GPU透传部署文档

- 环境:
  - 三控制三计算, 其中一个计算节点为gpu节点
  - 计算节点均使用intel cpu
- 版本: animbus 6.5

准备工作

1. 确认硬件支持虚拟化技术及PCI passthrough

由于需要硬件支持，先要确认CPU及主板（motherboard）是否支持Intel或AMD的硬件辅助虚拟化功能，可以查看官方的硬件支持列表，或者在BIOS中查看相关选项，还需要支持PCI passthrough的PCI硬件设备。

2. 在BIOS中打开硬件辅助虚拟化功能支持

- 对于intel cpu, 在主板中开启VT-x及VT-d选项
  - VT-x为开启虚拟化需要
  - VT-d为开启PCI passthrough

  这两个选项一般在BIOS中Advance下CPU和System或相关条目中设置，例如:

  - VT： Intel Virtualization Technology
  - VT-d: Intel VT for Directed I/O
- 对于 amd cpu, 在主板中开启SVM及IOMMU选项
  - SVM为开启虚拟化需要
  - IOMMU为开启PCI passthrough

3. 确认内核支持iommu

```
cat /proc/cmdline | grep iommu
```

如果没有输出, 则需要修改kernel启动参数

- 对于intel cpu
  1. 编辑`/etc/default/grub`文件, 在`GRUB_CMDLINE_LINUX`行后面添加:

     > intel_iommu=on

     例如:

     > GRUB_CMDLINE_LINUX="crashkernel=auto rd.lvm.lv=centos/root rd.lvm.lv=centos/swap rhgb quiet intel_iommu=on"

     如果没有`GRUB_CMDLINE_LINUX`, 则使用`GRUB_CMDLINE_LINUX_DEFAULT`

2. 更新grub

```
grub2-mkconfig -o /boot/grub2/grub.cfg
```

- 对于amd cpu

  与intel cpu的区别为, 添加的不是`intel_iommu=on`, 而是`iommu=on`, 其他步骤一样

4. 确认pci设备驱动信息

确认pci设备驱动信息并从host默认驱动程序中解绑, 以备虚拟机透传使用, 查看pci设备信息, 此处为nvidia显卡

```
lspci -nn | grep -i nvidia
```

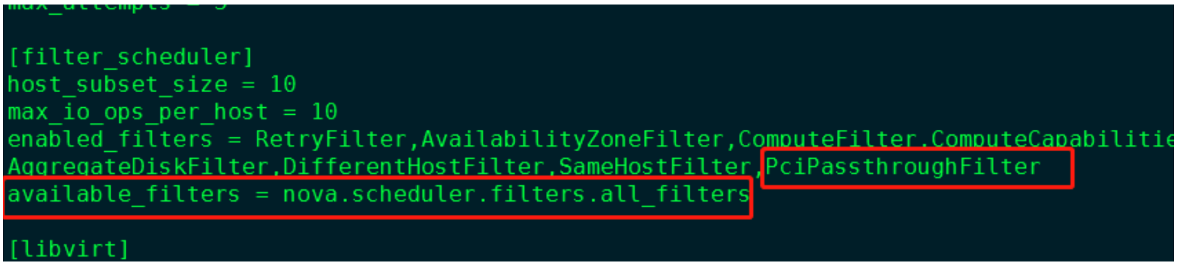
> c1:00.0 3D controller [0302]: NVIDIA Corporation GK110BGL [Tesla K40m][10de:1023] (rev a1)

其中[10de:1023]的10de为NVIDIA pci设备的vendor id, 1023为product id

配置openstack

1. 配置nova-scheduler

- 在filter_scheduler中加入`PciPassthroughFilter`, 同时添加`available_filters = nova.scheduler.filters.all_filters`

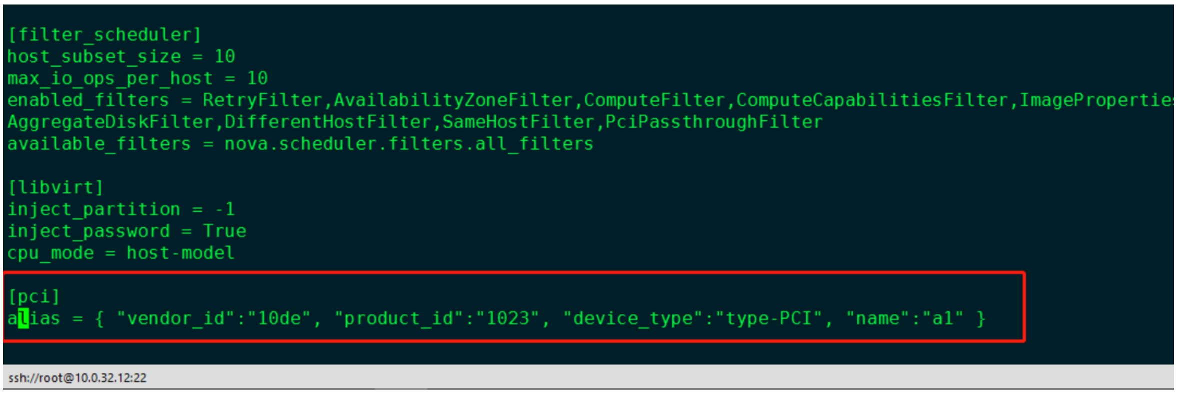


```
max_attempts = 3

[filter_scheduler]
host_subset_size = 10
max_io_ops_per_host = 10
enabled_filters = RetryFilter,AvailabilityZoneFilter,ComputeFilter,ComputeCapabilitie
AggregateDiskFilter,DifferentHostFilter,SameHostFilter,PciPassthroughFilter
available_filters = nova.scheduler.filters.all_filters

[libvirt]
```

2. 配置nova-api

- 添加新的块pci

  > [pci]
  >
  > alias = { "vendor_id":"10de", "product_id":"1023", "device_type":"type-PCI", "name":"a1" }



```
[filter_scheduler]
host_subset_size = 10
max_io_ops_per_host = 10
enabled_filters = RetryFilter,AvailabilityZoneFilter,ComputeFilter,ComputeCapabilitiesFilter,ImagePropertie
AggregateDiskFilter,DifferentHostFilter,SameHostFilter,PciPassthroughFilter
available_filters = nova.scheduler.filters.all_filters

[libvirt]
inject_partition = -1
inject_password = True
cpu_mode = host-model

[pci]
alias = { "vendor_id":"10de", "product_id":"1023", "device_type":"type-PCI", "name":"a1" }
```

ssh://root@10.0.32.12:22

3. reconfigure

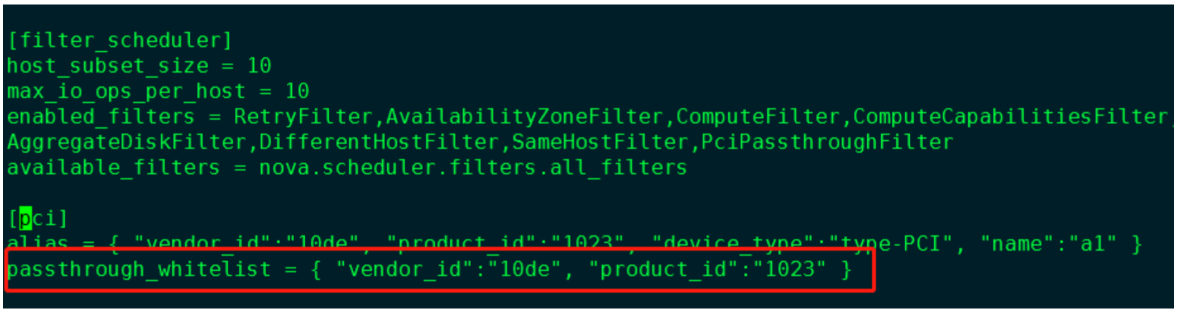
reconfigure nova-scheduler和nova-api, 重启nova-api, nova-scheduler节点

```
kolla-ansible -i ~/multinode reconfigure -t nova
```

4. 配置gpu所在计算节点的nova-compute

- 添加需要直通的pci设备信息

> [pci]
>
> passthrough_whitelist = { "vendor_id": "10de", "product_id": "1023" }



```
[filter_scheduler]
host_subset_size = 10
max_io_ops_per_host = 10
enabled_filters = RetryFilter,AvailabilityZoneFilter,ComputeFilter,ComputeCapabilitiesFilter
AggregateDiskFilter,DifferentHostFilter,SameHostFilter,PciPassthroughFilter
available_filters = nova.scheduler.filters.all_filters

[pci]
alias = { "vendor_id":"10de", "product_id":"1023", "device_type":"type-PCI", "name":"a1" }
passthrough_whitelist = { "vendor_id":"10de", "product_id":"1023" }
```

5. 创建带pci标签的flavor

```
openstack flavor set ml.large --property "pci_passthrough:alias"="a1:1"
```

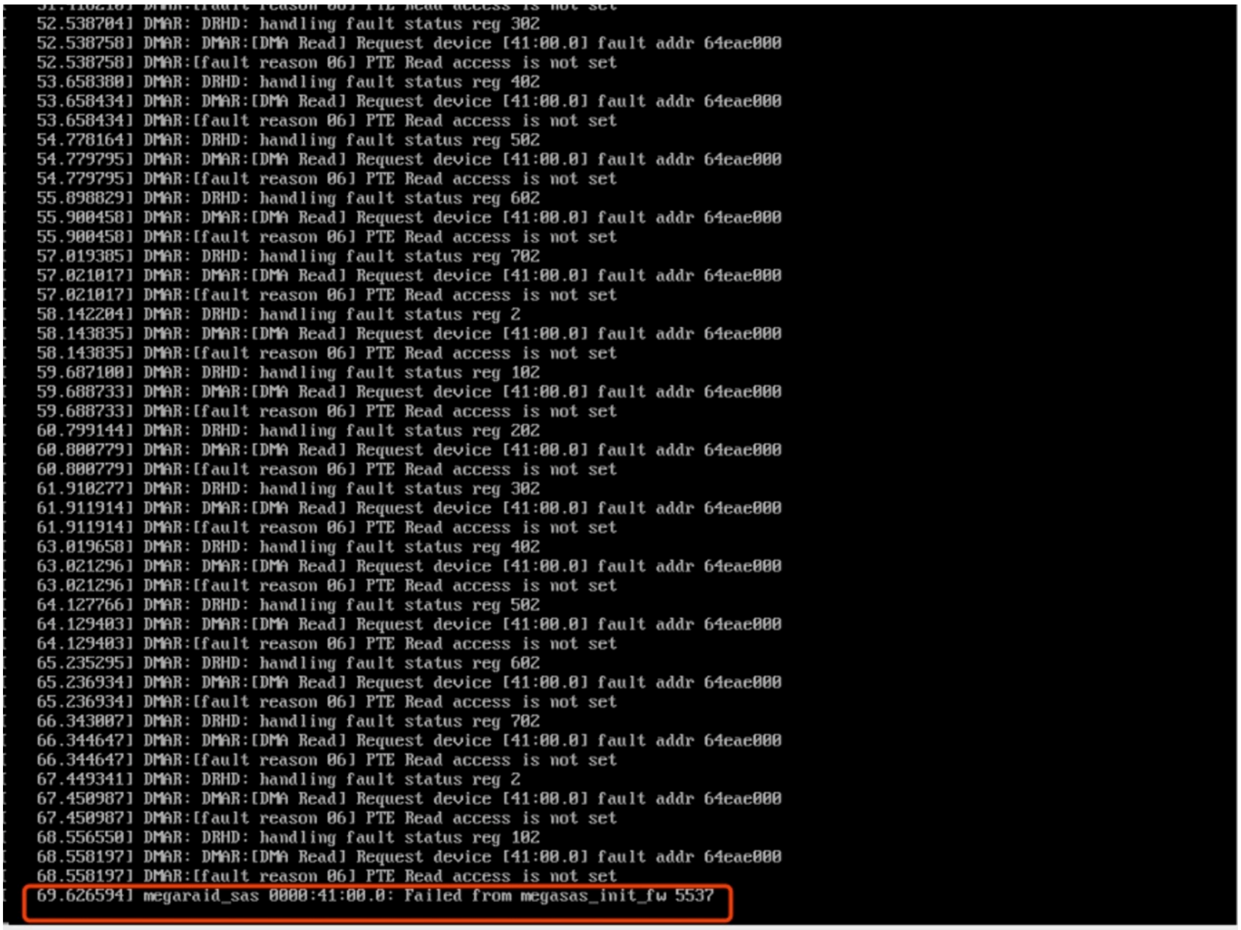
使用该flavor创建虚拟机, 虚拟机会自动调度到gpu节点上

链接

<https://docs.openstack.org/nova/pike/admin/pci-passthrough.html>

Issue:

在向kernel添加'intel_iommu=on'参数后, 系统启动失败



```
[   52.538704] DMAR: DRHD: handling fault status reg 302
[   52.538758] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   52.538758] DMAR:[fault reason 06] PTE Read access is not set
[   53.658380] DMAR: DRHD: handling fault status reg 402
[   53.658434] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   53.658434] DMAR:[fault reason 06] PTE Read access is not set
[   54.778164] DMAR: DRHD: handling fault status reg 502
[   54.779795] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   54.779795] DMAR:[fault reason 06] PTE Read access is not set
[   55.898829] DMAR: DRHD: handling fault status reg 602
[   55.900458] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   55.900458] DMAR:[fault reason 06] PTE Read access is not set
[   57.019385] DMAR: DRHD: handling fault status reg 702
[   57.021017] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   57.021017] DMAR:[fault reason 06] PTE Read access is not set
[   58.142204] DMAR: DRHD: handling fault status reg 2
[   58.143835] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   58.143835] DMAR:[fault reason 06] PTE Read access is not set
[   59.687100] DMAR: DRHD: handling fault status reg 102
[   59.688733] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   59.688733] DMAR:[fault reason 06] PTE Read access is not set
[   60.799144] DMAR: DRHD: handling fault status reg 202
[   60.800779] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   60.800779] DMAR:[fault reason 06] PTE Read access is not set
[   61.910277] DMAR: DRHD: handling fault status reg 302
[   61.911914] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   61.911914] DMAR:[fault reason 06] PTE Read access is not set
[   63.019658] DMAR: DRHD: handling fault status reg 402
[   63.021296] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   63.021296] DMAR:[fault reason 06] PTE Read access is not set
[   64.127766] DMAR: DRHD: handling fault status reg 502
[   64.129403] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   64.129403] DMAR:[fault reason 06] PTE Read access is not set
[   65.235295] DMAR: DRHD: handling fault status reg 602
[   65.236934] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   65.236934] DMAR:[fault reason 06] PTE Read access is not set
[   66.343007] DMAR: DRHD: handling fault status reg 702
[   66.344647] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   66.344647] DMAR:[fault reason 06] PTE Read access is not set
[   67.449341] DMAR: DRHD: handling fault status reg 2
[   67.450987] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   67.450987] DMAR:[fault reason 06] PTE Read access is not set
[   68.556550] DMAR: DRHD: handling fault status reg 102
[   68.558197] DMAR: DMAR:[DMA Read] Request device [41:00.0] fault addr 64eae000
[   68.558197] DMAR:[fault reason 06] PTE Read access is not set
[   69.626594] megaraid_sas 0000:41:00.0: Failed from megasas_init_fw 5537
```

原因: 此raid卡没有iommu模块, 导致添加iommu参数后, raid驱动初始化错误, 导致系统启动失败, 找服务商更换支持透传功能的raid, 问题解决