DELL Technologies
Server and Infrastructure Engineering

Direct from Development

Virtualized GPU Instances on Dell EMC PowerEdge Platforms for Compute Intensive Workloads

Tech Note by

Ramesh Radhakrishnan
Janet Morss
Mike Bennett
Matt Ogle

Summary

In this DfD we address a common problem that is faced by IT teams across different organizations – being able to efficiently share and utilize NVIDIA GPU resources across different teams and projects.

AI adoption is growing in many organizations leading to increased demand of GPU accelerated compute instances. We explore how IT teams can leverage existing investment in virtualized infrastructure combined with NVIDIA Virtual GPU software to provide optimized and secure GPU-ready compute environments for AI researcher and engineers.

Motivation for GPU Virtualization

The requirement and demand for GPU accelerated compute instances is steadily rising in all organizations, driven primarily by rise of AI and Deep Learning (DL) techniques to realize increased efficiencies and improve customer interactions. IT environments continue to adopt virtualization to run all workloads and address requirements of providing secure and agile compute capabilities to end users. NVIDIA Virtual GPU software (previously referred to as GRID) enables virtualizing a physical GPU and allows it to be shared across multiple virtual machines. The rising demand for GPU accelerated compute instances can be achieved by virtualizing GPUs and deploying cost effective GPU accelerated VM instances. Enabling a centralized and hosted solution in the data center provides the security and scalability that is critical to enterprise customers.

NVIDIA Virtual GPU software enables virtual GPUs to be created on a Dell EMC server with NVIDIA GPUs that can be shared across multiple virtual machines. Better utilization and sharing are achieved by transforming a one-to-one relationship from GPU to user to one-to-many.

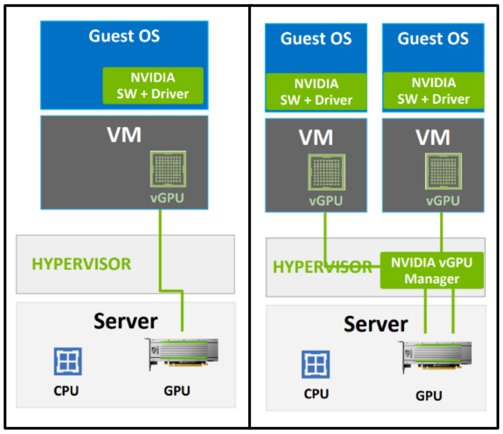

Figure 1. GPU enabled VM instances using GPU Pass-Though and GPU Virtualization (vGPU)

Traditionally, the IT best practices for compute-intensive (non-graphical) VM instances leveraged GPU pass-through shown in the left half of Figure 1. In a VMware environment, this is referred to as the VM DirectPath I/O mode of operation. It allows the GPU device to be accessed directly by the guest operating system, bypassing the ESXi hypervisor. This provides a level of performance of a GPU on vSphere that is very close to its performance on a native system (within 4-5%).

The main reasons for using the passthrough approach to expose GPUs on vSphere are:

(i) Simplicity: It is straightforward to allocate GPUs to a VM using pass-though and offer GPU acceleration benefits to end users
(ii) Dedicated use: there is no need for sharing the GPU among different VMs, because a single application will consume one or more full GPUs
(iii) Replicate public cloud instances: public cloud instances use GPU pass-through, and end user wants the same environment in an on-premises datacenter
(iv) A single virtual machine can make use of multiple physical GPUs in passthrough mode

An important point to note is that the passthrough option for GPUs works without third-party software driver being loaded into the ESXi hypervisor.

Disadvantages of GPU passthrough is as follows:

(i) The entire GPU is dedicated to that VM and there is no sharing of GPUs amongst the VMs on a server.
(ii) Advanced vSphere features of vMotion, Distributed Resource Scheduling (DRS) and Snapshots are not allowed with this form of using GPUs with a virtual machine.

Overview of NVIDIA vGPU Platform

GPU virtualization (NVIDIA vGPU) addresses limitations of pass-through but was traditionally deployed to accelerate virtualized profession graphics applications, virtual desktop instances or remote desktop solutions. NVIDIA added support for AI, DL and high-performance computing (HPC) workloads in GRID 9.0 that was released in summer 2019. It also changed vGPU licensing to make it more amenable for compute use cases. GRID vPC/vApps and Quadro vDWS are licensed by concurrent user, either as a perpetual license or yearly subscription. Since vComputeServer is for server compute workloads, the license is tied to the GPU rather than a user and is therefore licensed per GPU as a yearly subscription. For more information about NVIDIA GRID software, see http://www.nvidia.com/grid.

Figure 2 shows the different components of the Virtual GPU software stack.

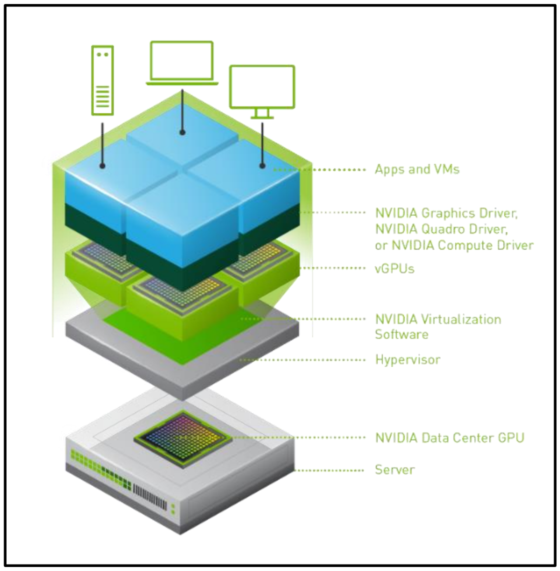
Apps and VMs
NVIDIA Graphics Driver, NVIDIA Quadro Driver, or NVIDIA Compute Driver
vGPUs
NVIDIA Virtualization Software
Hypervisor
NVIDIA Data Center GPU
Server

Figure 2. GPU enabled VM instances using GPU Pass-Though and GPU Virtualization (vGPU)

NVIDIA GPU Virtualization software transforms a physical GPU installed on a server to create virtual GPUs (vGPU) that can be shared across multiple virtual machines. The focus in this paper is on the use of GPUs for compute workloads using vComputeServer profile introduced in GRID 9. We are not looking at GPU usage for professional graphics or virtual desktop infrastructure (VDI) that will leverage Quadro vDWS or GRID vPC and vAPP profiles. GRID vPC/vApps and Quadro vDWS are client compute products for virtual graphics designed for knowledge workers and professional graphics use. vComputeServer is targeted for compute-intensive server workloads, such as AI, deep learning, and Data Science.

In an ESXi environment, the lower layers of the stack include the NVIDIA Virtual GPU Manager, that is loaded as a VMware Installation Bundle (VIB) into the vSphere ESXi hypervisor. An additional guest OS NVIDIA vGPU driver is installed within the guest operating system of your virtual machine.

Using the NVIDIA vGPU technology with vSphere provides options during creation of the VMs to dedicate a full GPU device(s) to one virtual machine or to allow partial sharing of a GPU device by more than one virtual machine.

IT admins will pick between the options depending on the application and user requirements:

- Partial GPUs: For AI dev environments a data scientist VM will not need the power of full GPU
- GPU sharing: IT admins want GPUs to be share by more than one team of users simultaneously
- High priority applications: dedicate a full GPU or multiple GPUs to one VM

The different editions of the vGPU driver are described next.

NVIDIA vComputeServer

For compute intensive server workloads such as artificial intelligence (AI), deep learning, and Data Science.

NVIDIA QUADRO vDWS

For professional graphics applications; includes an NVIDIA Quadro driver.

NVIDIA GRID vPC

For virtual desktops delivering standard PC applications, browser, and multimedia.

NVIDIA GRID vAPPS

Use with Citrix XenApp or other RDSH solutions like VMware Horizon Apps.

NVIDIA virtual GPU Software is available in four editions that deliver accelerated virtual desktops to support the needs of different workloads.

vGPU Profile	Use Case
NVIDIA vComputeServer	compute intensive workloads like AI, DL and HPC.
NVIDIA QUADRO vDWS	professional graphics applications; includes NVIDIA Quadro driver
NVIDIA GRID vPC	virtual desktops delivering standard PC applications, browser and multimedia
NVIDIA GRID vAPPS	Citrix XenApp or other RDSH solutions like VMware Horizon Apps

IT administrators can configure VMs using vComputeServer (vCS) profiles to deploy GPU compute instances on top of Dell EMC PowerEdge servers configured with NVIDIA V100 or T4 GPUs. Details of vCS GPU profile and list of Dell EMC Servers that can be used to run VMs accelerated using vCS GPU profiles is provided in the following tables. IT teams have a range of options in terms of vGPU profiles, GPU Models and supported Dell platforms to accommodate the compute requirements of their customer workloads.

USE CASE	AI, ML/DL, Data Science & HPC
Compute Type	Server Workloads
Virtual GPU Profile	vComputeServer
Recommended GPU	NVIDIA V100 or T4

	V100	V100S	T4
Architecture	Volta	Volta	Turing
CUDA Cores	5120	5120	2560
Tensor Cores	640	640	320
RT Cores	n/a	n/a	40
Memory	32GB/16GB HBM2	32GB HBM2	16GB GDDR6
vGPU Profiles	1GB, 2GB, 4GB, 8GB, 16GB, 32GB	1GB, 2GB, 4GB, 8GB, 16GB, 32GB	1GB, 2GB, 4GB, 8GB, 16GB
Form Factor	PCIe dual slot/SXM2	PCIe dual slot	PCIe single slot
Power	250W/300W	250W	70W
Supported Dell EMC Servers	R740, R740xd, T640, C4140, R940xa, R7525, DSS8440	R740, R740xd, C4140, R940xa, R7525, DSS8440	R640, R7515, R740, R740xd, C4140, R7525, DSS8440

vComputeServer Features and Deployment Patterns

vComputeServer was designed to complement existing GPU virtualization capabilities for graphics and VDI and address the needs of the data centers to virtualize compute-intensive workloads such as AI, DL and HPC. As part of addressing the needs of compute-intensive workloads, vCS introduced GPU aggregation inside a VM (multi vGPU support in a VM), GPU P2P support for NVLink, container support using NGC and support for application, VM, and host-level monitoring. A few of the key features are:

Management and monitoring: Admins can use the VMware management tools like VMware vSphere to manage GPU servers, with visibility at the host, VM and app level. GPU-enabled virtual machines can be migrated with minimal disruption or downtime.

Multi vGPU support: Administrators can now combine management benefits of vGPU and leverage the compute capability of scaling-out jobs across multiple GPUs by leveraging multi vGPU support in vComputeServer. Multiple vGPUs can now be deployed in a single virtual machine to scale application performance and speed up production workflows.

Support for NGC Software: vComputeServer supports NVIDIA NGC GPU-optimized software for deep learning, machine learning, and HPC. NGC software includes containers for the popular AI and data science software, validated and optimized by NVIDIA, as well as fully-tested containers for HPC applications and data analytics. NGC also offers pre-trained models for a variety of common AI tasks that are optimized for NVIDIA Tensor Core GPUs. This allows data scientists, developers, and researchers to reduce deployment times focus on building solutions, gathering insights, and delivering business value.

Deploying Virtualized GPU Instances for Compute Intensive Workloads

In this paper we covered the benefits of deploying virtualized VMs that can leverage GPU compute for accelerating emerging workloads like AI, Deep Learning and HPC. Customers that care about highest performance can leverage virtualized instances of NVIDIA V100 GPU in their VMs and also aggregate multiple vGPUs on Dell PE-C4140 server to get increased performance using GPU aggregation capability of vComputeServer profile. Customers concerned about cost can share a GPU between multiple users by leveraging smaller vGPU profiles (upto 16 vGPU profiles can be created from a single V100 or T4 GPU).

PowerEdge DfD Repository
For more technical learning

Contact Us
For feedback and requests

Follow Us
For PowerEdge news