VGpus with OpenStack

Accelerating Science
SYLVAIN BAUZA
[sil-vã]
@sylvainbauza
IRC: bauzas

MOHAMMED NASER
@_mnaser
IRC: mnaser
GPU & ACCELERATION CAPABILITIES IN NOVA

× PCI passthrough
× SR-IOV
× vGPUS
GPU
PASSTHROUGH
GPU PASSTHROUGH

- Direct guest usage
- Maximum performance
- Supported in Nova since Havana
Caveats

- Capacity management challenges
- Device exposure to the guest
  - Security of devices
- PCI-E lanes limitations per CPU
- Guest UEFI firmware is recommended
USE CASES

- Access to a physical device
  - GPUs
  - NICs
  - FPGAs
  - Offloading cards
    - i.e. Intel QAT for OpenSSL
THE OPS OF PCI PASSTHROUGH

× Enable kernel configs (iommu)
× Say which PCI devices should be used for each compute
× Create PCI aliases for each PCI device
× Use the PCI alias(es) in the flavor extra spec
× ... and don’t forget the filter!
SR-IOV FOR GPUS
SR-IOV

- Multiple virtual functions from different physical devices
- Hardware accelerated virtualization
- Heavily used in network offloading
- Supported in Nova since Juno
- Emerging use in GPUs (only with AMD)
Caveats

- Limited number of virtual functions
- Bypass security features of OpenStack
- Complicated setup process
- Device specific
  - GPUs: very new, not heavily tested/supported
  - NICs: Unable to use with VxLAN/GRE networks
USE CASES

× High performance networking
  × High throughput/pps scenarios
  × Scaling 40G/100G + Infiniband networks
× GPUs
  × Ideally a better grained resource allocation
  × ... but not tested hence not supported in Nova yet!
VIRTUAL GPUS
VIRTUAL GPUS

× vCPUs.. but for GPUs
× Improves efficiency of GPU usage (no 1:1 mapping anymore)
× Brings all the advantages of CPU virtualization to GPUs
× In Nova since Queens
VIRTUAL GPUs

- Manageable host API for allocating resources on demand
- Device management only on the host
- Introspectable hardware
Caveats

- Vendor-specific driver
  - Intel GVT-g
  - Nvidia GRID

- Product lines specific
  - Mainly GRID (Tesla, Volta) for Nvidia
  - Intel Core 5th Gen SoC and higher

- Guest driver limitations
NVIDIA VIRTUAL GPU DRIVER LIMITATIONS

- Unified Memory
- Dynamic page retirement
- Error-correcting code (ECC)
- Peer-to-peer (between GPU cards)
- GPUDirect remote direct memory access (RDMA)
USE CASES

× Virtual desktops for science
  × High performance server-side rendering
  × Improved efficiency by sharing GPUs
× General-Purpose GPU usage
  × GPUs are not permanently reserved
× GPU profiling per use case
  × You can select what you want
**GPU Profiles (aka. 'MDEV' Types)**

Example for the NVidia V100 PCIe card (one GPU per board):

<table>
<thead>
<tr>
<th>Virtual GPU Type</th>
<th>Frame Buffer (Gbytes)</th>
<th>Maximum vGPUs per GPU</th>
<th>Maximum vGPUs per Board</th>
<th>Required License Edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100D-32Q</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>Quadro vDWS</td>
</tr>
<tr>
<td>V100D-16Q</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>Quadro vDWS</td>
</tr>
<tr>
<td>V100D-8Q</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>Quadro vDWS</td>
</tr>
<tr>
<td>V100D-4Q</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>Quadro vDWS</td>
</tr>
<tr>
<td>V100D-2Q</td>
<td>2</td>
<td>16</td>
<td>16</td>
<td>Quadro vDWS</td>
</tr>
<tr>
<td>V100D-1Q</td>
<td>1</td>
<td>32</td>
<td>32</td>
<td>Quadro vDWS</td>
</tr>
</tbody>
</table>
VFIO MEDIATED DEVICES

|- [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]

THREE WORDS:
VFIO MEDIATED DEVICES

Requires Linux kernel >= 4.10
ALLOCATING AN INSTANCE

The libvirt driver creates a new mediated device or uses a non-allocated one.

```
echo <UUID> > /sys/class/mdev_bus/<device>/mdev_supported_types/<type>/create
```
EARLY STEPS FOR PLAYING WITH VGPUS
THE INSTALL AND CONFIGURE STEPS

1. Edit nova.conf

[devices]
enabled_vgpu_types = nvidia-35

[devices]
enabled_vgpu_types = i915-GVTg_V5_1

2. Create a flavor

$ openstack flavor create --vcpus 4 --ram 4096 --disk 20 vgpu_1
$ openstack flavor set vgpu_1 --property "resources:VGPU=1"

https://docs.openstack.org/nova/latest/admin/virtual-gpu.html
USING NVIDIA GRID AND RHEL?

× Remove the **nouveau** driver

```
$ echo -e "blacklist nouveau\noptions nouveau modeset=0" > /etc/modprobe.d/nouveau.conf
$ dracut -v --force
```

× Install the **nvidia GRID** driver

× Reboot
**WHAT’S NEXT FOR VGPUS IN NOVA?**

- Fix bugs!
- A specific quota class (e.g. VGPU)
- NUMA affinity with CPU resources
- Mixing GPU profiles per host
THANKS! QUESTIONS?
HOST AND GUEST OUTPUTS

[root@virtlab606 nvidia]# nvidia-smi
Wed Jul  4 17:35:23 2018

+-----------------------------------------------------------------------------+
| NVIDIA-SMI 390.42                 Driver Version: 390.42                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla M10           On   | 00000000:84:00.0 Off |                  N/A |
| N/A   28C    P8    10W /  53W |    521MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M10           On   | 00000000:85:00.0 Off |                  N/A |
| N/A   28C    P8    10W /  53W |     13MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla M10           On   | 00000000:86:00.0 Off |                  N/A |
| N/A   23C    P8    10W /  53W |     10MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla M10           On   | 00000000:87:00.0 Off |                  N/A |
| N/A   23C    P8    10W /  53W |     10MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0     93890    C+G   vgpu                                         508MiB |
+-----------------------------------------------------------------------------+

[cloud-user@instance0 ~]$ sudo lspci -nn | grep NVIDIA
00:05.0 VGA compatible controller [0300]: NVIDIA Corporation GM107GL [Tesla M10]
[10de:13bd] (rev a2)

[cloud-user@instance0 ~]$ nvidia-smi
Wed Jul  4 18:13:30 2018

+-----------------------------------------------------------------------------+
| NVIDIA-SMI 390.42                 Driver Version: 390.42                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GRID M10-0B         On   | 00000000:00:05.0 Off |                  N/A |
| N/A   N/A    P8    N/A /  N/A |     48MiB /   512MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+