OpenNebula - Bug #5306
Inconsistent view of GPU resource usage leading to deployment failure. (how to repair?)

08/07/2017 11:49 AM - Hans Feringa

Status:	Pending	**Start date:**	08/07/2017
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Core & System	**Estimated time:**	0.00 hour
Target version:			
Resolution:		**Pull request:**	
Affected Versions:	OpenNebula 5.2		

Description

Opennebula 5.2.1
using PCI GPU passthrough.
There are 4 GPU cards in each (GPU) host.

The information for the host in OpenNebula is not consistent with the actual use of the GPU resource on the host. The schedular will schedule the new VM to be deployed on the host for that particular PCI device and will keep on trying this. As a work-around I disabled the node so it will not be used to deploy new VM's on it.

When looking at the data (body field) in the vm_pool table for the particular VM, I could see the information in the body field that the PCI device was actually in use (indicated by the address for the PCI device).
Looking at the host_pool table in the body field for the host the VM is deployed on, the resource is indicated as available.

This inconsistency is also shown with the onehost and onevm commands.

The annoying thing is that a failed VM is never tried on another host, and is actually stuck on the node where ONE thinks that the allocated device/resource is still available. To get out of this situation I had to disable the host so another GPU host was tried.

The information for the host is in the table host_pool in field body. It shows that for the XML tag VMID has a value of ![CDATA[-1] while it should have (in this case) ![CDATA[25400].

In the body field of the vm_pool table (information for the VM), in the XML blob the information regarding the usage of the PCI device is present, with the correct address, bus etc info. So clearly this information between the two tables is out of sync.

I expected that onedb fsck could "repair" things, but this particular error was not reported nor fixed.

To my knowledge the only thing that remains at this moment is to edit the data in the database directly and correct the information in the host_pool table for this host.

I also expect that the inconsistency can exist the other way arround, where in the host_pool table a resource is recorded as being in use while it actually is available and is not recorded in any of the records in the vm_pool table.

History

#1 - 08/07/2017 12:21 PM - Hans Feringa

Look also at
https://forum.opennebula.org/t/vm-pool-and-host-pool-table-out-of-sync-resulting-in-error-requested-operation-is-not-valid-gpu-pci-device-in-use-by-drier-qemu/4652
I reported it there first but was advised by collegues to report it here instead.