Mellanox Unstructured Data Acceleration (UDA) Quick Start Guide

Rev 3.0
NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT ("PRODUCT(S)") AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER’S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR 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 FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
# Table of Contents

List of Tables ................................................................. 4
Revision History ............................................................... 5
About This Manual .............................................................. 6
   Intended Audience ......................................................... 6
   Typographical Conventions ............................................ 6
   Common Abbreviations and Acronyms .............................. 7
   Glossary ........................................................................... 8
   Related Documentation .................................................. 10
   Support and Updates Webpage ........................................ 10

Chapter 1 Overview .............................................................. 11
   1.1 Mellanox UDA Solution .............................................. 11
   1.2 Mellanox OFED for Linux ......................................... 11

Chapter 2 Hardware Setup .................................................... 12
   2.1 Setting up the Adapter Cards .................................... 12
   2.2 Setting up the Switch System .................................... 12

Chapter 3 Installing, Configuring and Running UDA Software .......... 13
   3.1 Supported Operating Systems ................................... 13
   3.2 Installation Prerequisites .......................................... 13
   3.3 Installing UDA ........................................................ 13
   3.4 UDA Configuration ................................................. 14
      3.4.1 RDMA Plug-in Parameters Basic Tuning Guidelines .... 17
List of Tables

Table 1: Typographical Conventions .................................................. 6
Table 2: Abbreviations and Acronyms .............................................. 7
Table 3: Glossary ........................................................................... 8
Table 4: Reference Documents ....................................................... 10
Revision History

Rev 3.0 (August 2012)
- Major updates in Chapter 1, “Overview”
- Major updates in Chapter 2, “Hardware Setup”
- Major updates in Chapter 3, “Installing, Configuring and Running UDA Software”

Rev 2.1 (April 11, 2012)
- Renamed the document title (was Mellanox Web 2.0 Acceleration Kit Quick Start Guide)
- Reorganized the sections in Chapter 1, “Overview” and updated links to the software
- Consolidated all adapter cards HW and SW installation into Section 2.1, “Setting up the Adapter Cards,” on page 12
- Consolidated all switch system HW and management SW installation into Section 2.2, “Setting up the Switch System,” on page 12 (the details of the HW installation have been removed; the reader is referred to the switch installation guide for the installation details)
- Added a prerequisite to increase the maximum number of memory translation table segments per HCA in Section 3.2, “Installation Prerequisites,” on page 13
- Updated EULA path in Section 3.3, “Installing UDA,” on page 13
- Updated Section 3.4, “UDA Configuration,” on page 14
- Added Section 3.4.1, “RDMA Plug-in Parameters Basic Tuning Guidelines,” on page 17
- Updated Section 3.6, “Killing Previous Hadoop Runs,” on page 39

Rev 1.1 (October 2011)
- Updated section 3.3, “Mellanox UDA Installation” for UDA 2.0

Rev 1.0 (June 22, 2011)
- First draft
About This Manual

This document describes the setup and configuration of Mellanox Unstructured Data Acceleration (UDA) software package for Hadoop Map Reduce frameworks.

Intended Audience

This manual is intended for system administrators responsible for the installation, configuration, management and maintenance of Mellanox UDA software. It is also intended for application developers.

Typographical Conventions

Table 1 - Typographical Conventions

<table>
<thead>
<tr>
<th>Description</th>
<th>Convention</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>File names</td>
<td>file.extension</td>
<td></td>
</tr>
<tr>
<td>Directory names</td>
<td>directory</td>
<td></td>
</tr>
<tr>
<td>Commands and their parameters</td>
<td>command param1</td>
<td></td>
</tr>
<tr>
<td>Optional items</td>
<td>[ ]</td>
<td></td>
</tr>
<tr>
<td>Mutually exclusive parameters</td>
<td>{ p1</td>
<td>p2</td>
</tr>
<tr>
<td>Optional mutually exclusive parameters</td>
<td>[ p1</td>
<td>p2</td>
</tr>
<tr>
<td>Prompt of a user command under bash shell</td>
<td>hostname$</td>
<td></td>
</tr>
<tr>
<td>Prompt of a root command under bash shell</td>
<td>hostname#</td>
<td></td>
</tr>
<tr>
<td>Prompt of a user command under tcsh shell</td>
<td>tcsh$</td>
<td></td>
</tr>
<tr>
<td>Environment variables</td>
<td>VARIABLE</td>
<td></td>
</tr>
<tr>
<td>Code example</td>
<td>if (a==b){};</td>
<td></td>
</tr>
<tr>
<td>Comment at the beginning of a code line</td>
<td>!, #</td>
<td></td>
</tr>
<tr>
<td>Characters to be typed by users as-is</td>
<td>bold font</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>bold font</td>
<td></td>
</tr>
<tr>
<td>Variables for which users supply specific values</td>
<td>Italic font</td>
<td></td>
</tr>
<tr>
<td>Emphasized words</td>
<td>Italic font</td>
<td>These are emphasized words</td>
</tr>
<tr>
<td>Pop-up menu sequences</td>
<td>menu1 --&gt; menu2 --&gt; ... --&gt; item</td>
<td></td>
</tr>
</tbody>
</table>
### Table 1 - Typographical Conventions

<table>
<thead>
<tr>
<th>Description</th>
<th>Convention</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note</td>
<td>&lt;text&gt;</td>
<td></td>
</tr>
<tr>
<td>Warning</td>
<td>&lt;text&gt;</td>
<td>.</td>
</tr>
</tbody>
</table>

### Common Abbreviations and Acronyms

#### Table 2 - Abbreviations and Acronyms  (Sheet 1 of 2)

<table>
<thead>
<tr>
<th>Abbreviation / Acronym</th>
<th>Whole Word / Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>(Capital) ‘B’ is used to indicate size in bytes or multiples of bytes (e.g., 1KB = 1024 bytes, and 1MB = 1048576 bytes)</td>
</tr>
<tr>
<td>b</td>
<td>(Small) ‘b’ is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)</td>
</tr>
<tr>
<td>FCoE</td>
<td>Fibre Channel over Ethernet</td>
</tr>
<tr>
<td>FW</td>
<td>Firmware</td>
</tr>
<tr>
<td>HCA</td>
<td>Host Channel Adapter</td>
</tr>
<tr>
<td>HW</td>
<td>Hardware</td>
</tr>
<tr>
<td>IB</td>
<td>InfiniBand</td>
</tr>
<tr>
<td>LSB</td>
<td>Least significant <em>byte</em></td>
</tr>
<tr>
<td>lsb</td>
<td>Least significant <em>bit</em></td>
</tr>
<tr>
<td>MSB</td>
<td>Most significant <em>byte</em></td>
</tr>
<tr>
<td>msb</td>
<td>Most significant <em>bit</em></td>
</tr>
<tr>
<td>NIC</td>
<td>Network Interface Card</td>
</tr>
<tr>
<td>SW</td>
<td>Software</td>
</tr>
<tr>
<td>VPI</td>
<td>Virtual Protocol Interconnect</td>
</tr>
<tr>
<td>IPoIB</td>
<td>IP over InfiniBand</td>
</tr>
<tr>
<td>PFC</td>
<td>Priority Flow Control</td>
</tr>
<tr>
<td>PR</td>
<td>Path Record</td>
</tr>
<tr>
<td>RDS</td>
<td>Reliable Datagram Sockets</td>
</tr>
<tr>
<td>RoCE</td>
<td>RDMA over Converged Ethernet</td>
</tr>
<tr>
<td>SDP</td>
<td>Sockets Direct Protocol</td>
</tr>
</tbody>
</table>
The following is a list of concepts and terms related to InfiniBand in general and to Subnet Managers in particular. It is included here for ease of reference, but the main reference remains the InfiniBand Architecture Specification.

### Glossary

The following is a list of concepts and terms related to InfiniBand in general and to Subnet Managers in particular. It is included here for ease of reference, but the main reference remains the InfiniBand Architecture Specification.

#### Table 3 - Glossary

<table>
<thead>
<tr>
<th>Abbreviation / Acronym</th>
<th>Whole Word / Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>Service Level</td>
</tr>
<tr>
<td>SRP</td>
<td>SCSI RDMA Protocol</td>
</tr>
<tr>
<td>MPI</td>
<td>Message Passing Interface</td>
</tr>
<tr>
<td>EoIB</td>
<td>Ethernet over InfiniBand</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>ULP</td>
<td>Upper Level Protocol</td>
</tr>
<tr>
<td>VL</td>
<td>Virtual Lanes</td>
</tr>
<tr>
<td>vHBA</td>
<td>Virtual SCSI Host Bus adapter</td>
</tr>
<tr>
<td>uDAPL</td>
<td>User Direct Access Programming Library</td>
</tr>
</tbody>
</table>

---

**Abbreviation / Acronym**

**Whole Word / Description**

<table>
<thead>
<tr>
<th>Abbreviation / Acronym</th>
<th>Whole Word / Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA (Channel Adapter)</td>
<td>A device which terminates an InfiniBand link, and executes transport level functions</td>
</tr>
<tr>
<td>CLI</td>
<td>Command Line Interface. A user interface in which you type commands at the prompt</td>
</tr>
<tr>
<td>DMA (Direct Memory Access)</td>
<td>Allows hardware to move data blocks directly to the memory, bypassing the CPU</td>
</tr>
<tr>
<td>DNS</td>
<td>Domain Name System. A hierarchical naming system for devices in a computer network</td>
</tr>
<tr>
<td>Fabric Management</td>
<td>The use of a set of tools (APIs) to configure, discover, and manage a group of devices organized as a connected fabric.</td>
</tr>
<tr>
<td>Gateway</td>
<td>A network node that interfaces with another network using a different network protocol</td>
</tr>
<tr>
<td>GUID (Globally Unique IDentifier)</td>
<td>A 64-bit number that uniquely identifies a device or component in a subnet</td>
</tr>
<tr>
<td>GID (Global IDentifier)</td>
<td>A 128-bit number used to identify a Port on a network adapter (see below), a port on a Router, or a Multicast Group.</td>
</tr>
<tr>
<td>HA (High Availability)</td>
<td>A system design protocol that provides redundancy of system components, thus enables overcoming single or multiple failures and minimal downtime</td>
</tr>
</tbody>
</table>
### Table 3 - Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host</td>
<td>A computer platform executing an Operating System which may control one or more network adapters</td>
</tr>
<tr>
<td>IB</td>
<td>InfiniBand</td>
</tr>
<tr>
<td>LID (Local IDentifier)</td>
<td>A 16 bit address assigned to end nodes by the subnet manager</td>
</tr>
<tr>
<td></td>
<td>Each LID is unique within its subnet.</td>
</tr>
<tr>
<td>MTU (Maximum Transfer Unit)</td>
<td>The maximum size of a packet payload (not including headers) that can be sent / received from a port</td>
</tr>
<tr>
<td>Network Adapter</td>
<td>A hardware device that allows for communication between computers in a network</td>
</tr>
<tr>
<td>QoS or Quality of Service</td>
<td>Quality of service is the ability to manage different applications or users by priority such that a required bit rate, delay, packet dropping probability, and/or other measures may be guaranteed.</td>
</tr>
<tr>
<td>RDMA (Remote Direct Memory Access)</td>
<td>Allows accessing memory on a remote side without involvement of the remote CPU</td>
</tr>
<tr>
<td>SA (Subnet Administrator)</td>
<td>The interface for querying and manipulating subnet management data</td>
</tr>
<tr>
<td>SSH</td>
<td>Secure Shell. A protocol (program) for securely logging in to and running programs on remote machines across a network. The program authenticates access to the remote machine and encrypts the transferred information through the connection.</td>
</tr>
<tr>
<td>Subnet Manager (SM)</td>
<td>An entity that configures and manages the subnet, discovers the network topology, assign LIDs, determines the routing schemes and sets the routing tables. There is only one master SM and possible several slaves (Standby mode) at a given time. The SM administers switch routing tables thereby establishing paths through the fabric</td>
</tr>
<tr>
<td>TCA (Target Channel Adapter)</td>
<td>A Channel Adapter that is not required to support verbs, usually used in I/ O devices</td>
</tr>
<tr>
<td>UDA</td>
<td>Unstructured Data Acceleration</td>
</tr>
<tr>
<td>WebUI</td>
<td>Web User Interface. A user interface in which you select commands from drop down menus or by clicking on icons</td>
</tr>
</tbody>
</table>
Related Documentation

Table 4 - Reference Documents

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfiniBand Architecture Specification, Vol. 1, Release 1.2.1</td>
<td>The InfiniBand Architecture Specification that is provided by IBTA</td>
</tr>
<tr>
<td>Mellanox OFED for Linux</td>
<td>Software and documentation can be found at <a href="http://www.mellanox.com/content/pages.php?pg=products_dyn&amp;product_family=26&amp;menu_section=34">http://www.mellanox.com/content/pages.php?pg=products_dyn&amp;product_family=26&amp;menu_section=34</a></td>
</tr>
<tr>
<td>Mellanox MLNX-OS™ Switch Management Software documents</td>
<td>Documentation collateral for MLNX-OS™ CLI, configuration and HowTOs. See <a href="http://www.mellanox.com/content/pages.php?pg=mlnx_os&amp;menu_section=55">http://www.mellanox.com/content/pages.php?pg=mlnx_os&amp;menu_section=55</a></td>
</tr>
<tr>
<td>Firmware Release Notes for Mellanox adapter devices</td>
<td>See the Release Notes PDF file relevant to your adapter device under docs/ folder of installed package.</td>
</tr>
<tr>
<td>ConnectX®-3 Dual Port 40GbE Adapter Card User Manual</td>
<td>This manual provides details of the interfaces of ConnectX-3 EN 40 Gb/s Ethernet adapter cards, specifications, required software and firmware for operating the boards, and relevant documentation. See <a href="http://www.mellanox.com/related-docs/user_manuals/ConnectX-3_Ethernet_Single_and_Dual_QSFP_Port_Adapter_Card_User_Manual.pdf">http://www.mellanox.com/related-docs/user_manuals/ConnectX-3_Ethernet_Single_and_Dual_QSFP_Port_Adapter_Card_User_Manual.pdf</a></td>
</tr>
<tr>
<td>SX6036 SwitchX® 1U 36 Port FDR 56Gb/s InfiniBand Switch Installation Guide Document No. 3489</td>
<td>This manual provides installation and set-up instructions for the SX6036 FDR top of rack InfiniBand Switch platforms. See <a href="http://www.mellanox.com/related-docs/user_manuals/SX60XX_Installation_Guide.pdf">http://www.mellanox.com/related-docs/user_manuals/SX60XX_Installation_Guide.pdf</a></td>
</tr>
<tr>
<td>SX1036 SwitchX® 1U 36 Port QSFP 40Gb/ E Switch Installation Guide Document No. 3468</td>
<td>This manual provides installation and set-up instructions for the SX1036 40Gb/s Ethernet top of rack Switch platforms. See <a href="http://www.mellanox.com/related-docs/user_manuals/SX10XX_Installation_Guide.pdf">http://www.mellanox.com/related-docs/user_manuals/SX10XX_Installation_Guide.pdf</a></td>
</tr>
</tbody>
</table>

Support and Updates Webpage

Please visit the following Web site for downloads, FAQ, troubleshooting, future updates to this manual, etc: http://support.mellanox.com/SupportWeb/software_products/hostacceler_products/UDA.
1 Overview

1.1 Mellanox UDA Solution

Mellanox UDA (Unstructured Data Accelerator) is a software plugin that accelerates Hadoop and improves the scaling of Hadoop clusters executing data-analytics intensive applications. A novel data shuffling protocol is provided for Hadoop to take advantage of RDMA in the network technologies InfiniBand and RoCE (RDMA over Converged Ethernet). Mellanox UDA is an RDMA based software plugin which combined with MLNX Linux (MLNX OFED) inbox driver and ConnectX® based adapter cards will accelerate tasks associated with Map/Reduce file transfer. UDA more than doubles the data processing throughput and reduces CPU utilization by half of Hadoop nodes. Mellanox UDA is developed in collaboration with Auburn University, Alabama.

1.2 Mellanox OFED for Linux

Mellanox OFED for Linux (MLNX_OFED_LINUX) is provided as ISO images, one per a supported Linux distribution, that includes source code and binary RPMs, firmware, utilities, and documentation. The ISO image contains an installation script (called mlnxofedinstall) that performs the necessary steps to accomplish the following:

- Discover the currently installed kernel
- Uninstall any InfiniBand stacks that are part of the standard operating system distribution or another vendor's commercial stack
- Install the MLNX_OFED_LINUX binary RPMs (if they are available for the current kernel)
- Identify the currently installed InfiniBand HCAs and perform the required firmware updates
2 Hardware Setup

2.1 Setting up the Adapter Cards

This manual assumes one or more of the Mellanox ConnectX® family adapter cards is installed in your host machine. Mellanox UDA package takes advantage of the silicon architectures of ConnectX®-3, ConnectX®-2 and ConnectX® based InfiniBand and Ethernet adapter cards. For details, please refer to the relevant adapter card user manual available under www.mellanox.com -> Products -> Adapters.

When using an OEM pre-installed card please refer to the OEM server user manual.

Mellanox UDA requires the installation of Mellanox OFED for Linux driver, version 1.5.3 or later. Mellanox UDA is currently supported on Linux based machines only. Visit the driver Web page below to access software and documents. The supported Linux distributions and kernels are listed in the release notes file; the installation instructions are provided in the user manual. See www.mellanox.com -> Products -> Adapter IB/VPI SW -> Linux SW/Drivers

2.2 Setting up the Switch System

Mellanox UDA benefits from lossless fabric characteristics and requires an RDMA based network. The RDMA capability is available on InfiniBand and RoCE (RDMA over Converged Ethernet) based networks. For the best performance of Mellanox UDA, it is recommended to use Mellanox Ethernet and InfiniBand switches as the software utilizes their architectures.

Visit www.mellanox.com -> Products -> Switches for the state-of-the-art switch portfolio Mellanox offers for Big Data clusters.
3 Installing, Configuring and Running UDA Software

3.1 Supported Operating Systems

Please refer to the product release notes.

3.2 Installation Prerequisites

Prior to installing UDA on a cluster node:

1. Make sure Java Development Kit (JDK) version 1.6.0_25 is already installed on the node.


2. Make sure `ulimit -l` is set to unlimited in all slaves and master nodes. If it is not set, add the following line to your `~/.bashrc` file:

   ```
   ulimit -l unlimited
   ```

3. Verify the current settings in the configuration file.

4. Increase the maximum number of memory translation table segments per HCA:

   ```
   echo "options mlx4_core log_num_mtt=24" > /etc/modprobe.d/mofed.conf
   ```

5. Reboot the server or restart the openibd.

3.3 Installing UDA

The following steps describe how to install the UDA distribution:

**Step 1.** Install Apache Hadoop 1.0.x or Apache Hadoop 1.1. The installation guide and configurations of Apache Hadoop are available at hadoop.apache.org.

**Step 2.** Test your vanilla Hadoop installation to make sure you have a successful and tuned installation. For tuning and configuration details, see http://hadoop.apache.org/common/docs/<Hadoop Version>/cluster_setup.html.

**Step 3.** Patch the plugin (currently supported versions are Hadoop 1.0, CDH3u4 and Hadoop 0.20.2).

   1. Download the appropriate patch from the UDA page on Mellanox’s Support Web site.

   2. Apply the patch as follows:

      ```
      # cd <hadoop extraction directory>
      # patch -p0 --verbose < <patch_name>
      ```

   3. Rebuild.

   4. Optional: Run a Terasort job.

**Step 4.** Install the UDA RPM.

   1. Use the following install command:

      ```
      # sudo rpm -ivh <rpm location>
      ```
2. Make sure that all the files were successfully installed by running the following query. Expected output is listed as command output below.

```
# rpm -ql libuda
/usr/lib64/uda/libuda.so
/usr/lib64/uda/set_hadoop_slave_property.sh
/usr/lib64/uda/uda-CDH3u4.jar
/usr/lib64/uda/uda-hadoop-0.20.2.jar
/usr/lib64/uda/uda-hadoop-1.x.jar
/usr/share/doc/libuda-3.0.1/LICENSE.txt
/usr/share/doc/libuda-3.0.1/README
```

3. Add at the end of your hadoop-env.sh a line containing the jar name matching your hadoop version.

- For Hadoop 1.x add:
  ```
  export HADOOPCLASSPATH=$HADOOPCLASSPATH:/usr/lib64/uda/uda-hadoop-1.x.jar
  ```
- For CDH3u4 add:
  ```
  export HADOOPCLASSPATH=$HADOOPCLASSPATH:/usr/lib64/uda/uda-CDH3u4.jar
  ```
- For Hadoop 0.20.2 add:
  ```
  export HADOOPCLASSPATH=$HADOOPCLASSPATH:/usr/lib64/uda/uda-hadoop-0.20.2.jar
  ```

3.4 UDA Configuration

Assume a cluster with 16 nodes, eagle1 through eagle16, where you wish to set eagle1 as the master of the InfiniBand cluster and the rest as slaves. Similar settings are needed for RoCE based deployments, replacing the InfiniBand host name with the corresponding Ethernet host name.

**Step 1.** For a single homed machines, skip to the next step.

For multi-homed machines, you first need to configure hadoop to use the right interface by setting the “slave.host.name” property to mapred.xml. Each host will have a unique property value along with the appropriate interface. There are two options to configure the host name property:

- **Option 1:** On all slaves and master, add to the file /etc/hosts the hadoop addresses of all hosts in the cluster (in the format: 40.0.0.1 eagle1-ib). In this case the value of the “slave.host.name” property will be “eagle1-ib” for host eagle1, “eagle2-ib” for host eagle2, etc.

- **Option 2:** If you do not wish to change file /etc/hosts, the value of “slave.host.name” property will be the IP address of the appropriate interface. For example, “40.0.0.1” will represent host eagle1, “40.0.0.2” will represent host eagle2, etc. In addition, in the file conf/slaves, slaves need to be listed as the relevant IP addresses of the desired interface. In all properties of conf/*.xml files where the master appears, the master needs to be listed as the IP address of the desired interface.

Under /usr/lib64/uda is a utility script called set_hadoop_slave_property.sh (provided by Mellanox) which edits the “slave.host.name” property. The utility script can help you configure hadoop properly and requires the following parameters:

```
--hadoop-conf-dir=DIR // path to conf directory of the existing Hadoop installation
```
and one of the following two options:

```bash
--host-suffix=HOST_SUFFIX // For Option 1 above: Suffix to be added to hostname
--interface=INTERFACE_NAME // For Option 2 above: Interface name as it appears in 'ifconfig' output
```

**Step 2. XML Configuration:**

1. **HDFS settings:**

Merge the following lines into your `hdfs-site.xml`:

```xml
<?xml version="1.0" standalone="yes"?
<configuration>
  <property>
    <name>dfs.datanode.dns.interface</name>
    <value>ib0</value>
    <description>The name of the Network Interface from which a data node should report its IP address.
  </property>
</configuration>
```

These lines must be in `mapred-*.xml` to be considered during TaskTracker initialization. Therefore, this step cannot be performed per job only.

2. **TaskTracker level settings:**

Merge the following lines into your `mapred-site.xml`:

```xml
<?xml version="1.0" standalone="yes"
<configuration>
  <property>
    <name>mapred.tasktracker.shuffle.provider.plugin</name>
    <value>com.mellanox.hadoop.mapred.UdaShuffleProviderPlugin</value>
    <description>represents plugin for shuffle at TaskTracker side. default value is: (empty string)
You can also try: com.mellanox.hadoop.mapred.UdaShuffleProviderPlugin</description>
  </property>
</configuration>
```
3. Job level settings:

The following additional settings are required for every UDA job. The settings can be provided either in a command line or in mapred-*.xml.

```xml
<?xml version="1.0"?>
<configuration>
  <property>
    <name>mapred.reducetask.shuffle.consumer.plugin</name>
    <value>com.mellanox.hadoop.mapred.UdaShuffleConsumerPlugin</value>
    <description>Represents plugin for shuffle at ReduceTask side. Default is: org.apache.hadoop.mapred.ReduceCopier
    You can also try:
    com.mellanox.hadoop.mapred.UdaShuffleConsumerPlugin</description>
  </property>
</configuration>
```

4. Optional settings:

The following are optional default parameter settings for UDA.

```xml
<?xml version="1.0"?>
<configuration>
  <property>
    <name>mapred.rdma.cma.port</name>
    <description>Port number to be used for the RDMA connection</description>
    <value>9011</value>
  </property>
  <property>
    <name>mapred.rdma.wqe.per.conn</name>
    <description>Number of allocated Work Queue Elements (WQEs) for Receive Queue per connection.</description>
    <value>256</value>
  </property>
</configuration>
```
3.4.1 RDMA Plug-in Parameters Basic Tuning Guidelines

- UdaShuffleProviderPlugin allocates buffers for reading MOFs from the disk and for writing them using RDMA to satisfy reduce task shuffle requests. Therefore, UdaShuffleProviderPlugin's buffer size determines the max buffer size to be used also by reduce tasks.

- When TaskTracker is spawned and the UdaShuffleProviderPlugin is initialized, it is essential that the mapred.rdma.buf.size parameter is properly configured to satisfy reducers. RDMA buffers from each reducer are allocated from mapred.child.java.opts * mapred.job.shuffle.input.buffer.percent.

When UDA is enabled, each reducer must allocate 2*#MOFs = 2*Dataset/Blocksize. Unless you have memory issues we recommend that each RDMA buffer will be of size 1024 KB for optimum performance.

For example, when running a job with a 100GB input size and a 256MB split size, 600 MOFs are created. This requires configuring at least 1200 buffers. Continuing with the above example, a configuration that runs 4 slots of reducers per node requires the allocation of 4 x 1200 = 4800 buffers for the job. By using mapred.rdma.buf.size=1024, a total of 4800MB is allocated per node.