



GPU VIRTUALIZATION TECHNICAL OVERVIEW

NVIDIA Multi-Instance GPU on NexaVM Cloud

How one data-center GPU is divided into hardware-isolated instances, and how NexaVM Cloud delivers those instances to virtual machines for AI and high-performance computing workloads.

NexaVM Cloud Platform Documentation

2026

Example: One NVIDIA A100 80GB Divided into Four Isolated Instances

Memory slices (8×10 GB):

M0	M1	M2	M3	M4	M5	M6	M7
10 GB	10 GB	10 GB	10 GB	10 GB	10 GB	10 GB	10 GB

GPU instances (● = one compute slice):

3g.40gb — VM-A	2g.20gb — VM-B	1g.10gb — VM-C	1g.10gb — VM-D
● ● ● (3 compute slices)	● ● (2 compute slices)	● (1 compute slice)	● (1 compute slice)

3 + 2 + 1 + 1 = all 7 compute slices in use, each instance assigned to its own VM.

What MIG Is

Multi-Instance GPU (MIG) is a hardware partitioning feature built into NVIDIA data-center GPUs from the Ampere generation onward (compute capability 8.0 and above). Instead of letting several workloads take turns on the whole GPU, MIG divides the GPU into fixed partitions. Each partition receives its own dedicated share of GPU memory and its own compute cores, so workloads in different partitions do not compete for the same resources.

MIG partitioning is static: compute cores and memory are divided in advance into partitions that are independent at the hardware level. The layout is planned up front and can be changed only when the affected instances are idle.

MIG is designed for high-performance computing and AI model training and inference, where predictable throughput and strong isolation matter most.

How a GPU Is Partitioned

NVIDIA describes MIG with a small set of building blocks. Once these are clear, profile names such as 3g.40gb are easy to read.

Term	Meaning
Streaming multiprocessor (SM)	The GPU cores that execute compute instructions.
Memory slice	About one-eighth of the GPU's memory, including its share of memory bandwidth and cache.
Compute slice	About one-seventh of the GPU's SMs when MIG mode is enabled.
GPU Instance (GI)	A combination of memory slices, compute slices and hardware engines (copy engines, video decoders). This is the unit that provides memory isolation and quality of service.
Compute Instance (CI)	An optional subdivision inside a GI. CIs have their own SMs but share the parent GI's memory and engines. VM deployments typically use one CI covering the whole GI.

Reading a Profile Name

3g.40gb means 3 compute slices and 40 GB of memory. Memory and compute slices are not always equal: on an A100 80GB, a 3g.40gb instance uses 4 memory slices but 3 compute slices, as shown in the diagram above.

Instance sizes are not arbitrary. The NVIDIA driver provides a fixed list of profiles for each GPU model, and each profile occupies specific physical positions on the card.

Profile (A100 80GB)	Memory	Compute	Max per GPU
1g.10gb	1/8	1/7	7
1g.20gb	2/8	1/7	4
2g.20gb	2/8	2/7	3
3g.40gb	4/8	3/7	2
4g.40gb	4/8	4/7	1
7g.80gb	Full	7/7	1

Other GPUs follow the same pattern with different memory sizes. For example, the H100 80GB uses the same profile names, while the H100 94GB uses 1g.12gb, 3g.47gb and similar. Profiles with a +me suffix include media engines (video decode, JPEG), and only one such small instance is allowed per GPU.

Placement and Fragmentation

Because each profile sits in a fixed physical position, creating and deleting instances in different places can leave gaps. A larger profile may then fail to fit even when total free capacity looks sufficient. The best practice is to plan a standard layout per GPU and change it as a whole.

What Is Isolated

NVIDIA compares MIG with the two earlier ways of sharing a GPU in CUDA: streams within a single process, and the Multi-Process Service (MPS).

	CUDA Streams	MPS	MIG
Partition type	Single process	Logical	Physical
Max partitions	Unlimited	48	7
Compute performance isolation	No	By percentage	Yes
Memory protection	No	Yes	Yes
Memory bandwidth QoS	No	No	Yes
Error isolation	No	No	Yes

	CUDA Streams	MPS	MIG
Reconfiguration	Any time	At process launch	Only when idle

In practice, a crash, memory fault or runaway job in one MIG instance does not affect the others, and each tenant receives a guaranteed share of memory throughput, not only memory capacity. The trade-off is flexibility: partitions have fixed sizes and cannot borrow idle capacity from neighboring instances.

MIG in NexaVM Cloud

NexaVM Cloud provides four GPU delivery modes for cloud VMs: full GPU passthrough, NVIDIA vGPU, MIG, and dGPU dynamic splitting. MIG sits between whole-card passthrough and time-sliced sharing, offering fine-grained allocation with hardware-level isolation.

NVIDIA supports MIG in virtualized environments in two ways:

- **MIG-backed vGPU (multiple VMs per GPU).** The host runs the NVIDIA vGPU host driver, the GPU is placed in MIG mode, and each GPU instance backs one vGPU. Several VMs share one card while keeping MIG's isolation. This mode requires NVIDIA vGPU / NVIDIA AI Enterprise licensing; please consult your hardware vendor for license details.
- **Whole-GPU passthrough with MIG inside the guest.** The entire card is passed through to a single Linux VM, and MIG is configured inside that VM using the same tools as on bare metal. This suits a VM that runs its own container platform and schedules MIG slices itself.

Architecture, from Hardware to VM

Layer	Description
Cloud VM	Guest OS with NVIDIA guest driver and license client; sees one GPU of a fixed size (VM-A, VM-B, VM-C, VM-D).
NexaVM Cloud GPU resources	Device presentation: on Ampere and later GPUs, the card exposes SR-IOV virtual functions; each MIG-backed vGPU is placed on one, and NexaVM Cloud assigns it to a VM.
GPU instances	3g.40gb, 2g.20gb, 1g.10gb, 1g.10gb
Physical GPU	MIG mode enabled (e.g. A100, A30, H100, H200)
NexaVM Cloud host	KVM compute node with the NVIDIA vGPU host driver, managed by NexaVM Cloud.

How instances appear. In NexaVM Cloud, each MIG instance is recognized as an independent device that can be assigned to a VM, similar to a physical PCI device. Physically, all instances are partitions of the same GPU; each VM receives a virtual function or vGPU device backed by its own dedicated, hardware-isolated partition.

Operating Workflow

The typical sequence is shown below. Menu names may vary slightly between NexaVM Cloud releases; GPU operations are found under Resource Center → Hardware → Compute → Physical Machine, with GPU specifications under Compute Configuration.

Step	Action	Details
1	Confirm hardware and licensing	Check that the GPU model is on NVIDIA's MIG support list, that the host driver supports MIG-backed vGPU for that model, and that NVIDIA vGPU / AI Enterprise licenses and a license server are available.
2	Prepare the host	Install the NVIDIA vGPU host driver on the NexaVM Cloud host and enable MIG mode on the GPU. Depending on the model, the driver resets the GPU to apply MIG mode, and some GPUs require a host reboot.
3	Reconnect the host	Reconnect the physical machine in NexaVM Cloud so the platform re-reads the GPU and recognizes it as MIG-capable.
4	Partition the GPU	On the physical GPU, run the split action and select the MIG profiles. The resulting instances appear as GPU devices and as entries in the GPU specification list.
5	Attach instances to VMs	When creating or editing a VM, attach a GPU either by specification (the platform selects a free device and can release it when the VM shuts down) or by choosing a specific device. Install the NVIDIA guest driver and license client in the VM.
6	Change the layout when needed	To change the number or size of instances, detach and release the affected instances, re-partition the GPU, then reconnect the host. Some GPUs require a reboot, so plan layout changes within a maintenance window.

Underlying NVIDIA Commands

For reference and troubleshooting, the platform workflow corresponds to these standard NVIDIA commands (example on an A100 40GB; profile IDs differ by model):

```
# enable MIG mode on GPU 0 (may trigger a GPU reset)
nvidia-smi -i 0 -mig 1

# list the GPU instance profiles supported by this GPU
nvidia-smi mig -lgip

# create 3g.20gb, 2g.10gb and 1g.5gb instances, each with a default compute instance
nvidia-smi mig -i 0 -cgi 9,14,19 -C

# verify the result
nvidia-smi mig -lgi
nvidia-smi -L
```

On hosts managed by NexaVM Cloud, use these commands for inspection only and make changes through the platform, so that the platform's view stays in sync with the hardware.

MIG vs. Time-Sliced vGPU

NVIDIA offers two GPU-splitting technologies, and both are supported in NexaVM Cloud. The table below summarizes the differences.

Category	Time-Sliced vGPU	MIG
Principle	Memory divided into fixed framebuffers; compute shared by time slicing	Compute cores and memory divided statically into hardware partitions
Isolation	Memory isolated; compute performance depends on neighboring VMs	Memory, compute, memory bandwidth and faults isolated
Device recognition	Virtual (mediated) GPU devices	Independent, separately assignable devices
Changing the split	Re-split after devices are detached	Release instances and reconnect the host; some GPUs require a reboot
Maximum per GPU	Depends on profile; can exceed 7	Up to 7 (4 on A30 and RTX PRO 6000 Blackwell; 2 on some smaller GPUs)
Additional licensing	NVIDIA vGPU license	NVIDIA vGPU license; consult your hardware vendor
Typical use cases	VDI, 3D design, cloud gaming, light inference, teaching labs	High-performance computing, AI model training and inference
Supported GPUs	Wide range, including L-series and RTX graphics GPUs	Ampere and later data-center GPUs; see NVIDIA's official list

Planning Considerations

Supported GPUs and Maximum Instances

According to NVIDIA's official support list (September 2026):

GPU	Architecture	Memory	Max Instances
GB200, B200	Blackwell	186 / 180 GB	7
RTX PRO 6000 Blackwell (Server, Workstation, Max-Q)	Blackwell	96 GB	4
RTX PRO 5000 Blackwell	Blackwell	48 GB	2
RTX PRO 4500 Blackwell	Blackwell	32 GB	2
H200 (SXM5, NVL)	Hopper	141 GB	7
H100 (PCIe, SXM5), H100 on GH200, H20	Hopper	80–96 GB	7
A100 (PCIe, SXM4)	Ampere	40 / 80 GB	7
A30	Ampere	24 GB	4

GPUs not on this list, such as the L40S, L4, A10 and A40, do not support MIG and can use GPU passthrough or time-sliced vGPU in NexaVM Cloud instead. Please confirm GPU model compatibility with NexaVM for the specific NexaVM Cloud release in your project.

Key Considerations

- **Licensing.** Multi-VM MIG requires NVIDIA vGPU / AI Enterprise licenses and a license server; include these in the project budget.
- **Layout changes are offline.** Changing the partition layout requires detaching the affected VMs, and some GPUs need a host reboot. A standard layout per GPU pool keeps operations simple.
- **Fixed instance sizes.** A workload needing 25 GB on an A100 80GB must use a 40 GB instance. Size profiles against the real memory footprint of your models.
- **Limited cross-instance communication.** Instances do not behave like separate NVLink-connected GPUs, so MIG suits many independent workloads rather than one job spread across several slices.
- **Best-fit scenarios.** Multi-tenant inference, serving many small models, development and notebook environments, and training platforms where each user needs guaranteed GPU resources.

References

- [NVIDIA MIG User Guide: Supported GPUs](#)
- [NVIDIA MIG User Guide: Concepts](#)
- [NVIDIA MIG User Guide: Supported MIG Profiles](#)
- [NVIDIA MIG User Guide: Virtualization](#)
- [NexaVM official website](#)