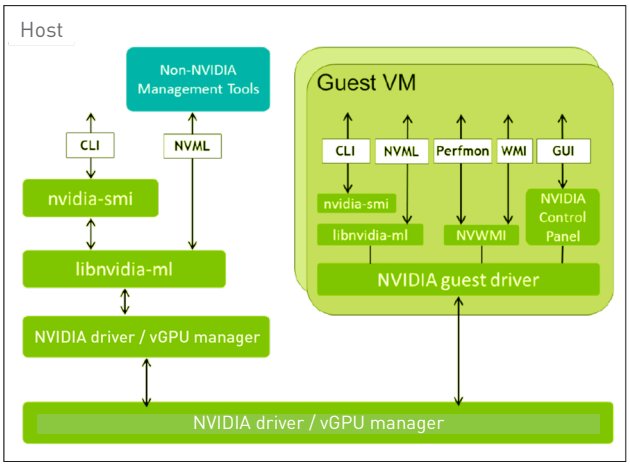


Figure 1: NVIDIA Virtual GPU Interface Architecture

The NVIDIA virtual GPU management solution provides many interfaces to get GPU utilization information. (See Figure 1: NVIDIA Virtual GPU Interface Architecture).

At the guest VM level, NVIDIA software comes with a built-in command line interface called NVIDIA System Management Interface (nvidia-smi). It's a very powerful tool for looking into static and dynamic GPU information.

For anyone writing GUI-based tools, NVIDIA Management Library (NVML) and Windows Management Instrumentation (WMI) interfaces provide great visibility into the GPU. In fact, Windows performance monitor uses WMI counters and shows the utilization metrics by default.

At the host level, data can be accessed through NVML and nvidia-smi. NVML exposes the NVIDIA virtual GPU management SDK metrics to third-party vendors, while nvidia-smi is a management interface/command line tool that allows software programs or management and monitoring tools to interact with GPUs on the server.

The data collected through these interfaces can be stored in a database for future analysis. For example, the nvidia-smi output can be stored into a .csv file and post-processed to create visually informative graphs.

THE METRICS IN DETAIL

NVIDIA virtual GPU management solution provide metrics at the host and guest layers. At the host layer, IT is able to capture physical GPU characteristics, as well as run-time vGPU discovery functions. When accessed inside a guest, these counters represent the particular VM's share of the common resources such as compute/graphics engine, encode, or decode engine. When accessed from the host, it gives a side-by-side view of all the guests running on a host. Insights on each application using GPU engines (3D, Encode, Decode) are also available.

Host Metrics:

- **Physical Characteristics.** Retrieve clocks, temperature, power usage, etc. Change some system parameters like GPU mode, ECC mode, and more.
- **vGPU discovery - Supported vGPU.** Retrieve the **supported** vGPU types on a physical GPU at runtime.
- **vGPU discovery - Creatable vGPU.** Retrieve the currently **creatable** vGPU types on a physical GPU at runtime.
- **vGPU discovery - active vGPU.** Retrieve the currently **active** vGPU types on a physical GPU at runtime.
- **vGPU properties**. Retrieve the properties of a vGPU profile, such as name, number of displays supported, maximum resolution supported, frame buffer (FB) size, current license status, etc.
- **vGPU utilization - 3D, FB, Encode, Decode.** Retrieve average 3D, Encode, Decode, FB utilizations since last monitoring cycle for active vGPU(s).
- **vGPU application utilization - 3D, Encode, Decode.** Retrieve average 3D, Encode, Decode utilizations for applications within a vGPU instance since last monitoring cycle for active vGPU(s).
- **pGPU utilization.** Utilization of the individual counters for the physical GPU.

Guest Metrics:

- **Graphics engine utilization.** A VM's share of the shared graphics engine.
- **Encoder utilization.** A VM's share of the shared encode engine.
- **Decoder utilization.** A VM's share of the shared decode engine.
- **Frame buffer utilization.** A VM's share of the shared memory bandwidth.
- **Frame buffer usage.** Amount of dedicated frame buffer in use.
- **vGPU application utilization - 3D, Encode, Decode.** Retrieve average 3D, Encode, Decode utilizations for applications within the vGPU instance since last monitoring cycle for active vGPU(s).

MIGRATION OF GPU-ACCELERATED VMS

The metrics and insights from NVIDIA virtual GPU management solution are only one part of delivering a high-quality user experience. Delivering high availability is the second.

Migration is the process of moving a running VM from one physical host system to another with minimal disruption or downtime. Some benefits include:

- More efficient server maintenance (hardware replace, software update, etc.)
- Movement of VMs to another host, without interruptions to users, to enable a host and GPU to be used for compute after hours
- Load balance of users without taking users offline (manually now with VMware Distributed Resources Scheduler [DRS] and workload balancing [WLB] in the future)
- Manual consolidation of common profile types for better density
- Maximized data center utilization by running multiple workloads on VDI infrastructure (VDI during the day, AI/deep learning at night)

Live migration has been around for years, but live migration for GPUs hasn't been possible until recently. NVIDIA virtual GPU is now supported with Citrix XenMotion and VMware vMotion delivering the industry's first, and to date the only, support for live migration of GPU-accelerated VMs. Now, IT can achieve improved agility and live migrate users in seconds, without end-user disruption or data loss.

Migrating a VM that includes GPU-acceleration technology is a very difficult task to accomplish. Whereas a CPU only contains a few cores, the GPU contains thousands of cores. Live migration must replicate the GPU on one server to another server and map its processes one to one, as well as copy the state of all active components in use. GPU live migration has been one of the most requested features since NVIDIA virtual GPU first came to market, and it is now a reality.

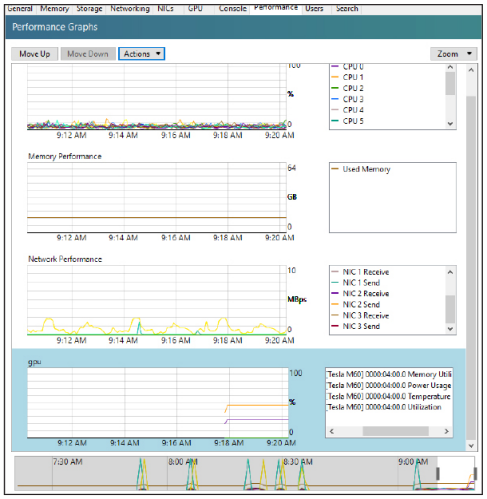
NVIDIA has also worked with VMware to deliver suspend and resume technology for VMware-enabled virtual environments. This feature lets IT suspend running NVIDIA vGPU–powered vSphere virtual desktops and resume them later on a compatible infrastructure with minimal end-user interruption and no data loss. At the same time, they can preserve desktop and application states.

NVIDIA VIRTUAL GPU MONITORING IN ACTION

NVIDIA Virtual GPU in the Citrix Ecosystem

Without any additional work or configurations, a system admin can add NVIDIA virtual GPU as a component into their Citrix XenCenter console. This lets them monitor not only the physical characteristics of their hosts, but also host-level GPU utilizations. Individual VM and vGPU utilization data is available with Citrix Director, providing live GPU monitoring, and is targeted towards help desk use cases.

Live migration of GPU accelerated VMs with Citrix XenMotion is also supported.



Host Monitoring in Citrix XenServer

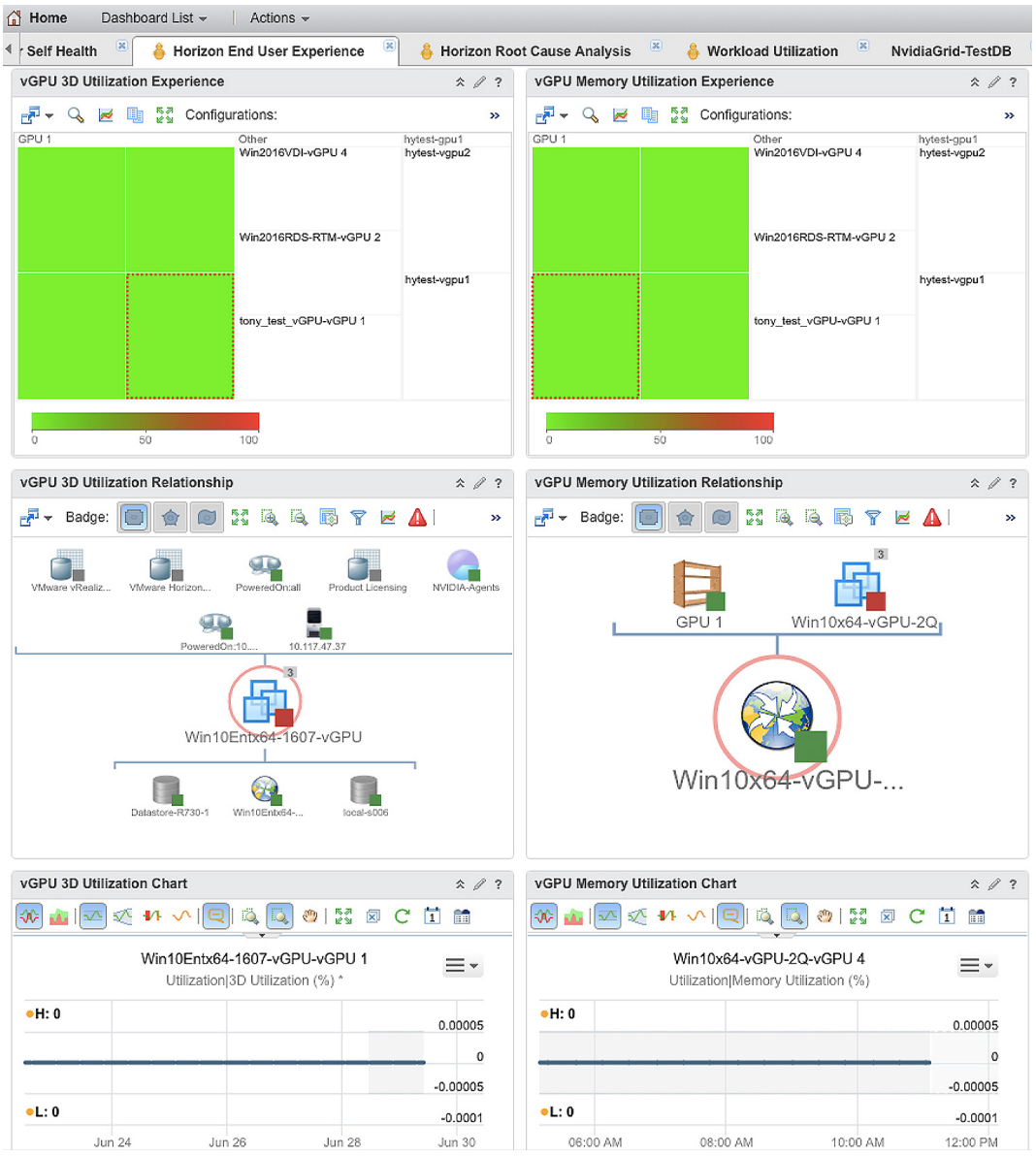
Guest Monitoring in Citrix Director

NVIDIA Virtual GPU in the VMware Ecosystem

NVIDIA vGPU and GPU integration with VMware comes in different forms. First, GPU information is reflected in the VMware Horizon admin-focused dashboards in V4H (vRealize Operations for Horizon). Second, NVIDIA is building GPU-focused dashboards that will address both system admin as well as help desk use cases.

Migration with VMware vMotion is also supported on NVIDIA GPU-enabled virtual environments.

NVIDIA GPU insights integrated into vROps for Horizon:



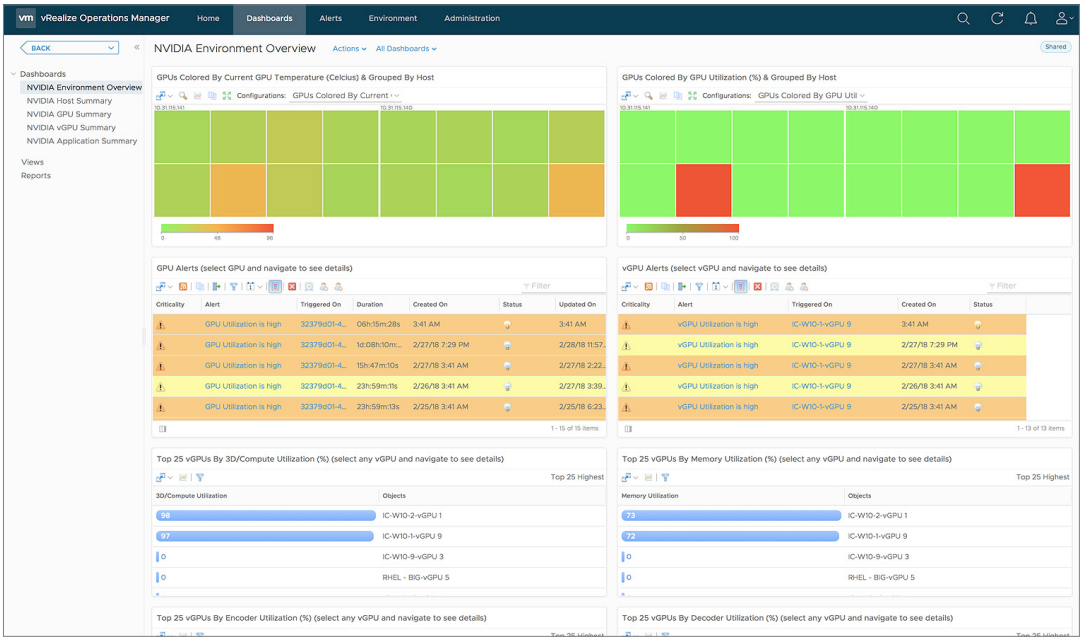
Heat Map, Relationship and Topology view, and User-Desktop Centric View

NVIDIA VIRTUAL GPU MANAGEMENT SDK

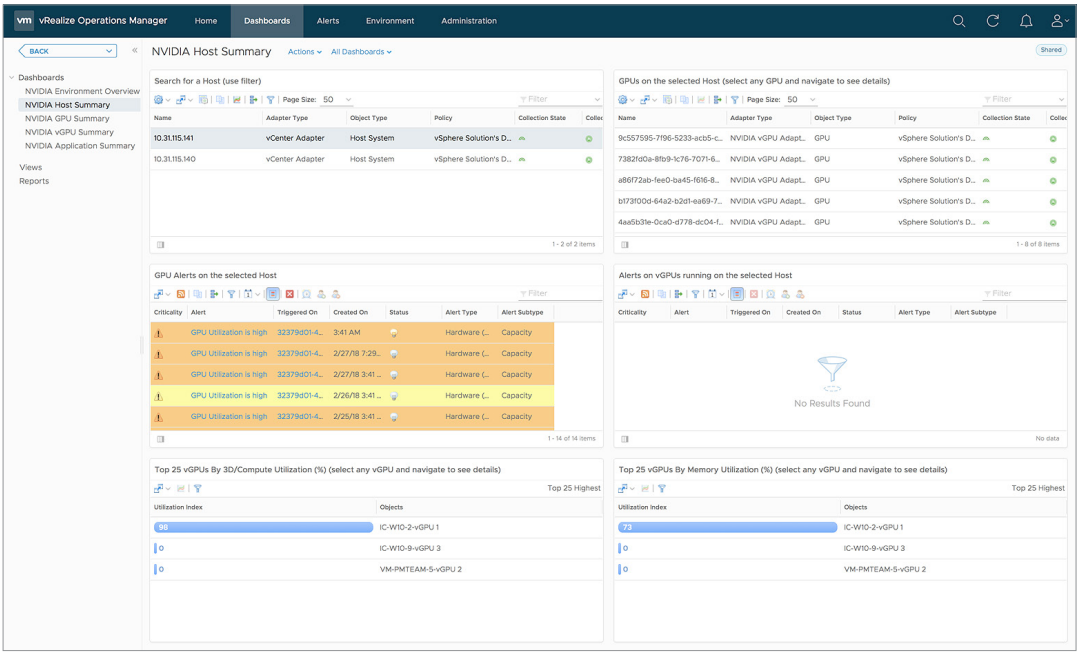
VMware Horizon or Citrix XenDesktop/XenApp customers who do not wish to use vRealize Operations for Horizon (V4H) can leverage the NVIDIA virtual GPU Management SDK free of charge.

The NVIDIA vGPU management SDK provides dashboards with end-to-end GPU insights in vROps for the most comprehensive view of the virtualized infrastructure, supporting architects, admin, and help desk use cases:

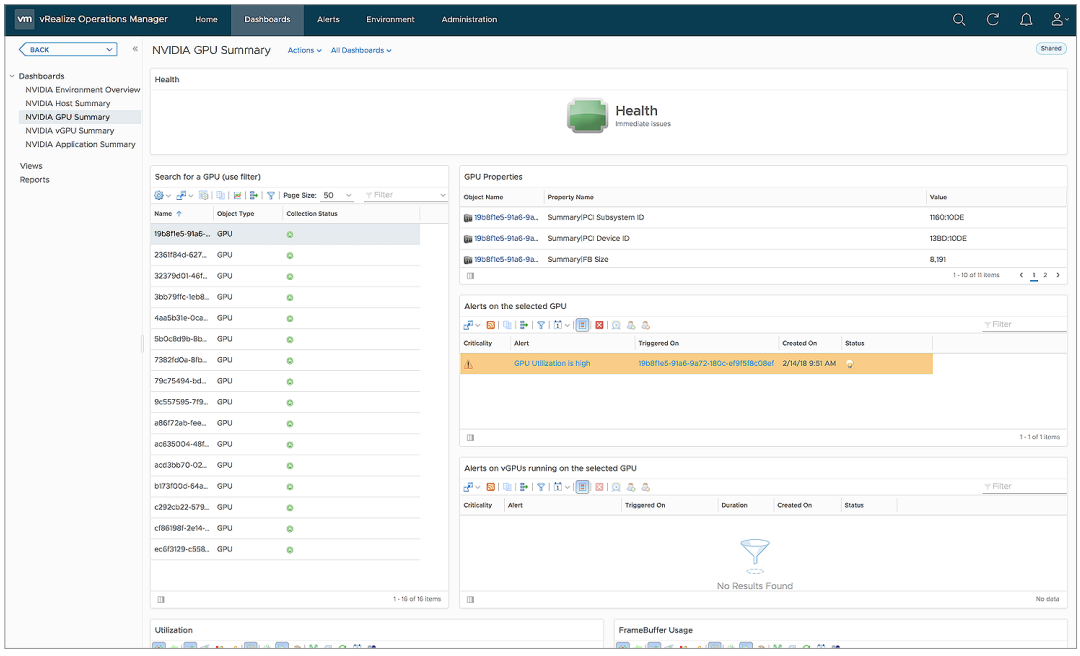
- **Environment view** for complete visibility of a cluster, with real-time GPU, memory, temperature, alerts, and top 25 GPUs and VMs for a cluster
- **Host view** for individual host drill down, with real-time view of a host including GPU, Encode, Decode, memory, temperature, alerts, and top 25 GPUs and VMs on that host
- **GPU view** for individual GPU drilldown, with real-time GPU, Encode, Decode, FB utilization, alerts, and top 10 apps per VM
- **vGPU view** gives insights on each GPU, with real-time GPU, Encode, Decode, FB utilization, alerts, and top 10 apps per VM
- **Application view** for the most granular visibility into GPU, with real-time GPU, Encode, Decode, FB utilization per app



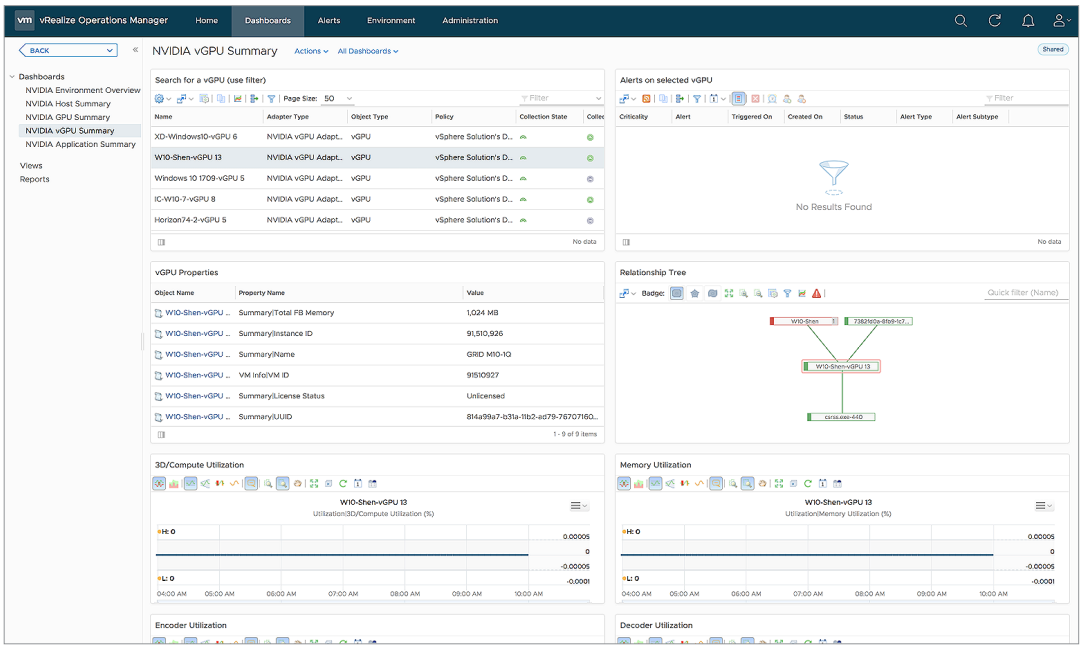
Environment View



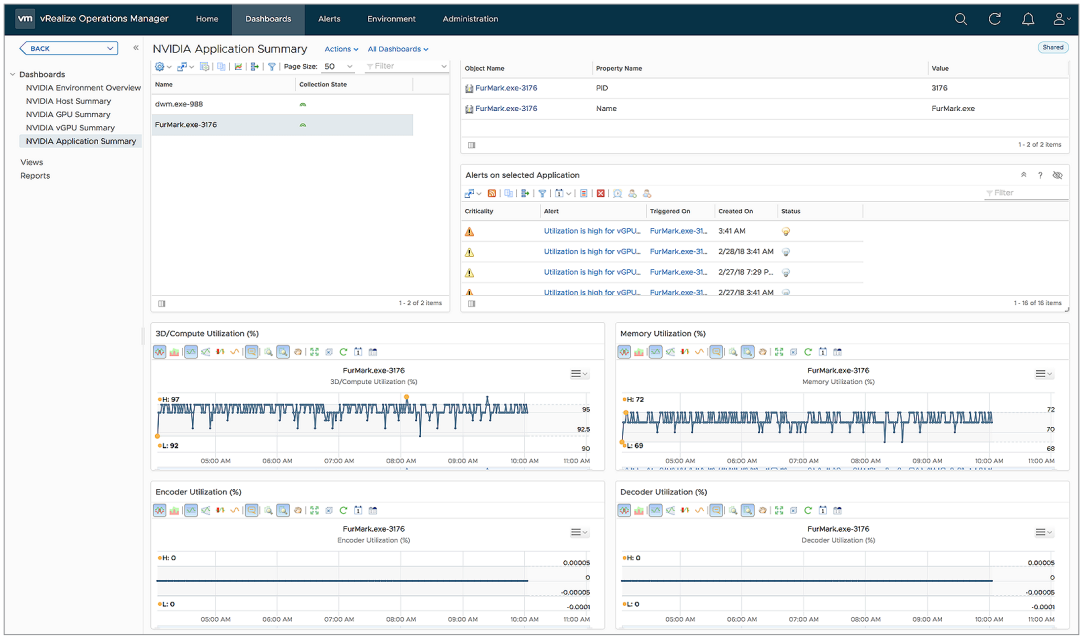
Host View



GPU View



vGPU View



Application View

Other Tools:

Customers also have a broad choice of monitoring tools outside of VMware and Citrix products. These include:

- **PerfMon.** Used by help-desk admins to get an understanding of a system (physical or VM), with FB, GPU, Encode, and Decode utilizations metrics
- **nvidia-smi.** Used by system admins and help desk support teams to understand their current GPU deployment and utilization data, at the guest or host level
- **Gpuprofiler.** Shows GPU and memory utilization inside a VM
- **Powershell.** Scripting becomes a breeze with WMI extensions and the python bindings NVIDIA provides
- **ControlUp.** NVIDIA virtual GPU metrics is integrated into the ControlUp console for efficient troubleshooting of issues and identifying trends in your VDI environment
- **eG Enterprise.** GPU-sourced insights that aid in design right-sizing, proactive infrastructure optimization, and help desk issue remediation
- **Lakeside Systrack.** Used for assessment, deployment, and help desks, with relevant utilization metrics and vGPU monitoring capabilities
- **Liquidware Stratusphere.** Virtual GPU visibility and metrics with Stratusphere UX for better management, optimization, monitoring, and troubleshooting of graphics-related challenges

Feature	NVIDIA Virtual GPU Management Pack for vROps	vROps for Horizon with NVIDIA Virtual GPU	Citrix Director with NVIDIA Virtual GPU	Nutanix Prism	Lakeside	Liquidware	ControlUp	eG Innovations
Host Monitoring								
3D	X			X	X			X
Encode/Decode	X				X			
Memory Usage and Utilization	X			X (only usage)	X			X
Guest Monitoring								
3D	X	X		X	X	X	X	X
Encode/Decode	X	X		X	X	X	X	X
Memory Usage and Utilization	X	X (only utilization)	X (only utilization)	X (only utilization)	X	X	X	X
Application Monitoring								
3D	X	X			X	X	X	
Encode/Decode	X	X			X	X	X	
Memory Usage and Utilization	X	X			X	X	X	

COMPLETE SOLUTION TO FIT ANY ENVIRONMENT

There are three different ways NVIDIA insights will be available to IT admins and end users:

1. **Standard GPU and VM Tools.** All existing Windows tools, such as PerfMon that rely on the WMI counter, will benefit out-of-the-box. Also, nvidia-smi is an NVIDIA-provided command line tool widely used by IT admins. Plus, the command line tool gets extended with these new functionalities.
2. **Building an Enterprise Partner Ecosystem.** Windows-based tools are limited to only showing single guest-level information. To achieve a single pane-of-glass management of the VDI infrastructure, from the client to the server to the GPU and beyond, a number of leading monitoring vendors have integrated NVIDIA GPU metrics into their popular products. These include Lakeside, Liquidware, eG Innovations, and ControlUp. VMware and Citrix have also extended their monitoring products with NVIDIA GPU and vGPU visibility.
3. **Open Platform for Development.** The NVIDIA SDK supports those IT admins or solution vendors who want to build their own infrastructure monitoring and management tools leveraging host, guest, and application graphics insights.

CONCLUSION

NVIDIA's management and monitoring stack provides complete insights into both the physical and virtual environments, along with deeper insights through application-level monitoring. These insights are integrated into the tools you're already familiar with today, such as Citrix Director and VMware vROps. Broad partner ecosystem adoption (i.e. Lakeside, Liquidware, eG Innovations, and ControlUp) and great out-of-the-box monitoring capabilities (i.e. PerfMon and nvidia-smi), empower IT to better manage large-scale deployments and complex IT environments. They also provide the agility to address the needs of demanding users. In addition, live migration features with NVIDIA GPU-accelerated VMs let IT deliver quality user experiences, as well as high availability.

Learn more about how to deploy NVIDIA virtual GPU metrics in your environment. Read our Virtual GPU Software Management SDK User Guide or the NVIDIA vROps Management Pack User Guide. Test these solutions for yourself by downloading the NVIDIA vROps Management Pack or the Virtual GPU Software Management SDK, or by deploying one of the solutions from our partners. Stayed tuned for information on general availability of migration features for your GPU-enabled virtualized environment.