



NVIDIA Multi-Instance GPU and NVIDIA Virtual Compute Server

GPU Partitioning

Technical Brief

Table of Contents

Solution Overview	1
GPU Partitioning.....	2
NVIDIA vCS Virtual GPU Types	4
MIG Backed Virtual GPU Types	5
Managing MIG – GPU Instances.....	6
NVIDIA Virtual Compute Server with MIG Mode Enabled	7
NVIDIA Virtual Compute Server with MIG Mode Disabled.....	9
Compute Workflows	10
Single User: Multiple Apps.....	10
Single Tenant: Multiple Users	11
Multiple Tenant: Multiple Users.....	12
Summary	13
Resources Links.....	14
NVIDIA GRID Resources.....	14
NVIDIA Virtual Compute Server Resources	14
NVIDIA Multi-Instance GPU Resources.....	14
Other Resources.....	14

Solution Overview

[NVIDIA A100 Tensor Core GPU](#) is based upon the NVIDIA Ampere architecture and accelerates compute workloads such as artificial intelligence (AI), data analytics, and HPC in the data center. The NVIDIA vGPU software that enables data centers to virtualize the NVIDIA A100 graphics processing unit (GPU) is the [NVIDIA Virtual Compute Server \(vCS\)](#) product. This NVIDIA vGPU solution extends the power of the NVIDIA A100 GPU to users allowing them to run any compute-intensive workload in a virtual machine (VM). NVIDIA vGPU 11.1 or later software releases offers support for Multi-Instance GPU (MIG) backed virtual GPUs and users have the flexibility to use the NVIDIA A100 in MIG mode or non-MIG mode. Combining MIG with vCS, enterprises can take advantage of management, monitoring and operational benefits of hypervisor-based server virtualization, running a VM on each MIG partition and Linux distribution.

GPU Partitioning

GPU partitioning is particularly beneficial for workloads which do not fully saturate the GPU's compute capacity. A lot of GPU workloads do not require a full GPU. For example, if you are giving a demo, you are building POC code or are testing out a smaller model, you do not need 40 GB of GPU memory which is offered by the NVIDIA A100 Tensor Core GPU. Without GPU partitioning, a user doing this type of work would have an entire GPU allocated, whether they are using it or not. Compute workloads which use Kubernetes clusters can benefit from GPU partitioning as well as multi-tenant use cases where one client cannot impact the work or scheduling of other clients, providing isolation for customers.

While NVIDIA vGPU software implemented shared access to the NVIDIA GPU's for quite some time, the new Multi-Instance GPU (MIG) feature allows the NVIDIA A100 GPU to be spatially partitioned into separate GPU instances for multiple users as well. The goal of this technical brief is to understand the similarities as well as differences between NVIDIA A100 MIG capabilities and NVIDIA vGPU software, while also highlighting the additional flexibility when they are combined.

The following table summarizes the concurrency mechanisms points which will be discussed.

Table 1. Concurrency Mechanisms Points

	NVIDIA A100 MIG Backed Virtual GPU Types	NVIDIA A100 with NVIDIA vCS Virtual GPU Types
GPU Partitioning	Spatial (hardware)	Temporal (software)
Number of Partitions	7	10
Compute Resources	Dedicated	Shared
Compute Instance Partitioning	Yes	No
Address Space Isolation	Yes	Yes
Fault Tolerance	Yes (highest quality)	Yes
Low Latency Response	Yes (highest quality)	Yes
NVIDIA® NVLink® Support	No	Yes
Multi-Tenant	Yes	Yes
NVIDIA® GPUDirect® RDMA	Yes (GPU instances)	Yes
Heterogenous Profiles	Yes	No
Management - requires Super User	Yes	No

NVIDIA vCS Virtual GPU Types

NVIDIA vGPU software uses temporal partitioning and has full IOMMU protection for the virtual machines that are configured with vGPUs. Virtual GPU provides access to shared resources and the execution engines of the GPU: Graphics/Compute, Copy Engines. A GPU hardware scheduler is used when VMs share GPU resources. This scheduler uses time slicing to impose limits on GPU processing cycles used by a vGPU and automatically dequeues work from channels onto the GPU's engines. If vGPUs are added or removed, the share of GPU processing cycles allocated can change accordingly (dependent of scheduling policy), resulting in performance to increase when utilization is low, and decrease when utilization is high. This type of scheduling dynamically harvests empty GPU cycles and allows for efficient use of GPU resources.

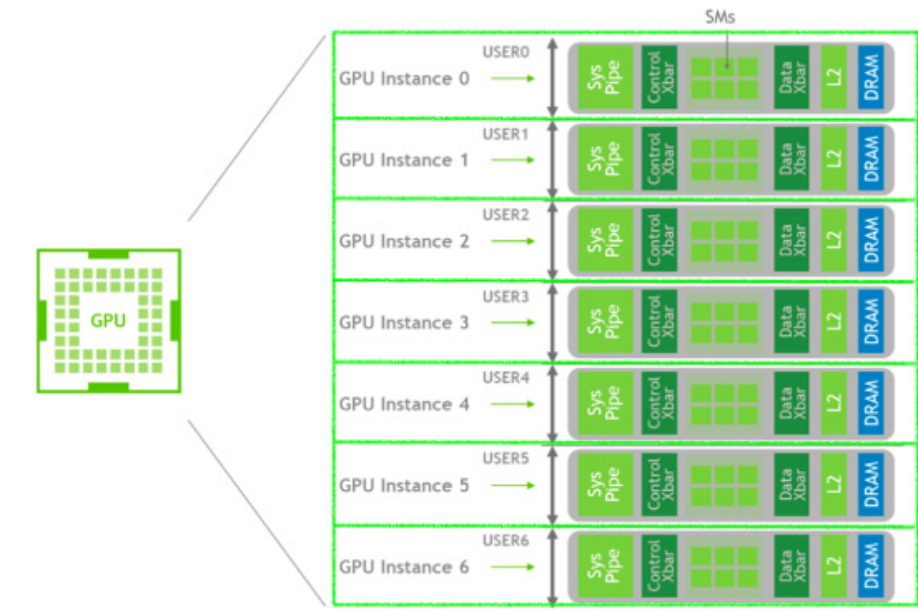
NVIDIA vGPU software, which uses temporal partitioning, can partition a NVIDIA A100 up to 10 vGPUs, thereby 10 VM's can access this shared resource (40 GB of GPU memory) with 4 GB GPU memory allocated per VM. A vGPU is assigned to VM's using vGPU profiles.

To enable vGPU support on a virtual machine, a shared PCIe device is added to the VM. Once this device is added, vGPU profiles are assigned using a centralized management utility, like VMware vSphere or Red Hat RHV/RHEL, which is provided by the hypervisor. Root privileges are not required for enabling vGPU support on a virtual machine as long as the named user is part of the administrator role.

MIG Backed Virtual GPU Types

The NVIDIA A100 is the first NVIDIA GPU to offer MIG. MIG enables multiple GPU instances to run in parallel on a single, physical NVIDIA A100 GPU. MIG mode spatially partitions the hardware of GPU so that each MIG can be fully isolated with its own streaming multiprocessors (SM's), high-bandwidth, and memory. MIG can partition available GPU compute resources as well.

Figure 1. MIG Enabled Multi-GPU Instances



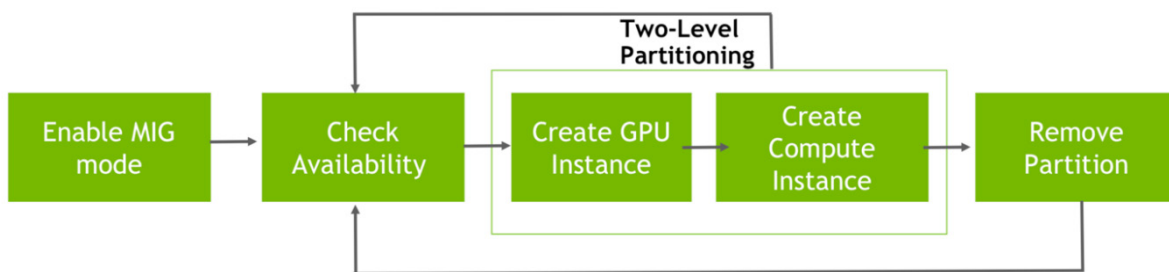
With MIG, each instance's processors have separate and isolated paths through the entire memory system - the on-chip crossbar ports, L2 cache banks, memory controllers, and DRAM address busses are all assigned uniquely to an individual instance. This ensures fault tolerance and an individual user's workload can run with predictable throughput and latency, with the same L2 cache allocation and DRAM bandwidth, even if other tasks are thrashing their own caches or saturating their DRAM interfaces.

A single NVIDIA A100 has up to 7 usable GPU memory slices, each with 5 GB of memory. MIG is configured (or reconfigured) using `nvidia-smi` and has instance profiles that can be chosen to meet the needs of HPC, Deep Learning, or Accelerated Computing workloads.

Managing MIG – GPU Instances

The workflow for managing MIG is executed using NVML/`nvidia-smi` commands. Creating a GPU instance requires `CAP_SYS_ADMIN` or root privileges. The following graphic illustrates the workflow.

Figure 2. Managing MIG Workflow



MIG instances can be created and destroyed dynamically and does not affect other GPU instances. However, if a portion of the GPU is not being used, the empty GPU processing cycles are not allocated to the actively used partition. Therefore, MIG does not have the flexibility to dynamically harvest empty GPU cycles. The following table illustrates the GPU instance sizes which are available to MIG and well as the number of instances which can be created.

Table 2. GPU Instance Sizes Available to MIG

GPU Instance Size	Number of Instances Available	SMs per GPU Instance	Memory
1g.5gb	7	14	5 GB
2g.10gb	3	28	10 GB
3g.20gb	2	42	20 GB
4g.20gb	1	56	20 GB
7g.40gb	1	98	40 GB

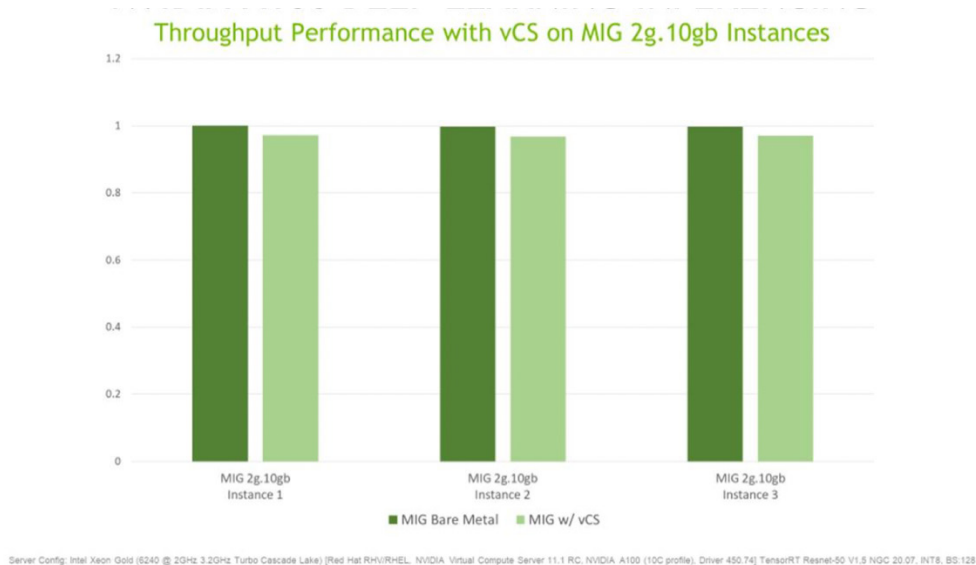
A single GPU compute instance resides within the GPU instance. However, more than one compute instance can be created and provides partial isolation to compute resources but allows independent workload scheduling.

NVIDIA Virtual Compute Server with MIG Mode Enabled

Combining NVIDIA vCS and NVIDIA A100 MIG enables additional flexibility on the Ampere architecture. This includes provisioning and orchestration benefits where end-to-end management tools are available providing real-time insights.

Use cases which require high quality of service with low latency response and error isolation are key workloads for MIG spatial partitioning. Since MIG offers separate and isolated paths through the entire memory system, MIG ensures that an individual user's workload can run with predictable throughput and latency. The extreme throughput and latency of MIG surpass vGPU temporal partitioning. The following graph illustrates an example of inferencing throughput differences between bare metal MIG and vCS using MIG backed virtual GPU's (mileage may vary according to dataset and workflows).

Figure 3. NVIDIA A100 Deep Learning Inferencing




NVIDIA vGPU software supports MIG GPU instances only with NVIDIA Virtual Compute Server and Linux guest operating systems. To support GPU instances with NVIDIA vGPU, a GPU must be configured with MIG mode enabled and GPU instances must be created and configured on the physical GPU. For more information, refer to the *vCS Deployment Guide for Red Hat RHEL*. For general information about the MIG feature, see the [NVIDIA Multi-Instance GPU User Guide](#).

One of the new features introduced to vGPU when VM's are using MIG backed virtual GPU's is the ability to have different sized (heterogenous) partitioned GPU instances. The following figure illustrates the 18 possible size combinations when NVIDIA A100 has MIG mode enabled.

Figure 4. NVIDIA A100 MIG Mode Enabled Possible Combinations

Slice #1	Slice #2	Slice #3	Slice #4	Slice #5	Slice #6	Slice #7
7						
4				2		1
4				1	1	1
2		2		3		
2		1	1	3		
1	1	2		3		
1	1	1	1	3		
3				3		
3				2		1
3				1	1	1
2		2		2		1
2		2		1	1	1
1	1	2		2		1
1	1	2		1	1	1
2		1	1	2		1
2		1	1	1	1	1
1	1	1	1	2		1
1	1	1	1	1	1	1

 **Note:** When using vCS and MIG mode is enabled, the vGPU software recognizes the MIG backed vGPU resource as if it were 1:1 or full GPU profile.

Not all hypervisors support GPU instances in NVIDIA vGPU deployments. To determine if your chosen hypervisor supports GPU instances in NVIDIA vGPU deployments, consult the release notes for your hypervisor at [NVIDIA Virtual GPU Software Documentation](#).

NVIDIA Virtual Compute Server with MIG Mode Disabled

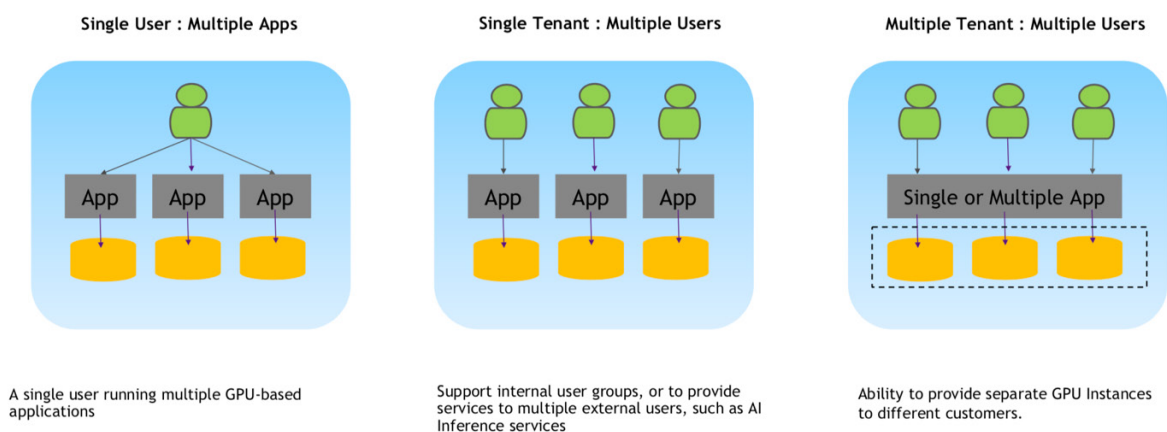
When the NVIDIA A100 is in non-MIG mode, NVIDIA vCS software uses vGPU temporal partitioning where VM's have shared access to compute resources which can be beneficial to certain workloads. Dynamic scheduling harvests empty GPU cycles and allows for efficient use of GPU resources during idle or less demand. During these times there is higher throughput potential for compute operations. However, during peak demand users may see a performance impact due to context switching. Since vCS offers up to 10 GPU partitions (MIG offers 7) and can harvest empty GPU cycles, a better total cost of ownership (TCO) can be achieved for certain workloads.

NVIDIA vCS, with MIG mode disabled, also offers access to non-compute engines (like NVENC, NVDEC, JPEF and OFA) when VM's are using vGPU fractional profiles. VM's do not have access to the full set of non-compute engines when MIG mode is enabled unless the NVIDIA A100 GPU is configured for 7 slice partitions. Peer-to-Peer NVIDIA® CUDA® transfers over NVLink are supported by vCS; this support is not offered to NVIDIA A100 when MIG is enabled.

Compute Workflows

Compute workloads can benefit from using separate GPU partitions where each GPU partition are isolated and protected. The flexibility of GPU partitioning allows a single GPU to be used by small, medium, and large sized workloads. The following graph illustrates use cases where a single user is running multiple applications, as well as single and multi-tenant workflows on a single NVIDIA GPU.

Figure 5. Compute Workflows

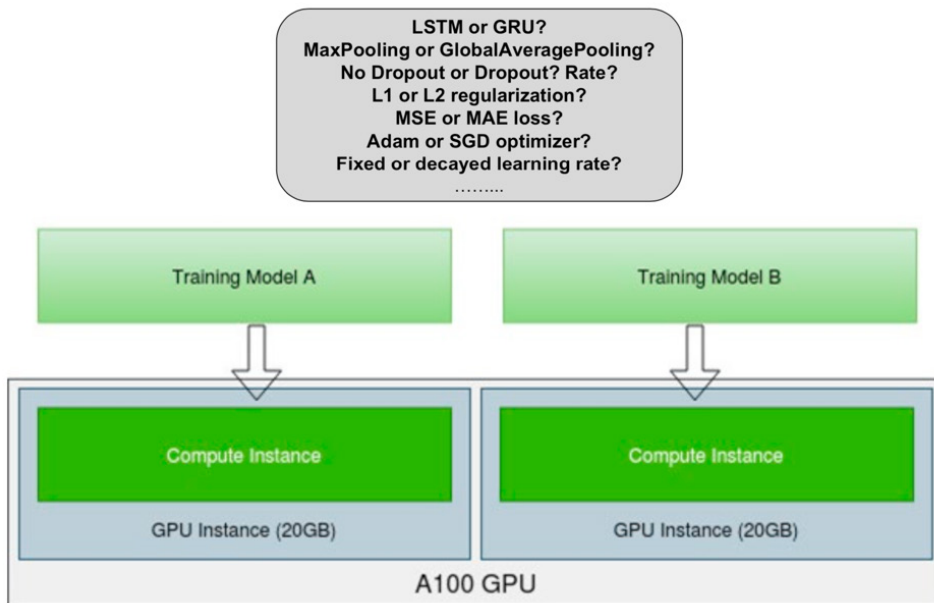


Single User: Multiple Apps

This use case improves the GPU utilization for smaller to medium sized workloads which underutilize the GPU. An example are Deep Learning training and inferencing workflows which utilize smaller datasets.

GPU Partitioning offers an efficient way to try different hyperparameters but is highly dependent on the size of the data/model, users may need to decrease batch sizes. The following graph illustrates training 2 models with different hyperparameters on two GPU partitions simultaneously.

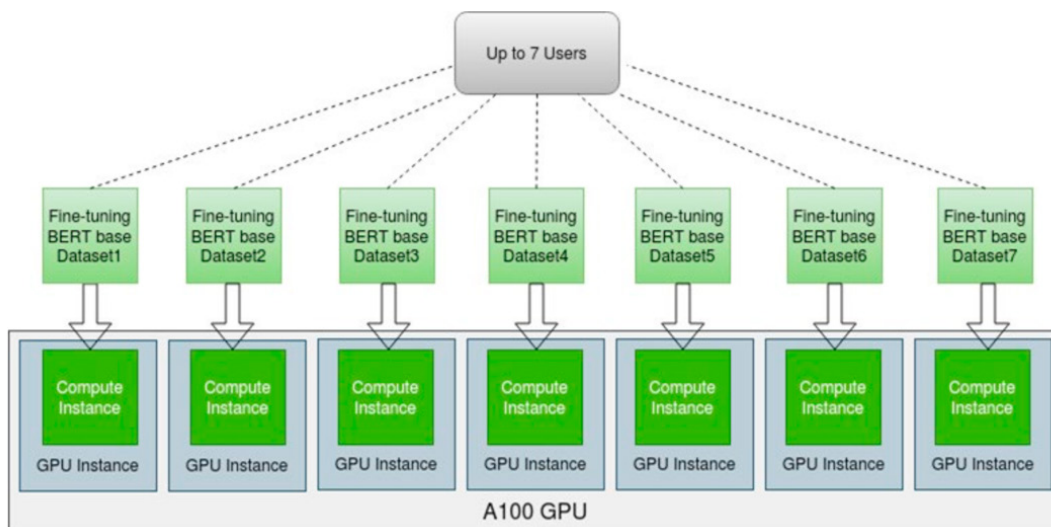
Figure 6. Training Models with Different Hyperparameters



Single Tenant: Multiple Users

In this use case a single NVIDIA A100 is enabled for MIG and it is serving multiple users for fine tuning 7 BERT base Pytorch models on 7 MIG instances for different datasets. In this example, MIG was enabled, 7 GPU instances were created, and each had its own compute instance.

Figure 7. Single NVIDIA A100 Enabled for MIG

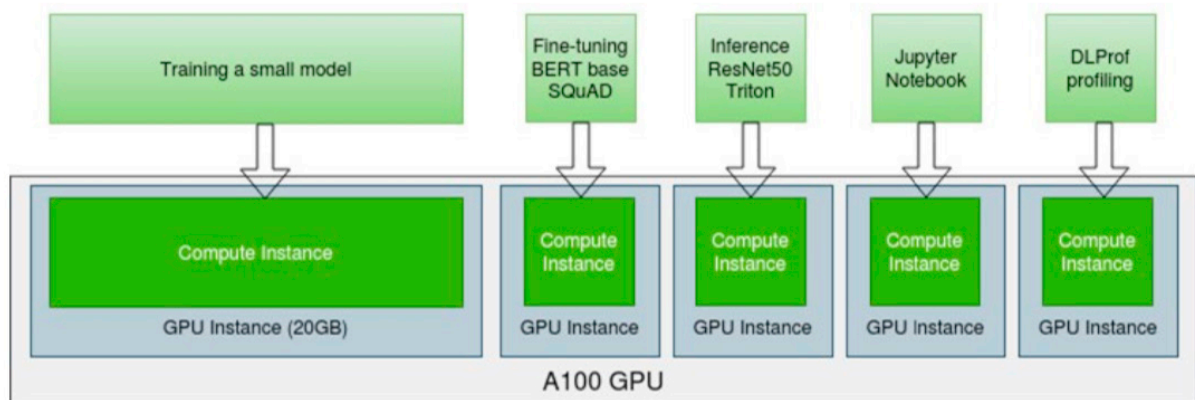


One NVIDIA A100 can also serve multiple users using different frameworks, models and/or datasets.

Multiple Tenant: Multiple Users

In this use case a single NVIDIA A100 can be used for multiple workloads such as Deep Learning training, fine-tuning, inference, Jupyter Notebook, profiling, debugging, etc. The following graph illustrates these multiple workload use cases.

Figure 8. Single NVIDIA A100 used for Multiple Workloads



Summary

NVIDIA A100 in virtualized environments using NVIDIA vCS enables additional flexibility on the Ampere architecture. NVIDIA vGPU software offers support for Multi-Instance GPU (MIG) backed vGPUs but users can choose to use the NVIDIA A100 in MIG mode or non-MIG mode. When the NVIDIA A100 in non-MIG mode, NVIDIA vCS provides additional software features as well shared access to compute resources where dynamic scheduling can harvest empty GPU cycles, resulting in higher throughput potential and better TOC per user. Use cases which require highest quality of service with low latency response and error isolation are key workloads for enabling MIG spatial partitioning. Combining MIG with vCS, enterprises can run a VM on each MIG partition while also taking advantage of provisioning and orchestration benefits as well as end-to-end management tools providing real-time insights.

Resources Links

NVIDIA GRID Resources

[Quantifying the Impact of NVIDIA Virtual GPUs](#)

[NVIDIA GRID Solution Overview](#)

[NVIDIA GRID webpage](#)

NVIDIA Virtual Compute Server Resources

[NVIDIA Virtual Compute Server webpage](#)

[NVIDIA Virtual Compute Server Solution Overview](#)

[Webinar: Introducing the Modern Data Center Powered by NVIDIA Virtual Compute Server](#)

NVIDIA Multi-Instance GPU Resources

[NVIDIA Multi-Instance GPU User Guide](#)

[Running CUDA Applications as Containers](#)

[Schedule Kubernetes pods on MIG instances](#)

Other Resources

[Try NVIDIA vGPU for free](#)

[Using NVIDIA Virtual GPUs to Power Mixed Workloads](#)

[NVIDIA Virtual GPU Software Documentation](#)

[NVIDIA vGPU Certified Servers](#)

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, CUDA, GPUDirect, NVIDIA GRID, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2020 NVIDIA Corporation. All rights reserved.