How can I give virtual GPU resources to my end users seamlessly?

OpenInfra Day France 2024

Sylvain Bauza
Red Hat

Sylvain [sil-vɛ̃] Bauza

Principal Software Engineer @ Red Hat

@sylvainbauza

IRC: bauzas

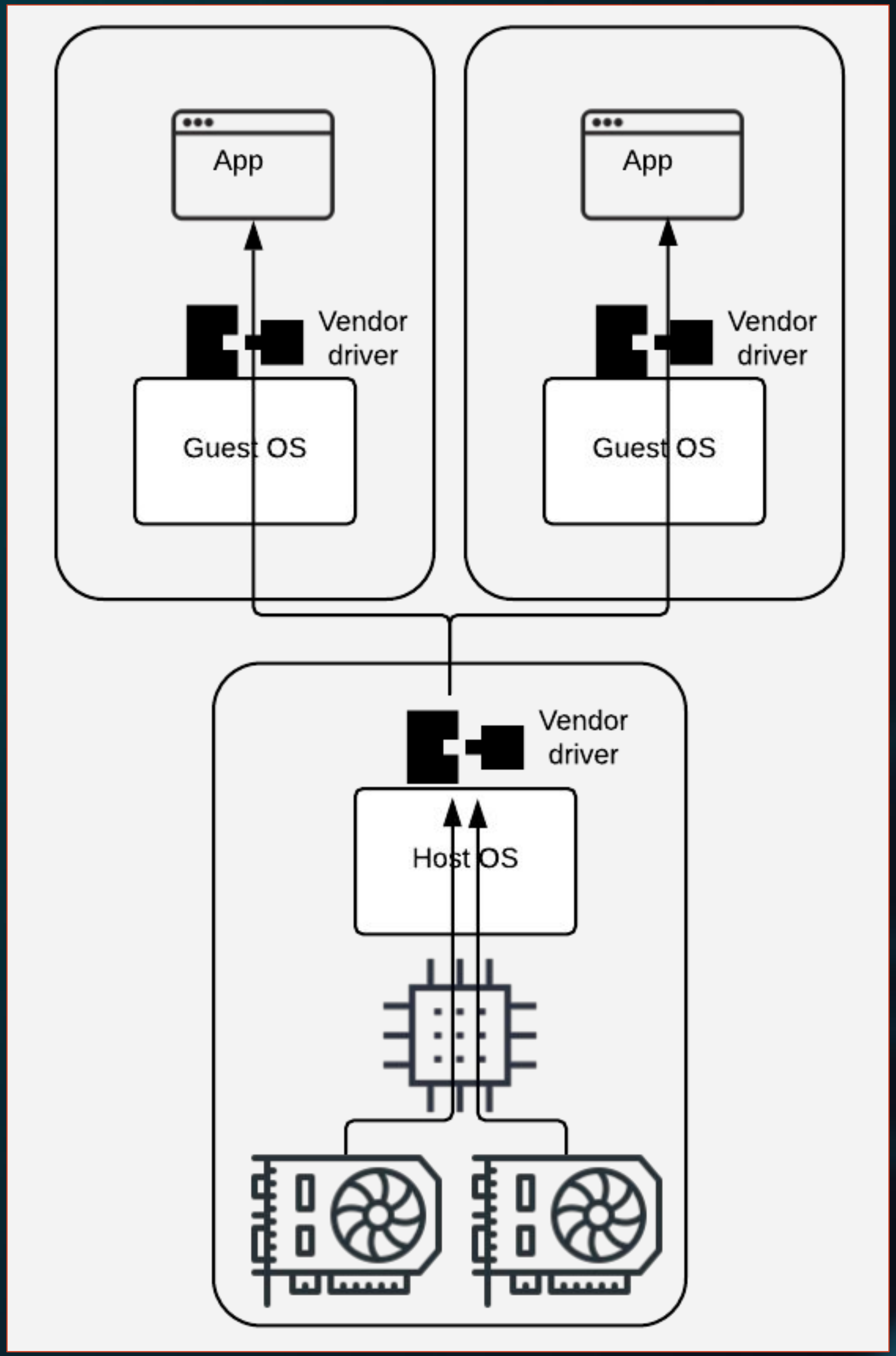
- Nova/Placement PTL
- Nova contributor since 2013
- Previously : Operator & DevOps

Virtual GPUs in Nova

How can I give virtual GPU resources to my end users seamlessly ?

OpenInfra Day France 2024

How this works ?



The kernel interface (VFIO-mdev)

```
|- [parent physical device]
 |--- Vendor-specific-attributes [optional]
 |--- [mdev_supported_types]
 |     |--- [<type-id>]
 |     |   |--- create
 |     |   |--- name
 |     |   |--- available_instances
 |     |   |--- device_api
 |     |   |--- description
 |     |   |--- [devices]
 |     |--- [<type-id>]
 |     |   |--- create
 |     |   |--- name
 |     |   |--- available_instances
 |     |   |--- device_api
 |     |   |--- description
 |     |   |--- [devices]
```

```
|- [parent phy device]
 |--- [$MDEV_UUID]
        |--- remove
        |--- mdev_type {link to its type}
        |--- vendor-specific-attributes [optional]
```

All of this is vendor specific !

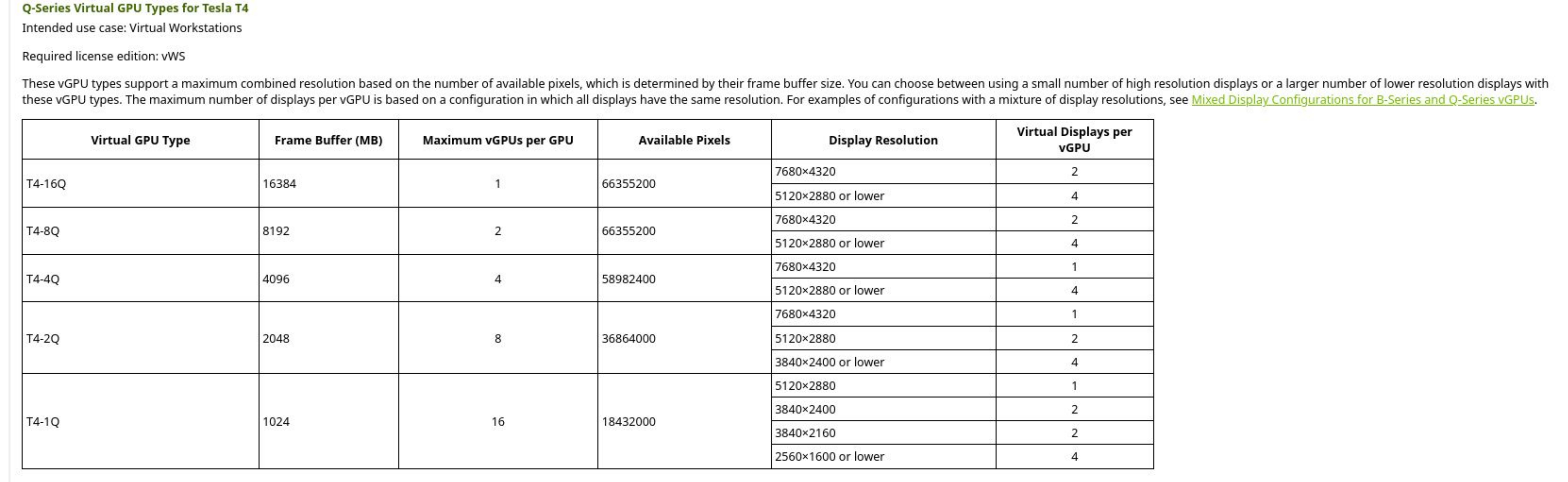
Proprietary or opensource
kernel module
(eg. nvidia.ko)

Usually licence-based
(eg. nvidia GRID & AIE)

Depending on the product line

Hardware capabilities (SR-IOV,
framebuffer dirty pages tracking...)

Mediated device types (aka. GPU profiles)



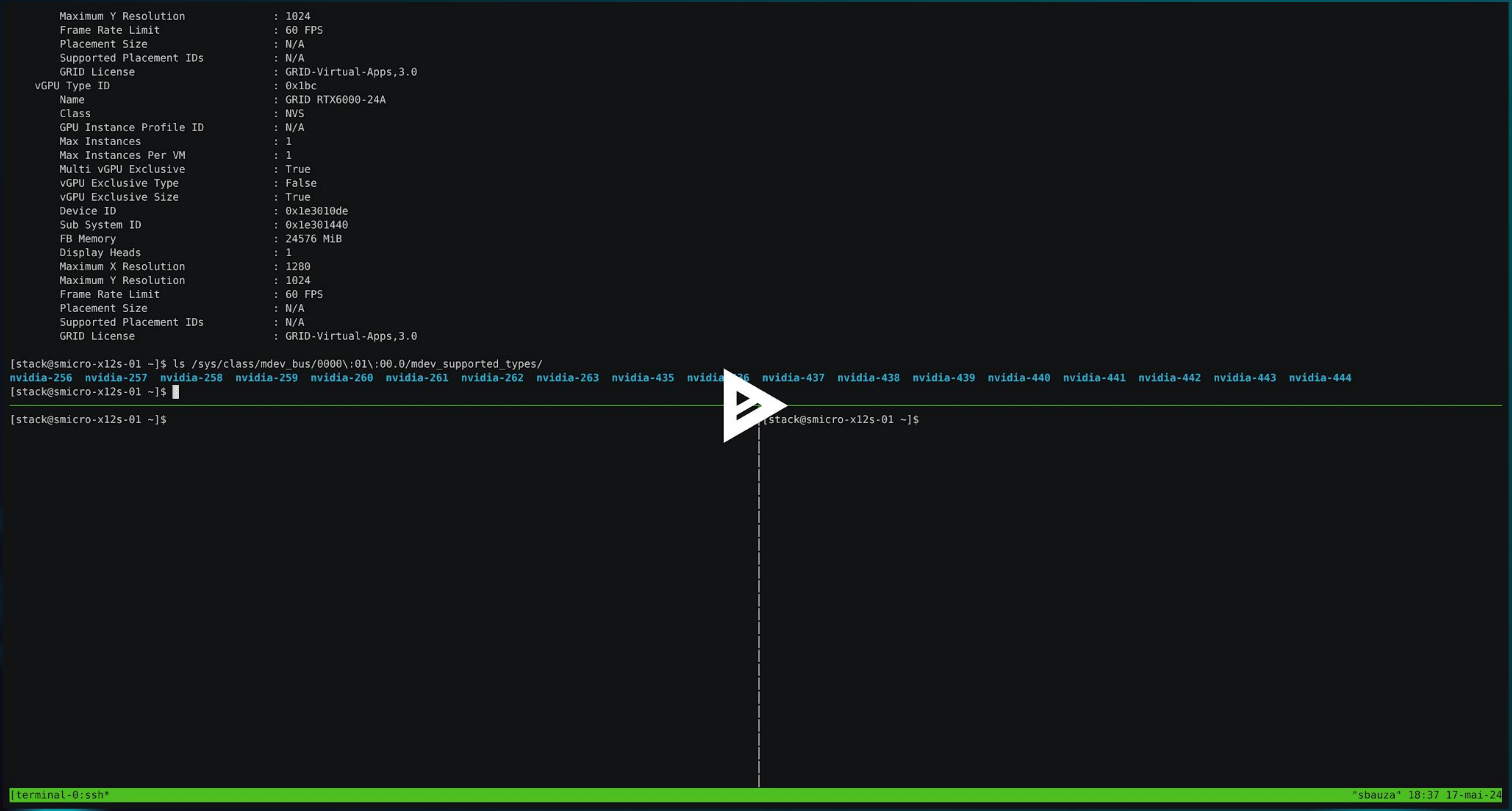
Q-Series Virtual GPU Types for Tesla T4

Intended use case: Virtual Workstations

Required license edition: vWS

These vGPU types support a maximum combined resolution based on the number of available pixels, which is determined by their frame buffer size. You can choose between using a small number of high resolution displays or a larger number of lower resolution displays with these vGPU types. The maximum number of displays per vGPU is based on a configuration in which all displays have the same resolution. For examples of configurations with a mixture of display resolutions, see Mixed Display Configurations for B-Series and Q-Series vGPUs.

Virtual GPU Type	Frame Buffer (MB)	Maximum vGPUs per GPU	Available Pixels	Display Resolution	Virtual Displays per vGPU
T4-16Q	16384	1	66355200	7680×4320	2
				5120×2880 or lower	4
T4-8Q	8192	2	66355200	7680×4320	2
				5120×2880 or lower	4
T4-4Q	4096	4	58982400	7680×4320	1
				5120×2880 or lower	4
T4-2Q	2048	8	36864000	7680×4320	1
				5120×2880	2
				3840×2400 or lower	4
T4-1Q	1024	16	18432000	5120×2880	1
				3840×2400	2
				3840×2160	2
				2560×1600 or lower	4



```
        Maximum Y Resolution              : 1024
        Frame Rate Limit                  : 60 FPS
        Placement Size                    : N/A
        Supported Placement IDs           : N/A
        GRID License                      : GRID-Virtual-Apps,3.0
    vGPU Type ID                          : 0x1bc
        Name                              : GRID RTX6000-24A
        Class                             : NVS
        GPU Instance Profile ID           : N/A
        Max Instances                     : 1
        Max Instances Per VM              : 1
        Multi vGPU Exclusive              : True
        vGPU Exclusive Type               : False
        vGPU Exclusive Size               : True
        Device ID                         : 0x1e3010de
        Sub System ID                     : 0x1e301440
        FB Memory                         : 24576 MiB
        Display Heads                     : 1
        Maximum X Resolution              : 1280
        Maximum Y Resolution              : 1024
        Frame Rate Limit                  : 60 FPS
        Placement Size                    : N/A
        Supported Placement IDs           : N/A
        GRID License                      : GRID-Virtual-Apps,3.0

[stack@smicro-x12s-01 ~]$ ls /sys/class/mdev_bus/0000\:01\:00.0/mdev_supported_types/
nvidia-256  nvidia-257  nvidia-258  nvidia-259  nvidia-260  nvidia-261  nvidia-262  nvidia-263  nvidia-435  nvidia-436  nvidia-437  nvidia-438  nvidia-439  nvidia-440  nvidia-441  nvidia-442  nvidia-443  nvidia-444
[stack@smicro-x12s-01 ~]$
```

```
[stack@smicro-x12s-01 ~]$
```

```
[stack@smicro-x12s-01 ~]$
```

```
[terminal-0:ssh*]                                                                                "sbauza" 18:37 17-mai-24
```

How to configure it in Nova

You need a flavor…

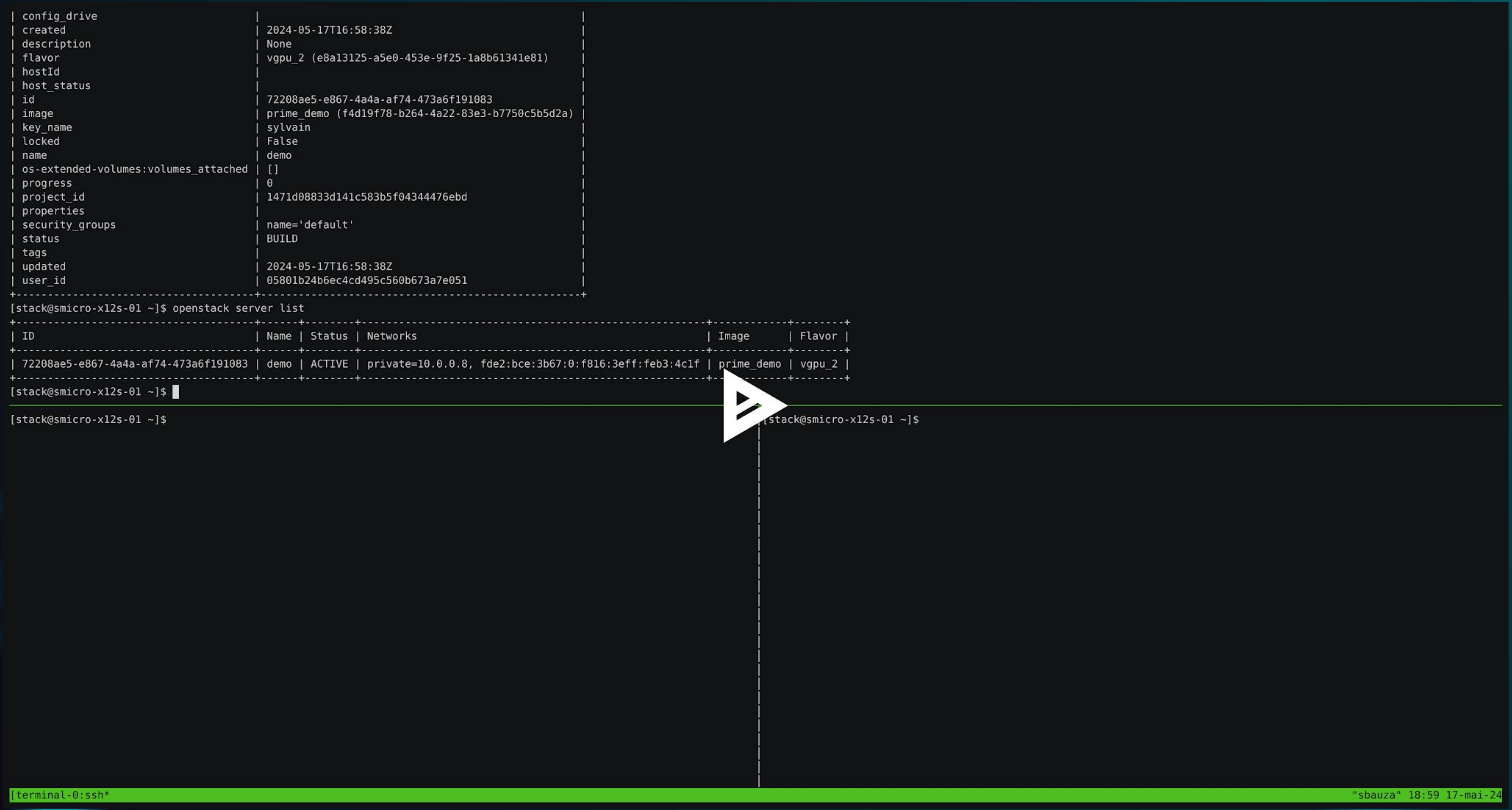
```
$ openstack flavor set myflavor --property
"resources:VGPU=1"
```

… and a compute node

```
[devices]
enabled_mdev_types = nvidia-256

[mdev_nvidia-256]
device_addresses = 0000:84:00.0,0000:85:00.0
```

https://docs.openstack.org/nova/latest/admin/virtual-gpu.html



```
| config_drive                        |                                                   |
| created                             | 2024-05-17T16:58:38Z                              |
| description                         | None                                              |
| flavor                              | vgpu_2 (e8a13125-a5e0-453e-9f25-1a8b61341e81)     |
| hostId                              |                                                   |
| host_status                         |                                                   |
| id                                  | 72208ae5-e867-4a4a-af74-473a6f191083              |
| image                               | prime_demo (f4d19f78-b264-4a22-83e3-b7750c5b5d2a) |
| key_name                            | sylvain                                           |
| locked                              | False                                             |
| name                                | demo                                              |
| os-extended-volumes:volumes_attached | []                                                |
| progress                            | 0                                                 |
| project_id                          | 1471d08833d141c583b5f04344476ebd                  |
| properties                          |                                                   |
| security_groups                     | name='default'                                    |
| status                              | BUILD                                             |
| tags                                |                                                   |
| updated                             | 2024-05-17T16:58:38Z                              |
| user_id                             | 05801b24b6ec4cd495c560b673a7e051                  |
+-------------------------------------+---------------------------------------------------+
[stack@smicro-x12s-01 ~]$ openstack server list
+--------------------------------------+------+--------+-------------------------------------------------------+------------+--------+
| ID                                   | Name | Status | Networks                                              | Image      | Flavor |
+--------------------------------------+------+--------+-------------------------------------------------------+------------+--------+
| 72208ae5-e867-4a4a-af74-473a6f191083 | demo | ACTIVE | private=10.0.0.8, fde2:bce:3b67:0:f816:3eff:feb3:4c1f | prime_demo | vgpu_2 |
+--------------------------------------+------+--------+-------------------------------------------------------+------------+--------+
[stack@smicro-x12s-01 ~]$
```

```
[stack@smicro-x12s-01 ~]$
```

```
[stack@smicro-x12s-01 ~]$
```

```
[terminal-0:ssh*                                                                            "sbauza" 18:59 17-mai-24
```

How can I give virtual GPU resources to my end users seamlessly ?

OpenInfra Day France 2024

Now, what's new in Caracal ?

SR-IOV GPUs

The case

Now some physical GPUs have virtual functions

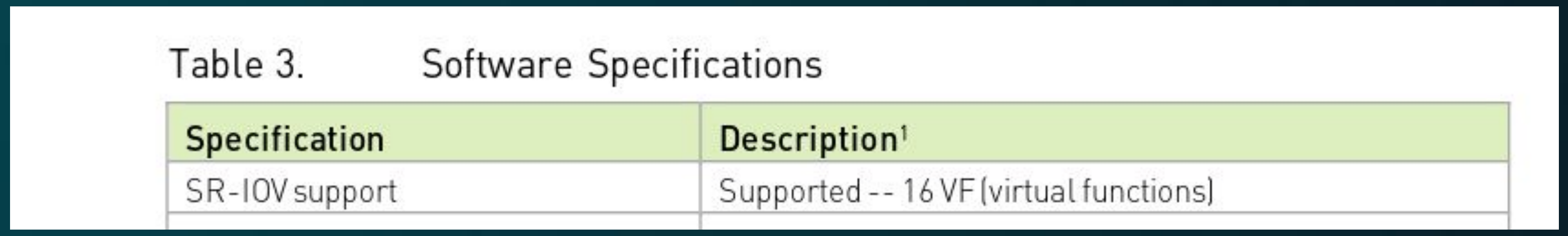
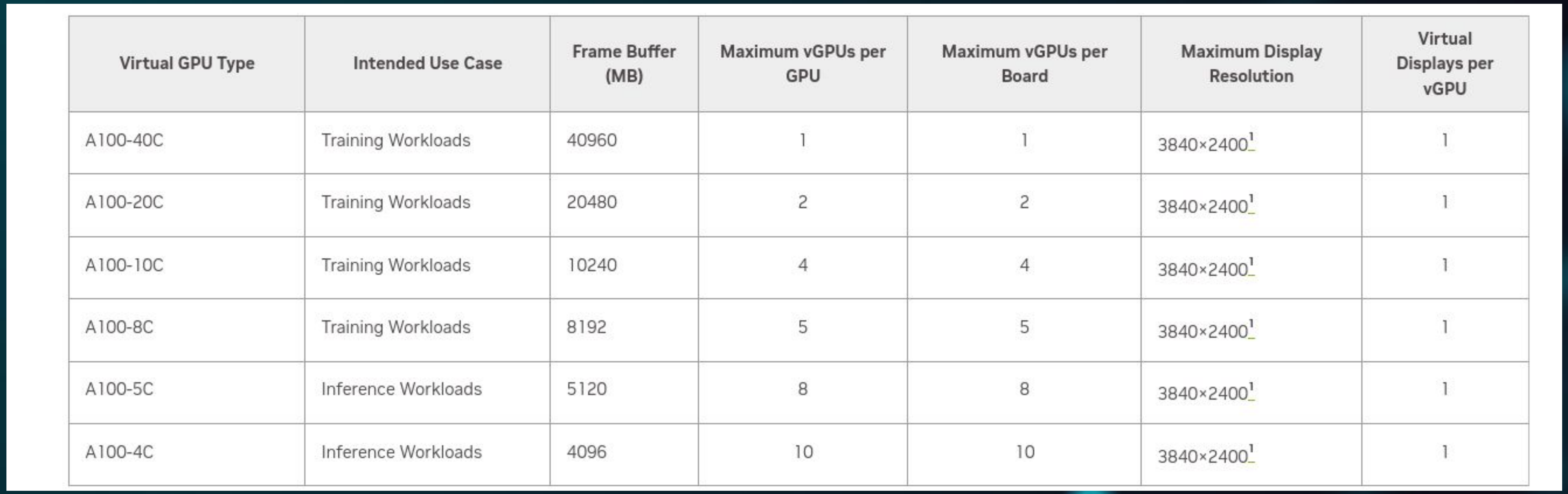


Table 3. Software Specifications

Specification	Description[1]
SR-IOV support	Supported -- 16 VF (virtual functions)

The usage

Nothing changes : each type supports less mdevs than the number of the VFs



Virtual GPU Type	Intended Use Case	Frame Buffer (MB)	Maximum vGPUs per GPU	Maximum vGPUs per Board	Maximum Display Resolution	Virtual Displays per vGPU
A100-40C	Training Workloads	40960	1	1	3840×2400[1]	1
A100-20C	Training Workloads	20480	2	2	3840×2400[1]	1
A100-10C	Training Workloads	10240	4	4	3840×2400[1]	1
A100-8C	Training Workloads	8192	5	5	3840×2400[1]	1
A100-5C	Inference Workloads	5120	8	8	3840×2400[1]	1
A100-4C	Inference Workloads	4096	10	10	3840×2400[1]	1

SR-IOV GPUs

The problem

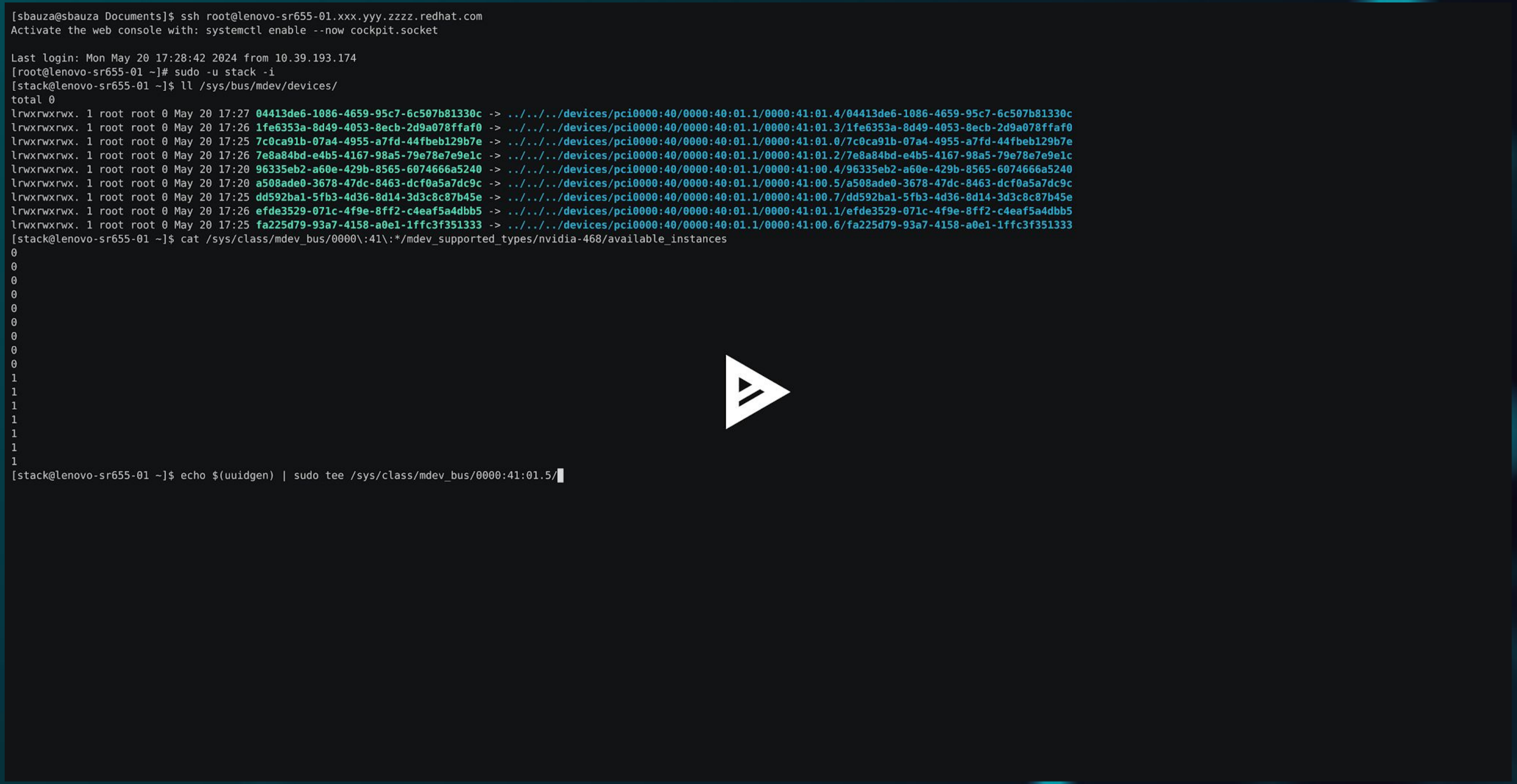
If all the mdevs are created, then all the VFs no longer have inventories

The fix

Nova now asks how many mdevs could be used

```
[devices]
enabled_mdev_types = nvidia-468

[mdev_nvidia-468]
max_instances = 10
```

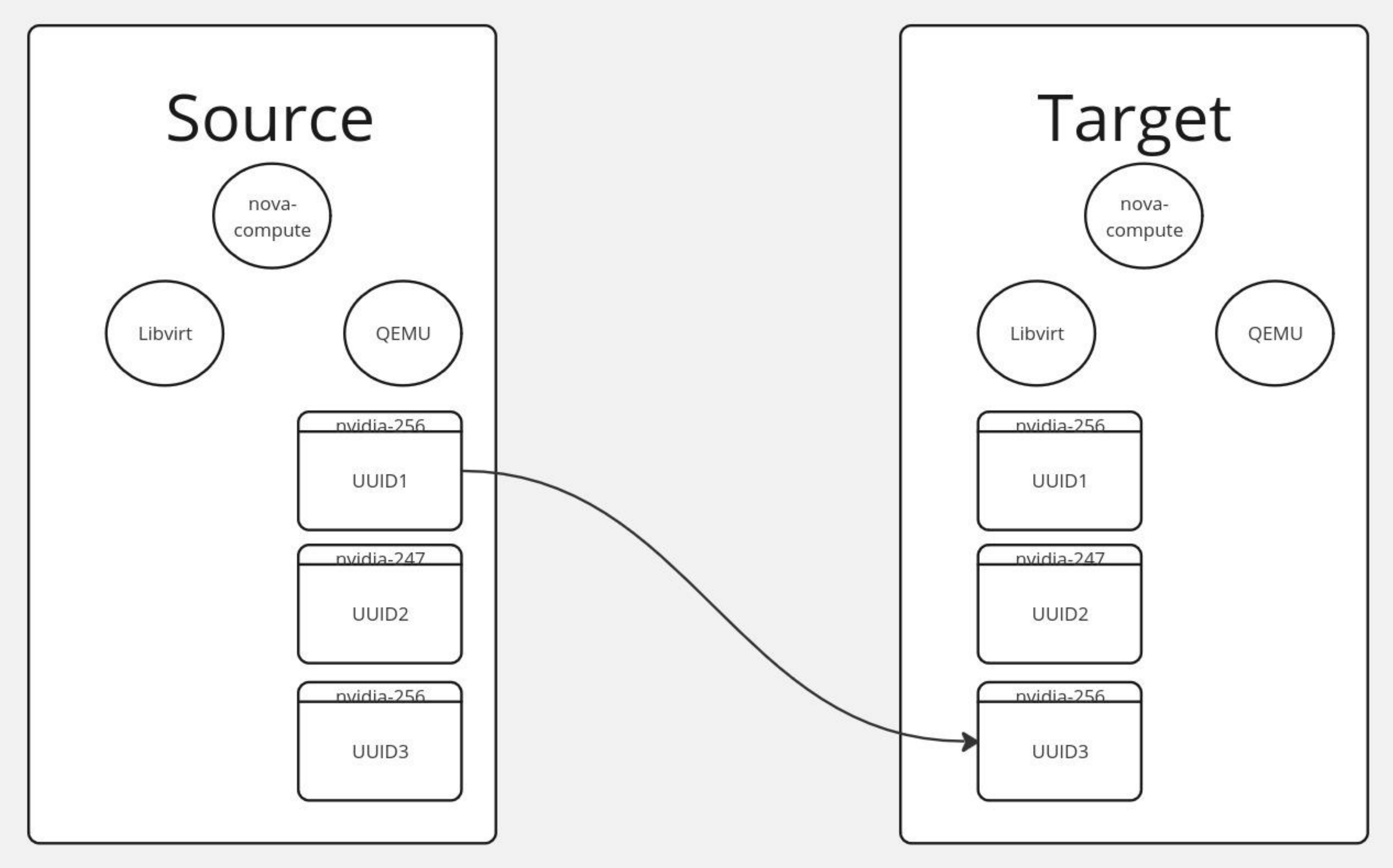


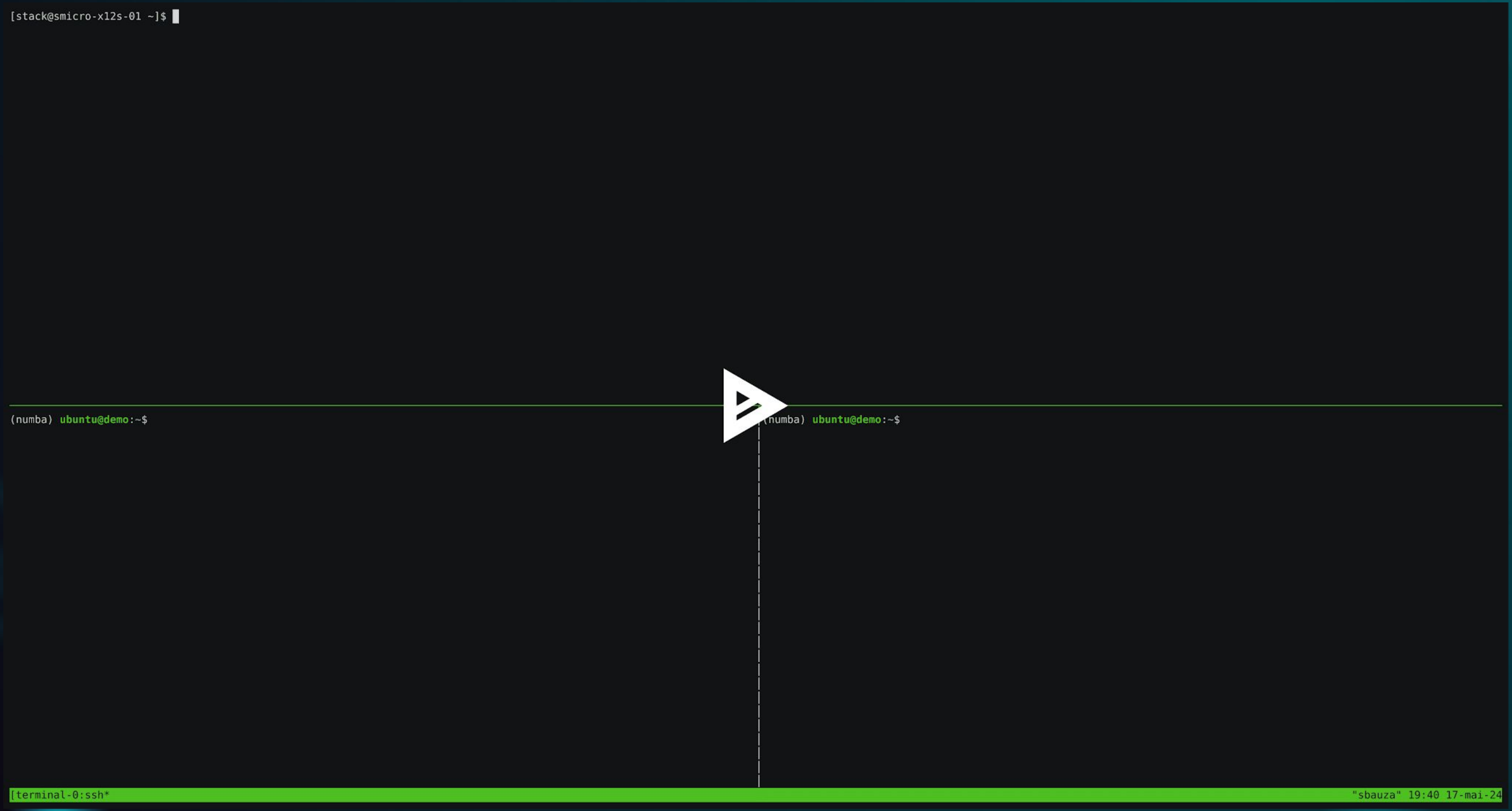
```
[sbauza@sbauza Documents]$ ssh root@lenovo-sr655-01.xxx.yyy.zzzz.redhat.com
Activate the web console with: systemctl enable --now cockpit.socket

Last login: Mon May 20 17:28:42 2024 from 10.39.193.174
[root@lenovo-sr655-01 ~]# sudo -u stack -i
[stack@lenovo-sr655-01 ~]$ ll /sys/bus/mdev/devices/
total 0
lrwxrwxrwx. 1 root root 0 May 20 17:27 04413de6-1086-4659-95c7-6c507b81330c -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:01.4/04413de6-1086-4659-95c7-6c507b81330c
lrwxrwxrwx. 1 root root 0 May 20 17:26 1fe6353a-8d49-4053-8ecb-2d9a078ffaf0 -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:01.3/1fe6353a-8d49-4053-8ecb-2d9a078ffaf0
lrwxrwxrwx. 1 root root 0 May 20 17:25 7c0ca91b-07a4-4955-a7fd-44fbeb129b7e -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:01.0/7c0ca91b-07a4-4955-a7fd-44fbeb129b7e
lrwxrwxrwx. 1 root root 0 May 20 17:26 7e8a84bd-e4b5-4167-98a5-79e78e7e9e1c -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:01.2/7e8a84bd-e4b5-4167-98a5-79e78e7e9e1c
lrwxrwxrwx. 1 root root 0 May 20 17:20 96335eb2-a60e-429b-8565-6074666a5240 -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:00.4/96335eb2-a60e-429b-8565-6074666a5240
lrwxrwxrwx. 1 root root 0 May 20 17:20 a508ade0-3678-47dc-8463-dcf0a5a7dc9c -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:00.5/a508ade0-3678-47dc-8463-dcf0a5a7dc9c
lrwxrwxrwx. 1 root root 0 May 20 17:25 dd592ba1-5fb3-4d36-8d14-3d3c8c87b45e -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:00.7/dd592ba1-5fb3-4d36-8d14-3d3c8c87b45e
lrwxrwxrwx. 1 root root 0 May 20 17:26 efde3529-071c-4f9e-8ff2-c4eaf5a4dbb5 -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:01.1/efde3529-071c-4f9e-8ff2-c4eaf5a4dbb5
lrwxrwxrwx. 1 root root 0 May 20 17:25 fa225d79-93a7-4158-a0e1-1ffc3f351333 -> ../../../devices/pci0000:40/0000:40:01.1/0000:41:00.6/fa225d79-93a7-4158-a0e1-1ffc3f351333
[stack@lenovo-sr655-01 ~]$ cat /sys/class/mdev_bus/0000\:41\:*/mdev_supported_types/nvidia-468/available_instances
0
0
0
0
0
0
0
0
0
1
1
1
1
1
1
1
[stack@lenovo-sr655-01 ~]$ echo $(uuidgen) | sudo tee /sys/class/mdev_bus/0000:41:01.5/
```

vGPU Live migration support

- Libvirt-8.6.0
- QEMU-8.1.0
- Linux kernel 5.18.0



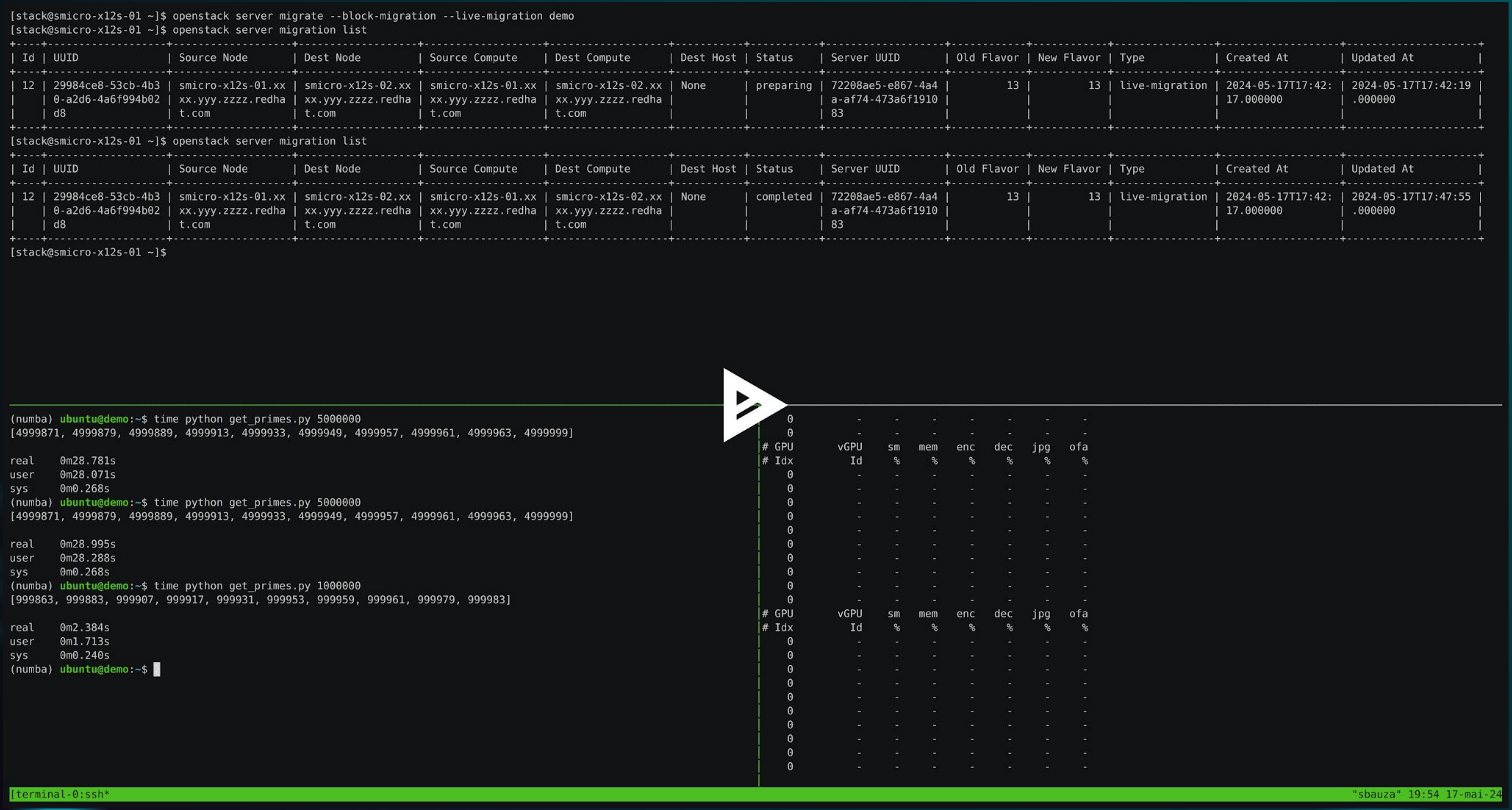


```
[stack@smicro-x12s-01 ~]$
```

```
(numba) ubuntu@demo:~$
```

```
(numba) ubuntu@demo:~$
```

```
[terminal-0:ssh*]                                    "sbauza" 19:40 17-mai-24
```



```
[stack@smicro-x12s-01 ~]$ openstack server migrate --block-migration --live-migration demo
[stack@smicro-x12s-01 ~]$ openstack server migration list
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
| Id | UUID              | Source Node       | Dest Node         | Source Compute    | Dest Compute      | Dest Host | Status    | Server UUID       | Old Flavor | New Flavor | Type           | Created At        | Updated At          |
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
| 12 | 29984ce8-53cb-4b3 | smicro-x12s-01.xx | smicro-x12s-02.xx | smicro-x12s-01.xx | smicro-x12s-02.xx | None      | preparing | 72208ae5-e867-4a4 |         13 |         13 | live-migration | 2024-05-17T17:42: | 2024-05-17T17:42:19 |
|    | 0-a2d6-4a6f994b02 | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha |           |           | a-af74-473a6f1910 |            |            |                | 17.000000         | .000000             |
|    | d8                | t.com             | t.com             | t.com             | t.com             |           |           | 83                |            |            |                |                   |                     |
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
[stack@smicro-x12s-01 ~]$ openstack server migration list
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
| Id | UUID              | Source Node       | Dest Node         | Source Compute    | Dest Compute      | Dest Host | Status    | Server UUID       | Old Flavor | New Flavor | Type           | Created At        | Updated At          |
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
| 12 | 29984ce8-53cb-4b3 | smicro-x12s-01.xx | smicro-x12s-02.xx | smicro-x12s-01.xx | smicro-x12s-02.xx | None      | completed | 72208ae5-e867-4a4 |         13 |         13 | live-migration | 2024-05-17T17:42: | 2024-05-17T17:47:55 |
|    | 0-a2d6-4a6f994b02 | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha | xx.yyy.zzzz.redha |           |           | a-af74-473a6f1910 |            |            |                | 17.000000         | .000000             |
|    | d8                | t.com             | t.com             | t.com             | t.com             |           |           | 83                |            |            |                |                   |                     |
+----+-------------------+-------------------+-------------------+-------------------+-------------------+-----------+-----------+-------------------+------------+------------+----------------+-------------------+---------------------+
[stack@smicro-x12s-01 ~]$
```

```
(numba) ubuntu@demo:~$ time python get_primes.py 5000000
[4999871, 4999879, 4999889, 4999913, 4999933, 4999949, 4999957, 4999961, 4999963, 4999999]

real    0m28.781s
user    0m28.071s
sys     0m0.268s
(numba) ubuntu@demo:~$ time python get_primes.py 5000000
[4999871, 4999879, 4999889, 4999913, 4999933, 4999949, 4999957, 4999961, 4999963, 4999999]

real    0m28.995s
user    0m28.288s
sys     0m0.268s
(numba) ubuntu@demo:~$ time python get_primes.py 1000000
[999863, 999883, 999907, 999917, 999931, 999953, 999959, 999961, 999979, 999983]

real    0m2.384s
user    0m1.713s
sys     0m0.240s
(numba) ubuntu@demo:~$
```

```
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
# GPU     vGPU     sm    mem    enc    dec    jpg    ofa
# Idx       Id      %      %      %      %      %      %
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
# GPU     vGPU     sm    mem    enc    dec    jpg    ofa
# Idx       Id      %      %      %      %      %      %
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
    0        -      -      -      -      -      -      -
```

```
[terminal-0:ssh*]                                                "sbauza" 19:54 17-mai-24
```

Limits with live-migration

You need to use the same mediated device type between the compute nodes

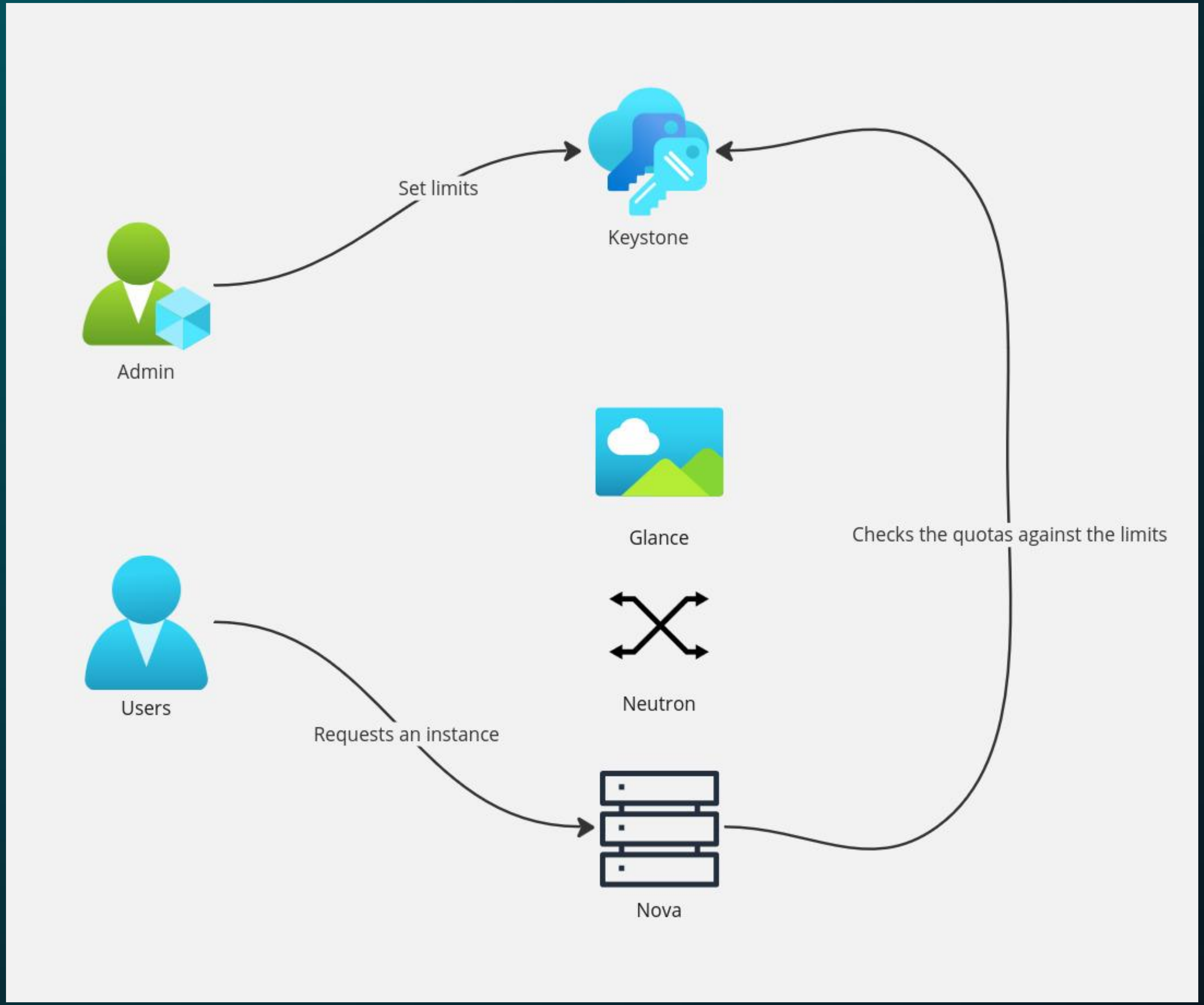
You need to use the same nvidia version between the compute nodes

Older <u>nvidia GPU architectures</u> (Ampere etc.) don’t support framebuffer dirty pages tracking

```
live_migration_completion_timeout = 0
live_migration_downtime = 500000
live_migration_downtime_steps = 3
live_migration_downtime_delay = 3
```

New quotas (aka. unified limits)

How this works, unified limits ?



The setup

API configuration

```
[quota]
driver = nova.quota.UnifiedLimitsDriver

[oslo_limit]
endpoint_id = <uuid>
auth_url = http://<keystone_url>/identity
auth_type = password
username = nova
password = <password>
system_scope = all
user_domain_name = Default
```

Add reader role to the nova user which is system scoped

```
$ openstack role add --user nova
  --user-domain <domain> --system all reader
```

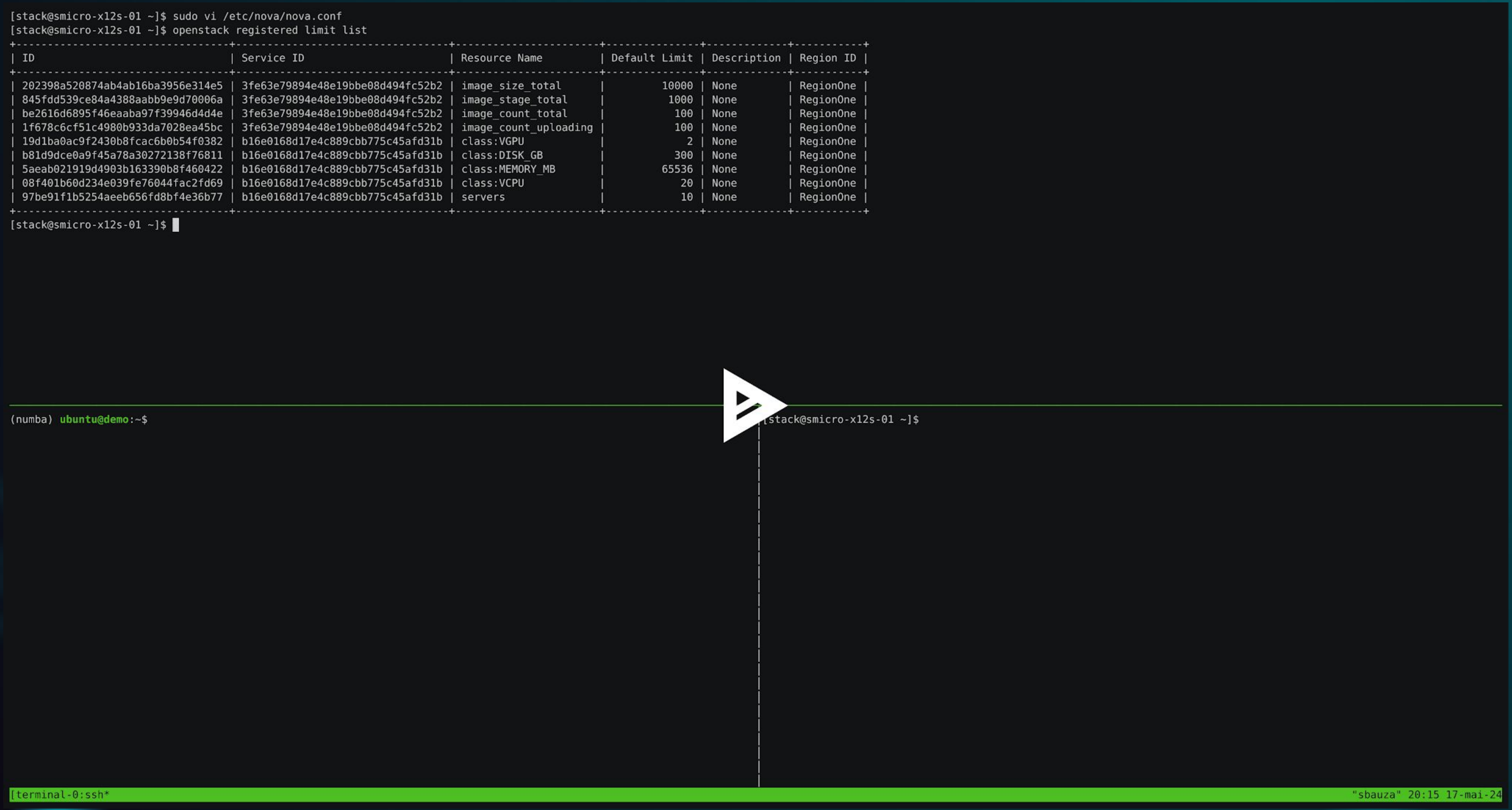
Import existing legacy quota limits

```
$ nova-manage limits migrate_to_unified_limits [--project-id
<project-id>] [--region-id <region-id>] [--verbose]
[--dry-run]
```

Create a specific VGPU limit

```
$ openstack registered limit create --service nova --default-limit <X> class:VGPU
```

https://docs.openstack.org/nova/latest/admin/unified-limits.html



```
[stack@smicro-x12s-01 ~]$ sudo vi /etc/nova/nova.conf
[stack@smicro-x12s-01 ~]$ openstack registered limit list
+----------------------------------+----------------------------------+-----------------------+---------------+-------------+-----------+
| ID                               | Service ID                       | Resource Name         | Default Limit | Description | Region ID |
+----------------------------------+----------------------------------+-----------------------+---------------+-------------+-----------+
| 202398a520874ab4ab16ba3956e314e5 | 3fe63e79894e48e19bbe08d494fc52b2 | image_size_total      |         10000 | None        | RegionOne |
| 845fdd539ce84a4388aabb9e9d70006a | 3fe63e79894e48e19bbe08d494fc52b2 | image_stage_total     |          1000 | None        | RegionOne |
| be2616d6895f46eaaba97f39946d4d4e | 3fe63e79894e48e19bbe08d494fc52b2 | image_count_total     |           100 | None        | RegionOne |
| 1f678c6cf51c4980b933da7028ea45bc | 3fe63e79894e48e19bbe08d494fc52b2 | image_count_uploading |           100 | None        | RegionOne |
| 19d1ba0ac9f2430b8fcac6b0b54f0382 | b16e0168d17e4c889cbb775c45afd31b | class:VGPU            |             2 | None        | RegionOne |
| b81d9dce0a9f45a78a30272138f76811 | b16e0168d17e4c889cbb775c45afd31b | class:DISK_GB         |           300 | None        | RegionOne |
| 5aeab021919d4903b163390b8f460422 | b16e0168d17e4c889cbb775c45afd31b | class:MEMORY_MB       |         65536 | None        | RegionOne |
| 08f401b60d234e039fe76044fac2fd69 | b16e0168d17e4c889cbb775c45afd31b | class:VCPU            |            20 | None        | RegionOne |
| 97be91f1b5254aeeb656fd8bf4e36b77 | b16e0168d17e4c889cbb775c45afd31b | servers               |            10 | None        | RegionOne |
+----------------------------------+----------------------------------+-----------------------+---------------+-------------+-----------+
[stack@smicro-x12s-01 ~]$
```

```
(numba) ubuntu@demo:~$
```

```
[stack@smicro-x12s-01 ~]$
```

```
[terminal-0:ssh*                                                            "sbauza" 20:15 17-mai-24
```

The limits of unified limits

This is experimental yet

Make sure you create all the requested limits

Thanks, questions ?