cloud STACK | COLLABORATION CONFERENCE

GPU as a first class feature

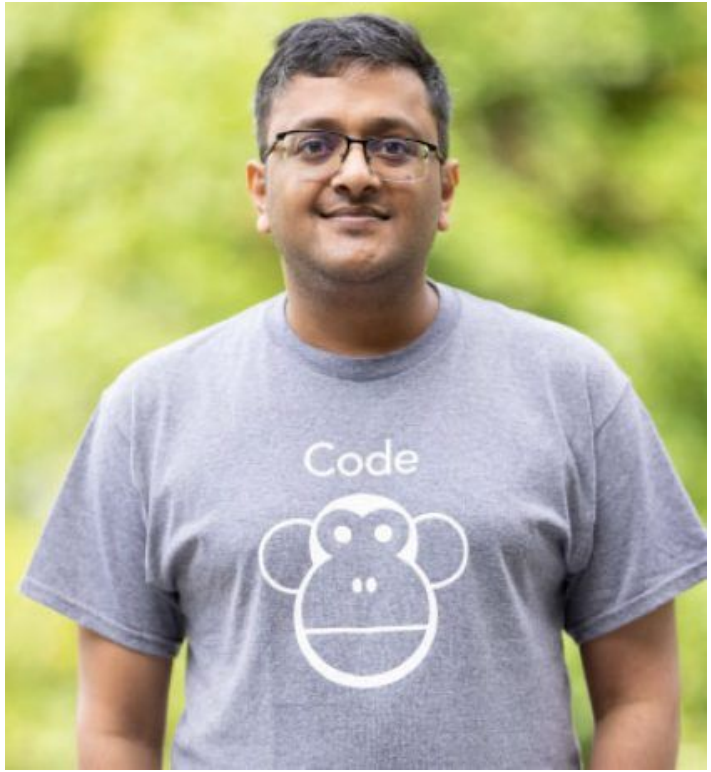
A proposal and an appeal for help

Vishesh Jindal
Software Engineer @ ShapeBlue

November 20 - 22, 2024 | Madrid, Spain

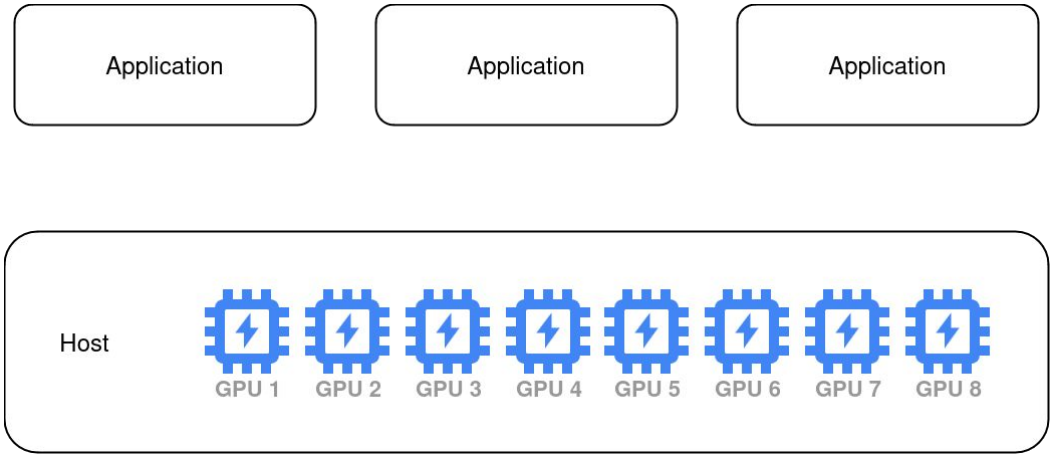
About me

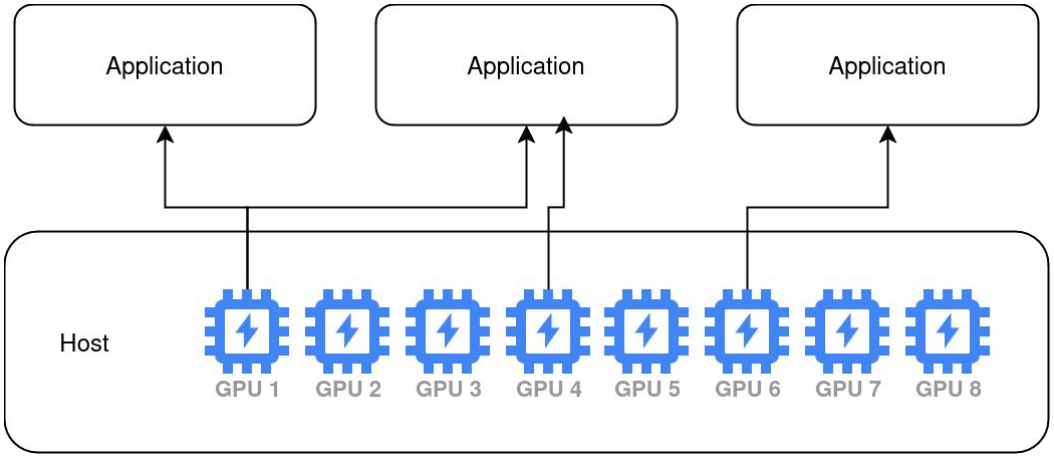
- Contributing to ACS since 2023
- Committer @ Apache CloudStack since 2024
- Software Engineer @ ShapeBlue
- Worked as a Backend & Infrastructure engineer in the past
- Experience managing and administering kubernetes, AWS, PostgreSQL in production
- Passionate about cloud infrastructure, security, and AI

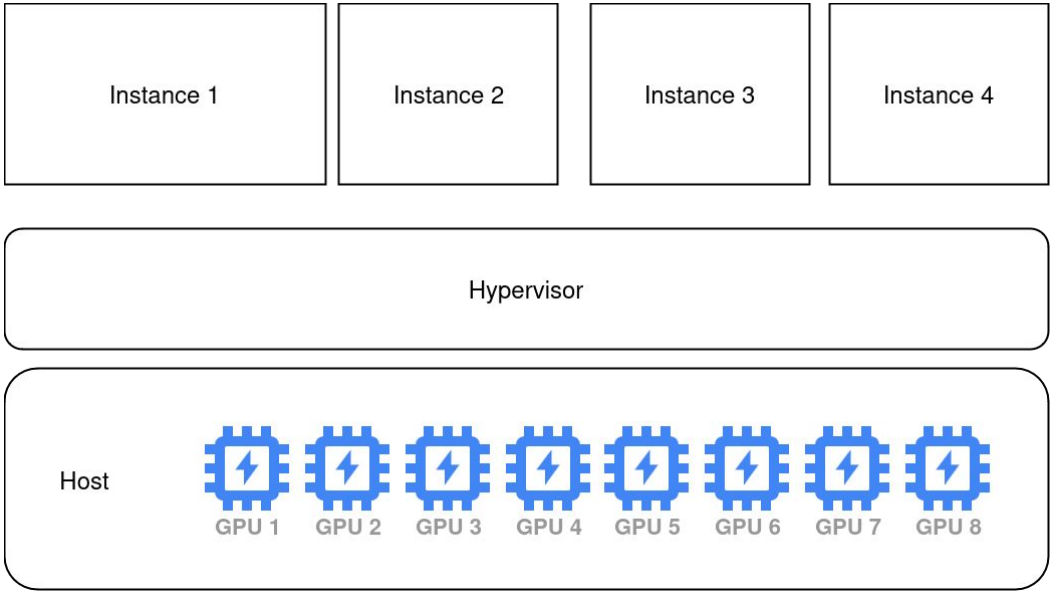


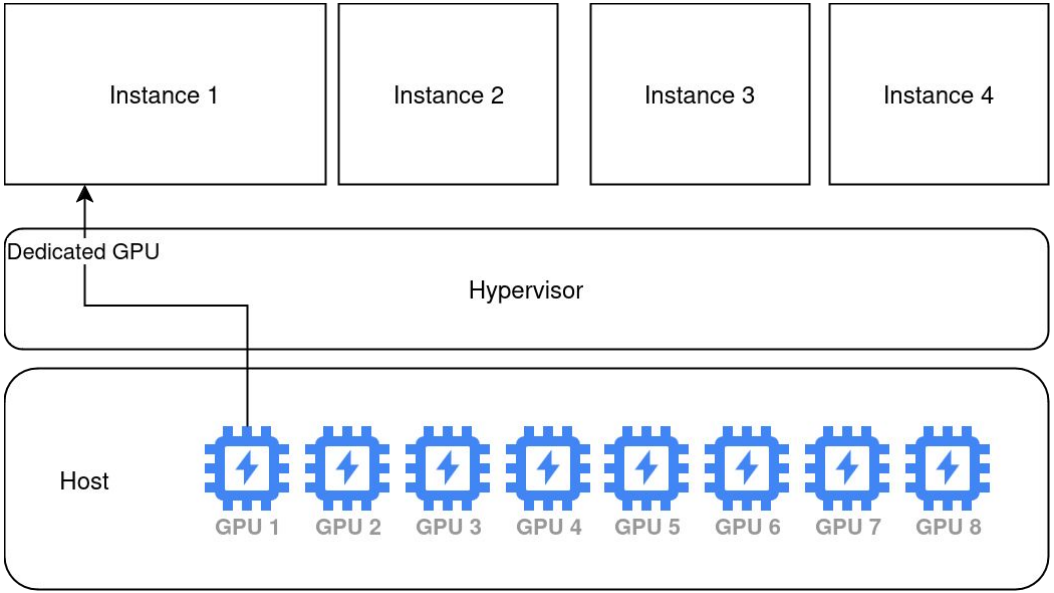
What's in it for you?

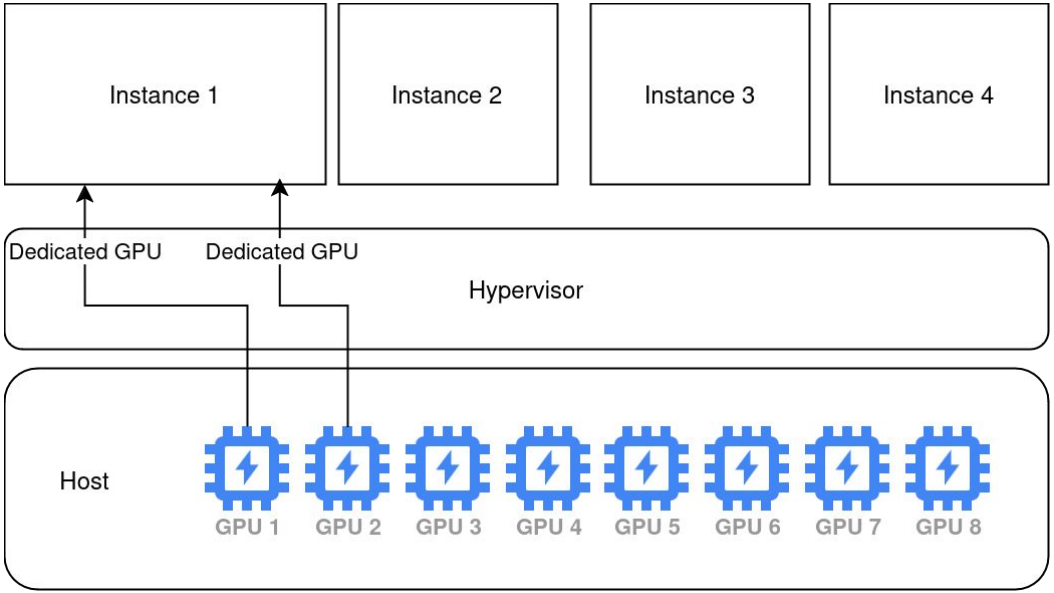
How are GPUs used in the cloud?

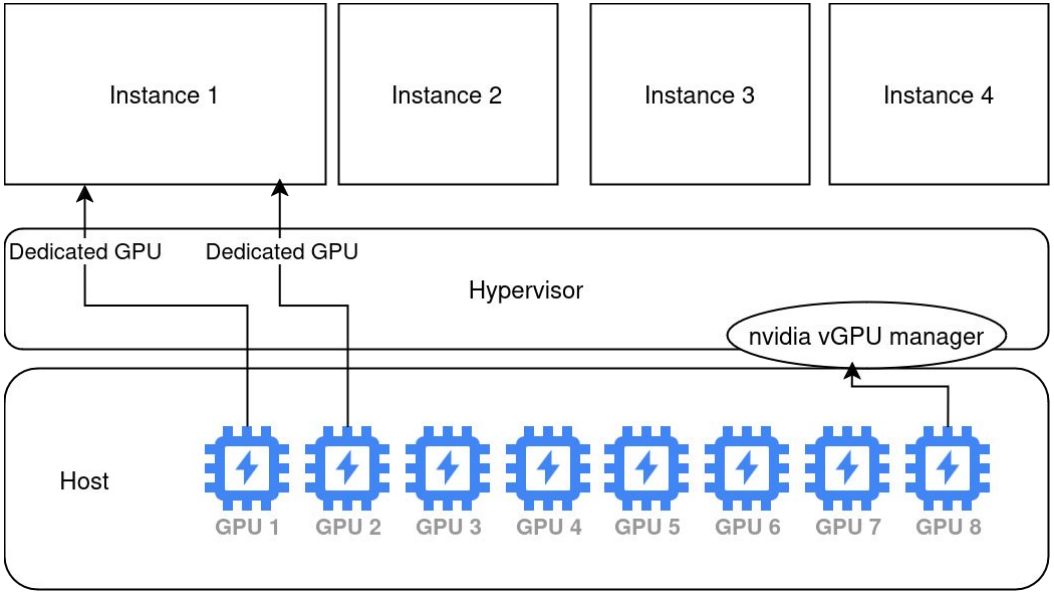

Application
Application
Application
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

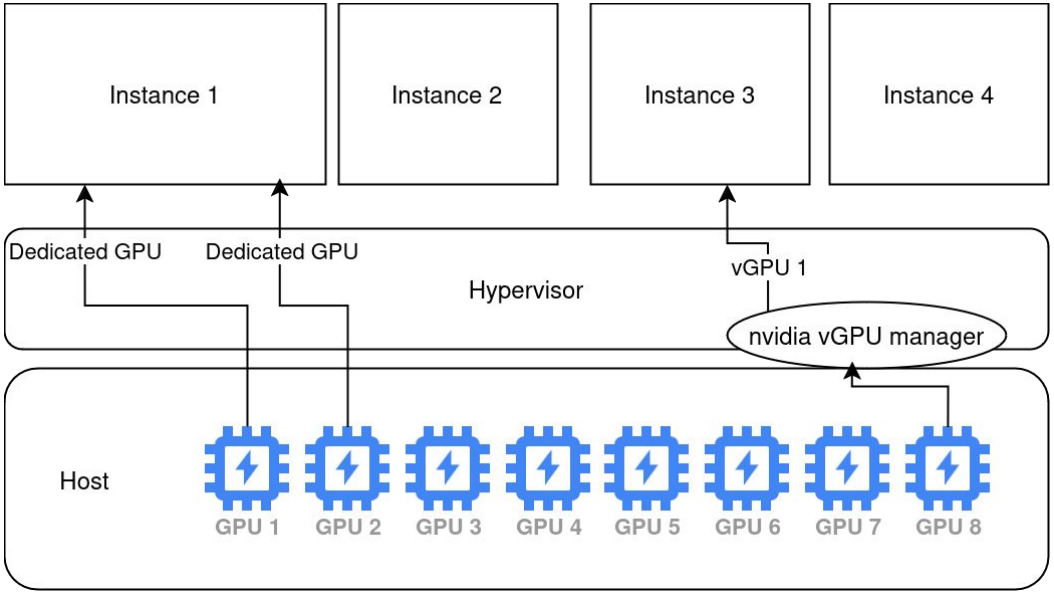

Application
Application
Application
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

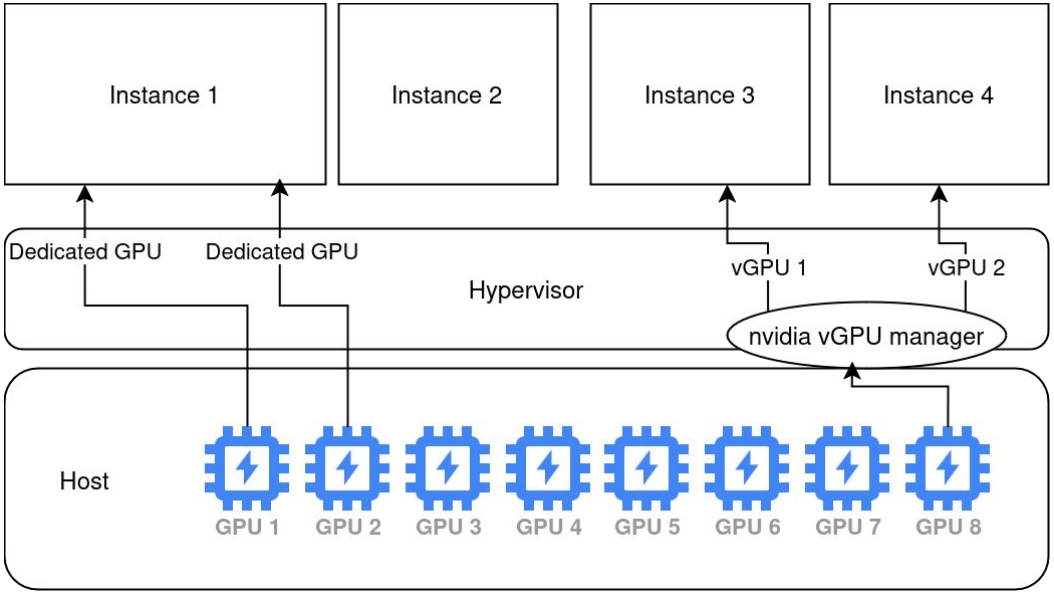

Instance 1
Instance 2
Instance 3
Instance 4
Hypervisor
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

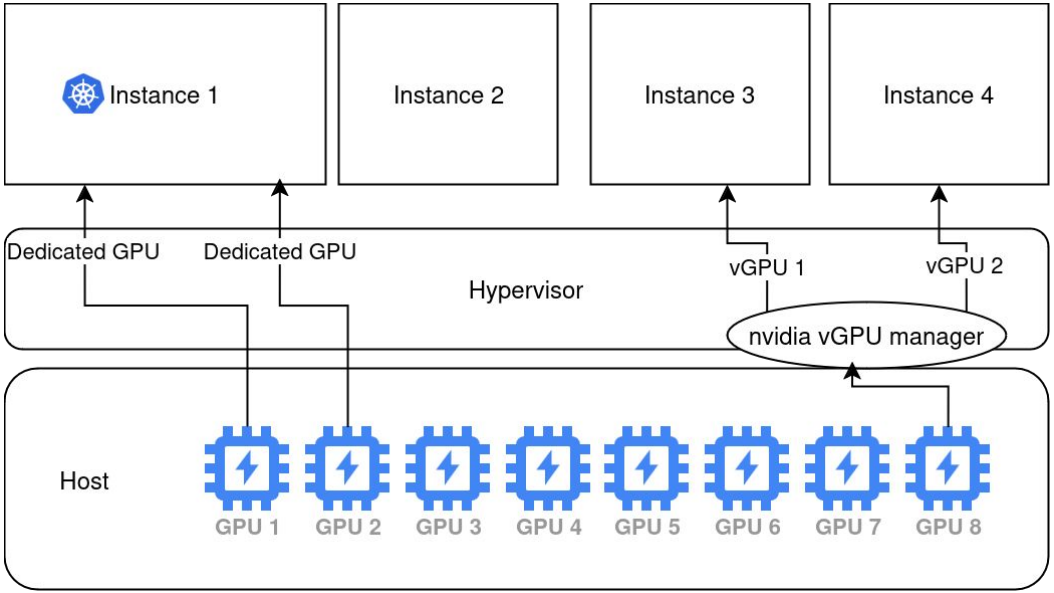

Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Hypervisor
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

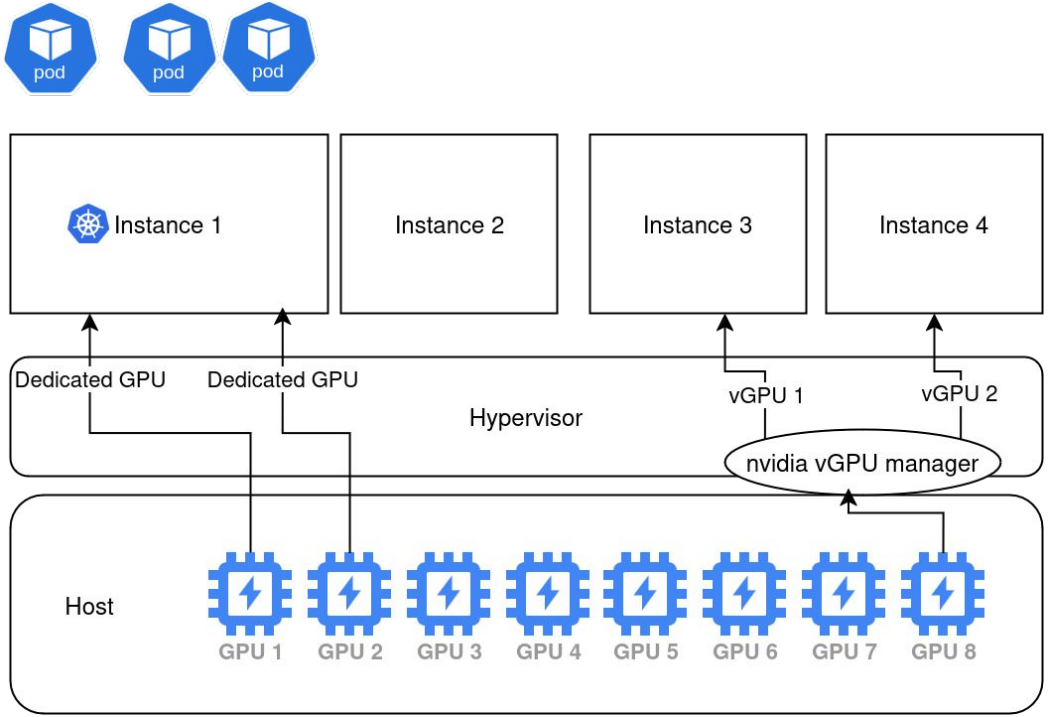

Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
Hypervisor
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

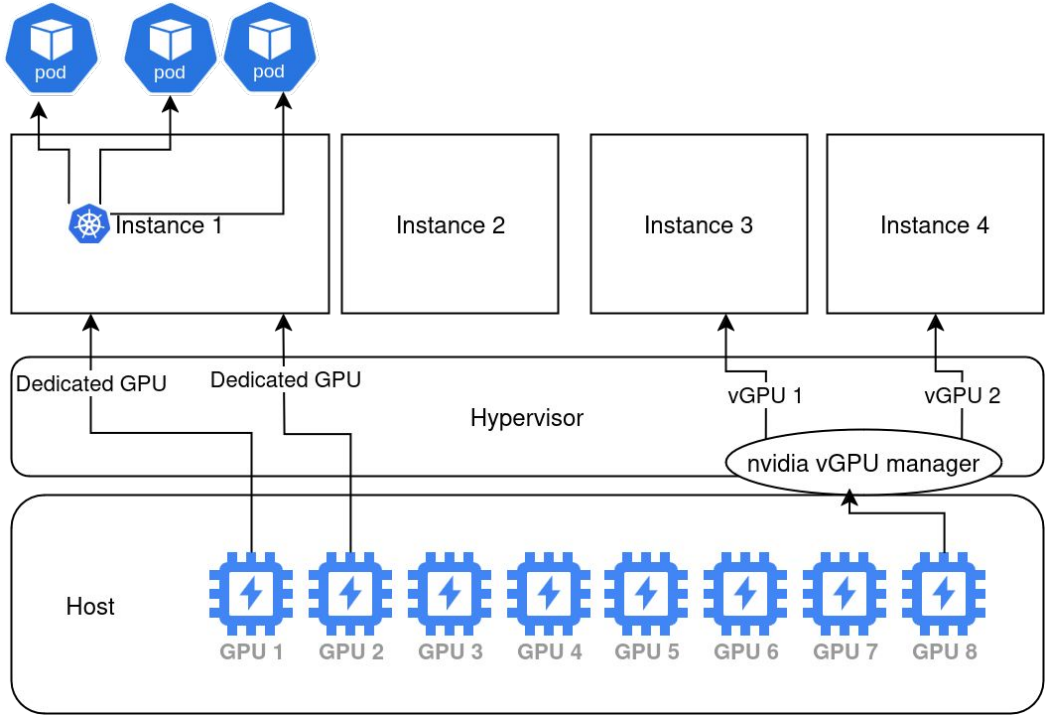

Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
Hypervisor
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8


Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
vGPU 1
Hypervisor
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

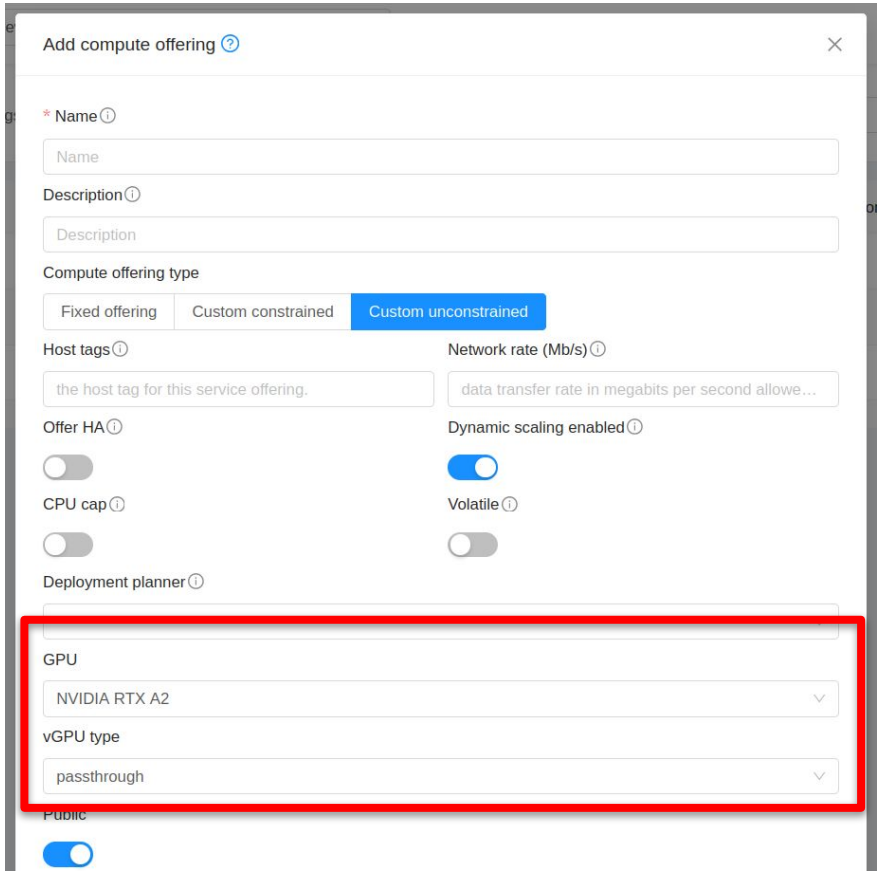

Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
vGPU 1
vGPU 2
Hypervisor
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8


Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
vGPU 1
vGPU 2
Hypervisor
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8


pod
pod
pod
Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
vGPU 1
vGPU 2
Hypervisor
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8


pod
pod
pod
Instance 1
Instance 2
Instance 3
Instance 4
Dedicated GPU
Dedicated GPU
Hypervisor
vGPU 1
vGPU 2
nvidia vGPU manager
Host
GPU 1
GPU 2
GPU 3
GPU 4
GPU 5
GPU 6
GPU 7
GPU 8

Current State of GPU support


Add compute offering
* Name
Name
Description
Description
Compute offering type
Fixed offering
Custom constrained
Custom unconstrained
Host tags
the host tag for this service offering.
Network rate (Mb/s)
data transfer rate in megabits per second allowe...
Offer HA
Dynamic scaling enabled
CPU cap
Volatile
Deployment planner
GPU
NVIDIA RTX A2
vGPU type
passthrough
Public

Workarounds with KVM

Configuration

```
[root@helium ~]# lspci -nn|grep NVIDIA
61:00.0 3D controller [0302]: NVIDIA Corporation GV100GL [Tesla V100 SXM2 32GB
62:00.0 3D controller [0302]: NVIDIA Corporation GV100GL [Tesla V100 SXM2 32GB
89:00.0 3D controller [0302]: NVIDIA Corporation GV100GL [Tesla V100 SXM2 32GB
8a:00.0 3D controller [0302]: NVIDIA Corporation GV100GL [Tesla V100 SXM2 32GB
[root@helium ~]#
```

```
[root@helium ~]# cat /etc/modprobe.d/vfio.conf
options vfio-pci ids=10de:1db5,10de:1db5,10de:1db5,10de:1db5
```

Source: https://lab.piszki.pl/cloudstack-kvm-and-running-vm-with-vgpu/

Additional libvirt XML for device passthrough

```
<devices>
   <hostdev mode='subsystem' type='pci' managed='yes'>
      <driver name='vfio'/>
      <source>
         <address domain='0x0000' bus='0x61' slot='0x00' function='0x0'/>
      </source>
     <alias name='nvidia0'/>
        <address type='pci' domain='0x0000' bus='0x00' slot='0x00' function='0x0'/>
   </hostdev>
   <hostdev mode='subsystem' type='pci' managed='yes'>
      <driver name='vfio'/>
      <source>
         <address domain='0x0000' bus='0x62' slot='0x00' function='0x0'/>
      </source>
     <alias name='nvidia1'/>
        <address type='pci' domain='0x0000' bus='0x00' slot='0x00' function='0x1'/>
   </hostdev>
</devices>
<devices>
   <hostdev mode='subsystem' type='pci' managed='yes'>
      <driver name='vfio'/>
      <source>
         <address domain='0x0000' bus='0x89' slot='0x00' function='0x0'/>
      </source>
     <alias name='nvidia2'/>
        <address type='pci' domain='0x0000' bus='0x00' slot='0x00' function='0x2'/>
   </hostdev>
   <hostdev mode='subsystem' type='pci' managed='yes'>
      <driver name='vfio'/>
      <source>
         <address domain='0x0000' bus='0x8a' slot='0x00' function='0x0'/>
      </source>
     <alias name='nvidia3'/>
        <address type='pci' domain='0x0000' bus='0x00' slot='0x00' function='0x3'/>
   </hostdev>
</devices>
```

Source: https://lab.piszki.pl/cloudstack-kvm-and-running-vm-with-vgpu/

Using extraconfig in the template

```
[root@sm101 ~]# cmk
Apache CloudStack 🐵 CloudMonkey 6.0.0
Report issues: https://github.com/apache/cloudstack-cloudmonkey/issues

(localcloud) 🐵 > update template id=d693aa4c-70e6-4f2e-9d6b-f6c3cd10a9b5 extraconfig=%3Cdevices%3E%0A%20%20%20
0A%20%20%20%20%20%20%3Cdriver%20name%3D%27vfio%27%2F%3E%0A%20%20%20%20%20%20%3Csource%3E%0A%20%20%20%20%20%20%2
%20function%3D%270x0%27%2F%3E%0A%20%20%20%20%20%20%3C%2Fsource%3E%0A%20%20%20%20%20%3Calias%20name%3D%27nvidia0
0x0000%27%20bus%3D%270x00%27%20slot%3D%270x00%27%20function%3D%270x0%27%2F%3E%0A%20%20%20%3C%2Fhostdev%3E%0A%20
27%3E%0A%20%20%20%20%20%20%3Cdriver%20name%3D%27vfio%27%2F%3E%0A%20%20%20%20%20%20%3Csource%3E%0A%20%20%20%20%2
x00%27%20function%3D%270x0%27%2F%3E%0A%20%20%20%20%20%20%3C%2Fsource%3E%0A%20%20%20%20%20%3Calias%20name%3D%27n
%3D%270x0000%27%20bus%3D%270x00%27%20slot%3D%270x00%27%20function%3D%270x1%27%2F%3E%0A%20%20%20%3C%2Fhostdev%3E
27yes%27%3E%0A%20%20%20%20%20%20%3Cdriver%20name%3D%27vfio%27%2F%3E%0A%20%20%20%20%20%20%3Csource%3E%0A%20%20%2
3D%270x00%27%20function%3D%270x0%27%2F%3E%0A%20%20%20%20%20%20%3C%2Fsource%3E%0A%20%20%20%20%20%3Calias%20name%
domain%3D%270x0000%27%20bus%3D%270x00%27%20slot%3D%270x00%27%20function%3D%270x2%27%2F%3E%0A%20%20%20%3C%2Fhost
ed%3D%27yes%27%3E%0A%20%20%20%20%20%20%3Cdriver%20name%3D%27vfio%27%2F%3E%0A%20%20%20%20%20%20%3Csource%3E%0A%2
0slot%3D%270x00%27%20function%3D%270x0%27%2F%3E%0A%20%20%20%20%20%20%3C%2Fsource%3E%0A%20%20%20%20%20%3Calias%2
%27%20domain%3D%270x0000%27%20bus%3D%270x00%27%20slot%3D%270x00%27%20function%3D%270x3%27%2F%3E%0A%20%20%20%3C%
```

Source: https://lab.piszki.pl/cloudstack-kvm-and-running-vm-with-vgpu/

Using libvirt hooks

```
class GpuDeviceAdder {
    String gpuXml = """
    <devices>
        <hostdev mode='subsystem' type='pci' managed='yes'>
            <driver name='vfio'/>
            <source>
                <address domain='0x0000' bus='0x61' slot='0x00' function='0x0'/>
            </source>
            <alias name='nvidia0'/>
            <address type='pci' domain='0x0000' bus='0x00' slot='0x00' function='0x0'/>
        </hostdev>
    </devices>
    """

    String transform(Object logger, String xml) {
        def vmDef = new XmlParser().parseText(xml)

        // Parse GPU XML
        def gpuDevices = new XmlParser().parseText(gpuXml)

        // Append GPU devices to the VM definition
        gpuDevices.hostdev.each { gpuDevice ->
            vmDef.devices[0].append(gpuDevice)
        }

        // Return updated XML definition
        return XmlUtil.serialize(vmDef)
    }

```

Source: https://gist.github.com/rajujith/f3b3854ed77f2cab8dc4fb5e3ee260c4

Is that good enough?

What's missing?

User

"I need to deploy instances with multiple GPUs to efficiently run demanding AI and ML workloads."

Cloud Service Provider (CSP)

"I want to offer instances with multiple GPUs to my users, complete with usage tracking and customizable limits, across different hypervisors."

Capabilities

- ❑ Discovery and Inventory Management
- ❑ Groupings and Offerings
- ❑ Allocation and Assignment
- ❑ Usage and Limits

Capabilities

- ➢ **Discovery and Inventory Management**
- ❑ Groupings and Offerings
- ❑ Allocation and Assignment
- ❑ Usage and Limits

GPUs on a Host

Default view

Create | Admin User

/ Hosts / Host 1 | Refresh

Host 1

Routing | Simulator | 4.20.0.0-SNAPSHOT

Status
Up

Resource state
Enabled

ID
4385d62d-c55e-4838-a680-d9d956fa1ee4

IP address
172.16.15.21

CPU
4 CPU x 8.00 Ghz | x86_64
0.00% Used
15.62% Allocated

Memory
8.00 GB Memory
0.00% Used
62.50% Allocated

Network
RX 32.00 MB | TX 16.00 MB

Cluster
C0

Details | GPU | Events | Comments

+ Add GPU | Discover GPU

DeviceId	Name	Type	Memory	State	Action
61:00.0	Nvidia A100	Passthrough	40GB	Used	
62:00.0	Nvidia A100	Passthrough	40GB	Used	
69:00.0	Nvidia A100	Passthrough	40GB	Used	
6a:00.0	Nvidia A100	Passthrough	40GB	Used	
6b:00.0	Nvidia A100	Passthrough	40GB	Available	
6c:00.0	Nvidia A100	Passthrough	40GB	Available	
6d:00.0	Nvidia A100	vGPU	40GB	Available	
6e:00.0	Nvidia A100	vGPU	40GB	Available	

1

GPUs on a Host

Default view

Create

/ Hosts Refresh All Metrics

Add host +

Search

Name	State	Resource state	IP Address	Hypervisor	Instances	Power state	Version	CPU cores	CPU total	CPU used	CPU allocated	Memory total	Memory used	Memory allocated	GPU devices	GPU devices used	GPU memory total	GPU memory used	Networ read
Host 1	Up	Enabled	172.16.15.21	Simulator	3 / 3	Disabled	4.21.0.0	4	32.00 Ghz	0.00 Ghz	5.00 Ghz	8.00 GB	0.00 GB	5.00 GB	8	4	320 GB	160 GB	0.03 G
Host 2	Up	Enabled	172.16.15.8	Simulator	0 / 0	Disabled	4.21.0.0	4	32.00 Ghz	0.00 Ghz	0.50 Ghz	8.00 GB	0.00 GB	0.50 GB	8	0	320 GB	0 GB	0.03 G
Host 3	Up	Enabled	172.16.15.23	Simulator	1 / 1	Disabled	4.21.0.0	4	32.00 Ghz	0.00 Ghz	3.00 Ghz	8.00 GB	0.00 GB	3.00 GB	0	0	0 GB	0 GB	0.03 G
Host 4	Up	Enabled	172.16.15.3	Simulator	0 / 0	Disabled	4.21.0.0	4	32.00 Ghz	0.00 Ghz	0.50 Ghz	8.00 GB	0.00 GB	1.00 GB	0	0	0 GB	0 GB	0.03 G

Showing 1-4 of 4 items 1 20 / page

Capabilities

- ✓ Discovery and Inventory Management
- ➢ **Groupings and Offerings**
- ❑ Allocation and Assignment
- ❑ Usage and Limits

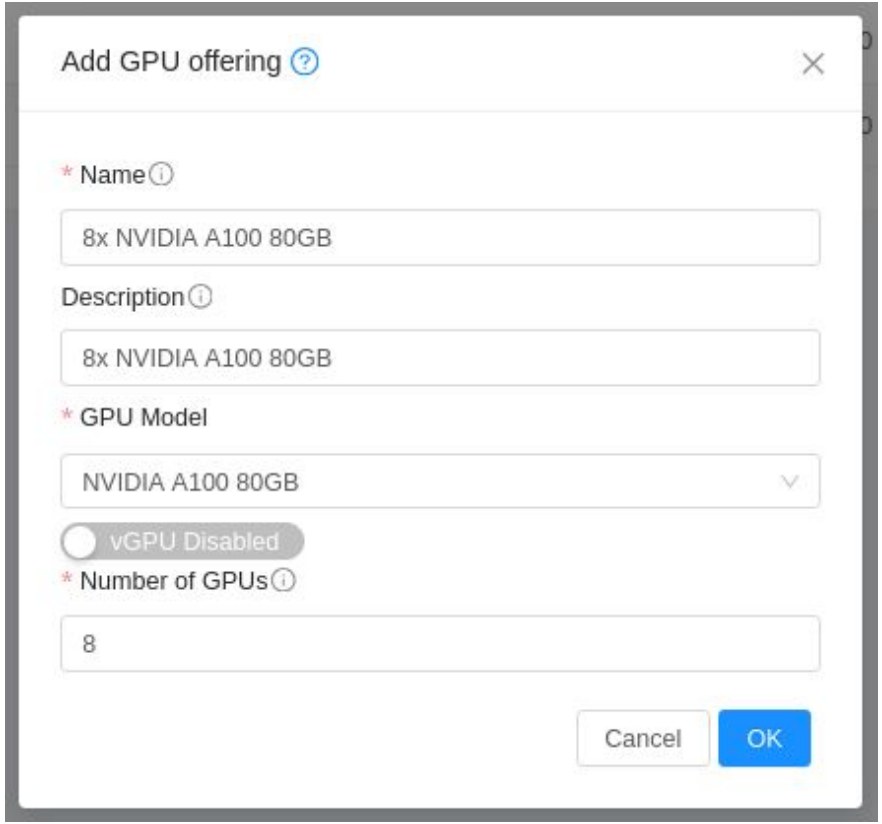
GPU Offerings

/ GPU Offerings Refresh Mine Add GPU offering + Search

Name	Description	State	Total memory	GPU Model	Count	Domain	Zone	Order
8x NVIDIA Tesla V100 32GB	NVIDIA Tesla V100	Active	256 GB	NVIDIA Tesla V100 32GB	8			
4x NVIDIA Tesla V100 32GB	NVIDIA Tesla V100	Active	128 GB	NVIDIA Tesla V100 32GB	4			
2x NVIDIA Tesla V100 32GB	NVIDIA Tesla V100	Active	64 GB	NVIDIA Tesla V100 32GB	2			
1x NVIDIA Tesla V100 32GB	NVIDIA Tesla V100	Active	32 GB	NVIDIA Tesla V100 32GB	1			

Showing 1-1 of 1 items 1 20 / page

Add GPU Offering



Capabilities

- ✓ Discovery and Inventory Management
- ✓ Groupings and Offerings
- ➢ Allocation and Assignment
- ❑ Usage and Limits

Deploying an Instance

Total 4 items < 1 > 10 / page

Override root disk offering

5 Data disk

Search

Disk offering	Disk size (in GB)	Min IOPS / Max IOPS
No thanks	-	-
Small	5 GB	-
Medium	20 GB	-
Large	100 GB	-
Custom	Custom disk size	-

6 GPU Offering

Enable GPU

Name	Total Memory	GPU Model
8x NVIDIA Tesla V100 32GB	256 GB	NVIDIA Tesla V100 32GB
4x NVIDIA Tesla V100 32GB	128 GB	NVIDIA Tesla V100 32GB
2x NVIDIA Tesla V100 32GB	64 GB	NVIDIA Tesla V100 32GB
1x NVIDIA Tesla V100 32GB	32 GB	NVIDIA Tesla V100 32GB

< 1 >

7 Networks

Simulator

OS type
CentOS 5.6 (64-bit)

CPU
1 CPU x 1.00 Ghz

Memory
1024 MB memory

Disk size (in GB)
5 GB (Root)

Networks
Test isolated (Default)

Template
CentOS 5.6 (64-bit) no GUI (Simulator)

Compute offering
Test

Zone
Sandbox-simulator

Service Offering with GPU

Add compute offering

* Name

Name

Description

Description

Compute offering type

Fixed offering | Custom constrained | Custom unconstrained

Host tags

the host tag for this service offering.

Network rate (Mb/s)

data transfer rate in megabits per second allowe...

Offer HA

Dynamic scaling enabled

CPU cap

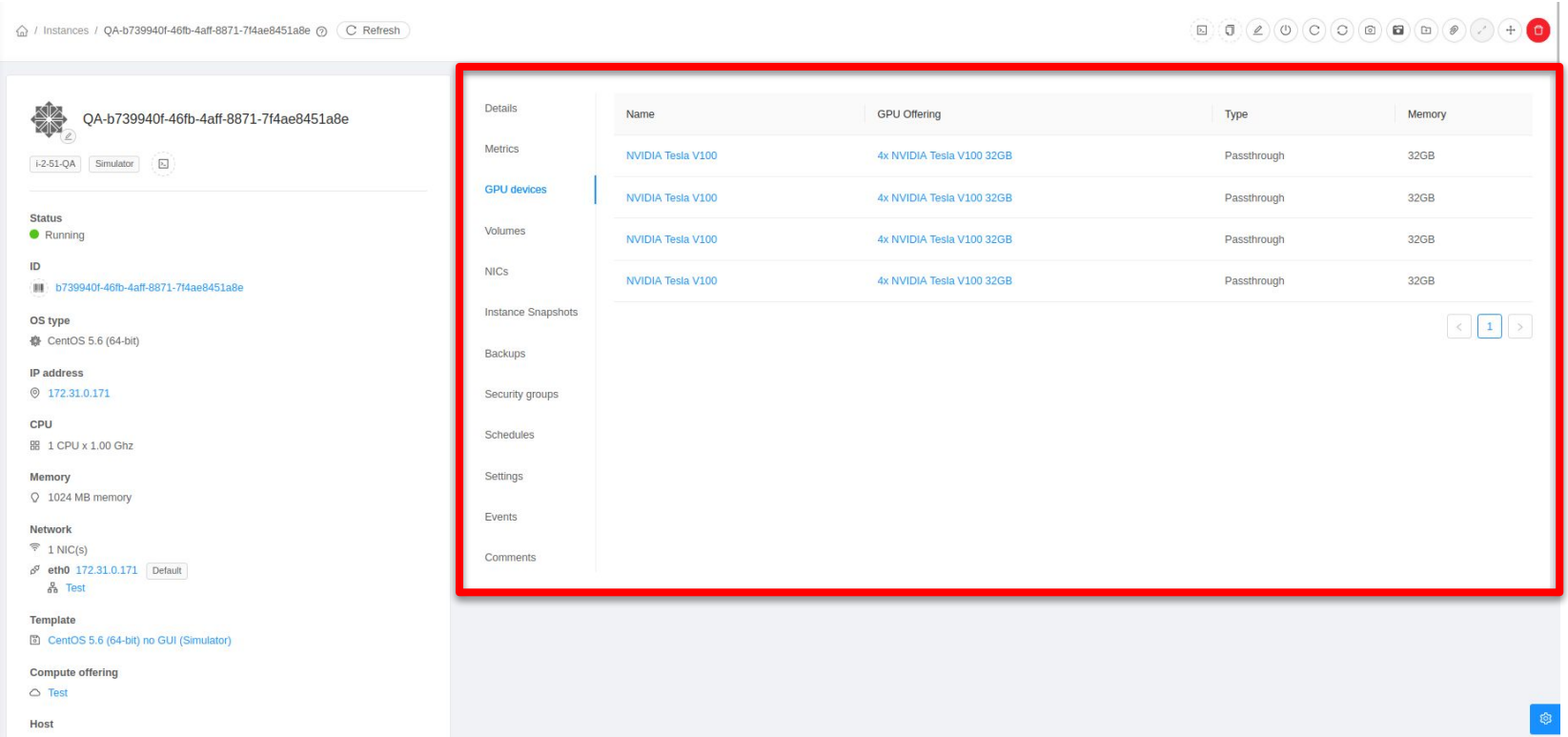
Volatile

Deployment planner

Attach GPU Offering

Name	Total Memory	GPU Model
8x NVIDIA Tesla V100 32GB	256 GB	NVIDIA Tesla V100 32GB
4x NVIDIA Tesla V100 32GB	128 GB	NVIDIA Tesla V100 32GB
2x NVIDIA Tesla V100 32GB	64 GB	NVIDIA Tesla V100 32GB
1x NVIDIA Tesla V100 32GB	32 GB	NVIDIA Tesla V100 32GB

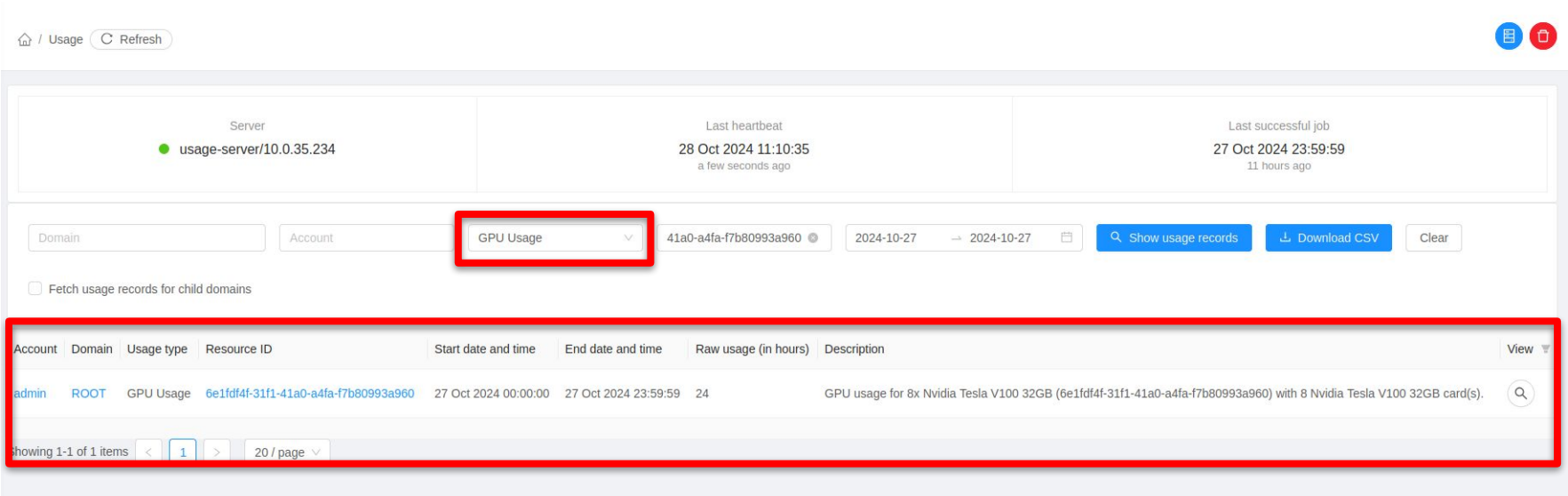
GPU details for an Instance



Capabilities

- ✓ Discovery and Inventory Management
- ✓ Groupings and Offerings
- ✓ Allocation and Assignment
- ➢ Usage and Limits

GPU Usage


/ Usage
Refresh
Server
usage-server/10.0.35.234
Last heartbeat
28 Oct 2024 11:10:35
a few seconds ago
Last successful job
27 Oct 2024 23:59:59
11 hours ago
Domain
Account
GPU Usage
41a0-a4fa-f7b80993a960
2024-10-27
2024-10-27
Show usage records
Download CSV
Clear
Fetch usage records for child domains
Account
Domain
Usage type
Resource ID
Start date and time
End date and time
Raw usage (in hours)
Description
View
admin
ROOT
GPU Usage
6e1fdf4f-31f1-41a0-a4fa-f7b80993a960
27 Oct 2024 00:00:00
27 Oct 2024 23:59:59
24
GPU usage for 8x Nvidia Tesla V100 32GB (6e1fdf4f-31f1-41a0-a4fa-f7b80993a960) with 8 Nvidia Tesla V100 32GB card(s).
Showing 1-1 of 1 items
1
20 / page

GPU Limits

Domain | Details | Limits | Configure limits | Settings | Events | Comments

Instance limits (Unlimited Available)

Used / Limit : 3 / Unlimited

CPU limits (Unlimited Available)

Used / Limit : 6 / Unlimited

Memory limits (MiB) (Unlimited Available)

Used / Limit : 6656 / Unlimited

GPU limits (Unlimited Available)

Used / Limit : 6 / Unlimited

GPU Memory limits (MiB) (Unlimited Available)

Used / Limit : 66560 / Unlimited

Primary storage limits (GiB) (Unlimited Available)

Used / Limit : 108 / Unlimited

Capabilities

- ✓ Discovery and Inventory Management
- ✓ Groupings and Offerings
- ✓ Allocation and Assignment
- ✓ Usage and Limits

This is just a proposal

Why hasn't the community built this yet?

Where do we go from here?

Questions?

Thank you!

#CSCollab24
@CloudStack