Frequently Asked Questions

NVIDIA GPU with VMware Cloud Director FAQ

Table of Contents

Business FAQs

Q. What is the GPU offering with VMware Cloud Director?

A. As of VMware Cloud Director 10.3.2, Cloud Providers can now leverage vSphere support for NVIDIA GPUs and NVIDIA AI Enterprise, a cloud-native software suite for the development and deployment of AI that has been optimized and certified for VMware vSphere. When you use NVIDIA GRID cards, installed on an x86 host, in a desktop and application virtualization solution running on vSphere 6.x and later, you can render application graphics with superior performance compared to non-hardware-accelerated environments. This capability is useful for graphics-intensive use cases such as designers in a manufacturing setting, architects, engineering labs, higher education, oil and gas exploration, clinicians in a healthcare setting, and for power users and knowledge workers who need access to rich 2D and 3D graphical interfaces. This enables vSphere capabilities like vMotion from within Cloud Director to now deliver multi-tenancy GPU services, which are key to maximizing GPU resource utilization. With Cloud Director support for the NVIDIA AI Enterprise software suite, customers now have access to best-in-class, GPU-optimized AI frameworks and tools and to deliver compute-intensive workloads, including artificial intelligence (AI) or machine learning (ML) applications within their data centers.

Q: Is VMware Cloud Director providing Self-Service vGPU as a Service?

A - VMware Cloud Director Customers can self-serve, manage, and monitor their vGPU-accelerated hosts and virtual machines within Cloud Director. Administrators can use vGPU policies loaded from vCenter to determine how many fixed share resources can be allocated to each VM from the total available memory. These policies must then be published to tenants so they can consume within their org virtual data center (VDC). The placement and size settings for VMs that require vGPU are managed by vGPU policies. Cloud Director customers have access to create vGPU virtual machines using the assigned vGPU profiles provided by the Cloud Director operational administrators. Cloud Providers can combine vGPU, CPU, RAM and Storage into specific t-shirt options that tenants can consume form the Cloud Director catalog, or they can be permitted to choose their own sizing policies, a normal operation in Cloud Director.

Q: What is AI/ML, and why is it important?

A - AI, which stands for Artificial Intelligence, and ML, which stands for Machine Learning, are closely related fields in computer science that have gained significant importance in recent years.

- Artificial Intelligence (AI): AI refers to the development of computer systems or software that can perform tasks that typically require human intelligence. These tasks include problem-solving, reasoning, learning, understanding natural language, and perceiving the environment. AI systems can range from simple rule-based algorithms to complex neural networks capable of deep learning.
- Machine Learning (ML): ML is a subset of AI that focuses on the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. ML algorithms allow systems to improve their performance on a specific task through experience and without being explicitly programmed.

AI and ML are important because they empower computers to perform tasks that were once considered exclusive to humans, revolutionizing industries, improving decision-making, and enhancing our daily lives in countless ways. Their applications continue to grow, making them fundamental technologies for the future.

Q: How is NVIDIA licensed with Cloud Director in a multi-tenant mode?

A - A fully managed service means the Cloud Provider partner needs to be a signed-up on the NVIDIA Partner Network (NPN) and handle all customer GPU workloads, all must have connectivity to an NVIDIA license server. If the offering is self-service, the customer must have a valid NVIDIA agreement and licensing in place, in which case the customer needs a licensing server in their infrastructure.

Q: What is the NVIDIA AI Enterprise License server?

A - NVIDIA products are licensed via the NVIDIA License System to serve a pool of floating licenses for NVIDIA-licensed products. The NVIDIA License System is configured with licenses from the NVIDIA Licensing Portal. The NVIDIA Delegated License Service (DLS) is a component of NVIDIA License System that serves licenses to licensed clients. A DLS instance is hosted on-premises at a location that is accessible from the customers / Cloud Provider private network(s), such as inside the organization virtual data center. More information can be found here.

Q: What is NVIDIA AI Enterprise?

A - NVIDIA AI Enterprise is a software and hardware solution offered by NVIDIA to help businesses deploy and manage artificial intelligence (AI) and deep learning workloads in their data centers. It combines various NVIDIA technologies and software components to create an integrated AI infrastructure that is optimized for enterprise use cases. NVIDIA AI Enterprise license is required to allow you to use vGPU or to slice resources of GPU. Some might not want the AI software offering but still need that to be able to slice GPU, and that still requires a license.

Q: What advantages does NVIDIA AI Enterprise bring?

A – These are some of the advantages NVIDIA AI Enterprise brings:

- AI Acceleration: NVIDIA AI Enterprise includes NVIDIA GPUs (Graphics Processing Units) specifically designed for AI and deep learning tasks. These GPUs provide the computational power to train and efficiently deploy AI models.
- Software Stack: It includes a comprehensive software stack that includes NVIDIA GPU drivers, CUDA (Compute Unified Device Architecture), cuDNN (CUDA Deep Neural Network), and other libraries that are essential for AI and deep learning workloads.
- Deep Learning Frameworks: NVIDIA AI Enterprise supports popular deep learning frameworks such as TensorFlow, PyTorch, and others, making it easier for data scientists and developers to build and train AI models.
- Containerization: The solution often includes containerization support, allowing organizations to create and manage containerized AI applications, which can simplify deployment and scaling.
- Management Tools: NVIDIA provides tools and software for deploying, monitoring, and managing AI workloads in an enterprise environment. This can help IT administrators ensure the optimal performance and reliability of AI applications.
- Enterprise-Grade Support: NVIDIA typically offers enterprise-level support and services to assist organizations in setting up, maintaining, and troubleshooting their AI infrastructure.
- Data Center Integration: NVIDIA AI Enterprise is designed to work seamlessly within existing data center environments, allowing organizations to leverage their existing infrastructure investments.
- Security and Compliance: For many industries, security and compliance are critical concerns. NVIDIA AI Enterprise may provide features and configurations that help meet these requirements.
- Scalability: Organizations can scale their AI infrastructure horizontally (adding more servers or GPUs) or vertically (upgrading GPU capabilities) to meet changing AI workload demands.
- Performance: With NVIDIA's hardware and software optimizations, AI workloads can often be executed more efficiently and with higher performance compared to general-purpose hardware.
- AI Applications: NVIDIA AI Enterprise can be used for a wide range of AI applications, including image recognition, natural language processing, recommendation systems, and more, making it suitable for various industries like healthcare, finance, manufacturing, and automotive.

Q: When should tenant customers use NVIDIA AI Enterprise vs VMware Marketplace applications?

A - NVIDIA AI applications are a production enterprise-grade solution and while customers can use the VMware Marketplace (Bitnami) solutions, there is no support for these apps they are not tuned to GPU cards. They are only tested for functional security and deliver applications and operators required for GPU to work with the application. NVIDIA AI Enterprise is required for AI computing for use with c-profiles. So, for running on GPUs like A100, A30, H100, you'd need NVIDIA AI Enterprise to run AI computing in a virtualized environment. NVIDIA AI Enterprise should be considered for production solutions to gain production-level support and performance - but remember that licenses are required for vGPU or resource slicing of GPU. Passthrough or bare metal can be used without a license.

Q: What type of NVIDIA AI license would be required?

A - NVIDIA AI Enterprise for AI compute, vWS/vPC/vApps for graphics/EUC/VDI, NVIDIA AI is licensed per GPU. Customers may, if requested, bring their own license for GPU to the CSP.

Q: How can I deliver GPU-based applications to my tenants/ customers via Cloud Director?

A - Cloud Service Providers using Cloud Director can use the Application Launch Pad (ALP) or the newer Cloud Director Content Hub to deliver customer applications.

App Launchpad is primarily a user interface (UI) and management tool for Kubernetes clusters and applications and includes built-in Helm chart synchronization capabilities. Helm is a separate package manager for Kubernetes that helps you manage and deploy applications as Helm charts. Cloud Service Providers or customers (multi-tenant managed service vs self-service) will need to sync the Helm chart of the NVIDIA AI Enterprise catalog to their App Launchpad or Content Hub, and then curate applications ready for users to consume.

Metering and Chargeback FAQs

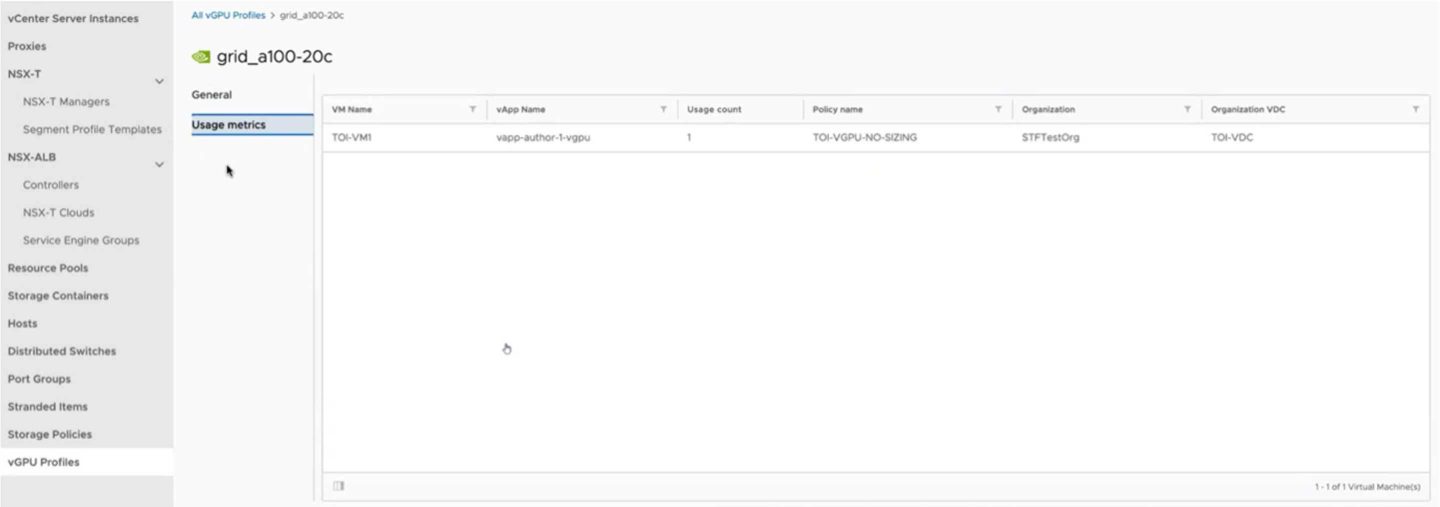
Q: What amount of memory is chargeable?

A - From a memory perspective, the memory for a VM must be 100% assigned by reservation. This is a requirement from NVIDIA and VMware vSphere. This limits the amount of overcommitment on a host as there can be no overcommit; unlike other non-GPU virtual machines, these reservations must be able to be fulfilled by the hosts; otherwise, virtual machines may not start. This will affect usage, including Usage Metering when applying vMotion restrictions to the target host GPU. This is why we advise utilizing a dedicated cluster for GPU workloads.

Q: What monitoring is available for vGPU in VMware Cloud Director?

A - Cloud Providers can monitor (through vCloud API and UI dashboard) NVIDIA vGPU allocation, usage per VDC and per VM to optimize utilization and meter/bill (through vCloud API) NVIDIA vGPU usage is averaged over a unit of time per tenant for tenant billing.

The UI will show all vGPU policies in every vGPU Organization Virtual Data Center.



VM Name	vApp Name	Usage count	Policy name	Organization	Organization VDC
TOI-VM1	vapp-author-1-vgpu	1	TOI-VGPU-NO-SIZING	STFTestOrg	TOI-VDC

1 - 1 of 1 Virtual Machine(s)

GET:
https://xyz/cloudapi/1.0.0/vgpuProfiles/consumers?page=1&pageSize=15&filterEncoded=true&filter=vgpuProfileName==grid_a100-10c&links=true

The auditTrails API can be used to obtain the same content as a VM power-on event captures **vm.vmPlacementPolicy.id**, which points to a vGPU policy in use.

Q: How does VMware Usage Meter work with GPU-enabled workloads?

A - Usage Meter charges for memory used; when a host is enabled for GPU, the entire amount of memory will be chargeable for the GPU card. For this reason, VMware recommends a separate cluster for customers needing GPU so the billing can be segregated to single customers. For more details, see the PUG.

Q: Is chargeback available for vGPU services?

A - Not currently (Sept 2023), it is on the roadmap for Chargeback capabilities.

Technical FAQs

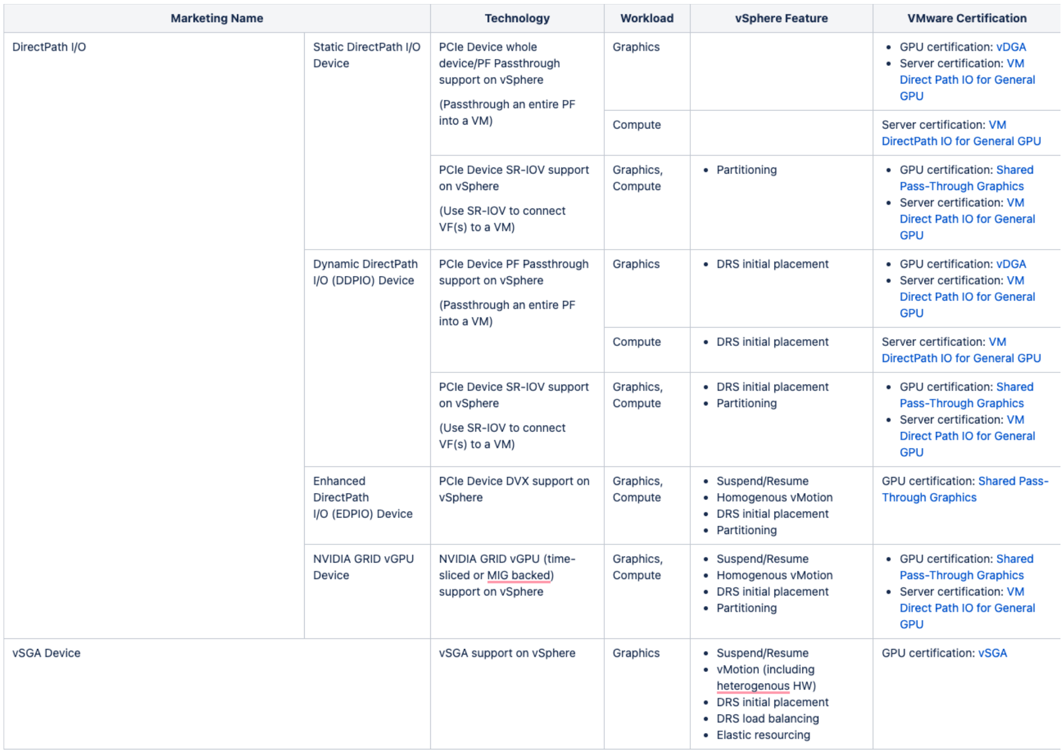
Q: What are the prerequisites for vCenter?

A - For GPU profiles to be visible in Cloud Director, at least one vCenter Server host must have an NVIDIA GPU installed, and all required vSphere Installation Bundles (VIBs) must be installed on the host.

Q: Which NVIDIA GRID vGPU are supported?

A - There are two ways for vSphere customers to use NVIDIA GPU. One is through NVAIE, which entitles the use of NVIDIA GRID vGPU, and the other is through DDIPO device passthrough. There are two types of workloads: compute and graphics. NVIDIA GRID vGPU is what allows the hypervisor to share a single GPU across multiple VMs, where the device passthrough will only allow us to pass that GPU device to a single VM.

Please see the following table:



Marketing Name		Technology	Workload	vSphere Feature	VMware Certification
DirectPath I/O	Static DirectPath I/O Device	PCIe Device whole device/PF Passthrough support on vSphere (Passthrough an entire PF into a VM)	Graphics		• GPU certification: vDGA • Server certification: VM Direct Path IO for General GPU
			Compute		Server certification: VM DirectPath IO for General GPU
		PCIe Device SR-IOV support on vSphere (Use SR-IOV to connect VF(s) to a VM)	Graphics, Compute	• Partitioning	• GPU certification: Shared Pass-Through Graphics • Server certification: VM Direct Path IO for General GPU
	Dynamic DirectPath I/O (DDPIO) Device	PCIe Device PF Passthrough support on vSphere (Passthrough an entire PF into a VM)	Graphics	• DRS initial placement	• GPU certification: vDGA • Server certification: VM Direct Path IO for General GPU
			Compute	• DRS initial placement	Server certification: VM DirectPath IO for General GPU
		PCIe Device SR-IOV support on vSphere (Use SR-IOV to connect VF(s) to a VM)	Graphics, Compute	• DRS initial placement • Partitioning	• GPU certification: Shared Pass-Through Graphics • Server certification: VM Direct Path IO for General GPU
	Enhanced DirectPath I/O (EDPIO) Device	PCIe Device DVX support on vSphere	Graphics, Compute	• Suspend/Resume • Homogenous vMotion • DRS initial placement • Partitioning	GPU certification: Shared Pass-Through Graphics
	NVIDIA GRID vGPU Device	NVIDIA GRID vGPU (time-sliced or MIG backed) support on vSphere	Graphics, Compute	• Suspend/Resume • Homogenous vMotion • DRS initial placement • Partitioning	• GPU certification: Shared Pass-Through Graphics • Server certification: VM Direct Path IO for General GPU
vSGA Device		vSGA support on vSphere	Graphics	• Suspend/Resume • vMotion (including heterogenous HW) • DRS initial placement • DRS load balancing • Elastic resourcing	GPU certification: vSGA

DirectPath I/O (or Passthrough internally) encompasses a family of technologies from VMware for providing virtual machines with direct, high-performance access to available I/O and accelerator devices. This includes supporting technologies such as SR-IOV, DVX and NVIDIA's GRID stack.

Unlike the vSGA vMotion, NVIDIA GRID vGPU vMotion and DVX vMotion have some limitations; hence they are described as "Homogenous vMotion".

- NVIDIA GRID vGPU vMotion supports vMotion between similar GPUs or identical GPUs, so-called like-hardware-vmotion or identical-hardware-vmotion.
- Depending on the device vendor's discretion and implementation, a DVX device might or might not support vMotion. DVX vMotion can support vMotion between similar GPUs or identical GPUs. It's theoretically possible for a vendor to allow cross-generation vMotion.

VMware Certification encompasses:

1. **Virtual Dedicated Graphics Acceleration [vDGA]:** (Server + GPU combination compatible with ESX/vSphere and Horizon) https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vdga
2. **VM DirectPath I/O for General GPU [DPIO GP GPU]:** (Server + GPU combination compatible with ESX/vSphere) https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vmdirect
3. **Shared Passthrough Graphics:** (GPU + driver combination compatible with ESx/vSphere and Horizon) (GPU + driver combination compatible with ESx/vSphere and Compute) https://www.vmware.com/resources/compatibility/search.php?deviceCategory=sptg
4. **Virtual Shared Graphics Acceleration [vSGA]:** (GPU + driver combination compatible with ESx/vSphere and Horizon) https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsga

DDPIO passthrough and vGPU technologies can support both Graphics and Compute workloads. It depends on the capability of the GPU. Some GPUs can only support compute workloads or graphics workloads, some GPUs can support both. Most of the enterprise GPUs that are supported by vSphere should also be supported by VCD. The NVIDIA GPUs supported by NVIDIA AI Enterprise (AI/compute) can be found here and for vGPU (graphics) here. It's at the system level where compatibility should be checked within the VMware HCL, and the NVIDIA Certified Systems list. Please ensure you validate this.

Q: What are vGPU Profiles, and how are vGPU Profiles created in the vCenter Server?

A - vGPU Profiles allow you to assign a GPU solely to one Virtual Machine's use or to be used in a shared mode with others. By choosing the vGPU profile that assigns the full memory of the GPU to a VM, we can dedicate that GPU in full to that VM.

Basically, vGPU profiles act as a structure for distributing fixed resource shares to virtual machines (VMs) from the overall pool of available memory, tailored to each user's requirements. Given GPUs' diverse features and resources, they also accommodate various types of profile support.

The process of profile creation is managed internally by the vCenter Server. Once hosts have the necessary NVIDIA VIBs installed, they will automatically communicate the vGPU profiles they support, and vCenter will subsequently report this information. Therefore, the key factor in this process is ensuring hosts have GPU cards and installing NVIDIA VIBs.

Q: Which methods can be used in vCenter Server to Share the GPU card?

A - There are two methods by which you can share the GPU card:

1. Shared – VMware shared virtual graphics.
2. Shared Direct – Vendor shared passthrough graphics.

Q: How many vGPU are supported per virtual machine?

A - Up to 16 vGPUs/VM are now supported.

Q: Which series of vGPU are supported for multiple vGPUs per VM?

A - As of VCD 10.5, only Q and C series vGPUs that are allocated with all of the physical GPU's frame buffer are supported for multiple vGPUs per virtual machine.

Q: Is Multi-Instance GPU (MIG) supported by VMware Cloud Director?

A - NVIDIA MIG (Multi-instance GPU) supports spatial workload segmentation between workloads at the physical level inside a single device, driving better hardware optimization and increased margins. Cloud Director supports MIG and relies on host pre-configuration for GPU services included in NVIDIA AI Enterprise and MIG which contains vGPU technology to enable deployment/configuration on hosts and GPU profiles. Cloud Director supports MIG by way of vSphere supporting this feature on GPUs such as NVIDIA A100, A30, H100. MIG is not required. It's a unique feature that's available to those who want to do spatial partition of the GPU, which may be suitable for running multiple small workloads (like inference) concurrently. GPU partitioning can also be accomplished with vGPU, it accomplishes this in a different way that's time-sliced vs. hardware based.

Q: Is vMotion supported with GPU enabled Virtual Machines?

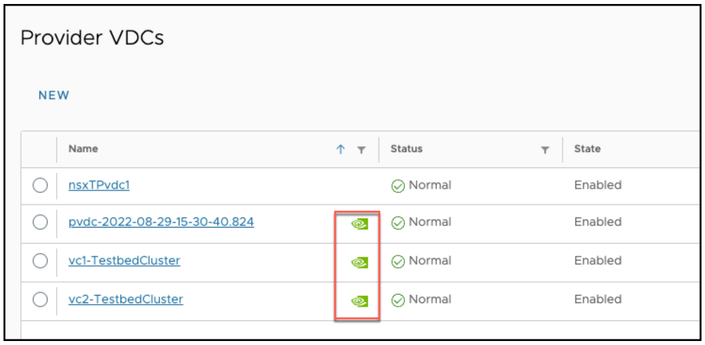
A - Yes, as long as the source and destination server have identical GPUs and matching drivers, it would be best practice to create a dedicated vGPU cluster to provide resilient services.

Q: Where can you access vGPU profiles within Cloud Director?

A - When the hosts within a cluster supporting the provider VDC resource pool possess vGPU capabilities, the VMware Cloud Director retrieves the vGPU profile details from the vCenter Server. Within the Infrastructure Resources section, you'll find the vGPU Profiles sourced by VMware Cloud Director from vCenter Server clusters equipped with virtual graphics processing unit (vGPU) capabilities. Each profile corresponds to a specific vGPU type, and these profiles serve as the basis for creating vGPU policies accessible to tenants for their virtual machines. NVIDIA vGPU profiles define the allocation of fixed resource shares for each VM based on the total available memory. Providers can review or modify vGPU profile information.

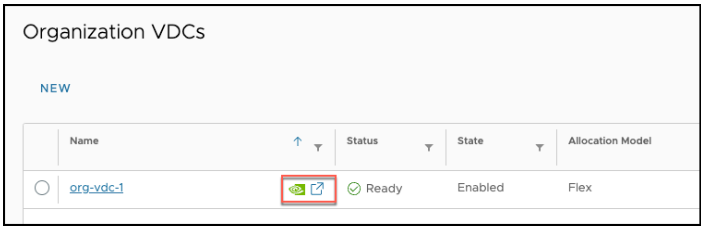
Q: How can I determine if the Provider VDC or the Organization VDC is backed by GPU?

A - There will be an NVIDIA icon/logo listed next to each of the Org VDC names, which indicates that the virtual GPU utilization policy has been associated with these OVDCs.



Provider VDCs

	Name	Status	State
○	nsxTPvdc1	✓ Normal	Enabled
○	pvdc-2022-08-29-15-30-40.824	✓ Normal	Enabled
○	vc1-TestbedCluster	✓ Normal	Enabled
○	vc2-TestbedCluster	✓ Normal	Enabled



Organization VDCs

	Name	Status	State	Allocation Model
○	org-vdc-1	✓ Ready	Enabled	Flex

Q: Which VDC Allocation Models are compatible with the vGPU Policies?

A - Flex, Reservation Pool, Allocation Pool, Pay-As-You-Go.

Q: What steps are involved in configuring a vGPU policy in Cloud Director?

A - You select a vGPU profile while creating a vGPU policy. When you publish a vGPU policy, the vGPU profile names and instructions of the profiles you added to the policy become visible to the tenants. See Creating and Managing vGPU Policies documentation.

You can establish and oversee vGPU policies either globally or at the provider level. Subsequently, you have the option to make individual policies available to one or more organization VDCs.

Each vGPU policy in Cloud Director can only be linked to one vGPU profile.

Q: How do you publish / add vGPU policy to a tenant organization?

A - When you generate a vGPU policy, it remains hidden from tenants' view. To make a vGPU policy accessible to tenants, you must publish it to an organization VDC. Publishing a vGPU policy to an organization VDC renders it visible to tenants. Tenants can then choose this policy when creating a new standalone VM or a VM from a template, making edits to a VM, adding a VM to a vApp, or creating a vApp from a vApp template.

Navigate to the top navigation bar and choose "Resources," then click on "Cloud Resources." In the left-hand panel, locate and select "Organization VDCs." Choose an organization VDC, and within the "Policies" section, go to the "vGPU" tab. Next, click on the "Add" button. From the available options, pick the vGPU policies you wish to apply to the organization VDC and confirm your selection by clicking "OK.

Q: GPU operators for Kubernetes?

A - To enable developers to deploy AI/ML workloads on TKG clusters, as a Cluster Operator, you configure the Kubernetes environment to support NVIDIA vGPU operations. The NVIDIA AI Enterprise edition of the GPU Operator is pre-configured and optimized for use with Tanzu Kubernetes Grid. The NVAIE GPU Operator differs from the GPU Operator that is available in the public NGC catalog; this makes up a part of the cluster pre-configuration required. See NVIDIA AI Enterprise for more information. To install a GPU ready cluster the following needs to be done:

- The Kubernetes (k8s) nodes, which are essentially virtual machines (VMs), should be equipped with vGPU capability. It's not mandatory for every node in the cluster to have GPU support from the outset. When a worker node-pool is introduced, either during the initial creation of the cluster or as part of a cluster update, there is an option to activate GPU functionality. If this option is turned on, it becomes necessary to select the appropriate GPU placement policy. Only nodes within this specific node-pool will adhere to the GPU placement policy and consequently have access to GPU resources.
- The installation of the NVIDIA GPU operator should be performed within the k8s cluster itself, rather than in individual k8s node VMs. This operator encompasses the NVIDIA GPU device plugin designed for k8s. It is capable of identifying the vGPU resources available in each k8s node (VM) and relaying this information to the k8s server. Once the GPU operator is successfully installed, the k8s cluster will have the capability to execute container workloads that leverage vGPU functionality.

Please be aware that it is also possible to install the GPU Operator on the k8s node in advance of the aforementioned operation.

Q: GPU operators for virtual machines?

A - vSphere incorporates a VM operator responsible for VM creation and possibly an operator that handles the configuration of NVIDIA drivers within the VMs. In contrast, VCD lacks both a VM operator and a GPU operator designed for VM management.

Q: What exactly does a GPU operator do, and why is it necessary?

A - Traditionally, when attempting to utilize an NVIDIA card within a virtual machine (VM), the process involves the following steps:

- Identifying the precise GPU version available.
- Obtaining the correct driver directly from NVIDIA.
- Installing the driver and running nvidia-smi within the VM in the hope that the GPU functions correctly.

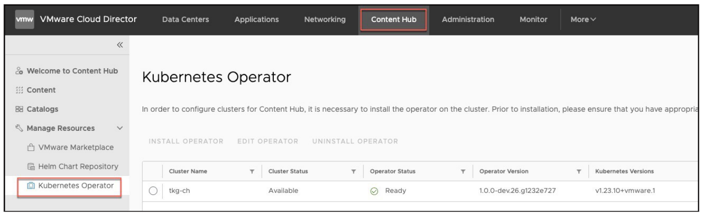
This method was susceptible to errors and complications, as GPU drivers were highly specific to each GPU card. Customers often encountered frustration while trying to locate the appropriate drivers, and the configuration process was entirely manual.

For Kubernetes clusters, the situation was even more challenging. While it was possible to use templates containing the correct drivers, a separate template was required for each GPU card. Managing a Kubernetes cluster with various GPU cards became an unwieldy and burdensome task, leading to significant challenges for users between 2018 and 2020.

To address these difficulties in Kubernetes clusters, NVIDIA initiated the development of a GPU operator. This operator serves the purpose of identifying the correct GPU card within the VM and automating the download and installation of the necessary drivers. Consequently, users can now significantly reduce their concerns about maintaining a variety of GPU drivers on their host systems.

Q: How do you install Kubernetes Operator from the VMware Cloud Director Portal?

A - As an organizational user, you have the option to install the Kubernetes operator using either the Content Hub section (available from VCD 10.5 onwards) or through the App Launchpad plug-in. To install the operator on your Kubernetes cluster, simply choose the specific Kubernetes cluster you wish to target and then click on the ‘Kubernetes Operator’ option, which is accessible through the Content Hub feature.



Kubernetes Operator

In order to configure clusters for Content Hub, it is necessary to install the operator on the cluster. Prior to installation, please ensure that you have appropria

	Cluster Name	Cluster Status	Operator Status	Operator Version	Kubernetes Versions
○	tkg-ch	Available	✓ Ready	1.0.0-dev.26.g1232e727	v1.23.10+vmware.1

Q: How will the end user use the vGPU policy on the Virtual Machine?

A - When creating a new Virtual Machine, the user can select a vGPU Policy instead of a General Purpose policy under the compute section. After creating the Virtual Machine, the user can verify the vGPU settings under the hardware, compute section. Once the Virtual Machine is created, the GPU hardware must be detected and seen inside the guest operating system.

Item No: NVIDIA GPU with Cloud Director FAQ_v1.0 9/23