GIGABYTE™

# Virtual GPU Technology in VDI

NVIDIA®

There is no better time than now for Virtual Desktop Infrastructure (VDI) using Virtual GPUs (vGPU) to take hold. Workers are more scattered, as are their devices, and these people often need to work remotely. This poses all sorts of problems in terms of security, productivity, management of data, and IT management as these workers become more and more dispersed. In addition, there is a shift to a greater need for engineers and designers to use 3D applications and other graphically intensive visualization applications with large amounts of data. In the past, CPU-only VDI environments deployed to centralize management; however, these solutions could only reach office workers, and power users were left out and had to stick with dedicated desktop machines that had to be upgraded and maintained individually, on top of data security concerns. Now, VDI has reached the point where these power users can have a solution by combining VDI with vGPU. Large data sets can be centralized, accessed, and processed using GIGABYTE servers with GPU virtualization handled by NVIDIA virtual GPU technology.

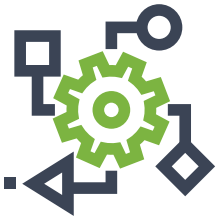
# The Evolution of the Digital Workforce

Not too long ago it was customary for workers to have their own PCs that operated independently from colleagues. As technology improved and realized the importance of collaboration, and not just in meetings, but also in the work produced via a computer, so did the need for a more complex structure of sharing resources, such as documents, files, computing resources, and so on.

Servers stored all sensitive data on site, and this solution worked for quite some time. However, it did not take long for something better to come along. Server virtualization allowed for better availability of resources, improved scalability, heightened security, and greater mobility.

This worked for most office workers, but that was exactly where it stopped. Users that required graphically intense applications were left out of VDI adoption because the user experience was not up to par. As GPU technology progressed, companies started optimizing virtual GPU software that could abstract GPU hardware at the hypervisor layer. Thus, it expanded the realm of profiles to include power users, designers, engineers and even AI scientists who required workstation-level performance for 3D graphics processing and HPC computing. It also did so with improved smoothness, and higher frames per second (fps), and thus upgraded the overall user experience.

# Driving VDI adoption



### Centralized Management

Application experience on user devices improves without device hardware upgrade; instead, tasks offload to the server, where there is a more robust, up-to-date system of hardware, to improve the user experience. Centralizing hardware also centralizes IT, which in turn, allows IT to adapt user devices and manage the server quickly.


### Cloud Solutions

Transition to the cloud started with data, and then applications. Virtual desktops soon became an extension of this concept. Now Virtual GPU (vGPU) is a key part of this picture of VDI adoptions as more companies support it and offer various virtualization solutions. The move to hybrid IT is inevitable.


### Remote Work

Flexibility in when and where work is done is increasingly desired as employees seek a better work-life balance outside the office. Compounded by global issues, employees may need to work outside the office while still remaining connected as if they never left.

### Security

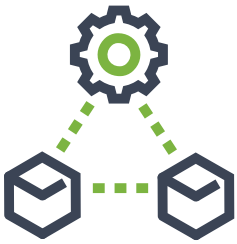
Data is stored and accessed in the data center, which mitigates the risk of losing sensitive company data or personal information from lost or stolen devices. At the same time, network speed has greatly improved with faster 4G or 5G Wi-Fi connectivity.

# Benefits for VDI

## Productivity & Flexibility

- Workers can work and access company documents (PDFs, Excel, photos…) from various endpoint devices at any location as long as they have internet access and VPN connection to transmit encrypted data.
- Support for different OS environments: Windows, iOS, Android, Linux. Any OS can be installed in virtual machines using the same GPU for operating independently and unknowing of each other's compute instance.
- Data are no longer limited by storage locality. Rather, data are stored, shared and pooled from a centralized server. Greater user access and resource control improves the efficiency of work.

## Ease of Management & Security



- Centralized servers allow IT administrators to quickly monitor, manage, upgrade, patch, and deploy compute resources from one location.
- New user instances deploy faster as there is no procurement process, and the preparation time takes less than an hour.
- Minimized risks of compromising business-critical and sensitive data from lost or stolen devices as data are safely stored in servers and not on user devices. Restricted user access also helps to prevent theft and intrusion.
- Backup servers can access and reclaim data in the event of a disaster as redundant servers have the same virtual resource abstraction running the same applications.
- Down time decreases significantly with users' ability to login into a virtual machine with vGPU resources. If a user instance stops working, the user can switch to another virtual machine and login to regain the access.

## Efficiency of Resource Utilization

- Computer hardware (CPU, RAM, GPU, storage, network connections) is allocated to virtual machines based on user application demand. It is not a one-size-fits-all model.
- Designers and power users receive the right amount of vCPU cores, vGPU memory, vRAM capacity, and virtual storage, compared to what needs to be allocated to knowledge workers, who need far less compute resources. Hardware is distributed on a basis of user profiling with transparency of actual hardware consumption levels.
- Pooling resources allows for flexibility in performance adjustments. By optimizing resources, scenarios where resources are underutilized or overprovisioned do not occur.

## Cost Savings

- Total Cost of Ownership (TCO) is reduced thanks to leaner or better organized IT staff, centralized software management, and improvements in power usage effectiveness (PUE).
- Power consumption decreases for running virtual desktops instead of traditional PCs. The pooled resources in the server are more efficient than individual devices because of economies of scale.
- Upgrading hardware entails expanding or replacing hardware on the host server. This method is much more efficient than tracking down all devices to upgrade or reclaim them individually.
- Multiple virtual environments can be built on one single infrastructure. With more high-density servers and less purpose-built ones, space saving decreases operating costs. Also, greater efficiency is achieved through easier management of virtual resources pools.

## User Experience

- User's workload will determine the appropriate virtual machine configuration. At the same time, the virtual machine will, and must, match the user's high expectations or performance. Virtual machines act like traditional PCs in the way that there are little to no perceivable differences in performance, thanks to vGPU acceleration.
- New virtual machines or instances can deploy immediately and be customized for different users. Updates are handled and pushed out centrally by the host server.
- Disaster recovery time only takes minutes after a failure, as a golden image is available for OS and programs to replace the destroyed environment. Data safely stored on the server can be quickly restored and a new environment created for access to the data.

# Application choices for VDI with Virtual GPU

The inclusion of virtual GPU into VDI allows users to have an improved experience for workloads that require using 3D graphics or other visualization applications, as well as for HPC and AI applications. In addition, users of these applications may need high-resolution monitors (4k or 8k) and multiple monitors at one time. Historically, latency hits occurred when high-resolution displays and multiple monitors were used; the inclusion of vGPU has solved that problem. To display graphically intense imagery a VDI instance must have a virtual GPU to give the user a great, and usable, experience with fast response time on top of a high-speed network. Some examples of vGPU applications: Computer-aided design (CAD), Adobe Premiere Pro, Geographical Information System (GIS), SOLIDWORKS, applications using CUDA, OpenGL or DirectX, and more. Other applications may not be seen as graphically demanding, but they actually are. For example, Office 2019, Skype, web browsers, PowerPoint, streaming video (YouTube), etc. all require accelerated graphics performance for a good user experience. After all, a VDI instance must be similar to or better than a desktop system for users to accept it.

# VDI Architecture

The concept of VDI is to make desktop stations from a server. To provide VDI instances, a server is outfitted with typical server hardware (CPU, RAM, storage, network interconnects, etc.) upon which a hypervisor is installed to abstract it.

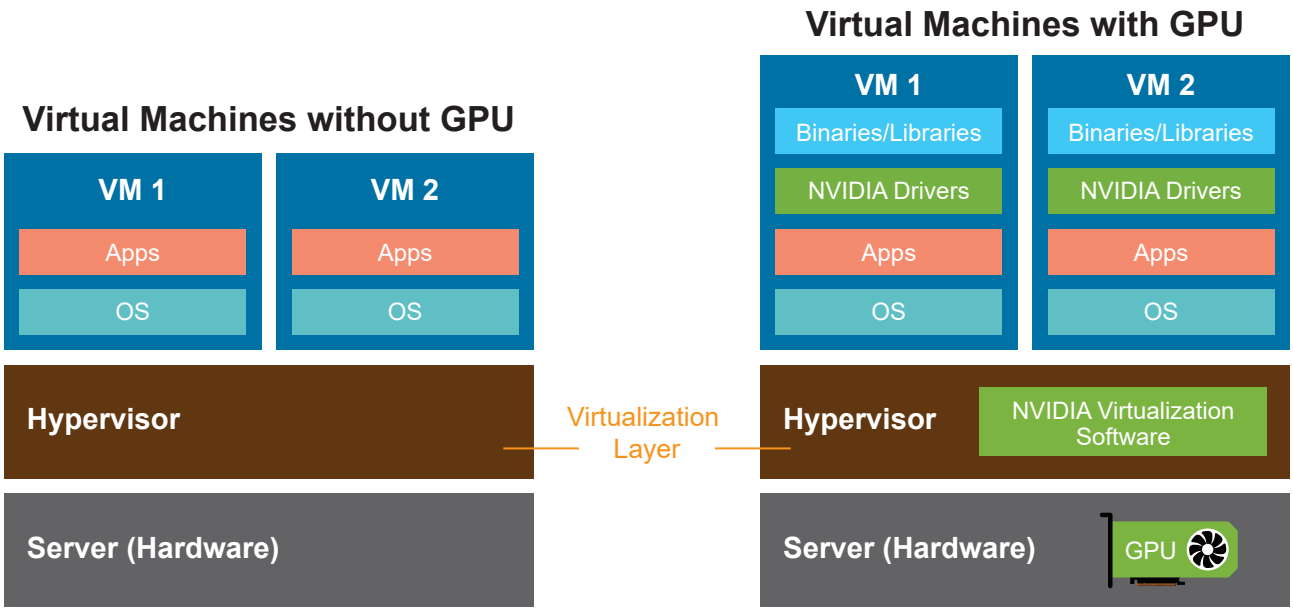
Popular choices of hypervisors are:

| Microsoft Hyper-V | Citrix XenServer | VMWare vSphere |
|---|---|---|

The hypervisor creates a virtualization layer for virtual machines. This hypervisor sits in between the virtualization layer and the hardware, and it contains a Virtual Machine Manger. On the virtualization layer, virtual machines reside, and each contain applications and an operating system (Windows, Linux, etc.). At this point a CPU only VDI has been created.

For VDI with Virtual GPU, software such as NVIDIA Virtual GPU Manager installs in the hypervisor. This software coupled with NVIDIA vCS, Quadro vDWS, GRID vPC, or GRID vApps allows customization of the virtual machine to fit the user type. On the vGPU layer are virtual machines, each with its own OS, applications, binaries/libraries and NVIDIA drivers.

The following figure compares different degrees of virtualization. The VDI on the left is a server that allocates its hardware into two virtual machines. However, this system does not include a GPU. On the right, is a server that is virtualized to include Virtual GPU (vGPU) that can be allocated into virtual machines.

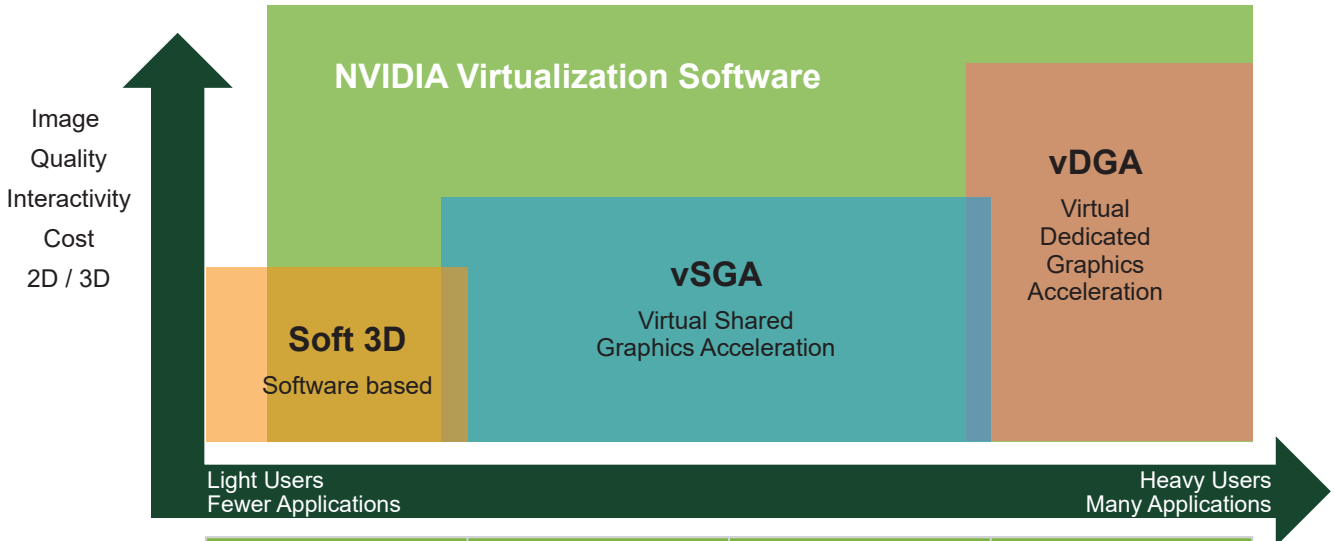


The following depicts use cases for employing VDI with Virtual GPU:

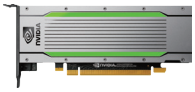
### Comparison of NVIDIA vGPU software:

| NVIDIA vGPU Software | | | | |
|---|---|---|---|---|
| **Product** | **vCS** | **Quadro vDWS** | **GRID vPC** | **GRID vApps** |
| **Used In** | Artificial Intelligence, Deep Learning, Data Science | CAD/DAE, 3D modeling, Scientific simulation, Data visualization, HPC applications | Productivity applications by office workers and knowledge users in all industries, with full desktop environment | All vGPU supported applications, without desktop environment |
| **Sample Applications** | TensorFlow, ONNX, mxnet | Autodesk 3ds Max, ANSYS Fluent, SOLIDWORKS | Bloomberg Autodesk AutoCAD PACS | All vGPU supported applications |
| **NVIDIA Accelerators** | Ampere A100 Quadro RTX8000 Turing T4 | Quadro RTX8000 Quadro RTX6000 Turing T4 | TuringT4 | Quadro RTX6000 Turing T4 |
| **Type of User** | Designer, Engineer, AI Scientist, Power User | Power User Knowledge Worker | Knowledge Worker | Task Worker, AI Scientist |

*Use cases vary, and the table gives typical applications.




| Task Worker | Knowledge Worker | Power User | Workstation User |
|---|---|---|---|
| Basic data input simple office app | Standard office apps & productivity tools | Some graphically demanding apps | Workstation level dedicated graphics |

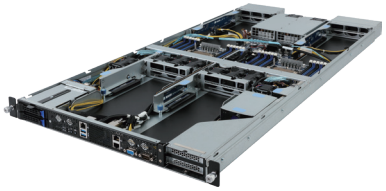
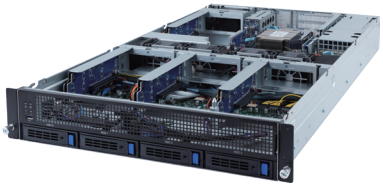
| NVIDIA Accelerators for Virtualized Environment | | | | | | |
|---|---|---|---|---|---|---|
| **Models** | **A100 PCIe** | **RTX A6000** | **A40** | **Quadro RTX 8000** | **Quadro RTX 6000** | **T4** |
| |  |  |  |  |  |  |
| **Architecture** | Ampere | Ampere | Ampere | Turing | Turing | Turing |
| **CUDA cores** | 6,912 | 10,752 | 10,752 | 4,608 | 4,608 | 2,560 |
| **Single-Precision (FP32)** | 19.5 TFLOPS | TBD | TBD | 16.3 TFLOPS | 16.3 TFLOPS | 8.1 TFLOPS |
| **GPU Memory** | 40 GB HBM2 | 48 GB GDDR6 | 48 GB GDDR6 | 48 GB GDDR6 | 24 GB GDDR6 | 16 GB GDDR6 |
| **Memory Bandwidth** | 1.6 TB/s | 768 GB/s | 696 GB/s | 672 GB/s | 624 GB/s | 320 GB/s |
| **Interface** | PCIe Gen 4 | PCIe Gen 4 | PCIe Gen 4 | PCIe Gen 3 | PCIe Gen 3 | PCIe Gen 3 |
| **Max Power** | 250W | 300W | 300W | 295W | 295W | 70W |
| **Form Factor** | dual-slot | dual-slot | dual-slot | dual-slot | dual-slot | single-slot |
| **Usage** | Ultra-high-end rend1ering, 3D design, AI and data science | High-end render-ing, 3D design, AI, and compute workloads | Mid-range to high-end 3D design and creative workflows | High-end render-ing, 3D design, and creative workflows | Mid-range to high-end render-ing, 3D design and engineering, AI and data science | Entry-level to high-end 3D design and engineering, AI and data science |

The following is a list of GIGABYTE servers that are NVIDIA vGPU Certified. These servers are built specifically to handle HPC, AI, and graphically demanding applications.

| GIGABYTE Servers (NVIDIA vGPU Certified) | | | | |
|---|---|---|---|---|
| **NVIDIA Models** | **A100 PCIe** | **Quadro RTX 8000** | **Quadro RTX 6000** | **NVIDIA T4** |
| **1U G-series** | G191-H44 | G191-H44 | G191-H44 | G191-H44 |
| **1U OCP-series** | - | - | - | T181-G23, T181-G24, T181-Z70 |
| **2U G-series** | G291-280, G291-281 G292-Z20, G292-Z40 | G291-280, G291-281 G242-Z10, G292-Z42 | G291-280, G291-281 G242-Z10, G292-Z42 | G291-280, G291-281, G291-Z30, G242-Z10, G291-Z20, G292-Z42 |
| **2U R-series** | R281-3C2, R281--G30 R282-Z93 | R282-Z93 | - | R281-G30, R281-3C2 R282-Z93 |
| **2U H-series** | - | - | - | H231-G20 |
| **4U G-series** | G481-HA0, G482-Z50 G492-Z50, G492--Z51 | G481-H80, G481-HA0 | G481-H80, G481-HA0, G481-HA1, G482-Z50, G481-Z51 | G481-H80, G481-HA0, G481-HA1, G482-Z51 |

# GIGABYTE Servers for Virtualization with vGPU

GIGABYTE has a range of servers designed for VDI. The G-series targets GPU dense systems and are designed for AI, deep learning, video streaming, and VDI workloads.

| Model | G191-H44 | G242-Z11 | G291-281 |
|---|---|---|---|
| |  |  |  |
| **Deployment & Benefits** | Ideal to scale up for 5G network infrastructure or deployment in a small space. Dual Intel Xeon Scalable processors and up to 4 full-length full-height GPUs. | Ideal for scale-out deployment in virtualization for GPU-centric workloads. High core count AMD EYPC™ processor and up to 4 GPUs with direct PCIe Gen4 x 16 connection to CPU. Also, 4 x 3.5" SATA and 2 x 2.5" U.2 (Gen 4) | Ideal for scale-up deployment in virtualization for GPU-centric workloads. Dual high-frequency Intel® Xeon® Scalable processors in a compact 2U chassis with balanced CPU-GPU ratio across roots. Support for up to 8 double slot GPUs. |

NVIDIA QVL https://www.nvidia.com/en-us/data-center/resources/vgpu-certified-servers/





2020/10 V1.0