WHITEPAPER

# USING NVIDIA VIRTUAL GPUs TO POWER MIXED WORKLOADS

## Accelerate your organization

APRIL 2019

NVIDIA

# CONTENTS

# OVERVIEW

The high demands of today's professional applications mean there are more use cases for GPUs across the enterprise than ever before. Designers and engineers rely on graphics-intensive applications featuring 3D visualization, many of which include AI-enhancements. Data scientists run compute-intensive applications powered by AI, deep learning, and inference. Even knowledge workers are using increasingly graphics-intensive office productivity applications—the most recent Windows 10 updates have resulted in a 20% increase in GPU consumption alone. Quickly processing data with GPUs exponentially speeds these applications, improving performance and helping companies innovate faster.

However, professionals using highly specialized applications still experience productivity setbacks. Many workflows are frequently slowed by inefficiencies in the data center and the limitations of physical workstations. Whenever users download or upload large files, submit files for rendering, or wait on results from jobs submitted to the High Performance Computing (HPC) cluster, they lose valuable time. Either they're waiting on processes that require compute power far beyond their endpoint, or they're waiting on work performed in the data center that would benefit from access to untapped resources. The end result is delays in time-to-market.

The enterprise is now looking for IT solutions that can significantly speed up these professional workflows so that decisions can be made before the end of meetings or the end of the workday. IT professionals need to enable them by deploying flexible IT infrastructure that seamlessly delivers better GPU utilization so they can satisfy more business needs with less resources and investment.

## ACCELERATE WORKFLOWS WITH NVIDIA VIRTUAL GPUs AND MIXED WORKLOADS.

The best way to accomplish these goals is with virtualization using NVIDIA virtual GPU (vGPU) software combined with the industry's most powerful NVIDIA® GPUs. Because the same NVIDIA GPUs can be used for both virtual machines (VMs) and deep learning, AI, inferencing, HPC, and other applications, IT can bring workloads together for greater data center efficiency. The resulting mixed workloads ensure GPU resources are maximized in the data center while professional workflows get faster.

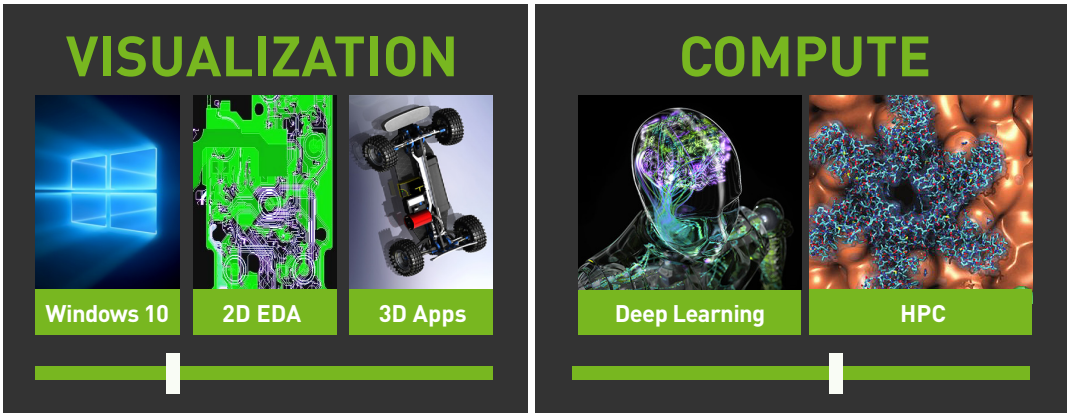
Using common server resources to run HPC and virtualization workloads adds flexibility and operational efficiency. With mixed workloads, compute resources can be used by engineers and designers for VDI during the day, then those same resources can be repurposed at night to run compute jobs, significantly increasing utilization. And since you're no longer managing physical workstations, IT costs decrease too.

Virtualizing workstations ensures users enjoy mobility and security with no compromise to performance. For example, a user on a VM doesn't actually download a model or data like they would need to do if they were working on a physical workstation. Instead, the VM is able to access large files and datasets stored in the data center in seconds, dramatically increasing productivity. Because data is not actually downloaded or uploaded, project teams also avoid issues with
version control.

From photorealistic rendering used in product design to data intensive simulation for manufacturing, and exponentially faster rendering for media and entertainment—mixed workloads accelerate professional workflows to create numerous operational efficiencies, including improved output accuracy, reduced time-to-market, and increased user productivity.

## WHAT ARE MIXED WORKLOADS?

Not all workloads are equal. The majority of today's applications fall into the broad segments of visualization and compute. Within these two segments, there are multiple sub-segments, each with their own unique requirements.




- **Office productivity applications** increasingly need GPU acceleration as applications become more demanding and enterprises migrate to Windows 10.
- **2D electronic design automation (EDA) applications** often require GPU acceleration to perform well in a VDI environment.
- **3D professional applications** need GPU acceleration (e.g., Dassault Systèmes CATIA and SOLIDWORKS).
- **Deep learning, AI, and inferencing** require powerful GPUs to enable better efficiencies.

With NVIDIA GPUs in the data center, businesses can now run multiple types of workloads—whether it's Monte Carlo simulations for assessing financial investments or 3D graphics and data-intensive workloads for oil and gas exploration—using the same infrastructure by repurposing hosts that run VDI during the day to run HPC and other compute workloads at night.

### MIGRATION OF GPU-ACCELERATED VMs

Mixed workloads are possible because of live migration. Live migration is the process of moving a running VM from one physical host system to another without end-user interruption or data loss. While live migration has been around for years, live migration for GPUs hasn't been possible until recently. NVIDIA virtual GPU software deliver the industry's first—and to date the only—support for live migration of GPU-accelerated VMs.

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.

Achieving mixed workloads using live migration for GPUs is possible because NVIDIA's virtualization software—NVIDIA Quadro® Virtual Data Center Workstation (Quadro vDWS) and NVIDIA GRID®—run on the same Tesla Turing™, Volta™ and Pascal™ GPUs as deep learning, inferencing, training and HPC workloads. Now, using NVIDIA vGPU software and VMware vMotion or Citrix XenMotion, IT can achieve improved data center agility and live migrate users in seconds. As a result, IT can maximize data center utilization while also achieving more efficient server maintenance.

**Maximize data center utilization.** IT admins can consolidate live VMs onto underutilized server nodes at any time. At the end of the day, when users start going home, IT can consolidate remaining VMs by live migrating them to a different host. Then they can repurpose the original host to run compute workloads, like HPC and deep learning, at night. When graphics resources are needed for VDI the next day, IT admins can simply repurpose the NVIDIA GPUs to virtual GPUs to support VDI again.

**Keep servers healthy.** Migrating VMs allows IT to perform critical services like workload leveling, infrastructure resilience, and server software upgrades on their own time, without any VM downtime. This ensures servers are always working at their peak and end-users never experience any disruption or data loss.

**Additional benefits:**

- **Minimal Business Disruptions.** Perform server maintenance like hardware replacement, upgrades, or software updates, without scheduling downtime.
- **Improved Server Density.** Manually load balance users and consolidate common profile types for better density.
- **Optimized Infrastructure Utilization.** Move VMs to another host to enable a host and GPU to be used for compute after hours.
- **Increased Agility.** Take advantage of insights from NVIDIA virtual GPU software monitoring coupled with migration to ensure a quality user experience with high availability.

## MIXED WORKLOADS IN ACTION

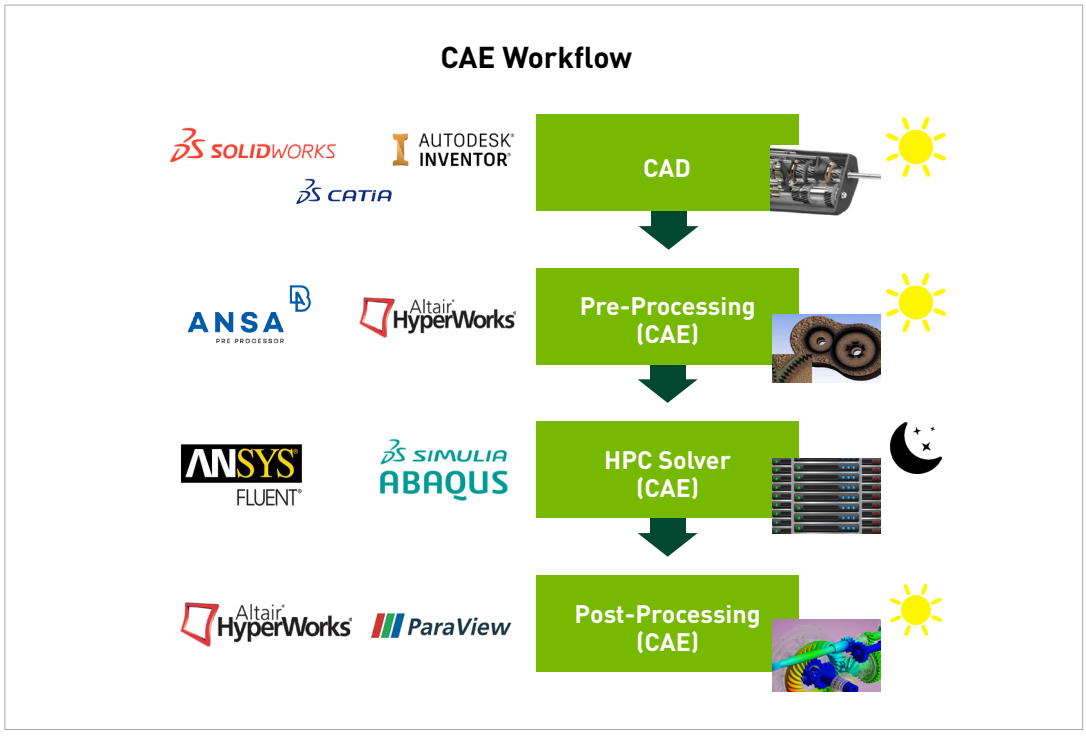
The advantages of implementing mixed workloads are best understood by reviewing a typical computer-aided engineering (CAE) workflow in-depth. In this example, an engineer is building a racecar and as part of the project, the engineer will determine whether the racecar's structure will be able to handle the forces of a real-life scenario with a factor of safety to validate the design with the lightest weight material possible. Here's his workflow:

1. **Start with a CAD design.** The complete assembly of the racecar is displayed in the CAD application. In this example, the racecar's spindle geometry is the structure that will be analyzed and validated.
2. **Pre-processing.** In this process, the imported geometry will be tested

for errors and simplified for simulation (to determine whether there are any sharp edges, etc.). The next step is to build a finite element mesh to discretize the geometry. This process is called meshing. Even the smallest parts of the model (e.g., a spindle) can have as many as two million elements, and this in turn will translate to many degrees of freedom. At this point, the engineer will assign a material and apply restraints and forces to the geometry.

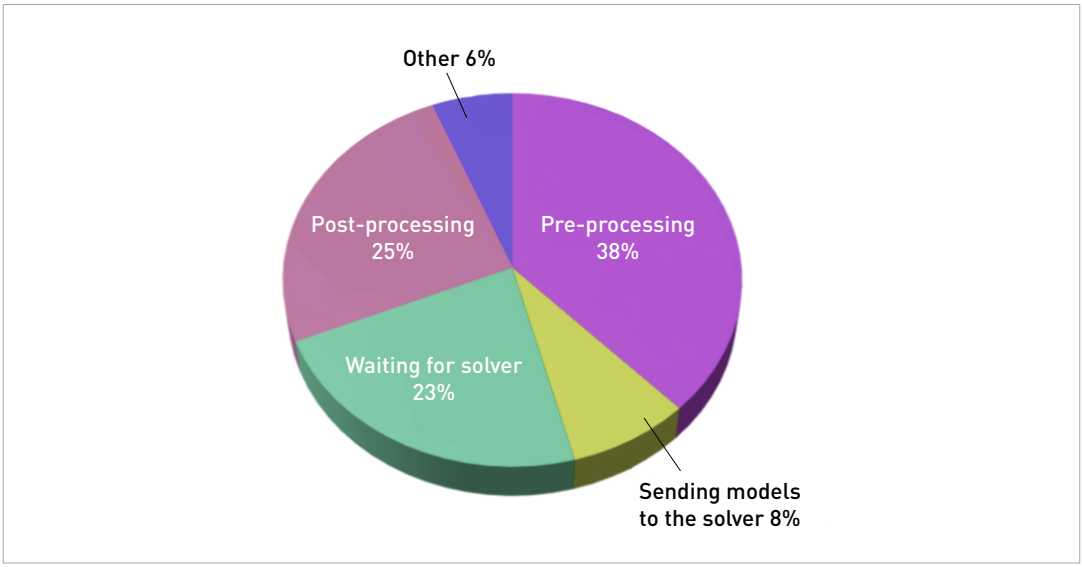
3. **HPC solver.** Next, the HPC solver solves the finite element model that was built. This is a highly computationally intensive job. A single simulation might take hours or even days. This is where NVIDIA GPU acceleration enables faster results for more efficient computation and job turnaround times, delivering more license utilization for the same investment.

4. **Post-processing.** Once analysis is complete, the engineer will analyze and infer the results to validate the design, checking for structural integrity and making necessary changes to the design/finite element model. Then, the simulation will be re-run.



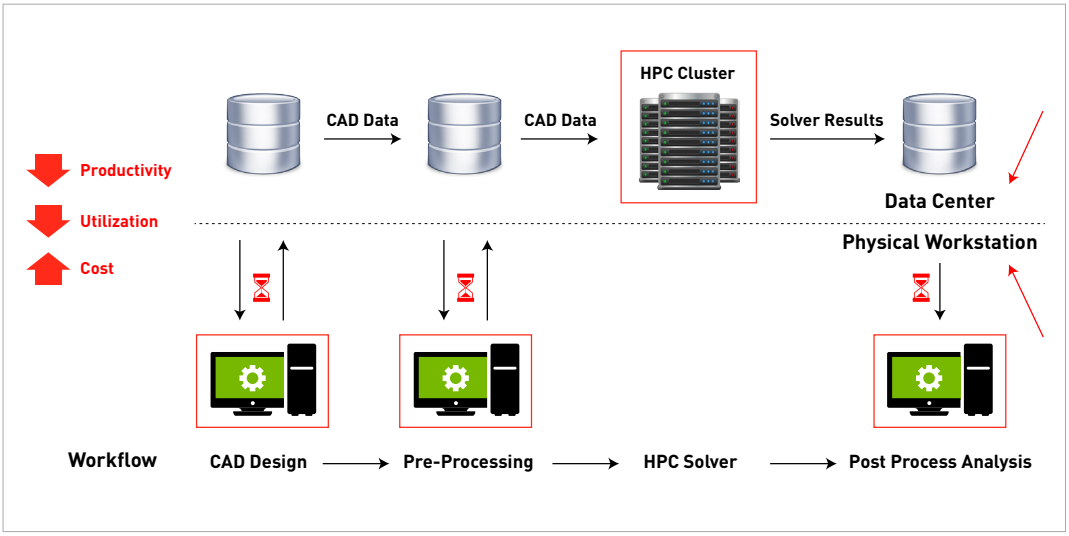

As engineers move through this workflow, they'll use a variety of applications. Before they leave one app to start another workflow, they'll save their files. Later, they'll go back to check them, switching back and forth to different applications that each require unique amounts of compute and visualization resources. To maximize efficiency, engineers typically design models and submit them for pre-processing multiple times during the workday. In the evening, they submit them to the HPC solver so that when they arrive at the office the next morning the results are in and ready for analysis.

How much time is spent on each phase of a typical workflow? Several studies have demonstrated that engineers spend about 66 percent of their time in the design, pre-processing, and post-processing phases. Basically this means that these phases are most interactive—where engineers sit at their workstations and actively work on models. The remaining one-third of the time, the HPC solver is working on analysis, which is not interactive.




## TRADITIONAL DEPLOYMENTS WASTE VALUABLE RESOURCES.

In a traditional deployment, engineers typically use physical workstations for CAD work. There's typically an HPC cluster and storage in the data center for purposes of collaboration. When engineers want to work on CAD data files, they'll check them out. Then, they'll check them back in when they're ready to do pre-processing. The data is then passed to the HPC cluster, and the HPC solver analyzes the results. Afterwards, these results are imported back into the engineers' workstations for further analysis.

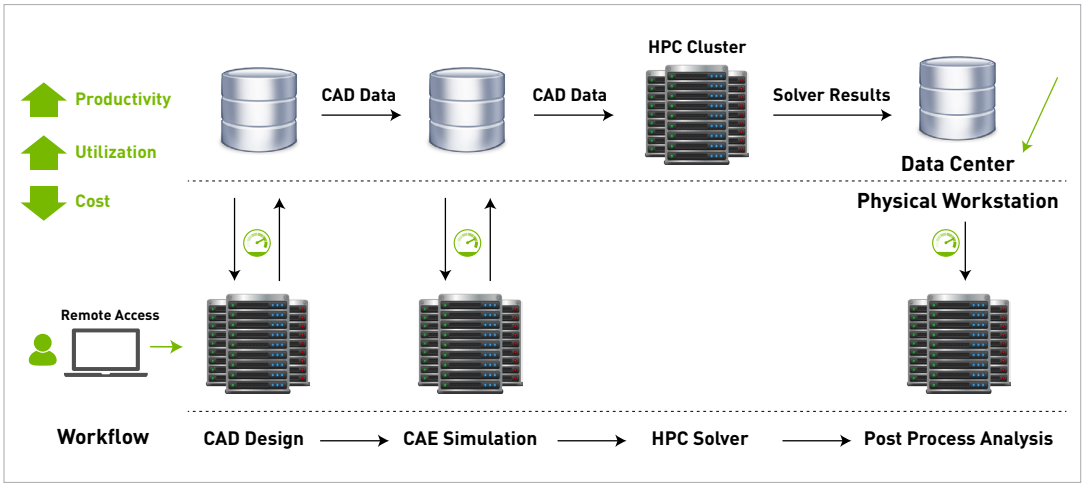


This type of deployment is rife with drawbacks and limitations. For starters, productivity is negatively impacted. Due to the type of analysis it performs, the HPC solver typically produces files with terabytes of data. Other stages produce tens to hundreds of megabytes. As a result, engineers can spend a long periods waiting to upload and download files. Depending on where workstations are located and the size of files, file access can take anywhere from one to 20+ minutes. As these models undergo numerous iterations throughout the day, engineers upload and download files repeatedly. Consider the cumulative impact on productivity for large engineering teams—a significant amount of time is spent waiting around.

And productivity isn't all that's wasted, so is utilization. During the workday, workstations are highly utilized while the HPC cluster is probably not utilized much at all. At the end of the day when engineers go home, their workstations are idle while the HPC cluster is being utilized a great deal. Moreover, costs are high. The IT department must maintain and manage both the data center as well as physical workstations. This means IT staff are spending all of their time procuring, deploying, upgrading, and patching two sets of hardware and software.

## VIRTUALIZED DEPLOYMENTS REMOVE LIMITATIONS.

Consider the impact on productivity, utilization, and costs if you move workstations to the data center and virtualize this environment.




Productivity increases instantly. With workstations now located close to the storage where CAD files are hosted, engineers don't have to wait to download files; they access them in seconds. There is also very high throughput to access storage, so even when engineers are iterating on files repeatedly they're working efficiently with minimal wait time.

Another advantage is that utilization goes up. During the workday when the HPC cluster is idle, compute resources can be used by engineers who are actively working on CAD designs. Essentially, the HPC cluster can be repurposed to do both pre-processing and post-processing. Lastly, the IT department is no longer managing physical workstations; it only needs to manage the data center. With very high utilization and less resources to manage, expenses are lower and costs go down.

Now let's take a look at how this mixed workloads concept plays out in the data center.

## BENEFITS OF GPU-ACCELERATED VIRTUALIZATION

1. Save time uploading and downloading large models and data files

   *A project that once required 8-16 hours of processing time can be reduced to approximately 40 minutes of processing time.*

2. Run mixed workloads for continuous resource utilization

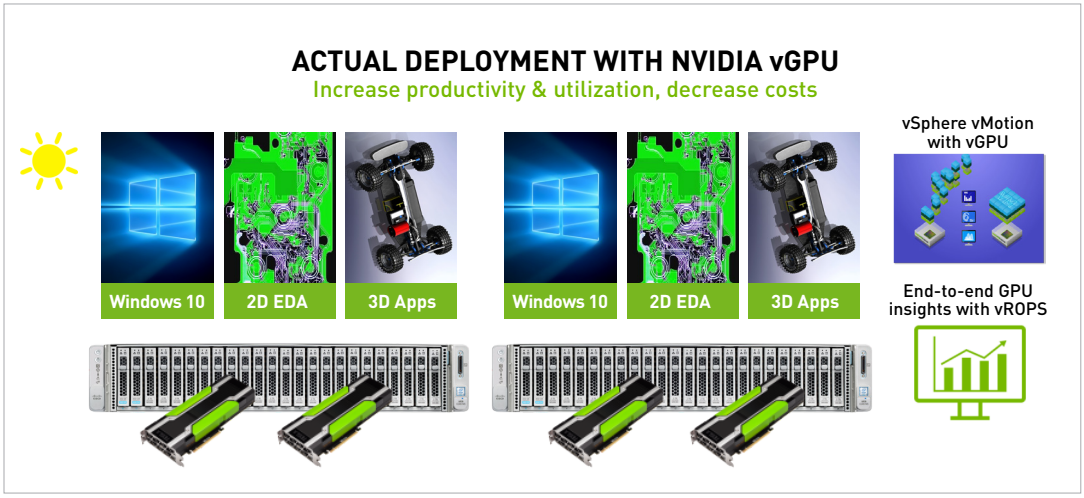
3. Lower CAPEX and OPEX

   *Physical workstations can cost $10,000+. That's more than double the cost of a virtual workstation.*

4. Keep data securely in the data center where it can be remotely accessed from anywhere on any device

## HOW MIXED WORKLOADS MAXIMIZE RESOURCES.

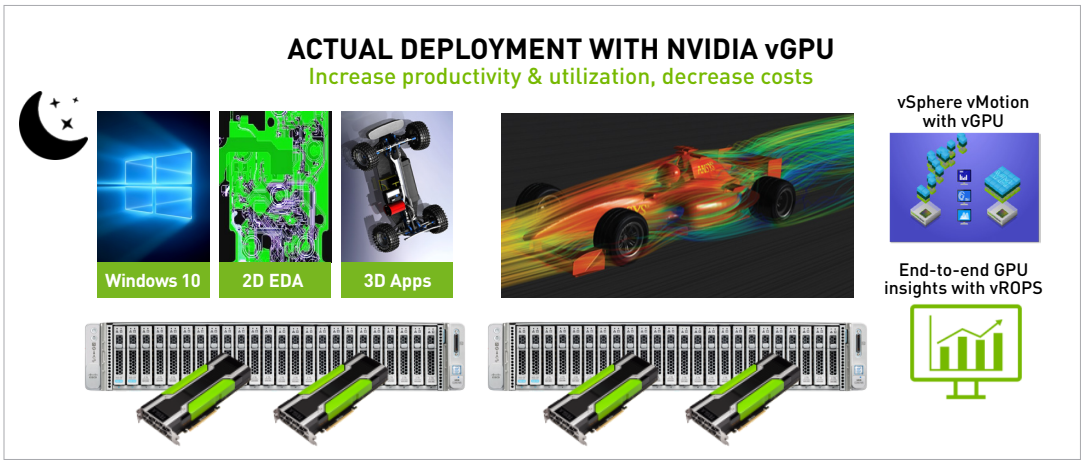
Here's how implementing mixed workloads in the data center works. In the illustration below, two server nodes are each installed with two Tesla data center GPUs running NVIDIA virtual GPU software. These servers are being used to host a VDI environment and are running multiple VMs during the workday.




The VMs are running Windows 10 and office productivity applications, 2D EDA applications, and 3D CAD and CAE workloads. Throughout the day, all of these workloads are hosted by the same compute nodes. When users begin to log off of their VMs at the end of the day, both nodes start to become underutilized. In the past, when compute resources were underutilized, it was impossible to use those spare resources. If you want to run an HPC workload and get quick results, harnessing those extra resources is crucial.

Now with NVIDIA vGPU live migration, you can live migrate any remaining VMs, consolidating them onto one node as users log off at the end of the day.


In doing so, you're freeing up the second node which you can now repurpose to run your HPC solver workload during the night.




The next morning, when the HPC solver finishes and staff return to work, VMs can go back online and the same nodes can be used for post-processing.

# DEPLOY MIXED WORKLOADS

## MAXIMIZE DATA CENTER EFFICIENCY AND AGILITY

Now that you've seen the benefits of running mixed workloads in your data center, here's what you'll need to get started:

- **NVIDIA data center GPUs.** Featuring the industry's highest performance available, NVIDIA data center GPUs can be used across both compute and virtualization workloads.
- **NVIDIA virtual GPU software.** NVIDIA vGPU software enables each VM to access the power of NVIDIA GPUs, ensuring better performance and a great user experience. It also provides the power to right-size your VMs.
- **Continuous uptime support.** Ensure you achieve full utilization and uptime for your environment with GPU enabled live migration, enabled with NVIDIA vGPU technology and supported with VMware vMotion or Citrix XenMotion.
- **End-to-end monitoring and insights.** Your workloads are complex. NVIDIA vGPU software offers integrations with several management and monitoring tools, such as VMware vROPS and Citrix Director, offering end-to-end insights across your entire virtualization stack. Get management and monitoring of the GPU at the host, VM, and application level for quick issue resolution and proactive management.

# CONCLUSION

Achieving mixed workloads is possible because NVIDIA vGPU software—NVIDIA Quadro vDWS and NVIDIA GRID—run on the same Turing, Volta and Pascal GPUs as deep learning, inferencing, training and HPC workloads. With NVIDIA GPUs in the data center, IT can use live migration to consolidate VMs onto underutilized server nodes at any time to repurpose hosts and ensure that HPC and other compute workloads are always accessing the most resources possible. Migrating VMs also allows IT to perform critical services like workload leveling, infrastructure resilience and server software upgrades without any VM downtime or data loss. With mixed workloads, IT can maximize data center resources while delivering high availability and a quality user experience.

To learn about how to deploy NVIDIA mixed workloads in your environment, read our Reference Design Guide: Mixed Workloads – Virtual Desktops and HPC on Common Architecture.

Learn more about how NVIDIA drives efficiency in the data center.