InfiniBand Cluster Bring-up Procedure

v1.0

Table of Contents

1 Overview

This document is intended for network operators responsible for the bring-up of InfiniBand (IB) clusters. The purpose of this document is to outline the necessary automation tools, required tests, and essential information needed when installing a new cluster. Additionally, the document provides recommendations and guidance on how to obtain the necessary inputs for these procedures and how to execute the bring-up operations effectively. The document's content is structured logically to facilitate easy reference and understanding.

To complete the bring-up process, please follow the checklist in the following link.

2 Related Documentation

Document	Link
ibdiagnet InfiniBand Fabric Diagnostic Tool User Manual	https://docs.nvidia.com/networking/software/management-software/index.html#infiniband-management-tools
MLNX_OFED User Manual	https://docs.nvidia.com/networking/software/adapter-software/index.html#mlnx-ofed
MLNX-OS User Manual	https://docs.nvidia.com/networking/software/switch-software/index.html#mlnx-os-infiniband
MFT User Manual	https://docs.nvidia.com/networking/software/firmware-management/index.html#mft
UFM User Manual	https://docs.nvidia.com/networking/software/management-software/index.html#nvidia-ufm
HPC-X User Manual	https://docs.nvidia.com/networking/software/accelerator-software/index.html#hpc-x

3 Document Revision History

For the list of changes made to this document, refer to Document Revision History.

4 Cluster Planning

Effective cluster planning lays the groundwork for an efficient system and good monitoring and debugging ability.

The planning procedures in this document are arranged in a sequential order for a new cluster installation. If you are not installing a new cluster, you might need to choose which procedures to use. However, you should still perform them in the order they appear in the Cluster planning section.

To plan your cluster, complete the following procedures:

1. Set the Topology type. For information on how to do so, see Setting the InfiniBand Cluster Topology.
2. Create a Point-to-Point excel file. For information on how to do so, see Creating a Point-to-Point Excel File.
3. Create a Topology file. For information on how to do so, see Creating a Topology File.
4. Save the Topology file for future usage. For information on how to do so, see Saving the Topology File.
5. Confirm the Topology was created and set as desired. For information on how to do so, see Confirming Topology.

When you are ready to install the components with which you plan to build your cluster, Confirm Components' Firmware and Software Versions. If update is needed to one of the components, review the information in section On-site Upgrade – Low scale to ensure that you have the latest information.

4.1 Setting the InfiniBand Cluster Topology

InfiniBand fabric components can be connected using different topologies and it should be decided before building the cluster.

Fat-Tree is NVIDIA's recommended topology, AI factory should based on rail optimized.
Rail-optimized design means a GPU node with multiple interfaces will put each GPU “rail” (IB network interface) onto a different first level (LEAF) switch for cluster Interconnect. This allows multiple nodes to utilize their internal NVSwitch path to talk across a NIC that is just one switch hop away (instead of having to cross multiple switches, incurring additional latency).

The diagram below shows an example of cluster topology for AI factory (based on rail optimized):

Figure 1: 3 Tier Topology

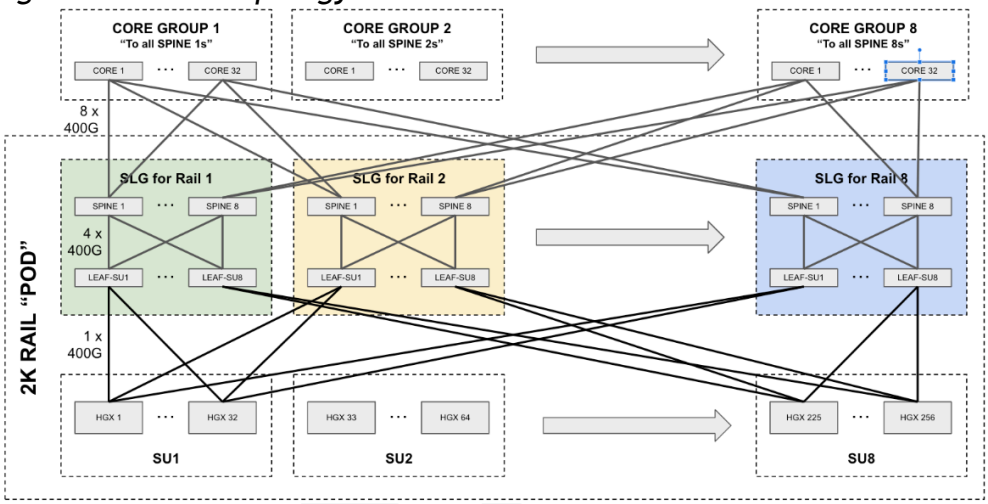
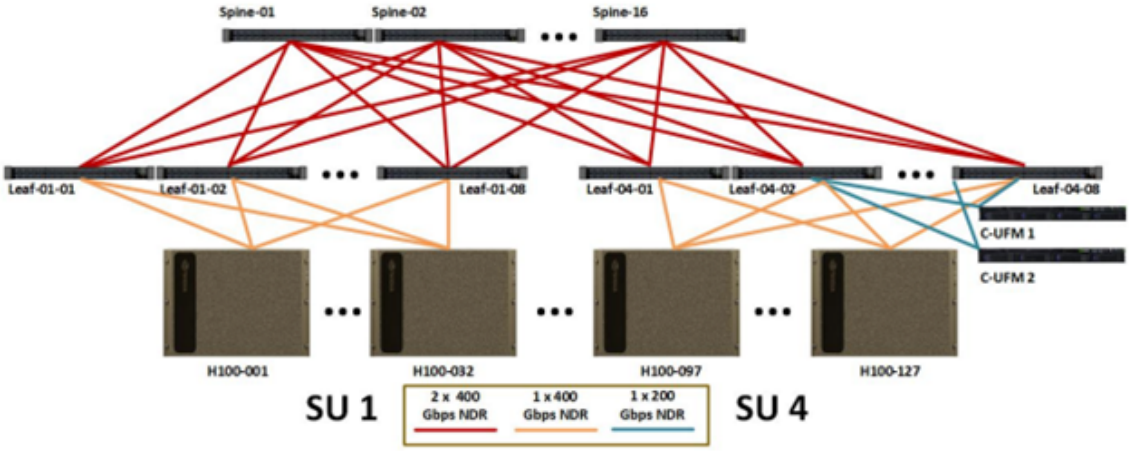


Figure 2: 2 Tier Topology



To choose the best cluster planning that fits the cluster needs, please contact NVIDIA Support.

After selecting the InfiniBand interconnect principles, create a PTP excel file to describe the cluster connectivity and to generate the Topology file.

> ⚠ It is imperative to verify that the cluster has been connected according to the cluster planning to ensure cluster maintenance.

4.2 Creating a Point-to-Point Excel File

The Point-to-Point Excel file centralizes all the physical information of the project and explicitly describes how to connect each cable. For the list of supported cables, see LinkX Cables and Transceivers | NVIDIA.

To create the Excel file:

1. Open an Excel file (use this template file).
2. Create 2 sheets as explained below:

- Legend – describes basic properties for each element of the cluster. Each element should include the following properties:
  - Name – describes the naming convention for each element, best practice is to include the element basic name and * before and after the name
  - Model – element model

    > ⚠ The Model is the device format as described inside the `"/usr/share/ibdm2.1.1/ibnl"`. If the model used is not part of the supported list, please create a new one as follow:
    >
    > https://linux.die.net/man/1/ibdm-topo-file
    >
    > https://linux.die.net/man/1/ibdm-ibnl-file

  - Switch/HCA - whether it is a switch or HCA
  - Speed – element speed
  - Comments – general comments

    Example:

Name	Model	Switch/HCA	Speed	Comments
dgx	HCA_12	hca	4x-100G	NDR
clf	MQM9700	switch	4x-100G	NDR
csp	MQM9700	switch	4x-100G	NDR

- PTP - explicitly describes how to connect each cable. The table has two main parts, Source and Destination, each one contains mostly the same columns. Each Line should include the following for each end of the cable:
  - Rack - device rack
  - U - device location in the rack
  - Name – name of the device (must comply with the naming convention as specified for the device type in the Label sheet)
  - HCA/port - HCA name and port (in Destination part only port)

    For example:

Source				Destination			
Rack	U	Name	HCA/port	Rack	U	Name	Port
SU1-1 A22	3	cl02s01dgx01	1	Leaves SU1 A38	25	cl02s01clf01	1
SU1-1 A22	3	cl02s01dgx01	2	Leaves SU1 A38	27	cl02s01clf02	1
SU1-1 A22	3	cl02s01dgx01	3	Leaves SU1 A38	29	cl02s01clf03	1
SU1-1 A22	3	cl02s01dgx01	4	Leaves SU1 A38	31	cl02s01clf04	1
SU1-1 A22	3	cl02s01dgx01	5	Leaves SU1 A38	33	cl02s01clf05	1
SU1-1 A22	3	cl02s01dgx01	6	Leaves SU1 A38	35	cl02s01clf06	1
SU1-1 A22	3	cl02s01dgx01	7	Leaves SU1 A38	37	cl02s01clf07	1
SU1-1 A22	3	cl02s01dgx01	8	Leaves SU1 A38	39	cl02s01clf08	1

Please note that:

- Destination device should always be a switch (HCAs should always be specified in source)
- For switches, use real/physical port numbers
- HCA ports can be named/enumerated as you wish, and you have to verify that there is a proper mapping from HCA port enumeration to real HCA interface name (will be referred in next step page)

⚠ In the provided examples, the element name *dgx* denotes the device with the identifier cl02s01dgx01.

⚠ Make sure to have clear and meaningful names, a well described element, its role, and its location in both the topology and in the cluster.

4.3 Creating a Topology File

The topology file describes the connection between the different cluster elements. It ensures standard documentation of the topology plan, verifies the implementation matches the topology plan (deployment/maintenance phases) and helps mapping and visualizing the topology when fixing problems.

To generate the topo file from the Point-To-Point (PTP) xls file, download the python script from here.

Within the script file (in the source code, at the top), there's a dictionary named 'hcaPortMapping'. It maps each HCA port enumeration (from previous section) to the real HCA interface name, in mlx format, for example, mlx5_1/P1.

The dictionary structure is as the following:

```
hcaPortMapping = {
    'U1/P<HCA-enumeration1>': '<real-HCA-interface1>/P1',
    'U1/P<HCA-enumeration2>': '<real-HCA-interface2>/P1',
    ...
}
```

For example:

```
hcaPortMapping = {
    'U1/P1': 'mlx5_3/P1',
    'U1/P2': 'mlx5_1/P1'
}
```

It means that:

- in PTP file, where we specified a connection of host with enumerated HCA port 1, it means that this connection is attached to interface mlx5_3 of the host
- in PTP file, where we specified a connection of host with enumerated HCA port 2, it means that this connection is attached to interface mlx5_1 of the host

Before running the parser script, make sure that 'hcaPortMapping' maps properly all the HCA ports enumeration from your PTP file to the proper HCA interfaces, as described above.

To create the topology file, use the `parse_ptp_file.py` script:

```
python parse_ptp_file.py parse -f ptp-data.xls -of -mhp
```

The script outputs the topo file "output.topo". Give it a meaningful name (name, date, etc.), and store/save it in a reachable location for a future use (this file is used for validations, etc.).

Example of a topology file:

```
MQM9700 c102s01c1f01 CFG : main=4x
    P1 -4x-100G-> HCA_mlx5_8 c10201dgx01 U1/P1
    P2 -4x-100G-> HCA_mlx5_8 c102s01dgx02 U1/P1
    P3 -4x-100G-> HCA_mlx5_8 c102s01dgx03 U1/P1
    P4 -4x-100G-> HCA_mlx5_8 c10201dgx04 U1/P1
    P5 -4x-100G-> HCA_mlx5_8 c102s01dgx05 U1/P1
    P6 -4x-100G-> HCA_mlx5_8 c102s01dgx06 U1/P1
    P7 -4x-100G-> HCA_mlx5_8 c10201dgx07 U1/P1
    P8 -4x-100G-> HCA_mlx5_8 c102s01dgx08 U1/P1
```

4.4 Saving the Topology File

After creating the Point-to-Point excel file and the topology file, make sure to save the file for future usage.

> ⚠ If you do not save the file, you will have to repeat the cluster panning steps the next time you bring-up the cluster.

Facilitating the topology enables you to apply the master topology settings to the current topology, when the final topology is reached. This ensures that the fabric is in a consistent and optimal state. The master topology is a reference point that represents the desired state of the fabric. It can be set by selecting the latest topology or by uploading a predefined custom topology from a file. Periodic comparison allows users to compare the current fabric topology with a preset master topology. By comparing the current topology with the master topology, users can detect any deviations, errors, or anomalies in the fabric and take corrective actions if needed.

4.5 Confirming Topology

The integrity of the topology is essential to ensure the reliability and performance of the network.

In this section we ensure the physically deployed topology matches the original design.

The confirmation process requires to:

- Install UFM
- Confirming topology using UFM

These steps are described in the sub-sections of this chapter.

For doing the process using ibtopodiff instead, see https://docs.nvidia.com/networking/display/ibdiagnetUserManualv211/Topology+Comparison.

4.5.1 UFM Enterprise Installation

NVIDIA Unified Fabric Manager (UFM) is a powerful platform for managing InfiniBand scale-out computing environments.
UFM enables data center operators to efficiently monitor and operate the entire fabric, boost application performance and maximize fabric resource utilization.

> ⚠ If you do not have a valid license, please fill out the NVIDIA Enterprise Account Registration form to get a UFM evaluation license.
>
> Save the license file on the master server at /tmp/license_file/

> ⚠ Before installing UFM server software in High Availability mode, ensure that the requirements at this link are met.

> ⚠ UFM HA package requires a dedicated partition with the same size and name for DRBD on both servers.

> ⚠ After installing the UFM server software, make sure to configure the fabric_interface parameter in gv.cfg.
>
> The fabric interface should be set to one of the InfiniBand IPoIB interfaces, which connect the UFM to the fabric.

4.5.1.1 Installing UFM on Docker Container - High Availability Mode

4.5.1.1.1 Pre-deployments requirements

- Install pacemaker, pcs, and drbd-utils on both servers
  For Ubuntu:

  ```
  apt install pcs pacemaker drbd-utils
  ```

  For CentOS/Red Hat:

  ```
  yum install pcs pacemaker drbd84-utils kmod-drbd84
  ```

  OR

  ```
  yum install pcs pacemaker drbd90-utils kmod-drbd90
  ```

- A partition for DRBD on each server (with the same name on both servers) such as `/dev/sdd1`. Recommended partition size is 10-20 GB, otherwise DRBD sync will take a long time to complete.
- CLI command `hostname -i` must return the IP address of the management interface used for pacemaker sync correctly (update `/etc/hosts/` file with machine IP)

- Create the directory on each server under `/opt/ufm/files/` with read/write permissions on each server. This directory will be used by UFM to mount UFM files, and it will be synced by DRBD.

4.5.1.1.2 Installing UFM Containers

On the main server, install UFM Enterprise container with the command below:

```
docker run -it --name=ufm_installer --rm \
-v /var/run/docker.sock:/var/run/docker.sock \
-v /etc/systemd/system/:/etc/systemd_files/ \
-v /opt/ufm/files/:/installation/ufm_files/ \
-v /tmp/license_file/:/installation/ufm_licenses/ \
mellanox/ufm-enterprise:latest \
--install
```

On the standby (secondary) server, install the UFM Enterprise container like the following example with the command below:

```
docker run -it --name=ufm_installer --rm \
-v /var/run/docker.sock:/var/run/docker.sock \
-v /etc/systemd/system/:/etc/systemd_files/ \
-v /opt/ufm/files/:/installation/ufm_files/ \
mellanox/ufm-enterprise:latest \
--install
```

4.5.1.1.3 Downloading UFM HA Package

Download the UFM-HA package on both servers using the following command:

```
wget https://www.mellanox.com/downloads/UFM/ufm_ha_5.5.0-9.tgz
```

For Sha256:

```
wget https://download.nvidia.com/ufm/ufm_ha/ufm_ha_5.5.0-9.sha256
```

4.5.1.1.4 Installing UFM HA Package

For more information on the UFM-HA package and all installation and configuration options, please refer to UFM High Availability User Guide.

1. [On Both Servers] Extract the downloaded UFM-HA package under /tmp/
2. [On Both Servers] Go to the extracted directory /tmp/ufm_ha_XXX and run the installation script. For example, if your DRBD partition is `/dev/sda5` run the following command:

   ```
   ./install.sh -l /opt/ufm/files/ -d /dev/sda5 -p enterprise
   ```

4.5.1.1.5 Configuring UFM HA

There are the three methods to configure the HA cluster:

- Configure HA with SSH Trust (Dual Link Configuration) - Requires passwordless SSH connection between the servers.

- Configure HA without SSH Trust (Dual Link Configuration) - Does not require passwordless SSH connection between the servers, but asks you to run configuration commands on both servers.
- Configure HA without SSH Trust (Single Link Configuration) - Can be used in cases where only one link is available among the two UFM HA nodes/servers.

4.5.1.1.5.1 Configure HA with SSH Trust (Dual Link Configuration)

1. On the master server only, configure the HA nodes. To do so, from /tmp, run the `configure_ha_nodes.sh` command as shown in the below example

```
configure_ha_nodes.sh \
--cluster-password 12345678 \
--master-primary-ip 10.10.50.1 \
--standby-primary-ip 10.10.50.2 \
--master-secondary-ip 192.168.10.1 \
--standby-secondary-ip 192.168.10.2 \
--no-vip
```

> ⚠ The script `configure_ha_nodes.sh` is is located under /usr/local/bin/, therefore, by default, you do not need to use the full path to run it.

> ⚠ The `--cluster-password` must be at least 8 characters long.

> ⚠ When using back-to-back ports with local IP addresses for HA sync interfaces, ensure that you add your IP addresses and hostnames to the `/etc/hosts` file. This is needed to allow the HA configuration to resolve hostnames correctly based on the IP addresses you are using.

> ⚠ `configure_ha_nodes.sh` requires SSH connection to the standby server. If SSH trust is not configured, then you are prompted to enter the SSH password of the standby server during configuration runtime

2. Depending on the size of your partition, wait for the configuration process to complete and DRBD sync to finish. To check the DRBD sync status, run:

```
ufm_ha_cluster status
```

4.5.1.1.5.2 Configure HA without SSH Trust (Dual Link Configuration)

If you cannot establish an SSH trust between your HA servers, you can use ufm_ha_cluster directly to configure HA. You can see all the options for configuring HA in the Help menu:

```
ufm_ha_cluster config -h
```

To configure HA, follow the below instructions:

Please change the variables in the commands below based on your setup.

1. [On Standby Server] Run the following command to configure Standby Server:

```
ufm_ha_cluster config -r standby -e <peer ip address> -l <local ip address> -p <cluster_password>
```

2. [On Master Server] Run the following command to configure Master Server:

```
ufm_ha_cluster config -r master -e <peer ip address> -l <local ip address> -p <cluster_password> -i
<virtual ip address>
```

Configure HA without SSH Trust (Single Link Configuration)

This is not the recommended configuration and, in case of network failure, it might cause HA cluster split brain.

If you cannot establish an SSH trust between your HA servers, you can use ufm_ha_cluster directly to configure HA. To configure HA, follow the below instructions:

Please change the variables in the commands below based on your setup.

a. [On Standby Server] Run the following command to configure Standby Server:

```
ufm_ha_cluster config \
-r standby \
-e 10.212.145.5 \
-l 10.212.145.6 \
--enable-single-link
```

b. [On Master Server] Run the following command to configure Master Server:

```
ufm_ha_cluster config -r master \
-e 10.212.145.6 \
-l 10.212.145.5 \
-i 10.212.145.50 \
--enable-single-link
```

You must wait until after configuration for DRBD sync to finish, depending on the size of your partition. To check the DRBD sync status, run:

```
ufm_ha_cluster status
```

IPv6 Example:

```
ufm_ha_cluster config -r standby -l fcfc:fcfc:209:224:20c:29ff:fee7:d5f2 -e fcfc:fcfc:209:224:20c:2
9ff:fecb:4962 --enable-single-link -p some_secret
```

Starting HA Cluster

- To start UFM HA cluster:

```
ufm_ha_cluster start
```

- To check UFM HA cluster status:

```
ufm_ha_cluster status
```

- To stop UFM HA cluster:

```
ufm_ha_cluster stop
```

- To uninstall UFM HA, first stop the cluster and then run the uninstallation command as follows:

```
/opt/ufm/ufm_ha/uninstall_ha.sh
```

4.5.2 Confirming Topology Using UFM GUI

The integrity of the topology is essential to ensure the reliability and performance of the network. In this step we ensure the physically deployed topology matches the original design.

For the bring-up procedure, we will do it as a custom comparison. For maintenance, you can do it also as a periodic comparison. For further information, see UFM user manual - Periodic Topology Comparison.

4.5.2.1 Custom Comparison

Custom comparison compares user-defined topology with the current fabric topology. UFM compares the current fabric topology to a topology snapshot (of the same setup) and reports any differences between them.

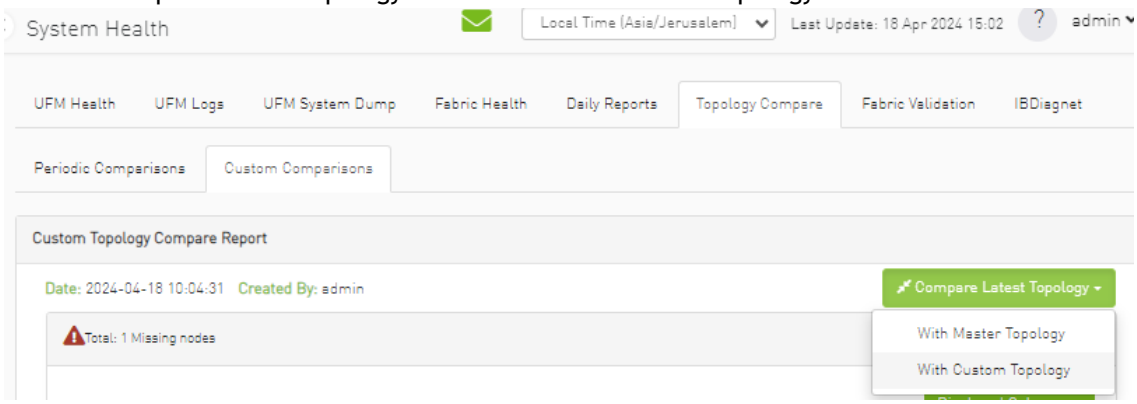
To be able to use the UFM topology comparison mechanism, first you need to create a TOPO file that defines the desired topology of the fabric. For further information refer to Creating a Topology File.

Once the TOPO file is created, you can use the topology comparison mechanism to compare the current fabric topology to the one in the TOPO file and view their differences (if found).

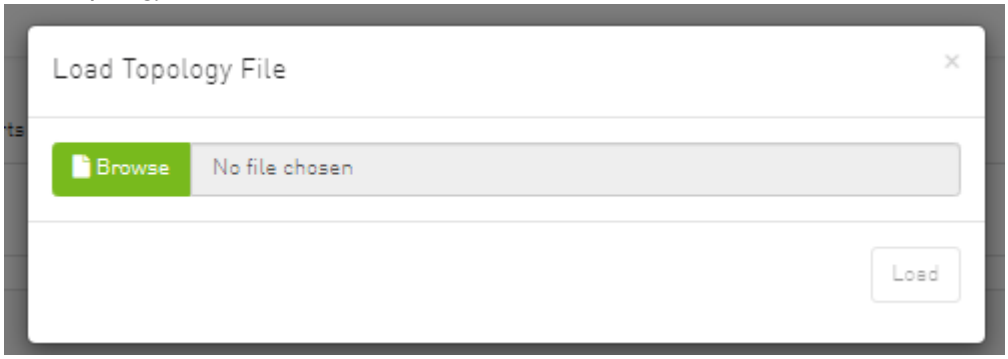
To perform topology comparison, do the following using Web UI:

- Access the System Health tab on the left menu
- Access the Topology Compare tab
- Access the Custom Comparisons tab

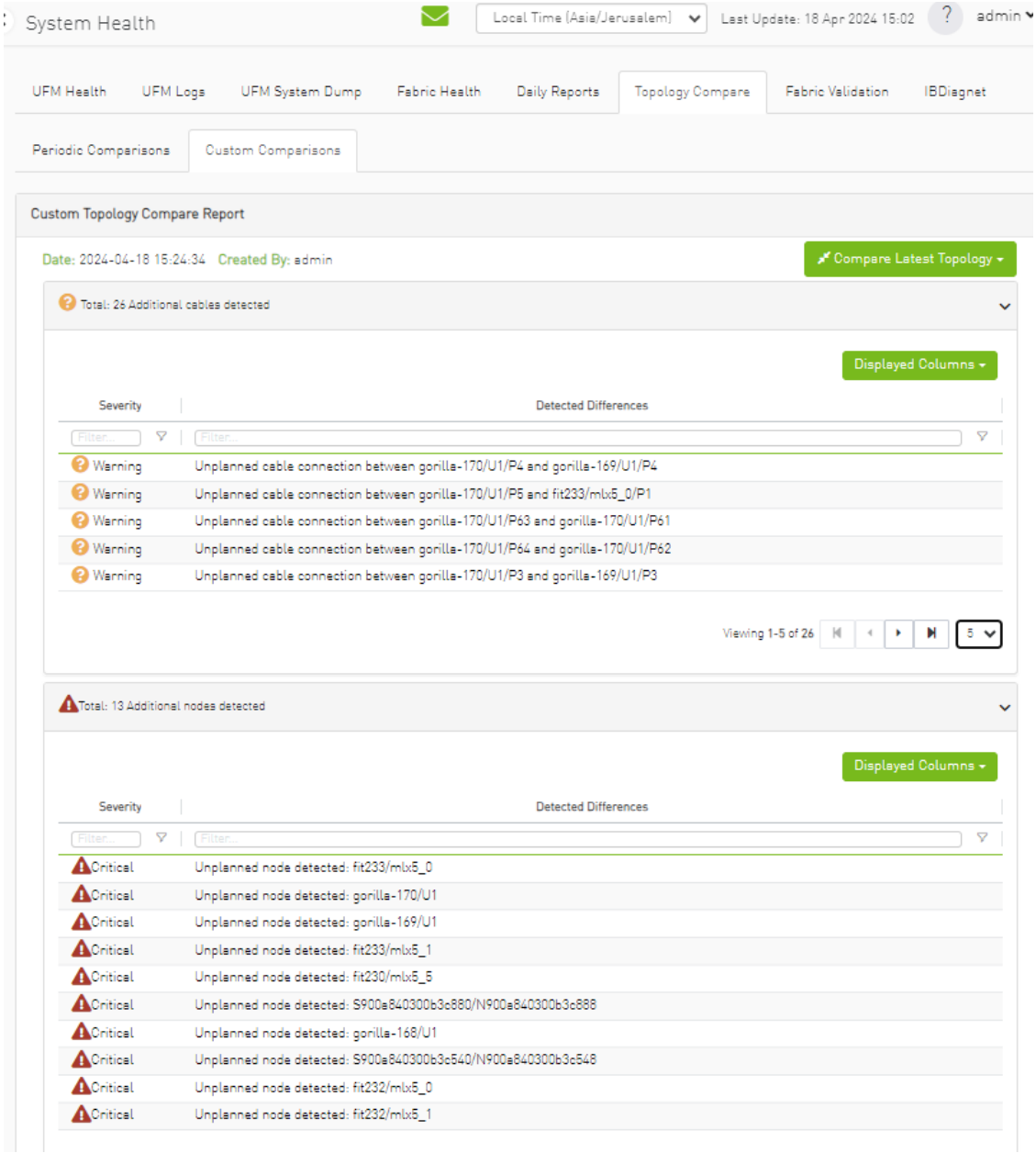
- Click on Compare Latest Topology and choose With Custom Topology



- Load topology file



- Review the report



For further information, see UFM user manual - Topology comparison.

For topology comparison using REST API, see UFM user manual - Topology comparison REST API.

5 Network Deployment

This chapter describes the required procedure to ensure that all network devices are running with the latest and greatest, most stable software packages and firmware versions.

5.1 Confirm Components' Firmware and Software Versions

This chapter will cover how to read firmware and software version for the following:

- Switch ASICs
- Transceivers
- HCA cards

The recommended guideline is to confirm that the versions among the cluster are aligned, or differ with up to 2 versions.

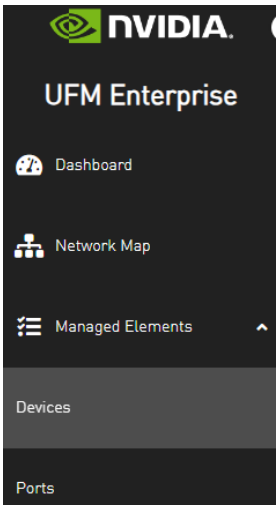
Information of the recommended NDR cluster bundle can be found here.

The process can be done using UFM GUI (which is recommended), or through MOFED commands.

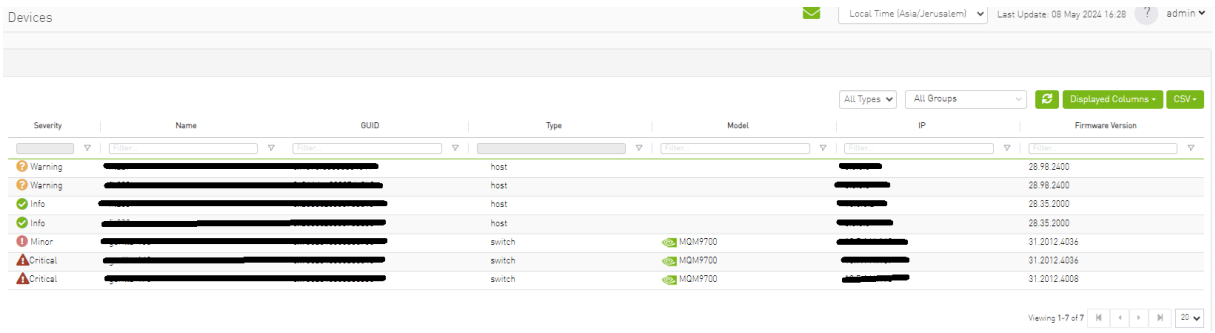
5.1.1 Verify versions using UFM GUI

5.1.1.1 ASICs and HCAs FW version

From the left side main menu, click on Managed Elements, and then on Devices.

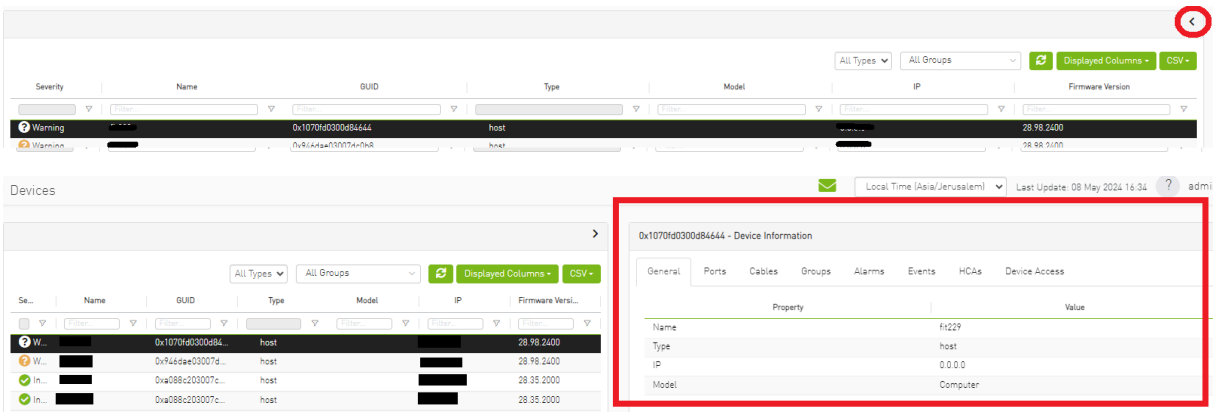


The Devices page opens and displays a table with all the managed switches/hosts in the cluster.



For switch ASIC, the FW version is listed in the main table.

For node HCA, select its row, Device Information section should pop up from the right side of the window, containing information about the selected device. If this section does not pop up, you should be able to open it by clicking on the left arrow on the top-right side of the table.



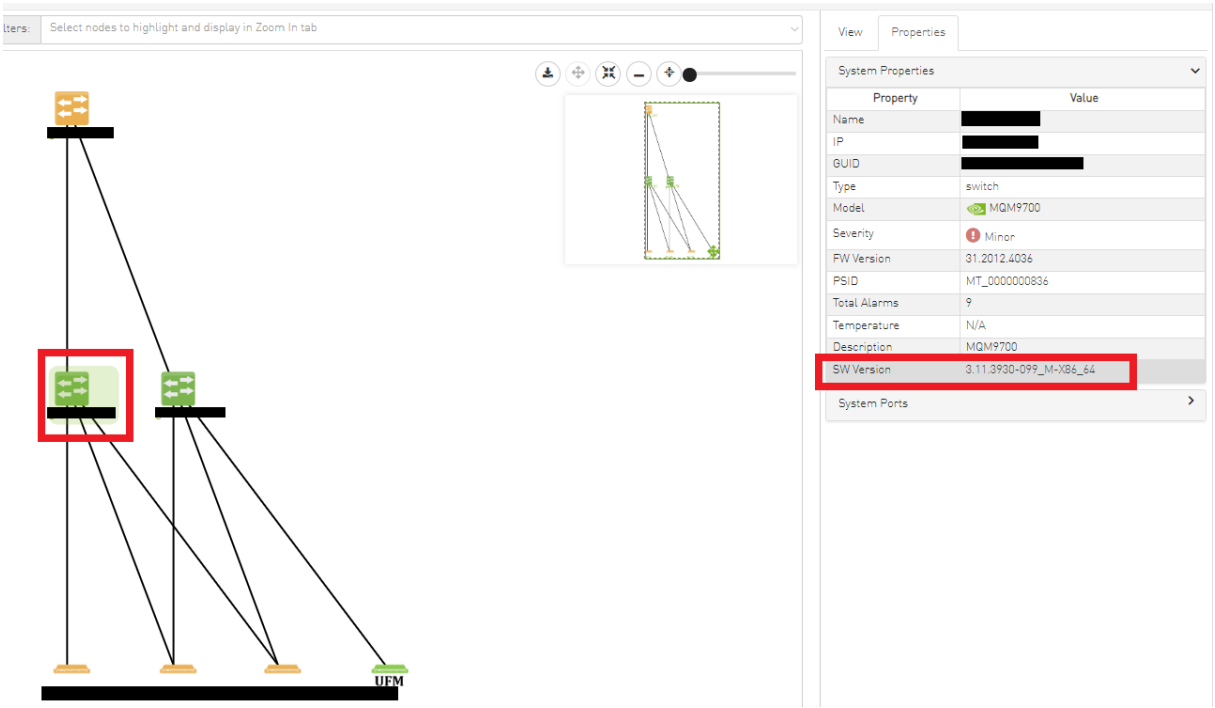
Click on the HCAs tab to see the device HCAs and the FW versions.

> ⚠ For HCAs only, click on HCAs from the left side main menu. All connected HCAs are listed there with the FW versions.

5.1.1.2 Managed switch SW (NOS) version

Click on Network Map from the left side main menu. The visualization of the cluster should display.

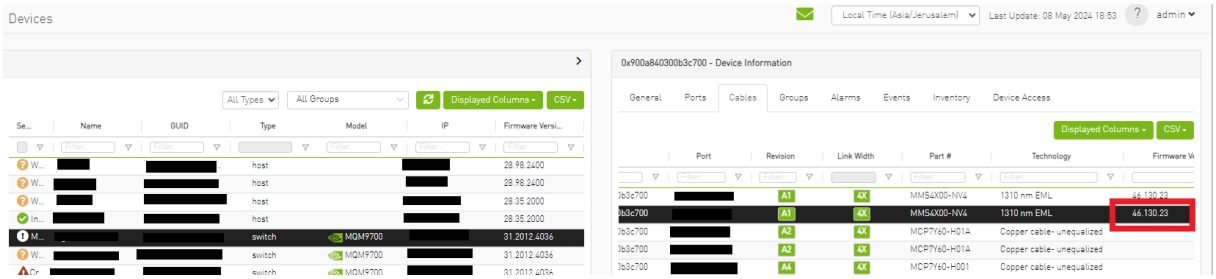
Select a switch. The switch information and the SW Version (NOS) should appear in the table on the left side.



5.1.1.3 Transceivers

From the Devices page, select a switch, and from the Device Information table on the right, click on Cables tab.

The page displays a table with the connected cables and the FW versions.



> Alternatively, go to Cables page from the left side main menu, which displays information on all the connected cables at once.

5.1.2 Optional Alternative - Verify Versions Using MOFED Tools

5.1.2.1 Prerequisite

- Make sure you have the latest MFT installed. If not, install it either as part of MLNX_OFED installation process or according to the instructions found here.

- Before using it, start the MST driver, run `mst start`
  This command will create files that represent NVIDIA devices in directory `/dev/mst`
  For the relevant devices, run "*mst status*"
  For further information, see the mst Service section in the MFT User Manual.

5.1.2.2 Identify the Switch Firmware Version

> ⚠ This section is applicable only to externally managed (unmanaged) switches (the ASIC firmware is bundled in NOS in managed systems).

1. Access the unmanaged switches via its LID.
2. Identify the switch LID, run `ibswitches.`

```
root@ufmx-qnt-02: #  ibswitches
Switch  0x900a8403006 f f780    ports   65  "MF0 ;grla -quanta -01:MQM9700/U 1"     enhanced    port    0
lid 1   lmc  0
Switch  0x900a8403006 f e0c0    ports   65  "MF0 ;grla -quanta -s2:MQM9700/U 1"     enhanced    port    0
lid 5   lmc  0
Switch  0x900a8403006 f f8c0    ports   65  "MF0 ;grla -quanta -s1:MQM9700/U 1"     enhanced    port    0
lid 14  lmc  0
Switch  0x900a8403006 f e040    ports   65  "MF0 ;grla -quanta -02:MQM9700/U 1"     enhanced    port    0
ltd 15  lmc  0
```

3. Check the firmware version, run `flint -d lid-X -qq q`.

```
root@ufmx -qnt-02: # flint -d lid-1 -qq q
Image type:          FS4
FW Version:          31.2012.3008
FW Release Date:     3.1.2024
Product Version:     31.2012.3008
Rom Info:            type=UEFI   version=skipped cpu=skipped
                     type=PXE    version=skipped devid=skipped
                     type=NVMe   version=skipped devid=skipped
Description:         UID     GuidsNumber
Base GUID:           900a8403006ff780    64
Base MAC:            900a846ff780    64
Image VSD:           N/A
Device VSD:          N/A
PSID:                MT 0000000577
Security Attributes:    secure-fw
```

5.1.2.3 Identify the Switch Version

1. Connect to your switch remotely with SSH: #ssh admin@my-switch-name(e.g. ssh admin@172.28.3.216)
2. Enter config mode.

```
switch> enable
switch# configure terminal
switch (config)#
```

3. Check the NOS' version.

```
switch (config)# show version
Product name:       MLNX-OS
Product release:    3.4.2002
Build ID:           #1-dev
Build date:         2015-07-30 20:13:19
Target arch:        x86_64
Target hw:          x86_64
Built by:           jenkins@fit74 _Version
summary:            X86_64 3.4.2002 2015-07-30 20:13:19 x86_64
```

5.1.2.4 Identify the HCA Firmware Version

1. Identify the HCA device, run `mst status`.

```
[root@fit229 ~]# mst status
MST modules:
------------
    MST PCI module is not loaded
    MST PCI configuration module loaded

MST devices:
------------
/dev/mst/mt4129_pciconf0         - PCI configuration cycles access.
                                   domain:bus:dev.fn=0000:04:00.0 addr.reg=88 data.reg=92
 cr_bar.gw_offset=-1
                                   Chip revision is: 00
```

2. Check the firmware version.

```
[root@fit229 ~]# flint -d /dev/mst/mt4129_pciconf0 -qq q
Image type:            FS4
FW Version:            28.98.2400
FW Release Date:       14.2.2022
Product Version:       28.98.2400
Rom Info:              type=UEFI version=14.25.21 cpu=AMD64,AARCH64
                       type=PXE version=3.6.502 cpu=AMD64
Description:           UID                GuidsNumber
Base GUID:             1070fd0300d84644        4
Base MAC:              1070fdd84644            4
Image VSD:             N/A
Device VSD:            N/A
PSID:                  MT_0000000798
Security Attributes:   N/A
```

3. For further details, see https://docs.nvidia.com/networking/display/mftv4270/Querying+the+Firmware+Image.

5.1.2.5 Identify the Transceiver Firmware Version

To check what is the transceiver firmware version, run `flint -d lid-1 --linkx --downstream_device_ids 1 q.`

```
[admin@gorilla-169 ~]# flint -d lid-1 --linkx --downstream_device_ids 1 q
Host : lid-1
 Device index 1
 Component Index 3
 Component Status NOT_PRESENT
 Component Update State IDLE
 Running state is :  Image A is running
Information block is :  FW image A is present
FW A Version : 46.130.0023
FW B Version : 00.00.0000
FW Factory Version : 00.00.0000
SupportedProtocol: CMIS 4.0 is implemented
Activation type: Self-activation with HW reset contained in the Run FW Image command. No additional actions
required from the host.
Serial number is 0
```

5.1.2.6 Identify the Driver Version

Make sure all the servers are using the latest driver version, run `- ofed_info -s.`

```
~ $ofed_info -s
MLNX_OFED_LINUX-23.04-0.5.3.3
```

5.2 On-site Upgrade – Low scale

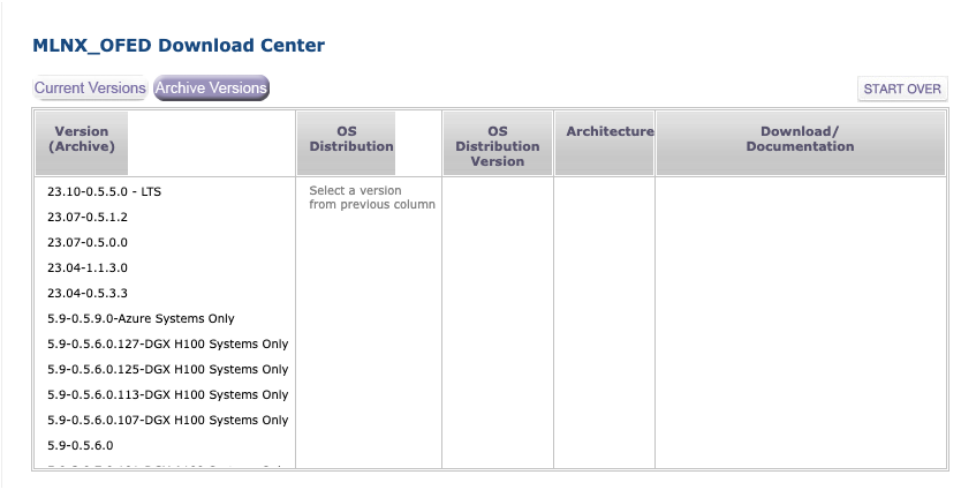
On-site service is an upgrade to the standard service level available on most systems.

To upgrade your cluster, complete the following:

- Install MLNX_OFED, see MLNX_OFED Installation
- Install managed switch OS, see Managed Switch Software Installation
- Install firmware (Host and Switch), see Firmware Installation

5.2.1 MLNX_OFED Installation

1. Download the desired MLNX_OFED Linux driver from here.
   Note: If another version is required, go to the Archive tab.



2. Log in to the installation machine as root.
3. Mount the ISO image on your machine.

```
host1# mount -o ro,loop MLNX_OFED_LINUX-<ver>-<OS label>-<CPU arch>.iso /mnt
```

4. Run the installation script.

```
/mnt/mlnxofedinstall
Logs dir: /tmp/MLNX_OFED_LINUX-x.x-x.logs
This program will install the MLNX_OFED_LINUX package on your machine.
Note that all other Mellanox, OEM, OFED, RDMA or Distribution IB packages will be removed.
Those packages are removed due to conflicts with MLNX_OFED_LINUX, do not reinstall them.
Starting MLNX_OFED_LINUX-x.x.x installation ...
........
........
Installation finished successfully.
Attempting to perform Firmware update...
Querying Mellanox devices firmware ...
```

For the installation instructions, refer to the User Manual.

5.2.2 Managed Switch Software Installation

5.2.2.1 MLNX-OS

MLNX-OS can be installed/upgraded using one of the methods below:

- 5.2.2.1.1 Via the Command Line Interface (CLI)
- 5.2.2.1.2 Via UFM

> Older versions of the software may require upgrading to one or more intermediate versions prior to upgrading to the latest. Missing an intermediate step may lead to errors.
>
> For further information, see MLNX-OS Release Notes.

5.2.2.1.1 Via the Command Line Interface (CLI)

1. Enter Config mode.

```
switch > enable
switch # configure terminal
switch (config) #
```

2. Display the currently available image (.img file).

```
switch (config) # show images
Installed images:

  Partition 1:
  <old_image>

  Partition 2:
  <old_image>

Last boot partition: 1
Next boot partition: 1

Images available to be installed:
 webimage.tbz
  <old_image>

Serve image files via HTTP/HTTPS: no

No image install currently in progress.

Boot manager password is set.

Image signing: trusted signature always required
Admin require signed images: yes

Settings for next boot only:
   Fallback reboot on configuration failure: yes (default)
```

3. Delete the image listed under “Images available to be installed” prior to fetching the new image. Use the command “image delete” for this purpose.

```
switch (config) # image delete <old_image>
```

> When deleting an image, it is recommended to delete the file, but not the partition, so as to not overload system resources.

4. Fetch the new software image.

```
switch (config) # image fetch scp://<username>:<password>@<ip-address>/var/www/html/<new_image>
Password (if required): ****** 100.0%[#####################################################################]
```

5. Display the available images again and verify that the new image now appears under “Images available to be installed”.

> ⚠ To recover from image corruption (e.g., due to power interruption), there are two installed images on the system. See the commands “image boot next” and “image boot location” for more information.

```
switch (config) # show images
Installed images:

  Partition 1:
  <old_image>

  Partition 2:
  <old_image>

Last boot partition: 1
Next boot partition: 1

Images available to be installed:
 webimage.tbz
  <new_image>

Serve image files via HTTP/HTTPS: no

No image install currently in progress.

Boot manager password is set.

Image signing: trusted signature always required
Admin require signed images: yes

Settings for next boot only:
   Fallback reboot on configuration failure: yes (default)
```

6. Install the new image.

```
switch (config) # image install <new_image>
Step 1 of 4: Verify Image
 100.0%  [#############################################################]
Step 2 of 4: Uncompress Image
 100.0%  [#############################################################]
Step 3 of 4: Create Filesystems
 100.0%  [#############################################################]
Step 4 of 4: Extract Image
 100.0%  [#############################################################]
```

> ⚠ CPU utilization may go up to 100% during image upgrade.

7. Have the new image activate during the next boot.

```
switch (config) # image boot next
```

8. Run “show images” to review your images.

```
switch (config) # show images
Installed images:

  Partition 1:
  <new_image>

  Partition 2:
  <old_image>

Last boot partition: 1
Next boot partition: 1

Images available to be installed:
 webimage.tbz
```

```
  <new_image>

Serve image files via HTTP/HTTPS: no

No image install currently in progress.

Boot manager password is set.

Image signing: trusted signature always required
Admin require signed images: yes

Settings for next boot only:
   Fallback reboot on configuration failure: yes (default)
```

9. Save current configuration.

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

10. Reboot to run the new image.

```
switch (config) # reload
Configuration has been modified; save first? [yes] yes
Configuration changes saved.
Rebooting...
switch (config)#
```

> ⚠ After software reboot, the software upgrade will also automatically upgrade the firmware version.

> ⚠ On systems with dual management, the software must be upgraded on both the host and the device modules.

For Further information, see MLNX-OS User Manual.

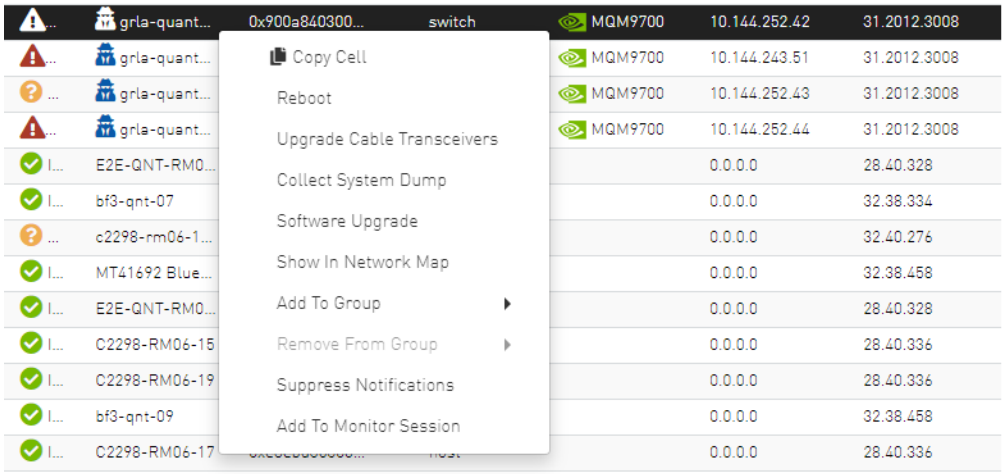
5.2.2.1.2 Via UFM

> ⚠ Upgrading MLNX-OS via UFM requires having MFT installed on the UFM server.

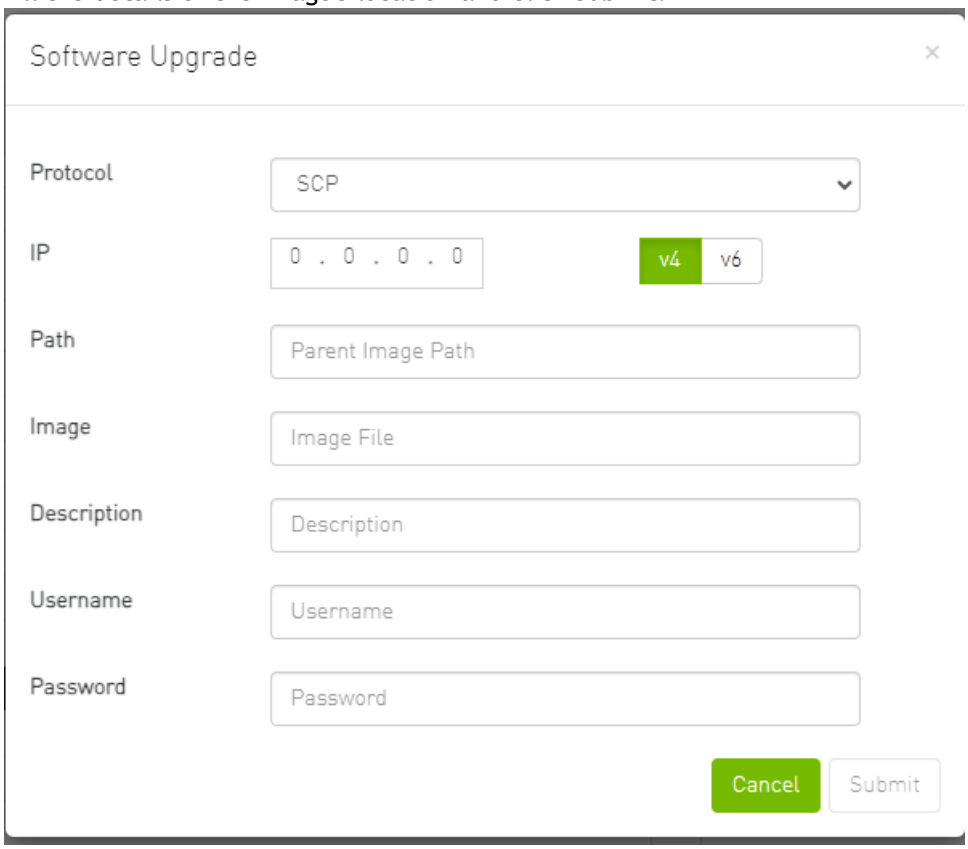
To upgrade MLNX-OS via UFM, follow the below steps:

1. Log into the UFM WEB UI.
2. Expand the "Managed Elements" and click on Devices.
3. Identify the switch.
4. Right-click on the chosen switch.

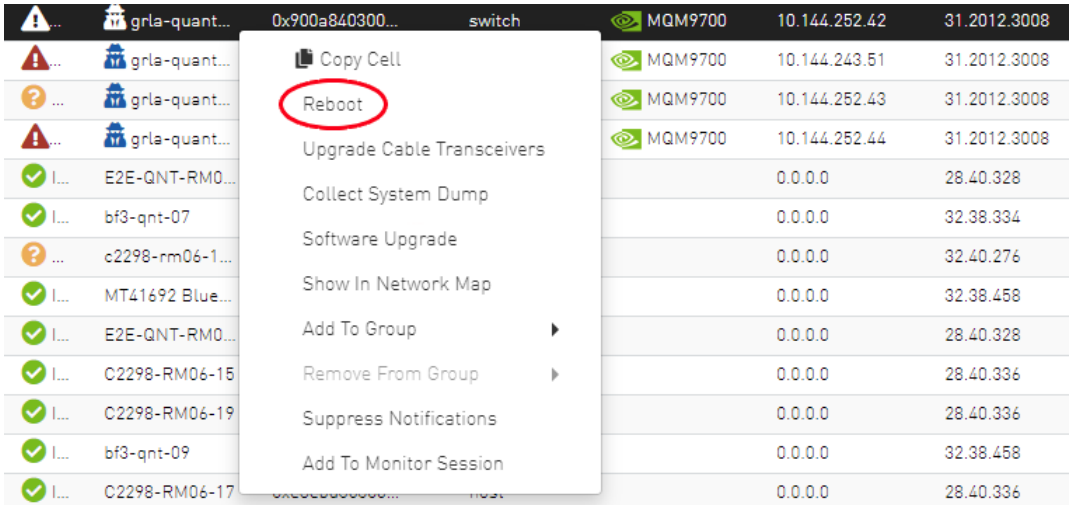
5. Click on "Software Upgrade".



6. Fill the details of the image's location and click Submit.



7. Reboot the switch.



5.2.3 Firmware Installation

5.2.3.1 HCA Firmware Installation

1. Download the desired firmware version from here. Choose the right OPN and PSID.
2. Install the firmware using `flint -d device_id -i firmware.bin burn`.
3. Reboot the device once the firmware burning is completed.

   For further information, see MFT User Manual.

5.2.3.2 Unmanaged Switch Firmware Installation

1. Download the desired firmware version from here.
2. Install switch firmware:
   a. Access the unmanaged switches via its LID.
   b. Identify the switch LID, run `ibswitches.`

```
root@ufmx-qnt-02: #  ibswitches
Switch  0x900a8403006 f f780    ports   65  "MF0 ;grla -quanta -01:MQM9700/U 1"     enhanced
port    0    lid 1    lmc  0
Switch  0x900a8403006 f e0c0    ports   65  "MF0 ;grla -quanta -s2:MQM9700/U 1"     enhanced
port    0    lid 5    lmc  0
Switch  0x900a8403006 f f8c0    ports   65  "MF0 ;grla -quanta -s1:MQM9700/U 1"     enhanced
port    0    lid 14   lmc  0
Switch  0x900a8403006 f e040    ports   65  "MF0 ;grla -quanta -02:MQM9700/U 1"     enhanced
port    0    ltd 15   lmc  0
```

   c. Install the firmware.

```
flint -d lid-xxx -i fw-Quantum-2-rel-31_2012_3008-MQM9700-NS2X_Ax.bin burn
```

   d. Reboot the device once the firmware burning is completed.

```
flint -d lid-xxx swreset
```

   For further information, see MFT User Manual.

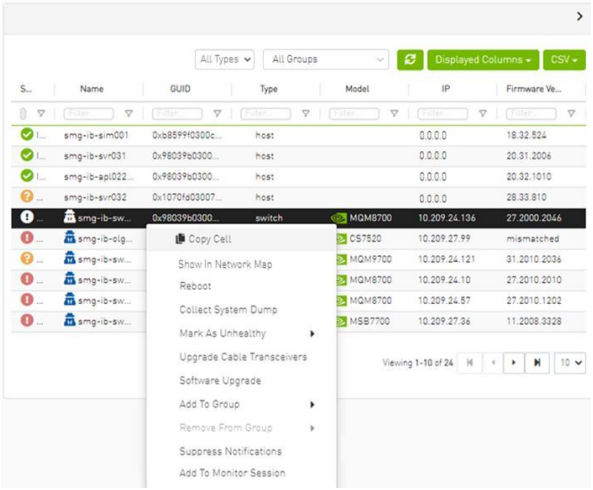
5.2.3.3 Transceivers Firmware Installation

For manage switch you must upgrade the transciever FW using UFM.

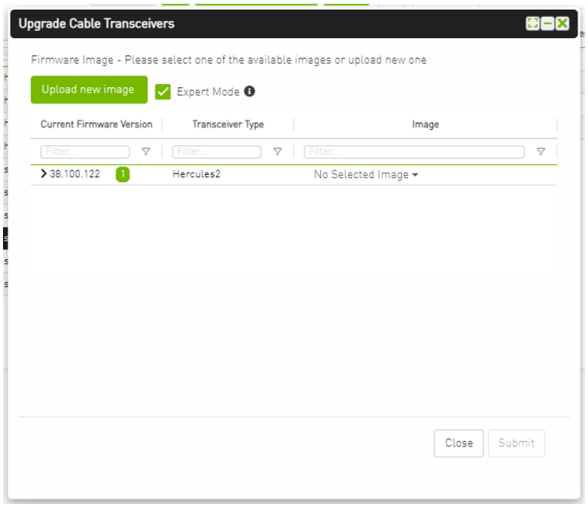
For unmanaged switch and servers you can upgrade using MFT.

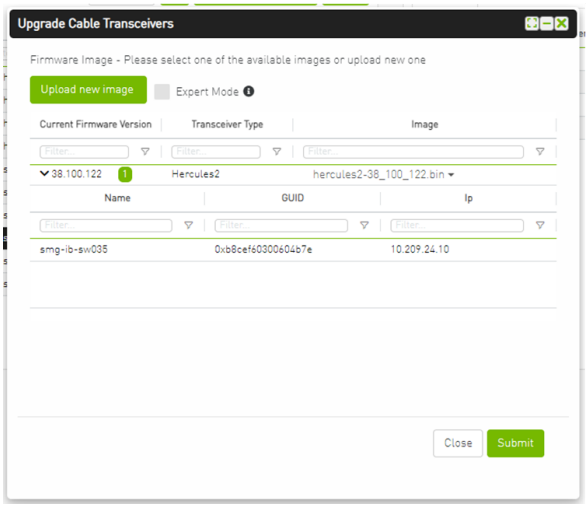
5.2.3.3.1 Via UFM:

1. Navigate to managed elements page.
2. select the target switches and click on Upgrade Cable Transceivers option.

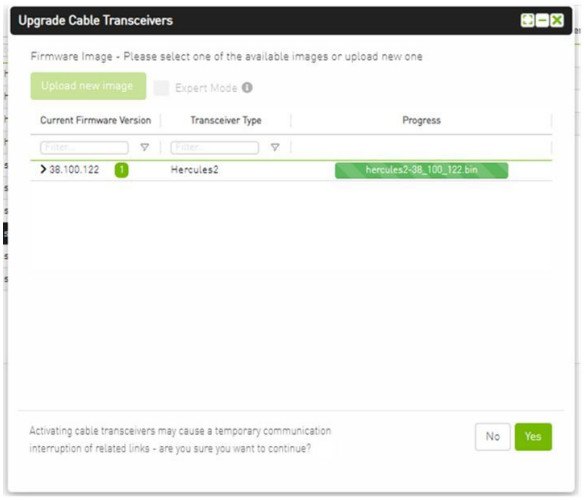


3. A model will be shown containing list of the active firmware versions for the cables of the selected switches, besides the version number, a badge will show the number of matched switches:

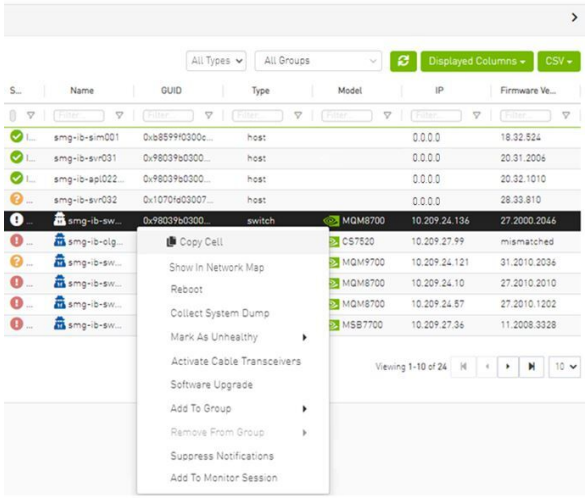




4. After the user clicks Submit, the GUI will start sending the selected binaries with the relevant switches sequentially, and a model with a progress bar will be shown (this model can be minimized):



5. After the whole action is completed successfully, you will be able to see the following message at the model bottom The upgrade cable transceivers completed successfully, do you want to activate it? by clicking the yes button it will run a new action on all the burned devices to activate the new uploaded binary image.
6. Another option to activate burned cables transceivers you can go to the Groups page and right click on the predefined Group named Devices Pending FW Transceivers Reset or you can right click on the upgraded device from managed element page and select Activate cable Transceivers action.



For further information, see UFM User Manual.

5.2.3.3.2 Via MFT:

1. Query the Transceiver firmware information.

```
flint -d lid-1 --linkx --downstream_device_ids 1 q
```

> When the cluster has many switches, multiple hosts may be engaged in the upgrade process. Each device (NIC, Switch) can update only the modules connected directly to it, not the far end. Updating the far-end transceiver requires the same operation to be done at the far-end switch(es).

2. Burn the cable using the Auto-update command.

```
flint -d <device> --linkx--linkx_auto_update--download_transfer-i<binary file> b
```

3. Activate the new firmware:

```
flint -d lid-2 --linkx--linkx_auto_update--activate b
```

More reading of flint for cables can be found in Cable Burn Command.

6 Configuration and Basic Features Activation

This chapter provides a comprehensive outline of the necessary steps and procedures for configuring a cluster. Each section will delve into detailed explanations and instructions for carrying out the configuration tasks effectively.

This section covers:

- Optional Network Configuration
- Switch Configuration
- UFM Configuration

6.1 Network Configuration (Optional)

This chapter covers the following additional optional configurations:

- Assign IP addresses to NICs (IPoIB):
  The support of communication using standard IP addresses through the IB cluster comes with several advantages and can be very common and useful.
  To support IPoIB, should configure IP addresses (via DHCP / static IPs) to the desired NICs. For more information, see: IP over InfiniBand (IPoIB).
- Configure the appropriate routing algorithm and enable the relevant subnet manager functions
  The following routing features are covered in this section
  - Adaptive routing
  - SHIELD (PFRN)
  - HBF (Hash-Based Forwarding)

6.1.1 Adaptive Routing

Adaptive Routing (AR) enables the switch to select the output port based on the port's load. It assumes there are no constraints on the output port selection (free adaptive routing). The subnet manager (SM) enables and configures the Adaptive Routing mechanism on the fabric switches. It scans all the fabric switches and identifies which ones support Adaptive Routing, then it configures the AR functionality on these switches. The subnet manager (SM) configures the AR groups and AR LFTs tables to allow switches to select an output port out of an AR group for a specific destination LID. The configuration of the AR groups relies on the selection of one of the following supported algorithm:

- LAG: All ports that are linked to the same remote switch are in the same AR group. This algorithm is suitable for any topology with multiple links between switches, especially Hypercube/3D torus/mesh, where there are several links in each direction of the X/Y/Z axis
- TREE: All ports with minimal hops to destination are in the same AR group. This algorithm is suitable for tree topologies such as fat-tree, quasi-fat-tree, parallel links fat-tree, etc
- DF_PLUS: This algorithm is designed for the Dragonfly plus topology

AR is enabled by default in the opensm.conf file.
routing_engine subnet manager configuration option needs to be adjusted based on the network's

topology. The default value is: `ar_updn` which is relevant for a fat tree topology.

To disable AR in the fabric, need to change the routing_engine subnet manager configuration option to a non-AR routing engine.
Fat tree topology example: routing_engine updn.

After changes in the opensm.conf file should run the "pkill -HUP opensm" bash command.

> ⚠ It is recommended to specify the correct root GUIDs file in the opensm.conf file.
>
> root GUIDs file contains all the root GUIDs, each root GUID in a new line.
>
> By default, the root GUIDs file is taken from the SM as you can see in the opensm.conf file
>
> root_guid_file /opt/ufm/files/conf/opensm/root_guid.conf

For further details and configuration options, see: https://enterprise-support.nvidia.com/s/article/Recommended-Topologies-for-Implementing-an-HPC-Cluster-with-NVIDIA-Quantum-InfiniBand-Solutions-Part-2

6.1.2 SHIELD

Self-Healing Interconnect Enhancement for Intelligent Datacenters, which referred to as Fast Link Fault Recovery (FLFR) throughout this document, enables the switch to select the alternative output port if the output port provided in the Linear Forwarding Table is not in Armed/Active state. This mode allows the fastest traffic recovery in case of switch-to-switch port failures due to link flaps, or neighbor switch reboots without intervention of Subnet Manager. The Fast Link Fault Notification (FLFN) enables the switch to report to neighbor switches that an alternative output port for the traffic to specific destination LID should be selected to avoid sending traffic to the switch. This is required when the FLFR on the switch has no alternative port to select for the destination LID. Adaptive Routing Notification (ARN) enables the switch to send a report to the neighbor switches, if its ports are congested above the threshold. This report causes neighbor switches to select another output port to deliver the traffic to the destination LID.

> ⚠ Fast Link Fault Notification (FLFN) is supported for fat-tree and quasi-fat-tree topologies only.

SHIELD is enabled by default in the `opensm.conf` file with `shield_mode` subnet manager configuration option.

To disable SHIELD in the fabric, you need to change the `shield_mode` subnet manager configuration option to 0.

For further details, see: https://enterprise-support.nvidia.com/s/article/Recommended-Topologies-for-Implementing-an-HPC-Cluster-with-NVIDIA-Quantum-InfiniBand-Solutions-Part-2

6.1.3 HBF

Hash-Based Forwarding (HBF) is an InfiniBand switch feature that enables selection of the switch outgoing port for statically routed packets based on the packet's parameters (ECMP like) as opposed to selection of the switch outgoing port based on Linear Forwarding Tables.

HBF configuration is done in the opensm.conf file (path: `/opt/ufm/files/conf/opensm/opensm.conf`).

For description of the possible HBF options, see https://enterprise-support.nvidia.com/s/article/Recommended-Topologies-for-Implementing-an-HPC-Cluster-with-NVIDIA-Quantum-InfiniBand-Solutions-Part-2.

6.2 Switch Configuration

> ⚠ This chapter is relevant for managed switch systems only.

6.2.1 Initial Management Configuration

The procedures described in this page assume that you have already installed and powered on your switch according to the instructions in the Hardware Installation Guide, which was shipped with the product.

6.2.1.1 Configuring the Switch for the First Time

To initialize the switch do the following:

1. Connect the host PC to the console (RJ-45) port of the switch system using the supplied cable.
2. Configure a serial terminal with the settings described below.

   > ⚠ This step may be skipped if the DHCP option is used and an IP is already configured for the MGT port.

Parameter	Setting
Baud Rate	115200
Data bits	8
Stop bits	1
Parity	None
Flow Control	None

3. Select the boot partition in the menu prompted.

   > ⚠ Select “0” to boot with software version installed on partition #1.
   > Select “1” to boot with software version installed on partition #2.

4. Login as admin and use admin as password. If the machine is still initializing, you might not be able to access the CLI until initialization completes. As an indication that initialization is ongoing, a countdown of the number of remaining modules to be configured is displayed in the following format: “<no. of modules> Modules are being configured”.
5. Go through the Switch Management configuration wizard.

Wizard Session Display (Example)	Comments
`Do you want to use the wizard for initial configuration? yes`	You must perform this configuration the first time you operate the switch or after resetting the switch to the factory defaults. Type “yes” and then press <Enter>.
`Step 1: Hostname? [switch-1]`	If you wish to accept the default hostname, then press <Enter>. Otherwise, type a different hostname and press <Enter>.
`Step 2: Use DHCP on mgmt0 interface? [yes]`	Perform this step to obtain an IP address for the switch. (mgmt0 is the management port of the switch.) - If you wish the DHCP server to assign the IP address, type “yes” and press <Enter>. If you type “no” (no DHCP), then you will be asked whether you wish to use the “zeroconf” configuration or not. If you enter “yes” (yes Zeroconf), the session will continue as shown in the "IP zeroconf configuration" table. If you enter “no” (no Zeroconf), then you need to enter a static IP, and the session will continue as shown in the "Static IP configuration" table.
`Step 3: Enable IPv6 [yes]`	Perform this step to enable IPv6 on management ports. The default is "yes" (enabled). If you enter “no” (no IPv6), then you will automatically be referred to Step 5.
`Step 4: Enable IPv6 autoconfig (SLAAC) on mgmt0 interface? [no]`	Perform this step to enable stateless address autoconfig on external management port. The default is "no" (disabled). If you wish to enable it, type “yes” and press <Enter>.
`Step 5: Use DHCPv6 on mgmt0 interface? [yes]`	Perform this step to enable DHCPv6 on the MGMT0 interface.
`Step 6: Update time?`	Perform this step to change the time configured. Press <enter> to leave the current time.
`Step 7: Enable password hardening? [yes]`	Perform this step to enable/disable password hardening on your machine. If enabled, new passwords will be checked upon configured restrictions. The default is "yes" (enabled). If you wish to disable it, enter “no”.
`Step 8: Admin password (Must be typed)? <new_password>`	To avoid illegal access to the machine, please type a password and then press <Enter>. Due to Senate Bill No. 327, this stage is required and cannot be skipped.
`Step 9: Confirm admin password? <new_password>`	Confirm the password by re-entering it. Note that password characters are not printed.

Recommended configuration:

```
Configuration wizard

Do you want to use the wizard for initial configuration? y

Step 1: Hostname? [grla-quanta-01]
Step 2: Use DHCP on mgmt0 interface? [yes]
```

```
Step 3: Enable IPv6? [yes]
Step 4: Enable IPv6 autoconfig (SLAAC) on mgmt0 interface? [yes]
Step 5: Enable DHCPv6 on mgmt0 interface? [yes]
Step 6: Update time? [2024/01/23 16:57:14]
Step 7: Enable password hardening: [no]
Step 8: Admin password (Enter to leave unchanged)?
Step 9: Monitor password (Enter to leave unchanged)?

You have entered the following information:

Hostname: grla-quanta-01
Use DHCP on mgmt0 interface: yes
Enable IPv6: yes
Enable IPv6 autoconfig (SLAAC) on mgmt0 interface: yes
Enable DHCPv6 on mgmt0 interface? yes
Update time: 2024/01/23 16:57:14
Enable password hardening: no
Admin password (Enter to leave unchanged): (unchanged)
Monitor password (Enter to leave unchanged): (unchanged)

To change an answer, enter the step number to return to.
Otherwise hit <enter> to save changes and exit.

Choice:

Configuration changes saved.
```

For further information on MLNX-OS, see MLNX-OS Getting Started.

6.2.2 Zero Touch Provisioning

Zero-Touch Provisioning (ZTP) automates the initial configuration of switch systems at boot time. It helps minimize manual operation and reduce customer initial deployment costs. ZTP allows for automatic upgrade of the switch with a specified OS image, setting up the initial configuration database, and to load and run a container from an image file.

ZTP is based on DHCP and is enabled by default. For ZTP to work, the software enables DHCP by default on all its management interfaces.
The initial configuration is applied using a regular text file. The user can create such a configuration file by editing the output of a “show running-config” command.

> ⚠ Only a textual configuration file is supported.

Flow

1. Switch OS sends request to the DHCP server, to get image to upgrade to, initial configuration to apply, and docker image to run a container.
    - For IPv4, the switch OS sends 'tftp-server-name' (option 66) and 'bootfile-name' (option 67) requests from the DHCPv4 server
    - For IPv6, the switch OS sends 'bootfile-url' (option 58) from the DHCPv6 server
2. DHCP server responds with URLs (e.g. scp url) to the required items (OS image, configuration file, docker image).
3. Switch OS downloads and uses the components using the URLs for the required actions (upgrade/apply configuration/run docker).

Configuration

To make ZTP work over the network, the DHCP server must be configured to respond with the desired URLs for the requests mentioned above.

For DHCPv4:

- Option 66 contains the URL prefix to the location of the files
- Option 67 contains the name of files
- Option 58 contains the complete URLs of files

DHCPv4:

Set options 66 and 67 in the DHCP server configuration file.

- option 66 URLs should be the prefix to the desired file location (without the filename itself)
- in option 67, specify the desired filenames accordingly

```
option tftp-server-name "<image server url>, <config server url>, <docker container server url>";
option bootfile-name "<image file>, <config file>, <docker container file>";
```

DHCPv6:

Set option 58 in the DHCP server configuration file.
Specify full URLs to the desired files

```
option dhcp6.bootfile-url "<image server url/image file>, <config server url/config file>, <docker container server
url/docher container file>"
```

The item value can be empty, but the comma shall not be omitted.

For further information on MLNX-OS, see Configuring the Switch with ZTP.

6.2.3 Configuring a Split Port

Split cables are cables that can connect one OSFP switch port to two separate devices, such as two HCAs (Host Channel Adapters). Split cables increase port's density and reducing the cost per port of the switch. However, using split cables requires explicit configuration of the switch, as the switch does not auto-sense the cable type and the number of lanes per port.

To configure the switch to use split cables:

1. Change the system's profile to "ib split-ready".

   ```
   system profile ib split-ready
   ```

2. Shut down the interface.

   ```
   interface ib 1/4 shutdown
   ```

3. Split the ports as desired.

   ```
   interface ib 1/1/1 port-type qsfp-split-2
   ```

For further information, see https://docs.nvidia.com/networking/display/mlnxosv3113002/InfiniBand+Interface

6.3 UFM Configuration

6.3.1 Configure a Static IPoIB on the IB Port:

1. From master, stop UFM.

```
ufm_ha_cluster stop
```

2. On both servers, set a static IP address for the main IB interface (using the proper tool, e.g. netplan for Ubuntu).
3. From master, start UFM.

```
ufm_ha_cluster start
```

4. Verify the configured IP address appears for the main IB interface using 'ifconfig'.

6.3.2 Recommended QoS Configuration:

Edit the following parameters in the OpenSM configuration file:

```
vi /opt/ufm/files/conf/opensm/opensm.conf
```

```
# Limit the maximal operational VLs
max_op_vls 2
# Enable QoS setup:
qos TRUE
# QoS policy file to be used:
qos_policy_file (null)
# Suppress QoS MAD status errors:
suppress_sl2vl_mad_status_errors FALSE
# Override multicast SL provided in join/create request:
override_create_mcg_sl 0xff
# QoS default options
qos_max_vls 0
qos_high_limit -1
qos_vlarb_high (null)
qos_vlarb_low (null)
qos_sl2vl 0,0,1,1,1,1,1,1,1,1,1,1,1,1
```

After the change of the OpenSM configuration file stop and restart the ufm_ha_cluster

```
ufm_ha_cluster stop
ufm_ha_cluster start
```

For additional optional configuration, see https://docs.nvidia.com/networking/display/ufmenterpriseumv6160/additional+configuration+(optional)

To upgrade UFM software, see https://docs.nvidia.com/networking/display/ufmenterpriseqsgv6160/upgrading+ufm+software

7 Cluster Verification

This chapter describes the required procedure to be executed toward the end of cluster bringup phase, just before the cluster operation. That includes files and logs to be reviewed and kept as reference when the cluster is signed off from the build phase to the operation phase and after performing UFM/OpenSM/Firmware upgrade procedure.

This chapter outlines the necessary steps that need to be taken as part of the final stages of InfiniBand cluster bring up and initialization, just before the cluster becomes operational.

Please adhere to the following steps:

1. Monitor SM Logs by:
   Looking for errors.
   Verifying the subnet is up.
2. Run UFM Fabric Health.
   Check the output summary.
   Check the port counters.
   Check the nodes information.
   Check the errors on the links.
3. UFM Telemetry - collects unique counters for each port in the InfiniBand fabric to ensure efficiency.
4. UFM Events and Alarms - allows to identify any problems including ports and device connectivity.

7.1 SM Logs

SM logs include details of reported errors, all errors reported in opensm.log should be treated as indicators of IB fabric health.

SM logs path:

- When only OpenSM is running without UFM: /var/log/opensm.log
- When OpenSM is running with UFM on a Docker, enter the container:

  ```
  docker exec -it ufm bash
  ```

  the path is: `/opt/ufm/files/log/opensm.log`

The SM log file should include the message "SUBNET UP" if OpenSM was able to set up the subnet correctly.

7.1.1 Logs Parameters

The SM log file size can be changed. You can choose how often a new SM log file will be created: daily, weekly (default), monthly.

The SM log file will reach its maximum log size, or it will obey the rotational periodically order.

1. Modify the OpenSM log maximum file size:

```
vi /opt/ufm/files/conf/opensm/opensm.conf
log_max_size
```

2. Modify the OpenSM log frequency rotation:

```
vi /etc/logrotate.d/opensm
```

7.1.2 Useful Commands

Locate the subnet manager:

```
[root@fit229 ~]# sminfo
sminfo: sm lid 8 sm guid 0xa088c203007cdd36, activity count 47086 priority 15 state 3 SMINFO_MASTER
```

Query node description:

```
[root@fit229 ~]# smpquery nd 8
Node Description:.................fit232 mlx5_0
```

7.1.3 Common Errors

Error	Description
TIMEOUT	Timeout in the network, look for a bad cable
trap128	The link state is changed. If this occurs too often on the same cable, make sure the cable is not corrupted
trap131	A bad cable connected
trap 144	Change in either link width/speed or node description
traps 257-259	Bad partitions

Example (Error trap 128):

Check the error by running the next command, if a port LinkDownedCounter is too big, it means the cable is corrupted.
for i in {1..<ports amount>};do echo Port:$i;perfquery <LID>$i | grep LinkDownedCounter;done

Apr 16 22:11:41 477567 [DA9C8640] 0x02 -> log_notice: Reporting Generic Notice type:1 num:128 (Link state change) from LID:4 GID:fe80::900a:8403:b3:c540

```
[root@l-qa-203 ~]# for i in {1..64};do echo Port:$i;perfquery 4 $i | grep LinkDownedCounter;done
Port:1
LinkDownedCounter:.............2
Port:2
LinkDownedCounter:.............0
Port:3
LinkDownedCounter:.............154222
Port:4
..
```

7.2 UFM Fabric Health

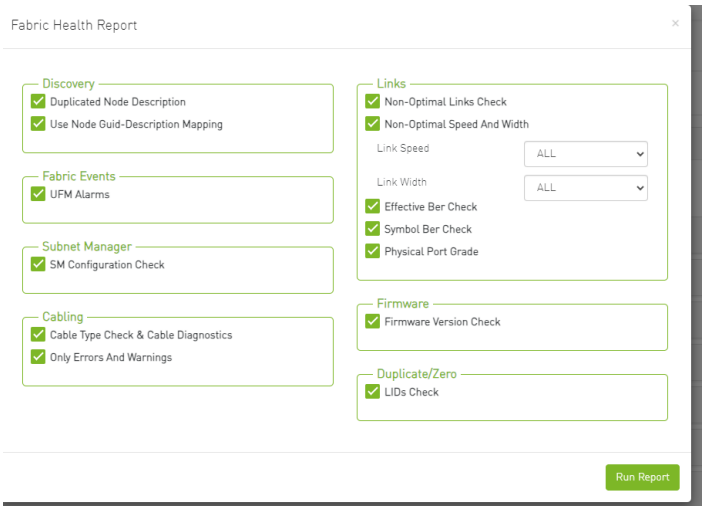
UFM fabric health report contains the results of a series of checks that run on the fabric.

The report displays, the following:

- A report summary table of the errors and warnings generated by the report
- A fabric summary of the devices and ports in the fabric
- Details of the results of each check run by the report

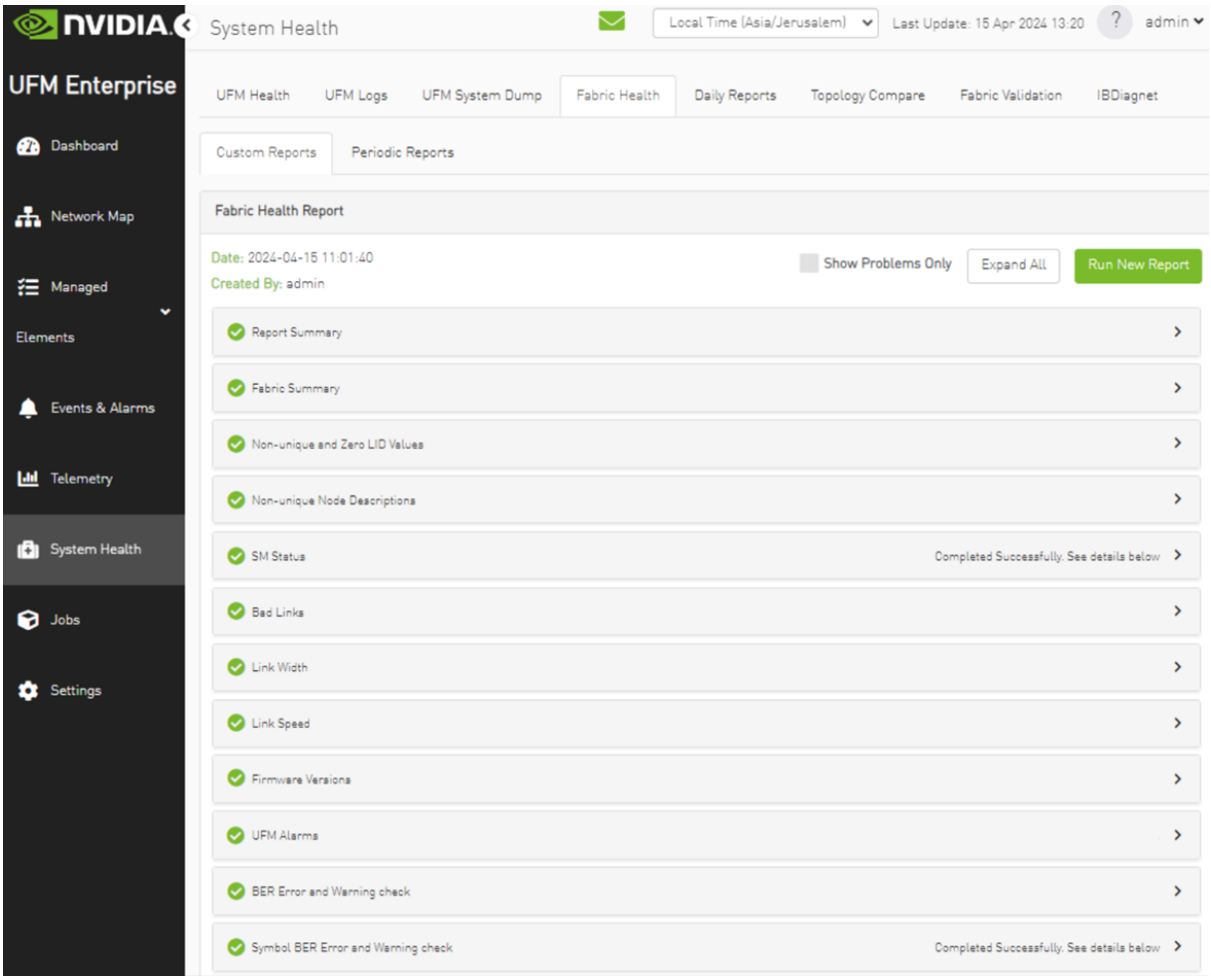
To generate fabric health report and verifying all sections are green, perform the following steps using Web UI:

- Access the "System Health" tab on the left menu
    - Under "Fabric Health"
        - Click on "Run New Report" under the "Fabric Health" section
        - check all checkboxes

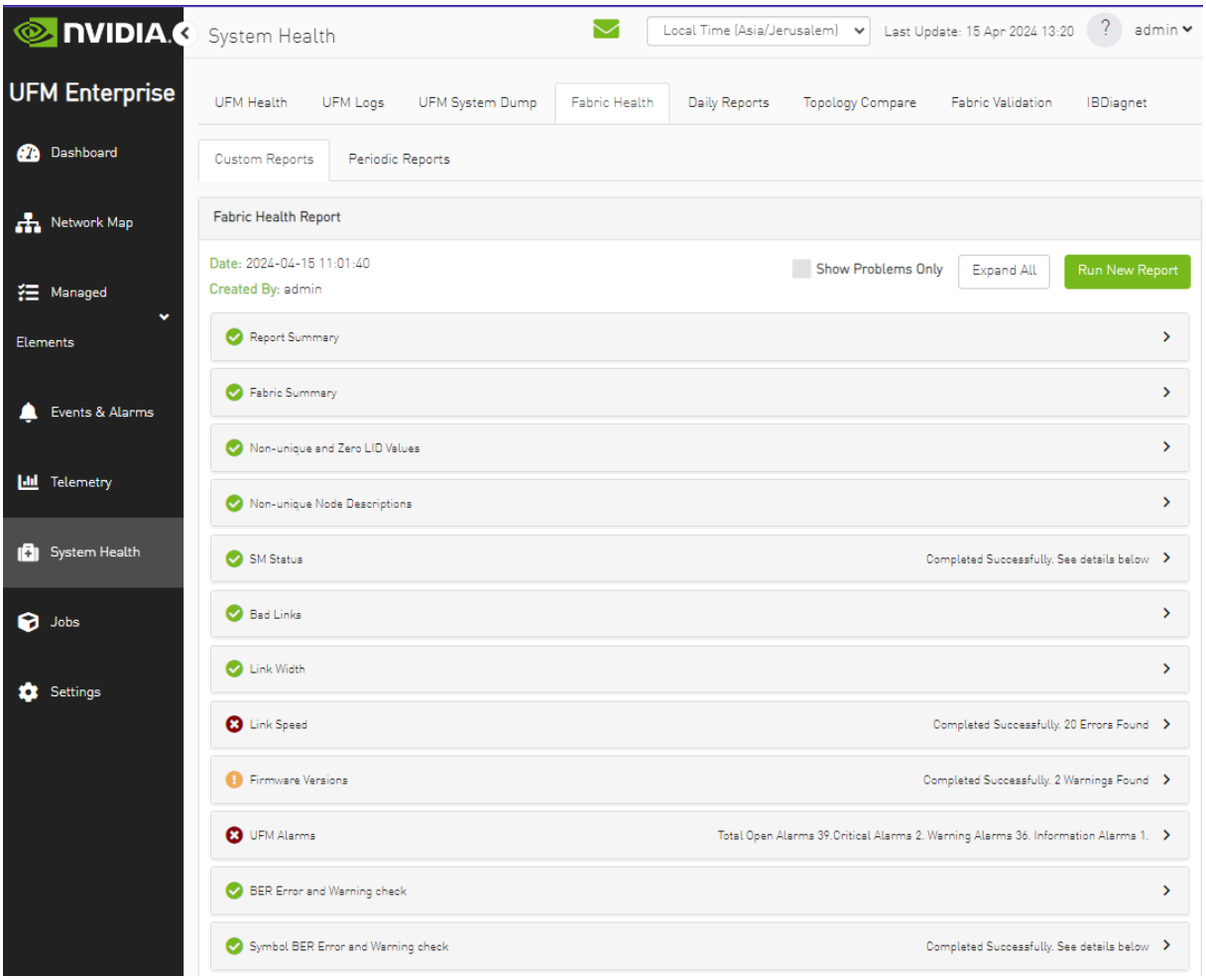


        - Confirm that all fields are indicating green status
        - For detailed instructions, refer Fabric Health Tab
    - Under "Fabric Validation"
        - Run the available tests
        - Verify the outcomes as either "Pass" or "Completed with No Errors"
        - For detailed instructions, see Fabric Validation Tab
    - Furthermore, it is recommended to conduct remote REST API tests from a remote node. This can be done using the REST APIs described in the following links:
        - Reports REST API

Expected report, without errors and alarms:



Example of errors and alarms in the health report:



For errors and alarms, see UFM Events and Alarms and contact NVIDIA Support.

7.3 UFM Telemetry

Unified Fabric Manager Telemetry collects over 120 unique counters (BER, Temperature, Histograms, Retransmissions, and many more) for each port in the InfiniBand fabric, enabling the user to predict which cables are marginal and should be replaced during the bring-up process to avoid malfunctions in the future.
The tool collects data samples from all ports over all the cluster and save the data in csv file.

To collect InfiniBand Link Quality metrics, perform the following:

```
curl http://{machine_ip}:9002/csv/xcset/low_freq_debug >> my_telemetry_file.csv
```

Example:
my_telemetry_file.csv

	D	E	F	G	H	I	J	K	L	M	N
1	port_guid	tag	Device_ID	node_description	lid	port_label	Phy_Manager_State	phy_state	logical_state	Link_speed_active	Link_width_active
2	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	01/05/2001	3	5	4	128	4
3	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	01/03/2002	3	5	4	128	4
4	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/32/2	3	5	4	128	4
5	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/31/2	3	5	4	128	4
6	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/30/2	3	5	4	128	4
7	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	01/03/2001	3	5	4	128	4
8	0xfc6a1c030067cc00	switch+cables	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/29/2	3	5	4	128	4
9	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/28/2	3	5	4	128	4
10	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/26/2	3	5	4	128	4
11	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/23/1	3	5	4	128	4
12	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/22/1	3	5	4	128	4
13	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/21/1	3	5	4	128	4
14	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/20/2	3	5	4	128	4
15	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/19/2	3	5	4	128	4
16	0xfc6a1c030067cc00	switch+cables	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/17/2	3	5	4	128	4
17	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	1/16/1	3	5	4	128	4
18	0xfc6a1c030067cc00	cables++switch	54002	HBHLDY_D7_501-04-05-51.VEIBS0-12-12	8411	01/01/2002	3	5	4	128	4

The following table lists the link monitoring key indicators and provides their descriptions and evaluation criteria.

<table>
<tr><th>Parameter</th><th>Description</th><th>Evaluation Criteria</th></tr>
<tr><td colspan="3">Link State</td></tr>
<tr><td>Phy_state</td><td>Physical link state</td><td>Verify link up (Enumeration value = 5)</td></tr>
<tr><td colspan="3">Link Quality</td></tr>
<tr><td>NDR Link Quality</td><td>Link Quality criteria depend of error correction scheme type.</td><td>
<table>
<tr><th rowspan="2">Error Correction Scheme TYPE</th><th rowspan="2">Media Type</th><th colspan="3">Post-FEC</th><th colspan="3">Symbol</th></tr>
<tr><th>Normal</th><th>Warning</th><th>Error</th><th>Normal</th><th>Warning</th><th>Error</th></tr>
<tr><td>Default for DAC/ACC/ AOC < 100m

Low_Latency_RS_FEC_PLR</td><td>DAC/ ACC/ AOC</td><td>1.00E-12</td><td>5.00E-12</td><td>1.00E-11</td><td>1.00E-15</td><td>5.00E-15</td><td>1.00E-14</td></tr>
<tr><td>Default for AOC> 100m

KP4_Standard_RS_FEC</td><td>AOC</td><td>1.00E-15</td><td>5.00E-15</td><td>1.00E-14</td><td>1.00E-15</td><td>5.00E-15</td><td>1.00E-14</td></tr>
</table>
DAC - directly attach copper
ACC - active copper cable
AOC - active optical cable

⚠ Minimum port up time for BER measurement - 125 minutes.
</td></tr>
<tr><td colspan="3">PHY Errors</td></tr>
</table>

Parameter	Description	Evaluation Criteria

Link_Down counter	Total number of link down occurred as a result of involuntary link shutdown.	If delta from last sample > 0:
		• Trace the event and include switch, port, date and time, link down counter.
		• If same switch and port has at least 2 link down occurrences within 24 hours, further investigation required.
		• Note:
		  • Make sure link down was due to involuntary port down from the partner side (e.g. not due to partner server reboot).
		  • The criteria intends to catch major link down events.

Cable Information		
Module_Temperature	Temperature of the transceiver - optic transceiver only	There is an alarm and threshold for each transceiver. Usually Warning [70c, 0c] and Alarm [80c, -10c]
rx_power_lane_x and tx_power_lane_x	Rx power and Tx power per transceiver lane - optic transceiver only	There is an alarm and threshold for each transceiver.

7.4 UFM Events and Alarms

Using UFM events and alarms, it allows you to identify any problems including ports and device connectivity.

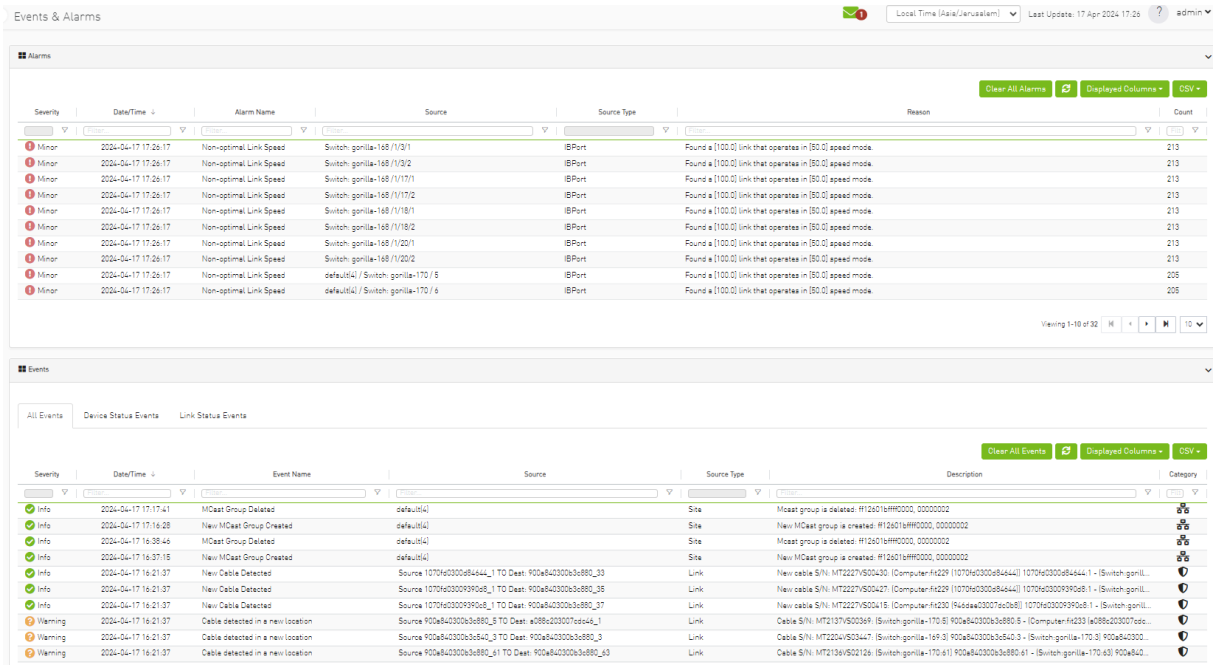
Problems can be detected both prior to running applications and during standard operation.

Events trigger alarms (except for "normal" events. i.e., Info events) when they exceed a predefined threshold.

For more information, see UFM user manual.

UFM alerts can detect a lot of scenarios, for example, bad link, low bandwidth, duplicate GUIDs, non-responsive switch, etc.

For the scenario list and explanation about how to detect and solve the issue, refer to list of scenarios.


Events & Alarms
Local Time (Asia/Jerusalem)
Last Update: 17 Apr 2024 17:26
admin
Alarms
Clear All Alarms
Displayed Columns
CSV
Severity
Date/Time
Alarm Name
Source
Source Type
Reason
Count
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/3/1 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/3/2 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/17/1 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/17/2 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/18/1 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/18/2 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/20/1 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed Switch: gorilla-168 /1/20/2 IBPort Found a [100.0] link that operates in [50.0] speed mode. 213
Minor 2024-04-17 17:26:17 Non-optimal Link Speed default(4) / Switch: gorilla-170 / 5 IBPort Found a [100.0] link that operates in [50.0] speed mode. 205
Minor 2024-04-17 17:26:17 Non-optimal Link Speed default(4) / Switch: gorilla-170 / 6 IBPort Found a [100.0] link that operates in [50.0] speed mode. 205
Viewing 1-10 of 32
10
Events
All Events
Device Status Events
Link Status Events
Clear All Events
Displayed Columns
CSV
Severity
Date/Time
Event Name
Source
Source Type
Description
Category
Info 2024-04-17 17:17:41 MCast Group Deleted default(4) Site Mcast group is deleted: ff12601bffff0000, 00000002
Info 2024-04-17 17:16:28 New MCast Group Created default(4) Site New MCast group is created: ff12601bffff0000, 00000002
Info 2024-04-17 16:38:46 MCast Group Deleted default(4) Site Mcast group is deleted: ff12601bffff0000, 00000002
Info 2024-04-17 16:37:15 New MCast Group Created default(4) Site New MCast group is created: ff12601bffff0000, 00000002
Info 2024-04-17 16:21:37 New Cable Detected Source 1070fd0300d84644_1 TO Dest: 900a840300b3c880_33 Link New cable S/N: MT2227VS00430. (Computer:ht229 (1070fd0300d84644)) 1070fd0300d84644:1 - (Switch:gorill...
Info 2024-04-17 16:21:37 New Cable Detected Source 1070fd03009390c8_1 TO Dest: 900a840300b3c880_35 Link New cable S/N: MT2227VS00427. (Computer:ht229 (1070fd03009390c8)) 1070fd03009390c8:1 - (Switch:gorill...
Info 2024-04-17 16:21:37 New Cable Detected Source 1070fd03009390c8_1 TO Dest: 900a840300b3c880_37 Link New cable S/N: MT2227VS00415. (Computer:ht230 (f46dae03007dc0b8)) 1070fd03009390c8:1 - (Switch:gorill...
Warning 2024-04-17 16:21:37 Cable detected in a new location Source 900a840300b3c880_9 TO Dest: a088c203007cdc46_1 Link Cable S/N: MT2137VS00369. (Switch:gorilla-170:5) 900a840300b3c880:5 - (Computer:ht233 (a088c203007cdc...
Warning 2024-04-17 16:21:37 Cable detected in a new location Source 900a840300b3c540_3 TO Dest: 900a840300b3c880_3 Link Cable S/N: MT2204VS03447. (Switch:gorilla-169:3) 900a840300b3c540:3 - (Switch:gorilla-170:3) 900a840300...
Warning 2024-04-17 16:21:37 Cable detected in a new location Source 900a840300b3c880_61 TO Dest: 900a840300b3c880_63 Link Cable S/N: MT2136VS02126. (Switch:gorilla-170:61) 900a840300b3c880:61 - (Switch:gorilla-170:63) 900a840...

8 Performance Testing

Performance testing verifies that the end-to-end solution works properly under different loads and traffic patterns.

Once the InfiniBand fabric is configured and is healthy based on the NVPS best practices, the network is ready to deliver maximum performance.

ClusterKit is the recommended tool for InfiniBand end-to-end performance validation. ClusterKit is a multifaceted node assessment tool for high-performance clusters. It is capable of testing latency, bandwidth, adequate bandwidth, memory bandwidth, GFLOPS by node, per-rack collective performance, and bandwidth and latency between GPUs and local/remote memory.

The tool employs well-known techniques and tests to achieve these performance metrics. It is intended to give the user a general look of the cluster's health and performance.

Setup

1. Create a shared directory that all systems under test (SUTs) can access.
2. Ensure that all SUTs have keyless SSH access to each other, with strict host key disabled.
3. Download the suitable HPC-X package from here - in the download tab under Resources at the bottom.
4. Untar the package.

```
tar jxf <package>
```

5. cd into ClusterKit sub-directory.

```
cd <shared-dir>/<package>/clusterkit
```

6. Create file 'hostfile.txt'.

```
vi hostfile.txt
```

7. Insert the SUTs names, and save. For example:

```
node1
node2
node3
...
```

8. Find active HCAs names to be tested.

```
ibnodes | grep <node-name>
```

> ⚠ Make sure that the tested HCAs appear as active HCAs for all the SUTs (run ibnodes and grep for each nodes).

> ⚠ Keep the list of tested HCAs in sorted order.

9. Create 'run_clusterkit.sh' script file.

```
vi run_clusterkit.sh
```

10. Paste the following script.

```
#!/bin/bash
HPCX_DIR=$HPCX_DIR
HOSTFILE=$HPCX_DIR/clusterkit/hostfile.txt
CK_DIR=$HPCX_DIR/clusterkit
##############################
# temporary - udpate output scripts
file=$CK_DIR/bin/output/run.sh
[ -e $file ] && (grep -q "\--no-cache-dir" $file || sed -i 's/pip install/pip install --no-cache-dir/g' $fi
le)
file=$CK_DIR/bin/clusterkit.sh
[ -e $file ] && sed -i "/^output_dir/s|=.*$|=\${CK_OUTPUT_SUBDIR:-\"\"}|g" $file
##############################
hcas=(<HCA_NUMBERS>)
declare -A hca_name
<HCA_NAMES_ASSIGNMENT>
mpi_opt+="-x CK_OUTPUT_SUBDIR -x CLUSTERKIT_HCA"
DATE=$(date "+%Y%m%d_%H%M%S")
CK_DATE_DIR=$CK_DIR/$DATE
for i in "${hcas[@]}"; do
  device=${hca_name[${i}]}
  export CK_OUTPUT_SUBDIR=${CK_DATE_DIR}_${device}
  export CLUSTERKIT_HCA=$i
  $CK_DIR/bin/clusterkit.sh --mpi_opt "$mpi_opt" --hpcx_dir $HPCX_DIR --hostfile $HOSTFILE --normalize --
mapper $CK_DIR/core_to_hca.sh --exe_opt --unidirectional
done
```

11. Substitute <HCA_NUMBERS> with list of tested HCAs numbers, separated with space.
    For example, if in step 8 we gathered these HCAs to be tested: mlx5_1, mlx5_2, mlx5_5, then replace <HCA_NUMBERS> with: 1 2 5
12. Substitute <HCA_NAMES_ASSIGNMENT> with insertion of the tested HCAs into the 'hca_name' array.
    Similar to the previous example (step 11), replace <HCA_NAMES_ASSIGNMENT> with:

```
hca_name["1"]=mlx5_1
hca_name["2"]=mlx5_2
hca_name["5"]=mlx5_5
```

    The final script (after all substitutions) for this example should look like this:

```
#!/bin/bash
HPCX_DIR=$HPCX_DIR
HOSTFILE=$HPCX_DIR/clusterkit/hostfile.txt
CK_DIR=$HPCX_DIR/clusterkit
##############################
# temporary - udpate output scripts
file=$CK_DIR/bin/output/run.sh
[ -e $file ] && (grep -q "\--no-cache-dir" $file || sed -i 's/pip install/pip install --no-cache-dir/g' $fi
le)
file=$CK_DIR/bin/clusterkit.sh
[ -e $file ] && sed -i "/^output_dir/s|=.*$|=\${CK_OUTPUT_SUBDIR:-\"\"}|g" $file
##############################
hcas=(1 2 5)
declare -A hca_name
hca_name["1"]=mlx5_1
hca_name["2"]=mlx5_2
hca_name["5"]=mlx5_5
mpi_opt+="-x CK_OUTPUT_SUBDIR -x CLUSTERKIT_HCA"
DATE=$(date "+%Y%m%d_%H%M%S")
CK_DATE_DIR=$CK_DIR/$DATE
for i in "${hcas[@]}"; do
  device=${hca_name[${i}]}
  export CK_OUTPUT_SUBDIR=${CK_DATE_DIR}_${device}
  export CLUSTERKIT_HCA=$i
  $CK_DIR/bin/clusterkit.sh --mpi_opt "$mpi_opt" --hpcx_dir $HPCX_DIR --hostfile $HOSTFILE --normalize --
mapper $CK_DIR/core_to_hca.sh --exe_opt --unidirectional
done
```

13. Save the script file, and make it executable.

```
chmod 777 run_clusterkit.sh
```

14. Create 'core_to_hca.sh' script file.

```
vi core_to_hca.sh
```

15. Paste the following script.

```
#!/bin/bash -x
case $CLUSTERKIT_HCA in
  <CORES>
esac
taskset -c $core $*
```

16. Replace <CORES> as the following instructions:
    a. For each tested HCA, find 2 core numbers across all SUTs

    ```
    cat /sys/class/infiniband/<HCA>/device/local_cpulist
    ```

    > Make sure to find 2 cores that appear in all nodes for the same HCA. For example, if outputs of this command with HCA 'mlx5_1' in all nodes contain cores 1, 2, then you can use them with that HCA

    Similar to the example we used before, if all hosts agreed that:
    - HCA 'mlx5_1' supported with cores 0, 10 (across all nodes)
    - HCA 'mlx5_2' supported with cores 0, 10 (across all nodes)
    - HCA 'mlx5_5' supported with cores 1, 11 (across all nodes)

    Then replace <CORES>

    b. Replace <CORES> in the script as the following:
    Similar to the example we used before, if all hosts agreed that (according to the previous sub-step 16.a.):
    - HCA 'mlx5_1' supported with cores 0, 10 (across all nodes)
    - HCA 'mlx5_2' supported with cores 0, 10 (across all nodes)
    - HCA 'mlx5_5' supported with cores 1, 11 (across all nodes)

    Then replace <CORES> with:

    ```
    1) core=0,10; export UCX_NET_DEVICES=mlx5_1:1;;
    2) core=0,10; export UCX_NET_DEVICES=mlx5_2:1;;
    5) core=1,11; export UCX_NET_DEVICES=mlx5_5:1;;
    ```

    The final script for that example should look like:

    ```
    #!/bin/bash -x
    case $CLUSTERKIT_HCA in
      1) core=0,10; export UCX_NET_DEVICES=mlx5_1:1;;
      2) core=0,10; export UCX_NET_DEVICES=mlx5_2:1;;
      5) core=1,11; export UCX_NET_DEVICES=mlx5_5:1;;
    esac
    taskset -c $core $*
    ```

17. Save the script file, and make it executable.

```
chmod 777 core_to_hca.sh
```

18. Verify/update output script configuration file.
    a. Check the configuration file.

    ```
    # while still in clusterkit/ sub-dir
    vi bin/output/output_config.ini
    ```

    b. Search for 'data_dir='.

c. Verify that the value is the actual path to ibdiagnet directory, usually at: /var/tmp/ibdiagnet2/.

Run ClusterKit test

1. Go to HPCX root directory.

```
cd <SHARED-DIR>/<HPCX-PACKAGE-DIR>/
```

2. Export environment variables (required in each new shell/session).

```
export HPCX_DIR=$PWD
export CK_DIR=$HPCX_DIR/clusterkit
```

3. Run the script.

```
./clusterkit/run_clusterkit.sh
```

4. For each of the tested HCAs, results should be available at: <HPCX-PACKAGE-DIR>/clusterkit/<test-timestamp>_<HCA>.

For further information, see ClusterKit.

Results Verification

Your cluster's performance is satisfactory when the minimum achieved result is at least 95% of the maximum available bandwidth, as illustrated in the table below.

For your convenience, the technology of your cluster interconnect is shown in the header of the bandwidth.txt file.

Expected InfiniBand Performance (for 4x Connections)

Technology	Speed, Gb/s	95% performance, MB/s
HDR	200	23,030
NDR	400	46,060

Visualize results

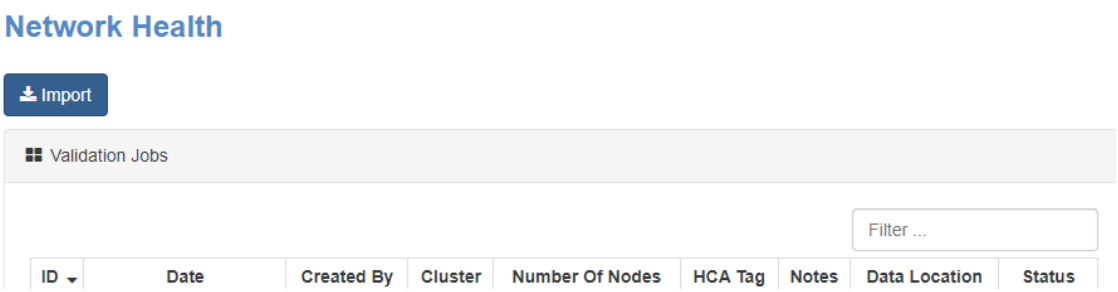
Analyzing the results involves using the UFM-Fabric Visualization Plugin, which is packaged as a docker image that can be run by any docker engine.

1. Install docker, if needed.
2. Pull and run the docker image to run the visualizer app.

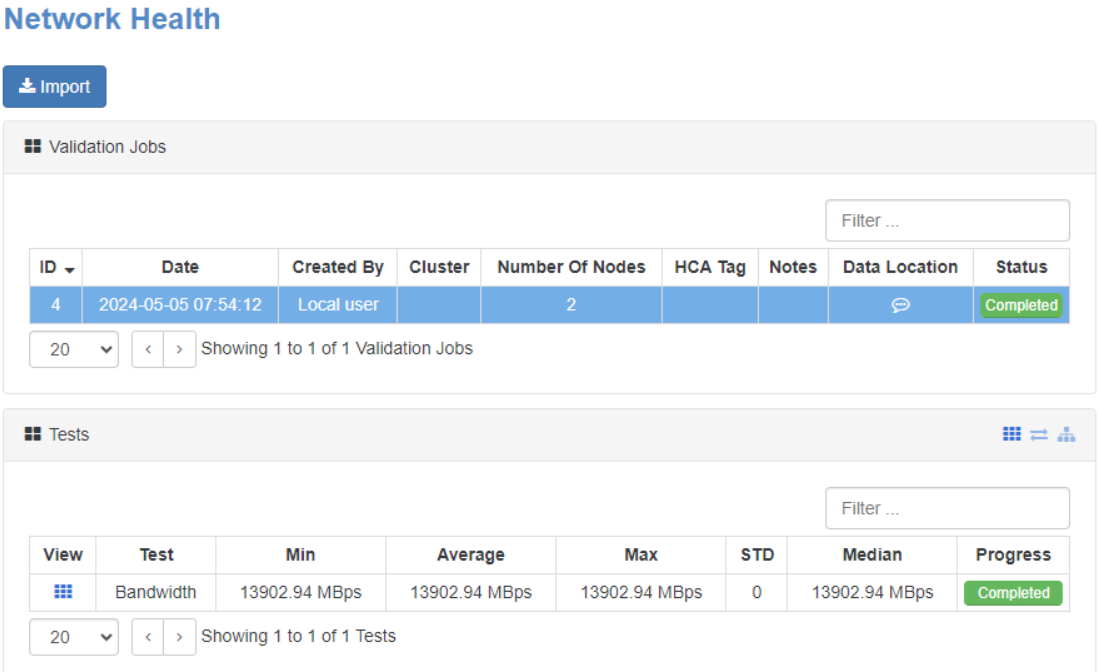
```
sudo docker pull mellanox/ufmfv
sudo docker run -dit --name ufmfv -p 9000:9000 mellanox/ufmfv
```

3. Enter to the server GUI at: http://<server-ip>:9000.
4. Upload/import the results, using the import button.

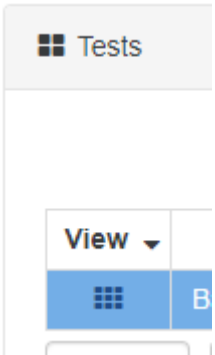
> ⚠ Upload the .json files only (not the .txt) of the latency/bandwidth tests.



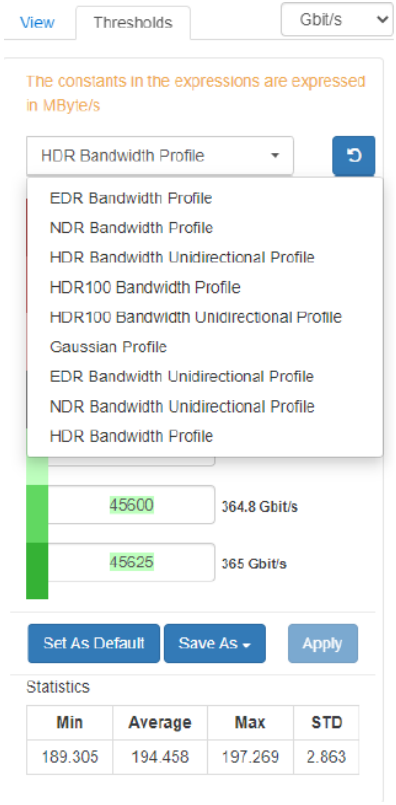
5. Select the new row that was added to the Validation Job table.
   Tests table should appear now.



6. Click on the small matrix icon in the 'View' cell of the new row under 'Tests' table to open the visualization of the uploaded test.

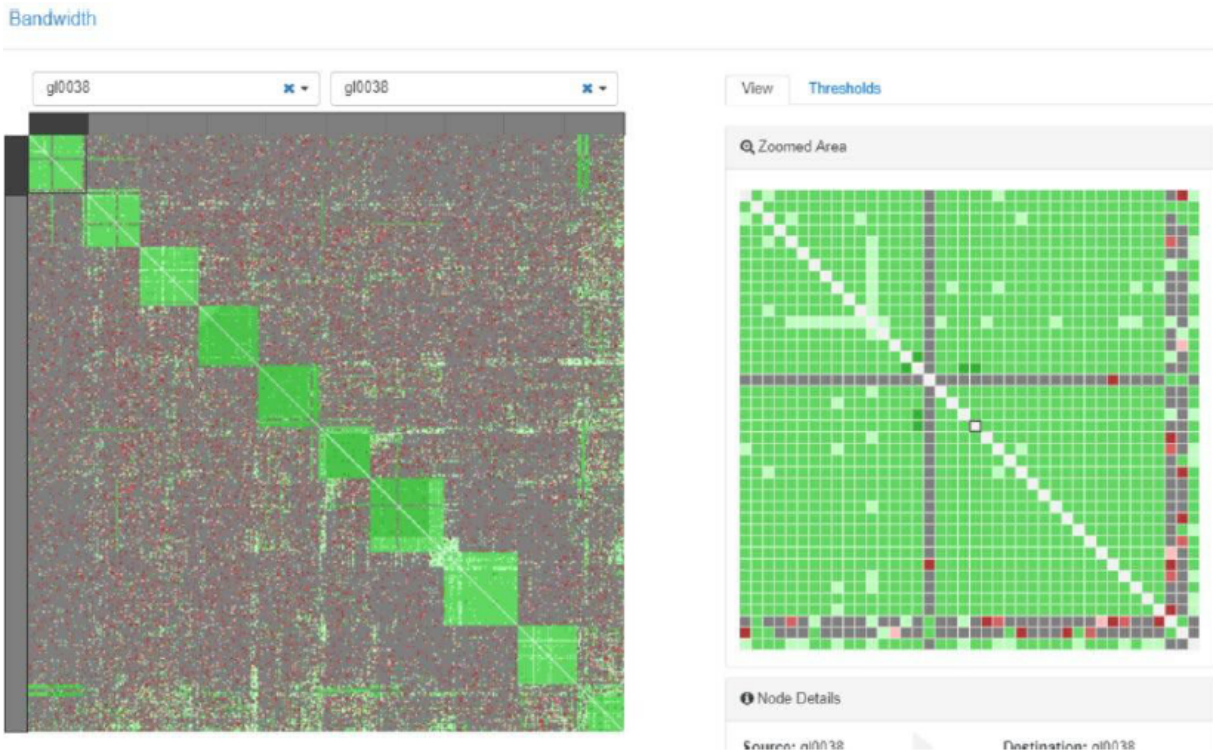


7. Choose the suitable thresholds from the panel on the right.



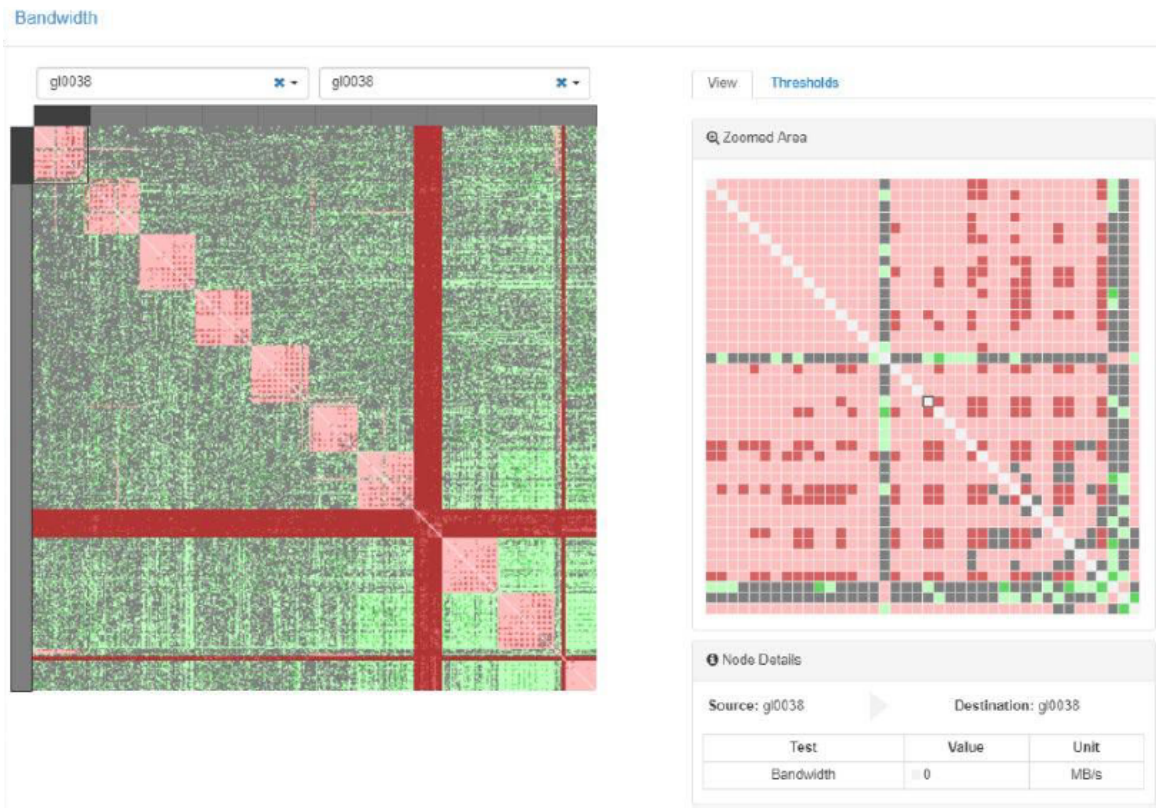
8.1 Normal Results

Normal results are shown in the figure below. The green ‘better’ latency results on the diagonal, as these are nodes that are on the same switch, so have the smallest possible latency. This is because there are no additional switch hops and the data goes through just two cables, and there is never congestion/blocking.



Abnormal results

Two actual abnormal results are shown in the figure below. The first is that nodes on the same switch exhibit poor behavior. This result was used to diagnose an issue with split cables.
The second issue is the red +, which is an indication that a given node has poor performance with all other nodes. In this particular case, quite a few nodes are problematic. The width of the '+' indicates that it is not just a single node but several. There were two poorly performing nodes in this example.


Bandwidth
gl0038
gl0038
View
Thresholds
Zoomed Area
Node Details
Source: gl0038
Destination: gl0038
Test
Value
Unit
Bandwidth
0
MB/s

9 Bring-up Process Checklist

#	Stage	Description	Entry Criteria	Done Criteria
1	Cluster Planning - Choose Topology	Setting the InfiniBand Cluster Topology	-	• Desired topology was defined
2	Cluster Planning - Create PTP File	Creating a Point-to-Point Excel File	• Desired topology was defined	• PTP file was created as described
3	Cluster Planning - Create & Save Topo File	Creating a Topology File Saving the Topology File	• Valid PTP file	• Run topo file successfully • Generated topo file renamed with meaningful name • Topo file is saved for future usage
4	Topology Confirmation - Install UFM	UFM Enterprise Installation	• Previous steps completed • Cluster components physically Installed/deployed	• UFM successfully installed and configured (including HA) • UFM status is running • UFM GUI works
5	Topology Confirmation using UFM	Topology Confirmation using UFM	• Generated topo file • UFM installation done criteria	• Custom Topology Compare Report is cleared from errors/ warnings
6	Network Deployment - SW & FW versions Alignment	Confirm Components' Firmware and Software Versions On-site Upgrade – Low scale	• UFM working with GUI	• All versions are aligned and confirmed
7	Configurations	Configuration and Basic Features Activation	• All previous sections are successfully completed	• Switch configuration done • UFM configuration done • Other optional configurations done

#	Stage	Description	Entry Criteria	Done Criteria
8	Cluster Verification	Cluster Verification	• All previous sections are successfully completed	• SM logs are verified and all issues were treated • UFM Fabric Health report is cleared of errors/alarms • The specified UFM Telemetry indicators comply the evaluation criteria as specified here • UFM events and alarms are monitored and treated • All links are up and with valid quality as described in UFM Telemetry
9	Performance	Performance Testing	• Cluster verification stage successfully completed	• HPC-X package successfully installed and ClusterKit is working • ClusterKit run results are as expected according to 'Results Verification' section • Optional - results were visualized and analyzed as described in 'Visualize results' section

10 Acronyms

Acronym	Description
CG	Acronym: Core Group. A collection of CORE switches used to connect to specific SPINE switches within different SLGs within a datacenter.
“CORE” Network Layer / Switch	The switches comprising the 3rd tier of a 3-tier Clos network. “SPINE” switches connect to “CORE” switches. The term applies to both Cluster Interconnected and Data/Storage networks and both Ethernet and IB networks. Other NVIDIA literature refers to this layer as “Super Spine”.
“LEAF” Network Layer / Switch	The switches comprising the 1st tier of a 3-tier Clos network. “Nodes” connect to “LEAF” switches. The term applies to both Cluster Interconnected and Data/Storage networks and both Ethernet and IB networks. Other NVIDIA literature refers to this layer as “TOR” or “Top-of-Rack”.
SLG	Acronym: Spine-Leaf Group. A collection of SPINE and LEAF layer devices that are interconnected (all SPINE connect to all LEAF within the group).
SU	Acronym: Scalable Unit.
“SPINE” Network Layer / Switch	The switches comprise the 2rd tier of a 3-tier Clos network. “LEAF” switches connect to “SPINE” switches. The term applies to both Cluster Interconnected and Data/Storage networks and both Ethernet and IB networks.

11 Document Revision History

Date	Version	Change
May 27, 2024	1.0	First release

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. Neither NVIDIA Corporation nor any of its direct or indirect subsidiaries and affiliates (collectively: “NVIDIA”) make any representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice. Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, and Mellanox are trademarks and/or registered trademarks of NVIDIA Corporation and/