StorPool
DISTRIBUTED STORAGE

Nested Virtualization with OpenNebula (and PCI Passthrough)

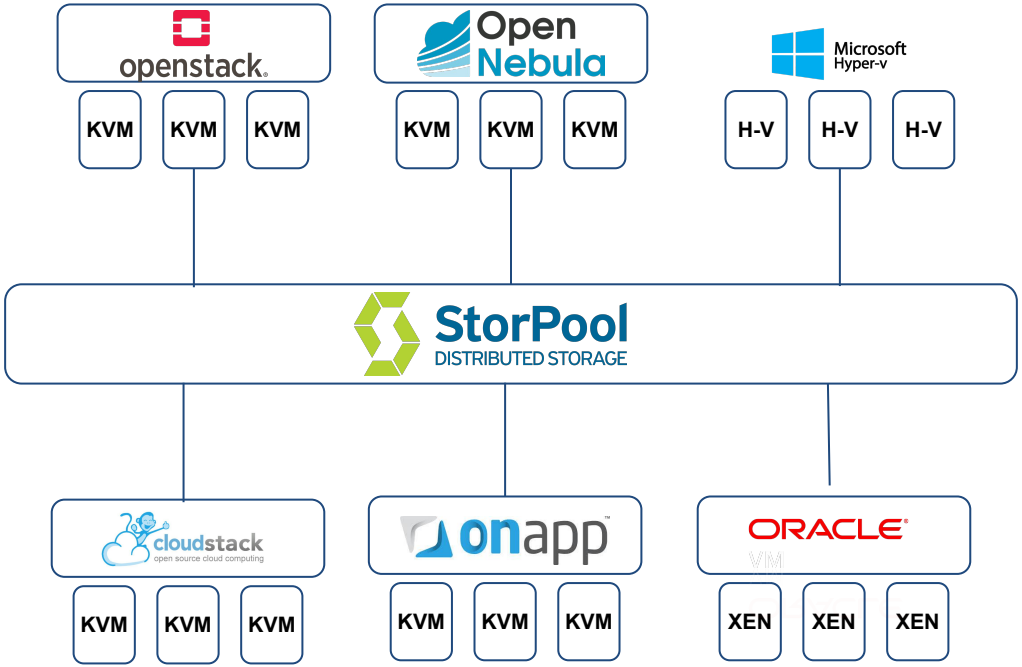
A Practical Approach

Venko Moyankov

OpenNebulaConf 2019

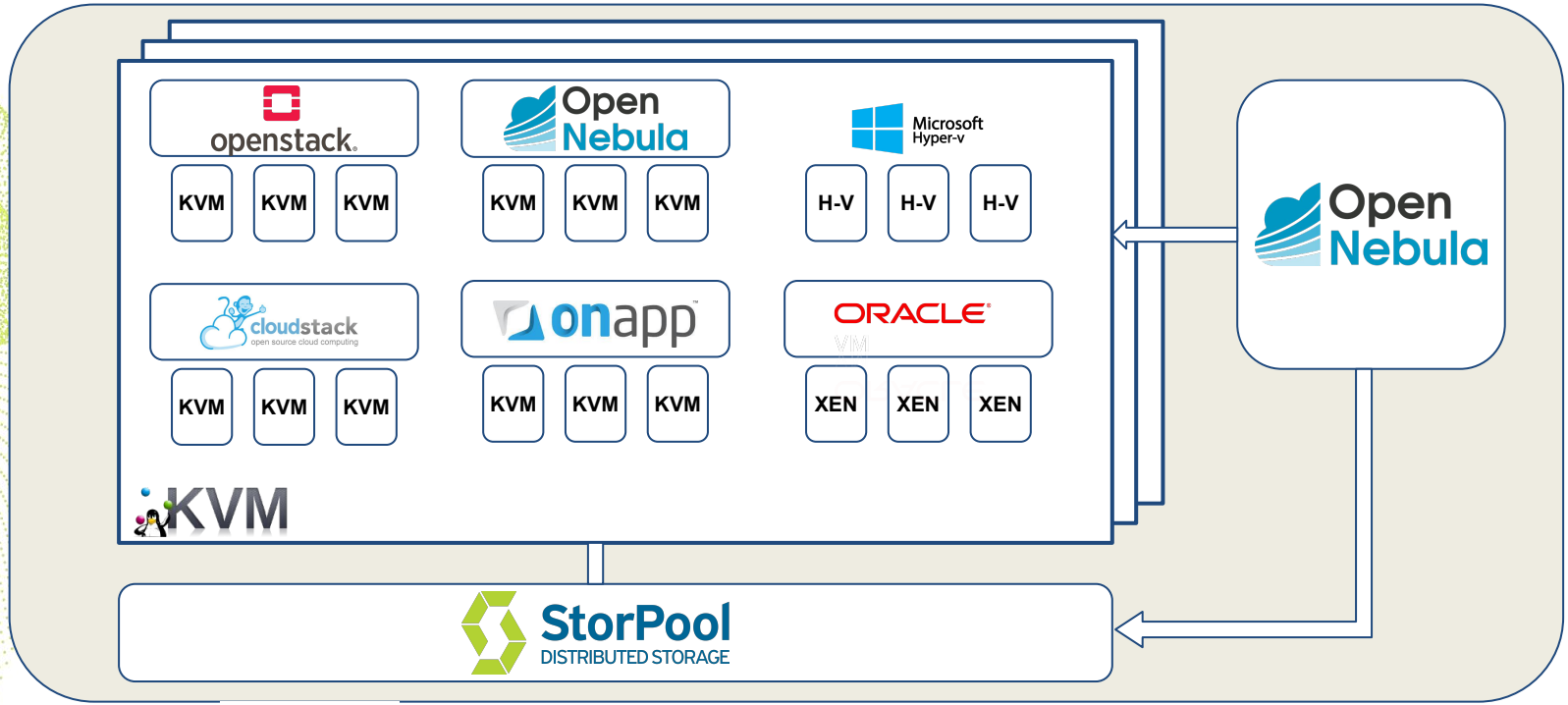
Barcelona, October 21-22

The Problem



StorPool
DISTRIBUTED STORAGE

StorPool Lab



also CITRIX® and more ...

StorPool
DISTRIBUTED STORAGE

The Technologies Behind

- Hardware virtualization (VT-x)
- Nested Virtualization
- VMCS Shadowing
- IOMMU (PCI Passthrough)
- SR-IOV
- ACS (IOMMU groups)

Host

- OpenNebula PCI Passthrough
- libvirt Domain XML tweaks (VF Net only)

OpenNebula

StorPool
DISTRIBUTED STORAGE

Hardware Virtualization

VT-x

AMD-V or SVM

All CPUs, but may need to enable it in BIOS

Nested Virtualization

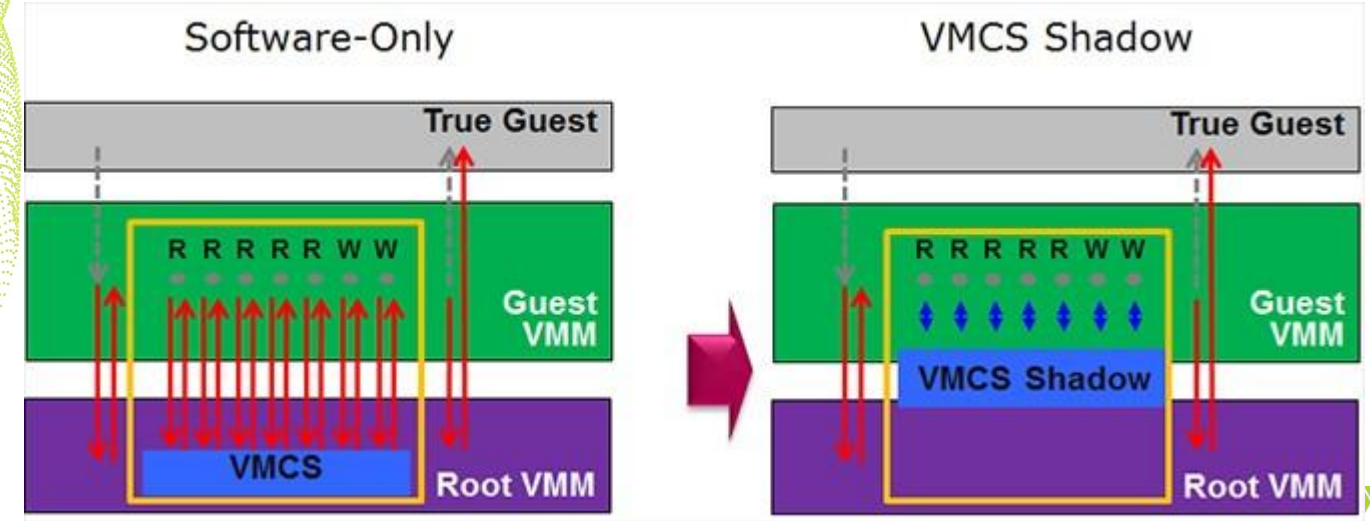
Enables hardware virtualization in the guest

KVM feature

StorPool
DISTRIBUTED STORAGE

VMCS Shadowing

- Hardware feature
- Accelerates nested virtualization
- Available in most CPUs since 2013 (Haswell)



Source: https://software.intel.com/en-us/blogs/2014/12/12/enabling-virtual-machine-control-structure-shadowing-on-a-nested-virtual-machine

StorPool DISTRIBUTED STORAGE

PCI Passthrough

Allows guests to have **direct**

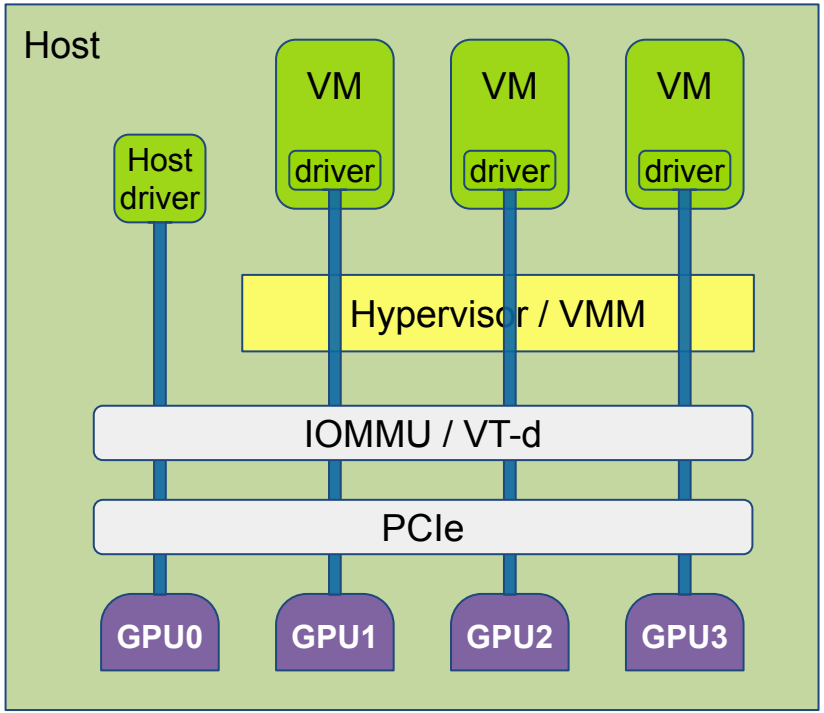
exclusive access to PCI devices

- I/O MMU virtualization (IOMMU)

intel VT-d

AMD AMD-Vi

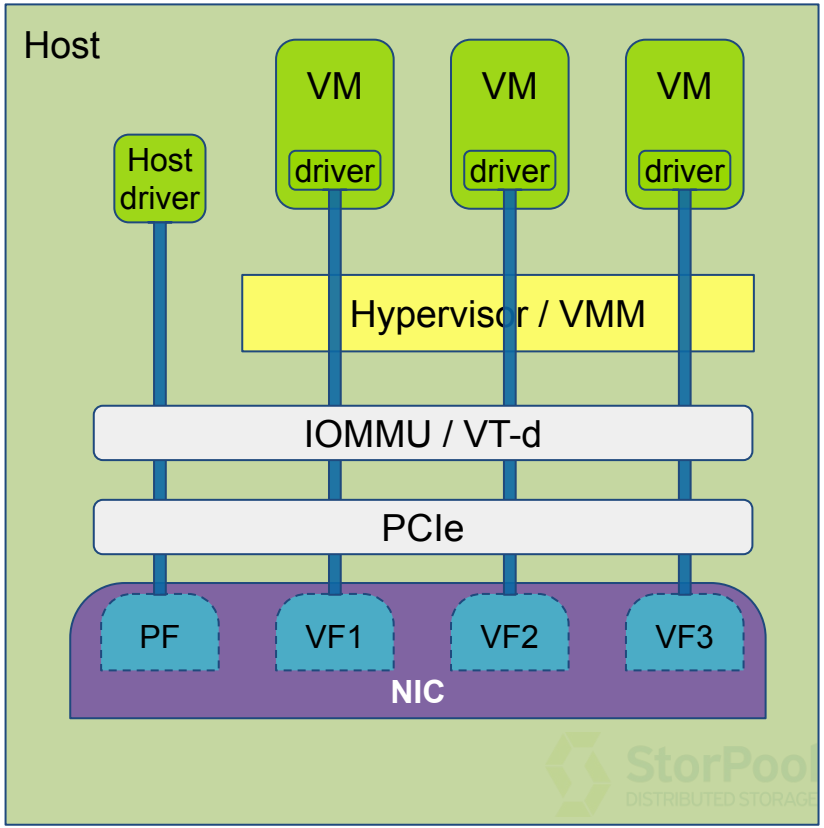
Mostly used for GPU and NIC



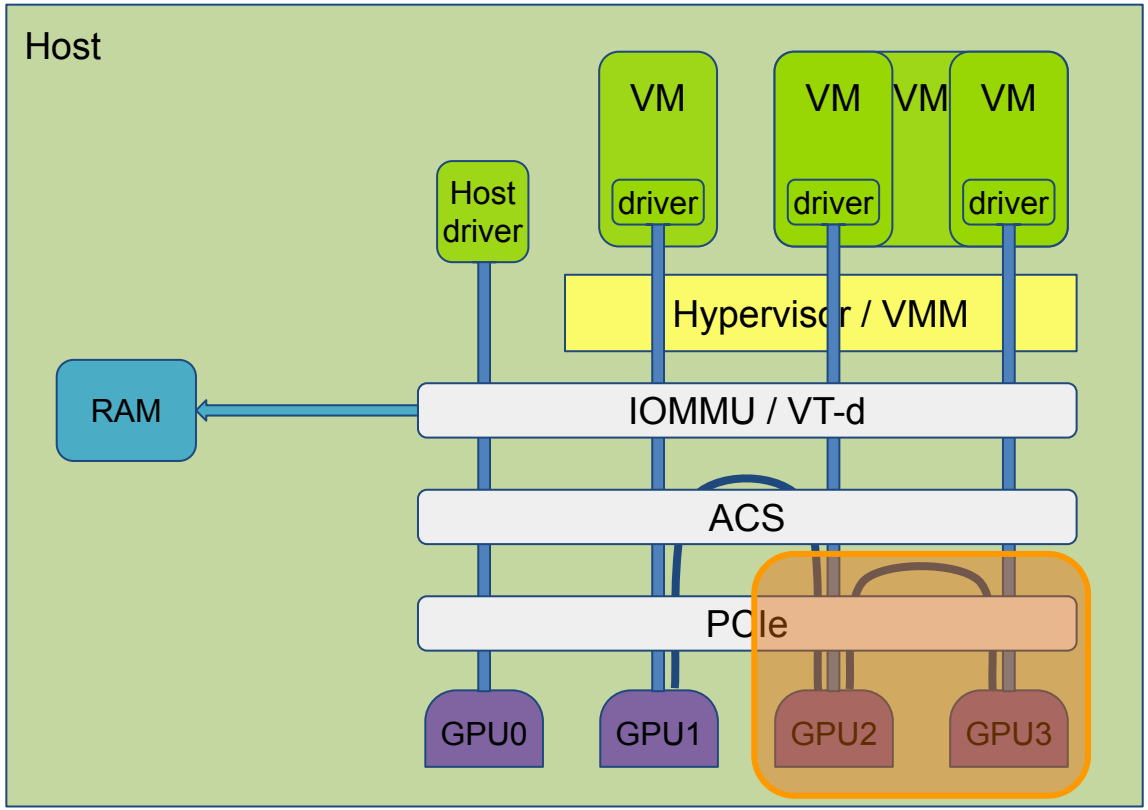
StorPool DISTRIBUTED STORAGE

SR-IOV

- Single Root I/O Virtualization
- One physical device appears as multiple virtual functions (VF)
- Allows different VMs to share a single PCIe hardware
- Mostly used for NIC



ACS and IOMMU Groups



StorPool
DISTRIBUTED STORAGE

Let's Do It

StorPool
DISTRIBUTED STORAGE

Host

Hardware Virtualization

Check it is available with: `lscpu | grep vmx`

Nested Virtualization

kvm-intel.conf: `options kvm-intel nested=1`

Check it in the *guest* with: `lscpu | grep vmx`

VMCS Shadowing

kvm-intel.conf: `options kvm-intel enable_shadow_vmcs=1`

StorPool
DISTRIBUTED STORAGE

PCI Passthrough & SR-IOV

PCI Passthrough

Enabled via ***kernel*** options:

Intel:

```
intel_iommu=on iommu=pt
```

AMD:

```
amd_iommu=pt
```

SR-IOV

- Ensure SR-IOV and VT-d are enabled in BIOS.
- Setup VFs:

```
# echo '8' > /sys/class/net/eth3/device/sriov_numvfs
```

StorPool DISTRIBUTED STORAGE

ACS and IOMMU Groups

- Check ACS is supported

```
lspci -vv | egrep "Access Control Services"
```

- Check IOMMU groups

```
for a in /sys/kernel/iommu_groups/*; do find $a -type l;
done | sort --version-sort
```

```
/sys/kernel/iommu_groups/20/devices/0000:18:00.0
/sys/kernel/iommu_groups/21/devices/0000:19:03.0
/sys/kernel/iommu_groups/22/devices/0000:1a:00.0
/sys/kernel/iommu_groups/23/devices/0000:1a:00.1
/sys/kernel/iommu_groups/24/devices/0000:3a:01.0
/sys/kernel/iommu_groups/25/devices/0000:3a:05.0
/sys/kernel/iommu_groups/25/devices/0000:3a:05.2
/sys/kernel/iommu_groups/25/devices/0000:3a:05.4
/sys/kernel/iommu_groups/26/devices/0000:3a:08.0
/sys/kernel/iommu_groups/27/devices/0000:3a:09.0
```

https://heiko-sieger.info/iommu-groups-what-you-need-to-consider/

StorPool DISTRIBUTED STORAGE

Congratulations!

You are ready to run

Nested Virtualization !

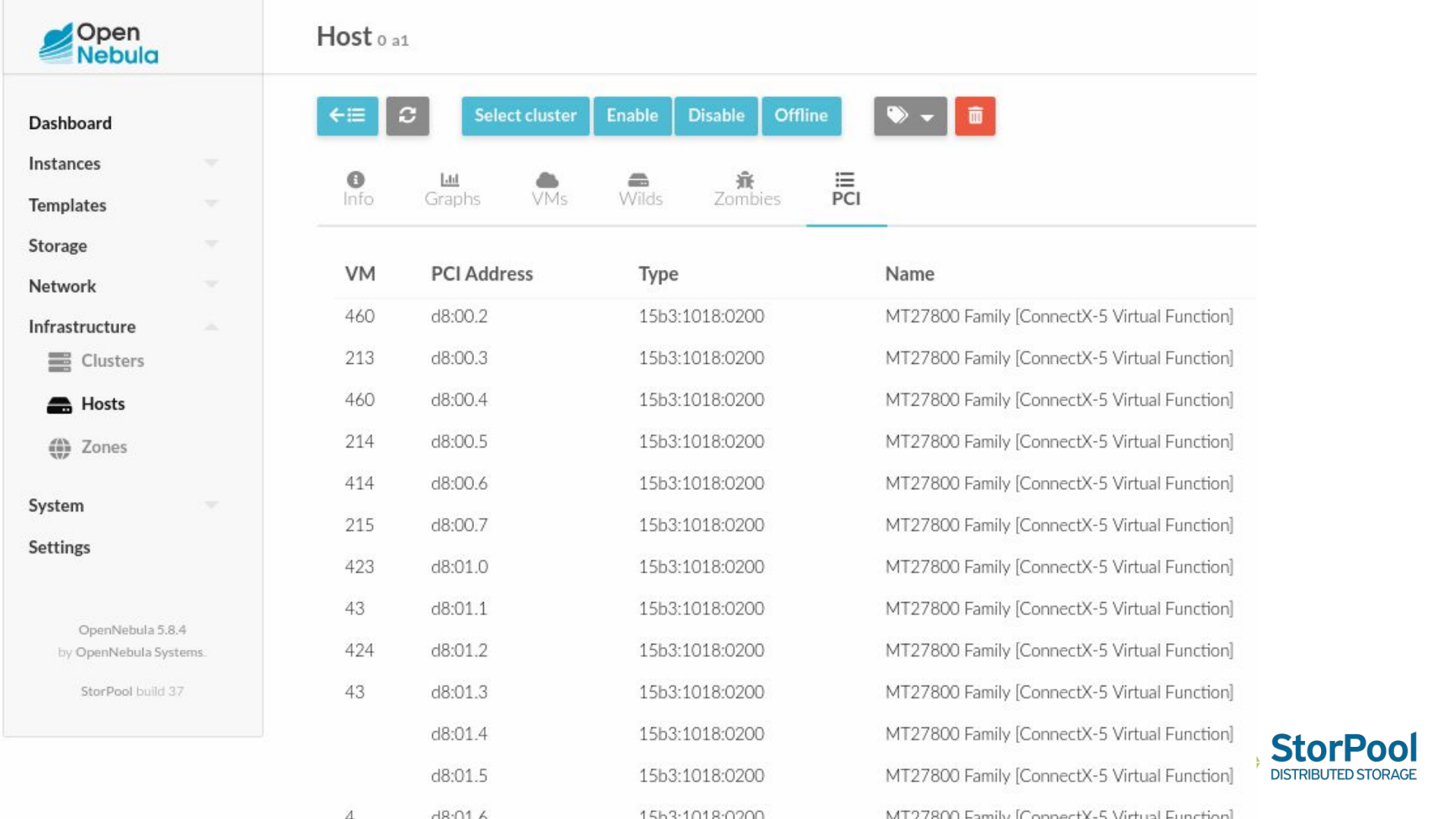
StorPool
DISTRIBUTED STORAGE

Enable PCI passthrough in OpenNebula

```
/var/lib/one/remotes/im/kvm-probes.d/pci.rb
/var/lib/one/remotes/etc/im/kvm-probes.d/pci.conf
```

```
:filter: '15b3:1018'
:short_address: []
:device_name: []
```

```
PCI = [
  TYPE = "15b3:1018:0200",
  VENDOR = "15b3",
  VENDOR_NAME = "Mellanox Technologies",
  DEVICE = "1018",
  DEVICE_NAME = "MT27800 Family [ConnectX-5 Virtual Function]",
  CLASS = "0200",
  CLASS_NAME = "Ethernet controller",
  ADDRESS = "0000:d8:17:5",
  SHORT_ADDRESS = "d8:17.5",
  DOMAIN = "0000",
  BUS = "d8",
  SLOT = "17",
  FUNCTION = "5"
]
```


Open Nebula
Host 0 a1
Dashboard
Instances
Templates
Storage
Network
Infrastructure
Clusters
Hosts
Zones
System
Settings
OpenNebula 5.8.4
by OpenNebula Systems
StorPool build 37
Select cluster
Enable
Disable
Offline
Info
Graphs
VMs
Wilds
Zombies
PCI
VM PCI Address Type Name
460 d8:00.2 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
213 d8:00.3 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
460 d8:00.4 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
214 d8:00.5 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
414 d8:00.6 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
215 d8:00.7 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
423 d8:01.0 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
43 d8:01.1 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
424 d8:01.2 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
43 d8:01.3 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
d8:01.4 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
d8:01.5 15b3:1018:0200 MT27800 Family [ConnectX-5 Virtual Function]
StorPool
DISTRIBUTED STORAGE

Tweak domain.xml

```
<hostdev mode='subsystem' type='pci' managed='yes'>
    <source>
      <address  domain='0x0000' bus='0xd8' slot='0x00' function='0x5'/>
    </source>
    <address type='pci' domain='0x0000' bus='0x01' slot='0x01' function='0'/>
</hostdev>
```

```
<interface managed="yes" type="hostdev">
  <driver name="vfio" />
  <mac address="02:00:11:ab:cd:01" />
  <source>
    <address bus="0xd8" domain="0x0000" function="0x5" slot="0x00" type="pci" />
  </source>
  <address bus="0x01" domain="0x0000" function="0" slot="0x01" type="pci" />
</interface>
```

https://github.com/OpenNebula/addon-storpool/blob/master/docs/advanced_configuration.md#vms-domain-xml-tweaking

StorPool DISTRIBUTED STORAGE

OpenNebula

141	cloudstackDevLocalKVM	slavka	oneadmin	RUNNING	16.1GB	a3	10.2.1.162	
140	cloudstackdev2KVM1	slavka	oneadmin	UNDEPLOYED	0KB	--	10.2.1.161	
138	OpenStack-dev-KVM-3	pp	oneadmin	RUNNING	1.6GB	a3	10.2.1.155	
137	OpenStack-dev-KVM-2	pp	oneadmin	RUNNING	1.5GB	a2	10.2.1.154 10.2.10.129	
135	OpenStack-dev-server	pp	oneadmin	RUNNING	1.6GB	a3	10.2.1.148	
127	OpenStack-dev-KVM-1	pp	oneadmin	POWEROFF	0KB	a1	10.2.1.160	
126	cloudstackdev1kvm2	ant	oneadmin	RUNNING	16.1GB	a2	10.2.1.159	
125	cloudstackdev1kvm1	ant	oneadmin	RUNNING	16.1GB	a2	10.2.1.158	
124	cloudstackdev1	ant	oneadmin	RUNNING	8.1GB	a2	10.2.1.157	
45	OpenStack-KVM-3	pp	oneadmin	RUNNING	16.1GB	a3	10.2.1.147	
44	OpenStack-KVM-2	pp	oneadmin	RUNNING	16.1GB	a2	10.2.1.145	
43	OpenStack-KVM-1	pp	oneadmin	RUNNING	16.1GB	a1	10.2.1.144	
38	OpenStack-Juju-controller	pp	oneadmin	RUNNING	16.1GB	a3	10.2.1.146	
28	CloudStackKvm2	ant	oneadmin	RUNNING	16.1GB	a2	10.2.1.135	
26	CloudStackKvm1	ant	oneadmin	RUNNING	16.1GB	a2	10.2.1.134	
20	CloudStack	ant	oneadmin	RUNNING	2.1GB	a3	10.2.1.140	

StorPool
DISTRIBUTED STORAGE

Summary

	CPU	BIOS	Kernel	KVM	OpenNebula
VT-x	□	□			
Nested Virt.				□	
VMCS Shadowing	□	□		□	
IOMMU (PCI Passthrough)	□	□	□		□
SR-IOV		□	□		
ACS (IOMMU groups)	□				
libvirt Domain XML (VF Net only)					□

Q&A

StorPool
DISTRIBUTED STORAGE

Thank you!

Venko Moyankov
venko@storpool.com

StorPool Storage
www.storpool.com
@storpool