**OFA Interoperability Working Group**

**OFA-IWG Interoperability Test Plan**

**Release 1.45**



October 09, 2012

DRAFT

Copyright © 2012 by OpenFabrics - All rights reserved.

**This document contains information proprietary to OpenFabrics. Use or disclosure without written permission from an officer of the OpenFabrics is prohibited.**

[OpenFabrics.org](http://www.openfabrics.org/)

|  |  |  |
| --- | --- | --- |
| **Revision** | **Release Date** |  |
| 0.50 | Apr 4, 2006 | • First FrameMaker Draft of the Interop Test Plan which was used in the March 2006 IBTA-OpenFabrics Plugfest. |
| 0.51 | Apr 25, 2006 | • Added DAPL and updated MPI. |
| 0.511 | June 1, 2006 | • Arkady Added iWARP. |
| 0.52 | May 30, 2006 | • Added Intel MPI. |
| 0.53 | June 6, 2006 | • Updated uDAPL section provided by Arkady. |
| 0.54 | June 13, 2006 | • Updated entire Test Spec based on changes made by  Arkady to incorporate iWARP into the Test Spec. |
| 0.80 | June 14, 2006 | • Updated for the OFA conference in Paris and for BoD  meeting. Added OFA logo and URL. |
| 1.0 | June 21, 2006 | • Released after review and approval at the OFA confer- ence in Paris. |
| 1.01 | Aug 17, 2006 | • Updated the iWARP Equipment requirements in the Gen- eral System Setup section. |
| 1.02 | Oct 31, 2006 | • Updated Table 4 for iSER, Table 5 for SRP, Table 10 for uDAPL and corresponding info in Tables 17,18 and 22 as per request by Arkady. |
|  |  | • Added new test section from Bob Jaworski for Fibre  Channel Gateway. |
| 1.03 | Dec 10, 2006 | • Updated test procedures based on the October 2006 OFA Interop Event. |
|  |  | • Updated Fibre Channel Gateway test based on changes submitted by Karun Sharma (QLogic). |
|  |  | • Added Ethernet Gateway test written by Karun Sharma  (QLogic). |
| 1.04 | Mar 6, 2007 | • Updated test procedures in preparation for the April 2007  OFA Interop Event |
| 1.05 | Mar 7, 2007 | • Updated iWARP test procedures based on review by Mik- kel Hagen of UNH-IOL. Added missing results tables. |
| 1.06 | April 3, 2007 | • Updated for April 2007 Interop Event based on review from OFA IWG Meeting on 3/27/07. |
| 1.07 | April 3, 2007 | • Updated for April 2007 Interop Event based on review from OFA IWG Meeting on 4/3/07 |
| 1.08 | April 4, 2007 | • Added list of Mandatory Tests for April 2007 Interop  Event. |
| 1.09 | April 9, 2007 | • Updated Intel MPI based on review by Arlin Davis. |
| 1.10 | April 10, 2007 | • Updated after final review by Arlin Davis and after the  OFA IWG meeting on 4/10/2007 |

**Revision History** 1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

1

**Revision Release Date**

2

1.11 Sep 7, 2007 • Updated with the latest scripts developed by UNH IOL 3

and based on the results from the April 2007 Interop

Event 4

5

1.12 Sep 12, 2007 • Updated the documents to embed the test scripts in the

document. 6

7

1.13 Jan 22, 2008 • Updated the documents for the March 2008 OFA Interop

event. IPoIB updated along with Cover Page and the Test 8

Requirements section. 9

1.14 Feb 11, 2008 • Added the following tests: 10

• 1. Ethernet Switch Tests 11

• 2. IPoIB Connected Mode 12

• 3. RDMA Interop 13

• 4. RDS 14

15

1.15 Feb 18, 2008 • Updates to the following tests: 16

• 1. Ethernet Switch Tests 17

• 2. IPoIB Connected Mode 18

• 3. RDMA Interop 19

1.16 Feb 25, 2008 • Removed all reference to Low Latency Ethernet Switches. 20

This is the version for the March 2008 Interop Event 21

1.17 March 3,

2008

• Added HP-MPI 22

23

1.18 July 22, 2008 • Updated HP-MPI based on results from the March 2008 24

Interop Event 25

1.19 July 28, 2008 • Updated HP-MPI URL for the tests. 26

• Added section for Open MPI 27

• Updated MPI based on feedback from UNH IOL 28

1.20 July 30, 2008 • Updated section for Open MPI and added tables 29

• Updated IB SM Failover as per Nick Wood 30

1.21 Aug 1, 2008 • Updated SRP call srp\_daemon -o -e -n 31

• Updated IB SM Failover as Bob Jaworski 32

• Updated HP-MPI 33

• Updated Intel MPI 34

• Updated Open MPI 35

36

1.22 Aug 29, 2008 • Added a section for MVAPICH 1 under OSU MPI 37

1.23 Feb 16, 2009 • Updated Link Init, Fabric Init, SRP, SDP, IPoIB CM, IPoIB 38

DM based on updates received from UNH-IOL 39

40

41

42

1

**Revision Release Date**

2

1.24 Feb 23, 2009 • Updated Intel MPI and Open MPI to reflect the fact that 3

they are not intended to work in a heterogeneous environ-

ment. 4

• Updated the RDS test procedure 5

• Updated the Test Glossary 6

• Updated the Mandatory test table for April 2009 7

8

1.25 Feb 24, 2009 • Updated the RDS Test after review by the OFA IWG

group. 9

10

1.26 Mar 13, 2009 • Restructured entire document to accommodate WinOF

and OFED 11

• Added NFS over RDMA to the test plan. 12

• Added WinOF tests 13

• Updated HP-MPI 14

• Add List of Contributors 15

16

1.27 Mar 17, 2009 • Updates based on the review from the OFA IWG 17

1.28 Mar 27, 2009 • Added links in Chapter 10 to the InfiniBand Test Scripts 18

• Added links to HP-MPI installation Packages 19

1.29 Aug 25, 2009 • Editorial & Technical updates based on April 2009 Interop 20

Event. 21

• Updated Mandatory tests for October 2009. 22

• Added Topology Check 23

• Added new Firmware Policy 24

1.30 Sep 4, 2009 • Updated Mandatory iWARP tests and several comments 25

based on the review from Harry Cropper 26

• Added changes suggested by Jess Robel from QLogic to 27

|  |  |  |
| --- | --- | --- |
| IPoIB DM and CM and Fabric Init. | | 28 |
| 1.31 | April 6, 2010 • Added definition of homogenous to Test Glossary | 29 |
|  | • Added updates from the November 2009 Interop Event | 30 |
| 1.32 | April 20, 2010 • Updated after the OFA IWG meeting on 4/6/2010 | 31 |
|  | • Updated MPI and MVAPICH based on changes request- | 32 |
|  | ed by Jeff Laird and Intel | 33 |
| 1.33 | April 23, 2010 • Major changes to Section 8 which describes the Software | 34 |
|  | and Firmware polices | 35 |
| 1.34 | July 20, 2010 • Changed uDAPL for iWARP to Beta for Aug 2010 GA | 36 |
| Event  • Removed HP MPI which is no longer supported  • Added -mca mpi\_leave\_pinned 0 for OpenMPI  • Add new parameters for MVAPICH2 for iWARP devices. | | 37 |
| 38 |
| 39 |
| 40 |
| 41 |
| 42 |

1

**Revision Release Date**

2

1.35 July 27, 2010 • Added new parameters for MVAPICH2 for iWARP devic- 3

es. The parameter is: MV2\_USE\_RDMA\_CM=1 4

1.36 Feb 22, 2011 • Added Link Init section as per changes provided by Chris 5

Hutchins and approved by OFA IWG. 6

• Updated Test Plan Status for April 2011 and October 2011 7

• Nick Wood from UNH-IOL updated NFSoRDMA 8

• Marty requested that we update SRP Results Table 6 and

remove the disconnect commands. 9

10

1.37 Oct 4, 2011 • Updated Test Plan Status for November 2011 11

• Added new Test Table for OS and OFED versions 12

• Nick Wood updated Link Init for IB 13

• Chris Hutchins updated RDMA Interop and RDMA Stress 14

• Removed XANSation testing 15

1.38 Oct 11, 2011 • Changed Link Init Section from Recommendation to MOI 16

• Updated Section 8 for Firmware, Software and Hardware 17

Policies to bring in line with Logo Program Document

18

• Updated InfiniBand Test Table 24

19

1.39 Oct 24, 2011 • Updated Open MPI as per changes submitted by Nick 20

Wood

• Updated RDMA Interop small test: drop iterations from 21

100000 to 25000 22

• Updated RDMA Interop large test, increase iterations 23

from 100 to 300 24

• Updated IPoIB Part A:, drop iterations (number of pings) 25

from 100 to 10.

26

1.40 Oct 25, 2011 • Modified the following sections 27

• 12.6.9 iwarp client 100000 -> 25000 28

• 12.6.13 29

• olarge read client 65536 -> 1000000

30

• olarge write client 65536 -> 1000000

• Added large send command (section c) 31

32

33

34

35

36

37

38

39

40

41

42

1

**Revision Release Date**

2

1.41 Mar 20, 2012 • **General Instructions:** Added note that the OpenSM will 3

be used to run all mandatory tests in the test plan and the

Vendor SM testing will include testing IPoIB, RDMA In- 4

terop and Open MPI testing. 5

• **General Instructions:** The OFILG decided as of April 6

2012 that the various ULPs contained in this test plan will 7

only be tested if it is supported by the Operating System.

• Logo Program Requirements: updated IB and iWARP. 8

Made NFSoRDMA Mandatory and MVAPICH Optional. 9

• **IPoIB:** Modified the way IPoIB is set to connected or data- 10

gram mode 11

**• IPoIB:** Changed the ping interval in IPoIB tests from 0.01 12

to 0.2

• **IPoIB:** Reduced number of frame sizes tested in the Ping 13

Test. 14

**• MVAPICH:** Made testing of MVAPICH 1 & 2 Optional 15

• **NFSoRDMA**: Eliminate the need to specify nfs-utils in the 16

NFSoRDMA installation section 17

• **NFSoRDMA:** Changed the way the servers are mounted 18

in NFSoRDMA

• **SDP:** Eliminated the need for vsftpd in SDP 19

• **SDP:** Eliminated the environment variables section in 20

SDP 21

• **SDP:** Changed the way the netperf server is started in 22

SDP 23

• **SDP:** Made SDP mandatory only for those Operating

Systems that support it. 24

• **SRP**: Mandated that Targets only advertise two volumes 25

in order to reduce the amount of time required to run the 26

tests 27

1.42 Apr 3, 2012 • Updated Ethernet Test requirements to move NFSoRDMA 28

to Beta for April 2012 29

• Changed the status of Intel MPI and OSU MVAPICH to 30

deprecated meaning the tests are no longer being run or

supported. 31

• Updated SRP notes as per Marty Schlining 32

33

|  |  |  |
| --- | --- | --- |
| 1.43 | Aug 14, 2012 | • Updated the definition for $NP in MVAPICH section |
|  |  | 12.10.2, 2, ii  • Updated Mandatory test tables for iWARP and IB  • Cleared all change bars for October 2012 Interop event |

34

35

36

37

38

39

40

41

42

1

**Revision Release Date**

2

1.44 Sep 18, 2012 • Removed Intel MPI because it is not Open Source 3

• Removed SDP because no longer supported in OFED 4

• Removed Ethernet Fabric Initialize, Failover and recon- 5

vergence. No longer applicable given DCB etc.

• Removed TI RDS for iWARP because RDS does not sup- 6

port iWARP 7

• Remove iWARP Connectivity - replaced by RDMA Interop 8

test section 9

|  |  |  |
| --- | --- | --- |
| • Added section 8 for OS Installation and OS Policy | | 10 |
| 1.45 | Oct 9, 2012 • Add second test of SRP | 11 |
| • Add RoCE test sections | | 12 |
| 13 |
| 14 |
| 15 |
| 16 |
| 17 |
| 18 |
| 19 |
| 20 |
| 21 |
| 22 |
| 23 |
| 24 |
| 25 |
| 26 |
| 27 |
| 28 |
| 29 |
| 30 |
| 31 |
| 32 |
| 33 |
| 34 |
| 35 |
| 36 |
| 37 |
| 38 |
| 39 |
| 40 |
| 41 |
| 42 |

**List of Contributors** 1

2

**Editor**: Rupert Dance 3

4

**Name Company** 5

Mark Alan HP 6

Harry Cropper Intel 7

8

Rupert Dance Software Forge 9

Sujal Das Mellanox 10

Arlin Davis Intel 11

Johann George QLogic 12

13

Mike Hagen UNH-IOL 14

Mitko Haralanov QLogic 15

Allen Hubbe UNH-IOL 16

17

Christopher Hutchins UNH-IOL

18

Bob Jaworski QLogic 19

Arkady Kanevsky NetApp 20

Llolsten Kaonga Software Forge 21

22

Amit Krig Mellanox

23

Jeff Laird UNH-IOL 24

Jon Mason Open Grid Computing 25

Edward Mossman UNH-IOL 26

27

Bob Noseworthy UNH-IOL 28

Yaroslav Pekelis Mellanox 29

Jess Robel Qlogic 30

Hal Rosenstock HNR Consulting 31

32

Martin Schlining DataDirect Networks 33

Karun Sharma QLogic 34

Stan Smith Intel 35

36

Dave Sommers Intel (NetEffect)

37

Jeff Squyres Cisco 38

Dennis Tolstenko Lamprey Networks 39

Steve Wise Open Grid Computing 40

41

Robert Woodruff Intel

42

1

**Name Company**

2

Nick Wood UNH-IOL 3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**LEGAL DISCLAIMER "This version of a proposed OpenFabrics Interop Test** 1

**Plan is provided "AS IS" and without any warranty of** 2

**any kind, including, without limitation, any express or** 3

4

**implied warranty of non-infringement, merchant-** 5

**ability or fitness for a particular purpose.** 6

7

**In no event shall OpenFabrics, IBTA or any member of** 8

**these groups be liable for any direct, indirect, special,** 9 **exemplary, punitive, or consequential damages, in-** 10 **cluding, without limitation, lost profits, even if ad-** 11 **vised of the possibility of such damages."** 12

13

14

Conditional text tag *Explanation* is shown in green. 15

~~Conditional text tag~~ *~~Deleted~~* ~~is shown in red with strike through.~~ 16

17

Conditional text tag *Proposal* is shown in turquoise (r0\_g128\_b128). 18

19

Conditional text tag *Author* is shown as is. 20

21

Conditional text tag Comment is shown in red with underline 22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1 INTRODUCTION**

**1.1 PURPOSE**

**1.2 INTENDED AUDIENCE**

1

Server OEM customers have expressed the need for RDMA hardware and soft- 2

ware to interoperate. 3

4

Specifically, InfiniBand HCA, OpenFabrics host software to interoperate with In- finiBand Switches, gateways, and bridges with management software provided 5

by OEMs, and IB integrated server OEM vendors. And, iWARP RNIC and Open- 6

Fabrics host software to interoperate with Ethernet Switches and management 7 software and hardware provided by Ethernet Switch OEMs and iWARP inte- 8 grated server OEM vendors. 9

It is necessary that the interoperability test effort be an industry-wide effort where 10 interoperability testing is conducted under the auspices of the appropriate net- 11 working organizations. For InfiniBand it is the IBTA, specifically within the charter 12

of the CIWG and for iWARP it is the IETF. 13

14

This document is intended to describe the production tests step by step ex- 15

plaining each test and its references. The purpose of this test plan is three fold: 16

17

1) Define the scope, equipment and software needs, and test procedures for 18

verifying full interoperability of RDMA HW and SW. For Infiniband HW it is

InfiniBand HCAs using the latest OpenFabrics OFED software with currently 19

available OEM Switches and their management software. The target OEM 20

IB Switch vendors are Intel and Mellanox. For iWARP HW it is iWARP 21

RNICs using the latest OpenFabrics OFED software with currently available 22

OEM Ethernet Switches, Bridges, Gateways, Edge Devices and so on with

their management software. 23

24

2) Serve as a basis for evaluating customer acceptance criteria for OFA host software interoperability and OFA Logo. 25

3) Serve as a basis for extensions to InfiniBand IBTA CIWG test procedures re- 26

lated to interoperability and use of these test procedures in upcoming 27

PlugFest events organized by IBTA. 28

Serve as a basis for extensions to iWARP test procedures for OpenFabrics 29 software related to interoperability and use of these test procedures in up- 30 coming PlugFest events organized by the UNH IOL OFILG testing service. 31

32

The following are the intended audience for this document: 33

34

1) Project managers in OEM Switch, Router, Gateway, Bridge Vendor compa- 35 nies to understand the scope of testing and participate in the extension of 36 this test plan and procedures as necessary to meet their requirements. 37

2) IBTA and CIWG, and iWARP and UNH IOL iWARP testing personnel and 38

companies to evaluate the scope of testing and participate in the extension

of this test plan and procedures as necessary to meet their requirements. 39

3) Test engineering and project leads and managers who will conduct the 40

testing based on this document. 41

42

**1.3 TEST PLAN STRUCTURE**

4) Customers and users of OFA host software who rely on OFA Logo for in- 1

teroperability. 2

5) Integrators and OEM of RDMA products. 3

4

This test plan is divided into two main sections. 5

6

1) Interoperability testing using OFED for Linux. 7 a) See Sections 10-12 8

2) Interoperability testing using WinOFED for Windows Platforms. 9

10 a) See Section 13 11

12

Sections 1.4 through 1.10 provide an overview of the tests which are described 13

in detail in sections 10 through 13. 14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.4 INFINIBAND ONLY - TEST OVERVIEW** 1

The tables below list all of the specific test procedures for InfiniBand Devices. 2

See the Transport Independent section for tests that apply to all transports. 3

4

5

**Table 1 - IB Link Initialize** 6

7

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Phy link up all ports | Check that all relevant LEDs are on for all HCAs and switches. |

8

9

10

11

**Table 2 - IB Fabric Initialization** 12

13

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Fabric Initialization | Run SM from each node in cluster and see that all ports are in Armed or Active state. |

14

15

16

**Table 3 - IB IPoIB - Connect Mode (CM)** 17

18

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Ping all to all | Run SM from one of the nodes and check all nodes responding. Repeat with all SMs. |
| 2 | Connect disconnect host | Run SM from one of the nodes and check all nodes responding. |
| 3 | FTP Procedure | Using a 4MB test file, put the file, then get the file and finally compare the file. |

19

20

21

22

23

24

25

**Table 4 - IB IPoIB - Datagram Mode (DM)** 26

27

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Ping all to all | Run SM from one of the nodes and check all nodes responding. Repeat with all SMs. |
| 2 | Connect disconnect host | Run SM from one of the nodes and check all nodes responding. |
| 3 | FTP Procedure | Using a 4MB test file, put the file, then get the file and finally compare the file. |

28

29

30

31

32

33

34

35

36

**Table 5 - IB SM Tests** 37

38

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Basic sweep test | verify that all SMs are NOT ACTIVE (after receiving  the SMSet of SMInfo to DISABLE) and that the selected SM (SM1) is the master ( |

39

40

41

42

**Table 5 - IB SM Tests** 1

2

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 2 | SM Priority test | Verify Subnet and SMs behavior according to the SMs priority. |
| 3 | Failover - Disable SM1 | Disable the master SM and verify that standby SM becomes master and configures the cluster. |
| 4 | Failover - Disable SM2 | Disable the master SM and verify that standby SM becomes master and configures the cluster. |

3

4

5

6

7

8

9

**Table 6 - IB SRP Tests** 10

11

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Basic dd application | Run basic dd application from SRP host connected to target. |
| 2 | IB SM kill | Kill the IB master SM while test is running and check that it completes properly. |
| 3 | Disconnect Host | Unload SRP Host and check SRP connection properly disconnected. |
| 4 | Disconnect Target | Unload SRP Target and check SRP connection properly disconnected. |

12

13

14

15

16

17

18

19

**Table 7 - IB Ethernet Gateway** 20

21

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Basic Setup | Connect the HCA of the IB host and Ethernet Gateway to the IB fabric. Connect the Ethernet gateway to the Ethernet network or Ethernet device. Start the SM to be used in this test. |
| 2 | Start ULP | Determine which ULP your ethernet gateway uses and be sure that ULP is running on the host. |
| 3 | Discover Gateway | Restart the ULP or using the tool provided by the ULP, make sure that the host “discovers” the  Ethernet Gateway. |
| 4 | SM Failover | While the ping is running, kill the master SM. Verify that the ping data transfer is unaffected. |
| 5 | Ethernet gateway reboot | Reboot the Ethernet Gateway. After the Ethernet Gateway comes up, verify that the host can dis- cover the Ethernet Gateway as it did before and we are able to configure the interfaces. |
| 6 | ULP restart | Restart the ULP used by Ethernet Gateway and verify that after the ULP comes up, the host can discover the Ethernet Gateway and we are able to configure the interfaces. |
| 7 | Unload/load ULP | Unload the ULP used by Ethernet Gateway and check that the Ethernet Gateway shows it discon- nected. Load the ULP and verify that the Ethernet gateway shows the connection. |

22

23

24

25

26

27

28

29

30

31

32

33

34

35

**Table 8 - IB Fibre Channel Gateway** 36

37

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Basic Setup | Connect the HCA of the IB host to the IB fabric. Connect the FC Gateway to the IB Fabric. Con- nect the FC Gateway to the FC network or FC device. Start the SM to be used in this test. |
| 2 | Configure Gateway | Configure the FC Gateway appropriately (how to do this is vendor specific). |

38

39

40

41

42

**Table 8 - IB Fibre Channel Gateway** 1

2

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 3 | Add Storage Device | Use ibsrpdm tool in order to have the host "see" the FC storage device. Add the storage device as target. |
| 4 | Basic dd application | Run basic dd application from SRP host connected to target. |
| 5 | IB SM kill | Kill the IB master SM while test is running and check that it completes properly. |
| 6 | Disconnect Host/Target | Unload the SRP host / SRP Target (target first/host first) and check that the SRP connection is properly disconnected. |
| 7 | Load Host/Target | Load the SRP host / SRP Target. Using ibsrpdm, add the target. |
| 8 | dd after SRP Host and  Target reloaded | Run basic dd application from the SRP host to the FC storage device. |
| 9 | Reboot Gateway | Reboot the FC Gateway. After FC Gateway comes up, verify using ibsrpdm tool that the host see the FC storage device. Add the storage device as target. |
| 10 | dd after FC Gateway reboot | Verify basic dd works after rebooting Gateway. |

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.5 ETHERNET ONLY - TEST OVERVIEW** 1

The tables below list all of the specific test procedures for iWARP and Ethernet 2

Devices. See the Transport Independent section for tests that apply to all trans- 3

ports. 4

5

**Table 9 - iWARP Link Initialize** 6

7

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Phy link up all ports | Check that all relevant green LEDs are on for all RN ICs and switches. |
| 2 | Verify basic IP connectiv- ity | Verify IP and RDMA connectivity can occur by driving minimum size ICMP echo requests and replies across the link or equivalent traffic. |

8

9

10

11

12

13

14

**Table 10 - RoCE Link Initialize** 15

16

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Phy link up all ports | Check that all relevant green LEDs are on for all RCAs and switches. |
| 2 | Verify basic IP connectiv- ity | Verify IP and RDMA connectivity can occur by driving minimum size ICMP echo requests and replies across the link or equivalent traffic. |

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.6 TRANSPORT INDEPENDENT - TEST OVERVIEW** 1

The tables below list the test procedures that apply to devices regardless of the 2

transport. 3

4

5

**Table 11 - TI iSER** 6

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Basic dd application | Run basic dd application from iSER host connected to target. |
| 2 | IB SM kill | [IB Specific] - Kill the IB master SM while test is running and check that it completes properly. |
| 3 | Disconnect Initiator | Unload iSER Host and check iSER connection properly disconnected. |
| 4 | Disconnect Target | Unload iSER Target and check iSER connection properly disconnected. |
| 5 | Repeat with previous SM Slave | [IB Specific Test] Repeat steps 1-4 now with the previous slave SM (we did not actually stop the target). |

7

8

9

10

11

12

13

14

15

16

17

**Table 12 - TI NFS Over RDMA** 18

19

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | File and directory creation | A total of six files and six directories are created |
| 2 | File and directory removal | removes the directory tree that was just created by test1 |
| 3 | Lookups across mount point | changes directory to the test directory and gets the file status of the working directory |
| 4 | Setattr, getattr, and lookup | Permissions are changed (chmod) and the file status is retrieved (stat) for each file |
| 5 | Read and write | Creates a file (creat), Gets status of file (fstat) , Checks size of file, Writes 1048576 bytes into the file (write) in 8192 byte buffers, Closes file (close), Gets status of file (stat) , Checks the size of the file |
| 6 | Readdir | The program creates 200 files (creat). The current directory is opened (opendir), the begin- ning is found (rewinddir), and the directory is read (readdir) in a loop until the end is found |
| 7 | Link and rename | This program creates ten files. For each of these files, the file is renamed (rename) and file statistics are retrieved (stat) for both the new and old names |
| 8 | Symlink and readlink | This program makes 10 symlinks (symlink). It reads (readlink), and gets statistics for (lstat)  each, and then removes them (unlink). |
| 9 | Statfs | This program changes directory to the test directory (chdir and/or mkdir) and gets the file system status on the current directory (statfs). |

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

**Table 13 - TI RDS** 38

39

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | rds-ping procedure | Run rds-ping and verify that you can reach all hosts in the cluster |

40

41

42

**Table 13 - TI RDS** 1

2

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 2 | rds-stress procedure | Set up passive receiving instance and an active sender and verify data is exchanged without error |

3

4

5

6

7

8

**Table 14 - TI uDAPL** 9

10

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Point-to-Point Topology | Connection and simple send receive. |
| 2 | Point-to-Point Topology | Verification, polling and scatter gather list. |
| 3 | Switched Topology | Verification and private data. |
| 4 | Switched Topology | Add multiple endpoints, polling, and scatter gather list. |
| 5 | Switched Topology | Add RDMA Write. |
| 6 | Switched Topology | Add RDMA Read. |
| 7 | Multiple Switches | Multiple threads, RDMA Read, and RDMA Write. |
| 8 | Multiple Switches | Pipeline test with RDMA Write and scatter gather list. |
| 9 | Multiple Switches | Pipeline with RDMA Read. |
| 10 | Multiple Switches | Multiple switches. |

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

**Table 15 - RDMA Basic Interop** 27

28

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Small RDMA READ | Create an RDMA command sequence to send a READ operation of one byte. |
| 2 | Large RDMA READ | Create an RDMA command sequence to send a READ operation of 10,000,000 bytes |
| 3 | Small RDMA Write | Create an RDMA command sequence to send a Write operation of one byte |
| 4 | Large RDMA Write | Create an RDMA command sequence to send a Write operation of 10,000,000 bytes |
| 5 | Small RDMA SEND | Create an RDMA command sequence to send a SEND operation of one byte. |
| 6 | Large RDMA SEND | Create an RDMA command sequence to send a SEND operation of one million bytes |
| 7 | Small RDMA Verify | Create an RDMA command sequence to send a VERIFY operation of one byte. |
| 8 | Large RDMA Verify | Create an RDMA command sequence to send a VERIFY operation of 10,000,000 bytes |

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**Table 16 - RDMA Stress Tests** 1

2

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | Switch Load | For one pair of endpoints generate a stream of RDMA READ operation in one direction and RDMA write operations in the opposite direction. For all remaining endpoint pairs configure an RDMA WRITE operation of 1 byte and have it sent 10000 times on both streams of the endpoint pair. |
| 2 | Switch Fan In | Connect all possible endpoint pairs such that data exchanges between pairs must traverse the pair of ports interconnecting the switch |

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.7 OPEN MPI - TEST OVERVIEW** 1

2

3

**Table 17 - TI - Open MPI Test Suite Description** 4

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Open MPI TESTs*** | ***Open MPI TESTs Suite Description*** |
| **Phase 1: "Short" tests** | | |
| 1 | 2 | OMPI built with OpenFabrics support |
| 2 | 3 | OMPI basic functionality (hostname) |
| 3 | 4.1 | Simple MPI functionality (hello\_c) |
| 4 | 4.2 | Simple MPI functionality (ring\_c) |
| 5 | 5 | Point-to-point benchmark (NetPIPE) |
| 6 | 6.1.1 | Point-to-point benchmark (IMB PingPong multi) |
| 7 | 6.1.2 | Point-to-point benchmark (IMB PingPing multi) |
| **Phase 2: "Long" tests** | | |
| 8 | 6.2.1 | Point-to-point benchmark (IMB PingPong) |
| 9 | 6.2.2 | Point-to-point benchmark (IMB PingPing) |
| 10 | 6.2.3 | Point-to-point benchmark (IMB Sendrecv) |
| 11 | 6.2.4 | Point-to-point benchmark (IMB Exchange) |
| 12 | 6.2.5 | Collective benchmark (IMB Bcast) |
| 13 | 6.2.6 | Collective benchmark (IMB Allgather) |
| 14 | 6.2.7 | Collective benchmark (IMB Allgatherv) |
| 15 | 6.2.8 | Collective benchmark (IMB Alltoall) |
| 16 | 6.2.9 | Collective benchmark (IMB Reduce) |
| 17 | 6.2.10 | Collective benchmark (IMB Reduce\_scatter) |
| 18 | 6.2.11 | Collective benchmark (IMB Allreduce) |
| 19 | 6.2.12 | Collective benchmark (IMB Barrier) |
| 20 | 6.3.1 | I/O benchmark (IMB S\_Write\_Indv) |
| 21 | 6.3.2 | I/O benchmark (IMB S\_IWrite\_Indv) |
| 22 | 6.3.3 | I/O benchmark (IMB S\_Write\_Expl) |
| 23 | 6.3.4 | I/O benchmark (IMB S\_IWrite\_Expl) |
| 24 | 6.3.5 | I/O benchmark (IMB P\_Write\_Indv) |
| 25 | 6.3.6 | I/O benchmark (IMB P\_IWrite\_Indv) |
| 26 | 6.3.7 | I/O benchmark (IMB P\_Write\_Shared) |

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Open MPI TESTs*** | ***Open MPI TESTs Suite Description*** |
| 27 | 6.3.8 | I/O benchmark (IMB P\_IWrite\_Shared) |
| 28 | 6.3.9 | I/O benchmark (IMB P\_Write\_Priv) |
| 29 | 6.3.10 | I/O benchmark (IMB P\_IWrite\_Priv) |
| 30 | 6.3.11 | I/O benchmark (IMB P\_Write\_Expl) |
| 31 | 6.3.12 | I/O benchmark (IMB P\_IWrite\_Expl) |
| 32 | 6.3.13 | I/O benchmark (IMB C\_Write\_Indv) |
| 33 | 6.3.14 | I/O benchmark (IMB C\_IWrite\_Indv) |
| 34 | 6.3.15 | I/O benchmark (IMB C\_Write\_Shared) |
| 35 | 6.3.16 | I/O benchmark (IMB C\_IWrite\_Shared) |
| 36 | 6.3.17 | I/O benchmark (IMB C\_Write\_Expl) |
| 37 | 6.3.18 | I/O benchmark (IMB C\_IWrite\_Expl) |
| 38 | 6.3.19 | I/O benchmark (IMB S\_Read\_Indv) |
| 39 | 6.3.20 | I/O benchmark (IMB S\_IRead\_Indv) |
| 40 | 6.3.21 | I/O benchmark (IMB S\_Read\_Expl) |
| 41 | 6.3.22 | I/O benchmark (IMB S\_IRead\_Expl) |
| 42 | 6.3.23 | I/O benchmark (IMB P\_Read\_Indv) |
| 43 | 6.3.24 | I/O benchmark (IMB P\_IRead\_Indv) |
| 44 | 6.3.25 | I/O benchmark (IMB P\_Read\_Shared) |
| 45 | 6.3.26 | I/O benchmark (IMB P\_IRead\_Shared) |
| 46 | 6.3.27 | I/O benchmark (IMB P\_Read\_Priv) |
| 47 | 6.3.28 | I/O benchmark (IMB P\_IRead\_Priv) |
| 48 | 6.3.29 | I/O benchmark (IMB P\_Read\_Expl) |
| 49 | 6.3.30 | I/O benchmark (IMB P\_IRead\_Expl) |
| 50 | 6.3.31 | I/O benchmark (IMB C\_Read\_Indv) |
| 51 | 6.3.32 | I/O benchmark (IMB C\_IRead\_Indv) |
| 52 | 6.3.33 | I/O benchmark (IMB C\_Read\_Shared) |
| 53 | 6.3.34 | I/O benchmark (IMB C\_IRead\_Shared) |
| 54 | 6.3.35 | I/O benchmark (IMB C\_Read\_Expl) |
| 55 | 6.3.36 | I/O benchmark (IMB C\_IRead\_Expl) |
| 56 | 6.3.37 | I/O benchmark (IMB Open\_Close) |

**Table 17 - TI - Open MPI Test Suite Description** 1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.8 OSU MPI - TEST OVERVIEW**

1

2

3

**Table 18 - TI - OSU MPI** 4

|  |  |  |
| --- | --- | --- |
| **Test #** | ***Test*** | ***Description*** |
| 1 | **Test** 1: PingPong |  |
| 2 | **Test** 1: PingPing point-to-point |  |
| 3 | **Test** 2: PingPong |  |
| 4 | **Test** 2: PingPing |  |
| 5 | **Test** 2: Sendrecv |  |
| 6 | **Test** 2: Exchange |  |
| 7 | **Test** 2: Bcast |  |
| 8 | **Test** 2: Allgather |  |
| 9 | **Test** 2: Allgatherv |  |
| 10 | **Test** 2: Alltoall |  |
| 11 | **Test** 2: Alltoallv |  |
| 12 | **Test** 2: Reduce |  |
| 13 | **Test** 2: Reduce\_scatter |  |
| 14 | **Test** 2: Allreduce |  |
| 15 | **Test** 2: Barrier |  |

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

[**1.9 REQUIREMENTS FOR OFA INTEROPERABILITY LOGO PROGRAM**](http://www.iol.unh.edu/services/testing/ofa/index.php) 1

The following table indicates the mandatory tests that will be used for Interop Val- 2 idation during the October 2012 Interop Debug Event and the Interop GA Event 3 using OFED 3.5 GA. Deprecated means that the test is no longer being actively 4

run during the OFA Interop Events. 5

**Table 19 - InfiniBand Transport Test Status for October 2012 Interop Event** 6

7

|  |  |  |
| --- | --- | --- |
| **Test Procedure** | **Linux** | **WinOF** |
| IB Link Initialize | **Mandatory** | **Mandatory** |
| IB Fabric Initialization | **Mandatory** | **Mandatory** |
| IB IPoIB Connected Mode | **Mandatory** | Not Available -1 |
| IB IPoIB Datagram Mode | **Mandatory** | **Beta** |
| IB SM Failover/Handover - OpenSM | **Mandatory** | **Beta** |
| IB SM Failover/Handover - Vendor SM | **Optional** | **Optional** |
| IB SRP | **Mandatory** | **Beta** |
| IB Ethernet Gateway | **Beta** | Not Available - 3 |
| IB Fibre Channel Gateway | **Beta** | Not Available - 3 |
| TI iSER | **Mandatory** | **Beta** |
| TI NFS over RDMA | **Mandatory** | Not Available - 1 |
| TI RDS | **Mandatory** | Not Available - 2 |
| TI uDAPL | **Mandatory** | **Beta** |
| TI Basic RDMA Interop | **Mandatory** | Not Available - 3 |
| TI RDMA Stress | **Mandatory** | Not Available - 3 |
| TI MPI Open MPI | **Mandatory** | Not Available - 2 |
| TI MVAPICH - OSU | **Deprecated** | Not Available - 2 |

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

Not Available means one of three things: 32

33

1) The feature is not currently supported by the WinOFED stack. 34

2) The ULP application has not been ported to the WinOFED Stack. 35

3) The test has not been updated for WinOFED. 36

**Optional** means that this test will not be made mandatory because it depends on 37

proprietary vendor capabilities. The test may be run during the OFA Interop 38

Events and reported in the results but it will not affect eligibility for the OFA Logo 39

List. 40

41

42

1

2

**Table 20 - iWARP Transport Test Status for October 2012 - OFED 3.5** 3

4

|  |  |
| --- | --- |
| **Test Procedure** | **Linux** |
| iWARP Link Initialize | **Mandatory** |
| TI iSER | **Beta** |
| TI NFS over RDMA | **Beta** |
| TI uDAPL | **Mandatory** |
| TI Basic RDMA Interop | **Mandatory** |
| TI RDMA Stress | **Mandatory** |
| TI MPI Open MPI | **Mandatory** |
| TI MVAPICH2 - OSU | **Deprecated** |

5

6

7

8

9

10

11

12

13

14

15

16

17

18

**Table 21 - RoCE Transport Test Status for October 2012 - OFED 3.5** 19

|  |  |
| --- | --- |
| **Test Procedure** | **Linux** |
| RoCE Link Initialize | **Beta** |
| RoCE Fabric Init | **TBD** |
| RoCE IPoCE | **TBD** |
| RoCE InfiniBand Gateway | **TBD** |
| RoCE Fibre Channel Gateway | **TBD** |
| TI iSER | **Beta** |
| TI NFS over RDMA | **Beta** |
| TI uDAPL | **Beta** |
| TI Basic RDMA Interop | **Beta** |
| TI RDMA Stress | **Beta** |
| TI MPI Open MPI | **Beta** |

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.10 SUBJECTS NOT COVERED**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | ***Subject/ Feature*** | ***Reason*** | ***Executor*** | ***Due Date*** |
| 1 | iWARP peer to peer | Future Testing |  | TBD |
| 2 | IPv6 testing | Future Testing |  | TBD |

**1.11 TEST GLOSSARY**

1

**Table 22 - SUBJECTS NOT COVERED** 2

3

4

5

6

7

8

9

**Table 23 - Test Glossary** 10

11

|  |  |
| --- | --- |
| **Technical Terms** |  |
| DCB | Data Center Bridging (used in RoCE) |
| HCA | IB Host Channel Adapter |
| IPoIB | IP over InfiniBand |
| iSER | iSCSI Extensions for RDMA |
| MPI | Message Passing Interface |
| RCA | RoCE Channel Adapter |
| RDF | Readme File |
| RDS | Reliable Datagram Sockets |
| RNIC | RDMA NIC (iWARP Network Interface Card) |
| RoCE | RDMA over Converged Ethernet |
| SA | IB Subnet Administration |
| SDN | Software Defined Network |
| SDP | Sockets Direct Protocol |
| SM | IB Subnet Manager |
| SPB | Shortest Path Bridging (used in RoCE) |
| SRP | SCSI RDMA Protocol |
| TD | Test Descriptions |
| TI | Transport Independent (tests) |
| TRILL | Transparent Interconnect of Lots of Links is a IETF Standard implemented by devices called  RBridges (Routing Bridges) or TRILL Switches (used in RoCE) |
| uDAPL | User Direct Access Programming Library |

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**1.12 HOMOGENOUS VERSUS HETEROGENEOUS** 1

Heterogeneous & homogeneous clusters are the same with one exception: the 2 end points must be from the same vendor in homogeneous clusters. The table 3 below defines the guidelines for building homogeneous and heterogeneous clus- 4

ters 5

6

7

8

|  |  |  |
| --- | --- | --- |
| **Description** | **Homogenous** | **Heterogeneous** |
| Mixing switches (both models and vendor products) | **Encouraged** | **Encouraged** |
| The use of any InfiniBand subnet manager | **Encouraged** | **Encouraged** |
| All devices of the same model number shall use the same firmware. | **Mandatory** | **Mandatory** |
| Any mix of products from the same vendor is acceptable - e.g. differ- ent model HCAs | **Encouraged** | **Encouraged** |
| A mix of end points (HCA/RNIC) from different OFA vendors | **Prohibited** | **Mandatory** |
| Mixing x86-32 (ix86) and x86\_64 Operating System - see notes | **Not-Tested** | **Not-Tested** |
| 32 bit architecture and 32 bit OS - see notes | **Not-Tested** | **Not-Tested** |
| Mixing x86-32 and x86-64 user-level application | **Optional** | **Optional** |
| Mixed system architecture - e.g. x86 servers mixed with IA-64 (Ita- nium) servers | **Prohibited** | **Prohibited** |
| Mixing endianness in system OS | **Prohibited** | **Prohibited** |
| Mixing the quantity of server RAM installed on the hosts | **Encouraged** | **Encouraged** |
| Mixing the server clock speeds | **Encouraged** | **Encouraged** |
| Mixing the number of server cores | **Encouraged** | **Encouraged** |
| Mixing PCIe generations | **Encouraged** | **Encouraged** |
| All servers shall run the same OFED version. | **Encouraged** | **Encouraged** |
| Mixing supported Operating Systems | **Encouraged** | **Encouraged** |

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

**Notes:** Intel drivers do not support 32 bit operating systems 36

37

38

39

40

41

42

**2 USE OF OPENFABRICS SOFTWARE FOR PRE-TESTING** 1

Depending on the schedule of testing and bugs or issues encountered, different 2 snapshots of latest OpenFabrics software will be used during pre-testing prior to

the Interoperability Event. Any changes that result in the OpenFabrics software 3 from interoperability testing per this test plan will be deposited back into the

OpenFabrics repository so that the OpenFabrics development community will 4 have full access to any bug fixes or feature additions that may result out of this

testing effort. The frequency of such deposits will be determined based on com- 5 pletion of adequate testing of the said fixes or feature additions.

6

**3 USE OF OPENFABRICS SOFTWARE FOR IBTA/CIWG COMPLIANCE PLUGFESTS**

During the pre-testing phase, UNH-IOL will apply all reasonable effort to ensure 7 that the OpenFabrics source and binary repositories are up-to-date with the lat-

est OFED release. This will enable cable interoperability testing at plugfests to 8

be conducted using software directly sourced from the OpenFabrics tree.

9

Should there be any issues with the OpenFabrics community not accepting cer-

tain bug fixes or features with the time frames matching with Compliance 10

Events, UNH-IOL will inform all participants about the same and offer those bug

fixes or features in source code and binary formats directly to the participants 11 and InfiniBand solution suppliers.

12

**4 USE OF OPENFABRICS SOFTWARE FOR OFA IWG INTEROPERABILITY EVENTS** 13

During the pre-testing phase, UNH-IOL will apply all reasonable effort to ensure 14 that the OpenFabrics source and binary repositories are up-to-date with the lat-

est OFED releases chosen by the OFA IWG for use in the Interoperability Event. 15

Should there be any issues with the OpenFabrics community not accepting cer- 16 tain bug fixes or features with the time frames matching with Interoperability

Events, UNH-IOL will inform all participants about the same and offer those bug 17 fixes or features in source code and binary formats directly to the participants

and InfiniBand solution suppliers. 18

19

20

21

22

23

24

25

26

**5 GENERAL SYSTEM SETUP**

**Configuration**

**5.1 IB HW UNITS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Equipment** | ***Amount*** | ***Details*** | ***Check*** |
| Servers with OS installed | 12 or more | The OS should be supported by OpenFabrics Software. |  |
| 4X IB Cables | 30 or more | Between 1 meter => 10 meters. |  |
| IB Switches | 4 | The number and types of switches needed from member com- panies or OEMs is dependent on variations in subnet manage- ment and other IBTA defined management software. For example if the software on Switch A is different from the software used in Switch B, both Switches will be needed. Note that it is not dependent on number of ports supported by a switch. |  |
| IB HCAs | 12 or more |  |  |

**5.2 IB SOFTWARE**

1

2

3

The test environment for the user interface contains: 4

5

6

**Table 24 - IB Equipment** 7

8

9

10

11

12

13

14

15

16

17

18

19

20

**5.2.1 LINUX/WINDOWS PLATFORMS** 21

**5.2.2 OFED - MOST CURRENT TESTED RELEASE** 22

**5.2.3 IB HCA FW – VERSION XXX - VENDOR SPECIFIC** 23

24

**5.2.4 IB SWITCH FW CANDIDATE – VERSION XXX - VENDOR SPECIFIC** 25

**5.2.5 IB SWITCH SW – VERSION XXX - VENDOR SPECIFIC** 26

27

**5.3 IWARP HW UNITS** 28

29

**Table 25 - iWARP Equipment** 30

31

|  |  |  |  |
| --- | --- | --- | --- |
| **Equipment** | ***Amount*** | ***Details*** | ***Check*** |
| Servers with OS installed | 5 or more | The OS should be supported by OpenFabrics Software. |  |
| 4X CX4 or SFP Cables | 10 or more | Between 1 meter => 10 meters. |  |
| 10 GbE Switches | 1 | At least one 10 GbE switch must be made available to support the various RNICs in the Fabric.. There is no need to have multiple switches if there are enough ports on the primary switches to support all the devices in the fabric. |  |
| iWARP RNIC | 5 or more | Each vendor must supply 5 or more RNICs in order to support  MPI testing. |  |

32

33

34

35

36

37

38

39

40

**5.4 IWARP SOFTWARE** 1

**5.4.1 LINUX PLATFORMS** 2

**5.4.2 OFED - MOST CURRENT TESTED RELEASE** 3

**5.4.3 IWARP RNIC FW – VERSION XXX - VENDOR SPECIFIC** 4

5

**5.4.4 10GBE SWITCH FW CANDIDATE – VERSION XXX - VENDOR SPECIFIC** 6

**5.4.5 10GBE SWITCH SW – VERSION XXX - VENDOR SPECIFIC** 7

**5.4.6 VENDOR SPECIFIC NOTES** 8

**Note**: Currently there is no interoperability between cxgb4 and nes if peer2peer is enabled. 9

Both nes and cxgb4 have their own proprietary ways of doing "client must send the first 10 fpdu". The Chelsio parameter file /sys/module/iw\_cxgb4/parameters/peer2peer should be 11 modified on all hosts to contain the appropriate value for each test. For example: the value 12 must be set to ’1’ for the uDAPL test. 13

14

Arlin Davis suggests the following given the current situation: 15

1)The dapltest -T P (performance tests) will always send data from server side first. This test will

NOT work reliably with iWARP vendors. 16

2)The dapltest -T T (transaction tests) should work fine with both IB and iWARP vendors given 17

that it always sends from client side first. 18

3)I recommend using only dapltest transaction mode (-T T) in your test plan and removing -T P 19

mode tests. 20

**5.5 ROCE HW UNITS** 21

22

**Table 26 - RoCE Equipment** 23

24

|  |  |  |  |
| --- | --- | --- | --- |
| **Equipment** | ***Amount*** | ***Details*** | ***Check*** |
| Servers with OS installed | 5 or more | The OS should be supported by OpenFabrics Software. |  |
| 4X QSFP+ Cables | 10 or more | Between 1 meter => 10 meters. |  |
| GbE DCB Switches | 1 | At least one 10 or 40 GbE DCB switch must be made avail- able to support the various RCAs in the Fabric. There is no need to have multiple switches if there are enough ports on the primary switches to support all the devices in the fabric. |  |
| RoCE RCA | 5 or more | Each vendor must supply 5 or more RCAs in order to support  MPI testing. |  |

25

26

27

28

29

30

31

32

33

**5.6 ROCE SOFTWARE** 34

35

**5.6.1 LINUX PLATFORMS** 36

**5.6.2 OFED - MOST CURRENT TESTED RELEASE** 37

**5.6.3 ROCE FW – VERSION XXX - VENDOR SPECIFIC** 38

**5.6.4 10/40 GBE DCB SWITCH FW CANDIDATE – VERSION XXX - VENDOR SPECIFIC** 39

**5.6.5 10/40 GBE DCB SWITCH SW – VERSION XXX - VENDOR SPECIFIC** 40

**5.7 MPI TESTING**

1

1)HCA/RCA/RNIC vendors must provide a minimum of five adapters. The adapters need 2 not be all the same model, but they can be. 3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**6 IB HW DESCRIPTION & CONNECTIVITY** 1

The test contains two major parts. This description is for each of those parts. 2

3

**6.1 BASIC CONNECTIVITY (P1P1)** 4

**6.1.1 HCA 1 SHOULD BE CONNECTED FROM PORT 1 TO LOWEST PORT NUMBER IN SWITCH** 5

**6.1.2 HCA 2 SHOULD BE CONNECTED FROM PORT 1 TO HIGHEST PORT NUMBER IN SWITCH** 6

**6.1.3 BOTH WITH COMPLIANT INFINIBAND CABLES** 7

**6.2 SWITCHES AND SOFTWARE NEEDED** 8

9

**6.2.1 SWITCHES PROVIDED BY OEMS** 10

It is necessary that Switches provided by OEMs cover the full breadth of software 11

versions supported by the Switch OEMs. Port count is not critical for the tests. It

is recommended that OEMs provide six switches covering all variations of soft- 12

ware supported on the Switches. 13

14

**6.2.2 OPENFABRICS SOFTWARE RUNNING ON HOSTS** 15

Where there are dependencies of OEM provided and IBTA defined management 16 software (such as subnet managers and agents, performance managers and 17 agents etc.) with OpenFabrics software running on Hosts, such software should

be provided to UNH-IOL for interoperability testing. Any known dependencies 18

should be communicated to UNH-IOL. 19

20

**6.3 CLUSTER CONNECTIVITY** 21

**6.3.1 HOSTS AND TARGETS 1-6 SHOULD BE CONNECTED FROM PORT 1 OR 2 TO PORTS X IN ALL SWITCHES** 22

**USING COMPLIANT INFINIBAND CABLES.** 23

**Figure 1 - Template for IB Interop Setup** 24

25

**Host or**

**Target 1**

**1**

**Host or**

**Target 2**

**1**

**1**

**Host or**

**Target 3**

**2**

**1**

**Host or**

**Target 4**

**1**

**Host or**

**Target 5**

**2**

**Host or** 26

**Target 6** 27

28

29

30

31

**1** 32

33

34

**Switch 1**

|  |  |
| --- | --- |
|  | 36 |
| 37 |
| 38 |
| 39 |
| 40 |
| 41 |
|  | 42 |

**Switch 2**

**Switch 3**

**Switch 4** 35

**Switch 5**

**7 IWARP HW DESCRIPTION & CONNECTIVITY** 1

2

3

**7.1 IWARP BASIC CONNECTIVITY (P1P1)** 4

**7.1.1 RNIC 1 ON ONE HOST SHOULD BE DIRECTLY CONNECTED TO RNIC 2 ON ANOTHER HOST OR TO A** 5

**10GBE SWITCH.** 6

**7.1.2 WITH 10GBE CABLES** 7

**7.2 SWITCHES AND SOFTWARE NEEDED** 8

**7.2.1 SWITCHES PROVIDED BY OEMS** 9

It is necessary that Switches provided by OEMs cover the full breadth of software 10 versions supported by the Switch OEMs. Port count is not critical for the tests. It 11 is recommended that OEMs provide a switch per variations of software supported 12

on the Switch. 13

**7.2.2 OPENFABRICS SOFTWARE RUNNING ON RNICS** 14

15

Where there are dependencies of OEM provided with OpenFabrics software run-

ning on RNICs, such software should be provided to UNH-IOL for interoperability 16

testing, and any known dependencies should be communicated to UNH-IOL. 17

18

**7.3 CLUSTER CONNECTIVITY** 19

**7.3.1 HOSTS AND TARGETS 1-6 SHOULD BE CONNECTED TO SWITCHES USING 10GBE CABLES.** 20

21

**Figure 2 Template for iWARP Interop Setup** 22

23

**Host or**

**Target 1**

**1**

**Host or**

**Target 2**

**1**

**1**

**Host or**

**Target 3**

**2**

**1**

**Host or**

**Target 4**

**1**

**Host or**

**Target 5**

**2**

**Host or** 24

**Target 6** 25

26

27

28

29

**1** 30

31

32

**Switch 1**

**Switch 2**

**Switch 5**

**Switch 3**

**Switch 4** 33

34

35

36

37

38

39

40

41

42

1

2

**7.4 GATEWAY, BRIDGES, ROUTERS CONNECTIVITY** 3

**TBD** 4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**8 ROCE HW DESCRIPTION & CONNECTIVITY** 1

**8.1 ROCE BASIC CONNECTIVITY (P1P1)** 2

**8.1.1 RCA 1 ON ONE HOST SHOULD BE DIRECTLY CONNECTED TO RCA 2 ON ANOTHER HOST OR TO A** 3

**10/40 GBE SWITCH DCB ENABLED.** 4

**8.1.2 CONNECTED WITH 10/40 GBE CABLES** 5

6

**8.2 SWITCHES AND SOFTWARE NEEDED** 7

**8.2.1 SWITCHES PROVIDED BY OEMS** 8

RoCE testing is being introduced as of October 2012 and the choice of Ethernet 9

Fabrics such as Fabric Path, QFabric, MLAG, SPB, TRILL and others are initially 10

not being addressed. This allows us to start Beta Testing RoCE with just one

10/40 GbE Ethernet Switch which is DCB enabled. In future Interop events we 11

will consider using multiple switches from vendors such as Brocade, Cisco, Ex- 12

treme, HP, Mellanox and others which will allow us to test various Ethernet Fabric 13

solutions. 14

**8.2.2 OPENFABRICS SOFTWARE RUNNING ON RCAS** 15

Where there are dependencies of OEM provided with OpenFabrics software run- 16

ning on RCAs, such software should be provided to UNH-IOL for interoperability 17

testing, and any known dependencies should be communicated to UNH-IOL. 18

19

**8.2.3 ROCE PRIORITY LEVELS** 20

Ethernet provides a construct, called a Priority Level which corresponds concep- 21 tually to InfiniBand’s SLs. Eight priorities, numbered zero through seven are sup- ported. As in InfiniBand, a verbs consumer accessing a RoCE port specifies its 22

desired service level, which is then mapped to a given Ethernet Priority. The de- 23

fault mapping is as follows: 24

25

• SL 0-7 are mapped directly to Priorities 0-7 respectively

• SL 8-15 are reserved. 26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**8.3 FABRIC CONNECTIVITY** 1

**8.3.1 HOSTS AND TARGETS 1-6 SHOULD BE CONNECTED TO SWITCHES USING 10/40 GBE CABLES.** 2

3

**Figure 3 Template for RoCE Interop Setup** 4

5

6

**Host or**

**Target 1**

**Host or**

**Target 2**

**1**

**Host or**

**Target 3**

**2**

**3**

**Host or**

**Target 4**

**5**

**4**

**Host or**

**Target 5**

**6**

**Host or** 7

**Target 6** 8

9

10

11

12

13

14

15

16

17

18

19

20

21

**10/40 GbE** 22

**DCB Switch**

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**9 FW & SW INSTALLATION**

**9.1 BURNING THE FW**

**9.1.1 FIRMWARE POLICY**

1

2

3

4

**Firmware Policy during the Interop Debug Event**

The firmware used during the Interop Debug Event is at the discretion of the de- 5

vice vendor. Vendors will be allowed to make changes to the firmware during the 6

Interop Debug Event. However changes should be made as early in the event pe- 7 riod as possible to reduce the amount of retesting which will result from these 8 changes.

9

**Firmware Policy during the Interop GA Event** 10

The firmware image used during the Interop GA Event must be provided to the 11

UNH-IOL at least one week prior to the event. No firmware changes of any kind 12 are allowed during the Interop GA Event. If the vendor does not provide updated firmware by the deadline, then the UNH-IOL will use the firmware from the Interop 13

Debug Event or from the vendor's website, whichever is more current. 14

15

**Firmware Policy after the Interop GA Event** 16

The firmware used to obtain the OFA Logo (or a child of this firmware with the

same base functionality) must be the default publicly available firmware on the 17

vendor's website and must be the default firmware that is shipped with the 18

product. This must be completed within six months of the Interop GA Event. 19

20

**9.1.2 PLEASE REFER TO FIRMWARE BURNING TOOLS AND PROCEDURES DOCUMENTATION FROM HCA IB VENDOR** 21

**9.2 OPERATING SYSTEM INSTALLATION** 22

**9.2.1 OPERATING SYSTEM POLICY** 23

The OS used during an Interop Debug Event will be determined by the OFA IWG 24 and will be none as the primary OS. All available updates will be installed prior to 25 the start of the Interop Debug Event and frozen in place for the duration of the

Interop Debug Event. 26

27

The OS used during an Interop GA Event will be the same agreed-upon version 28 of the primary OS tested during the Interop Debug Event. The updates applied at 29 the start of the Interop Debug Event will remain frozen in place for the duration of

the Interop GA Event. 30

31

In addition to the mandatory testing performed using the primary OS, beta testing 32 using the secondary operating systems is performed after completion of manda- 33 tory testing. The secondary operating systems are deployed in a similar manner

to the primary OS, in that updates are applied at the beginning of the Interop 34

Debug Event and frozen in place for the duration of the Interop GA Event. 35

36

**9.2.2 OPERATING SYSTEM INSTALLATION** 37

Install the primary OS on all hosts in the cluster. Use a package manager to up- 38

date all installed packages to their latest versions available as of the start of the 39

Interop Debug Event.

40

41

42

**9.3 SW INSTALLATION**

**9.3.1 SOFTWARE POLICY**

Install the secondary operating systems on all hosts in the cluster. Use a package 1 manager to update all installed packages to their latest versions available as of 2 the start of the Interop Debug Event. Install and test as many secondary oper-

ating systems as time permits. 3

4

5

6

**Software Policy during an Interop Debug Event** 7

The software used during an Interop Debug Event will be an agreed-upon RC re- 8

lease of the subsequent OFED version. During the Interop Debug Event vendors 9 will be allowed to make changes to the software, provided that the changes are 10 based on the same RC release. Vendors are not allowed to extensively modify

the software or completely replace it. 11

12

**Software Policy during the Interop GA event** 13

The software used during an Interop GA Event will be the GA release of the same 14

OFED version as was used during the Interop Debug Event. No software

changes of any kind are allowed during the Interop GA Event. It is the vendor's 15

responsibility to ensure that any changes made during the Interop Debug Event 16

are present in the OFED GA release. Vendors whose products do not use firm- 17 ware may request that patches be applied to an OFED GA release if that release 18 has known defects that prevent the vendor product from being interoperable. The

Arbitration Committee will be responsible for approving the requested patches. 19

20

**Software Policy after the Interop GA event** 21

All products that are granted the OFA Logo must be distributed by default with the 22

OFED GA version (or a later revision of OFED with the same base functionality).

23

**9.3.2 PLEASE REFER TO SOFTWARE INSTALLATION MANUAL FROM HCA IB VENDOR.** 24

**9.3.3 PLEASE REFER TO SOFTWARE INSTALLATION MANUAL FROM RNIC VENDOR.** 25

**9.4 SUMMARY** 26

27

• For the Interop GA Event the vendor cannot update or change any part of

the device under test - this includes hardware, firmware and software. The 28

only exception is for an outright hardware failure in which case the hardware 29

may be replaced with an identical piece of hardware with the same SW and 30

FW. 31

• If an end user requests customized firmware or a modified version of OFED, 32

then the vendor must disclose that this is not an OFA certified configuration. 33

• The OFA reserves the right to revoke the OFA Logo for products that do not 34

follow these policies.

35

• These policies will be in effect for the April 2011 Interop Events and all

events thereafter. 36

**9.5 HARDWARE POLICY** 37

38

For MPI testing, HCA/RNIC vendors must provide at least five adapters. The 39

adapters need not be all the same model, but they can be.

40

**9.6 OFED USAGE** 41

42

**10 GENERAL INSTRUCTIONS**

**10.1 FIRST STEP INSTRUCTIONS**

• OFED Release Candidates (RC) should be used during the Interop Debug 1

Event. This allows vendors to resolve bugs and issues and commit them to 2

the OFED tree before the OFED General Availability (GA) is released. 3

• OFED GA versions shall be used for the Interop GA Events. 4

5

6

7

1) Burn the FW release XXX on all HCAs and RNICs using the above proce- 8

dure as required by vendor. 9

2) Host and Target Configuration 10

a) Install OFED software on host systems (using a 64 bit OS) configured to 11

run OFED. 12

b) Install WinOF software on host systems (using a 64 bit OS) configured 13

to run WinOF. 14

c) Configure non-OFED systems for use in the cluster as per the vendors 15

instructions.

16

d) Configure iSER/SRP targets for use in the cluster as per the vendors in- 17

structions.

18

3) Install the switch or gateway with the candidate SW stack as required by

vendor. 19

20

4) Burn the switch or gateway with the released FW as required by vendor.

21

5) Connect the Hosts and Targets to an appropriate switch following the basic 22

connectivity.

23

**10.2 INFINIBAND SUBNET MANAGERS** 24

1) The OpenSM will be used to run all mandatory tests in the test plan 25

2) Vendor SM testing will include testing IPoIB, RDMA Interop and Open MPI 26

testing. In order to reduce the scope of testing, iSER, NFS over RDMA, 27

RDS, SDP, SM Failover and SRP will not be performed using vendor SMs.

28

**10.3 OPERATING SYSTEM CONSIDERATIONS** 29

1) The OFILG decided as of April 2012 that the various ULPs contained in this 30

test plan will only be tested if it is supported by the Operating System. 31

2) As a requirement for the OFILG Logo, a vendor's DUT must pass all man- 32

datory testing using an agreed upon primary OS and OpenSM. Additional

beta testing is performed using secondary Operating Systems. This beta 33

testing has no bearing on whether the OFILG Logo is granted to a device It 34

is purely informative. 35

36

37

38

39

40

41

42

**11 INFINIBAND SPECIFIC INTEROP PROCEDURES USING OFED** 1

**Note**: UNH-IOL has created automated scripts to run many of the OFED based . 2 tests. Please contact them at [ofalab@iol.unh.edu](mailto:ofalab@iol.unh.edu) if you wish to obtain copies of 3 the latest scripts 4

**11.1 IB LINK INITIALIZE USING OFED FOR LINUX** 5

6

**11.1.1 Procedure** 7

1) Select a pair of devices to test from the created topology 8

2) Determine the maximum port width and lane speed supported by both de- 9

vices

10

3) Select a cable to use which has been certified for the link parameters deter- mined by step 2 of section 10.1.1 during an IBTA Plugfest held within the 11

last 6 months 12

4) Disconnect all IB cables from the selected devices 13

5) Shutdown all SMs running on the selected devices 14

6) Connect the selected devices back to back using the cable selected during 15

step 3 of section 10.1.1 16

7) Wait for a physical indication that a link has been established 17

8) Verify that the link created in step 6 of section 10.1.1 has come up with the 18

parameters determined in step 2 of section 10.1.1 19

9) Repeat steps 1-8 with a different device pairing 20 a) All unique device pairs present in the created topology must be tested; 21 except SRP target to SRP target and gateway to SRP target. 22

b) Each device must link at the maximum port width and lane speed sup- 23 ported by both devices in all pairings for said device to pass link initial- 24 ization testing 25

**11.1.2 Method of Implementation for all Linux OSs** 26

1) To perform step 7 of section 10.1.1: 27

a) Look for link LEDs on the ports you are using 28

2) To perform step 8 of section 10.1.1: 29 a) ssh into a device supporting such remote connections and is running 30 the OFED stack; usually a compute node with an HCA 31

b) Run "ibdiagnet -wt <desired-topology-file-name>" 32 c) Check the topology file created by the previous command: 33 i) Match the GUIDs to the devices in the selected pair 34

ii) Verify link width is the highest common denominator of pair capabil- 35

ities (1x, 4x, 12x) 36

iii) Verify link speed is the highest common denominator of pair capa- 37

bilities (2.5G, 5G, 10G, 14G) 38

3) To determine switch to SRP target and switch to switch link parameters 39 a) Run the commands outlined by step 2 of section 10.1.2 from a third de- 40 vice 41

42

i) Should be a compute node with an HCA that is linked to a switch 1

that is part of the desired pairing 2

ii) Carefully match the GUIDS as you now have more than just two in 3

the topology file 4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**11.2 IB FABRIC INITIALIZATION USING OFED** 1

**11.2.1 Architect the Network we want to build.** 2

1) Develop a cluster diagram based on the devices that have been submitted

for Interop Testing and assign IP addresses to the IPoIB interfaces and the 3 ethernet management interfaces.

2) See Figure 4- Sample Network Configuration below. 4

**11.2.2 Procedure**

**11.2.3 Verification Procedures**

5

1) Connect the HCAs and switches as per the Architected Network and make 6 sure that no SM/SA is running on the Fabric.

2) Start an SM on a device and let it initialize (all SM's will need to be tested) 7

3) Visually verify that all devices are in the active state. Verify that the LED is

on when the port is active. 8

4) Run "ibdiagnet -wt <file>" to generate a topology file

5) Run "ibdiagnet -pc" to clear all port counters 9

6) Wait 17 seconds as per the specifications requirements.

7) Run "ibdiagnet -c 1000" to send 1000 node descriptions. 10

8) Run "ibdiagnet" to generate fabric report.

a) Use /tmp/ibdiagnet.sm file to determine running sm 11 b) sminfo can also be used to determine the master SM or saquery -s to

find all SMs. 12

**Note**: "ibdiagnet -r" seg faulted but was fixed in OFED 1.5 according to

Bug 1618 13

9) Run "ibchecknet" to build guid list.

10) Run "ibdiagnet -t <file>" to compare current topology to the previously gen- 14 erated topology file

15

1) Review "PM Counters" section of the fabric report. There should be no il- 16 legal PM counters. The Specification says there should be no errors in 17

seconds. 17

2) Review "Subnet Manager " section of the fabric report. Verify that the

running SM is the one you started and verify number of nodes and switches 18

in the fabric.

3) Review the ibchecknet report and verify that there are no duplicate GUIDs in 19

the fabric

4) Verify that step 10 above indicates that the topology before the test and the 20

topology after the test are the same.

21

Restart all devices in the fabric and follow Sections 10.2.2 and 10.2.3. Run the 22

SM from a different device in the fabric until all SMs present have been used. All

SMs on managed switches (including those switches running **opensm**) should 23

be tested and at least one instance of **opensm** on an HCA must be tested. If there

are HCAs from more than one vendor, then **opensm** should be run from each 24

vendor’s HCA.

25

Each device must pass all verification procedures with every SM to pass Fabric

Initialization test. 26

1

2

**Table 27 - ibdiagnet commands** 3

|  |  |
| --- | --- |
| **Commands** | **Description** |
| Ibdiagnet -c 1000 | Send 1000 node descriptions |
| ibdiagnet -h | Help |
| Ibdiagnet -lw 4x - ls 2.5 | Specify link width and speed |
| Ibdiagnet - pc | Clear counters |
| ibdiagnet -t <file> | Compare current topology to saved topology |
| ibdiagnet -wt | Writes the topology to a file |

4

5

6

7

8

9

10

**Note**: The topology file is being generated after the SM starts but before any

testing has started. The topology comparison is being performed after testing has 11

been completed but before the systems get rebooted. A topology check is per- formed during every part of every test section that does not specifically state 12

"change the topology". For example Fabric Init only has 1 part so there is only 1 check but RDS has 2 parts so 2 checks are performed. However, IPoIB has 3 13

parts for each of 2 modes but 1 of those parts specifically says to change the to- pology so only 4 checks occur. 14

15

16

17

18

19

20

21

22

23

24

25

26

**Figure 4 - Sampie Network Configuration**

Ethernet Addressing

IP Address 172.16.xxx.xxx

Netmask 255.255.0.0

IPolB Addressing

IP Address tO.t O.xxx.xxx

Netmask 255.0.0.0

172.16.30.20

10 103020

172.16.20.24 172.16.20.20 172.16.20.3 172.16.130.2

10 10 20 24 10 10 20 20 10 10 20 3 10 10 100 2

**IYIC'IIanox IYIC'II800X lvtt:ll8flOX** Qloglc

|  |  |  |
| --- | --- | --- |
| '""''"'"ox  Connect)(  QDR Mlcroglga | |  |
| Farbau | d IG1:i |
|  |  | |
|  | |

Connect)( **ConnextX** Cougar

ee

DDR SDR

172.16.20.6 172.16.20.7

I Ska

Atlas Janus Skoll

h . I

172.16.20.25 172.16.20.21

10.10.20.6 10.10.20.7

-·

LSIDDR

6 5 11 10 2 1 10.10.20.25 10.10.20.21

**IYiena.nox Mellanox** SRP Target DON DDR "'euanox **Mellanox**

**MiniSDR lsix·bb2!·op** r- 1-

Cheelah Mellanox Shark SRP Target ConnextX **LionMini**

SOR 36 Port QDR DDR G2 DDR

Tilan **r**

**l**

**172.16.50.3 <lctl-&289000**

**hli-llbb 172. 18.90.1** Dione(G21 Mimas )

0. 10.20.22

**172.16.50.4**

24 13 12 I

172.16.20.22 I

1

6 *7* 8 **MeUanox** - 2

**Flextronlcs** CUb DDR

8 Port SDR **Pandora**

2 1 -

**J**

2 1 - 3

**Qlogie 24 Port**

**9024- DOR**

**ql9024 172.16.30.24**

"" •

+ 2 1 5

•

**Flextronics** • )

•

3 2 1

**QLogle 40Pori**

I

**904000R**

I **ql9040 172. 16.30.23**

**24Port 0DR**

5

)

LSISDR 6 5 • I 1

SRPTarget '"""•nux o..oogiC

**ConnextX**

**ll!JK· hbt.Jop** 1 -1 I I I

**172.16.50.1** -----1 SDR SDR

**19 xbb1-bottom**

|  |  |  |
| --- | --- | --- |
| ':'0510180  Cafri'US  **¢bfKafi'4)Uf>1**  **.n.16.7<ll** | |  |
|  |  | |
| **Melianox** Cub SDR Tarvoe | |  |

- ... c:U o

**vluaan** I Hvoerion Miranda

**1721650.2**

**Campus** I I

**ohs-eamp. d** I I

Cheelah

172.16.130.4 172.16.30.1

**172.16.70.4** I I DDR 10.10.100.1 10.10.30.1 3

•

I NarviIG2l Telesto

I

I 172.16.30.21 172.16.20.23 )

**Fibre**

**Channel** I 10.10.30.21 10.10.20.23

Switch ·I -- **v1u1an uu:s1u1an** 1

-

I LongbowXR

**1-**

**('.i)$..1Qnglxl>lf.l(r1**

Longbow XR

**ob5-klnglxJw..la'2**

+ **172.16.70.1 172. 16.70.2** 2

172.16.20.4 **Fibre** *;*

Gatewaylo

10.1020.4 **Channel Ethernet** 3

Disk Array **Fabric Mellanox**

TigerSDR

Calypso

December 2008 Topology 172.16.20.5 )

10.10.20.5

26

**11.3 IB IPOIB CONNECT MODE (CM) USING OFED** 1

**11.3.1 SETUP** 2

Connect the HCAs and switches as per the Architected Network and make sure 3

that no SM is running on the Fabric. 4

5

This procedure, as the previous ones, will be based on the cluster connectivity. 6

An SM/SA which supports IPoIB (sufficient IB multicast support) will be running

on the HCAs, or on a switch with an embedded SM/SA or a third HCA which 7

would only run SM/SA for the partner pair (with a switch in the middle). This pro- 8

cedure has been developed for Linux and may be ported to Windows if there is 9

sufficient vendor support. 10

**Optional**: In the procedures below, an IB analyzer can be inserted in the appro- 11 priate link to obtain traces and validate the aspects of the procedures specifically 12 detailed below in subsequent sections. 13

14

**11.3.2 IPOIB INTERFACE CREATION AND IPOIB SUBNET CREATION** 15

1) Configure IPoIB address. All addresses must reside on the same subnet. 16

a) Set interfaces to 10.0.0.x/24 (10.0.0.x/netmask 255.255.255.0) using 17

the command *ifconfig ib0 10.0.0.x netmask 255.255.255.0* 18

**11.3.3 .BRINGING THE IPOIB IN CONNECTED MODE** 19

1) echo 'connected' > /sys/class/net/ib0/mode 20

2) Validate CM mode by checking that "/sys/class/net/<I/F name>/mode" equal 21

to '**connected**’ 22

3) Repeat steps 1-2 in section 10.3.3 on all nodes being tested. 23

24

25

**11.3.4 PING PROCEDURES** 26

**Step A** 1) Stop all SM's and verify that none are running 27

2) Power cycle all switches in the fabric (this insures that the new SM will con- 28

figure all the links and create the multi-cast join). 29

3) Start an SM (All SM's will need to be tested) and let it initialize 30 a) Visually verify that all devices are in the active state. Verify that the LED 31 is on when the port is active. 32

b) Run "ibdiagnet -r" and verify that the SM you started is the one that is 33 running and and that it is the master. You will need to know the GUID of 34 the device since the SM will be reassigned on each reboot. 35

c) Verify that all nodes and switches were discovered. 36

**Note**: Ibdiagnet may show more switches than indicated by the physical 37 number of switch platforms present. This is because some switches have 38 multiple switch chips.

39

4) Examine the arp table (via arp -a) and remove the destination node’s ib0 ad- dress from the sending node’s arp table (via arp -d). 40

41

42

5) Ping every HCA except localhost with packet sizes of 511, 1025, 2044, 1

8192, 32768 and 65507. 2

a) ping -i 0.2 -t 3 -c 10 -s <ping size> <destination> 3 i) "-i" - interval 0.2 seconds 4 ii) "-t" - IP Time to Live equals 3 seconds 5 iii) "-c" - count equals 100 6 iv) "-s" - size of the ping 7

v) "destination" - the IP address of the IPoIB interface being pinged. 8

b) Repeat step #4 before issuing each ping command. Every packet size is 9

a new ping command. 10

6) In order to pass Step A, a reply must be received for every ping sent (without 11 losing a single packet) while using each one of the SMs available in the 12 cluster. 13

**Step B** 1) Bring up all HCAs but one. 14

2) Start an SM (all SMs will need to be tested). 15

3) Check for ping response between all node (All to All). 16

a) A response from the disconnected HCA should not be returned. 17

18

4) Disconnect one more HCA from the cluster. 19

5) Ping to the newly disconnected HCA from all nodes (No response should be 20

returned).

21

6) Connect the first machine (the one that was not connected) and check for 22

ping response from all nodes that are still connected.

23

7) Connect the disconnected HCA to a different switch on the subnet which will change the topology. 24

25

8) Ping again from all nodes (this time we should get a response).

26

9) Follow Step B, this time bring the interface down and then back up using if- 27 config ibX down and ifconfig ibX up commands instead of physically discon- necting the HCAs. 28

**Note**: Each step must exhibit the expected behavior while using each SM in 29

order for the device to pass Step B overall. 30

**Step C** Follow Step A and B using a different SM until all SM's have been used. Only one 31 instance of each available SM is required. Steps A, B, and C must pass in order 32 for the device to pass 10.3.4 overall. 33

34

**11.3.5 SFTP PROCEDURE** 35

SFTP procedures require an SFTP server to be configured on each machine in 36

the partner pair. An SFTP client needs to be available on each machine as well.

The default RHEL install includes both. 37

38

A 4 MB file will be SFTP'd to the partner and then SFTP'd back and binary com- 39 pared to the original file, this will be done in each direction and then bidirectional 40 using every SM available.

41

42

**11.3.5.1 SETUP**

**11.3.5.2 PROCEDURE**

1

1) Make sure vsftpd is installed on each node for SFTP application. 2

3

2) A special account for this should be created as follows:

4

b) Username: Interop 5

c) Password: openfabrics 6

7

1) Run SFTP server on all nodes. 8

2) Start an SM (all SM's will need to be tested) and let it initialize 9

10

a) Verify that the running SM is the one you started. 11

3) SFTP: 12

a) Connect an HCA pair via SFTP on IPoIB using the specified user name 13

and password. 14

b) Put the 4MB file to the /tmp dir on the remote host. 15

c) Get the same file to your local dir again. 16 d) Compare the file using the command *cmp tfile tfile.orig.* 17 i) The two must be identical 18

4) Repeat the procedure with a different SM. 19

**Note**: Every node must SFTP the 4MB file to all others using all SM's and the 20 files must be identical as determined by the binary compare in order for the 21 device to pass 10.3.5 overall. 22

**Note**: Sections 10.3.4 and 10.3.5 must pass using the configuration deter- 23

mined by sections 10.3.1, 10.3.2, and 10.3.3 for the device to pass IPoIB 24

Connected mode overall. 25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**11.4 IB IPOIB DATAGRAM MODE (DM) USING OFED** 1

**11.4.1 SETUP** 2

Connect the HCAs and switches as per the Architected Network and make sure 3

that no SM is running on the Fabric. 4

5

This procedure, as the previous ones, will be based on the cluster connectivity. 6

An SM/SA which supports IPoIB (sufficient IB multicast support) will be running

on the HCAs, or on a switch with an embedded SM/SA or a third HCA which 7

would only run SM/SA for the partner pair (with a switch in the middle). This pro- 8

cedure has been developed for Linux and may be ported to Windows if there is 9

sufficient vendor support. 10

**Optional**: In the procedures below, an IB analyzer can be inserted in the appro- 11 priate link to obtain traces and validate the aspects of the procedures specifically 12 detailed below in subsequent sections. 13

14

**11.4.2 IPOIB INTERFACE CREATION AND IPOIB SUBNET CREATION** 15

1) Configure IPoIB address. All addresses must reside on the same subnet. 16

a) Set interfaces to 10.0.0.x/24 (10.0.0.x/netmask 255.255.255.0) using 17

the command *ifconfig ib0 10.0.0.x netmask 255.255.255.0* 18

**11.4.3 .BRINGING THE IPOIB IN DATAGRAM MODE** 19

1) echo 'datagram' > /sys/class/net/ib0/mode 20

2) Validate DM mode by checking that "/sys/class/net/<I/F name>/mode" equal 21

to '**datagram**’ 22

3) Repeat steps 1-2 in section 10.4.3 on all nodes being tested. 23

24

25

**11.4.4 PING PROCEDURES** 26

**Step A** 1) Stop all SM's and verify that none are running 27

2) Power cycle all switches in the fabric (this insures that the new SM will con- 28

figure all the links and create the multi-cast join). 29

3) Start an SM (All SM's will need to be tested) and let it initialize 30 a) Visually verify that all devices are in the active state. Verify that the LED 31 is on when the port is active. 32

b) Run "ibdiagnet -r" and verify that the SM you started is the one that is 33 running and and that it is the master. You will need to know the GUID of 34 the device since the SM will be reassigned on each reboot. 35

c) Verify that all nodes and switches were discovered. 36

**Note**: Ibdiagnet may show more switches than indicated by the physical 37 number of switch platforms present. This is because some switches have 38 multiple switch chips.

39

4) Examine the arp table (via arp -a) and remove the destination node’s ib0 ad- dress from the sending node’s arp table (via arp -d). 40

41

5) Issue the command: sysctl net.ipv4.neigh.ib0.unres\_qlen=33

42

a) This sets the qlen variable to 33 which increases the buffer size so that 1 you do not get an initial dropped packet when using ping sizes 8192 and 2 greater. 3

6) Ping every HCA except localhost with packet sizes of 511, 1025, 2044, 4

8192, 32768 and 65507.

5

a) ping -i 0.2 -t 3 -c 10 -s <ping size> <destination> 6

i) "-i" - interval 0.2 seconds 7

ii) "-t" - IP Time to Live equals 3 seconds 8 iii) "-c" - count equals 100 9 iv) "-s" - size of the ping 10 v) "destination" - the IP address of the IPoIB interface being pinged. 11

b) Repeat step #4 before issuing each ping command. Every packet size is 12

a new ping command. 13

7) In order to pass Step A, a reply must be received for every ping sent (without 14 losing a single packet) while using each one of the SMs available in the 15 cluster. 16

**Step B** 1) Bring up all HCAs but one. 17

2) Start an SM (all SMs will need to be tested). 18

3) Check for ping response between all node (All to All). 19

a) A response from the disconnected HCA should not be returned. 20

4) Disconnect one more HCA from the cluster. 21

5) Ping to the newly disconnected HCA from all nodes (No response should be 22

returned). 23

6) Connect the first machine (the one that was not connected) and check for 24

ping response from all nodes that are still connected. 25

7) Connect the disconnected HCA to a different switch on the subnet which will 26

change the topology. 27

8) Ping again from all nodes (this time we should get a response). 28

9) Follow Step B, this time bring the interface down and then back up using if- 29 config ibX down and ifconfig ibX up commands instead of physically discon- 30 necting the HCAs. 31

**Note**: Each step must exhibit the expected behavior while using each SM in 32

order for the device to pass Step B overall. 33

**Step C** 1) Follow Step A and B using a different SM until all SM's have been used. 34

Only one instance of each available SM is required. Steps A, B, and C must 35

pass in order for the device to pass 10.4.4 overall. 36

2) Issue the command: sysctl net.ipv4.neigh.ib0.unres\_qlen=3 37

a) This sets the qlen variable back to the default. 38

**11.4.5 SFTP PROCEDURE** 39

SFTP procedures require an SFTP server to be configured on each machine in 40

the partner pair. An SFTP client needs to be available on each machine as well. 41

The default RHEL install includes both. 42

**11.4.5.1 SETUP**

**11.4.5.2 PROCEDURE**

A 4 MB file will be SFTP'd to the partner and then SFTP'd back and binary com- 1 pared to the original file, this will be done in each direction and then bidirectional 2 using every SM available. 3

4

5

1) Make sure vsftpd is installed on each node for SFTP application. 6

2) A special account for this should be created as follows: 7

b) Username: Interop 8

c) Password: openfabrics 9

10

11

Run SFTP server on all nodes.

12

1) Start an SM (all SM's will need to be tested) and let it initialize 13

a) Verify that the running SM is the one you started. 14

2) SFTP: 15

16

a) Connect an HCA pair via SFTP on IPoIB using the specified user name

and password. 17

b) Put the 4MB file to the /tmp dir on the remote host. 18

19

c) Get the same file to your local dir again. 20

d) Compare the file using the command *cmp tfile tfile.orig.* 21

i) The two must be identical 22

3) Repeat the procedure with a different SM. 23

**Note**: Every node must SFTP the 4MB file to all others using all SM's and the 24 files must be identical as determined by the binary compare in order for the 25 device to pass 10.4.5 overall. 26

**Note**: Sections 10.4.4 and 10.4.5 must pass using the configuration deter- 27

mined by sections 10.4.1, 10.4.2, and 10.4.3 for the device to pass IPoIB Datagram mode overall. 28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**11.5 IB SM FAILOVER AND HANDOVER PROCEDURE USING OFED** 1

**11.5.1 SETUP** 2

1) Connect HCAs per the selected topology. 3

2) In this test, all active SMs on the fabric which are going to be tested, must 4

be from the same vendor. They will be tested pairwise; two at a time. 5

**11.5.2 PROCEDURE** 6

7

1) Disable all SMs in the cluster then start a SM on either machine in a chosen pair. 8

2) Run "saquery" on a node in the fabric. 9

10

a) Verify that all nodes in the cluster are present in the output 11

3) Using the ibdiagnet tool with the -r option, verify that the running SM is the 12

master.

13

4) Start a SM on the second machine in the current pair. 14

5) Verify that the SMs behave according to the SM priority rules. Use "ibdi- 15 agnet -r" again. 16 a) SM with highest numerical priority value is master and the other is in 17

standby.

18

a) If both SMs have the same priority value then the SM with the smallest 19

guid is master and the other is in standby.

20

6) Run "saquery" on either machine in the current pair. 21

|  |  |
| --- | --- |
| a) Verify that all nodes in the cluster | are present in the output. 22 |
| 7) Shutdown the master SM. | 23 |
| 8) Verify the other active SM goes into t | he master state using "ibdiagnet -r" 24 |
| again. | 25 |
| 9) Run "saquery" on either machine in th | e current pair. 26 |
| a) Verify that all nodes in the cluster | are present in the output. 27 |

10) Start the SM you just shutdown. 28

11) Verify that the newly started SM resumes it's position as master while the 29

other goes into standby again. 30

12) Run "saquery" on either machine in the current pair. 31

a) Verify that all nodes in the cluster are present in the output. 32

33

13) Shutdown the standby SM.

34

14) Verify that the previous master SM is still the master. 35

15) Run "saquery" on either machine in the current pair. 36

a) Verify that all nodes in the cluster are present in the output. 37

16) Repeat steps 1-15 above 2 more times, ensuring that the below criteria is 38 met (total of 3 tests per pair which can be run in any order): 39 a) First SM to be started having highest numerical priority value. 40 b) Second SM to be started having highest numerical priority value. 41

42

c) Both SMs having equal numerical priority values. 1

17) Repeat steps 1-16 until all possible SM pairs from identical vendors in the 2

cluster have been tested. 3

18) All of the "saquery" commands must return the expected list of nodes in 4

order for the SMs in this test to receive a passing grade. 5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**11.6 IB SRP USING OFED**

**11.6.1 SETUP**

1

2

Connect the HCAs and switches as per the Architected Network and make sure 3

that no SM is running on the Fabric. 4

5

**Note**: As of the April 2012 Interop events, one SRP target (i.e.target port) should 6

present 2 or more volumes. All other target ports may be limited to one volume

per port. This decision was made in order to reduce the amount of time required 7

to run the tests. 8

9

**Note**: As of October 2012, the SRP Extended Procedure is a Beta test 10

**11.6.2 SRP CORE PROCEDURE - MANDATORY** 11

1) Start an SM (all SM's will need to be tested) and let it initialize 12

13

a) Verify that the running SM is the one that you started 14

2) Choose a node to work with 15

3) Unload the srp module 16

4) Load srp module with cmd\_sg\_entries=255 17 a) **Example**: modprobe ib\_srp cmd\_sg\_entries=255 18 b) Let it initialize 19

5) Verify that the module loaded correctly 20

21

a) **Example**: lsmod | grep ib\_srp

22

6) Load srp\_daemon with -e -o -n options 23

a) **Example**: srp\_daemon -e -o -n 24

b) Let it initialize 25

7) Find all volumes from all targets 26

a) Use lsscsi 27

**Note**: As of April 2012, the OFILG mandated that the target only include 28

two volumes when doing mandatory testing. 29

8) Perform 6GB read from srp volume to null 30

a) **Example**: dd if=$drive of=/dev/null count=600 bs=10M 31

32

9) Perform 6GB write from zero to srp volume

33

a) **Example**: dd if=/dev/zero of=$drive count=600 bs=10M 34

10) Perform steps #8 and #9 for both volumes found from each target as deter- 35

mined by step #7 36

11) Unload srp module 37

12) Repeat steps 2 through 9 for all HCAs 38

13) Reboot all devices in the fabric and repeat the procedure using a different 39

SM. 40

**Note**: An HCA must successfully complete all DD operations to and from all 41

volumes on all targets using all available SM's in order to pass SRP testing. 42

1

**11.6.3 SRP EXTENDED PROCEDURE - BETA** 2

1) Start an SM (all SM's will need to be tested) and let it initialize 3

a) Verify that the running SM is the one that you started 4

2) Choose a node to work with 5

6

3) Unload the srp module 7

4) Load srp module with cmd\_sg\_entries=255 allow\_ext\_sg=1 8

indirect\_sg\_entries=2048

9

a) **Example**: modprobe ib\_srp cmd\_sg\_entries=255 allow\_ext\_sg=1 10

indirect\_sg\_entries=2048

11

b) Let it initialize 12

5) Verify that the module loaded correctly 13

a) **Example**: lsmod | grep ib\_srp 14

6) Load srp\_daemon with -e -o -n options 15 a) **Example**: srp\_daemon -e -o -n 16 b) Let it initialize 17

7) Find all volumes from all targets 18

19

a) Use lsscsi

20

**Note**: As of April 2012, the OFILG mandated that the target only include 21

two volumes when doing mandatory testing.

22

8) Perform 6GB read from srp volume to null 23

a) **Example**: dd if=$drive of=/dev/null count=600 bs=10M 24

9) Perform 6GB write from zero to srp volume 25

a) **Example**: dd if=/dev/zero of=$drive count=600 bs=10M 26

10) Perform steps #8 and #9 for both volumes found from each target as deter- 27

mined by step #7 28

11) Unload srp module 29

12) Repeat steps 2 through 9 for all HCAs 30

13) Reboot all devices in the fabric and repeat the procedure using a different 31

SM. 32

**Note**: An HCA must successfully complete all DD operations to and from all 33

volumes on all targets using all available SM's in order to pass SRP testing 34

35

36

37

38

39

40

41

42

**11.7 IB ETHERNET GATEWAY USING OFED** 1

**11.7.1 PROCEDURE** 2

1) Connect the HCA of the IB host to the IB fabric. Connect the Ethernet 3

Gateway to the IB fabric. Connect the Ethernet gateway to the Ethernet net- 4

work or Ethernet device. Start the SM to be used in this test. 5

2) Determine which ULP your ethernet gateway uses and be sure that ULP is 6

running on the host (VNIC or IPoIB). 7

3) Restart the ULP or using the tool provided by the ULP, make sure that the 8 host “discovers” the Ethernet Gateway. Configure the interfaces and make 9 sure they are up. 10

4) Run ping from the host to the Ethernet device. While the ping is running, kill 11

the master SM. Verify that the ping data transfer is unaffected.

12

5) Reboot the Ethernet Gateway. After the Ethernet Gateway comes up, verify 13

that the host can discover the Ethernet Gateway as it did before and we are

able to configure the interfaces. 14

6) Restart the ULP used by Ethernet Gateway and verify that after the ULP 15 comes up, the host can discover the Ethernet Gateway and we are able to 16 configure the interfaces. 17

7) Unload the ULP used by Ethernet Gateway and check that the Ethernet 18

Gateway shows it disconnected. Load the ULP and verify that the Ethernet 19

gateway shows the connection. 20

8) Repeat step 4 by using ssh and scp instead of ping. 21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**11.8 IB FIBRECHANNEL GATEWAY USING OFED** 1

**11.8.1 PROCEDURE** 2

1) Connect the HCA of the IB host to the IB fabric. Connect the FC Gateway 3

to the IB Fabric (how to do this is determined by the FC Gateway vendor). 4

Connect the FC Gateway to the FC network or FC device. Start the SM to 5

be used in this test. 6

2) Configure the FC Gateway appropriately (how to do this is vendor specific). 7

3) Use ibsrpdm tool in order to have the host "see" the FC storage device. Add 8

the storage device as target. 9

4) Run basic dd application from the SRP host to the FC storage device. 10

5) Run basic dd application from the SRP host to the FC storage device. 11

While the test is running, kill the master SM. Verify that the test completes 12

properly. 13

6) Unload the SRP host / SRP Target (target first/host first) and check that the 14

SRP connection is properly disconnected. 15

7) Load the SRP host / SRP Target. Using ibsrpdm, add the target. 16

8) Run basic dd application from the SRP host to the FC storage device. 17

9) Reboot the FC Gateway. After FC Gateway comes up, verify using ibsrpdm 18 tool that the host see the FC storage device. Add the storage device as 19 target. 20

10) Run basic dd application from the SRP host to the FC storage device. 21

11) Follow steps 1-10 above with each SM to be tested and with each HCA to 22

be tested, until each HCA and each SM has been tested with the FC 23

Gateway.

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**12 ETHERNET SPECIFIC INTEROP PROCEDURES USING OFED** 1

**12.1 IWARP LINK INITIALIZE USING OFED** 2

**12.1.1 PURPOSE** 3

4

The iWARP Link Initialize test is a validation that all iWARP devices receiving the

OFA Logo can link and pass traffic under nominal (unstressed) conditions. 5

6

**12.1.2 RESOURCE REQUIREMENTS** 7

1) Gigabit or 10Gigabit iWARP RNIC, 8

2) Gigabit or 10Gigabit Ethernet Switch 9

3) Compliant Cables 10

**12.1.3 DISCUSSION** 11

12

The validation of the underlying transport infrastructure is essential to the end-

users experience of the operation of the OFED software stack. To this end, this 13

test confirms that iWARP devices receiving the OFA Logo can suitably link and 14

pass traffic in any configuration. Exhaustive compliance testing of BER perfor- 15 mance of the channel or electrical signaling of the ports is not performed; how- 16 ever, successful completion of this test provides further evidence of the

robustness of the OFA logo bearing device. 17

18

**12.1.4 PROCEDURE** 19

1) Connect the two link partners together utilizing compliant cables. 20

2) Check all relevant LEDs on both ends of the link. 21

3) Verify that basic IP connectivity can occur by driving minimum size ICMP 22

echo requests and replies across the link or equivalent traffic (including 23

RDMA traffic if readily configured, in which case an additional RNIC re- 24 sponder station is required). To verify that an RDMA link has been initialized 25 between Host A and Host B run the following commands: 26

a) Start a server in verbose mode on Host A: 27

i) rping -sv 28 b) Start a client on Host B to ping Host A. 29 i) rping -cv -a Host A *RNIC\_IP\_Address* 30

c) Optional Command for the client 31

i) rping -cv -a Host A *RNIC\_IP\_Address* -C 4 -S 50 32

**Note**: This sends a count of 4 pings and character strings of size 50 33

4) Repeat steps 1-3 for all combinations of 2 RNICs to switches, switch to 34

switch, and RNIC to RNIC link partner combinations. Previously tested com- 35

binations resident in the OFILG cluster may be omitted. 36

**12.1.5 OBSERVABLE RESULTS** 37

1) Link should be established on both ends of the channel. 38

2) Traffic should pass in both directions. Error rates of 10e-5 or better should 39

be readily confirmed (no lost frames in 10,000). 40

41

42

**12.1.6 POSSIBLE PROBLEMS**

1

1) Traffic directed to a switches IP management address may not be pro- 2 cessed at high speed, in such cases, traffic should be passed across the 3 switch to a remote responder. 4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**12.2 ROCE LINK INITIALIZE USING OFED** 1

**12.2.1 PURPOSE** 2

The RoCE Link Initialize test is a validation that all RoCE devices receiving the 3

OFA Logo can link and pass traffic under nominal (unstressed) conditions. 4

5

**12.2.2 RESOURCE REQUIREMENTS** 6

1) 10 or 40 Gigabit RoCE Channel Adapter (RCA) 7

2) 10 or 40 Gigabit RoCE Switch (DCB Enabled) 8

3) Compliant Cables 9

**12.2.3 DISCUSSION** 10

The validation of the underlying transport infrastructure is essential to the end- 11

users experience of the operation of the OFED software stack. To this end, this 12

test confirms that RoCE devices receiving the OFA Logo can suitably link and 13 pass traffic in any configuration. Exhaustive compliance testing of BER perfor- 14 mance of the channel or electrical signaling of the ports is not performed; how-

ever, successful completion of this test provides further evidence of the 15

robustness of the OFA logo bearing device. 16

17

**12.2.4 PROCEDURE** 18

1) Connect the two link partners together utilizing compliant cables. 19

2) Check all relevant LEDs on both ends of the link. 20

3) Verify that basic IP connectivity can occur by driving minimum size ICMP 21

echo requests and replies across the link or equivalent traffic (including 22

RDMA traffic if readily configured, in which case an additional RoCE re- 23

sponder station is required). To verify that an RDMA link has been initialized between Host A and Host B run the following commands: 24

25

a) Start a server in verbose mode on Host A:

26

i) rping -sv 27

b) Start a client on Host B to ping Host A. 28

i) rping -cv -a Host A *RCA\_IP\_Address* 29 c) Optional Command for the client 30 i) rping -cv -a Host A *RCA\_IP\_Address* -C 4 -S 50 31

**Note**: This sends a count of 4 pings and character strings of size 50 32

4) Repeat steps 1-3 for all combinations of 2 RCAs to switches, switch to 33 switch, and RCA to RCA link partner combinations. Previously tested combi- 34 nations resident in the OFILG cluster may be omitted. 35

**12.2.5 OBSERVABLE RESULTS** 36

1) Link should be established on both ends of the channel. 37

2) Traffic should pass in both directions. Error rates of 10e-5 or better should 38

be readily confirmed (no lost frames in 10,000). 39

40

41

42

**12.3 ROCE FABRIC INIT USING OFED** 1

This test will be developed for the April 2013 Interop Debug event 2

3

**12.4 ROCE IPOCE** 4

The test for IP over Converged Ethernet will be developed for the April 2013 In- 5

terop Debug event 6

**12.5 ROCE INFINIBAND GATEWAY** 7

8

This test will be developed for the April 2013 Interop Debug event 9

**12.6 ROCE FIBRE CHANNEL GATEWAY** 10

This test will be developed for the April 2013 Interop Debug event 11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13 TRANSPORT INDEPENDENT INTEROP PROCEDURES USING OFED** 1

**13.1 TI ISER USING OFED** 2

**13.1.1 IB SETUP** 3

4

Connect initiator/target to switch as well as run one or more SMs (embedded in

the switch or host based). If more than one SM, let the SMs split into master and 5

slave. 6

7

**Optional**: In the procedures below, an IB analyzer can be inserted in the appro- 8

priate link to obtain traces and validate the aspects of the procedures specifically detailed below in subsequent sections. 9

10

**13.1.2 IWARP SETUP** 11

Connect iSER host initiator and target RNICs to an 10GbE switch. 12

13

**13.1.3 ROCE SETUP** 14

Connect iSER host initiator and target RCA to a 10/40 GbE switch which is DCB 15

Enabled. 16

**13.1.4 PROCEDURE** 17

1) Load iSER target and iSER initiator to hosts from OpenFabrics tree, check 18

iSER connection. 19

2) Run basic dd application from iSER initiator host connected to target. 20

3) [IB Specific Test] Run basic dd application from iSER initiator host con- 21 nected to target. Kill the master SM while test is running and check that it 22 completes properly. 23

4) Unload iSER initiator from a Host and check iSER connection properly dis- 24

connected on a target host. 25

5) Unload iSER target from a Host and check iSER connection properly dis- 26

connected on an initiator host. 27

6) [IB Specific Test] Repeat steps 2-5 now with the previous slave SM (we did 28

not actually stop the target). 29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13.2 TI NFS OVER RDMA USING OFED** 1

**13.2.1 Installation** 2

**Note:** Steps 2-4 are unneeded if an OFED supported OS is used along with an 3

official OFED release downloaded from [http://www.openfabrics.org](http://www.openfabrics.org/) 4

5

1) Verify that you are using a Linux kernel with NFS/RDMA on every system

used 6

7

a) The NFS/RDMA client and server are both included in the mainline

Linux kernel version 2.6.25 and later. This and other versions of the 2.6 8

Linux kernel can be found at: ftp://ftp.kernel.org/pub/linux/kernel/v2.6/ 9

**Note:** OFED supported OS releases of lower kernel revision than men- 10

tioned above have been updated by their respected maintainers to allow 11

NFS RDMA to function. Check the nfs-rdma.release-notes.txt provided 12

with the OFED release you are using for supported OS releases. 13

**Note:** As of OFED 1.5.3 rc2 NFSoRDMA is not installed by default. To 14

do so you must have built OFED from src with nfsrdma=y directive con- tained within the ofed.conf file used by the OFED installer. 15

16

i) To generate an ofed.conf file run the following from within the down- loaded OFED src. 17

1. $ ./install.pl -p 18

ii) Add the following directives to the generated ofed-all.conf file 19

20

1. nfsrdma=y

21

iii) Install OFED 22

1. ./install.pl -c ofed-all.conf 23

2) Configure the RDMA stack on every system used 24

a) Make sure your kernel configuration has RDMA support enabled. Under 25

Device Drivers => InfiniBand support, update the kernel configuration to 26

enable InfiniBand support.

27

**Note**: the option name is misleading. Enabling InfiniBand support is re- quired for all RDMA devices (IB, iWARP, etc.). 28

29

b) Enable the appropriate IB HCA support (mlx4, mthca, ehca, ipath, qib,

etc.) or iWARP adapter support (amso, cxgb3, etc.). 30

c) If you are using InfiniBand, be sure to enable IP-over-InfiniBand (IPoIB) 31

support. 32

3) Configure the NFS client 33

34

a) Your kernel configuration must also have NFS file system support

and/or NFS server support enabled. These and other NFS related con- 35

figuration options can be found under File Systems => Network File 36

Systems. 37

4) Build, install, reboot 38 a) The NFS/RDMA code will be enabled automatically if NFS and RDMA 39 are turned on. The NFS/RDMA client and server are configured via the 40

hidden SUNRPC\_XPRT\_RDMA config option that depends on SUN- 41

RPC and INFINIBAND. The value of SUNRPC\_XPRT\_RDMA will be: 42

i) - N if either SUNRPC or INFINIBAND are N, in this case the 1

NFS/RDMA client and server will not be built 2

ii) - M if both SUNRPC and INFINIBAND are on (M or Y) and at least 3 one is M, in this case the NFS/RDMA client and server will be built 4 as modules

5

iii) - Y if both SUNRPC and INFINIBAND are Y, in this case the

NFS/RDMA client and server will be built into the kernel 6

b) If you have followed the steps above and turned on NFS and RDMA, the 7

NFS/RDMA client and server will be built. 8

c) Build a new kernel, install it and boot it 9

5) Check RDMA Setup 10

11

a) If you are using InfiniBand, make sure there is a Subnet Manager (SM)

running on the network. 12

13

b) Use IPoIB to ping two hosts.

14

6) Configure NFS exports, start NFS server 15

a) Use two machines, one to act as the client and one to act as the server. 16

b) On the server system, configure the /etc/exports file and start the 17

NFS/RDMA server. Export entries with the following formats have been 18

tested:

19

i) /vol0 192.168.0.47(fsid=0,rw,async,insecure,no\_root\_squash) 20

ii) /vol0 192.168.0.0/255.255.255.0(fsid=0,rw,async,inse- 21

cure,no\_root\_squash)

22

c) The IP address(es) is (are) the client's IPoIB address for an InfiniBand

HCA or the client's iWARP address(es) for an RNIC. 23

24

**Note**: The "insecure" option must be used because the NFS/RDMA client

does not use a reserved port. This does not interfere with normal NFS over 25

TCP/IP operations. 26

d) The remainder of this section will assume an export of /server 27

e) Start the NFS server 28

29

i) If the NFS/RDMA server was built as a module

(CONFIG\_SUNRPC\_XPRT\_RDMA=m in kernel config), load the 30

RDMA transport module: 31

1. $ modprobe svcrdma 32 ii) Regardless of how the server was built (module or built-in), start the 33 server: 34

1. $ /etc/init.d/nfs start or service nfs start 35

iii) Instruct the server to listen on the RDMA transport: 36

1. $ echo rdma 20049 > /proc/fs/nfsd/portlist 37

7) Check NFS Setup 38 a) For the NFS components enabled above (client and/or server), test their 39 functionality over standard Ethernet using TCP/IP or UDP/IP. 40

b) On the client system: 41

42

i) Use this command to mount the NFS server export: 1

1. $ mount <server name or TCP/IP address>:/<export> /<mount 2

path> 3

ii) To verify that the mount is using TCP, run "cat /proc/mounts" and 4

check the "proto" field for the given mount. 5

8) Check NFS/RDMA Setup 6

a) For the NFS components enabled above (client and/or server), test their 7

functionality over RDMA. 8

b) On the client system: 9

i) If the NFS/RDMA client was built as a module 10

(CONFIG\_SUNRPC\_XPRT\_RDMA=m in kernel config), load the 11

RDMA client module: 12

1. $ modprobe xprtrdma 13

ii) Regardless of how the client was built (module or built-in), use this 14

command to mount the NFS server export:

15

1. $ /sbin/mount.rnfs <IPoIB server name or address>:/<export>

/<mount path> -o \ rdma,port=20049 16

**Note**: OFED will build and install the mount utility needed. The binary 17 is called mount.rnfs. Either this binary or the mount binary provided 18 with nfs-utils revision greater than version 1.1 can be used. The re- 19

mainder of this section will assume mount.rnfs built by OFED is 20

used. 21

iii) To verify that the mount is using RDMA, run "cat /proc/mounts" and 22

check the "proto" field for the given mount.

23

9) Connectathon 24

a) Download the Connectathon test suite from http://www.connec- 25

tathon.org/nfstests.html

26

b) Install Connectathon on every client to be used 27

i) Modify tests.init within the connectathon tarball to suit your environ- 28

ment. 29

1. The MOUNTCMD, UMOUNTCMD and MNTOPTIONS direc- 30 tives are unimportant as we will be calling the runtests connec- tathon binary directly. 31

2. Be sure to remove the -fwritable-strings option from the 32

CFLAGS variable. Your build will fail if this is used. 33

ii) Run make to build the connectathon binaries. 34

10) Test the connectathon runtests binary 35 a) Run sudo ./runtests -a -t /mnt/ to test the binary against the local file sys- 36 tem. 37

b) All tests should pass but you will see 1 warning. This is ok. 38

39

40

41

42

**13.2.2 NFSoRDMA Test Procedure**

1

1) **Note**: IB Only 2

a) Start an SM 3

2) Server setup 4

a) Add nfs rdma server support to the running kernel if not already present. 5

6

i) $ modprobe svcrdma

7

b) Start the server 8

i) $ /etc/init.d/nfs start 9

c) Tell the server to listen for rdma connection requests on port 20049 10

i) $ echo rdma 20049 > /proc/fs/nfsd/portlist 11

3) Client setup 12 a) Add nfs rdma client support to the running kernel if not already present. 13 i) $ modprobe xprtrdma 14

b) Mount the servers export using rdma 15

i) $ /sbin/mount -t nfs <server IPoIB address>:/server /<mount path> - 16

o \ rdma,port=20049 -i 17

**Note**: <mount path> is assumed to be /mnt/<servername> for the re- 18

mainder of this section 19

c) Verify that the mount is using the rdma protocol 20 i) Verify that the mount is using RDMA, run "cat /proc/mounts" and 21 check the "proto" field for the given mount. 22

4) Run Connectathon's runtests binary 23

a) $ ./runtests -a -t /mnt/<servername>/<hostname> 24

25

5) Repeat steps 2-4 using a new client-server pair until all nodes have acted as

both a server and a client. 26

6) Repeat steps 2-5 using a new SM until all registered SM's have been used. 27

28

7) All tests run by the connectathon runtests binary must pass on all client

nodes rdma mount points from all server nodes using all SM's in order for 29

the device to pass NFSoRDMA Test Procedure overall. 30

31

32

33

34

35

36

37

38

39

40

41

42

**13.3 TI RELIABLE DATAGRAM SERVICE (RDS) USING OFED** 1

**13.3.1 RDS-PING PROCEDURE** 2

**Note**: RDS does not support iWARP 3

4

1) Use the command *modprobe rds\_rdma* to add RDS support to the kernel 5

2) Verify that the kernel supports RDS by issuing the *rds-info* command. 6 a) The rds-info utility presents various sources of information that the RDS 7 kernel module maintains. When run without any optional arguments rds- 8

info will output all the information it knows of. 9

3) [**For IB**] Start one of the Subnet Managers in the cluster 10

**Note**: RDS is IP based so you need to provide a host address either through 11

an out of band Ethernet connection or through IPoIB. RDS also requires the 12

LIDs to be set in an InfiniBand Fabric and therefore an SM must be run. 13

**Note**: All SMs in the fabric should be tested. 14

4) Choose a host and use *rds-ping host* to communicate with every other end 15

point in the fabric. 16

**Note**: Be sure that you identify the correct host when using the command *rds-* 17 *ping host*. 18 a) rds-ping is used to test whether a remote node is reachable over RDS. 19

Its interface is designed to operate in a similar way to the standard

ping(8) utility, even though the way it works is pretty different. 20

21

b) rds-ping opens several RDS sockets and sends packets to port 0 on the indicated host. This is a special port number to which no socket is 22

bound; instead, the kernel processes incoming packets and responds to 23

them. 24

5) Verify that all nodes respond without error. 25

**Note**: To avoid losing packets, do not run this while RDS-Stress is running. 26

**13.3.2 RDS-STRESS PROCEDURE** 27

1) Choose a host and start a passive receiving session for the RDS Stress test. 28

It only needs to be told what port to listen on. 29

a) $ rds-stress -p 4000 30

31

2) Chose a second host and start an active sending instance giving it the ad-

dress and port at which it will find a listening passive receiver. In addition, it 32

is given configuration options which both instances will use. 33

a) $ rds-stress -T 5 -s recvhost -p 4000 -t 1 -d 1 34

**Note**: If you repeat the test in less than one minute you may get the error 35 message "Cannot assign requested address" since the port numbers are 36 not immediately reusable. Either wait or change the port number using 37

the *-p* option 38

**Note**: The *-t* option is for the number of tasks (child processes), which de- 39 faults to 1 so "-t 1" is optional. The *-d* option is for the message queue 40 depth, which also defaults to 1 so "-d 1" is optional.

41

42

3) Every second, the parent process will display statistics of the ongoing stress 1 test. If the -T option is given, the test will terminate after the specified time 2 and a summary is printed. 3

4) Verify that the test completes without error. 4

5) Repeat steps 1-4 until all end points in the cluster have been tested. 5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13.4 TI UDAPLTEST COMMANDS USING OFED** 1

Server Command: dapltest -T S -D <ia\_name> 2

3

**13.4.1 SETUP** 4

• The /etc/dat.conf needs to be verified to be sure that the correct interface is 5 used. By default the dapl interface for IB is ib0 and for iWARP is eth2. If these 6 are not correct for the current cluster then errors will occur.

7

• It is also important to verify that the desired dapl library is being used. 8

• [For IB] an SM needs to be running. 9

10

• [For iWARP hosts with Chelsio RNICs] Ensure that

/sys/module/iw\_cxgb3/parameters/peer2peer contains '1' on all hosts. 11

12

**13.4.2 GROUP 1: POINT-TO-POINT TOPOLOGY** 13

[1.1] 1 connection and simple send/recv: 14

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -R BE 15

• client SR 256 1 server SR 256 1 16

[1.2] Verification, polling, and scatter gather list: 17

• dapltest -T T -s <sever\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R BE 18

• client SR 1024 3 -f \ 19

• server SR 1536 2 -f 20

**13.4.3 GROUP 2: SWITCHED TOPOLOGY** 21

**InfiniBand Switch**: Any InfiniBand switch 22

23

**iWARP Switch**: 10 GbE Switch 24

25

**RoCE Switch**: 10/40 GbE DCB Enabled switch 26

[2.1] Verification and private data: 27

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R BE 28

• client SR 1024 1 \ 29

• server SR 1024 1 30

[2.2] Add multiple endpoints, polling, and scatter gather list: 31

32

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 10 -V -P -R

33

• BE client SR 1024 3 \ 34

• server SR 1536 2 35

[2.3] Add RDMA Write : 36

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R BE 37

• client SR 256 1 \ 38

• server RW 4096 1 server SR 256 1 39

[2.4] Add RDMA Read: 40

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R BE 41

42

• client SR 256 1 \ 1

• server RR 4096 1 server SR 256 1 2

**13.4.4 GROUP 3: SWITCHED TOPOLOGY WITH MULTIPLE SWITCHES** 3

**Note**: This test is **not applicable to RoCE** for the October 2012 Events 4

5

[3.1] Multiple threads, RDMA Read, and RDMA Write: 6

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 4 -w 8 -V -P -R BE 7

• client SR 256 1 \ 8

• server RR 4096 1 server SR 256 1 client SR 256 1 server RR 4096 1 \ 9

• server SR 256 1 10

[3.2] Pipeline test with RDMA Write and scatter gather list: 11

• dapltest -T P -s <server\_name> -D <ia\_name> -i 1024 -p 64 -m p RW 12

8192 2 13

[3.3] Pipeline with RDMA Read: 14

• **InfiniBand**: dapltest -T P -s <server\_name> -D <ia\_name> -i 1024 -p 64 15

-m p RR 4096 2 16

• **iWARP**: dapltest -T P -s <server\_name> -D <ia\_name> -i 1024 -p 64 -m 17

p RR 4096 1 18

[3.4] Multiple switches: 19

• dapltest -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 10 -V -P -R 20

• BE client SR 1024 3 \ 21

• server SR 1536 2 22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13.5 TI RDMA BASIC INTEROP** 1

**13.5.1 Purpose** 2

To demonstrate the ability of endpoints to exchange core RDMA operations 3

across a simple network path. This test procedure validates the operation of end- 4

points at the RDMA level, in a simple network configuration. 5

6

The Basic RDMA interop test identifies interoperability issues in one of four ways:

7

• The inability to establish connections between endpoints 8

• The failure of RDMA operations to complete 9

• Incorrect data after the completion of RDMA exchanges 10

• Inconsistent performance levels.

**13.5.2 General Setup** 11

The RDMA interop procedure can be carried out using the OFA Verbs API to 12

create RDMA Connections and send RDMA operation. 13

14

**13.5.3 Topology** 15

The topology of the network that interconnects the switches can be changed to 16 validate operation of the endpoints over different networks paths. It is recom- 17 mended that this procedure first be executed between endpoints connected by a

single switch, and then the process repeated for more complex network configu- 18

rations. 19

20

**13.5.4 IB Setup** 21

Connect endpoints to switch and run one or more SMs (embedded in the switch 22

or host based). 23

**13.5.5 iWARP Setup** 24

Connect iWARP RDMA endpoints to an 10GbE switch. 25

26

**13.5.6 RoCE Setup** 27

Connect RoCE RCAs to a 10/40 GbE switch which is DCB Enabled. 28

29

**13.5.7 RDMA Connectivity Setup** 30

Each of the tests described below must be run twice with Host A being the server 31 and then Host B being the server. This ensures that the different semantics as- 32 sociated with active and passive sides of the connection are exercised. This way

each RDMA interface tested will be sending RDMA data (Requestor) in one test 33

and receiving RDMA data (Target) in the next. 34

35

**13.5.8 Small RDMA READ Procedure** 36

1) Select the two devices that will be tested: 37

2) On the server device issue the following command on command line: 38 a) [**For IB & RoCE**] ib\_read\_bw -d <dev\_name> -i <port> -m 2048 39 b) [**For iWARP**] - Not applicable - see 12.6.9 40

3) On the client device issue the following command on command line: 41

42

a) [**For IB & RoCE**] ib\_read\_bw -d <dev\_name> -i <port> -s 1 -n 25000 -m 1

2048 2

b) [**For iWARP**] - Not applicable - see 12.6.9 3

4) Verify that the operation completed without error and the level of perfor- 4

mance achieved is reasonable and as expected. 5

**13.5.9 Large RDMA READ Procedure** 6

1) Select the two devices that will be tested: 7

2) On the server device issue the following command on command line: 8

a) [**For IB & RoCE**] ib\_read\_bw -d <dev\_name> -i <port> -m 2048 9

10

b) [**For iWARP**] - Not applicable - see 12.6.10

11

3) On the client device issue the following command on command line: 12

a) [**For IB & RoCE**] ib\_read\_bw -d <dev\_name> -i <port>-s 1000000 -n 13

300 -m 2048 14

b) [**For iWARP**] - Not applicable - see 12.6.10 15

4) Verify that the operation completed without error and the level of perfor- 16

mance achieved is reasonable and as expected. 17

**13.5.10 Small RDMA Write Procedure** 18

1) Select the two devices that will be tested: 19

2) On the server device issue the following command on command line: 20 a) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -m 2048 21 b) [**For iWARP**] rdma\_bw -c -s 1 -n 25000 22

3) On the client device issue the following command on command line: 23

24

a) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -s 1 -n 25000 -

m 2048 25

26

b) [**For iWARP**] rdma\_bw -c -s 1 -n 25000 *RNIC\_IP\_Address*

27

4) Verify that the operation completed without error and the level of perfor- 28

mance achieved is reasonable and as expected.

29

**13.5.11 Large RDMA Write Procedure** 30

1) Select the two devices that will be tested: 31

2) On the server device issue the following command on command line: 32 a) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -m 2048 33 b) [**For iWARP**] rdma\_bw -c -s 1000000 -n 300 34

3) On the client device issue the following command on command line: 35

a) [**For IB & RoCE**] ib\_ write \_bw -d <dev\_name> -i <port>-s 1000000 -n 36

300 -m 2048 37

b) [**For iWARP**] rdma\_bw -c -s 1000000 -n 300 *RNIC\_IP\_Address* 38

4) Verify that the operation completed without error and the level of perfor- 39

mance achieved is reasonable and as expected. 40

41

42

**13.5.12 Small RDMA SEND Procedure** 1

This procedure may fail due to the inability of a endpoint to repost the consumed 2

buffers. 3

4

1) Select the two devices that will be tested:

5

2) On the server device issue the following command on command line: 6

a) [**For IB & RoCE**] ib\_ send \_bw -d <dev\_name> -i <port> -m 2048 7

b) [**For iWARP**] - Not applicable - see 12.6.9 8

3) On the client device issue the following command on command line: 9

a) [**For IB & RoCE**] ib\_writesend\_bw -d <dev\_name> -i <port> -s 1 -n 10

25000 -m 2048 11

b) [**For iWARP**] - Not applicable - see 12.6.9 12

4) Verify that the operation completed without error and the level of perfor- 13

mance achieved is reasonable and as expected. 14

**13.5.13 Large RDMA SEND Procedure** 15

This procedure may fail due to the inability of a endpoint to repost the consumed 16

buffers. 17

18

1) Select the two devices that will be tested: 19

2) On the server device issue the following command on command line: 20 a) [**For IB & RoCE**] ib\_ send \_bw -d <dev\_name> -i <port> -m 2048 21 b) [**For iWARP**] - Not applicable - see 12.6.10 22

3) On the client device issue the following command on command line: 23

a) [**For IB & RoCE**] ib\_ send \_bw -d <dev\_name> -i <port>-s 1000000 -n 24

300 -m 2048 25

b) [**For iWARP**] - Not applicable - see 12.6.10 26

4) Verify that the operation completed without error and the level of perfor- 27

mance achieved is reasonable and as expected. 28

**13.5.14 Additional IB Notes** 29

30

1) Alternate read commands available

31

a) Server command: ib\_read\_bw -m 2048 32

b) Client command (small): ib\_read\_bw -s 1 -n 25000 *IPoIB Address for* 33

*server* -m 2048 34

c) Client command (large): ib\_read\_bw -s 1000000 -n 300 *IPoIB Address* 35

*for server* -m 2048

36

2) Alternate write commands available 37

a) Server command: ib\_write\_bw -m 2048 38

b) Client command (small): ib\_write\_bw -s 1 -n 25000 *IPoIB Address for* 39

*server* 40

c) Client command (large): ib\_write\_bw -s 1000000 -n 300 *IPoIB Address* 41

*for server* -m 2048 42

3) Alternate send commands available 1 a) Server command: ib\_send\_bw -m 2048 2 b) Client command: ib\_send\_bw -s 1 -n 25000 *IPoIB Address for server* -m 3

2048 4

c) Client command (large): ib\_send\_bw -s 1000000 -n 300 *IPoIB Address* 5

*for server* -m 2048 6

4) Explanation of parameters 7 a) "-d" allows you to specify the device name which may be obtained from 8 the command lane: ***ibv\_devinfo*** 9

b) "-i" allows you to specify the port number. This may be useful if you are 10 running the tests consecutively because a port number is not immedi- 11 ately released and this will allow you to specify another port number to 12

run the test. 13

c) "-s" - this is the size of the operation you wish to complete 14

d) "-n" - this it the number of operations you wish to complete. 15

e) "-m" - this specifies the IB PMTU size. AS of 10/3/2011 some devices 16

did not support greater than 2048 17

**13.5.15 Additional iWARP Notes** 18

1) The "-c" option specifies to use the rdma\_cm for connection 19

20

**IB Example**: 21

**DevInfo - Server** 22

hca\_id: mthca0 23 fw\_ver: 1.2.0 24 node\_guid: 0002:c902:0020:b4dc 25

sys\_image\_guid: 0002:c902:0020:b4df 26

vendor\_id: 0x02c9

vendor\_part\_id: 25204 27

hw\_ver: 0xA0 28

board\_id: MT\_0230000001 29

phys\_port\_cnt: 1 30

port: 1

|  |  |  |
| --- | --- | --- |
| state: | PORT\_ACTIVE (4) | 31 |
| max\_mtu: | 2048 (4) | 32 |
| active\_mtu: | 2048 (4) | 33 |
| sm\_lid: | 1 | 34 |
| port\_lid: | 2 | 35 |

port\_lmc: 0x00

36

**Command Line**: ib\_read\_bw -d mthca0 -i 1 37

38

**DevInfo - Client**

hca\_id: mlx4\_0 39

fw\_ver: 2.2.238 40

node\_guid: 0002:c903:0000:1894 41

sys\_image\_guid: 0002:c903:0000:1897 42

vendor\_id: 0x02c9 1 vendor\_part\_id: 25418 2 hw\_ver: 0xA0

board\_id: MT\_04A0110002 3

phys\_port\_cnt: 2 4

port: 1 5 state: PORT\_ACTIVE (4) 6 max\_mtu: 2048 (4)

active\_mtu: 2048 (4) 7

sm\_lid: 1 8

port\_lid: 1 9

port\_lmc: 0x00 10

**Command Line**: ib\_send\_bw -d mlx4\_0 -i 1 10.0.0.1 -s 1 -n 300 11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13.6 TI RDMA STRESS TEST**

**13.6.1 Purpose**

**13.6.2 Topology**

**13.6.3 Switch Load**

1

2

This test is designed to identify problems that arise when RDMA operations are 3 performed over interconnection devices in the fabric. The test is not designed to 4 measure the forwarding rate or switching capacity of a device, but does use per- 5 formance measures to identify failures. 6

Test failures are identified by the following events: 7

8

• The inability to establish connections between endpoints 9

• The failure of RDMA operations to complete 10

• Incorrect data after the completion of RDMA exchanges

• Inconsistent performance levels. 11

12

13

14

This test does not define a detailed topology and can be used either on a single 15 switch or across a RDMA fabric that may include gateways to and from other 16 technologies. The test configuration depends on the number of endpoints avail-

able to perform the testing. 17

18

19

The switch load test validates proper operation of a switch when processing a 20 large number of small RDMA frames. This test is analogous to normal switch 21 testing. 22

1) Attach a device to each port on the switch. 23

2) Select two ports on the switch to test (This will be your control stream) 24

3) Generate RDMA WRITE Operations of size 1024 bytes 100, 000 times on 25 each device by issuing the following commands 26 a) On the server device issue the following command on command line: 27

28 i) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -m

2048 29

30 ii) [**For iWARP**] rdma\_bw -c -s 1024 -n 25000 31

b) On the client device issue the following command on command line: 32 i) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -s 33

1024 -n 25000 -m 2048 34

ii) [**For iWARP**] rdma\_bw -c -s 1024 -n 25000 *RNIC\_IP\_Address* 35

4) This must be done on both devices at the same time. 36

5) On all other pairs generate RDMA WRITE Operations of size 1 byte continu- 37

ously until the control stream completes. 38

6) Repeat above steps until all port pairs are tested. 39

40

41

42

**13.6.4 Switch FAN in**

7) Repeat the above steps with all endpoint pairs, except the control stream 1 changed such that the size of the RDMA WRITE operation is 1,000,000 2 bytes (~1 MB) 3

4

The switch fan in test attempts to validate proper operation of RDMA exchanges 5

in the presence of traffic loads that exceed the forwarding capacity of the switch. 6

The test requires a minimum of two switches that are interconnected by one port

pair. 7

8

1) Connect all possible endpoint pairs such that data exchanges between pairs 9 must traverse the pair of ports interconnecting the switch. The control con- 10 nections must be across the interconnect network.

11

2) Select two ports such that it has to cross both switches. (This will be your

control stream) 12

13

3) Generate RDMA WRITE Operations of size 1024 bytes 100, 000 times on

each device by issuing the following commands 14

a) On the server device issue the following command on command line: 15

16 i) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -m 17

2048

18

ii) [**For iWARP**] rdma\_bw -c -s 1024 -n 25000 19

b) On the client device issue the following command on command line: 20 i) [**For IB & RoCE**] ib\_write\_bw -d <dev\_name> -i <port> -s 21

1024 -n 25000 -m 2048 22

ii) [**For iWARP**] rdma\_bw -c -s 1024 -n 25000 *RNIC\_IP\_Address* 23

4) This must be done on both devices at the same time. 24

5) On all other pairs generate RDMA WRITE Operations of size 1 byte continu- 25

ously until the control stream completes. 26

6) Repeat above steps until all port pairs are tested. 27

7) Repeat the above steps with all endpoint pairs, except the control stream 28 changed such that the size of the RDMA WRITE operation is 1,000,000 29 bytes (~1 MB) 30

31

32

33

34

35

36

37

38

39

40

41

42

**13.7 TI MPI - OPEN MPI USING OFED** 1

The following values are used in examples below: 2

3

• $MPIHOME: The absolute directory location of the Open MPI installation 4

that is common to all systems under test. 5

• $NP: The number of MPI processes to use in the test. 6

• $HOSTFILE: The absolute filename location of the hostfile 7

• $IMBHOME: The absolute directory location of the Intel MPI Benchmark 8

(IMB) tools installation that is common to all systems under test. 9

**13.7.1 CLUSTER SETUP** 10

1) Network configuration requirements 11

a) All systems must be reachable by each other over IPoIB. 12

b) All nodes must agree on the IPoIB IP addresses of all systems (e.g., via 13

/etc/hosts, DNS, or some other mechanism). 14

2) The same version of OFED must be installed in the same filesystem location 15

on all systems under test. 16

3) The same version of the Intel MPI Benchmark (IMB) tools must be installed 17 in the same filesystem location on all systems under test. 18 a) IMB can be used from the OFED installation or, if a later version of Open 19

MPI is to be used, IMB can be downloaded from Intel's web site: 20

[http://software.intel.com/en-us/articles/intel-mpi-bench-](http://software.intel.com/en-us/articles/intel-mpi-benchmarks/?wapkw=intel%20mpi%20benchmarks) 21

[marks/?wapkw=intel%20mpi%20benchmarks](http://software.intel.com/en-us/articles/intel-mpi-benchmarks/?wapkw=intel%20mpi%20benchmarks) 22

4) The same version of Open MPI must be available in the same filesystem lo- 23

cation on all systems under test.

24

a) Open MPI can be used from the OFED installation, or, if a later version 25

is required, can be downloaded and installed from the main Open MPI

web site: 26

<http://www.open-mpi.org/>27

28

i) If building Open MPI from source, and if the OpenFabrics libraries

and headers are installed in a non-default location, be sure to use 29

the --with-openib=<dir> option to configure to specify the OpenFab- 30

rics filesystem location. 31

ii) Open MPI can be installed once on a shared network filesystem that 32 is available on all nodes, or can be individually installed on all sys- 33 tems. The main requirement is that Open MPI's filesystem location

is the same on all systems under test. 34

iii) If Open MPI is built from source, the --prefix value given to configure 35 should be the filesystem location that is common on all systems un- 36 der test. For example, if installing to a network filesystem on the file- 37

system server, be sure to specify the filesystem location under the 38 common mount point, not the "native" disk location that is only valid 39 on the file server.

40

41

42

**13.7.2 TEST SETUP**

iv) **Note** that Open MPI is included in some Linux distributions and oth- 1 er operating systems. Multiple versions of Open MPI can peacefully 2 co-exist on a system as long as they are installed into separate file-

system locations (i.e., configured with a different --prefix argument). 3

All MPI tests must be built and run with a single installation of Open 4

MPI. 5 v) Ensure that the Open MPI installation includes OpenFabrics sup- 6 port: 7 shell$ $MPIHOME/bin/ompi\_info | grep openib MCA btl: openib 8

(MCA v1.0, API v1.0.1, Component v1.4) 9

The exact version numbers displayed will vary depending on your 10

version of Open MPI. The important part is that a single "btl" line ap- pears showing the openib component. 11

b) Basic Open MPI run-time functionality can first be verified by running 12 simple non-MPI applications. This ensures that the test user's rsh 13 and/or ssh settings are correct, etc. 14

shell$ $MPIHOME/bin/mpirun -ssh -np $NP --hostfile $HOSTFILE host- 15 name 16 i) The output should show the hostname of each host listed in the 17 hostfile; the hostname should appear as many times as there are 18

lines in the hostfile. The list of hostnames may appear in random or- 19

der; this is normal

20

ii) Note that any serial application can be run; "hostname" is a good,

short test that clearly identifies that specific hosts were used 21

5) All systems must be setup with at least one identical user account. This user 22 must be able to SSH or RSH to all systems under test from the system that 23 will launch the Open MPI tests with no additional output to stdout or stderr 24

(e.g., all SSH host keys should already be cached, no password/passphrase 25

prompts should be emitted, etc.). 26

6) The lockable memory limits on each machine should be set to allow un- 27

limited locked memory per process. 28

7) The underlying OpenFabrics network used in the test should be stable and 29

reliable.

30

8) No other fabric interoperability tests should be running during the Open MPI 31

tests.

32

9) MPI tests should be run across at least 5 separate systems to force the use

of the OpenFabrics network (vs. using just shared memory for in-system 33

communication). 34

35

1) Create a hostfile ($HOSTFILE) listing the hostname of each system that will 36 be used in the test. If a system under test can run more than one MPI pro- 37 cess (such as multiprocessor or multicore systems), list the hostname as 38

many times as MPI processes are desired. For example, for two systems 39 named node1.example.com and node2.example.com that are each able to 40 run 4 processes:

41

shell$ cat hostfile.txt 42

**13.7.3 TEST PROCEDURE**

node1.example.com 1 node1.example.com 2 node1.example.com 3

node1.example.com 4

5

node2.example.com

6

node2.example.com 7

node2.example.com 8

node2.example.com 9

10

2) Determine the number of Open MPI processes ($NP) that are to be run de- 11

termined by the number of host entries in the created hostfile. 12

3) Open MPI defaults to probing all available networks at run-time to determine 13

which to use. OpenFabrics testing must specifically force Open MPI to 14

\*only\* use its OpenFabrics stack for testing purposes (e.g., do not fail over 15 to TCP if the OpenFabrics stack is unavailable). To do this add an extra command line parameter; both iWarp and InfiniBand: 16

--mca btl openib,self 17

18

4) It has been discovered that the following Open MPI command line pa-

rameter is required to facilitate multi RDMA adaptor vendor MPI rings; both 19

iWarp and InfiniBand: 20

--mca pml ob1 --mca btl\_openib\_flags 306 21

5) It has been discovered that the following Open MPI command line pa- 22 rameter is required to facilitate multi RNIC adaptor vendors MPI rings; iWarp 23 specific: 24

--mca btl\_openib\_receive\_queues P,65536,256,192,128 25

26

1) Create a hostfile listing the MPI ring nodes, process distribution, and total 27

number of processes to use as indicated in steps 1 and 2 of section 12.11.2. 28

The filesystem location of this hostfile is irrelevant. 29

2) Locate the "mpirun" binary that will be used. This determines the version of 30

Open MPI that will be used. 31

3) Locate the "IMB-MPI1" IMB binary. This must have been built against the 32 version of Open MPI selected above. If using an OFED distribution this build 33 process has already been performed. 34

4) Verify that a subnet manager has configured the fabric. If not, start one. 35

5) Verify that all hosts present within the hostfile are online and accessible. 36

6) Run the IMB-MPI1 benchmarks 37

7) Repeat steps 4-6 using a different subnet manager until all subnet man- 38

agers under test have been used. 39

8) All IMB benchmarks must pass successfully using all subnet managers 40

under test in order for the devices under test defined within the hostfile pass. 41

42

**13.7.4 METHOD OF IMPLEMENTATION FOR ALL LINUX OS'S** 1

1) To perform step 4 of section 12.9.3 use "ibdiagnet -r" from a host defined in 2

the mpi hostfile and look for an "SM - Master" entry in the output 3

2) To perform step 5 of section 12.9.3 ping the IPoIB address of all hosts de- 4

fined in the mpi hostfile from a host defined in said hostfile. 5

3) To perform step 6 of section 12.9.3 use the following command from a host 6 that can access all hosts defined within the hostfile; this host can be part of 7 the hostfile 8

a) For **InfiniBand & RoCE**: 9

$MPIHOME/bin/mpirun --mca btl openib,self,sm --mca pml ob1 -mca 10

btl\_openib\_flags \ 306 -np $NP -hostfile $HOSTFILE $IMBHOME/IMB- 11

MPI1

12

a) For **iWarp**:

13

$MPIHOME/bin/mpirun --mca btl openib,self,sm --mca pml ob1 --mca \ 14 btl\_openib\_flags 306 --mca btl\_openib\_receive\_queues P,65536,256,192,128 -np \ $NP -hostfile $HOSTFILE $IMBHOME/IMB- 15

MPI1 16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**13.8 TI MPI - OHIO STATE UNIVERSITY USING OFED** 1

**13.8.1 MVAPICH - SETUP** 2

1) Network configuration requirements 3 a) All systems must be reachable by each other a common network that 4 supports TCP (Ethernet, IPoIB, etc.) 5

b) All nodes must agree on the IP addresses for all TCP networks on all 6

systems (e.g., via /etc/hosts, DNS, or some other mechanism). 7

2) The same version of OFED must be installed in the same filesystem location 8

on all systems under test. 9

3) MVAPICH is included in OFED distributions. The updated versions of 10

MVAPICH can be obtained from OpenFabrics website. 11

4) The same version of MVAPICH must be available in the same filesystem lo- 12 cation on all systems under test. 13 a) MVAPICH can be installed once on a shared network filesystem that is 14 available on all nodes, or can be individually installed on all systems. 15

The main requirement is that MVAPICH filesystem location is the same 16

on all systems under test.

17

5) All systems must be setup with at least one identical user account. This user 18

must be able to SSH or RSH to all systems under test from the system that

will launch the MVAPICH tests with no additional output to stdout or stderr 19

(e.g., all SSH host keys should already be cached, no password/passphrase 20

prompts should be emitted, etc.). 21

6) The lockable memory limits on each machine should be set to allow un- 22 limited locked memory per process. This can be achieved by using ulimit 23 command. 24

7) The underlying IB network(s) used in the test should be stable and reliable. 25

No other fabric interoperability tests should be running during the MVAPICH

tests. 26

8) Multiple versions of MVAPICH can peacefully co-exist on a system as long 27 as they are installed into separate filesystem locations (i.e., configured with 28 a different --prefix argument). All tests must be built and run with a single in- 29

stallation of MVAPICH. 30

9) MVAPICH tests should be run across at least 5 separate systems to force 31 the use of the IB networks (vs. using just shared memory for in-system com- 32 munication). 33

**Note**: MVAPICH is commonly referred to as MVAPICH1 to distinguish it from 34

the new and updated MVAPICH2

35

**13.8.2 MVAPICH - TEST SETUP AND PROCEDURE** 36

1) Test Setup 37

a) Create a hostfile listing the hostname of each system that will be used in 38 the test. If a system under test can run more than one MPI process 39 (such as multiprocessor or multicore systems) list the hostname as

many times as MPI processes are desired. For example, for two 2 pro- 40

cessor systems named host1 and host2 41

42

$ cat hostfile.txt 1

host1 2

host1

host2 3

host2 4

b) Download and install Intel® MPI Benchmarks on all nodes from: 5 [http://www.intel.com/cd/software/products/asmo-](http://www.intel.com/cd/software/products/asmo-na/eng/cluster/mpi/219848.htm) 6 [na/eng/cluster/mpi/219848.htm](http://www.intel.com/cd/software/products/asmo-na/eng/cluster/mpi/219848.htm) 7

Follow the instructions below to install: 8

i) untar downloaded archive 9

ii) open <natured directory>/src/make\_mpich and fill in the following 10

variables: 11

• MPI\_HOME=<path to mvapich1 directory> #mine was 12

/usr/mpi/gcc/mvapich-1.0.1 13

• CPPFLAGS= -DCHECK 14

iii) gmake -f make\_mpich 15

This will install the benchmarks inside the MPI\_HOME/tests directory 16

**Note**: Intel® MPI Benchmarks are installed with OFED installation by de- 17

fault 18

c) Enter all nodes and run the following commands: 19 i) echo "PATH=\$PATH:<path to mvapich1 directory>/bin:<path to 20 mvapich1 directory>/tests/IMB-3.0" >> /<username>/.bashrc # or 21

.cshrc 22 ii) echo "ulimit -l unlimited" >> /<username>/.bashrc # or .cshrc 23 iii) source /<username>/.bashrc # or .cshrc 24

**Note**: these commands may fail or produce unexpected results with a 25

shared $HOME 26

2) Testing Procedure 27 a) The following values are used in the examples below 28 i) $MPIHOME - The absolute directory location of the MVAPICH in- 29

stallation that is common to all systems under test 30

ii) $NP - The number of MPI processes that are to be run determined 31

by the number of host entries in the created hostfile. 32

iii) $HOSTFILE - The absolute location of the hostfile 33 b) Run Intel® MPI Benchmarks: 34 i) Run the PingPong and PingPing point-to-point tests 35

$MPIHOME/bin/mpirun\_rsh -ssh -np $NP IMB-MPI1 -multi 0 Ping- 36

Pong PingPing -hostfile $HOSTFILE 37

ii) Run all the tests (PingPong, PingPing, Sendrecv, Exchange, Bcast, 38

Allgather, Allgatherv, Alltoall, Reduce, Reduce\_scatter, Allreduce, 39

Barrier), in non-multi mode. 40

$MPIHOME/bin/mpirun\_rsh -ssh -np $NP IMB-MPI1 -multi 0 -hostfile 41

$HOSTFILE 42

**13.8.3 MVAPICH2 - SETUP**

1

1) Download and install OFED on all nodes from: 2

<http://www.openfabrics.org/downloads/OFED>3

2) Download and install Intel® MPI Benchmarks on all nodes from: 4

5 [http://www.intel.com/cd/software/products/asmo- na/eng/cluster/mpi/219848.htm](http://www.intel.com/cd/software/products/asmo-na/eng/cluster/mpi/219848.htm) 6

You will have to accept a license. Follow the instructions below to install. 7

8

a) untar downloaded archive 9

b) open <untarred directory>/src/make\_mpich and fill in the following vari- 10

ables:

11

i) MPI\_HOME=<path to mvapich2 directory> #mine was 12

/usr/mpi/gcc/mvapich2-1.0.3

13

ii) CPPFLAGS= -DCHECK

14

c) gmake -f make\_mpich 15

This will install the benchmarks inside the MPI\_HOME/tests directory 16

3) All nodes should be physically connected. 17

4) Enter all nodes and run the following cmds: 18 a) echo "PATH=\$PATH:<path to mvapich2 directory>/bin:<path to 19 mvapich2 directory>/tests/IMB-3.0" >> /<username>/.bashrc # or .cshrc 20

b) echo "ulimit -l unlimited" >> /<username>/.bashrc; 21

c) source /<username>/.bashrc # or .cshrc 22

5) Create an mpi ring: 23 a) Construct a file called hosts that has the following format. Include as 24 many lines as you have hosts. Be sure to leave a blank line at the end of 25

the file: 26

i) <host>ifhn=<infiniband ip address> 27 b) Run the following commands 28 i) mpdboot -n `cat hosts|wc -l` -f hosts --ifhn=<localhost infiniband ip 29 address> 30

ii) mpdtrace -l #OPTIONAL, shows current ring members. 31

6) MVAPICH tests should be run across at least 5 separate systems to force 32

the use of the IB networks (vs. using just shared memory for in-system com- 33

munication). 34

**13.8.4 MVAPICH2 - TEST PROCEDURE** 35

**Step A:** [**For IB**] Run a subnet manager from one node only. 36

37

**Step B** Run Intel® MPI Benchmarks: 38

1) Two sets of tests should be run, with these command lines 39

40

[**For IB**]

41

42

a) mpirun\_rsh -ssh -np <number of nodes X number of processors/node> 1

IMB-MPI1 -multi 0 PingPong PingPing 2

b) mpirun\_rsh -ssh -np <number of nodes X number of processors/node> 3

IMB-MPI1 4

[**For iWARP**] 5

a) mpirun\_rsh -ssh -np <number of nodes X number of processors/node> 6

MV2\_USE\_IWARP\_MODE=1 MV2\_USE\_RDMA\_CM=1 IMB-MPI1 - 7

multi 0 PingPong PingPing 8

b) mpirun\_rsh -ssh -np <number of nodes X number of processors/node> 9

MV2\_USE\_IWARP\_MODE=1 MV2\_USE\_RDMA\_CM=1 IMB-MPI1

10

The first command runs just the PingPong and PingPing point-to-point tests, 11

but makes all tasks active (pairwise).

12

The second command runs all the tests (PingPong, PingPing, Sendrecv, Ex- change, Bcast, Allgather, Allgatherv, Alltoall, Reduce, Reduce\_scatter, Allre- 13

duce, Barrier), in non-multi mode. 14

2) [**For IB**] If the test passes shutdown current subnet manager and start an- 15

other one on a different node; run both tests again. 16

3) [**For IB**] Repeat until all nodes have run a subnet manager and passed all 17

tests. 18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**14 INFINIBAND SPECIFIC INTEROP PROCEDURES USING WINOF** 1

**14.1 IB LINK INITIALIZE USING WINOF** 2

3

**14.1.1 Setup** 4

**Note**: The WinOF Subnet Manager and diagnostics are still evolving as com- 5 pared to OFED. Therefore, you must include an OFED Linux node along with the 6

Win 7

8

1) Disconnect the full topology and select a cable whose length should be a 9 maximum of 15 meters for SDR and 10 meters for DDR when using copper 10 cables.OF node to run diagnostics for this test.

11

2) Verify that no SM is running 12

3) Connect two devices back to back 13

4) ssh to the OFED node. 14 a) Run "ibdiagnet -lw 4x" to verify portwidth 15

16

b) Run "ibdiagnet -ls 2.5" to check link speed. Interpret output and com- 17 pare to advertised speed. 18

**Note**: This command will only produce output if the link speed is anything 19 other than SDR. Keep this in mind during your interpretation of the output. 20

5) Repeat steps 1-3 with a different device pairing. 21 a) All device pairs must be tested except SRP target to SRP target. 22

23

i) HCA to HCA 24

ii) HCA to Switch 25

iii) HCA to Target 26

27

iv) Switch to Switch 28

v) Switch to Target 29

30

**Note**: HCA to Target and HCA to HCA cannot be tested under 31

WinOF 2.0.2 because there are no utilities available. Switches can 32 be tested by using a Linux Host and the OFED Utilities.

33

b) Each device must link to all other devices in order for the device to pass 34 link init over all. 35

|  |  |
| --- | --- |
|  | 36 |
| In order to determine Switch to Target and Switch to Switch link parameters, run | 37 |
| commands from an HCA linked to the switch under test. This does require more | 38 |
| interpretation of the output to differentiate the reported parameters. | 39 |
|  | 40 |
|  | 41 |
|  | 42 |
|  | 28 |
|  | 29 |
|  | 30 |
|  | 31 |

**14.1.2 Recommendations**

**14.2 IB FABRIC INITIALIZATION USING WINOF** 1

**14.2.1 Architect the Network we want to build.** 2

**Note**: The WinOF Subnet Manager and diagnostics are still evolving as com- 3

pared to OFED. Therefore, you must include an OFED Linux node along with the 4

WinOF node to run diagnostics for this test. 5

6

1) Design and implement a Cluster Topology.

2) End to end IPoIB connectivity is required between all end points. Therefore 7

you must create and assign IP addresses to each IB end point. 8

3) See Figure 5- Sample Network Configuration below. 9

10

**14.2.2 Procedure** 11

1) Connect the HCAs and switches as per the Architected Network and make 12

sure that no SM/SA is running on the Fabric. 13

2) Start an SM on a device and let it initialize (all SMs will need to be tested)

3) Visually verify that all devices are in the active state using LEDs (however 14

the vendor decided to implement it). 15

4) The following step s must be done using a Linux OFED end point. 16 a) Run "ibdiagnet -pc" to clear all port counters 17 b) Wait 17 seconds as per the specifications requirements. 18

c) Run "ibdiagnet -c 1000" to send 1000 node descriptions. 19

d) Run "ibdiagnet" to generate fabric report and open report to see results. 20

/tmp/ibdiagnet.sm 21

e) Run "ibchecknet" to build guid list. 22

23

**14.2.3 Verification Procedures** 24

1) Review "PM Counters" section of the fabric report. There should be no il- 25 legal PM counters. The Specification says there should be no errors in 17 26 seconds. 27

2) Review "Subnet Manager " section of the fabric report. Verify that the

running SM is the one you started and verify number of nodes and switches 28

in the fabric. 29

3) Review the ibchecknet report and verify that there are no duplicate GUIDs in 30

the fabric 31

**Note**: the reports are located in the /tmp directory 32

Restart all devices in the fabric and follow Sections 13.2.2 and 13.2.3. Run the 33

SM from a different device in the fabric until all SMs present have been used. All

SMs on managed switches and one instance of **opensm** must be used. 34

35

Each device must pass all verification procedures with every SM to pass Fabric 36

Initialization test. 37

38

39

40

41

42

1

2

**Table 28 - ibdiagnet commands** 3

|  |  |
| --- | --- |
| **Commands** | **Description** |
| Ibdiagnet -c 1000 | send 1000 Node Descriptions |
| ibdiagnet -h | Help |
| Ibdiagnet -lw 4x - ls 2.5 | Specify link width and speed |
| Ibdiagnet - pc | Clear Counter |
| ibdiagnet -t <file> | Compare current topology to saved topology |
| ibdiagnet -wt | Writes the topology to a file |

4

5

6

7

8

9

10

11

12

13

14

15

16

**Note**: The topology file is being generated after the SM starts but before any 17 testing has started. The topology comparison is being performed after testing has 18 been completed but before the systems get rebooted. A topology check is per-

formed during every part of every test section that does not specifically state 19

"change the topology". For example Fabric Init only has 1 part so there is only 1 20

check but RDS has 2 parts so 2 checks are performed. However, IPoIB has 3 21 parts for each of 2 modes but 1 of those parts specifically says to change the to- 22 pology so only 4 checks occur. 23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**Figure 5 - Sampie Network Configuration** 1

2

Ethernet Addressing

IP Address 172.16.xxx.xxx

Netmask 255.255.0.0

IPolB Addressing

IP Address tO.t O.xxx.xxx

Netmask 255.0.0.0

172.16.30.20

10 103020

172.16.20.24 172.16.20.20 172.16.20.3 172.16.130.2

10 10 20 24 10 10 20 20 10 10 20 3 10 10 100 2

**IYIC'IIanox IYIC'II800X lvtt:ll8flOX** Qloglc

|  |  |  |
| --- | --- | --- |
| '""''"'"ox  Connect)(  QDR Mlcroglga | |  |
| Farbau | d IG1:i |
|  |  | |
|  | |

Connect)( **ConnextX** Cougar

ee

DDR SDR

I Ska

Atlas Janus Skoll

h . I 0

1

172.16.20.6 172.16.20.7 172.16.20.25 172.16.20.21 2

10.10.20.6

10.10.20.7

LSIDDR

6 5 11 10 2 1 10.10.20.25 10.10.20.21 3

-· **hli-llbb 172. 18.90.1** Dione(G21 Mimas 5

**IYiena.nox Mellanox** SRP Target DON DDR

**MiniSDR lsix·bb2!·op** r- 1-

"'euanox **Mellanox**

Cheelah Mellanox Shark SRP Target ConnextX **LionMini** 4

SOR 36 Port QDR DDR G2 DDR

**l r 172.16.50.4**

24 13 12 I 6

Tilan

**172.16.50.3 <lctl-&289000**

172.16.20.22 7

10.10.20.22

6 *7* 8 **MeUanox** 8

**Flextronlcs**

CUb DDR - 9

8 Port SDR **Pandora** 0

2 1 -

2 1 -

**Qlogie 24 Port** 1

**J 9024- DOR**

-

+

2

-

5 4

3 2 1

2 1

**Flextronics**

**24Port 0DR**

5 4

I

**ql9024 172.16.30.24**

"" 3

4

**QLogle 40Pori**

**904000R**

I **ql9040 172. 16.30.23** 5

LSISDR 6 5 4 I 1

|  |  |  |
| --- | --- | --- |
|  | | R |
| '"""•nux  **ConnextX**  I  Hvoerion | o..oogiC  SDR Miranda | 3 |

**172.16.50.1** -----1 SDR

**ll!JK· hbt.Jop** 1 -1 I I

**19 xbb1-bottom**

|  |  |  |
| --- | --- | --- |
| ':'0510180  Cafri'US  **¢bfKafi'4)Uf>1**  **.n.16.7<ll** | |  |
|  |  | |
|  |  |  |
| **Melianox**  Cub SDR  Tarvoe | |
|  | |  |

- ... c:U o

**1721650.2**

**vluaan** I

**Campus** I I

**ohs-eamp. d** I I

Cheetah

172.16.130.4 172.16.30.1 9

**172.16.70.4**

•I I NarviIG2l Telesto

I DDR 10.10.100.1 10.10.30.1 0

I

**Fibre** I

**Channel** I

172.16.30.21 172.16.20.23 1

10.10.30.21 10.10.20.23 2

Switch ·I -- **v•u•an \,,lr,SlUI3n** 3

- **1-** 4

I LongbowXR Longbow XR

|  |  |  |
| --- | --- | --- |
|  |  | |
| **Fibre Channel** Disk Array | |  |

**<:'.b$.1r,nglxwf.xr1 ob5-klngi:<Jw..lo"2**

172.16.20.4

+

Gatewaylo

**172.16.70.1 172. 16.70.2**

5

*;*

10.1020.4

**Ethernet** 6

**Fabric Mellanox**

7

TigerSDR

Calypso 8

December 2008 Topology 172.16.20.5 9

10.10.20.5 0

41

42

**14.3 IB IPOIB DATAGRAM MODE (DM) USING WINOF** 1

**14.3.1 SETUP** 2

**Note**: WinOF 2.0.2 only supports IPoIB Datagram Mode. Future WinOF releases 3

will support IPoIB Connected-Mode. 4

5

Connect the HCAs and switches as per the Architected Network and make sure 6

that no SM is running on the Fabric.

7

This procedure, as the previous ones, will be based on the cluster connectivity. 8

An SM/SA which supports IPoIB (sufficient IB multicast support) will be running 9 on the HCAs, or on a switch with an embedded SM/SA or a third HCA which 10 would only run SM/SA for the partner pair (with a switch in the middle). This pro-

cedure has been developed for the Windows environment. 11

12

**Optional**: In the procedures below, an IB analyzer can be inserted in the appro- 13 priate link to obtain traces and validate the aspects of the procedures specifically 14 detailed below in subsequent sections.

15

**14.3.2 IPOIB INTERFACE CREATION AND IPOIB SUBNET CREATION** 16

1) Configure IPoIB address. All addresses must reside on the same subnet. 17

2) Verify which 'Local Area Connection' the IPoIB interfaces are bound to: 18

19

a) Start | Server Manager | View Network Connections.

20

b) Find the OpenFabrics IPoIB interfaces (one per HCA port). If your plat- 21

form has two Ethernet ports, then IPoIB interfaces likely will be assigned

'**Local Area Connection 3**' & '**Local Area Connection 4**' as the Ether- 22

net ports are assigned '**Local Area Connection**' and '**Local Area Con-** 23

**nection 2**' . 24

3) Set interfaces to 10.0.0.x/24 (10.0.0.x/netmask 255.255.255.0) using the fol- 25 lowing commands: 26 a) netsh interface ip set address "Local Area Connection 3" static 27

10.10.4.x 255.255.255.0 28

b) netsh interface ip set address "Local Area Connection 4" static 29

10.10.4.y 255.255.255.0 30

4) View the IPoIB IP address using the following command 31

a) netsh interface ip show address "Local Area Connection 3" 32

**14.3.3 PING PROCEDURES** 33

**Step A** 1) Stop all SM's and verify that none are running 34

2) Power cycle all switches in the fabric (this insures that the new SM will con- 35

figure all the links and create the multi-cast join). 36

3) Start an SM (All SM's will need to be tested) and let it initialize 37

**Note**: For link testing it is recommended to use an OFED Linux OpenSM as 38 the Windows version of OpenSM does not support all SA queries and func- 39 tionality of the OFED 1.4 OpenSM. 40

41

42

**Note**: All WinOF installed systems contain a disabled OpenSM windows ser- 1

vice. A WinOF installation option/feature is to automatically 'start/enable' the 2

OpenSM service on the local node. 3

• Start | Server Manager | Configuration | Services | InfiniBand Subnet 4

Manager | Automatic | apply

5

• Start | Apply will enable the local OpenSM to start and be started upon system boot. 6

a) Visually verify that all devices are in the active state. Orange led will be 7

on if the port is active. 8

b) From a Linux system, Run "ibdiagnet" and verify that the SM you started 9

is the one that is running and and that it is the master. You will need to 10

know the GUID of the device since the SM will be reassigned on each 11

reboot; the Windows 'vstat' command displays HCA info. 12

c) Verify that all nodes and switches were discovered. 13

d) WinOF 2.0.2 does not provide a ibdiagnet utility. 14

**Note**: Ibdiagnet may show more switches than indicated by the physical 15 number of switch platforms present. This is because some switches have 16 multiple switch chips. 17

4) Examine the arp table (via arp -a) and remove the destination node’s ib0 ad- 18

dress from the sending node’s arp table (via arp -d). 19

5) Issue the command: sysctl net.ipv4.neigh.ib0.unres\_qlen=18 20

a) This sets the qlen variable to 18 which increases the buffer size so that 21 you do not get an initial dropped packet when using ping sizes 8192 and 22 greater. 23

6) Ping every IPoIB interface IPv4 address except localhost with packet sizes 24

of 511, 1025, 2044, 8192, 32768 and 65500. 'ping /?' displays ping help.

25

a) 10 packets of each size will be sent 26

b) Every packet size is a new ping command. 27

**Note**: Windows does not support 65507 so we used 65500. 28

**Note**: This is done from the Head Node utility "Run a Command" using 29 the following command: 30 for %i in (64, 511, 2044, 8192, 32768 and 65500) DO %d arp -d %d & 31

ping -i 1 -n 10 -l %i %d & arp -d %d 32

7) In order to pass Step A, a reply must be received for every ping sent (without 33 losing a single packet) while using each one of the SMs available in the 34 cluster.

35

36

**Step B** 1) Bring up all HCAs but one. 37

2) Start an SM (all SMs will need to be tested). 38

3) Check for ping response between all node (All to All). 39

a) A response from the disconnected HCA should not be returned. 40

4) Disconnect one more HCA from the cluster. 41

42

5) Ping to the newly disconnected HCA from all nodes (No response should be 1

returned). 2

6) Connect the first machine (the one that was not connected) and check for 3

ping response from all nodes that are still connected. 4

7) Connect the disconnected HCA to a different switch on the subnet which will 5

change the topology. 6

8) Ping again from all nodes (this time we should get a response). 7

9) Follow Step B, this time bring the interface down and then back up: Start | 8

Server Manager | View Network Connections | IPoIB(Local Area connection) 9

disable and enable commands instead of physically disconnecting the

HCAs. 10

**Note**: Each step must exhibit the expected behavior while using each SM in 11

order for the device to pass Step B overall. 12

13

14

**Step C** 1) Follow Step A and B using a different SM until all SM's have been used.

Only one instance of each available SM is required. Steps A, B, and C must 15

pass in order for the device to pass 13.3.3 overall. 16

2) Issue the command: sysctl net.ipv4.neigh.ib0.unres\_qlen=3 17

a) This sets the qlen variable back to the default. 18

19

**14.3.4 FTP PROCEDURE** 20

FTP procedures requires an FTP server to be configured on each machine in the 21

partner pair. An FTP client needs to be available on each machine as well; an

FTP client is a standard Windows component. 22

23

An FTP server is a component of the IIS '**Internet Information Services**' manger 24

which **not** a part of a standard Windows installation: 25

See Start | Server Manager | Roles | Add IIS. Configure FTP server via IIS man- 26

ager. 27

28

**14.3.4.1 SETUP** 29

1) Make sure ftpd is installed on each node for the FTP application. 30

2) A special account for this should be created as follows: 31 b) Username: Interop 32 c) Password: openfabrics 33

34

**14.3.4.2 PROCEDURE** 35

Run FTP server on all nodes. 36

37

1) Start an SM (all SMs will need to be tested) and let it initialize (ref MS Net-

work utilities docs) 38

39

a) Verify that the running SM is the one you started.

40

2) FTP: 41

42

a) Connect an HCA pair via FTP on IPoIB using the specified user name 1

and password. 2

b) Put the 4MB file to the %windir%\temp folder (generally C:\Win- 3

dows\Temp) on the remote host. 4

c) Get the same file to your local dir again. 5

d) Binary compare the file using the Windows command 'fc /B tfile 6 tfile.orig'. 7 i) The two must be identical 8

3) Repeat the procedure with a different SM. 9

**Note**: Every node must FTP the 4MB file to all others using all SMs and the 10 files must be identical as determined by the binary compare in order for the 11 device to pass 13.3.4 overall. 12

**Note**: Sections 13.3.3 and 13.3.4 must pass using the configuration deter- 13 mined by sections 13.3.1 and 13.3.2 for the device to pass IPoIB Datagram 14 mode overall.

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**14.4 IB SM FAILOVER AND HANDOVER PROCEDURE USING WINOF** 1

**14.4.1 SETUP** 2

1) Connect HCAs per the selected topology. 3

2) In this test, all active SMs on the fabric which are going to be tested, must 4

be from the same vendor. They will be tested pairwise: two at a time. 5

**14.4.2 PROCEDURE** 6

7

1) Disable all SMs in the cluster.

8

2) Start a SM on either machine in a chosen pair. 9

a) Start | Server Manager | Configuration | Services | InfiniBand Subnet 10

Manager | start | apply 11

3) Run "vstat" on all Windows nodes in the fabric. 12

a) Verify HCA link active in vstat output. 13

4) Verify IPoIB is active on each node 14 a) Verify Local Area Connection assigned to IPoIB interface: 15 i) Start | Control Panel | Network and Sharing Center | Manage Net- 16

work Connections. 17

b) Show IPv4 address assigned to IPoIB Interface(s): 18 i) netsh interface ip show address "Local Area Connection 3" 19 ii) netsh interface ip show address "Local Area Connection 4" 20

c) Verify the IPoIB devices (one per cabled connected HCA port) are visi- 21

ble & operational from a device driver perspective using Device Manag- 22

er 23

i) Start | Run | devmgmt.msc 24 d) Ping the IPoIB interface IPv4 address local and remote, verify traffic is 25 actually going in/out over IPoIB 'local area connection x'. 26

5) Start an Open SM on the second machine in the current pair. 27

6) Verify that the SMs behave according to the SM priority rules. 28

a) The Windows OpenSM log file is located at '%windir%\temp\osm.log'. 29

30

**Note**: The SM with highest numerical priority value is master and the

other is in standby.If both SMs have the same priority value then the SM 31

with the smallest guid is master and the other is in standby. 32

33

7) Verify that all nodes in the cluster are present - ping all IPoIB interfaces

34

8) Shutdown the master SM. 35

9) Verify the other active SM goes into the master state: see osm.log file. 36

10) Verify that all nodes in the cluster are present - ping all IPoIB interfaces 37

11) Start the SM you just shutdown. 38

12) Verify that the newly started SM resumes it's position as master while the 39

other goes into standby again; see '%windir%\temp\osm.log'. 40

13) Verify that all nodes in the cluster are present - ping all IPoIB interfaces 41

42

14) Shutdown the standby SM. 1

15) Verify that the previous master SM is still the master; view 2

'%windir%\temp\osm.log'. 3

16) Verify that all nodes in the cluster are present - ping all IPoIB interfaces 4

17) Repeat proceeding steps [1-16] 2 more times with the same node pair, en- 5 suring that the below criteria is met (total of 3 tests per pair which can be run 6 in any order): 7 a) First SM to be started having highest numerical priority value. 8

b) Second SM to be started having highest numerical priority value. 9

c) Both SMs having equal numerical priority values. 10

18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the 11

cluster have been tested. 12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**14.5 IB SRP USING WINOF**

**14.5.1 SETUP**

**14.5.2 WINDOWS PROCEDURE**

1

2

1) Connect the HCAs and switches as per the Architected Network and make 3

sure that no SM is running on the Fabric. 4

2) [Configure and Start a Linux OFED SRP target - VDISK BLOCKIO mode;](https://wiki.openfabrics.org/tiki-index.php?page=SRPT%2BInstallation) 5 [(some assembly required) -](https://wiki.openfabrics.org/tiki-index.php?page=SRPT%2BInstallation) https://wiki.openfabrics.org/tiki- 6 index.php?page=SRPT+Installation 7

a) assume /dev/sdb1 & /dev/sdc1 are formatted with /sbin/mkfs.msdos 8

b) Setting SRPT\_LOAD=yes in /etc/infiniband/openib.conf is not good 9 enough. It only loads ib\_srpt module and does not load scst and its 10 dev\_handlers. 11

c) modprobe scst 12 d) modprobe scst\_vdisk 13 e) echo "open vdisk0 /dev/sdb BLOCKIO" > /proc/scsi\_tgt/vdisk/vdisk 14 f) echo "open vdisk1 /dev/sdc BLOCKIO" > /proc/scsi\_tgt/vdisk/vdisk 15

g) echo "add vdisk0 0" >/proc/scsi\_tgt/groups/Default/devices 16

17

h) echo "add vdisk1 1" >/proc/scsi\_tgt/groups/Default/devices

18

**Note**: For the April 2012 Interop events, the OFILG decided that each target 19 should only advertise two volumes in order to reduce the amount of time required 20 to run the tests 21

22

23

1) Start an SM (all SM's will need to be tested) and let it initialize. 24

a) Verify that the running SM is the one that you started 25

2) Choose a node to work with 26

3) Verify the SRP driver loaded correctly; locate the SRP Miniport. 27

a) Start |Control Panel | Device Manager | Storage Controllers [InfiniBand 28

SRP Miniport] 29

4) Discover + Enable (bring online) the SRP drive(s) 30

a) Start | Server Manager | Storage | Disk Management 31

32

5) You will find a basic ‘unknown’ and ‘offline’ disk; this one of your SRP

volume(s). 33

6) Right-click the offline disk and select ‘online’. 34

35

7) Right-click the volume space, assign the drive letter ‘T’. 36

8) Right-click the volume space, format the volume. 37

9) Access the SRP drive via assigned drive letter. From a Windows/DOS 38 command prompt window, execute the following commands. 39 a) vol T: 40

b) dir T:\ (should be empty) 41

42

c) mkdir T:\tmp 1 d) copy /B WinOF\_wlh\_x64.msi T:\tmp 2 e) fc /B WinOF\_wlh\_x64.msi T:\tmp\WinOF\_wlh\_x64.msi 3

f) copy /B T:\tmp\WinOF\_wlh\_x64.msi T:\tmp\WOF2.msi 4

5

g) fc /B T:\tmp\WinOF\_wlh\_x64.msi T:\tmp\WOF2.msi

6

h) fc /B WinOF\_wlh\_x64.msi T:\tmp\WOF2.msi 7

i) copy /B T:\tmp\WOF2.msi WOF3.msi 8

j) fc /B WinOF\_wlh\_x64.msi WOF3.msi 9 k) del T:\tmp\WOF2.msi 10 l) del T:\tmp\WinOF\_wlh\_x64.msi 11

m) dir T:\tmp (should be empty) 12

n) rmdir T:\tmp 13

14

o) dir T:\ (should be empty) 15

p) del WOF3.msi 16

10) For each SRP target located in Procedure #4 17

a) Perform step 9 for each volume found for all targets as determined by 18

Windows Procedure step #4 - see Discover + Enable (bring online) the 19

SRP drive(s) 20

11) Take SRP drive offline 21

a) Start | Server Manager | Storage | Disk Management 22 b) Right-click the online disk and select ‘offline’ 23 c) dir T:\ (should fail). 24

12) Reboot all devices in the fabric and repeat the procedure using a different 25

SM. 26

**Note**: An HCA must successfully complete all operations to and from all volumes 27 on all targets using all available SM's in order to pass SRP testing. Two volumes 28 per target are all that is required. 29

30

31

32

33

34

35

36

37

38

39

40

41

42

**14.6 IB UDAPLTEST COMMANDS USING WINOF** 1

Server Command: dapl2test -T S -D <ia\_name> 2

3

**14.6.1 IB SETUP** 4

• The %SystemDrive%\DAT\dat.conf needs to be verified to be sure that the 5

correct interface is used. The DAPL interface for IB is ibnic0v2. 6

• It is also important to verify that the desired dat/dapl libraries are available 7

• %windir%\dat2.dll 8

• %windir%\dapl2.dll 9

10

• To run dapl2test on IB, an SM needs to be running.

11

**14.6.2 GROUP 1: POINT-TO-POINT TOPOLOGY** 12

[1.3] 1 connection and simple send/recv: 13

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -R BE 14

• client SR 256 1 server SR 256 1 15

[1.4] Verification, polling, and scatter gather list: 16

• dapl2test -T T -s <sever\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R 17

BE 18

• client SR 1024 3 -f \ 19

• server SR 1536 2 -f 20

**14.6.3 GROUP 2: SWITCHED TOPOLOGY** 21

InfiniBand Switch: Any InfiniBand switch 22

23

[2.5] Verification and private data: 24

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R 25

BE 26

• client SR 1024 1 \ 27

• server SR 1024 1 28

[2.6] Add multiple endpoints, polling, and scatter gather list: 29

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 10 -V -P -R 30

• BE client SR 1024 3 \ 31

• server SR 1536 2 32

[2.7] Add RDMA Write : 33

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R 34

BE 35

• client SR 256 1 \ 36

• server RW 4096 1 server SR 256 1 37

[2.8] Add RDMA Read: 38

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 1 -V -P -R 39

BE 40

• client SR 256 1 \ 41

42

• server RR 4096 1 server SR 256 1 1

**14.6.4 GROUP 3: SWITCHED TOPOLOGY WITH MULTIPLE SWITCHES** 2

[3.5] Multiple threads, RDMA Read, and RDMA Write: 3

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 4 -w 8 -V -P -R 4

BE 5

• client SR 256 1 \ 6

• server RR 4096 1 server SR 256 1 client SR 256 1 server RR 4096 1 \ 7

• server SR 256 1 8

[3.6] Pipeline test with RDMA Write and scatter gather list: 9

• dapl2test -T P -s <server\_name> -D <ia\_name> -i 1024 -p 64 -m p RW 10

8192 2 11

[3.7] Pipeline with RDMA Read: 12

• dapl2test -T P -s <server\_name> -D <ia\_name> -i 1024 -p 64 -m p RR 13

4096 2 14

[3.8] Multiple switches: 15

• dapl2test -T T -s <server\_name> -D <ia\_name> -i 100 -t 1 -w 10 -V -P -R 16

• BE client SR 1024 3 \ 17

• server SR 1536 2 18

**14.6.5 WINOF DAPL2TEST WRAPPER SCRIPTS** 19

20

All the specified DAPL tests are conveniently located in the WinOF distributed

DAPL test server & client scripts. 21

22

• %ProgramFiles(x86)%\WinOF\dt-svr.bat 23

• To run the dapl2test Server, to a Windows cmd-prompt window 24 type ‘dt-svr’. Only one server is necessary – multiple clients can 25 communicate with a single dapl2test server; multiple servers on 26 different nodes can exist. A single dapl2test client communicates 27

with only one dapl2test server at a time. 28

• No further server action is required as the dapl2test server is per- 29 sistent; looping waiting for dapltest client requests. 30

• %ProgramFiles(x86)%\WinOF\dt-cli.bat 31

• ‘dt-cli’ no arguments, will display dt-cli command args & options. 32

• Dapl2test client invocation: ‘dt-cli IPoIB\_IPv4\_server\_address 33 cmd’ 34

• If the dt-svr command was executed on a system where the IPoIB 35 interface address is 10.10.4.200 then 36

• ‘dt-cli 10.10.4.200 interop’ would run the above dap2tests be- 37 tween the client and server. 38

39

• ‘dt-cli 10.10.4.200 conn’ is a simple, quick test to verity dapl2test

client | server connection is operational. 40

41

42

**14.7 IB MPI - INTEL MPI USING WINOF** 1

**14.7.1 Requirements** 2

1) Intel MPI is not part of the WinOF installation; acquire Intel MPI installer file 3

from Intel. 4

2) Install same O/S version (Windows Server 2008-HPC) on homogenous 5

x86\_64 systems. 6

3) MPI testing requires a reliable IB fabric without other fabric interop testing 7

occurring. 8

4) Private Ethernet Network configuration 9

a) DNS names must match hostnames in hosts file. 10

5) WinOF Installation requirements 11

a) Install the latest version of WinOF on all systems (double-click 12

WinOF\_wlh\_x64.msi); see 13

i) <http://www.openfabrics.org/downloads/WinOF/README>14 ii) Select the ‘default’ set of install features; includes uDAPL. 15 iii) Run OpenSM either on the headnode OR from one of the IB switch- 16 es. 17

iv) If OpenSM on the headnode, select WinOF install feature ‘OpenSM 18

Started’. 19

b) Once WinOF installation on all nodes has completed, configure IPoIB 20 interfaces. 21 i) %windir%\system32\Drivers\etc\hosts should be setup with IB host- 22 names and static IP addresses. 23

ii) Assign IPv4 address, from hosts file, to each IPoIB interface; Exam- 24

ple: Local Area Connection 3 is the 1st IPoIB interface. 25

• netsh interface ip set address "Local Area Connection 4" static 26

10.10.4.y 255.255.255.0 27

This allows you to **set** the IPoIB IP address. 28

• netsh interface ip show address "Local Area Connection 3" 29

This allows you to **view** the IPoIB IP address. 30

iii) Verify by pinging IPoIB interface addresses on all nodes. 31

32

33

**14.7.2 Setup information for Intel MPI** 34

Install Intel MPI on every cluster node: 35

1) [Intel MPI runtime environment kit](http://www.intel.com/cd/software/products/asmo-na/eng/308295.htm) 36

a) <http://www.intel.com/cd/software/products/asmo-na/eng/308295.htm>37

2) [Intel MPI Benchmarks](http://www.intel.com/cd/software/products/asmo-na/eng/cluster/mpi/219848.htm) , 38

39

a) <http://www.intel.com/cd/software/products/asmo-na/eng/clus-> ter/mpi/219848.htm 40

41

3) Add identical user account (%SystemDrive%\users\test) on every node.

42

**14.7.3 Additional Information**

4) Headnode mount points (%SystemDrive%\test\export) on user accounts. 1

2

3

1) Go to the individual test directories and follow the steps in the respective 4

README-\*.txt files. 5

2) For Intel MPI Support Services go to: 6 a) [http://software.intel.com/en-us/articles/intel-mpi-library-for-win-](http://software.intel.com/en-us/articles/intel-mpi-library-for-windows/all/1/) 7 [dows/all/1/](http://software.intel.com/en-us/articles/intel-mpi-library-for-windows/all/1/) 8

b) See Intel MPI Reference Manual for Additional information 9

10

11

**14.7.4 Intel MPI (MVAPICH 2) - Test Procedure**

12

1) Run a subnet manager from one node only.

13

2) Run Intel® MPI Benchmarks from the HPC head-node: 14

a) Two sets of tests should be run, with these command lines 15

• mpiexec -np <number of nodes X number of proces- 16 sors/node> IMB-MPI1 -multi 0 PingPong PingPing 17

• mpiexec -np <number of nodes X number of proces- 18 sors/node> IMB-MPI1 19

The first command runs just the PingPong and PingPing point-to- 20

point tests, but makes all tasks active (pairwise). 21

The second command runs all the tests (PingPong, PingPing, Send- 22

recv, Exchange, Bcast, Allgather, Allgatherv, Alltoall, Reduce, 23

Reduce\_scatter, Allreduce, Barrier), in non-multi mode.

24

b) If the test passes shutdown current subnet manager and start another

one on a different node; run both tests again. 25

26

3) Repeat until all nodes have run a subnet manager and passed all tests.

27

28

**14.7.5 Interpreting the results** 29

1) TBA 30

31

32

33

34

35

36

37

38

39

40

41

42

**15 BUG REPORTING METHODOLOGY DURING PRE-TESTING** 1

The following bug reporting methodology will be followed during the execution of 2

interoperability pre-testing at UNH-IOL. 3

4

1) UNH-IOL and the OEMs (e.g. Chelsio, Data Direct, Intel, NetApp, Mellanox)

will assign a focal point of contact to enable fast resolution of problems. 5

6

2) Bug reports will include:

7

a) Detailed fail report with all relevant detail (Test/Application, Topology.). 8

b) [**For IB**] IB trace if needed. 9

c) [**For iWARP**] iWARP, TCP and SCTP traces if needed. 10

3) Bug reports will be sent via email by UNH-IOL to the focal point assigned by 11

the OEM 12

4) Bug reports and suggested fixes will be sent to the OpenFabrics devel- 13

opment community - [OFA Bugzilla](https://bugs.openfabrics.org/). When such reports are communicated, 14

UNH-IOL will ensure that confidentiality between UNH-IOL and the OEM will

be maintained. Bug reports will be generalized and not include any company 15

specific proprietary information such as product name, software name, 16

version etc. 17

5) All bug fixes/issues that are found during testing will be uploaded to the 18

OpenFabrics repository. Documentation related to fixes will not mention any 19

company specific proprietary information. 20

**Note**: This test plan does not cover how bugs will be reported by IBTA/CIWG or 21

IETF iWARP during or after interoperability testing at plugfests. 22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**16 RESULTS SUMMARY** 1

**16.1 INFINIBAND SPECIFIC TEST RESULTS** 2

Please add a check mark whenever a test case passes and when the system is 3

behaving according to the criteria mentioned below. Otherwise indicate a failure 4

along with a comment explaining the nature of the failure. 5

6

**Results Table 1 - IB Link Initialize** 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Phy link up all ports |  |  |  |

8

9

10

11

12

**Results Table 2 - IB Fabric Initialization** 13

14

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Verify that all ports are in Armed or  Active state |  |  |  |

15

16

17

18

19

**Results Table 3 - IB IPoIB - Connected Mode (CM)** 20

21

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Ping all to all - Ping using SM 1 |  |  |  |
| 2 | Ping all to all - Ping using SM 2 |  |  |  |
| 3 | Ping all to all - Ping using SM 3 |  |  |  |
| 4 | Ping all to all - Ping using SM 4 |  |  |  |
| 5 | Ping all to all - Ping using SM 5 |  |  |  |
| 6 | Ping all to all - Ping using SM 6 |  |  |  |
| 7 | Ping all to all - Ping using SM x |  |  |  |
| 8 | Connect/Disconnect Host |  |  |  |
| 9 | FTP Procedure |  |  |  |

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**Results Table 4 - IB IPoIB - Datagram Mode (DM)** 1

2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Ping all to all - Ping using SM 1 |  |  |  |
| 2 | Ping all to all - Ping using SM 2 |  |  |  |
| 3 | Ping all to all - Ping using SM 3 |  |  |  |
| 4 | Ping all to all - Ping using SM 4 |  |  |  |
| 5 | Ping all to all - Ping using SM 5 |  |  |  |
| 6 | Ping all to all - Ping using SM 6 |  |  |  |
| 7 | Ping all to all - Ping using SM x |  |  |  |
| 8 | Connect/Disconnect Host |  |  |  |
| 9 | FTP Procedure |  |  |  |

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

**Table 5 - IB SM Failover/Handover** 18

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Basic sweep test |  |  |  |
| 2 | SM Priority test |  |  |  |
| 3 | Failover test - Disable SM1 |  |  |  |
| 4 | Failover test - Disable SM2 |  |  |  |

19

20

21

22

23

24

25

26

27

**Results Table 6 - IB SRP** 28

29

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Basic dd application |  |  |  |
| 2 | IB SM kill |  |  |  |

30

31

32

33

34

35

36

37

38

39

40

41

42

1

2

**Results Table 7 - Fibre Channel Gateway - (IB Specific)** 3

4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Basic Setup |  |  |  |
| 2 | Configure Gateway |  |  |  |
| 3 | Add Storage Device |  |  |  |
| 4 | Basic dd application |  |  |  |
| 5 | IB SM kill |  |  |  |
| 6 | Disconnect Host/Target |  |  |  |
| 7 | Load Host/Target |  |  |  |
| 8 | dd after SRP Host and Target reloaded |  |  |  |
| 9 | Reboot Gateway |  |  |  |
| 10 | dd after FC Gateway reboot |  |  |  |

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

**Results Table 8 - Ethernet Gateway - (IB Specific)** 21

22

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Basic Setup |  |  |  |
| 2 | Start ULP |  |  |  |
| 3 | Discover Gateway |  |  |  |
| 4 | SM Failover |  |  |  |
| 5 | Ethernet gateway reboot |  |  |  |
| 6 | ULP restart |  |  |  |
| 7 | Unload/load ULP |  |  |  |

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**16.2 ETHERNET SPECIFIC TEST RESULTS** 1

**Results Table 9 - iWARP Link Initialize** 2

3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Phy link up all ports |  |  |  |
| 2 | Verify basic IP connectivity |  |  |  |

4

5

6

7

8

9

**Table 10 - RoCE Link Initialize** 10

11

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Phy link up all ports |  |  |  |
| 2 | Verify basic IP connectivity |  |  |  |

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**16.3 TRANSPORT INDEPENDENT TEST RESULTS** 1

2

3

**Results Table 11 - TI iSER** 4

5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Basic dd application |  |  |  |
| 2 | IB SM kill |  |  |  |
| 3 | Disconnect Initiator |  |  |  |
| 4 | Disconnect Target |  |  |  |
| 5 | Repeat with previous SM Slave |  |  |  |

6

7

8

9

10

11

12

13

14

15

**Results Table 12 - TI NFS Over RDMA** 16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | File and directory creation |  |  |  |
| 2 | File and directory removal |  |  |  |
| 3 | Lookups across mount point |  |  |  |
| 4 | Setattr, getattr, and lookup |  |  |  |
| 5 | Read and write |  |  |  |
| 6 | Readdir |  |  |  |
| 7 | Link and rename |  |  |  |
| 8 | Symlink and readlink |  |  |  |
| 9 | Statfs |  |  |  |

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

**Results Table 13 - TI RDS** 32

33

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | rds-ping procedure |  |  |  |
| 2 | rds-stress procedure |  |  |  |

34

35

36

37

38

39

40

41

42

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | P2P - Connection & simple send receive |  |  |  |
| 2 | P2P - Verification, polling & scatter gather list |  |  |  |
| 3 | Switched Topology -Verification and private data |  |  |  |
| 4 | Switched Topology - Add multiple endpoints, polling, &  scatter gather list |  |  |  |
| 5 | Switched Topology - Add RDMA Write |  |  |  |
| 6 | Switched Topology - Add RDMA Read |  |  |  |
| 7 | Multiple Switches - Multiple threads, RDMA Read, & RDMA Write |  |  |  |
| 8 | Multiple Switches - Pipeline test with RDMA Write &  scatter gather list |  |  |  |
| 9 | Multiple Switches - Pipeline with RDMA Read |  |  |  |
| 10 | Multiple Switches - Multiple switches |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Small RDMA READ |  |  |  |
| 2 | Large RDMA READ |  |  |  |
| 3 | Small RDMA Write |  |  |  |
| 4 | Large RDMA Write |  |  |  |
| 5 | Small RDMA SEND |  |  |  |
| 6 | Large RDMA SEND |  |  |  |
| 7 | Small RDMA Verify |  |  |  |
| 8 | Large RDMA Verify |  |  |  |

|  |  |
| --- | --- |
|  | 1 |
| 2 |
| **Results Table 14 - TI uDAPL** | 3 |
|  | 4 |
|  | 5 |
|  | 6 |
|  | 7 |
|  | 8 |
|  | 9 |
|  | 10 |
|  | 11 |
|  | 12 |
|  | 13 |
|  | 14 |
|  | 15 |
|  | 16 |
|  | 17 |
|  | 18 |
|  | 19 |
|  | 20 |
|  | 21 |
|  | 22 |
| **Results Table 15 - TI RDMA Basic Interop** | 23 |
|  | 24 |
|  | 25 |
|  | 26 |
|  | 27 |
|  | 28 |
|  | 29 |
|  | 30 |
|  | 31 |
|  | 32 |
|  | 33 |
|  | 34 |
|  | 35 |
|  | 36 |
|  | 37 |
|  | 38 |
|  | 39 |
|  | 40 |
|  | 41 |
|  | 42 |

1

2

**Results Table 16 - TI RDMA Stress Tests** 3

4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | Switch Load |  |  |  |
| 2 | Switch Fan In |  |  |  |

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**16.4 OPEN MPI TEST RESULTS**

1

2

3

**Results Table 17 - TI MPI - Open MPI** 4

5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test Suite** | **Pass** | **Fail** | **Comment** |
| **Phase 1: "Short" tests** | | | | |
| 2 | OMPI built with OpenFabrics support |  |  |  |
| 3 | OMPI basic functionality (hostname) |  |  |  |
| 4.1 | Simple MPI functionality (hello\_c) |  |  |  |
| 4.2 | Simple MPI functionality (ring\_c) |  |  |  |
| 5 | Point-to-point benchmark (NetPIPE) |  |  |  |
| 6.1.1 | Point-to-point benchmark (IMB PingPong multi) |  |  |  |
| 6.1.2 | Point-to-point benchmark (IMB PingPing multi) |  |  |  |
| **Phase 2: "Long" tests** | | | | |
| 6.2.1 | Point-to-point benchmark (IMB PingPong) |  |  |  |
| 6.2.2 | Point-to-point benchmark (IMB PingPing) |  |  |  |
| 6.2.3 | Point-to-point benchmark (IMB Sendrecv) |  |  |  |
| 6.2.4 | Point-to-point benchmark (IMB Exchange) |  |  |  |
| 6.2.5 | Collective benchmark (IMB Bcast) |  |  |  |
| 6.2.6 | Collective benchmark (IMB Allgather) |  |  |  |
| 6.2.7 | Collective benchmark (IMB Allgatherv) |  |  |  |
| 6.2.8 | Collective benchmark (IMB Alltoall) |  |  |  |
| 6.2.9 | Collective benchmark (IMB Reduce) |  |  |  |
| 6.2.10 | Collective benchmark (IMB Reduce\_scatter) |  |  |  |
| 6.2.11 | Collective benchmark (IMB Allreduce) |  |  |  |
| 6.2.12 | Collective benchmark (IMB Barrier) |  |  |  |
| 6.3.1 | I/O benchmark (IMB S\_Write\_Indv) |  |  |  |
| 6.3.2 | I/O benchmark (IMB S\_IWrite\_Indv) |  |  |  |
| 6.3.3 | I/O benchmark (IMB S\_Write\_Expl) |  |  |  |
| 6.3.4 | I/O benchmark (IMB S\_IWrite\_Expl) |  |  |  |
| 6.3.5 | I/O benchmark (IMB P\_Write\_Indv) |  |  |  |
| 6.3.6 | I/O benchmark (IMB P\_IWrite\_Indv) |  |  |  |

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test Suite** | **Pass** | **Fail** | **Comment** |
| 6.3.7 | I/O benchmark (IMB P\_Write\_Shared) |  |  |  |
| 6.3.8 | I/O benchmark (IMB P\_IWrite\_Shared) |  |  |  |
| 6.3.9 | I/O benchmark (IMB P\_Write\_Priv) |  |  |  |
| 6.3.10 | I/O benchmark (IMB P\_IWrite\_Priv) |  |  |  |
| 6.3.11 | I/O benchmark (IMB P\_Write\_Expl) |  |  |  |
| 6.3.12 | I/O benchmark (IMB P\_IWrite\_Expl) |  |  |  |
| 6.3.13 | I/O benchmark (IMB C\_Write\_Indv) |  |  |  |
| 6.3.14 | I/O benchmark (IMB C\_IWrite\_Indv) |  |  |  |
| 6.3.15 | I/O benchmark (IMB C\_Write\_Shared) |  |  |  |
| 6.3.16 | I/O benchmark (IMB C\_IWrite\_Shared) |  |  |  |
| 6.3.17 | I/O benchmark (IMB C\_Write\_Expl) |  |  |  |
| 6.3.18 | I/O benchmark (IMB C\_IWrite\_Expl) |  |  |  |
| 6.3.19 | I/O benchmark (IMB S\_Read\_Indv) |  |  |  |
| 6.3.20 | I/O benchmark (IMB S\_IRead\_Indv) |  |  |  |
| 6.3.21 | I/O benchmark (IMB S\_Read\_Expl) |  |  |  |
| 6.3.22 | I/O benchmark (IMB S\_IRead\_Expl) |  |  |  |
| 6.3.23 | I/O benchmark (IMB P\_Read\_Indv) |  |  |  |
| 6.3.24 | I/O benchmark (IMB P\_IRead\_Indv) |  |  |  |
| 6.3.25 | I/O benchmark (IMB P\_Read\_Shared) |  |  |  |
| 6.3.26 | I/O benchmark (IMB P\_IRead\_Shared) |  |  |  |
| 6.3.27 | I/O benchmark (IMB P\_Read\_Priv) |  |  |  |
| 6.3.28 | I/O benchmark (IMB P\_IRead\_Priv) |  |  |  |
| 6.3.29 | I/O benchmark (IMB P\_Read\_Expl) |  |  |  |
| 6.3.30 | I/O benchmark (IMB P\_IRead\_Expl) |  |  |  |
| 6.3.31 | I/O benchmark (IMB C\_Read\_Indv) |  |  |  |
| 6.3.32 | I/O benchmark (IMB C\_IRead\_Indv) |  |  |  |
| 6.3.33 | I/O benchmark (IMB C\_Read\_Shared) |  |  |  |
| 6.3.34 | I/O benchmark (IMB C\_IRead\_Shared) |  |  |  |
| 6.3.35 | I/O benchmark (IMB C\_Read\_Expl) |  |  |  |
| 6.3.36 | I/O benchmark (IMB C\_IRead\_Expl) |  |  |  |

**Results Table 17 - TI MPI - Open MPI** 1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**Results Table 17 - TI MPI - Open MPI** 1

2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test Suite** | **Pass** | **Fail** | **Comment** |
| 6.3.37 | I/O benchmark (IMB Open\_Close) |  |  |  |

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

**16.5 OSU MPI TEST RESULTS**

1

2

3

**Results Table 18 - TI MPI - OSU** 4

5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test #** | **Test** | **Pass** | **Fail** | **Comment** |
| 1 | **Test** 1: PingPong |  |  |  |
| 2 | **Test** 1: PingPing point-to-point |  |  |  |
| 3 | **Test** 2: PingPong |  |  |  |
| 4 | **Test** 2: PingPing |  |  |  |
| 5 | **Test** 2: Sendrecv |  |  |  |
| 6 | **Test** 2: Exchange |  |  |  |
| 7 | **Test** 2: Bcast |  |  |  |
| 8 | **Test** 2: Allgather |  |  |  |
| 9 | **Test** 2: Allgatherv |  |  |  |
| 10 | **Test** 2: Alltoall |  |  |  |
| 11 | **Test** 2: Alltoallv |  |  |  |
| 12 | **Test** 2: Reduce |  |  |  |
| 13 | **Test** 2: Reduce\_scatter |  |  |  |
| 14 | **Test** 2: Allreduce |  |  |  |
| 15 | **Test** 2: Barrier |  |  |  |

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

**Results Table 19 Remarks** 28

29

**General Remarks:** Comments about the set-up, required updates to the TD, and any other issues that came up 30

during the testing. 31

32

33

34

35

36

37

38

39

40

41

42