The Mellanox ConnectX-4 Plugin for Fuel Documentation

Release 3.2-3.2.0-1

Mellanox

July 04, 2016

1 User documentation **1**
1.1 Definitions, Acronyms and abbreviations 1
1.2 Mellanox plugin 2
1.3 Release notes 3
1.4 Mellanox plugin configuration 3
1.5 Installation Guide 4
1.6 Post-deployment SR-IOV test scripts 8
1.7 Known issues 9
1.8 Supported images 10
1.9 Contact Support 10
1.10 Troubleshooting notes 10
1.11 Appendix 10

CHAPTER

ONE

USER DOCUMENTATION

1.1 Definitions, Acronyms and abbreviations

SR-IOV SR-IOV stands for "Single Root I/O Virtualization". It is a specification that allows a PCI device to appear virtually in multiple virtual machines (VMs), each of which has its own virtual function. The specification defines virtual functions (VFs) for the VMs and a physical function for the hypervisor. Using SR-IOV in a cloud infrastructure helps reaching higher performance since traffic bypasses the TCP/IP stack in the kernel.

iSER iSER stands for "iSCSI Extensions for RDMA". It is an extension of the data transfer model of iSCSI, a storage networking standard for TCP/IP. iSER enables the iSCSI protocol to take advantage of the RDMA protocol suite to supply higher bandwidth for block storage transfers (zero time copy behavior). To that fact, it eliminates the TCP/IP processing overhead while preserving the compatibility with iSCSI protocol.

RDMA RDMA stands for "Remote Direct Memory Access". It is a technology that enables to read and write data from remote server without involving the CPU. It reduces latency and increases throughput. In addition, the CPU is free to perform other tasks.

ConnectX-3 Pro ConnectX-3 Pro adapter cards with Virtual Protocol Interconnect (VPI) supporting InfiniBand and Ethernet connectivity provide the highest performing and most flexible interconnect solution for PCI Express Gen3 servers used in Enterprise Data Centers, High-Performance Computing, and Embedded environments.

Infiniband A computer-networking communications standard used in high-performance computing, features very high throughput and very low latency. It is used for data interconnect both among and within computers. InfiniBand is also utilized as either a direct, or switched interconnect between servers and storage systems, as well as an interconnect between storage systems.

VXLAN offload Virtual Extensible LAN (VXLAN) is a network virtualization technology that attempts to improve the scalability problems associated with large cloud computing deployments

QoS QoS is defined as the ability to guarantee certain network requirements like bandwidth, latency, jitter and reliability in order to satisfy a Service Level Agreement (SLA) between an application provider and end users.

VF VF is virtual NIC that will be available for VMs on Compute nodes.

OpenSM OpenSM is an InfiniBand compliant Subnet Manager and Administration, and runs on top of OpenIB. It provides an implementation of an InfiniBand Subnet Manager and Administration. Such a software entity is required to run for in order to initialize the InfiniBand hardware (at least one per each InfiniBand subnet).

PKey PKEY stands for partition key. It is a 16 bit field within the InfiniBand header called BTH (Base Transport Header). A collection of endnodes with the same PKey in their PKey Tables are referred to as being members of a partition.

ConnectX-4 ConnectX-4 adapter cards with Virtual Protocol Interconnect (VPI), supporting EDR 100Gb/s InfiniBand and 100Gb/s Ethernet connectivity, provide the highest performance and most flexible solution for high-performance, Web 2.0, Cloud, data analytics, database, and storage platforms.

NEO Mellanox NEO™ is a powerful platform for managing scale-out computing networks. Mellanox NEO™ enables data center operators to efficiently provision, monitor and operate the modern data center fabric.

1.2 Mellanox plugin

The Mellanox Fuel plugin is a bundle of scripts, packages and metadata that will extend Fuel and add Mellanox features such as SR-IOV for networking and iSER protocol for storage.

Fuel can configure Mellanox ConnectX-4 network adapters to accelerate the performance of compute and storage traffic.

This implements the following performance enhancements:

- **Compute nodes network enhancements:**
  - SR-IOV based networking
  - QoS for VM traffic
  - VXLAN traffic offload
- Cinder nodes use iSER block storage as the iSCSI transport rather than the default iSCSI over TCP.

These features reduce CPU overhead, boost throughput, reduce latency, and enable network traffic to bypass the software switch layer (e.g. Open vSwitch).

Mellanox Plugin integration with Mellanox NEO SDN Controller enables switch VLAN auto provisioning and port configuration for Ethernet and SM PK auto provisioning for InfiniBand networks, over private VLAN networks.

1.2.1 Developer's specification

Please refer to: HowTo Install Mellanox OpenStack Plugin for Mirantis Fuel 8.0

1.2.2 Requirements

Requirement	Version/Comment
Mirantis OpenStack compatibility	8.0

The Mellanox ConnectX-4 adapters family supports up to 100 Gb/s. To reach 100 Gb/s speed in your network with ConnectX-4 adapters, you must use Mellanox Ethernet / Infiniband switches supporting 100 Gb (e.g. SN2700 (ETH), SB7700 (IB)). The switch ports should be configured specifically to use 100 Gb speed. No additional configuration is required on the adapter side.

1.3 Release notes

1.4 Mellanox plugin configuration

If you plan to enable VM to VM RDMA and to use iSER storage transport you need to configure switching fabric to support the features.

1.4.1 Ethernet network:

1. Configure the required VLANs and enable flow control on the Ethernet switch ports.
2. All related VLANs should be enabled on the Mellanox switch ports (for relevant Fuel logical networks).
3. Login to the Mellanox switch by ssh and execute following commands:

Note: In case of using NEO auto provisioning, private network VLANs can be considered as dynamically configured.

```
switch > enable
switch # configure terminal
switch (config) # vlan 1-100
switch (config vlan 1-100) # exit
switch (config) # interface ethernet 1/1 switchport mode hybrid
switch (config) # interface ethernet 1/1 switchport hybrid allowed-vlan all
switch (config) # interface ethernet 1/2 switchport mode hybrid
switch (config) # interface ethernet 1/2 switchport hybrid allowed-vlan all
...
switch (config) # interface ethernet 1/36 switchport mode hybrid
switch (config) # interface ethernet 1/36 switchport hybrid allowed-vlan all
```

Flow control is required when running iSER (RDMA over RoCE - Ethernet). On Mellanox switches, run the following command to enable flow control on the switches (on all ports in this example)::

```
switch (config) # interface ethernet 1/1-1/36 flowcontrol receive on force
switch (config) # interface ethernet 1/1-1/36 flowcontrol send on force
```

save the configuration (permanently), run::

```
switch (config) # configuration write
```

Note: When using an untagged storage network for iSER over Ethernet - please add the following commands for Mellanox switches or use trunk mode instead of hybrid.

```
interface ethernet 1/1 switchport hybrid allowed-vlan add 1
interface ethernet 1/2 switchport hybrid allowed-vlan add 1
...
```

1.4.2 Infiniband network:

Mellanox **UFM** is a pre-requisite for using the Mellanox plugin for Fuel 8.0 with InfiniBand fabrics. Mellanox.s Unified Fabric Manager (UFM®) is a powerful platform for managing scale-out computing environments. UFM

enables data center operators to monitor, efficiently provision, and operate the modern data center fabric. UFM is licensed per managed fabric node. For more information on how to obtain UFM, please visit Mellanox.com.

Update OpenSM configurations on UFM node as follows:

1. Update opensm.conf file and make sure of the following:

```
vim /opt/ufm/conf/opensm/opensm.conf
- virt_enabled 2
- no_partition_enforcement TRUE
- part_enforce off
- allow_both_pkeys FALSE
```

2. Update the partitions.conf file:

```
vim /opt/ufm/conf/partitions.conf.user_ext

Example:

vlan1=0x1, ipoib, sl=0, defmember=full: ALL_CAS;
vlan2=0x2, ipoib, sl=0, defmember=full: ALL_CAS;
vlan3=0x3, ipoib, sl=0, defmember=full: ALL_CAS;

vlan4=0x4, ipoib, sl=0, defmember=full: SELF;
vlan5=0x5, ipoib, sl=0, defmember=full: SELF;
vlan6=0x6, ipoib, sl=0, defmember=full: SELF;
vlan7=0x7, ipoib, sl=0, defmember=full: SELF;
vlan8=0x8, ipoib, sl=0, defmember=full: SELF;
vlan9=0x9, ipoib, sl=0, defmember=full: SELF;
. . .
vlan20=0x14, ipoib, sl=0, defmember=full: SELF;
```

Note: In this example

- Infra networks VLANs are 1-3 so VLAN2 is assigned to PK 0x2 and will be used for Openstack Management network and VLAN3 is assigned to PK 0x3 and will be used for Openstack Storage network.
- Private VLANs are 4-20 so VLANs 4 through 20 are assigned to PKs 0x4 to 0x14 will be used for Tenant networks.
- VLAN1 is defined, but not used for consistency with Ethernet setup installation.
- The maximum number of VLANs is 128.

1. Restart ufmd:

```
/etc/init.d/ufmd restart
```

1.5 Installation Guide

To install Mellanox plugin, follow these steps:

1. Install Fuel Master node. For more information on how to create a Fuel Master node, please see Mirantis Fuel 8.0 documentation.
2. Download the plugin rpm file for MOS 8.0 from Fuel Plugin Catalog.
3. Copy the plugin on already installed Fuel Master. scp can be used for that.:

```
# scp mellanox-plugin-3.2-3.2.0-1.noarch.rpm root@<Fuel_Master_ip>:/tmp
```

4. Install the plugin:

```
# cd /tmp
# fuel plugins --install mellanox-plugin-3.2-3.2.0-1.noarch.rpm
```

5. Verify the plugin was installed successfully by having it listed using `fuel plugins` command:

```
# fuel plugins
#  id | name              | version | package_version
#  ---|-------------------|---------|----------------
#  1  | mellanox-plugin   | 3.2.0   | 3.0.0
```

6. Define bootstrap discovery parameters to be burnt on Mellanox Adapters cards:
   - **link_type** , available link_type values are:
     - `eth` for changing link type to Ethernet
     - `ib` for changing link type to Infiniband
     - `current` for leaving link type as is
   - **max_num_vfs** as integer, default is set to 16.
7. Create Bootstrap discovery image for detecting Mellanox HW and support related configurations with pre-defined parameters:

```
[root@fuel ~]# create_mellanox__bootstrap --link_type $link_type --max_num_vfs $max_num_vfs
[root@fuel ~]# create_mellanox_bootstrap --help
```

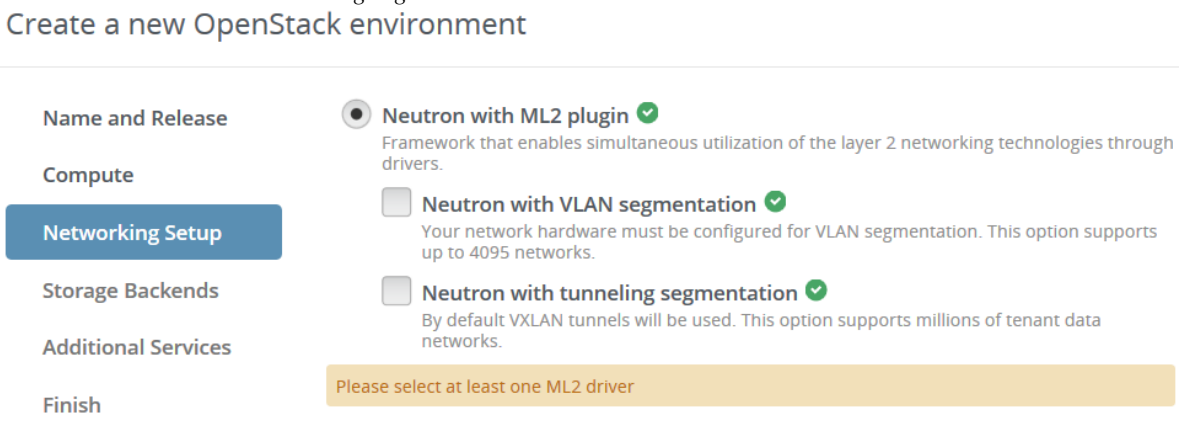
```
usage: create_mellanox_bootstrap [-h] [--link_type {eth,ib,current}]
                                 [--max_num_vfs MAX_NUM_VFS]
Available link_type values are:
------------------------------
- eth for changing link type to Ethernet
- ib for changing link type to Infiniband
- current for leaving link type as is

optional arguments:
  -h, --help            show this help message and exit
  --link_type {eth,ib,current}
  --max_num_vfs MAX_NUM_VFS
                        an integer for the maximum number of vfs to be burned in bootstrap


::

  Try to build image with data:
  bootstrap:
  certs: null
  container: {format: tar.gz, meta_file: metadata.yaml}
  . . .
  . . .
  . . .
  Bootstrap image f790e9f8-5bc5-4e61-9935-0640f2eed949 has been activated.
```

1. Reboot nodes after installing plugin:

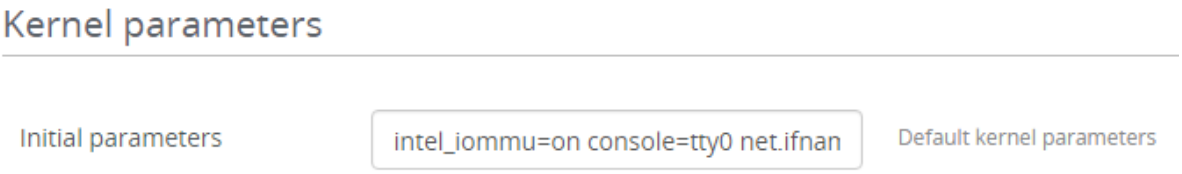
```
[root@fuel ~]# reboot_bootstrap_nodes -a
[root@fuel ~]# reboot_bootstrap_nodes -h
```

```
Usage: reboot_bootstrap_nodes [-e environment_id] [-h] [-a]
   This script is used to trigger reboot for nodes in 'discover' status,
   of a given environment (if given) or of all environments.
   Please wait for nodes to boot again after triggering this script.

Options:

-h          Display the help message.
-e <env>    Reboot all nodes in state 'discover' of the given environment.
-a          Reboot all nodes in state 'discover' of all environments.
```

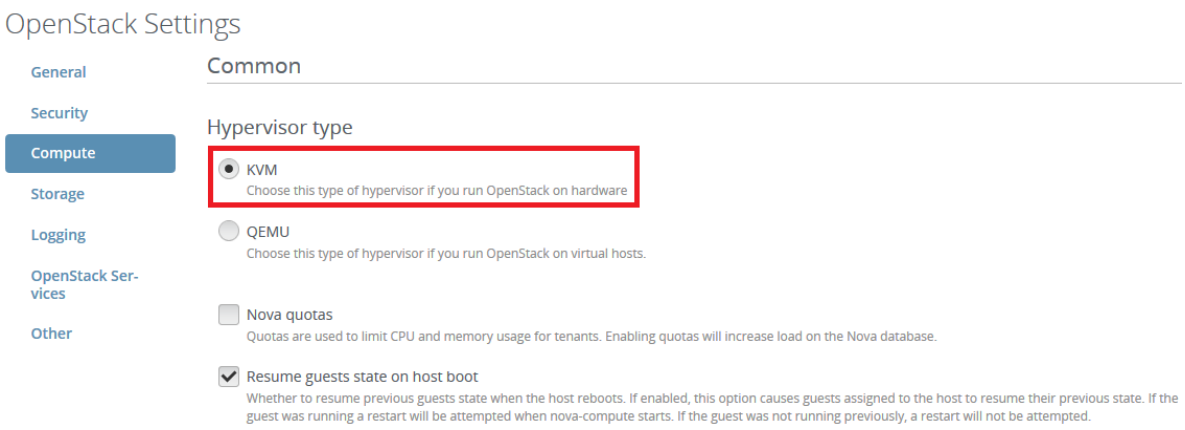
1. Create an environment - for more information please see how to create an environment. We support both main network configurations:
   - *Neutron with VLAN segmentation*
   - *Neutron with tunneling segmentation*



2. Adjust the kernal parameters in the settings tab which is a condition for both iSER and SRIOV. Open the Settings tab, select General section and then add `intel_iommu=on` at the beginning of the initial parameters.



3. Enable KVM hypervisor type. KVM is required to enable Mellanox Openstack features. Open the Settings tab, select Compute section and then choose KVM hypervisor type.



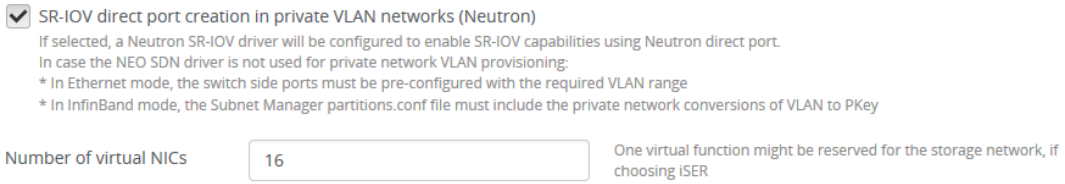
4. Enable desired Mellanox Openstack features. Open the Other tab. Enable Mellanox features by selecting Mellanox Openstack features checkbox. Select relevant plugin version if you have multiple versions installed.

Now you can enable one or more features relevant for your deployment:

(a) Support SR-IOV direct port creation in private VLAN networks **Note**: Relevant for *VLAN segmentation* only

- This enables Neutron SR-IOV support.
- **Number of virtual NICs** is amount of virtual functions (VFs) that will be available on Compute node.

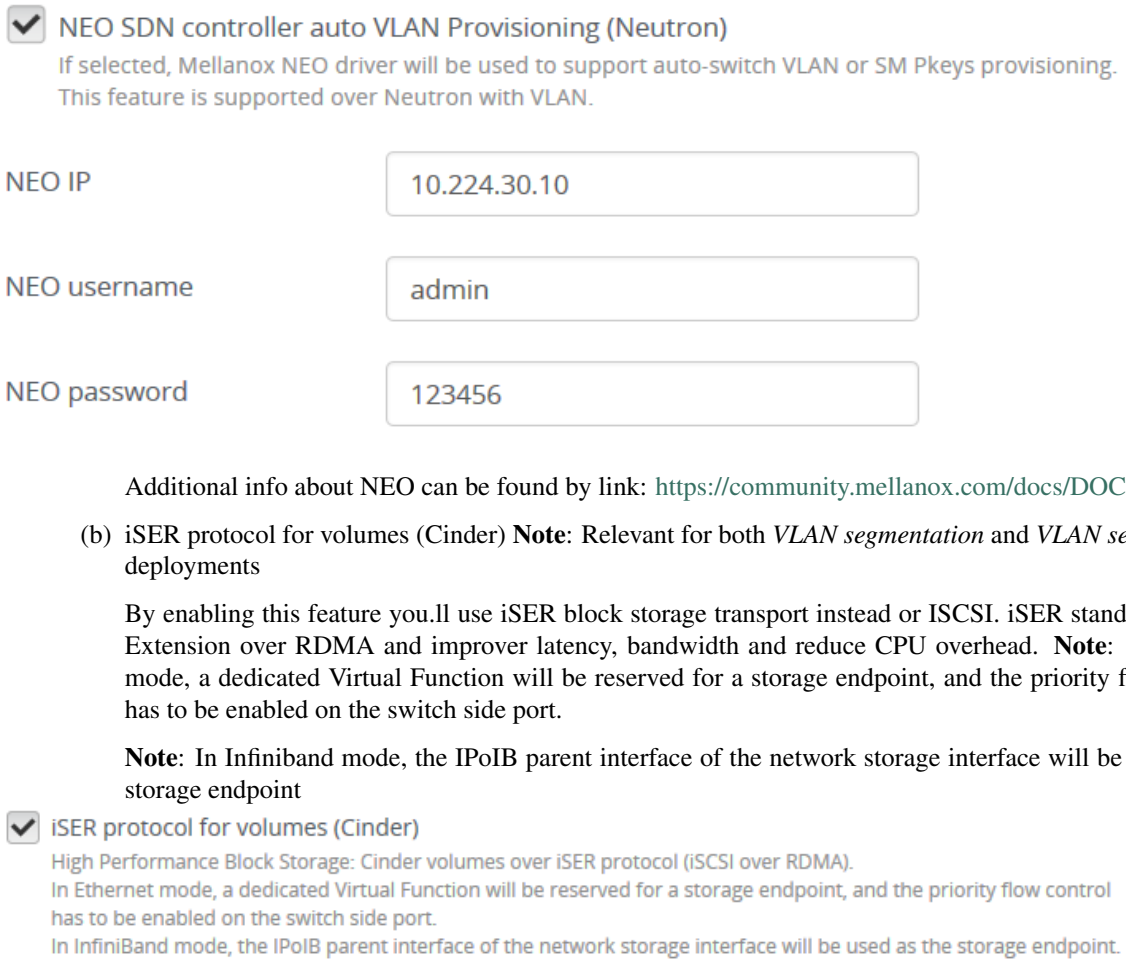
Note: One VF will be utilized for iSER storage transport if you choose to use iSER. In this case you will get 1 VF less for Virtual Machines.



(a) Support NEO SDN controller auto VLAN Provisioning (Neutron) **Note**: Relevant for *VLAN segmentation* only

If selected, Mellanox NEO Mechanism driver will be used in order to support Auto switch VLAN auto-provisioning for Ethernet network

To use this feature please provide IP address, username and password for NEO SDN controller.



Additional info about NEO can be found by link: https://community.mellanox.com/docs/DOC-2155

(b) iSER protocol for volumes (Cinder) **Note**: Relevant for both *VLAN segmentation* and *VLAN segmentation* deployments

By enabling this feature you.ll use iSER block storage transport instead or ISCSI. iSER stands for ISCSI Extension over RDMA and improver latency, bandwidth and reduce CPU overhead. **Note**: In Ethernet mode, a dedicated Virtual Function will be reserved for a storage endpoint, and the priority flow control has to be enabled on the switch side port.

Note: In Infiniband mode, the IPoIB parent interface of the network storage interface will be used as the storage endpoint

---

Note: When configuring Mellanox plugin, please mind the following:

---

1. You *cannot* install a plugin for existing environment without the plugin support. That means, the plugin will appear in the certain environment only if the plugin was installed before creating the environment. You can upgrade the plugin for existing non-deployed environments.
2. Enabling the mellanox Openstack features hardware support on your environment, regardless of the chosen Mellanox features.
3. In Ethernet cloud, when using SR-IOV & iSER, one of the virtual NICs for SR-IOV will be reserved to the storage network.
4. When using SR-IOV you can set the number of virtual NICs (virtual functions) to up to 31 if your hardware and system capabilities like memory and BIOS support it). In any case of SR-IOV hardware limitation, the installation will try to fallback a VF number to the default of 16 VFs.

1.6 Post-deployment SR-IOV test scripts

In order to test that SR-IOV is working properly **after** deploying an OpenStack environment with SR-IOV support **successfully**, a couple of scripts have been added under /sbin/:

Note: Please use the 2 last commands with caution, since they can delete some of your environment ports and image resources.

- **upload_sriov_cirros**

  Uploads a pre-configured Mellanox Cirros image to glance images list.

- **start_sriov_vm**

  For starting a VM with direct port from previous image. In order to test that SR-IOV is working properly, start two SR-IOV VM.s and make sure you have ping between these nodes. Assumes upload_sriov_cirros was executed before.

- **delete_sriov_ports**

  Deletes all SR-IOV ports created in previous scripts.

- **delete_all_glance_images**

  Deletes all Glance images.

1.7 Known issues

Issue 1

- Description: For custom (OEM) adapter cards based on Mellanox ConnectX-4 ICs, adapter firmware must be manually burnt prior to the installation with SR-IOV support
- Workaround: See the firmware installation instructions.

Issue 2

- Description: The number of SR-IOV virtual functions supported by Mellanox adapters is up to 31 on ConnectX-4 adapters (depends on your HW capabilities).
- Workaround: NA

Issue 3

- Description: When using a dual port physical NIC for SR-IOV over Ethernet, the Openstack private network has to be allocated on the first port.
- Workaround: NA

Issue 4

- Description: Changing port type in bootstrap stage over a single port HCA is not supported
- Workaround: Create a bootstrap image with link type current, and change the port type manually.

Issue 5

- Description: Starting large amount (>15) of IB VMs with normal port at once may result in some VMs not getting DHCP over InfiniBand networks.
- Workaround: Reboot VMs that didn't get IP from DHCP on time or start VMs in smaller chunks (<10).

Issue 6

- Description: Network verification for IB network is not supported over untagged networks or after deployment.
- Workaround: NA

Issue 7

- Description: When using NEO auto provisioning, network verification should fail for the private network VLANs
- Workaround: NA

Issue 8

- Description: When deploying an Infiniband cluster with iSER over VLAN, all controllers should be deployed at once.
- Workaround: Use untagged storage network when using Infiniband with iSER over VLAN, or deploy all controllers at once.

1.8 Supported images

Issue	Supported OS	Tested kernel
1	CentOS7	3.10.0-327.13.1.e17.x86_64
2	ubuntu14.04	3.13.0-85-generic
3	Cirros Mellanox	3.11.0-26-generic

This Fuel Mellanox plugin ver. 3.2-3.2.0-1 is using MLNX_OFED_LINUX version 3.3-1.5.0.

1.9 Contact Support

For reporting a problem, suggesting an enhancements or new features please contact Mellanox support for help (support@mellanox.com).

1.10 Troubleshooting notes

- Please verify your network configurations prior to the deployment.
- Please make sure all your Health check Tests are passing.
- To make sure SR-IOV is working properly, please refer to user scripts mentioned previously.
- **Mellanox Plugin log file is located on each slave node on the following path:**
  - /var/log/Mellanox-plugin.log
- **For further information you can check the relevant logs too:**
  - /var/log/docker-logs/astute/astute.log (fuel-master)
  - /var/log/dmesg (target nodes)
  - /var/log/messages (target nodes)
- For debugging Ethernet or Infinband driver issues, please deploy with Openstack debug logging enabled.

1.11 Appendix

Mellanox site where users can read about possible configurations:

- Mellanox ConnectX-4

- Mellanox ConnectX-3 pro
- HowTo Install Mirantis Fuel OpenStack with Mellanox
- Mellanox InfiniBand Switches
- Mellanox NEO