Table of Contents

Preface ......................................................................................................................................................... iv

1. Requirements ........................................................................................................................................... 1
   1.1. Rocks Version ................................................................................................................................... 1
   1.2. Other Rolls ..................................................................................................................................... 1

2. Installing the mlnx-ofed Roll ................................................................................................................. 2
   2.1. Adding the Roll .............................................................................................................................. 2
   2.2. Adding the mlnx-ofed Roll to an already-installed Frontend ...................................................... 2

3. Using the mlnx-ofed Roll ....................................................................................................................... 4
   3.1. Using the mlnx-ofed Roll ................................................................................................................ 4
   3.2. Using MPI-Selector with mlnx-ofed roll .................................................................................... 5
   3.3. Debugging tools included with mlnx-ofed roll .......................................................................... 6

4. Copyrights ............................................................................................................................................... 8
   4.1. Copyrights .................................................................................................................................... 8
Preface

The mlnx-ofed Roll installs and configures the driver stack and MPI layer as provided in the MLNX_OFED ISO from Mellanox Technologies.

Please visit the Mellanox Technologies site¹ to learn more about their release and the individual software components.

Notes

Chapter 1. Requirements

1.1. Rocks Version

The mlnx-ofed Roll is for use with Rocks version 5.1 ("V.I").

1.2. Other Rolls

The mlnx-ofed Roll is does not require any other Rolls (other than the HPC Roll) to be installed on the Frontend. Compatibility has been verified with the following Rolls.

• base
• kernel
• os
• intel-developer
• intel-icr
• pgi
• sge
• torque
• moab
• lsfhpcl
Chapter 2. Installing the mlnx-ofed Roll

2.1. Adding the Roll

The mlnx-ofed Roll can be installed during the Frontend installation step of your cluster (refer to section 1.2 of the Rocks usersguide), or onto a running system.

The mlnx-ofed Roll is added to a Frontend installation in exactly the same manner as the HPC Roll. Specifically, after the HPC Roll is added the installer will once again ask if you have a Roll (see below). Select ‘Yes’ and insert the mlnx-ofed Roll.

2.2. Adding the mlnx-ofed Roll to an already-installed Frontend

The mlnx-ofed Roll can be added onto an already-installed frontend. You must be the root user to perform this procedure. For sake of discussion, we assume that you have an iso image of the mlnx-ofed roll called
mlnx-ofed.iso. The following procedure will allow you to add the roll.

```bash
# rocks add roll mlnx-ofed.iso
# rocks enable roll mlnx-ofed
# cd /export/rocks/install
# rocks create distro
# kroll mlnx-ofed | sh
```

Reboot your frontend to complete configuration:

```bash
# init 6
```

Reinstall your compute nodes:

```bash
# cluster-fork /boot/kickstart/cluster-kickstart
or
# tentakel /boot/kickstart/cluster-kickstart
```
3.1. Using the mlnx-ofed Roll

The mlnx-ofed roll installs Infiniband drivers, MVAPICH, Open MPI and includes an auto firmware flashing routine during the node kickstart process. A debug file is provided containing output from the auto-installation and update process. This can be a helpful tool in diagnosing HCA issues. The file is located at /root/mlnx-ofed-debug.out. The following is an example of information included in the debug file:

Below is the list of OFED packages that you have chosen (some may have been added by the installer due to package dependencies):

- kernel-ib
- kernel-ib-devel
- ofed-scripts
- libibverbs
- libibverbs-devel
- librdmacm
- librdmacm-devel
- dapl-v2
- tvflash

Uninstalling the previous version of OFED
Build ofa_kernel RPM
Running rpmbuild --rebuild --define '_topdir /var/tmp/OFED_topdir' --define 'configure_options ' --
Install kernel-ib RPM:
Running rpm -iv /opt/mlnx-ofed/src/OFED-1.4/RPMS/centos-release-5-3.el5.centos/x86_64/kernel-ib-1.4-
Install kernel-ib-devel RPM:
Running rpm -iv /opt/mlnx-ofed/src/OFED-1.4/RPMS/centos-release-5-3.el5.centos/x86_64/kernel-ib-devel

The auto firmware flashing routine prints out information at the end of the debug file. If the process fails to update in a failsafe mode, useful information may be provided. Here is an example of an error message related to an invariant sector found on an HCA that will not allow a failsafe update:

probing devices

discovered dev: mt25204_pci_cr0
standard firmware for PSID[MT_03B0120002] is 1.2.0
probed fw version 1.2.917
need to update firmware
mlxburn -dev /dev/mst/mt25204_pci_cr0 -fw fw-25204-rel-1_2_000/fw-25204-rel.mlx -conf fw-25204-rel-1_2_000/MHGS18-XTC_A2-A3.ini -force
-I- Generating image ...

Current FW version on flash: 1.2.917
New FW version: 1.2.0

Note: The new FW version is not newer than the current FW version on flash.

Do you want to continue ? (y/n) [n] : y
Read and verify Invariant Sector - DIFF DETECTED

Invariant sector mismatch. Address 0x40 in image: 0x15000720, while on flash: 0x14000720

The invariant sector can not be burnt in a failsafe manner. You can perform the FW update without burning the invariant sector by specifying the --skip_is flag. See FW release notes for details on invariant sector updates.

*** ERROR *** Failsafe burn error: Invariant sector mismatch
-E- Image burn failed: child process exited abnormally

The firmware update failed due to an Invariant Sector error.

This could be due to Cisco firmware loaded on the hca or perhaps old Mellanox firmware. If you would like to update the firmware, you may use following command. This command will update the firmware in NON FAILSAFE mode. If the system loses power during firmware update it could render the hca inoperable. Proceed with the firmware update at your own risk.

cd /opt/mlnx-ofed/firmware
mlxburn -dev /dev/mst/mt25204_pci_cr0 -fw fw-25204-rel-1_2_000/fw-25204-rel.mlx -conf fw-25204-rel-1_2_000/MHGS18-XTC_A2-A3.ini -force -nofs

In the case above, you have the option of updating the firmware using the provided commands. The instructions provided in the above output are only guidelines on updating software images for HCAs. Please consult with your hardware supplier support team for more information. Additional information about firmware flashing can be found on the Mellanox Technologies Firmware Support and Downloads Site.

3.2. Using MPI-Selector with mlnx-ofed roll

The mlnx-ofed Roll allows users to easily switch between default MPI implementations. Here are a few simple commands that users can execute:

Display current default MPI version:

[user@frontend ~]$ mpi-selector --query
openmpi_intel-1.2.8
level:user

Display available versions:

[user@frontend ~]$ mpi-selector --list
mvapich_gcc-1.1.0
mvapich_intel-1.1.0
mvapich_pgi-1.1.0
openmpi_gcc-1.2.8
openmpi_intel-1.2.8
openmpi_pgi-1.2.8

Switch to another version:

[user@frontend ~]$ mpi-selector --set mvapich_intel-1.1.0
Chapter 3. Using the mlnx-ofed Roll

Defaults already exist; overwrite them? (y/N) y
[user@frontend ~]$ mpi-selector --query
default:mvapich_intel-1.1.0
level:user

A menu driven version of mpi-selector is also available:

[user@frontend ~]# mpi-selector-menu
Current system default: none
Current user default: none

"u" and "s" modifiers can be added to numeric and "U"
commands to specify "user" or "system-wide".

1. mvapich_gcc-1.1.0
2. mvapich_intel-1.1.0
3. mvapich_pgi-1.1.0
4. openmpi_gcc-1.2.8
5. openmpi_intel-1.2.8
6. openmpi_pgi-1.2.8
U. Unset default
Q. Quit

Selection (1-6[us], U[us], Q):

More information about using the mpi-selector tool can be found in the man pages ([user@frontend ~] man
mpi-selector). Detailed information on adding and removing your own MPI implementations is also referenced in
the man page.

3.3. Debugging tools included with mlnx-ofed roll

The mlnx-ofed Roll contains a number of tools that can be used to debug issues with the HCAs and Infiniband fabric.
The ibstat tool will query basic status of Infiniband device(s):

[root@compute-0-0 ~]# ibstat
CA 'mthca0'
    CA type: MT25204
    Number of ports: 1
    Firmware version: 1.2.0
    Hardware version: a0
    Node GUID: 0x0005ad00000c8a40
    System image GUID: 0x0005ad000100d050
    Port 1:
        State: Active
        Physical state: LinkUp
        Rate: 20
        Base lid: 138
        LMC: 0
        SM lid: 2
        Capability mask: 0x02510a68
        Port GUID: 0x0005ad00000c8a41
Chapter 3. Using the mlnx-ofed Roll

The mlnx-ofed roll provides an Infiniband HCA Self Test Utility hca_self_test.ofed which provides firmware version, host driver version & state, link status and other useful information:

```
[root@compute-0-0 ~]# hca_self_test.ofed

---- Performing InfiniBand HCA Self Test ----
Number of HCAs Detected ................ 1
PCI Device Check ........................  PASS
Kernel Arch .............................. x86_64
Host Driver Version ...................... OFED-1.4-2.6.18_128.el5
Host Driver RPM Check ................... PASS
HCA Firmware Check for HCA #0 .......... NA
   REASON: NO required fw version
Host Driver Initialization .............. PASS
Number of HCA Ports Active ............. 1
Port State of Port #0 on HCA #0 ........ UP 4X DDR
Error Counter Check on HCA #0 .......... PASS
Kernel Syslog Check ...................... PASS
Node GUID on HCA #0 ..................... 00:05:ad:00:00:0c:8a:40
------------------ DONE ------------------
```

Output of the auto install and upgrade process can be helpful when troubleshooting issues with Infiniband HCAs. The following output shows the firmware was up-to-date during the auto install and upgrade process:

```
[root@compute-0-0 ~]# tail /root/mlnx-ofed-debug.out

121 ini files registered

probing devices

discovered dev: mt25204_pci_cr0
standard firmware for PSID[MT_03B0120002] is 1.2.0
probed fw version 1.2.0
firmware up-to-date
```

Notes

Chapter 4. Copyrights

4.1. Copyrights

Copyright (c) 2006 - 2009 Clustercorp Inc. All rights reserved. This product includes software developed by Clustercorp Inc., these portions may not be modified, copied, or redistributed without the express written consent of Clustercorp Inc. This product includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center at the University of California, San Diego and its contributors.

Copyright (c) 2006 The Regents of the University of California. All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: 1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. 2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. 3. All advertising materials mentioning features or use of this software must display the following acknowledgement: "This product includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center at the University of California, San Diego and its contributors." 4. Neither the name or logo of this software nor the names of its authors may be used to endorse or promote products derived from this software without specific prior written permission. The name of the software includes the following terms, and any derivatives thereof: "Rocks", "Rocks Clusters", and "Avalanche Installer". THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Copyright (c) 2006 Mellanox Technologies. All rights reserved. This Software is licensed under one of the following licenses: 1) under the terms of the "Common Public License 1.0" a copy of which is available from the Open Source Initiative, see http://www.opensource.org/licenses/cpl.php. 2) under the terms of the "The BSD License" a copy of which is available from the Open Source Initiative, see http://www.opensource.org/licenses/bsd-license.php. 3) under the terms of the "GNU General Public License (GPL) Version 2" a copy of which is available from the Open Source Initiative, see http://www.opensource.org/licenses/gpl-license.php. Licensee has the right to choose one of the above licenses. Redistributions of source code must retain the above copyright notice and one of the license notices. Redistributions in binary form must reproduce both the above copyright notice, one of the license notices in the documentation and/or other materials provided with the distribution.