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

OpenInfra Meetup @ CERN 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

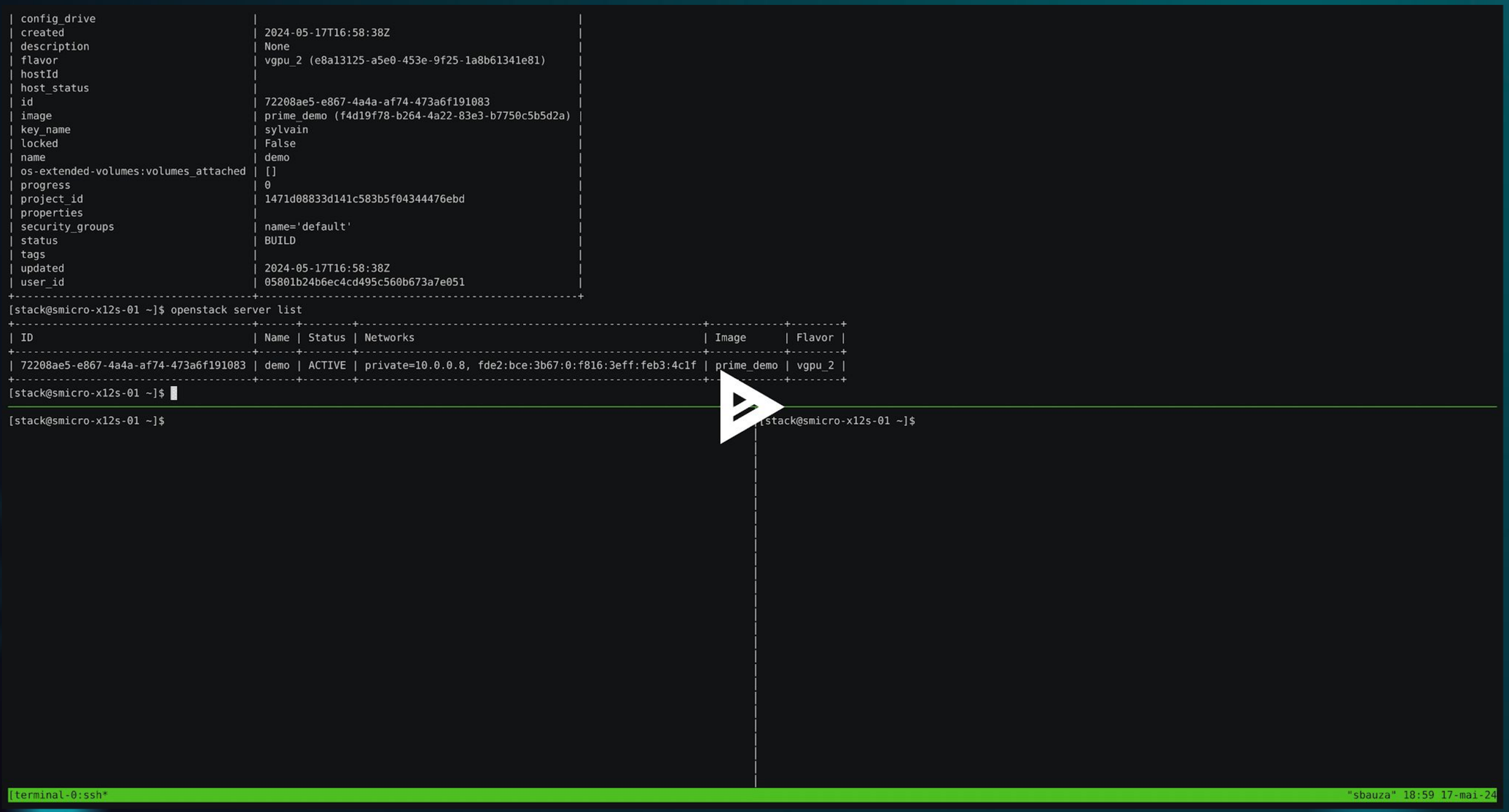
How can I give virtual GPU resources to my end users seamlessly ?

OpenInfra Meetup @CERN 2024

Virtual GPUs in Nova

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

OpenInfra Meetup @CERN 2024



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

OpenInfra Meetup @CERN 2024

Now, what's new in Caracal ?

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

OpenInfra Meetup @CERN 2024

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

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

SR-IOV GPUs

The problem

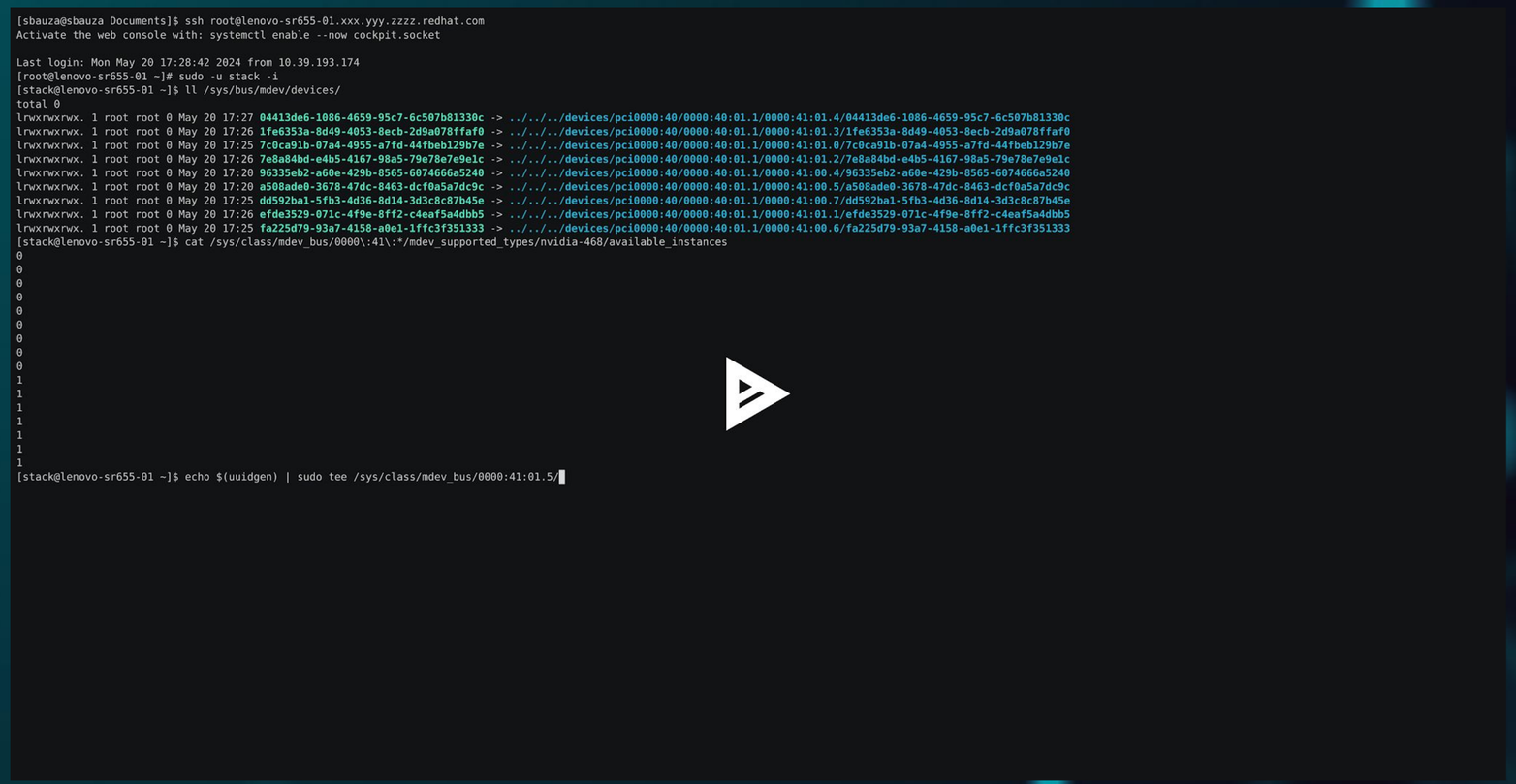
If all the mdevs are created, then all the VFs no longer have inventories

The fix

Nova now asks how many mdevs could be used

```
[devices]
enabled_mdev_types = nvidia-468

[mdev_nvidia-468]
max_instances = 10
```

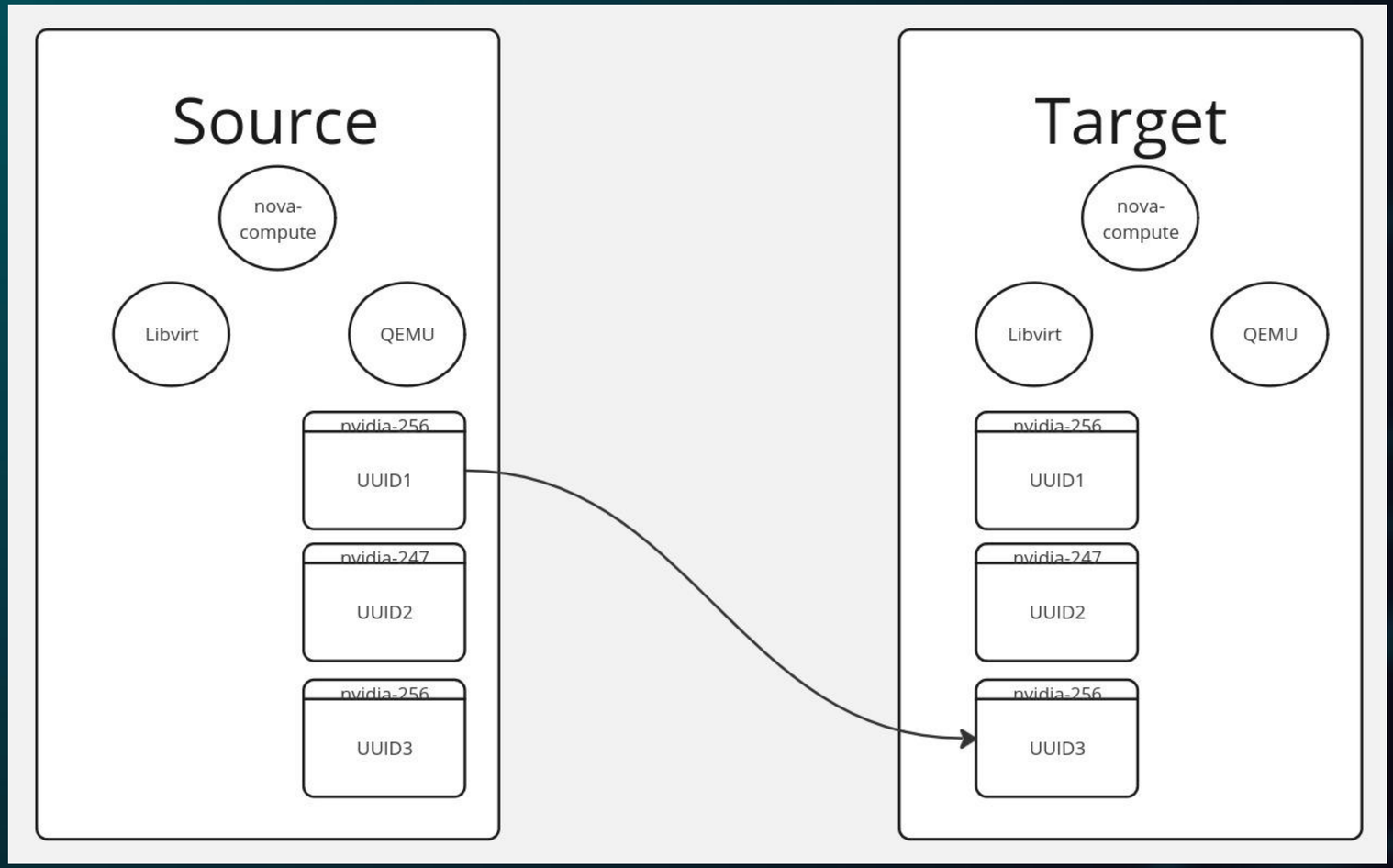


How can I give virtual GPU resources to my end users seamlessly ?

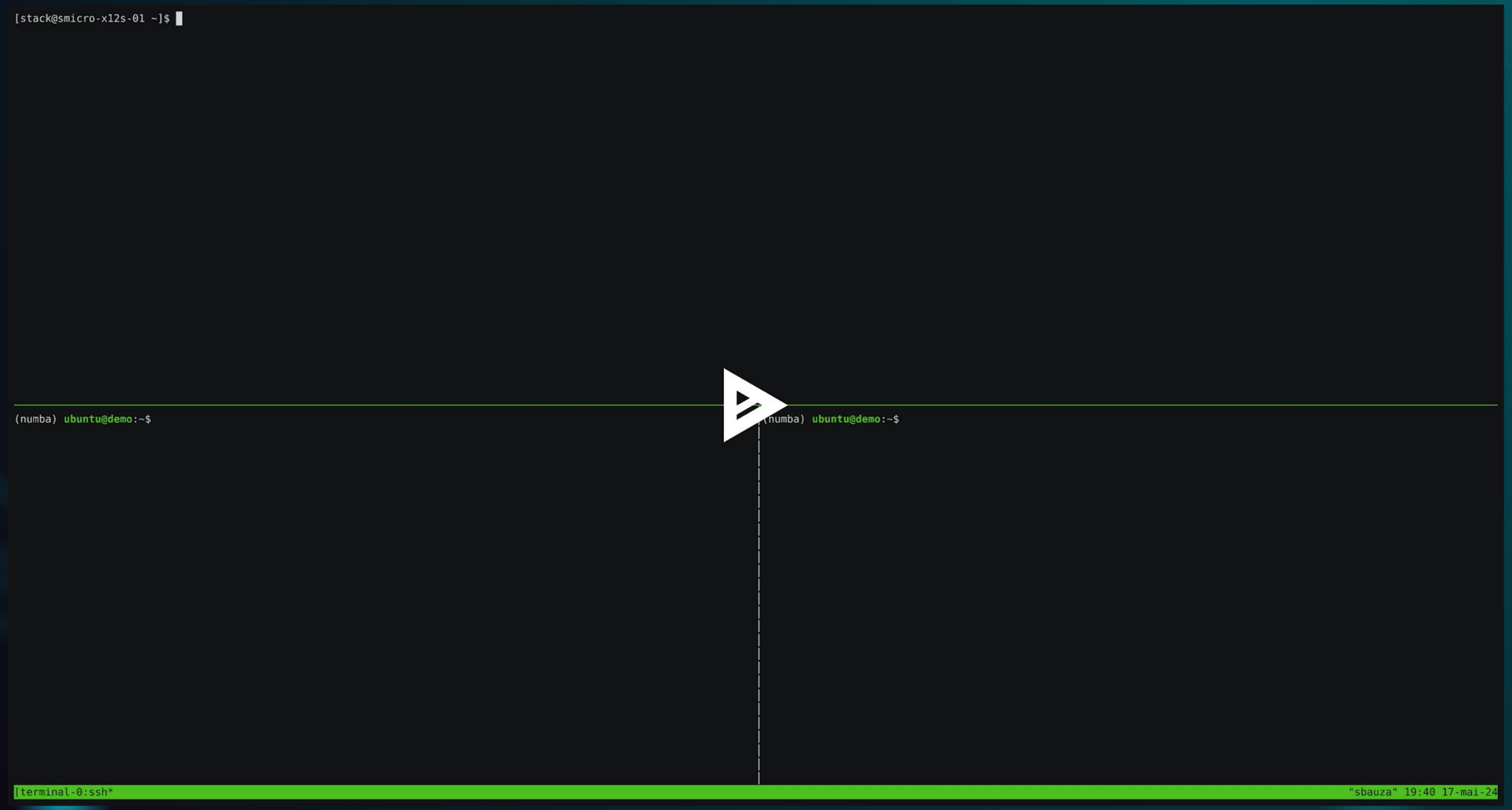
OpenInfra Meetup @CERN 2024

vGPU Live migration support

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



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



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

OpenInfra Meetup @CERN 2024

```
[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 nvidia GPU architectures (Ampere etc.) don't support framebuffer dirty pages tracking

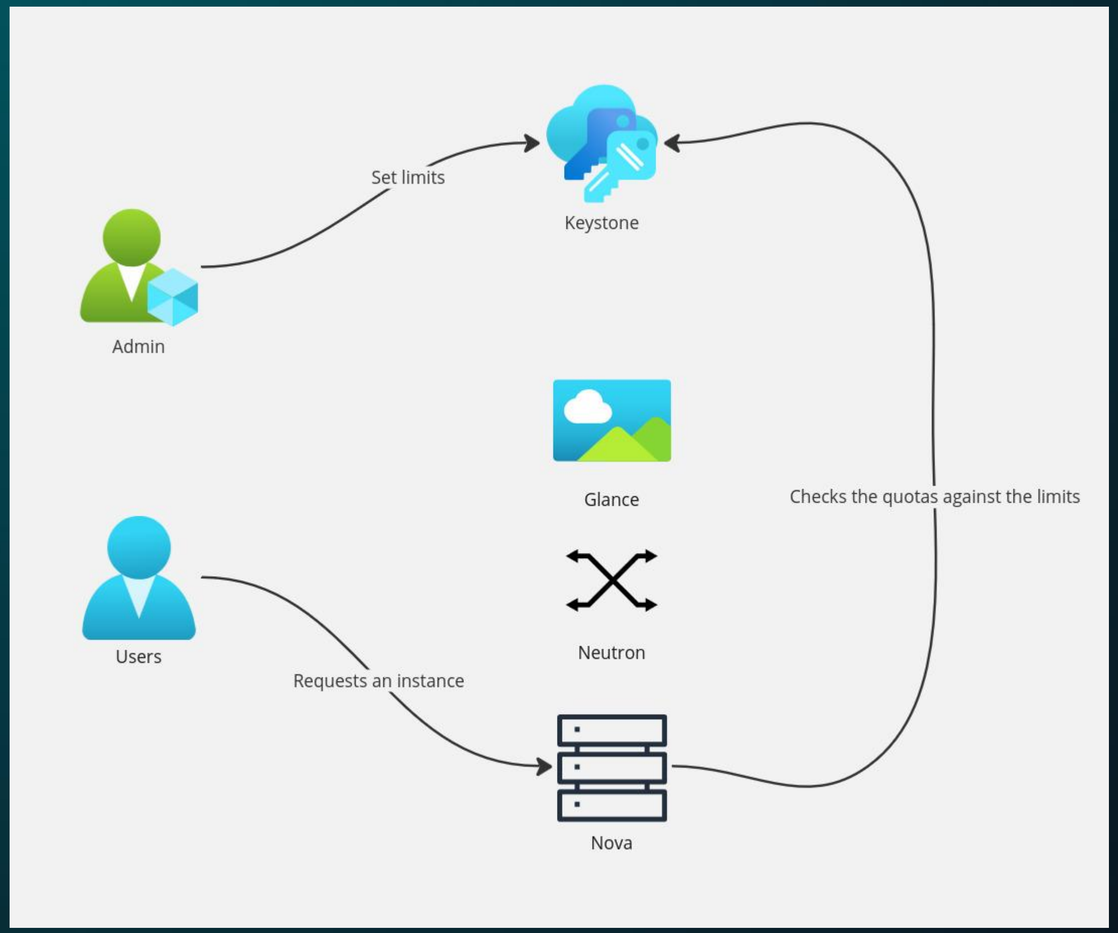
```
live_migration_completion_timeout = 0
live_migration_downtime = 500000
live_migration_downtime_steps = 3
live_migration_downtime_delay = 3
```

New quotas (aka. unified limits)

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

OpenInfra Meetup @CERN 2024

How this works, unified limits ?



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

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
```

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

OpenInfra Meetup @CERN 2024

The limits of unified limits

This is experimental yet

Make sure you create all the requested limits

Thanks, questions ?