



cloud STACK

INDIA USER GROUP
MEETUP

CLOUDSTACK GPU INTEGRATION

Bringing AI, Graphics & VDI to IaaS Clouds

ROHIT YADAV
CTO, SHAPEBLUE

JULY 11, 2025 | GREATER NOIDA, INDIA

\$ whoami

- Committer, PMC Member & Ex-VP Apache CloudStack
- Chair, CloudStack India User Group
- ASF Member & Kubernetes SIG Member
- CTO, ShapeBlue
- Tinkerer, Homelabber, Renewables Champion



CloudStack & Legacy GPU Support

- ExtraConfig
- KVM Groovy-based Agent Hooks
- Out-of-band configuration (hacking hypervisors)
- Old GPU/vGPU support with XenServer

<https://cwiki.apache.org/confluence/display/CLOUDSTACK/GPU+and+vGPU+support+for+CloudStack+Guest+VMs>

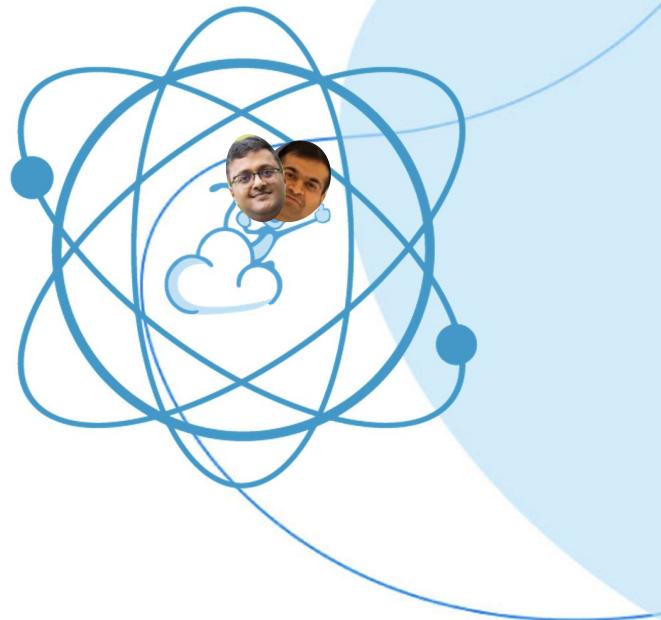
GPU Integration Proposal



Vishesh Jindal @ CCC24 Madrid
Staff Software Engineer, ShapeBlue

<https://www.youtube.com/watch?v=UyJTcy69ncg>

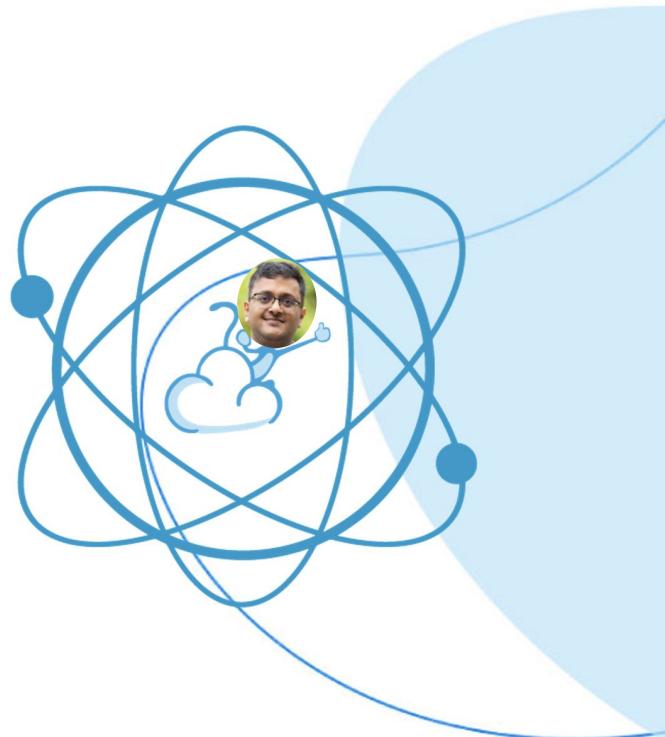
CloudStack GPU Integration Collaborative Effort



CloudStack GPU Integration - Technical Preview

- Pull Request submitted by Vishesh:
<https://github.com/apache/cloudstack/pull/11143>
- Feature to ship as Technical Preview in ACS 4.21**

*** depends on community acceptance*



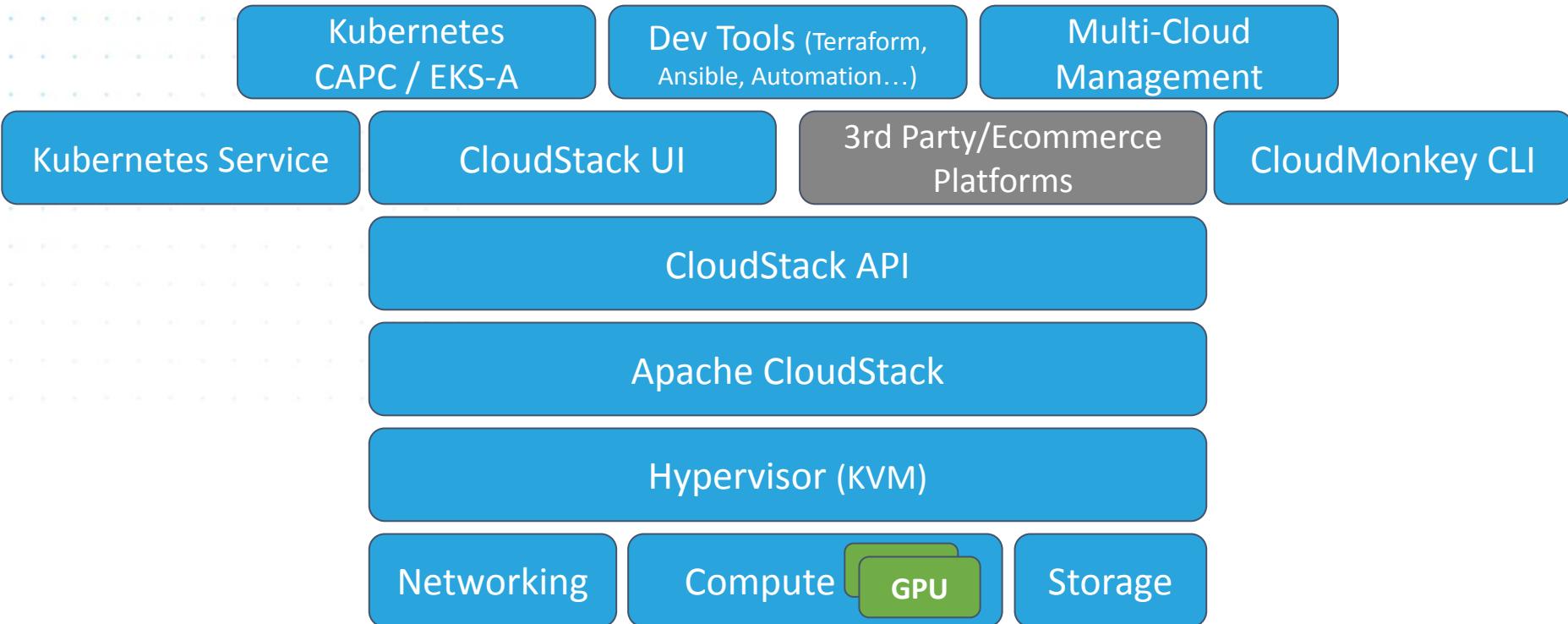
Why this feature?

Why Have this Feature?

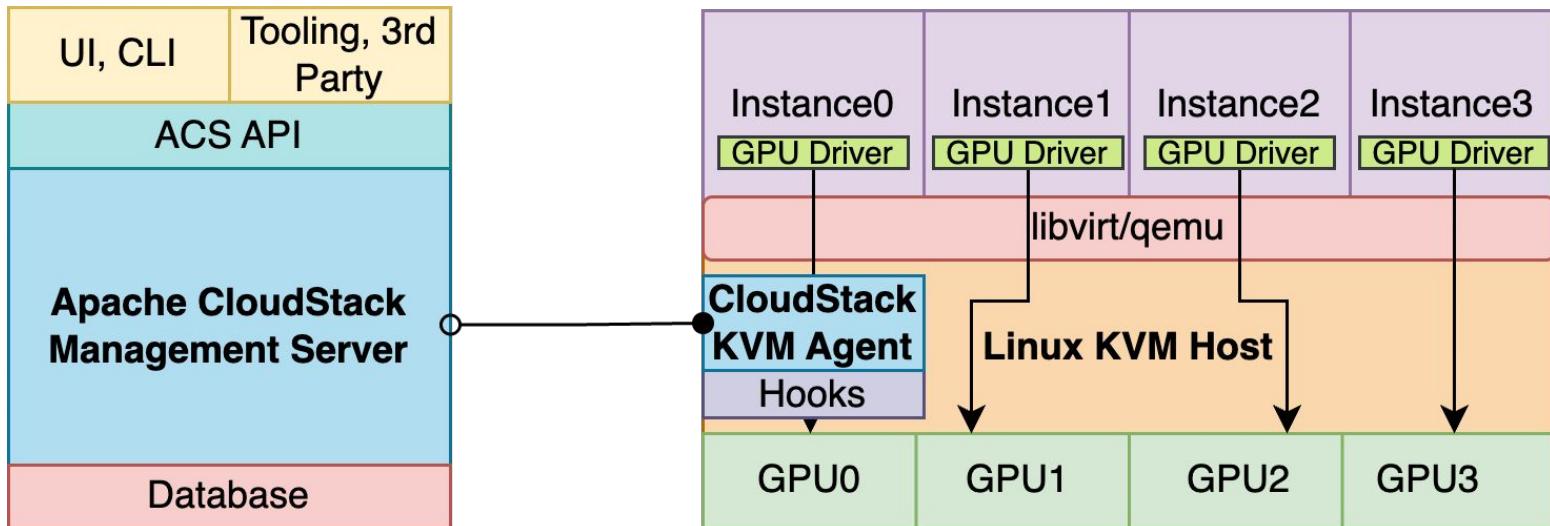
- Satisfy variety of GPU use-cases
- Enable CSPs to offer GPU-backed compute offerings
- Data Locality & Sovereignty
- Self-service access to GPUs
- Benefit end-users: AI/ML engineers, VDI users, R&D Teams

GPU-Integration Design & Architecture

Apache CloudStack Architecture



Feature Design & Architecture



GPU-Integration Demo

GPU-Enabled Instance Deployment

3 Image

Type

Template ISO

OS image that can be used to boot Instances.

Operating System

--	--	--

Filter by Search

GPU Ubuntu 24.04 Featured Public HVM

Override root disk size

Total 1 items < **1** > 10 / page

Your Instance

Simulator

OS type Other Ubuntu (64-bit)

CPU 2 CPU x 1.00 GHz

Memory 2048 MB memory

GPU Cards 2 x ACS Simulated Graphics Card Pro (sim-8q)

Disk size (in GB) 1 GB (Root)

Networks Admin Isolated Network (Default)

Template GPU Ubuntu 24.04

Compute offering Sim 2xMedium GPU Offering

Launch Instance

GPU-Enabled Instance Deployment

3 Image

Type

Template ISO

OS image that can be used to boot Instances.

Operating System

 GPU Temp...	 CentOS	 User
--	--	--

Filter by Search

<input checked="" type="radio"/>  GPU Ubuntu 24.04	Featured	Public	HVM
---	----------	--------	-----

Total 1 items < 1 > 10 / page

Override root disk size

4 Compute offering

Show only GPU enabled offerings

Search

Compute offering	CPU	Memory	GPU Cards
<input type="radio"/> Small Instance	1 CPU x 0.50 Ghz	512 MB	
<input type="radio"/> Medium Instance	1 CPU x 1.00 Ghz	1024 MB	
<input checked="" type="radio"/> Sim 2xMedium GPU Offering	2 CPU x 1.00 Ghz	2048 MB	2 x ACS Simulated Graphics Card Pro (sim-8q)

Override root disk offering

Total 3 items < 1 > 10 / page

Your Instance

 Simulator

OS type Other Ubuntu (64-bit)

CPU 2 CPU x 1.00 Ghz

Memory 2048 MB memory

GPU Cards 2 x ACS Simulated Graphics Card Pro (sim-8q)

Disk size (in GB) 1 GB (Root)

Networks Admin Isolated Network (Default)

Template GPU Ubuntu 24.04

Compute offering Sim 2xMedium GPU Offering

Launch Instance



GPU-Enabled Instance Deployment

3 Image

Type

Template ISO

OS image that can be used to boot Instances.

Operating System

	CentOS	User
---	--------	------

Filter by Search

<input checked="" type="radio"/>  GPU Ubuntu 24.04	Featured	Public	HVM
---	----------	--------	-----

Total 1 items < 1 > 10 / page

Override root disk size

4 Compute offering

Show only GPU enabled offerings

Compute offering	CPU	Memory	GPU Cards
Small Instance	1 CPU x 0.50 Ghz	512 MB	
Medium Instance	1 CPU x 1.00 Ghz	1024 MB	
<input checked="" type="radio"/> Sim 2xMedium GPU Offering	2 CPU x 1.00 Ghz	2048 MB	2 x ACS Simulated Graphics Card Pro (sim-8q)

Override root disk offering

Your Instance

Simulator

OS type Other Ubuntu (64-bit)

CPU 2 CPU x 1.00 Ghz

Memory 2048 MB memory

GPU Cards 2 x ACS Simulated Graphics Card Pro (sim-8q)

Disk size (in GB) 1 GB (Root)

Networks Admin Isolated Network (Default)

Template GPU Ubuntu 24.04

Compute offering Sim 2xMedium GPU Offering

Launch Instance

GPU-Enabled Instance (User View)

Instances Refresh Mine Metrics Projects Add Instance + Search

Name	State	IP Address	Arch	Compute offering	Zone
User-GPU1 GPU Enabled	Running	10.1.1.235		Sim 2xMedium GPU Offering	Sandbox-simulator

Instances / User-GPU1 Refresh

User-GPU1

Status
Running

ID
52c792cc-7eb7-4bbb-9c80-c4d09fd19454

IP address
10.1.1.235

CPU
2 CPU x 1.00 GHz

Memory
2048 MB memory

GPU Cards
2 x ACS Simulated Graphics Card Pro (sim-8q)

Details Metrics Volumes GPU Cards NICs Instance Snapshots Schedules Settings Events Comments

Name Total

ACS Simulated Graphics Card Pro 2

GPU-Enabled Instance (Admin View)

Instances / Test-GPU-VM1  

 **Test-GPU-VM1**

i-2-11-QA x86_64 Simulator 

Status: Running

ID: c8601926-c273-4cf3-afb0-7cd7ed7237f9

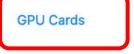
IP address: 10.1.1.250

CPU: 2 CPU x 1.00 GHz x86_64

Memory: 2048 MB memory

GPU Cards (2 x ACS Simulated Graphics Card Pro (sim-8q))

Network: 1 NIC(s) eth0 10.1.1.250 Default Admin Isolated Network

Details Metrics Volumes GPU Cards 

Summary GPU Devices

Name	Total	Allocated	Available
ACS Simulated Graphics Card Pro	2	2	0

NICs Instance Snapshots Schedules Settings Events Comments

GPU-Enabled Instance (Admin View)

Instances / Test-GPU-VM1 

 Test-GPU-VM1

i-2-11-QA x86_64 Simulator 

Status
● Running

ID
 c8601926-c273-4cf3-afb0-7cd7ed7237f9

IP address
 10.1.1.250

CPU
 2 CPU x 1.00 GHz 

Memory
 2048 MB memory

GPU Cards
 2 x ACS Simulated Graphics Card Pro (sim-8q)

Details

Metrics

Volumes

GPU Cards

NICs

Instance Snapshots

Schedules

Settings

Events

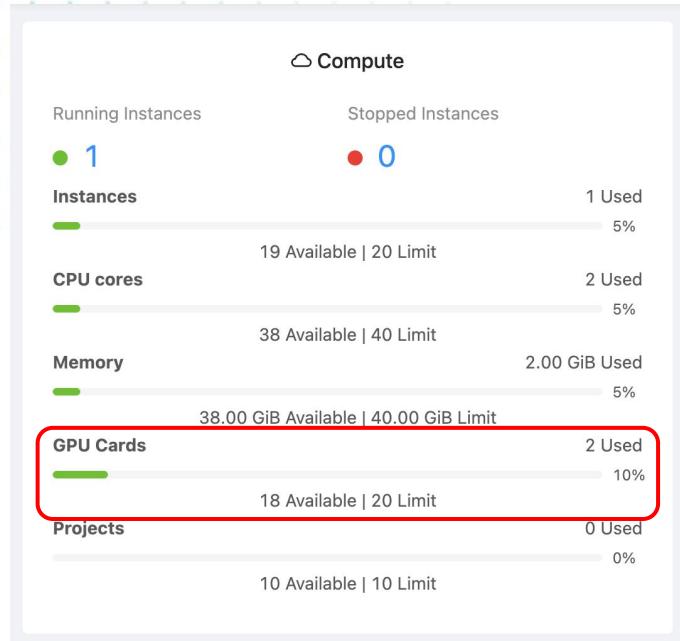
Comments

Summary GPU Devices

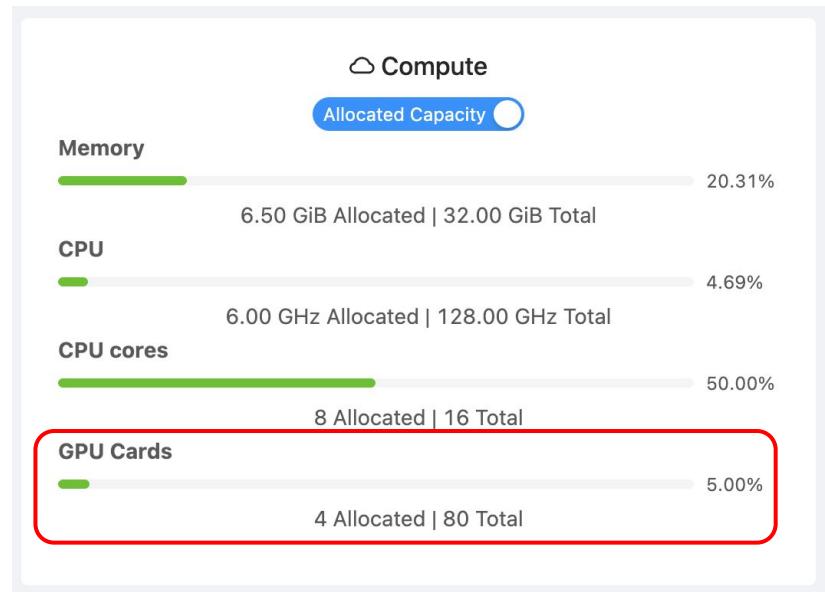
GPU Card	Profile	Host	Address	Managed state	State
<input type="checkbox"/> ACS Simulated Graphics Ca... sim-8q	H4	4399b5e6-4527-42f3-a384-d41f3f4297da	 Managed	 Allocated	
<input type="checkbox"/> ACS Simulated Graphics Ca... sim-8q	H4	b80ed272-e0ac-480f-b6a5-2b9fc7fd746	 Managed	 Allocated	

Dashboard GPU Allocation / Limit



User Dashboard



Admin Dashboard



Tenant GPU Limits

[Home](#) / Accounts / user  Refresh

 user

Status: Enabled

ID: 5fa3a3d2-19a9-4b95-94ad-e8c85f927611

CPU: 2

Memory: 2048 Memory

GPU Cards: 2 GPU Cards

Role: User

Domain: ROOT

Details **Limits** Configure limits Certificate Settings Events

Instance limits (19 Available)

Used / Limit : 1 / 20 5.00%

CPU limits (38 Available)

Used / Limit : 2 / 40 5.00%

Memory limits (MiB) (38912 Available)

Used / Limit : 2048 / 40960 5.00%

GPU limits (18 Available)

Used / Limit : 2 / 20 10.00% (This row is highlighted with a red box)

Primary storage limits (GiB) (200 Available)

Used / Limit : 0 / 200 0.00%

Volume limits (19 Available)

Used / Limit : 1 / 20 5.00%

Public IP Limits (19 Available)

Used / Limit : 1 / 20 5.00%

Network limits (19 Available)

- `max.account.gpus` - The default maximum number of GPU devices that can be used for an account.
- `max.domain.gpus` - The default maximum number of GPU devices that can be used for a domain.
- `max.project.gpus` - The default maximum number of GPU devices that can be used for a project.
- `gpu.detach.on.stop` (default: false, scope: domain): When true the GPU devices are detached from stopped instances, available for allocation to other instances.

GPU-Enabled Service Offerings

Add compute offering 

* Name 

Description 

Compute offering type

Fixed offering Custom constrained Custom unconstrained

* CPU cores  * CPU (in MHz)  * Memory (in MB) 

Host tags  Network rate (Mb/s) 

Offer HA  Dynamic scaling enabled 

CPU cap  Volatile 

Deployment planner 

GPU Card 

GPU Profile 

GPU Count  GPU Display 

Public 

GPU Card

ACS Simulated Graphics Card Pro

GPU Profile

passthrough

sim-4q

sim-2q

GPU-Enabled Service Offerings

Home / Compute offerings   Active    Search

<input type="checkbox"/>	Name	Description	State	CPU cores	CPU (in MHz)	Memory	GPU Cards
<input type="checkbox"/>	Small Instance	Small Instance	 Active	1	500	512	
<input type="checkbox"/>	Medium Instance	Medium Instance	 Active	1	1000	1024	
<input type="checkbox"/>	Sim 2xGPU Offering 	Sim 2xGPU Offering	 Active	16	2000	40960	2x ACS Simulated Graphics Card Pro(sim-8q)

Showing 1-3 of 3 items <  > 20 / page 

Usage Tracking: GPU-Instances

GPU-enabled instance usage for tenants is tracked (for billing and other purposes) via GPU-enabled service offerings for allocated and running VMs:

Usage Type 1 - RUNNING_VM

Usage Type 2 - ALLOCATED_VM

GPU Inventory Management

Hosts  All  Metrics 

 Search

Name	State	Resource state	IP Address	Arch	Hypervisor	System VMs	Version	Cluster name	Zone	Management Server
H1	 Up	 Enabled	172.16.15.2	x86_64	Simulator		4.21.0.0-SNAPSHOT	C0	Sandbox-simulator	ref-trl-8904-k-mol8-vishesh-jindal-mgmt1.sofia.shapeblue.com
H2	 Up	 Enabled	172.16.15.10	x86_64	Simulator		4.21.0.0-SNAPSHOT	C0	Sandbox-simulator	ref-trl-8904-k-mol8-vishesh-jindal-mgmt1.sofia.shapeblue.com
H3	 Up	 Enabled	172.16.15.23	x86_64	Simulator		4.21.0.0-SNAPSHOT	C1	Sandbox-simulator	ref-trl-8904-k-mol8-vishesh-jindal-mgmt1.sofia.shapeblue.com
H4	 Up	 Enabled	172.16.15.14	x86_64	Simulator		4.21.0.0-SNAPSHOT	C1	Sandbox-simulator	ref-trl-8904-k-mol8-vishesh-jindal-mgmt1.sofia.shapeblue.com

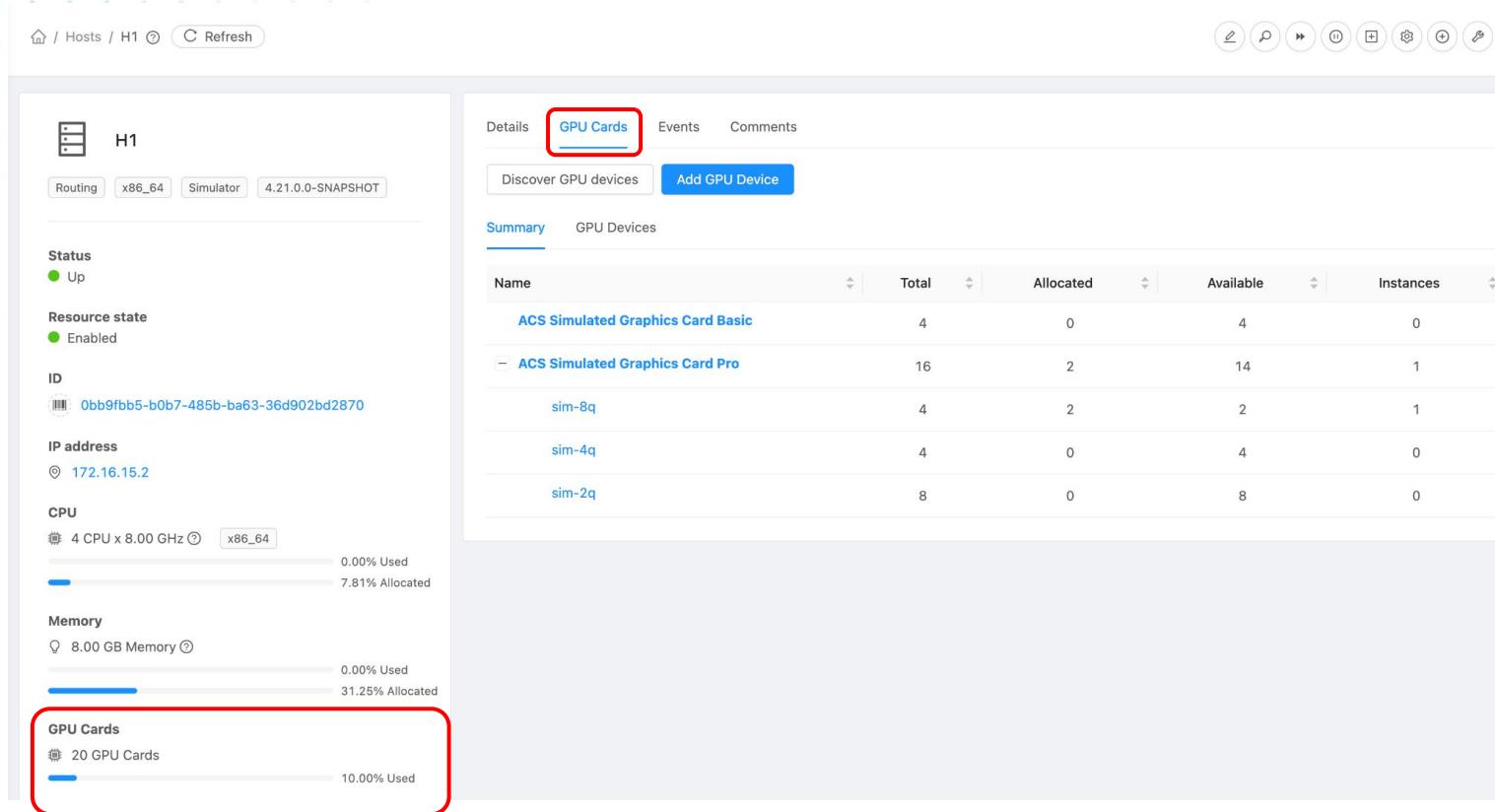
Showing 1-4 of 4 items < [1](#) > 20 / page ▾

Host BIOS Setup

Preparing GPU-based Hosts

- **Enable SR-IOV:** Required for creating virtual function (VF) for sharing PCI GPU
- **Enable VT-d (Intel) or IOMMU (AMD):** Required for VFIO-PCI and shared GPU, for secure DMA and device isolation
- **Enable Virtualisation:** Intel VMX or AMD SVM
- **PCI settings:** GPU card(s) specific tunings
 - Above 4G Decoding
 - PCIe ARI Support
 - ...

GPU Inventory Management



The screenshot shows a web-based interface for managing GPU inventory. On the left, a sidebar displays host details for 'H1'. The 'GPU Cards' tab is selected in the main navigation bar. A table below lists various GPU devices with their total, allocated, available, and instance counts.

Host Details:

- H1**
- Status:** Up
- Resource state:** Enabled
- ID:** 0bb9fb5-b0b7-485b-ba63-36d902bd2870
- IP address:** 172.16.15.2
- CPU:** 4 CPU x 8.00 GHz (x86_64)
- Memory:** 8.00 GB Memory
- GPU Cards:** 20 GPU Cards

GPU Cards Tab:

Summary GPU Devices

Name	Total	Allocated	Available	Instances
ACS Simulated Graphics Card Basic	4	0	4	0
- ACS Simulated Graphics Card Pro	16	2	14	1
sim-8q	4	2	2	1
sim-4q	4	0	4	0
sim-2q	8	0	8	0



GPU Inventory Management

Hosts / H1 ⏪ C Refresh

H1

Routing x86_64 Simulator 4.21.0.0-SNAPSHOT

Status Up

Resource state Enabled

ID 0bb9fb5b-b0b7-485b-ba63-36d902bd2870

IP address 172.16.15.2

CPU 4 CPU x 8.00 GHz x86_64 0.00% Used 7.81% Allocated

Memory 8.00 GB Memory 0.00% Used 31.25% Allocated

GPU Cards 20 GPU Cards 10.00% Used

GPU Cards (highlighted)

Details GPU Cards Events Comments

Discover GPU devices Add GPU Device

Summary GPU Devices

<input type="checkbox"/> Address	GPU Card	Profile	Managed state	State	Instance name	Actions	
00:01.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
00:02.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
00:03.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
00:04.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
00:05.0	ACS Simulated Graphics C...						
1b5b09f7-3bfc-4ed0-b998-42883cab69c	ACS Simulated Graphics C...	sim-8q	Managed	Allocated	i-4-13-QA		
1eb93f63-2108-4c05-99af-2e1cd0782cd7	ACS Simulated Graphics C...	sim-4q	Managed	Free			
09c8f8b5-3685-4e91-9be5-fdf6aa7a6f15	ACS Simulated Graphics C...	sim-2q	Managed	Free			
88d48ac1-a506-417a-8128-a316404aac55	ACS Simulated Graphics C...	sim-2q	Managed	Free			



GPU Inventory Management

Hosts / H1 ⏪ C Refresh

H1

Status: Up

Resource state: Enabled

ID: 0bb9fb5-b0b7-485b-ba63-36d902bd2870

IP address: 172.16.15.2

CPU: 4 CPU x 8.00 GHz (x86_64) 0.00% Used 7.81% Allocated

Memory: 8.00 GB Memory 0.00% Used 31.25% Allocated

GPU Cards: 20 GPU Cards 10.00% Used

GPU Cards (Tab): Details, GPU Cards (highlighted), Events, Comments

Action Buttons: Discover GPU devices, Add GPU Device (highlighted)

GPU Devices (Table):

<input type="checkbox"/>	Address	GPU Card	Profile	Managed state	State	Instance name	Actions	
<input type="checkbox"/>	00:01.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/>	00:02.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/>	00:03.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/>	00:04.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/>	00:05.0	ACS Simulated Graphics C...						
<input type="checkbox"/>	1b5b09f7-3bfc-4ed0-b998-42883cab69c	ACS Simulated Graphics C...	sim-8q	Managed	Allocated	i-4-13-QA		
<input type="checkbox"/>	1eb93f63-2108-4c05-99af-2e1cd0782cd7	ACS Simulated Graphics C...	sim-4q	Managed	Free			
<input type="checkbox"/>	09c8f8b5-3685-4e91-9be5-fdf6aa7a6f15	ACS Simulated Graphics C...	sim-2q	Managed	Free			
<input type="checkbox"/>	88d48ac1-a506-417a-8128-a316404aac55	ACS Simulated Graphics C...	sim-2q	Managed	Free			



GPU Inventory Management

Hosts / H1 ⏪ C Refresh

H1

Details GPU Cards Events Comments

Discover GPU devices Add GPU Device

GPU Devices

<input type="checkbox"/> Address	GPU Card	Profile	Managed state	State	Instance name	Actions	
<input type="checkbox"/> 00:01.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/> 00:02.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/> 00:03.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/> 00:04.0	ACS Simulated Graphics C...	passthrough	Managed	Free			
<input type="checkbox"/> 00:05.0	ACS Simulated Graphics C...						
<input type="checkbox"/> 1b5b09f7-3bfc-4ed0-b998-42883cab69c	ACS Simulated Graphics C...	sim-8q	Managed	Allocated	i-4-13-QA		
<input type="checkbox"/> 1eb93f63-2108-4c05-99af-2e1cd0782cd7	ACS Simulated Graphics C...	sim-4q	Managed	Free			
<input type="checkbox"/> 09c8f8b5-3685-4e91-9be5-fdf6aa7a6f15	ACS Simulated Graphics C...	sim-2q	Managed	Free			
<input type="checkbox"/> 88d48ac1-a506-417a-8128-a316404aac55	ACS Simulated Graphics C...	sim-2q	Managed	Free			

GPU Cards

20 GPU Cards 10.00% Used

GPU Classification: GPU Card Types

Dashboard

Compute

Storage

Network

Images

Events

Projects

Roles

Accounts

Domains

Infrastructure

Service offerings

Configuration

- Global Settings
- LDAP configuration
- OAuth configuration
- Hypervisor capabilities
- Guest OS Categories
- Guest OS
- Guest OS mappings
- GPU Card**

Default view

/ GPU Card / ACS Simulated Graphics Card Pro Refresh

ACS Simulated Graphics Card Pro

ID: f1595671-6f3a-42a6-9d83-e3fde623bdd

View GPU Device

View GPU Profile

Profile

Add profile

GPU Profile	Video RAM	Max heads	Resolution	Max. vGPUs per physical GPU
passthrough	0	1		1
sim-8q	8.00 GB	4	4096x2160	2
sim-4q	4.00 GB	2	1920x1080	4
sim-2q	2.00 GB	1	1920x1080	8

Licensed under the Apache License, Version 2.0.
CloudStack 4.21.0.0-SNAPSHOT Ask a question or Report an issue

Host Allocation & Orchestration

- Finds a set of host(s) based on host-GPU inventory, service offering, GPU limits and other input parameters
- Allocates GPU(s) on host to the instance based on service offering, GPU(s) availability and other parameters (such as NUMA)

Customisable Agent Hooks

GPU Auto-Discovery Script

Operator customisable GPU
auto-discovery script

`/etc/cloudstack/agent/hooks/gpudiscovery.sh`

```
{  
    "gpus": [  
        {  
            "pci_address": "01:00.0",  
            "vendor_id": "10de",  
            "device_id": "25a0",  
            "vendor": "NVIDIA Corporation",  
            "device": "GA107M [GeForce RTX 3050 Ti Mobile]",  
            "driver": "nvidia",  
            "pci_class": "3D controller [0302]",  
            "iommu_group": "16",  
            "pci_root": "0000:01:00.0",  
            "numa_node": -1,  
            "sriov_totalvfs": 0,  
            "sriov_numvfs": 0,  
            "full_passthrough": {  
                "enabled": 1,  
                "libvirt_address": {  
                    "domain": "0x0000",  
                    "bus": "0x01",  
                    "slot": "0x00",  
                    "function": "0x0"  
                },  
                "used_by_vm": null  
            },  
            "vgpu_instances": [],  
            "vf_instances": []  
        }  
    ]  
}
```

Customisable Agent Hooks

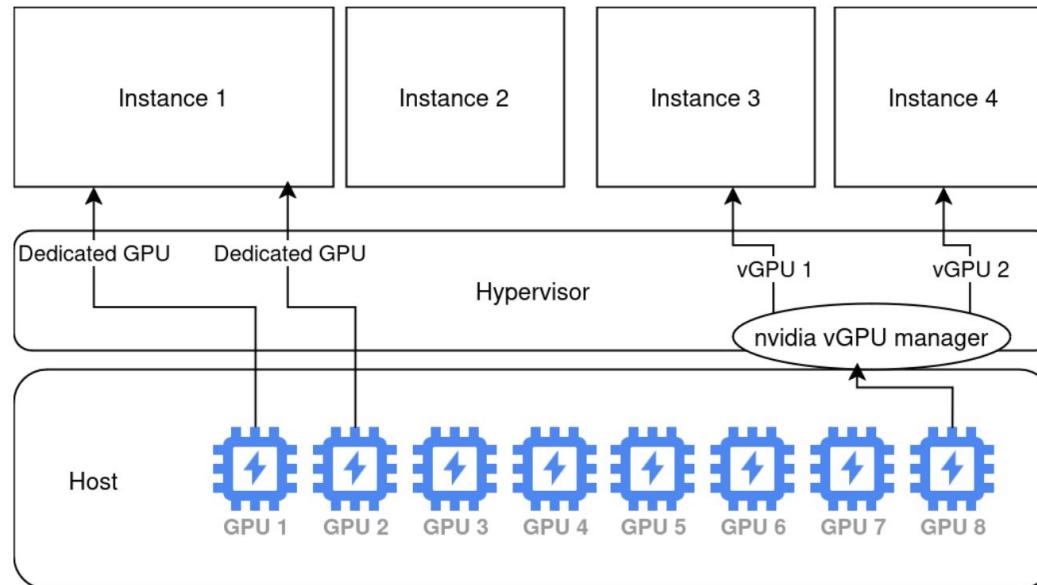
Domain XML Transformer & States

In addition to the Groovy-based agent hooks, following are supported:

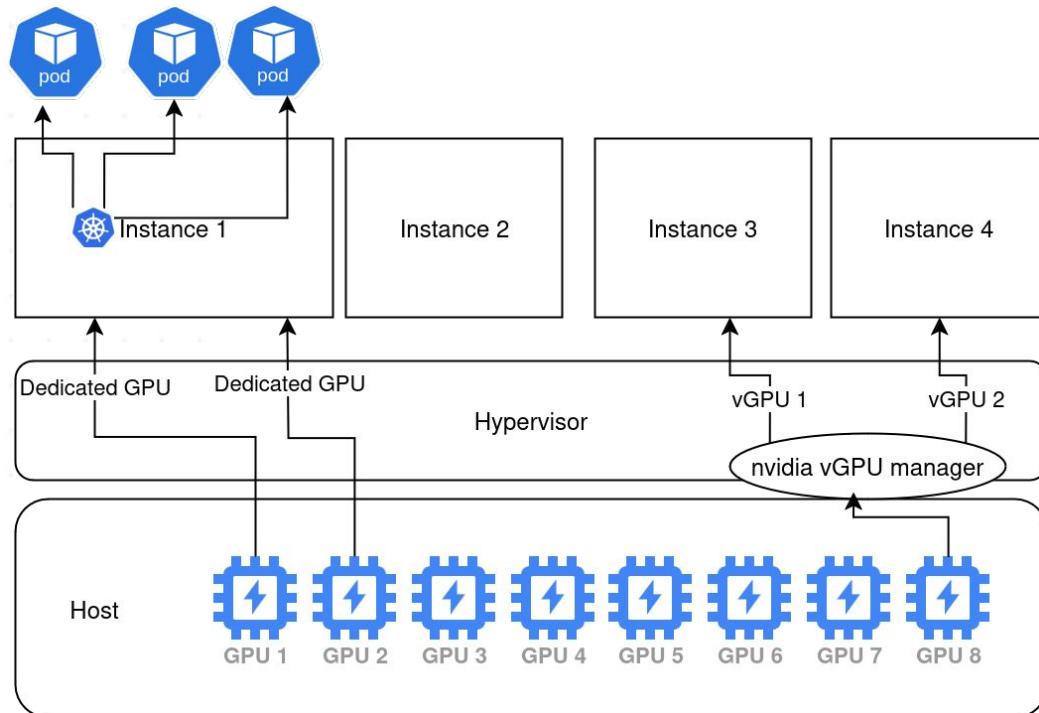
- Libvirt Domain XML Transformer `libvirt-vm-xml-transformer.sh`
- VM Start hook `libvirt-vm-state-change.sh`
- VM Stop hook `libvirt-vm-state-change.sh`

Location: `/etc/cloudstack/agent/hooks`

GPUs with Instances



GPUs with K8S (CKS/CAPC/EKS-A)



GPUs with Baremetal?

Yes - Possible!

Checkout the new
CloudStack Extensions Framework

Assumption & Limitations

- The role of CloudStack is limited to discover the devices on the host and assign/unassign them to instances. Admin/operator must configure the GPU devices properly on the KVM hosts.
- Admins/users must build custom templates with GPU tools and drivers specific to their environments and GPUs.
- Unsupported actions (as of now):
 - VM snapshot with memory for GPU enabled VM is not supported.
 - Dynamic scaling of VMs with GPU attached.
 - Live Migration of GPU enabled VM is not supported.
 - Cold migration is supported.
 - Migration with vGPU may work, but not verified/tested.

Understanding Current GPU Technologies

GPU Vendors (Currently)

	NVIDIA	AMD	Intel
Platform	CUDA	ROCM	oneAPI
Virtualisation	vGPU (GRID)	MxGPU (SR-IOV)	Basic/Flex
Current Standing	Market Leader	Growing Contender	Catching up
Ecosystem	PyTorch, TensorFlow...	Improving support in PyTorch, etc.	Laggard

Other contenders: Apple (Metal), Qualcomm (Adreno)...

GPU Types

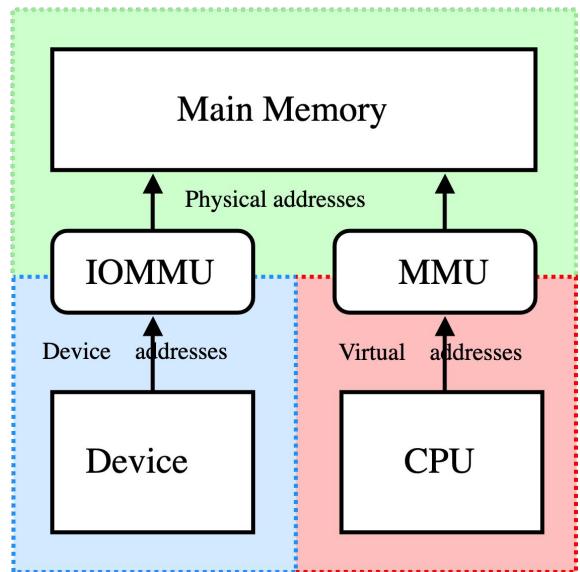
- **Passthrough GPU:** full access of GPU card
- **Shared GPU:** sliced / time-shared / partitioned GPU access
 - SR-IOV based (AMD MxGPU, Intel Flex)
 - vGPU (Nvidia)
 - MIG (Nvidia)
 - ...

IOMMU: What is that?

The I/O Memory Management Unit (IOMMU) provides memory remapping services for I/O devices.

It adds support for address translation and system memory access protection on direct memory access (DMA) transfers from peripheral devices.

IOMMU maps device-visible virtual addresses (also called device addresses or memory mapped I/O addresses in this context) to physical addresses.



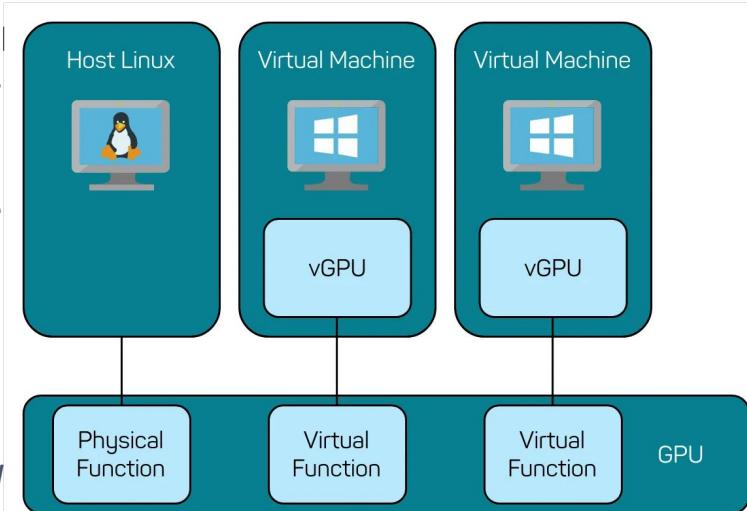
SR-IOV: What is that?

Single Root IO Virtualization (SR-IOV) allows the isolation of PCI Express resources for manageability and performance reasons. A single physical PCI Express bus can be shared in a virtual environment using the SR-IOV specification.

- Physical Functions: ability to move data in and out of the device.
- Virtual Functions (VF): lightweight PCIe functions that support data flowing but also have a restricted set of configuration resources.

The SR-IOV allows different virtual machines (VMs) in a virtual environment to share a single PCI Express hardware interface.

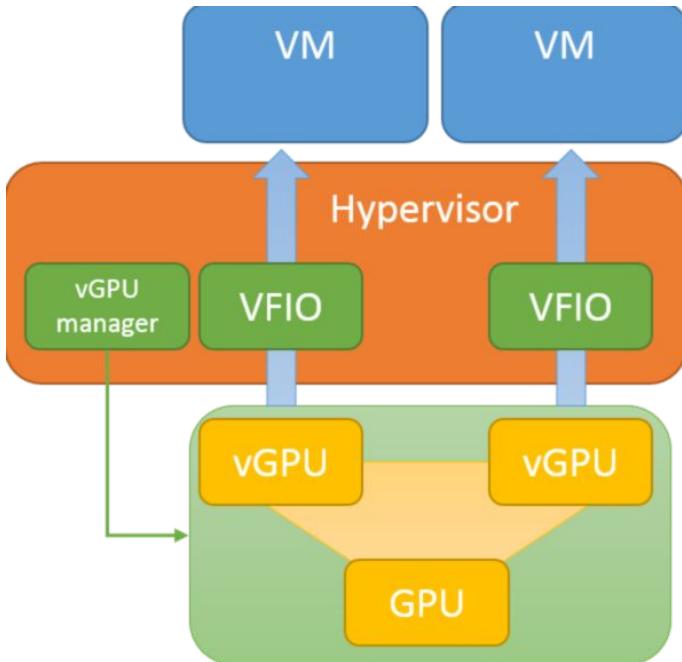
TL;DR: SR-IOV enables VMs to access shared GPU(s) via Virtual Function



VFIO: What is that?

Virtual Function I/O (VFIO) *driver* is an IOMMU/device agnostic (Linux) framework for exposing direct device access to userspace, in a secure, IOMMU protected environment.

TL;DR: VFIO enables direct access of GPU Device to VM

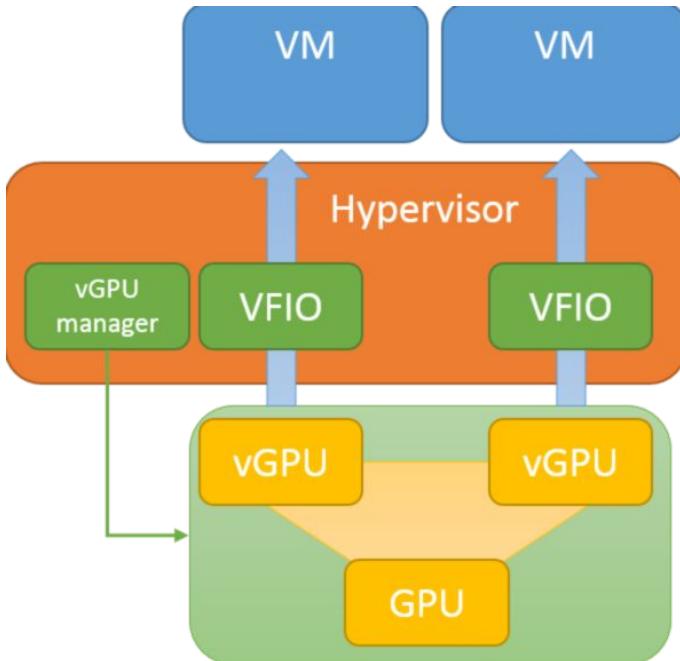


MDEV: What is that?

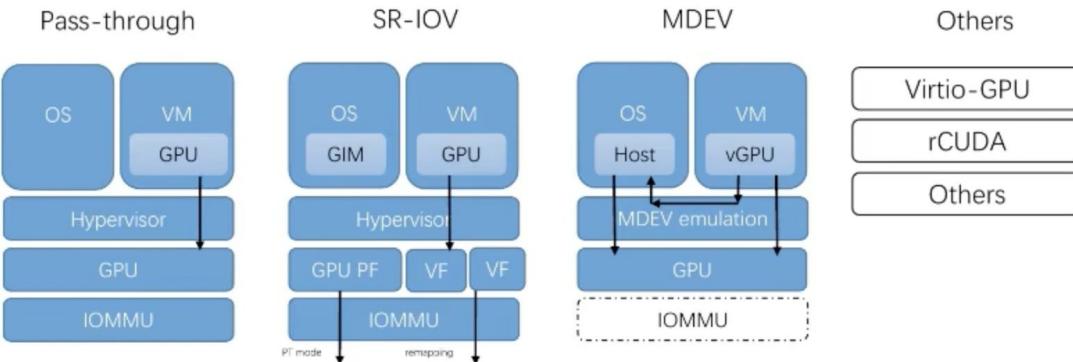
Mediated device (mdev) is a virtual device created by the host kernel from a physical PCI device, like a GPU, and exposed to VMs as a virtual PCI device.

The mdev framework allows the host device driver (e.g., NVIDIA, Intel, AMD) to partition and manage hardware resources, while still allowing each guest to use its own slice with near-native performance.

TL;DR: MDEV creates GPU partitions that can be used with VM using VFIO-mdev



GPU Virtualisation



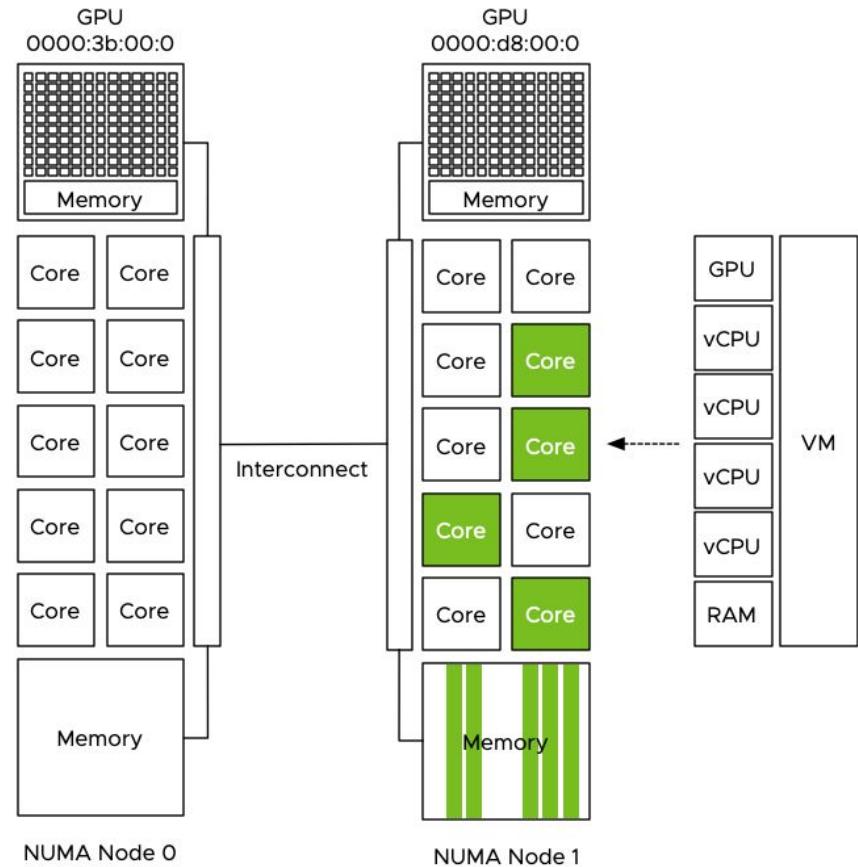
GPU Virtualisation

Feature	MDEV	SR-IOV	VFIO Passthrough
Sharing	Yes	Yes	No
Uses VFIO	Yes (<code>vfio_mdev</code>)	Yes (<code>vfio-pci</code>)	Yes (<code>vfio-pci</code>)
Granularity	Fine-grained virtual functions	Hardware-based VFs	Entire physical device
Device support	Software-defined (via driver)	Hardware-defined	Full passthrough
Needs IOMMU	Yes	Yes	Yes
Examples	NVIDIA vGPU	AMD MxGPU, Intel Flex	Full GPU passthrough
Guest Driver	Vendor vGPU driver	Vendor driver	Vendor driver

NUMA (Non-uniform memory access)

- NUMA systems are server platforms with more than one system bus.
- These platforms can utilize multiple processors on a single motherboard, and all processors can access all the memory on the board.

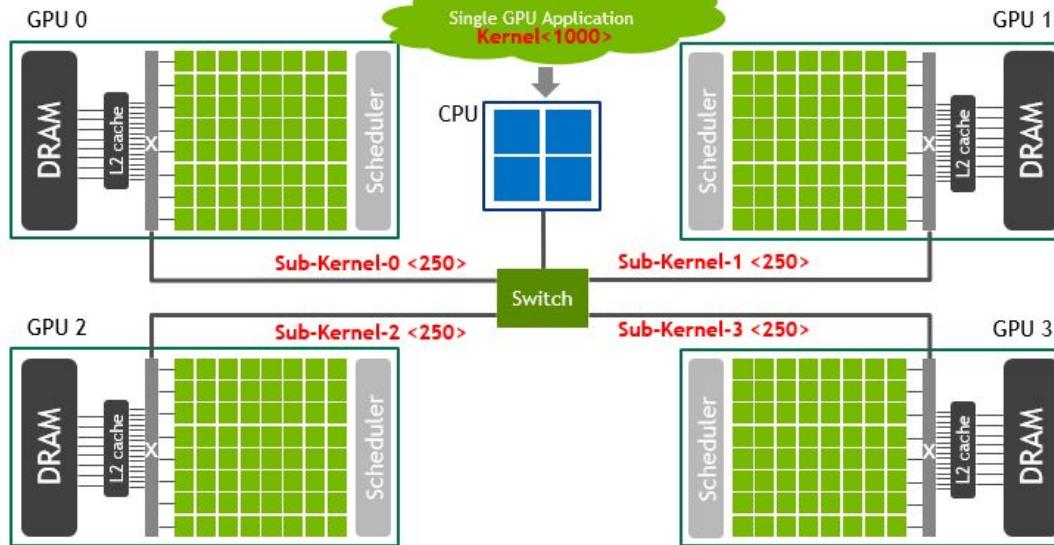
TL;DR: GPU placements on different NUMA nodes can have severe performance impact.



NUMA (Non-uniform memory access)

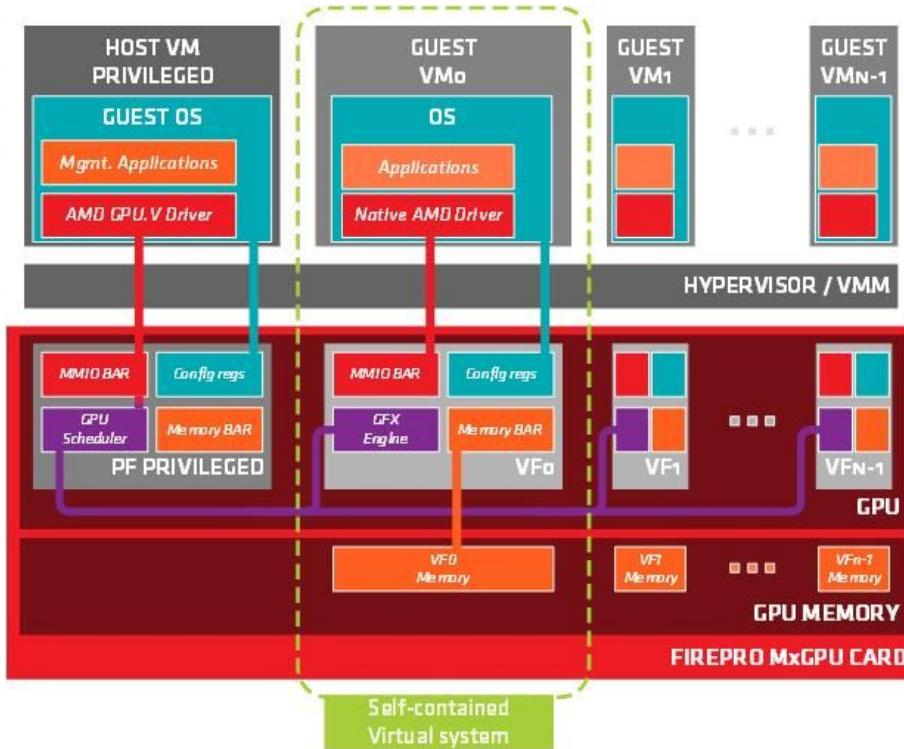
NUMA AWARE GPUS

Enable Scaling of Single GPU Applications Transparently on Multi-GPUs



- Dynamic and asymmetric NVLINK bandwidth reconfiguration
- Dynamic NUMA-aware caching strategies

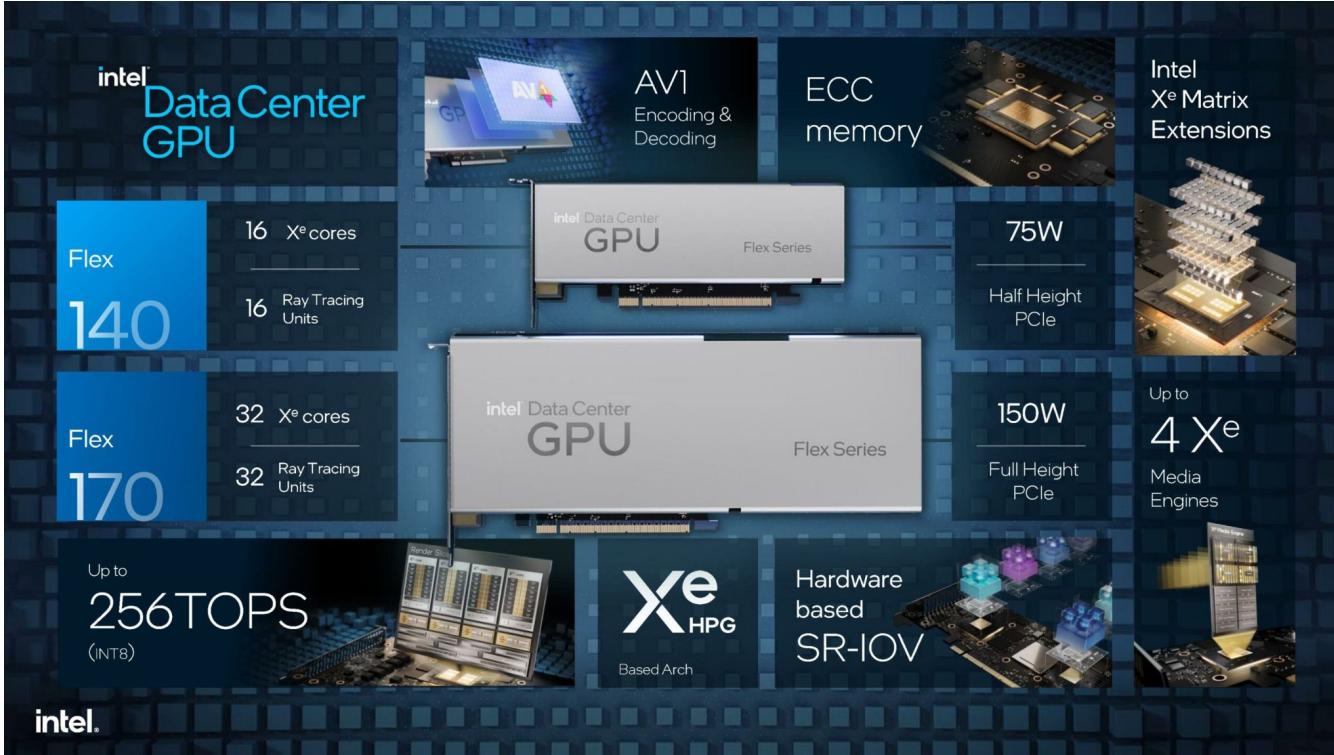
Example: AMD MxGPU (SR-IOV)



AMD GPU Cards

- MI210X
- MI300X
- MI350X

Example: Intel DC GPU Flex (SR-IOV)

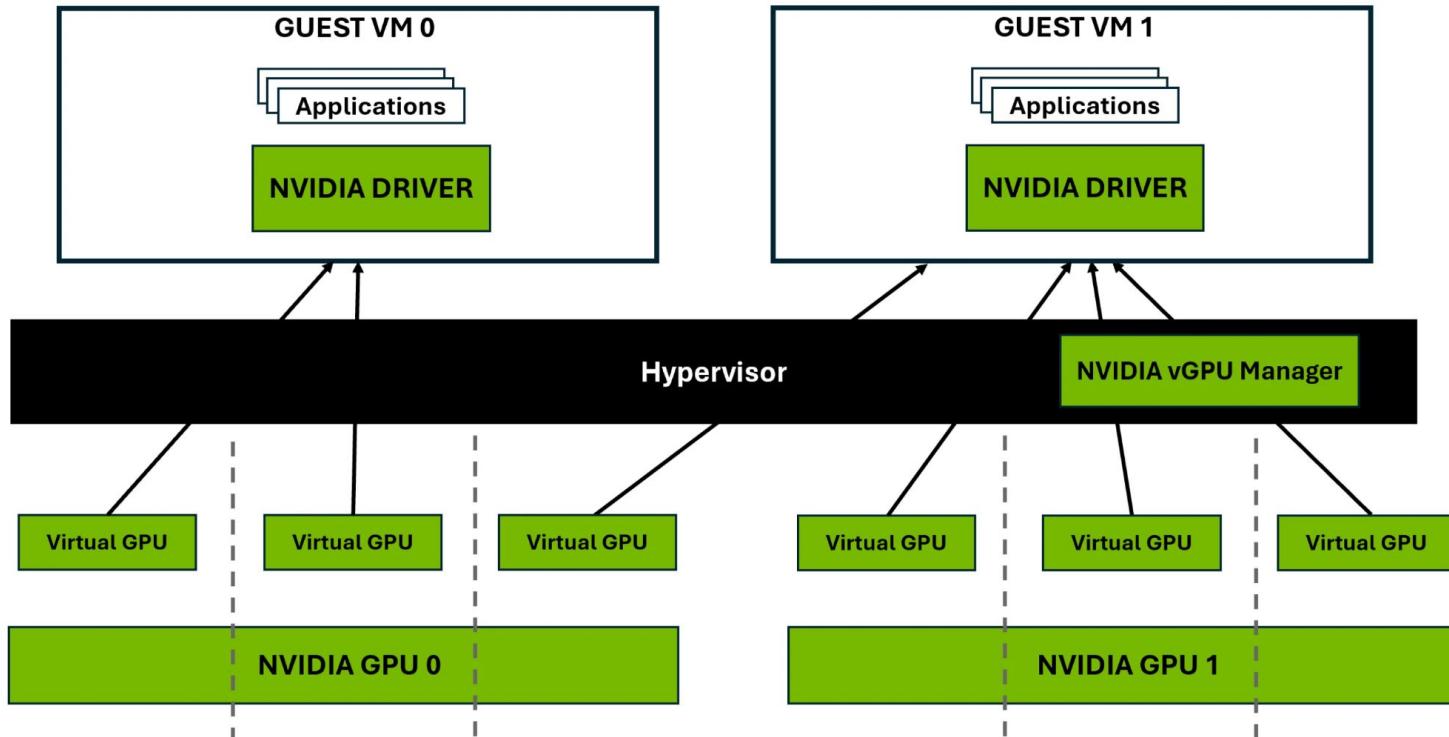


The image displays the Intel Data Center GPU Flex Series, featuring two PCIe cards and various technical specifications:

- intel DataCenter GPU**
- Flex 140**: 16 Xe cores, 16 Ray Tracing Units.
- Flex 170**: 32 Xe cores, 32 Ray Tracing Units.
- Up to 256TOPS (INT8)**: Render Slices image.
- Xe HPG Based Arch**
- AV1 Encoding & Decoding**
- ECC memory**
- 75W**: Half Height PCIe
- 150W**: Full Height PCIe
- Hardware based SR-IOV**
- Intel Xe Matrix Extensions**
- Up to 4 Xe Media Engines**: Server rack image.

intel.

Example: NVIDIA vGPU (MDEV)

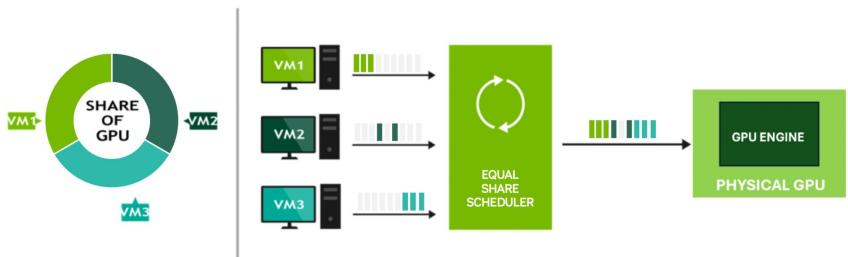


Example: NVIDIA vGPU Schedulers



Best Effort Round-Robin Scheduler

- VRAM Partitioned
- Time-sliced Sharing of Entire GPU



Equal-Share Scheduler



Fixed-Share Scheduler

Example: NVIDIA vGPU Schedulers



Best Effort Round-Robin Scheduler

- VRAM Partitioned
- Time-sliced Sharing of Entire GPU



Equal-Share Scheduler

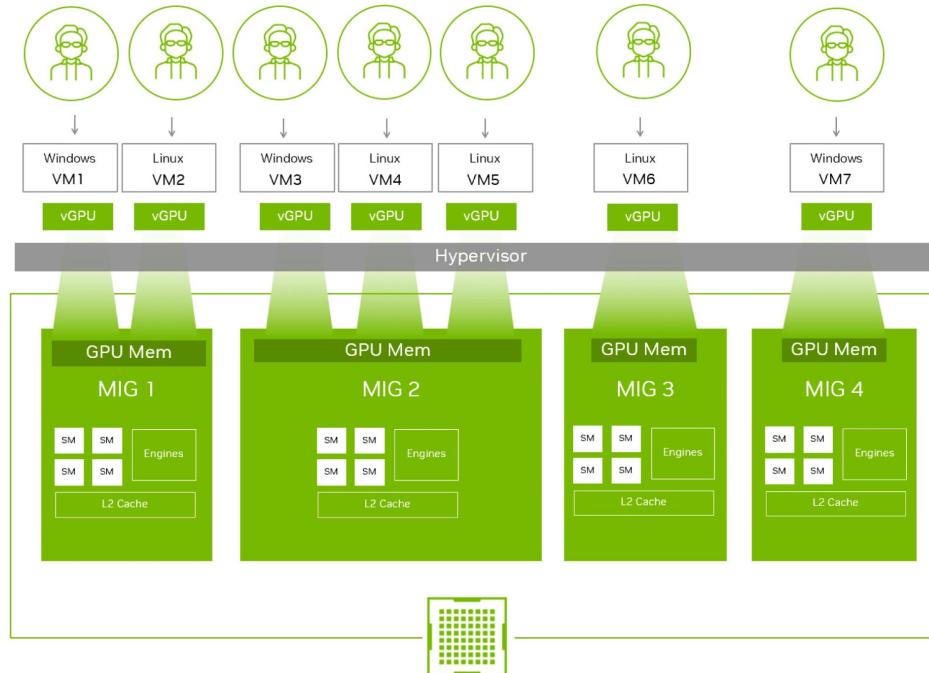


Fixed-Share Scheduler

Case Study: NVIDIA Multi-Instance GPU (MIG) (VFIO-PCI)



Case Study: MIG based Multi-Tenancy



RTX PRO 6000 Blackwell Server Edition

Shared GPU Type	Vendor(s)	Concurrent?	Isolation Level	Notes
vGPU (NVIDIA GRID)	NVIDIA	Yes	Medium	Licensed, great for VDI/AI-Infer.
SR-IOV / MxGPU	AMD, Intel	Yes	High	Hardware-dependent
Time-Sliced Sharing	NVIDIA, others	No	High	Simple, but not parallel
MIG (NVIDIA)	NVIDIA A100, etc.	Yes	Very High	Each instance fully isolated
API/Driver Level Sharing	All (via OS)	Yes	Low	Easiest for containers

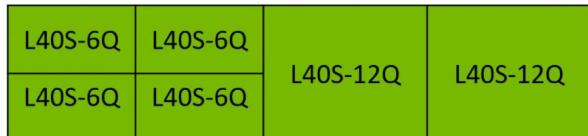
Understanding NVIDIA vGPUs

Series	Optimal Workload
Q-series	Virtual workstations for creative and technical professionals who require the performance and features of Quadro technology
B-series	Virtual desktops for business professionals and knowledge workers
A-series	App streaming or session-based solutions for virtual applications users

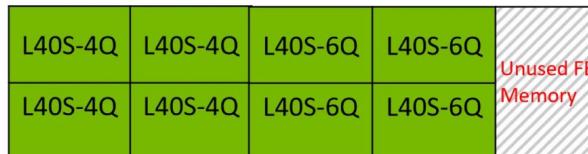
<https://docs.nvidia.com/vgpu/latest/grid-vgpu-user-guide/index.html>

NVIDIA Heterogeneous vGPU Profiles

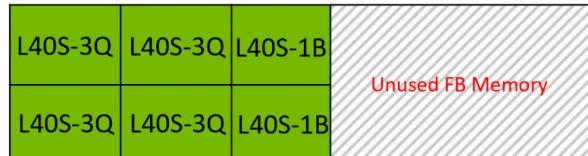
4 x L40S-6Q and 2 x L40S-12Q on GPU1:



4 x L40S-4Q and 4 x L40S-6Q on GPU1:

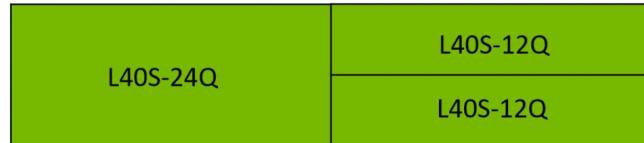


4 x L40S-3Q and 2 x L40S-1B on GPU1:

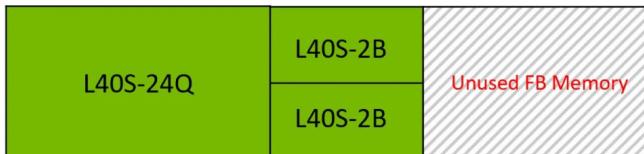


Physical GPU 1

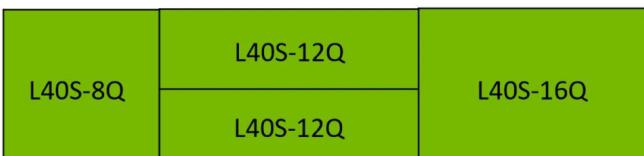
1 x L40S-24Q and 2 x L40S-12Q on GPU2:



1 x L40S-24Q and 2 x L40S-2B on GPU2:



1 x L40S-8Q, 2 x L40S-12Q and 1 x L40S-16Q on GPU2:



Physical GPU 2

Physical Host

Example: NVIDIA L40S GPU Card

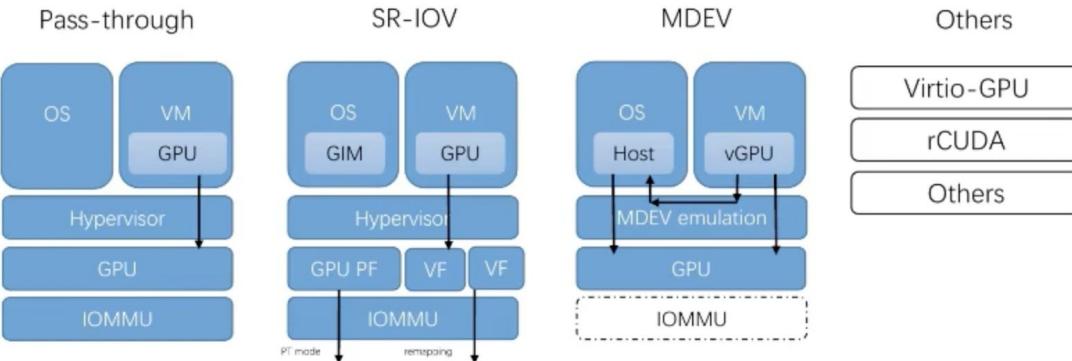
Designing & Deploying GPU-Enabled IaaS Cloud

Popular GPU Vendors (Currently)

	NVIDIA	AMD	Intel
Platform	CUDA	ROCM	oneAPI
Virtualisation	vGPU (GRID)	MxGPU (SR-IOV)	Basic/Flex
Current Standing	Market Leader	Growing Contender	Catching up
Ecosystem	PyTorch, TensorFlow...	Improving support in PyTorch, etc.	Laggard

Other contenders: Apple (Metal), Qualcomm (Adreno)...

GPU Usage & Virtualisation



GPU Virtualisation

Feature	MDEV	SR-IOV	VFIO Passthrough
Sharing	Yes	Yes	No
Uses VFIO	Yes (<code>vfio_mdev</code>)	Yes (<code>vfio-pci</code>)	Yes (<code>vfio-pci</code>)
Granularity	Fine-grained virtual functions	Hardware-based VFs	Entire physical device
Device support	Software-defined (via driver)	Hardware-defined	Full passthrough
Needs IOMMU	Yes	Yes	Yes
Examples	NVIDIA vGPU	AMD MxGPU, Intel Flex	Full GPU passthrough, NVIDIA MIG
Guest Driver	Vendor vGPU driver	Vendor driver	Vendor driver



Supported KVM Hypervisor: NVIDIA GPUs

- Red Hat Enterprise Linux 9 & 8 with KVM hypervisor
- Ubuntu 24.04 LTS and 22.04 LTS with KVM hypervisor

<https://docs.nvidia.com/vgpu/latest/product-support-matrix/index.html>

GPU IaaS Clouds Use-Cases

AI / Machine Learning / Deep Learning

Model training, inferencing, Computer Vision, NLP...

Virtual Desktops (VDI w/GPU Acceleration)

Access GPU-intensive apps remotely

Cloud Gaming & Streaming

Game and video streaming services

Graphics & Media

Video and 3D Rendering, Production

High Performance Computing (HPC)

Scientific simulation, fluid dynamics, molecular modeling...

World Simulation

Industrial simulations, urban planning, metaverse apps

Data Analytics & Big Data

GPU accelerated analytics

Cryptocurrency / Blockchain

GPU-based mining, blockchain compute



Picking the appropriate GPUs Example

Shared GPU Workloads: (vGPU/MxGPU/VF SR-IOV)

- NVIDIA L2
- NVIDIA L4
- NVIDIA L20
- NVIDIA L40
- NVIDIA L40S
- NVIDIA A2
- NVIDIA A10
- NVIDIA A16
- NVIDIA A40
- Tesla V100S, Tesla V100
- AMD MI210X, MI300X, MI350X
- Intel DC Flex Series...

More at:

<https://www.nvidia.com/en-us/data-center/graphics-cards-for-virtualization/>

https://instinct.docs.amd.com/projects/virt-drv/en/latest/userguides/Getting_started_with_MxGPU.html

<https://www.intel.com/content/www/us/en/products/details/dissecrete-gpus/data-center-gpu/flex-series.html>

AI Workloads:

- NVIDIA H100
- NVIDIA H800
- NVIDIA A800
- NVIDIA A40
- NVIDIA A30
- NVIDIA A10
- NVIDIA A16
- NVIDIA A2
- NVIDIA L4
- NVIDIA L40
- NVIDIA L40S
- NVIDIA V100, ...

More at:

<https://docs.nvidia.com/ai-enterprise/4.1/pdf/nvidia-ai-enterprise-release-notes.pdf>

Considering Passthrough GPUs

- Any workload requiring large VRAM
 - High-end AI/ML training
 - 3D work or game development
- Latency-Sensitive or Real-Time Workloads
 - Video Transcoding
 - Robotics Control and Industrial Automation
- Security and Isolation needs
- GPU Virtualisation not available

Considering Shared GPUs

- VDI or Desktop as a Service (Graphics Acceleration)
- Moderate-workloads:
 - VRAM intensive AI/ML Inferencing
 - Cloud gaming or media streaming
 - SaaS/Remote apps requiring Short-Medium GPU Bursts
 - Dev/Testing or CI/CD for GPU based Apps
 - Education & Learning Apps, such as Jupyter notebooks
- Media Transcoding and Encoding at scale
- Workloads requiring high-availability and uptime

Challenges & Future Work

Challenges & Future Work

- Accessibility of GPU-based hosts for feature development and QA
 - Test vGPU, MIG and passthrough: Nvidia, AMD, Intel...
 - Test Live VM Migration for vGPU case
- Validate and improve built-in GPU auto-discovery and domain transformer scripts against different GPU cards/models
- Improve feature implementation and aim for production readiness with ACS 4.22.
- GPU Metrics & Expand support for other hypervisors.

Thank You

#CSIndiaUG2025