Mellanox Unstructured Data Acceleration (UDA) Quick Start Guide

Rev 3.2.2-0

www.mellanox.com

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 PRODUCTO(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.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Mellanox Technologies, Ltd.
Beit Mellanox
PO Box 586 Yokneam 20692
Israel
www.mellanox.com
Tel: +972 (0)74 723 7200
Fax: +972 (0)4 959 3245

Table of Contents

List of Tables

Document Revision History

Table 1 - Document Revision History

Revision	Date	Description
3.2	September 2013	Updated the following sections: • Section 3.3, "Installing UDA", on page 17 • Section 3.4, "UDA Configuration", on page 18 • Appendix A: "Patching and Building Hadoop," on page 23
3.1.11	June 2013	Updated the following sections: • Section 3.2, "Installation Prerequisites", on page 16 • Section 3.3, "Installing UDA", on page 17 • Section 3.4, "UDA Configuration", on page 18 Added the following sections: • Section 3.5, "UDA Log Setting", on page 22 • Appendix A: "Patching and Building Hadoop," on page 23
3.0	August 2012	Major updates to the following chapters: • Chapter 1, "Overview" • Chapter 2, "Hardware Setup" • Chapter 3, "Installing, Configuring and Running UDA Software"

Table 1 - Document Revision History

Revision	Date	Description
2.1	April 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 15 • Consolidated all switch system HW and management SW installation into Section 2.2, "Setting up the Switch System", on page 15 (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 16 • Updated EULA path in Section 3.3, "Installing UDA", on page 17 • Updated Section 3.4, "UDA Configuration", on page 18 • Added Section 3.4.1, "RDMA Plug-in Parameters Basic Tuning Guidelines", on page 21 • Updated Section 3.6, "Killing Previous Hadoop Runs," on page 39
1.1	October 2011	Updated section 3.3, "Mellanox UDA Installation" for UDA 2.0
1.0	June 2011	Initial release

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 2 - Typographical Conventions

Description	Convention	Example
File names	`file.extension`	
Directory names	`directory`	
Commands and their parameters	`command param1`	
Optional items	`[ ]`	
Mutually exclusive parameters	`{ p1 | p2 | p3 }`	
Optional mutually exclusive parameters	`[ p1 | p2 | p3 ]`	
Prompt of a *user* command under bash shell	`hostname$`	
Prompt of a *root* command under bash shell	`hostname#`	
Prompt of a *user* command under tcsh shell	`tcsh$`	
Environment variables	VARIABLE	
Code example	`if (a==b){};`	
Comment at the beginning of a code line	`!, #`	
Characters to be typed by users as-is	bold font	
Keywords	bold font	
Variables for which users supply specific values	Italic font	

Table 2 - Typographical Conventions

Description	Convention	Example
Emphasized words	Italic font	These are emphasized words
Pop-up menu sequences	menu1 --> menu2 -->... --> item	
Note	 <text>	
Warning	 <text>	.

Common Abbreviations and Acronyms

Table 3 - Abbreviations and Acronyms (Sheet 1 of 2)

Abbreviation / Acronym	Whole Word / Description
B	(Capital) 'B' is used to indicate size in bytes or multiples of bytes (e.g., 1KB = 1024 bytes, and 1MB = 1048576 bytes)
b	(Small) 'b' is used to indicate size in bits or multiples of bits (e.g., 1Kb = 1024 bits)
FCoE	Fibre Channel over Ethernet
FW	Firmware
HCA	Host Channel Adapter
HW	Hardware
IB	InfiniBand
LSB	Least significant *byte*
lsb	Least significant *bit*
MSB	Most significant *byte*
msb	Most significant bit
NIC	Network Interface Card
SW	Software
VPI	Virtual Protocol Interconnect
IPoIB	IP over InfiniBand
PFC	Priority Flow Control

Table 3 - Abbreviations and Acronyms (Sheet 2 of 2)

Abbreviation / Acronym	Whole Word / Description
PR	Path Record
RDS	Reliable Datagram Sockets
RoCE	RDMA over Converged Ethernet
SDP	Sockets Direct Protocol
SL	Service Level
SRP	SCSI RDMA Protocol
MPI	Message Passing Interface
EoIB	Ethernet over InfiniBand
QoS	Quality of Service
ULP	Upper Level Protocol
VL	Virtual Lanes
vHBA	Virtual SCSI Host Bus adapter
uDAPL	User Direct Access Programming Library

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 4 - Glossary

CA (Channel Adapter)	A device which terminates an InfiniBand link, and executes transport level functions
CLI	Command Line Interface. A user interface in which you type commands at the prompt
DMA (Direct Memory Access)	Allows hardware to move data blocks directly to the memory, bypassing the CPU
DNS	Domain Name System. A hierarchical naming system for devices in a computer network
Fabric Management	The use of a set of tools (APIs) to configure, discover, and manage and a group of devices organized as a connected fabric.
Gateway	A network node that interfaces with another network using a different network protocol

Table 4 - Glossary

GUID (Globally Unique IDentifier)	A 64-bit number that uniquely identifies a device or component in a subnet
GID (Global IDentifier)	A 128-bit number used to identify a Port on a network adapter (see below), a port on a Router, or a Multicast Group.
HA (High Availability)	A system design protocol that provides redundancy of system components, thus enables overcoming single or multiple failures and minimal downtime
Host	A computer platform executing an Operating System which may control one or more network adapters
Hadoop	Open source, distributed, big data processing application. (an Apache project)
IB	InfiniBand
LID (Local IDentifier)	A 16 bit address assigned to end nodes by the subnet manager Each LID is unique within its subnet.
MTU (Maximum Transfer Unit)	The maximum size of a packet payload (not including headers) that can be sent /received from a port
Network Adapter	A hardware device that allows for communication between computers in a network
QoS or Quality of Service	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.
RDMA (Remote Direct Memory Access)	Allows accessing memory on a remote side without involvement of the remote CPU
SA (Subnet Administrator)	The interface for querying and manipulating subnet management data
SSH	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.
Subnet Manager (SM)	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
TCA (Target Channel Adapter)	A Channel Adapter that is not required to support verbs, usually used in I/O devices
UDA	Unstructured Data Acceleration

Table 4 - Glossary

UDA Plugin	A software plugin that plugs into the Hadoop application
WebUI	Web User Interface. A user interface in which you select commands from drop down menus or by clicking on icons

Related Documentation

Table 5 - Reference Documents

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

Table 5 - Reference Documents

Document Name	Description
SX1036 SwitchX® 1U 36 Port QSFP 40Gb/E Switch Installation Guide Document No. 3468	This manual provides installation and set-up instructions for the SX1036 40Gb/s Ethernet top of rack Switch platforms. See http://www.mellanox.com/related-docs/user_manuals/SX10XX_Installation_Guide.pdf

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 -> InfiniBand/VPI Drivers -> 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 You have a Hadoop environment installed and running on the node.
2. Make sure `ulimit -l` is set to unlimited in all slaves and master nodes.

   If it is not set:

   a. Add the following line to your `~/.bashrc` file.

   ```
   ulimit -l unlimited
   ```

   b. Set the parameters below as follow in the `/etc/security/limits.conf` file.

   ```
   * soft memlock unlimited
   * hard memlock unlimited
   ```

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

   ```
   # echo "options mlx4_core log_num_mtt=24 log_mtts_per_seg=0[1]" > /etc/modprobe.d/
   mofed.conf
   ```

   1. If you need more than 64GB, you can increase the maximum amount of available RDMA memory by increasing the value of log_mtts_per_seg.

   a. Reboot the server or restart the openibd.

   To restart the openibd:

   ```
   # sudo service openibd restart
   ```

   b. Verify the changes took effect.

   ```
   # cat /sys/module/mlx4_core/parameters/log_num_mtt
   # cat /sys/module/mlx4_core/parameters/log_mtts_per_seg
   ```

4. Disable swap on all the nodes in the cluster.

   The swap option can be disabled in one of the following ways:

   a. Edit the `/etc/fstab` file. Remove the swap file system and run ones the command below.[1]

   ```
   # swapoff -a
   ```

   b. Run the below after every reboot.

   ```
   # swapoff -a
   ```

1. We recommend using this option as it is a one time operation.

3.3 Installing UDA

The following steps describe how to install the UDA distribution:

Step 1. Install Apache Hadoop 1.x.y or 2.x.y. 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 Hadoop with Mellanox plugin ability patch (currently supported versions are Hadoop 1.x.y and 2.x.y, CDH 4.1.2 MRv1, CDH 4.2.1 MRv1 CDH 4.3.1 MRv1, CDH4.4.0 MRv1 and HDP 1.1).

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 <patch_name>
# echo $?[2]
```

2. Please verify the output of this command is 0.

3. Build according to the instruction in Appendix A: "Patching and Building Hadoop," on page 23.
4. [Optional] Run a Terasort job.

Step 4. Install the UDA RPM.

1. Use the following install command:

```
# sudo rpm -Uvh <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/LICENSE.txt
/usr/lib64/uda/README
/usr/lib64/uda/journal.txt
/usr/lib64/uda/libuda.so
/usr/lib64/uda/source.tgz
/usr/lib64/uda/uda-hadoop-1.x-cdh-4.2.jar
/usr/lib64/uda/uda-hadoop-1.x-v1.jar
/usr/lib64/uda/uda-hadoop-1.x-v2.jar
/usr/lib64/uda/uda-hadoop-1.x-v3.jar
/usr/lib64/uda/uda-hadoop-2.x.jar
/usr/lib64/uda/uda-hadoop-3.x.jar
/usr/lib64/uda/utils.tgz
```

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

- For Hadoop 1.x.y, HDP 1.1, CDH 4.1.2 MRv1 and CDH 4.2.1 MRv1 add[1]:

```
export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:/usr/lib64/uda/uda-hadoop-1.x.jar
```

- For Hadoop 2.x.y add:

```
export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:/usr/lib64/uda/uda-hadoop-2.x.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. Note, this is a special property and requires each node host to have a unique property value along with the appropriate interface. The host name can be configured as follow:

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.cluster). In this case, configure as follow:

- hadoop-env.sh:

```
export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -DHADOOPHOSTNAME=`host-
name`.ib.cluster ${HADOOP_OPTS}"
```

- core-site.xml:

```
<property>
      <name>slave.host.name </name>
      <value>${HADOOPHOSTNAME}</value>
</property>
```

Step 2. XML Configuration:

1. HDFS settings:

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

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
    <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.
        </description>
      </property>
    </configuration>
```

2. TaskTracker level settings:

1. When using the old v1 patch and plugin, add: `export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:/usr/lib64/uda/uda-hadoop-1.x-v1.jar`

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

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

```
<?xml version="1.0"?>
    <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
    <configuration>
      <property>[1]
       <name>mapreduce.shuffle.provider.plugin.classes</name>
       <value>com.mellanox.hadoop.mapred.UdaShuffleProviderPlugin,
       org.apache.hadoop.mapred.TaskTracker$DefaultShuffleProvider</value>
       <description>A comma-separated list of classes that should be loaded as
       ShuffleProviderPlugin(s).
       A ShuffleProviderPlugin can serve shuffle requests from reducetasks.
       Each class in the list must be an instance of
       org.apache.hadoop.mapred.ShuffleProviderPlugin.
       </description>
      </property>

      <property>
        <name>mapred.tasktracker.dns.interface</name>
        <value>ib0</value>
      </property>
    </configuration>
```

1. The example below is used for the old v1 version of Mellanox plugin (uda-hadoop-1.x-v1.jar).

```
<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>
```

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 version="1.0"?>
    <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
    <configuration>
      <property>
        <name>mapreduce.job.reduce.shuffle.consumer.plugin.class</name>[1]
        <value>com.mellanox.hadoop.mapred.UdaShuffleConsumerPlugin</value>
        <description>Represents plugin for shuffle at ReduceTask side.
        </description>
      </property>
    </configuration>
```

1. Use `<name>mapred.reducetask.shuffle.consumer.plugin</name>` when using the old v1 version of Mellanox plugin (uda-hadoop-1.x-v1.jar).

4. Optional settings:

The following are optional default parameter settings for UDA.

```
<?xml version="1.0"?>
    <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
    <configuration>
     <property>
      <name>mapred.rdma.compression.buffer.ratio</name>
      <description>The ratio in which memory is divided between RDMA buffer and
       decompression buffer (used only with intermediate data compression)
      </descrition>
      <value>0.20</value>
      </property>
      <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>
      <property>
        <name>mapred.rdma.buf.size</name>
        <value>1024</value>
        <description>Used by both UdaShuffleProvider and
        UdaShuffleConsumer:
        - UdaShuffleProvider (TaskTracker): determines the RDMA&AIO
          Buffers size to satify Map Output's RDMA fetch requests
        - UdaShuffleConsumer (Reducer): user preferred RDMA buffer
          size for fetching map outputs. Size is in KB and must be
          aligned to page size.
        </description>
      </property>
```

```
    <property>
     <name>mapred.rdma.compression.buffer.ratio</name>
     <description>The ratio in which memory is divided between RDMA buffer and
      decompression buffer (used only with intermediate data compression)
     </description>
     <value>0.20</value>
    </property>
    <property>
      <name>mapred.rdma.buf.size.min</name>
      <value>32</value>
      <description>UDA reducer allocates RDMA buffers according to
      'mapred.rdma.buf.size'. If the buffer size is too big then a
      smaller buffer will be used while 'mapred.rdma.buf.size.min'
      is the limit.
      Bigger RDMA buffers improve the shuffle performance.
      Too small buffer sizes can significantly reduce performance.
      The task will fail if the reducer needs to use a buffer size
      smaller than 'mapred.rdma.buf.size.min'.
      </description>
    </property>
  </configuration>
```

5. YARN configuration, for Hadoop-2.x.y. It requires modifying the yarn-site.xml.

```
    <property>
      <name>yarn.nodemanager.aux-services[1]</name>
      <value>mapreduce.shuffle,uda.shuffle</value>
    </property>
    <property>
      <name>yarn.nodemanager.aux-services.uda.shuffle.class</name>
      <value>com.mellanox.hadoop.mapred.UdaShuffleHandler</value>
    </property>
```

1. If you already have entry with then name "yarn.nodemanager.aux-services", you will need to replace it with the new entry

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 x1200 = 4800 buffers for the job. By using mapred.rdma.buf.size=1024, a total of 4800MB is allocated per node.

3.5 UDA Log Setting

UDA Consumer and Provider logs are now integrated with the Hadoop log system and their properties are configured via the Hadoop's `log4j.properties` file (in your <hadoop-conf-dir>).

The Consumer logs are integrated into the ReduceTask whereas the Provider logs are integrated into the TaskTracker.

To configure these modules, add the lines below to the `log4j.properties` files:

- log4j.logger.org.apache.hadoop.mapred.ShuffleConsumerPlugin=<log_level>
- log4j.logger.org.apache.hadoop.mapred.ShuffleProviderPlugin=<log_level>

  The ShuffleProviderPlugin logging level can be changed at runtime. To do so, type:

```
#> bin/hadoop daemonlog -setlevel <hostname>:50060 org.apache.hadoop.mapred.ShuffleProviderPlugin <log level>
```

In the example above, 50060 is the default port value of mapreduce.tasktracker.http.address

If logs are not modified, UDA log level will set to the default setting of the distribution (default is INFO).

Appendix A: Patching and Building Hadoop

This section provides the procedure to add to supported Hadoop distributions the plug-in ability, enabling the application to utilize or disable UDA.

➢ ***To patch and build Hadoop:***

Step 1. Download Hadoop from http://hadoop.apache.org/releases.html.

The URL for hadoop-1.1.2 is: http://apache.mivzakim.net/hadoop/common/hadoop-1.1.2/hadoop-1.1.2.tar.gz

Step 2. Extract the tarball on all the nodes and test your installation.

Step 3. Download Mellanox's patch from http://uda-plugin.googlecode.com.

The URL is: https://uda-plugin.googlecode.com/files/HADOOP-1.x.y-v2.patch

Step 4. Patch hadoop.

1. Extract hadoop-1.1.2.tar.gz in a temp directory.

2. Change directory into the extraction directory.

3. Run the Mellanox patch.

```
# patch -p0 < HADOOP-1.x.y-v2.patch
```

4. Verify the previous operation was successful.

The expected result should be 0.

```
# echo $?
```

Step 5. Build your patched Hadoop.

Hadoop-1.x.y example:

```
ant -Djava5.home=/usr/lib64/java/jdk1.6.0_25 clean tar
```

This will create you a new tar.gz file under the ./build/ dir. (notice that the result will be called hadoop-1.1.3-SNAPSHOT since it is not a default 1.1.2 version).

Hadoop-2.x.y example:

```
mvn package -Pdist -DskipTests -Dtar
```