

---

# **The Mellanox ConnectX-3 pro Plugin for Fuel Documentation**

*Release 3.0-3.0.0-1*

**Mellanox**

March 31, 2016

## CONTENTS

<b>1</b>	<b>User documentation</b>	<b>1</b>
1.1	Definitions, Acronyms and abbreviations	1
1.2	Mellanox plugin	2
1.3	Mellanox plugin configuration	3
1.4	Installation Guide	4
1.5	Post-deployment SR-IOV test scripts	7
1.6	Known issues	8
1.7	Supported images	9
1.8	Contact Support	9
1.9	Troubleshooting notes	9
1.10	Appendix	9

## 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.

## 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-3 Pro](#) 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-3 Pro adapters family supports up to 40/56 Gb. To reach 56 Gb speed in your network with ConnectX-3 Pro adapters, you must use Mellanox Ethernet / Infiniband switches supporting 56 Gb (e.g. SX1710, SX6710). The switch ports should be configured specifically to use 56 Gb speed. No additional configuration is required on the adapter side. For additional information about how to run in 56GbE speed, see [HowTo Configure 56GbE Link on Mellanox Adapters and Switches](#).

### 1.2.3 Limitations

- Mellanox SR-IOV is supported only when choosing Neutron with VLAN segmentation.
- ConnectX-3 Pro adapters are required in order to enable VXLAN HW offload over Ethernet networks.
- QoS feature is implemented only for Ethernet VLAN SR-IOV ports using ConnectX-3 Pro adapters.

- Infiniband is configured by using OpenSM only.

## 1.3 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.

### 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
...
```

**Infiniband network:** If you use OpenSM you need to enable virtualization and allow all PKeys:

1. Create a new opensm.conf file:

```
opensm -c /etc/opensm/opensm.conf
```

2. Enable virtualization by editing /etc/opensm/opensm.conf and changing the allow\_both\_pkeys value to TRUE.:

```
allow_both_pkeys TRUE
```

3. Define the partition keys which are analog for Ethernet VLAN. Each VLAN will be mapped to one PK. Add/Change the following with the command

```
vi /etc/opensm/partitions.conf file:
(Example)
management=0x7fff,ipoib, sl=0, defmember=full : ALL, ALL_SWITCHES=full,SELF=full;
vlan1=0x1, ipoib, sl=0, defmember=full : ALL;
vlan2=0x2, ipoib, sl=0, defmember=full : ALL;
. . .
vlan100=0x64, ipoib, sl=0, defmember=full : ALL;
```

4. Restart OpenSM:

```
/etc/init.d/opensmd restart
```

## 1.4 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-3.0-3.0.0-1.noarch.rpm root@<Fuel_Master_ip>:/tmp
```

4. Install the plugin:

```
# cd /tmp
# fuel plugins --install mellanox-plugin-3.0-3.0.0-1.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 | 3.0.0 | 3.0.0
```

6. Create new bootstrap image for supporting infiniband networks (create\_mellanox\_vpi\_bootstrap can be used)::

```
[root@fuel ~]# create_mellanox_vpi_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.

7. In case of using the customized bootstrap image, you must reboot your target nodes with the new bootstrap image you just created. If you already have discovered nodes you can either reboot them manually or use :bash: `reboot_bootstrap_nodes` command. Run :bash: `reboot_bootstrap_nodes -h` for help.

8. 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*

### Create a new OpenStack environment

The screenshot shows the 'Networking Setup' step of the OpenStack environment creation process. On the left, a sidebar lists steps: Name and Release, Compute, Networking Setup (highlighted), Storage Backends, Additional Services, and Finish. The main area shows three radio button options, each with a green checkmark icon:
 

- Neutron with ML2 plugin** (selected): Framework that enables simultaneous utilization of the layer 2 networking technologies through drivers.
- Neutron with VLAN segmentation**: Your network hardware must be configured for VLAN segmentation. This option supports up to 4095 networks.
- Neutron with tunneling segmentation**: By default VXLAN tunnels will be used. This option supports millions of tenant data networks.

 A yellow warning bar at the bottom states: "Please select at least one ML2 driver".

9. 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.

### OpenStack Settings

The screenshot shows the 'OpenStack Settings' page with the 'Compute' tab selected. The 'Common' section is active, showing the 'Hypervisor type' configuration. The 'KVM' option is selected (radio button) and highlighted with a red rectangular box. Below it, the text reads: "Choose this type of hypervisor if you run OpenStack on hardware". Other options include 'QEMU' (unselected), 'Nova quotas' (checkbox), and 'Resume guests state on host boot' (checkbox).

10. 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.

**Mellanox Openstack Features**

Versions  3.0.0

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

- 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.

- SR-IOV direct port creation in private VLAN networks (Neutron)  
 If selected, a Neutron SR-IOV driver will be configured to enable SR-IOV capabilities using Neutron direct port.  
 In case the NEO SDN driver is not used for private network VLAN provisioning:  
 \* In Ethernet mode, the switch side ports must be pre-configured with the required VLAN range  
 \* In InfinBand mode, the Subnet Manager partitions.conf file must include the private network conversions of VLAN to PKey

Number of virtual NICs  One virtual function might be reserved for the storage network, if choosing iSER

- (a) Support quality of service over VLAN networks with Mellanox SR-IOV direct ports (Neutron) **Note:** Relevant for *VLAN segmentation* only  
 If selected, Neutron “Quality of service” (QoS) will be enabled for VLAN networks and ports over Mellanox HCAs. **Note:** This feature is supported only if:

- Ethernet mode is used
- SR-IOV is enabled

- Quality of Service over Mellanox SR-IOV ports (Neutron)  
 If selected, Neutron "Quality of Service" (QoS) will be enabled for SR-IOV ports over Mellanox HCAs.  
 This feature is supported in Ethernet mode over Neutron with VLAN after SR-IOV direct port creation.

- (b) 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.

- NEO SDN controller auto VLAN Provisioning (Neutron)  
 If selected, Mellanox NEO driver will be used to support auto-switch VLAN or SM Pkeys provisioning.  
 This feature is supported over Neutron with VLAN.

NEO IP

NEO username

NEO password

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

- (c) Support VXLAN Offloading (Neutron) **Note:** Relevant for *tunneling segmentation* only

If selected, Mellanox hardware will be used to achieve a better performance and significant CPU overhead reduction using VXLAN traffic offloading.

- VXLAN Offloading (Neutron)  
 If selected, Mellanox hardware will be used to achieve a better performance and significant CPU overhead reduction using VXLAN traffic offloading. This feature is supported in Ethernet mode over Neutron with tunneling.

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



By enabling this feature you will use iSER block storage transport instead of iSCSI. iSER stands for iSCSI Extension over RDMA and improves latency, bandwidth and reduces 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

iSER protocol for volumes (Cinder)

High Performance Block Storage: Cinder volumes over iSER protocol (iSCSI over RDMA).

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.

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 62 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 8 VFs.

## 1.5 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 VMs 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.

## 1.6 Known issues

### Issue 1

- Description: For custom (OEM) adapter cards based on Mellanox ConnectX-3 / ConnectX-3 Pro 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 16 on ConnectX-3 adapters and up to 62 on ConnectX-3 Pro 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: A single port HCA might not be supported for SRIOV and iSER over Ethernet network.
- Workaround: NA

### Issue 5

- Description: SR-IOV QoS is supported only with updating SR-IOV existing ports with a policy. QoS-policy detach might result in non accurate bandwidth limit. (<https://bugs.launchpad.net/neutron/+bug/1504165>).
- Workaround: Delete port / instance and attach a new port.

### Issue 6

- 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 7

- Description: After large InfinBand deployment of more than ~20 nodes at once with Controllers HA, it might take time for controllers services to stabilize.
- Workaround: Restart openibd service on controller nodes after the deployment, or deploy with phases.

### Issue 8

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

## 1.7 Supported images

Issue	Supported OS	Tested kernel
1	CentOS7	3.10.0-229.14.1.el7.x86_64
2	ubuntu14.04	3.13.0-67-generic
3	Cirros Mellanox	3.11.0-26-generic

This Fuel Mellanox plugin ver. 3.0-3.0.0-1 is using MLNX\_OFED version 3.1-1.5.5.

## 1.8 Contact Support

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

## 1.9 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 Infiniband driver issues, please deploy with Openstack debug logging enabled.

## 1.10 Appendix

Mellanox site where users can read about possible configurations:

- [Mellanox ConnectX-3 pro](#)
- [HowTo Install Mirantis Fuel OpenStack with Mellanox](#)
- [Mellanox InfiniBand Switches](#)