Chelsio iWARP Installation and Setup Guide.

Contents

1 Revision History 3
2 Installing OFED Driver 4
3 Chelsio iWARP Drivers compatibility with Chelsio Linux drivers 5
3.1 Installing Chelsio cxgb3toe-W.X.YY.ZZZ driver with OFED-X.Y.Z package 6
3.2 Loading Firmware and EEPROM 7
3.2.1 Loading Firmware 7
3.2.2 Loading EEPROM 8

4 Testing connectivity with ping and rping 9
5 Enabling various MPIs 11
5.1 DAPL Library configuration for Intel MPI, HP MPI and Scali MPI 11
5.2 Setting shell for Remote Logon 11
5.3 Configuration of various MPIs 12
5.3.1 MVAPICH2 12
5.3.2 HPMPI 14
5.3.3 Intel -MPI 15
5.3.4 Open-MPI 16

6 Additional Notes and issues 19

1 Revision History

Last Updated: 05/19/2010
Last Updated: 05/04/2010
Last Updated: 02/03/2010
Last Updated: 03/09/2009
Last Updated: 11/11/2008
Last Updated: 9/16/2008
Last Updated: 9/28/2007

2 Installing OFED Driver

1. Download OFED distro from http://service.chelsio.com/ site

```
# tar -xzf path-to/OFED-X.Y.Z.tgz
```

2. Goto OFED-X.Y.Z directory.

```
# cd OFED-X.Y.Z
```

3. Run install.pl

```
# ./install.pl
```

i) Choose option 2 to install OFED package.
ii) Then choose option 3 to install all OFED libraries.
iii) Then choose default options in which come while executing ./install.pl script to build and install OFED
OR
iv) If you are familiar with OFED installation you can choose option 2 then option 4 for customized installation.

4. After installation reboot system for changes to take effect.

5. Set Chelsio driver option for MPI connection changes. Give the below command on all systems

```
# echo 1 > /sys/module/iw_cxgb3/parameters/peer2peer
```

OR
you can add the following line to /etc/modprobe.conf to set the
option at module load time:

```
options iw_cxgb3 peer2peer=1
```

6. The option setting in file /etc/modprobe.conf shall take effect upon system reboot.

3 Chelsio iWARP Drivers compatibility with Chelsio Linux drivers

Following table shows the supported combination of compatible OFED-X.Y.Z package, cxgb3toe-W.X.YY.ZZZ driver and Chelsio Firmware.

OFED Package	Cxgb3toe-W.X.YY.ZZZ driver	Firmware	Supported/Not Supported/Not Tested
OFED-1.5.1	Cxgb3toe-1.4.1.2	7.10.0	Not Supported
OFED-1.5.1	Cxgb3toe-1.4.1.2	7.8.0	Supported
OFED-1.5.1	Cxgb3toe-1.4.1.2	7.4.0	Not Supported
OFED-1.5.1	Cxgb3toe-1.4.0.8	7.10.0	Not Supported
OFED-1.5.1	Cxgb3toe-1.4.0.8	7.8.0	Supported
OFED-1.5.1	Cxgb3toe-1.4.0.8	7.4.0	Supported
OFED-1.5.1	Cxgb3toe-1.3.1.10	7.10.0	Not Supported
OFED-1.5.1	Cxgb3toe-1.3.1.10	7.8.0	Not Supported
OFED-1.5.1	Cxgb3toe-1.3.1.10	7.7.0	Not Tested
OFED-1.5.1	Cxgb3toe-1.3.1.10	7.4.0	Not Supported
OFED-1.5	Cxgb3toe-1.4.1.2	7.10.0	Not Supported
OFED-1.5	Cxgb3toe-1.4.1.2	7.8.0	Not Tested
OFED-1.5	Cxgb3toe-1.4.1.2	7.4.0	Not Supported
OFED-1.5	Cxgb3toe-1.4.0.8	7.10.0	Not Supported
OFED-1.5	Cxgb3toe-1.4.0.8	7.8.0	Supported
OFED-1.5	Cxgb3toe-1.4.0.8	7.4.0	Not Supported
OFED-1.5	Cxgb3toe-1.3.1.10	7.10.0	Not Tested
OFED-1.5	Cxgb3toe-1.3.1.10	7.8.0	Not Supported
OFED-1.5	Cxgb3toe-1.3.1.10	7.7.0	Supported
OFED-1.5	Cxgb3toe-1.3.1.10	7.4.0	Not Supported
OFED-1.4.2	Not Tested	Not Tested	Not Tested
OFED-1.4.1	Cxgb3toe-1.4.1.2	7.10.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.4.1.2	7.8.0	Not Tested
OFED-1.4.1	Cxgb3toe-1.4.1.2	7.4.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.4.0.8	7.10.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.4.0.8	7.8.0	Not Tested
OFED-1.4.1	Cxgb3toe-1.4.0.8	7.4.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.3.1.10	7.10.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.3.1.10	7.8.0	Not Supported
OFED-1.4.1	Cxgb3toe-1.3.1.10	7.7.0	Not Tested
OFED-1.4.1	Cxgb3toe-1.3.1.10	7.4.0	Not Tested
OFED-1.4.1	Cxgb3toe-1.3.0	7.4.0	Supported

Table 1- Chelsio iWARP drivers compatibility with Chelsio Linux drivers

3.1 Installing Chelsio cxgb3toe-W.X.YY.ZZZ driver with OFED-X.Y.Z package.

You can also install Chelsio's cxgb3toe-W.X.YY.ZZZ driver after installing OFED-X.Y.Z package as mentioned in section 1. However it requires specific options to be given while loading the Chelsio iWARP drivers of OFED-X.Y.Z on top of Chelsio's cxgb3toe-W.X.YY.ZZZ driver.

Following are the steps to install cxgb3toe-W.X.YY.ZZZ driver and cxgbtool which comes with driver package.

1. Download cxgb3toe-W.X.YY.ZZZ driver from http://service.chelsio.com

2. Untar cxgb3toe-W.X.YY.ZZZ driver in shared directory

```
# tar -xzf path-to/cxgb3toe-W.X.YY.ZZZ.tgz
```

3. Build/install:

```
# cd cxgb3toe-W.X.YY.ZZZ
# (cd src; make && make install)
# (cd tools/cxgbtool; make && make install)
```

Note: To automatically load Chelsio iWARP drivers on different Linux platforms, please do the following

---

4. On supported RHEL 4.xx and RHEL 5.xx, add this to /etc/modprobe.conf:

```
options iw_cxgb3 peer2peer=1
install cxgb3 /sbin/modprobe -i cxgb3; /sbin/modprobe -f iw_cxgb3;
/sbin/modprobe rdma_ucm
alias eth2 cxgb3 # assume eth2 is used by the Chelsio interface.
```

5. On supported SLES 10 SP-X platform add this to /etc/modprobe.conf:

```
 options iw_cxgb3 peer2peer=1
 install cxgb3 /sbin/modprobe -i cxgb3; /sbin/modprobe --force-modversion
 iw_cxgb3; /sbin/modprobe rdma_ucm
alias eth2 cxgb3 # assume eth2 is used by the Chelsio interface.
```

6. On supported SLES 11 platform add this to /etc/modprobe.conf:

```
options iw_cxgb3 peer2peer=1
 install cxgb3 /sbin/modprobe -i --allow cxgb3; /sbin/modprobe --force
 --allow iw_cxgb3; /sbin/modprobe --allow rdma_ucm
alias eth2 cxgb3 # assume eth2 is used by the Chelsio interface.
```

7. Reboot the system to load the new modules

Note: To manually load Chelsio iWARP drivers on different Linux platform do the following

---

8. On RHEL 4.7, 4.8, 5.2, 5.3 and 5.4, add this to

```
# modprobe cxgb3 && modprobe -f iw_cxgb3 && modprobe rdma_ucm
```

9. On SLES 10 SP2 and SLES 10 SP3 add this to

```
# modprobe cxgb3 && modprobe --force-modversion iw_cxgb3 &&
modprobe rdma_ucm
```

10. On SLES 11 add this to

```
# modprobe --allow cxgb3 && modprobe --allow --force iw_cxgb3 &&
 modprobe --allow rdma_ucm
```

Note: Installation of the cxgb3toe kit driver is required if user wants to use the Chelsio Adapter as a RDMA, TOE and iSCSI adapter at the same time.

3.2 Loading Firmware and EEPROM

3.2.1 Loading Firmware

If your OFED-X.Y.Z distro/kernel supports firmware loading, you can place the Chelsio firmware images in /lib/firmware/cxgb3, then unload and reload the cxgb3 module to get the new images loaded. If this does not work, then you can load the firmware image manually:

To obtain the cxgbtool tool download cxgb3toe-W.X.YY.ZZZ driver from http://service.chelsio.com/

To build cxgbtool:

```
# cd cxgb3toe-W.X.YY.ZZZ/tools/cxgbtool
# make && make install
```

Then load the cxgb3 driver:

```
# modprobe cxgb3
```

Now Note the ethernet interface name for the T3 device. This can be done by typing 'ifconfig -a' and noting the interface name for the interface with a HW address that begins with "00:07:43". Then load the new firmware as shown below:

cxgbtool <chelsio interface name> loadfw <firmware file name>

e.g.

cxgbtool eth2 loadfw t3fw-7.8.0

Note: Please Note that above command is used for both upgrading the firmware as well as downgrading the firmware, i.e. if Chelsio adapter has a higher version of firmware, user can downgrade it using above command giving lower version of firmware file name in cxgbtool command. User can upgrade the firmware in the same way by giving higher version of firmware file name in cxgbtool command. Many times downgrading of firmware may not work since from cxgb3toe-1.4.0.8 and above, the driver automatically flashes the firmware with which it requires to work with, e.g. cxgb3toe-1.4.0.8 driver flashes the firmware 7.8.0 automatically if firmware file t3fw-7.8.0.bin is present in the /lib/firmware directory. Please refer to **Table 1** for correct combination of driver and firmware version to work with.

3.2.2 Loading EEPROM

To obtain the update_eeprom.sh tool download cxgb3toe-W.X.YY.ZZZ driver from http://service.chelsio.com/

There are two types of EEPROM images for two different series of Chelsio adapters. Currently Chelsio adapters are categorized into T3B and T3C series. To load the EEPROM image identify the Chelsio adapters as shown below.

Read the value of register 0x6f4 using cxgbtool

cxgbtool eth0 reg 0x6f4

0x4 [4]

If it shows the output as 0x4 then your Chelsio adapter is of T3C series.

cxgbtool eth0 reg 0x6f4

0x3 [3]

If it shows the output as 0x3 then your Chelsio adapter is of T3B series.

Based on either T3B and T3C adapters download either t3b_tp_eeprom-1.1.0.bin.gz or t3c_tp_eeprom-1.1.0.bin.gz image respectively from the http://service.chelsio.com/

Execute the following command to load the EEPROM image.

```
# update_eeprom.sh <Chelsio interface name> <eeprom file name>
# reboot
```

e.g.

```
# update_eeprom.sh eth2 t3c_tp_eeprom-1.1.0.bin
```

4 Testing connectivity with ping and rping

Load cxgb3, iw_cxgb3 and rdma_ucm modules. After you load iw_cxgb3 and rdma_ucm, you will see one or two ethernet interfaces for the T3 device. Configure them with an appropriate ip address, netmask, etc. You can use the Linux ping command to test basic connectivity via the T3 interface.

To test RDMA, use the rping command that is included in the librdmacm-utils rpm:

On the server machine:

```
# rping -s -a server_ip_addr -p 9999
```

On the client machine:

```
# rping -c –Vv -C10 -a server_ip_addr -p 9999
```

You should see ping data like this on the client:

```
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
client DISCONNECT EVENT...
#
```

5 Enabling various MPIs

5.1 DAPL Library configuration for Intel MPI, HP MPI and Scali MPI

You must set the iw_cxgb3 module option peer2peer=1 on all systems. This can be done by writing to the /sys/module/ file system during boot.
e.g:
For RHEL 5 and SLES platforms use following commands:

```
# echo 1 > /sys/module/iw_cxgb3/parameters/peer2peer
```

For RHEL 4 platforms use following commands:

```
# echo 1 > /sys/module/iw_cxgb3/peer2peer
```

OR

You can add the following line to /etc/modprobe.conf to set the option at module load time:

```
options iw_cxgb3 peer2peer=1
```

To run Intel MPI, HP MPI, and Scali MPI over RDMA interface, DAPL 1.2 or 2.0 should be set up as follows:

Enable the Chelsio device by adding an entry at the beginning of the /etc/dat.conf file for the Chelsio interface. For instance, if your Chelsio interface name is eth2, then the following line adds a DAT version 1.2 and 2.0 devices named "chelsio1" and "chelsio2" for that interface:

```
chelsio1 u1.2 nonthreadsafe default libdaplcma.so.1 dapl.1.2 "eth2 0" ""
chelsio2 u2.0 nonthreadsafe default libdaplofa.so.2 dapl.2.0 "eth2 0" ""
```

5.2 Setting shell for Remote Logon

User needs to set up authentication on the user account on all systems in the cluster to allow user to remotely logon or executing commands without password.

Quick steps to set up user authentication:

- Change to user home directory

```
# cd
```

- Generate authentication key

```
# ssh-keygen -t rsa
```

- Press enter upon prompting to accept default setup and empty password phrase.

- Create authorization file

```
# cd .ssh
# cat *.pub > authorized_keys
# chmod 600 authorized_keys
- Copy directory .ssh to all systems in the cluster
# cd
# scp -r .ssh remotehostname-or-ipaddress:
```

5.3 Configuration of various MPIs

5.3.1 MVAPICH2

The following env vars need to be set for enable MVAPICH2-X.Y.Z version. Place these in your user env after installing and setting up MVAPICH2 MPI:

```
export MVAPICH2_HOME=<path to MVAPICH2-X.Y.Z home directory>
export MV2_USE_IWARP_MODE=1
export MV2_USE_RDMA_CM=1
```

e.g. To set MVAPICH2-1.4.1 for root user add following lines in /root/.bashrc file

```
export MVAPICH2_HOME=/usr/mpi/gcc/mvapich2-1.4.1/
export MV2_USE_IWARP_MODE=1
export MV2_USE_RDMA_CM=1
```

On each node, add this to the end of /etc/profile.

```
ulimit -l 999999
```

On each node, add this to the end of /etc/init.d/sshd and restart sshd.

```
ulimit -l 999999
% service sshd restart
```

Verify the ulimit changes worked. These should show '999999':

```
% ulimit -l
% ssh ulimit –l
```

Note: You may have to restart sshd a few times to get it to work.

In root's home directory, create .mpd.conf and .mpdpasswd with one line in them:

```
secretword=f00bar
```

Note: The secrete word can be anything.

---

chmod both files

```
% chmod 600 .mpd.conf
% chmod 600 .mpdpasswd
```

Create mpd.hosts with list of hostname or ipaddrs in the cluster. They should be names/addresses that you can ssh to without passwords. (See Passwordless SSH Setup).

```
% cp .mpd.conf mpd.conf
% cp .mpd.conf /etc/.mpd.conf
% cp .mpd.conf /etc/mpd.conf
% cp mpd.hosts .mpd.hosts
% cp mpd.hosts /etc/mpd.hosts
% cp mpd.hosts /etc/.mpd.hosts
```

On each node, create /etc/mv2.conf with a single line containing the IP address of the local T3 interface. This is how MVAPICH2 picks which interface to use for RDMA traffic.

On each node, edit /etc/hosts file. Comment the entry if there is an entry with 127.0.0.1 IP Address and local host name. Add an entry for corporate IP address and local host name (name that you have given in mpd.hosts file) in /etc/hosts file.

To run MVAPICH2 application:

```
mpirun_rsh -ssh -np <no. of processes> -hostfile mpd.hosts <MVAPICH2 application path>
/usr/mpi/gcc/mvapich2-1.4.1/tests/IMB-3.2/IMB-MPI1
```

e.g.

```
mpirun_rsh -ssh -np 8 -hostfile mpd.hosts /usr/mpi/gcc/mvapich2-1.4.1/tests/IMB-3.2/IMB-
```

MPI1

5.3.2 HPMPI

Installation and Setup

HP MPI application and license can be obtained from HP website. HP MPI is released in Linux "rpm" package and can be installed with command:

```
# rpm -ivh <hpmpi>.rpm
```

Note: There are 2 doc files in /opt/hpmpi/doc with detailed explanation regarding this topic once the rpm is installed.

By default HP-MPI shall be installed in directory "/opt/hpmpi". The application should be installed on all systems of a test cluster. The license file which is granted by HP should be named as "license.lic" or "your-prefer-name.lic" and copied to directory "/opt/hpmpi/licenses" of all systems.

Installing License File

To start license server issue command

```
# /opt/hpmpi/bin/licensing/lmgrd -c /opt/hpmpi/licenses/license.lic
```

Note: For lmgrd to work, uncomment the loopback address (127.0.0.1) from the /etc/hosts file only on the headnode.

The command should return without any error for a valid license file.

Setting up HP-MPI environment (Applicable to all systems in the cluster):

The following env vars enable HP MPI version 2.03.01.00. Place these in your user env after installing and setting up HP MPI:

```
export MPI_ROOT=/opt/hpmpi
export PATH=$MPI_ROOT/bin:/opt/bin:$PATH
export MANPATH=$MANPATH:$MPI_ROOT/share/man
```

Log out & log back in.

To run HP MPI applications, use these mpirun options:

-prot -e DAPL_MAX_INLINE=64 -UDAPL

e.g:

$ mpirun -prot -e DAPL_MAX_INLINE=64 -UDAPL -hostlist r1-iw,r2-iw
/opt/hpmpi/tests/presta-1.4.0/glob

Where r1-iw and r2-iw are hostnames mapping to the Chelsio interfaces.

Also this assumes your first entry in /etc/dat.conf is for the Chelsio device.

Contact HP for obtaining their MPI with DAPL support.

The performance is best with NIC MTU set to 9000 bytes

5.3.3 Intel -MPI

Installation and Setup

Download latest Intel MPI from the Intel website

Copy COM_L___CF8J-98P6MBWL.lic into l_mpi_p_x.y.z directory

Create machines.LINUX (list of node names) in l_mpi_p_x.y.z

Install software on every node.

#./install.sh

Register and set IntelMPI with mpi-selector (do this on all nodes).

#mpi-selector --register intelmpi --source-dir /opt/intel/impi/3.1/bin/
#mpi-selector --set intelmpi

Edit .bashrc and add these lines:

export RSH=ssh
export DAPL_MAX_INLINE=64
export I_MPI_DEVICE=rdssm:chelsio

```
export MPIEXEC_TIMEOUT=180
export MPI_BIT_MODE=64
```

Logout & log back in.

Populate mpd.hosts with node names.

Note: The hosts in this file should be Chelsio interface IP addresses.

Note: I_MPI_DEVICE=rdssm:chelsio assumes you have an entry in /etc/dat.conf named "chelsio".

Note: MPIEXEC_TIMEOUT value might be required to increase if heavy traffic is going across the systems.

Contact Intel for obtaining their MPI with DAPL support.

To run Intel MPI applications:

```
mpdboot -n -r ssh --ncpus=
mpiexec -ppn -n 2 /opt/intel/impi/3.1/tests/IMB-3.1/IMB-MPI1
```

The performance is best with NIC MTU set to 9000 bytes

5.3.4 Open-MPI

Installation and Setup

The latest release of Open-MPI-1.4.1 comes with OFED package only. Select Open-MPI package while installing OFED-1.5.1 package.

OpenMPI iWARP support is only available in OpenMPI version 1.3 or greater.

Open MPI will work without any specific configuration via the openib btl. Users wishing to performance tune the configurable options may wish to inspect the receive queue values. Those can be found in the "Chelsio T3" section of mca-btl-openib-hca-params.ini.

NOTE: OpenMPI version 1.3 does not support newer Chelsio card with device ID 0x0035 and 0x0036. To use those cards add the device id of the cards in the "Chelsio T3" section of mca-btl-openib-hca-params.ini file.

To run OpenMPI applications:

```
mpirun --host , -mca btl openib,sm,self /opt/ompi/openmpi-install/tests/IMB-3.1/IMB-MPI1
```

5.3.1 Scali MPI

Installation and Setup
Install the license key on the head node:

```
# ./lminstall -p
```

Install scampi on each node (including the head node):

```
# ./smcinstall -a -b -n
```

Create mpivars files:

```
#cp /opt/scali/etc/scampivars.sh /opt/scali/etc/mpivars.sh
#cp /opt/scali/etc/scampivars.csh /opt/scali/etc/mpivars.csh
```

Register and set scampi with mpi-selector

```
# mpi-selector --register scampi --source-dir /opt/scali/etc/
# mpi-selector --set scampi
```

-> Scali MPI Environment Variables:
The following env vars enable Scali MPI. Place these in your user env
after installing and setting up Scali MPI for running over IWARP:

```
export DAPL_MAX_INLINE=64
export SCAMPI_NETWORKS=chelsio
export SCAMPI_CHANNEL_ENTRY_COUNT="chelsio:128"
```

Log out & log back in.

Note: SCAMPI_NETWORKS=chelsio assumes you have an entry in /etc/dat.conf named "chelsio".

Note: SCAMPI supports only dapl 1.2 library not dapl 2.0

Contact Scali for obtaining their MPI with DAPL support.

To run SCALI MPI applications:

mpimon -networks chelsio -- /opt/scali/tests/IMB-3.1/IMB-MPI1 -- <node1_IP> <procs> <node2_IP> <procs>

Note: <procs> is the number of processes to run on the node

Note: <node1 IP> and< node2 IP> should be the IP of Chelsio's interface

---

6 Additional Notes and issues

1) To run uDAPL over the Chelsio device, you must export this environment variable:

```
export DAPL_MAX_INLINE=64
```

2) If you have a multi-homed host and the physical ethernet networks are bridged, or if you have multiple Chelsio rnics in the system, then you need to configure arp to only send replies on the interface with the target ip address:

```
sysctl -w net.ipv4.conf.all.arp_ignore=2
```

3) If you are building OFED against a kernel.org kernel later than 2.6.20, then make sure your kernel is configured with the cxgb3 and iw_cxgb3 modules enabled. This forces the kernel to pull in the genalloc allocator, which is required for the OFED iw_cxgb3 module. Make sure these config options are included in your .config file:

```
CONFIG_CHELSIO_T3=m
CONFIG_INFINIBAND_CXGB=m
```

4) If you run the RDMA latency test using the ib_rdma_lat program, make sure you use the following command lines to limit the amount of inline data to 64:

```
server: ib_rdma_lat -c -I 64
client: ib_rdma_lat -c -I 64 server_ip_addr
```