# **OFA Interoperability Working Group**

## OFA-IWG Interoperability Test Plan Release 1.28



March 27, 2009 DRAFT

Copyright © 2009 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

#### **Revision History**

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 Ark- ady 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 conference in Paris.
1.01	Aug 17, 2006	Updated the iWARP Equipment requirements in the General 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 Chan- nel 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 Mikkel 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	Release Date	
1.11	Sep 7, 2007	Updated with the latest scripts developed by UNH IOL and based on the results from the April 2007 Interop Event
1.12	Sep 12, 2007	Updated the documents to embed the test scripts in the document.
1.13	Jan 22, 2008	Updated the documents for the March 2008 OFA Interop event. IPoIB updated along with Cover Page and the Test Requirements section.
1.14	Feb 11, 2008	Added the following tests: 1. Ethernet Switch Tests 2. IPoIB Connected Mode 3. RDMA Interop 4. RDS
1.15	Feb 18, 2008	Updates to the following tests: 1. Ethernet Switch Tests 2. IPoIB Connected Mode 3. RDMA Interop
1.16	Feb 25, 2008	Removed all reference to Low Latency Ethernet Switches. This is the version for the March 2008 Interop Event
1.17	March 3, 2008	Added HP-MPI
1.18	July 22, 2008	Updated HP-MPI based on results from the March 2008 Interop Event
1.19	July 28, 2008	Updated HP-MPI URL for the tests. Added section for Open MPI Updated MPI based on feedback from UNH IOL
1.20	July 30, 2008	Updated section for Open MPI and added tables Updated IB SM Failover as per Nick Wood
1.21	Aug 1, 2008	Updated SRP call srp_daemon -o -e -n Updated IB SM Failover as Bob Jaworski Updated HP-MPI Updated Intel MPI Updated Open MPI
1.22	Aug 29, 2008	Added a section for MVAPICH 1 under OSU MPI
1.23	Feb 16, 2009	Updated link init, fabric init, srp, sdp, ipoib cm, ipoib dm based on updates received from UNH-IOL
1.24	Feb 23, 2009	Updated Intel MPI and Open MPI to reflect the fact that they are not intended to work in a heterogeneous environ- ment. Updated the RDS test procedure
		Updated the Test Glossary
		opulled the manualory lest lable for April 2000

Revision	Release Date	
1.25	Feb 24, 2009	Updated the RDS Test after review by the OFA IWG group.
1.26	Mar 13, 2009	Restructured entire document to accommodate WinOF and OFED Added NFS over RDMA to the test plan. Added WinOF tests Updated HP-MPI Add List of Contributors
1.27	Mar 17, 2009	Updates based on the review from the OFA IWG
1.28	Mar 27, 2009	Added links in Chapter 10 to the InfiniBand Test Scripts Added links to HP-MPI installation Packages

#### List of Contributors

Name	Company
Mark Alan	HP
Rupert Dance	Lamprey Networks
Sujal Das	Mellanox
Arlin Davis	Intel
Johann George	QLogic
Mike Hagen	UNH-IOL
Allen Hubbe	UNH-IOL
Bob Jaworski	QLogic
Arkady Kanevsky	NetApp
Llolsten Kaonga	Lamprey Networks
Amit Krig	Mellanox
Jon Mason	Open Grid Computing
Bob Noseworthy	UNH-IOL
Yaroslav Pekelis	Mellanox
Hal Rosenstock	Obsidian
Martin Schlining	DataDirect Networks
Karun Sharma	QLogic
Stan Smith	Intel
Dave Sommers	Intel (NetEffect)
Jeff Squyres	Cisco
Dennis Tolstenko	Lamprey Networks
Steve Wise	Open Grid Computing
Robert Woodruff	Intel
Nick Wood	UNH-IOL

#### Editor: Rupert Dance

3	9
4	0
4	1

LEGAL DISCLAIMER	"This version of a proposed OpenFabrics Interop Test Plan is provided "AS IS" and without any warranty of any kind, including, without limitation, any express or implied warranty of non-infringement, merchant- ability or fitness for a particular purpose.	1 2 3 4 5 6
	In no event shall OpenFabrics, IBTA or any member of these groups be liable for any direct, indirect, special, exemplary, punitive, or consequential damages, in- cluding, without limitation, lost profits, even if ad- vised of the possibility of such damages."	7 8 9 10 11 12 13
	Conditional text tag Explanation is shown in green.	14 15
	Conditional text tag Deleted is shown in red with strike through.	16 17
	Conditional text tag <i>Proposal</i> is shown in turquoise (r0_g128_b128).	18 19
	Conditional text tag Author is shown as is.	20
	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

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	Introduction RELEASE 1.28	March 27, 2009 DRAFT
1 INTRODUCTION		1
	Server OEM customers have expressed the need for RE	MA hardware and soft- 2
	ware to interoperate.	3
	Specifically, InfiniBand HCA, OpenFabrics host software	to interoperate with In-
	finiBand Switches, gateways, and bridges with managen	nent software provided
	Fabrics host software to interoperate with Ethernet Switch	hes and management 7
	software and hardware provided by Ethernet Switch OEI	Is and iWARP inte- 8
	grated server OEM vendors.	9
	It is necessary that the interoperability test effort be an inc	lustry-wide effort where 1
	interoperability testing is conducted under the auspices of working organizations. For InfiniBand it is IBTA, specified	of the appropriate net- 1
	the CIWG. And for iWARP it is IETF, and specifically with	in UNH IOL iWARP
	Consortium.	1
1.1 PURPOSE		1
	This document is intended to describe the production tes	ts step by step ex-
	plaining each test and its references. The purpose of this	test plan is three fold: 1
	1) Define the scope, equipment and software needs, and	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
	verifying full interoperability of RDMA HW and SW. F	for Infiniband HW it is
	InfiniBand HCAs using the latest OpenFabrics IB OF	ED software with cur- 2
	OEM IB Switch vendors are Flextronics, Mellanox, C	bsidian, QLogic and 2
	Voltaire. For iWARP HW it is iWARP RNICs using th	e latest OpenFabrics
	Gateways, Edge Devices and so on with their manage	gement software. 2
	2) Serve as a basis for evaluating customer acceptance	e criteria for OFA host 2
	software interoperability and OFA Logo.	2
	3) Serve as a basis for extensions to InfiniBand IBTA C	IWG test procedures 2
	PlugFest events organized by IBTA.	aures in upcoming 2
	Serve as a basis for extensions to iWARP test proce	dures for OpenFabrics 3
	software related to interoperability and use of these	est procedures in up- 3
	coming PlugFest events organized by UNH IOL iWA	RP Consortium. 3
<b>1.2 INTENDED AUDIENCE</b>		3
	The following are the intended audience for this docume	nt: 3
	1) Project managers in OEM Switch, Router, Gateway,	ن Bridge Vendor compa-
	nies to understand the scope of testing and participa	te in the extension of
	this test plan and procedures as necessary to meet	neir requirements.
	<ol> <li>IB IA and CIWG, and IWARP and UNH IOL iWARP to companies to evaluate the scope of testing and participation.</li> </ol>	esting personnel and cipate in the extension 3
	of this test plan and procedures as necessary to me	et their requirements.
		4
		4

Last Modified: 3/29/09 8:19 am

OFA Interoperability Working Group		Introduction	March 27, 2009
JIA-IWO INTEROPERABILITT TEST I LAN		RELEASE 1.20	
	B) Test engine	eering and project leads and manag	ers who will conduct the
	testing bas	sed on this document.	
	<ul> <li>Customers teroperabil</li> </ul>	s and users of OFA host software wh lity.	no rely on OFA Logo for in-
	5) Integrators	and OEM of RDMA products.	
.3 TEST PLAN STRUCTURE			
	This test plan i	s divided into two main sections.	
	I) Interopera	ability testing using OFED for Lin	ux.
	a) See S	Section 10	
	2) Interoperal	bility testing using WinOF for Windo	ws Platforms.
	a) See S	Section 11	
	u) 0000		

2

3

1

#### **1.4 INFINIBAND ONLY - TEST OVERVIEW** 1 2 The tables below list all of the specific test procedures for InfiniBand Devices. See the Transport Independent section for tests that apply to all transports. 3 4 5 Table 1 - IB Link Initialize 6 7 Test **Description Overview** Test # 8 Phy link up all ports Check that all relevant green LEDs are on for all HCAs and switches. 9 10 Logical link up all ports All vendors should check that the link state is up and the port width and link speed is as adverswitch SM 11 tised by the vendor. 12 Logical link up all ports All vendors should check that the link state is up and the port width and link speed is as adver-HCA SM tised by the vendor. 13 14 15 16 Table 2 - IB Fabric Initialization 17 Test **Description Overview** Test # 18 19 Fabric Initialization Run SM from each node in cluster and see that all ports are in Armed or Active state. 20

#### Table 3 - IB IPoIB - Connect Mode (CM)

Test #	Test	Description Overview	
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.	

#### Table 4 - IB IPoIB - Datagram Mode (DM)

Test #	Test	Description Overview	
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.	

#### Table 5 IB SM Tests

			3
Test #	Test	Description Overview	4
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 (	6
2	SM Priority test	Verify Subnet and SMs behavior according to the SMs priority.	8
3	Failover - Disable SM1	Disable the master SM and verify that standby SM becomes master and configures the cluster.	Ç
4	Failover - Disable SM2	Disable the master SM and verify that standby SM becomes master and configures the cluster.	1

#### Table 6 - IB SRP Tests

Test #	Test	Description Overview
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.

#### Table 7 - IB Ethernet Gateway

Test #	Test	Description Overview	
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 discover 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.	

Test #	Test	Description Overview	
1	Basic Setup	Connect the HCA of the IB host to the IB fabric. Connect the FC Gateway to the IB Fabric. Connect 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).	
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.	

 Table 8 - IB Fibre Channel Gateway

Ethernet Only - Test Overview RELEASE 1.28

#### **1.5 ETHERNET ONLY - TEST OVERVIEW**

The tables below list all of the specific test procedures for iWARP and Ethernet2Devices. See the Transport Independent section for tests that apply to all transports.3

#### Table 9 - Ethernet Link Initialize

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

#### Table 10 - Ethernet Fabric Initialize

Test #	Test	Description Overview
1	Fabric Initialization	Verify IP connectivity to all IP attached stations in the Cluster. Source 1000 minimum size ICMP echo requests from all RNICs to all other IP entities to verify cluster connectivity.

#### **Table 11 - Ethernet Fabric Reconvergence**

Test #	Test	Description Overview
1	Fabric Reconvergence	Run SM from each node in cluster and see that all ports are in Armed or Active state.

#### Table 12 - Ethernet Fabric Failover

Test #	Test	Description Overview
1	Fabric Failover	Kill root RSTP switch of the primary VLAN, ensure there is a fully redundant path through the fabric and verify recovery occurs

Table 13	iWARP	Connections
Table 15		Connections

Test #	Test	Description Overview
1	UNH iWARP interop tests group 1	Verify that each single iWARP operation over single connection works.
2	UNH iWARP interop tests group 2	Verify that multiple iWARP operations over a single connection work.
3	UNH iWARP interop tests group 3	Verify that multiple iWARP connections work.
4	UNH iWARP interop tests group 4	Verify that disconnect/reconnect physical connections work.
5	UNH iWARP interop tests group 5	Verify that IP Speed negotiation work.
6	UNH iWARP interop tests group 6	Verify that iWARP error ratio work.
7	UNH iWARP interop tests group 7	Verify that stress pattern over iWARP work.
8	UNH iWARP interop tests group 8	Verify that iWARP parameter negotiation work.

#### Transport Independent - Test Overview RELEASE 1.28

1

4 5

18

37 38

#### **1.6 TRANSPORT INDEPENDENT - TEST OVERVIEW**

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

#### Table 14 - TI iSER Tests

Test #	Test	Description Overview
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).

#### Table 15 - TI NFS Over RDMA

Test #	Test	Description Overview
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).

#### Table 16 TI RDS

			39
Test #	Test	Description Overview	40
1	rds-ping procedure	Run rds-ping and verify that you can reach all hosts in the cluster	41

#### Table 16 TI RDS

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

#### Table 17 - TI SDP Tests

Test #	Test	Description Overview
1	netperf procedure	Run netperf where message size is 10, 100, 1000, 10000 and local buffer size is 1024, 6000.
2	FTP procedure	Using a 4MB test file, put the file, then get the file and finally compare the file.
3	IB SCP Procedure	Connect via SCP on IPoIB address from all other nodes uploading and downloading a file.
4	iWARP SCP Procedure	Connect via SCP from all other nodes uploading and downloading a file.

#### Table 18 - TI uDAPL

Test #	Test	Description Overview	
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.	

#### Table 19 - Basic RDMA Interop

Test #	Test	Description Overview	3
1	Small RDMA READ	Create an RDMA command sequence to send a READ operation of one byte.	- 30
2	Large RDMA READ	Create an RDMA command sequence to send a READ operation of 10,000,000 bytes	4(
3	Small RDMA Write	Create an RDMA command sequence to send a Write operation of one byte	41
			- 42

#### Table 19 - Basic RDMA Interop

Test #	Test	Description Overview
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

#### Table 20 - RDMA operations over Interconnect Components

Test #	Test	Description Overview
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

### 1.7 HP-MPI - TEST OVERVIEW

Test #	HP-MPI TESTs	HP-MPI TESTs Suite Description
1	IMB	This is the Intel MPI Benchmark. If this passes, then the basic interoperability of HP-MPI with the installed OFED is confirmed.
2	rings2	This is a proprietary HP test which has a good history of stressing interconnects to the point of failure. It also includes 1sided operations.
3	fork	New RDMA implementations often have fork issues, As new OS kernels come out the fork problems sometimes re-appear. This test makes a point of stressing that code path.
4	exitpath	The purpose of this test is simply to make sure machines and OFED drivers etc remain stable when applications repeatedly terminate abnormally.
5	alltoone	This test has all the ranks send a flood of messages to rank 0, to make sure the intercon- nect can handle heavy load in that message pattern.

24

25

#### **1.8 INTEL MPI - TEST OVERVIEW**

#### Table 22a - Intel MPI Benchmark Summary

Test #	Test	Description Overview	
1	Test 1: PingPong		
2	Test 1: PingPing		
3	Test 1: Sendrecv		
4	Test 1: Exchange		
5	Test 1: Allreduce		
6	Test 1: Reduce		
7	Test 1: Reduce_scatter		
8	Test 1: Allgather		
9	Test 1: Allgatherv		
10	Test 1: Alltoall		
11	Test 1: Alltoallv		
12	Test 1: Bcast		
13	Test 1: Barrier		

#### Table 22b - TI - Intel MPICH2 Suite - (Not part of OFA Stack)

Test #	MPICH2 (16 sections, 290 tests)	Intel - MPICH2 Test Suite Section Description
1	attr	Test programs for attribute routines
2	coll	Test programs for various collective operations
3	comm	Test programs for communicator operations
4	datatype	Test programs for various datatype operations
5	errhan	Test programs for error handling operations
6	group	Test programs for the group operations
7	info	Test programs for various info operations
8	init	Test programs for init operations
9	pt2pt	Test programs for various point to point routines (send, isend, probe, etc.)
10	rma	Test programs for memory access operations
11	spawn	Test programs for comm_spawn, intercom operations

#### Table 22b - TI - Intel MPICH2 Suite - (Not part of OFA Stack)

Test #	MPICH2 (16 sections, 290 tests)	Intel - MPICH2 Test Suite Section Description
12	topo	Test programs for various topology routines
13	io	Test programs for file i/o read/write, sync and async
14	F77	Test programs for f77
15	схх	Test programs for c++
16	threads	Test programs for threaded send/recv

#### Table 22c - TI - Intel MPI Test Suite Description - (Not part of OFA Stack)

Test #	IntelMPITEST (5 sections, 1371 tests)	IntelMPITest Suite Description	
1	testlist21 (1085 tests)	c - blocking, coll, datatype, env, group, misc, non-blocking	1
2	testlist2-21 (23 tests)	c, fortran – datatype create	1
3	testlist4 (216 tests)	fortran – grp, topo, blocking, coll, datatype, non-blocking, persist, probe, send/recv	
4	testlist4lg (1 test)	c - collective overlap	
5	testlist6 (46 tests)	c, fortran – topo cart/graph	

#### 1.9 OPEN MPI - TEST OVERVIEW

Test #	Open MPI TESTs Open MPI TESTs Suite Description			
		Phase 1: "Short" tests	7	
1	2	OMPI built with OpenFabrics support	8	
2	3	OMPI basic functionality (hostname)		
3	4.1	Simple MPI functionality (hello_c)	1	
4	4.2	Simple MPI functionality (ring_c)	1	
5	5	Point-to-point benchmark (NetPIPE)	1	
6	6.1.1	Point-to-point benchmark (IMB PingPong multi)		
7	6.1.2	Point-to-point benchmark (IMB PingPing multi)	1	
		Phase 2: "Long" tests	1	
8	6.2.1	Point-to-point benchmark (IMB PingPong)	1	
9	6.2.2	Point-to-point benchmark (IMB PingPing)	2	
10	6.2.3	Point-to-point benchmark (IMB Sendrecv)	2	
11	6.2.4	Point-to-point benchmark (IMB Exchange)	2	
12	6.2.5	Collective benchmark (IMB Bcast)	2	
13	6.2.6	Collective benchmark (IMB Allgather)	2	
14	6.2.7	Collective benchmark (IMB Allgatherv)		
15	6.2.8	Collective benchmark (IMB Alltoall)		
16	6.2.9	Collective benchmark (IMB Reduce)	2	
17	6.2.10	Collective benchmark (IMB Reduce_scatter)	3	
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)	3	
21	6.3.2	I/O benchmark (IMB S_IWrite_Indv)	3	
22	6.3.3	I/O benchmark (IMB S_Write_Expl)	3	
23	6.3.4	I/O benchmark (IMB S_IWrite_Expl)	3	
24	6.3.5	I/O benchmark (IMB P_Write_Indv)	3	
25	6.3.6	I/O benchmark (IMB P_IWrite_Indv)	4	
26	6.3.7	I/O benchmark (IMB P_Write_Shared)	4	

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 23 - TI - Open MPI Test Suite Description

4

### 1.10 OSU MPI - TEST OVERVIEW

#### Table 24 - TI - OSU MPI

Test #	Test	Description Overview	5
1	Test 1: PingPong		7
2	Test 1: PingPing point-to-point		8
3	Test 2: PingPong		9
4	Test 2: PingPing		10
5	Test 2: Sendrecv		12
6	Test 2: Exchange		13
7	Test 2: Bcast		14
8	Test 2: Allgather		16
9	Test 2: Allgatherv		17
10	Test 2: Alltoall		18
11	Test 2: Alltoallv		19
12	Test 2: Reduce		21
13	Test 2: Reduce_scatter		22
14	Test 2: Allreduce		23
15	Test 2: Barrier		24

1.11 REQUIREMENTS FOR OFA INTEROF The follo erability j pated tha following	<b>PERABILITY LOGO PROGRA</b> wing table indicates the manda Logo Program during the April at some of the Beta tests will b Interop Event.	M atory tests to qualify for the OFA Intero 2009 Interoperability Event. It is antici- e moved to Mandatory status for the	1 p-2 -3 4 5
Table 25       - InfiniBand Tra         Test Procedure	Linux	ril 2009 Interop Event WinOF	7
IB Link Initialize	Mandatory	Beta	9
IB Fabric Initialization	Mandatory	Beta	10 11
IB IPoIB Connected Mode	Mandatory	Not Available -1	12
IB IPoIB Datagram Mode	Mandatory	Beta	13
IB SM Failover/Handover	Beta	Beta	14 15
IB SRP	Mandatory	Beta	16
IB Ethernet Gateway	Beta	Not Available - 3	17
IB Fibre Channel Gateway	Beta	Not Available - 3	18
TI iSER	Mandatory	Beta	20
TI NFS over RDDMA	Beta	Not Available - 1	21
TI RDS	Beta	Not Available - 2	22
TI SDP	Mandatory	Not Available - 1	24
TI uDAPL	Beta	Beta	25
TI Basic RDMA Interop	Beta	Not Available - 3	20
TI RDMA Operations	Beta	Not Available - 3	28
ТІ МРІ НР	Beta	Not Available - 2	29
TI MPI Intel	Beta	Beta	31
TI MPI Open MPI	Beta	Not Available - 2	32
TI MPI OSU	Beta	Not Available - 2	33 34

Not Available means one of three things:

- 1) The feature is not currenlty supported by the WinOF stack
- 2) The ULP appliaction has not been ported to the WinOF Stack
- 3) The test has not updated for WinOF

35 36

37 38

39

40

Table 20 - Ethernet Transport Test Status for April 2009 Interop Event			
Test Procedure	Linux		
Ethernet Link Initialize	Beta		
Ethernet Fabric Initialization	Beta		
Ethernet Fabric Failover	Beta		
Ethernet Fabric Reconvergence	Beta		
iWARP Connectivity	Mandatory		
TI iSER	Beta		
TI NFS over DRDMA	Beta		
TI RDS	Beta		
TI SDP	Beta		
TI uDAPL	Beta		
TI Basic RDMA Interop	Beta		
TI RDMA Operations	Beta		
TI MPI HP	Beta		
TI MPI Intel	Beta		
TI MPI Open MPI	Beta		
TI MPI OSU	Beta		

#### + Test Statu . Annil 2000 Int T-11. 10 . c. .

#### **1.12 SUBJECTS NOT COVERED**

#### Table 27 - SUBJECTS NOT COVERED

Number	Subject/ Feature	Reason	Executor	Due Date
1	iWARP peer to peer	Future Testing		September 2009
2	IPv6 testing	Future Testing		September 2009

#### **1.13 TEST GLOSSARY**

#### Table 28 Test Glossary

Technical Terms	
НСА	IB Host Channel Adapter.
IPoIB	IP over InfiniBand
iSER	iSCSI Extension for RDMA
MPI	Message Passing Interface
RDF	Readme File.
RDS	Reliable Datagram Socket
RNIC	RDMA NIC (iWARP Network Interface Card).
SA	IB Subnet Administration.
SM	IB Subnet Manager.
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol
TD	Test Descriptions.
TI	Transport Independent (tests).
uDAPL	User Direct Access Programming Library

IB HW Units RELEASE 1.28

1 2

3

4 5

6 7

8

21

### 2 GENERAL SYSTEM SETUP Configuration

The test environment for the user interface contains:

#### 2.1 IB HW UNITS

Equipment	Amount	Details	Check
Operating System	12 or more	The OS should be supported by OpenFabrics.	
4X IB Cables	30 or more	Between 1M => 10M.	
IB Switch from a 3rd Party Vendor	6	The number and types of switches needed from OEM is dependent on variations in embedded and subnet management 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.	
InfiniBand 4X Analyzer	1		
IB HCAs	12 or more		

 Table 29
 IB Equipment

#### 2.2 IB SOFTWARE

2.3 2.4

ID OOI WARE	22
2.2.1 LINUX/WINDOWS PLATFORMS	23
2.2.2 OFED/WINOF - MOST CURRENT TESTED RELEASE	24
2.2.3 IB HCA FW – VERSION XXX - VENDOR SPECIFIC	25
2.2.4 IB SWITCH FW CANDIDATE – VERSION XXX - VENDOR SPECIFIC	26
2.2.5 IB Switch SW – Version XXX - Vendor Specific	27
	28
	29
IWARP HW UNITS	30
IWARP SOFTWARE	31
2.4.1 LINUX/WINDOWS PLATFORMS	32
2.4.2 OFED - MOST CURRENT TESTED RELEASE	33
2.4.3 IWARP RNIC FW – VERSION XXX - VENDOR SPECIFIC	34
2.4.4 10GBE SWITCH FW CANDIDATE – VERSION XXX - VENDOR SPECIFIC	35
	30
2.4.5 TUGBE SWITCH SW - VERSION AAA - VENDOR SPECIFIC	37
	38
	39
	40

41

Use of OpenFabrics Software for Pre-Testing RELEASE 1.28

3 USE OF OPENFABRICS SOFTWARE FOR PRE-TESTING				
[	Depending on the schedule of testing and bugs or issues encountered, different snapshots of latest OpenFabrics software will be used during pre-testing prior to	2		
t f	from interoperability Event. Any changes that result in the OpenFabrics software from interoperability testing per this test plan will be deposited back into the OpenFabrics repository so that the OpenFabrics development community will	3		
ŕ	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-	4 5		
ķ	pletion of adequate testing of the said fixes or feature additions.	6		
4 USE OF OPENFABRICS SOFTW	VARE FOR IBTA/CIWG INTEROPERABILITY PLUGFEST	-		
[ t	During the pre-testing phase, UNH-IOL will apply all reasonable effort to ensure that the OpenFabrics source and binary repositories are up-to-date with the	7		
r	results of interoperability testing prior to IBTA/CIWG sponsored interoperability	8		
i i i i i i i i i i i i i i i i i i i	ducted using software directly sourced from the OpenFabrics tree. 9			
5 +	Should there be any issues with the OpenFabrics community not accepting cer-	10		
r	tain bug fixes or features with the time frames matching with plugfest occur- rences, UNH-IOL will inform all participants about the same and offer those bug <sup>11</sup>	11		
t	fixes or features in source code and binary formats directly to the plug fest par- ticipants and InfiniBand solution suppliers.	12		
5 USE OF OPENFABRICS SOFTW	VARE FOR UNH IOL IWARP INTEROPERABILITY PLUGFESTS	13		
[	During the pre-testing phase, UNH IOL will apply all reasonable effort to ensure	14		
r	results of interoperability testing prior to UNH IOL iWARP sponsored interopera- bility plug fest events. This will enable interoperability testing at plug fests to be	15		
C	conducted using software directly sourced from the OpenFabrics tree.	16		
s	Should there be any issues with the OpenFabrics community not accepting cer- tain bug fixes or features with the time frames matching with plug fest occur-	17		
r f	rences, UNH IOL will inform all participants about the same and offer those bug	18		
t	ticipants and iWARP solution suppliers.	19		
		20		
		21		
		22		
		23		
		24		
		25		

6 IB HW DESCRIPTION & CONI	NECTIVITY		1
	The Test contains 2 major parts - this de	escription is for each of those parts.	2
6.1 BASIC CONNECTIVITY (P1F	21)		3
6.1.1 HCA 1 SHOULD BE	CONNECTED FROM PORT 1 TO LOWEST I	PORT NUMBER IN SWITCH	4 5
6.1.2 HCA 2 SHOULD BE	CONNECTED FROM PORT 1 TO HIGHEST		6
6.1.3 BOTH WITH COMPLI	ANT INFINIBAND CABLES		7
6.2 SWITCHES AND SOFTWARE	NEEDED		8
6.2.1 Switches provide	D BY OEMs		9
	It is necessary that Switches provided by versions supported by the Switch OEMs is recommended that OEMs provide six ware supported on the Switches.	OEMs cover the full breadth of software b. Port count is not critical for the tests. It switches covering all variations of soft-	10 11 12 13 14
6.2.2 OPENFABRICS SOF	WARE RUNNING ON HOSTS		15
	Where there are dependencies of OEM	provided and IBTA defined management	16
	agents etc.) with OpenFabrics software	running on Hosts, such software should	17
	be provided to UNH-IOL for interoperab	ility testing. Any known dependencies	18
	should be communicated to ONH-IOL.		20
6.3 CLUSTER CONNECTIVITY			21
6.3.1 HOSTS AND TARGET	S 1-6 SHOULD BE CONNECTED FROM POR	RT 1 OR 2 TO PORTS X IN ALL SWITCHES	22
	Figure 1 - Template for IB Interon S	etun	23
			24
Host orHost oTarget 1Target	r Host or Host or 2 Target 3 Target 4	Host or Host or Target 5 Target 6	26 27
			28
			29
			30
	1	<b>1</b>	32
			33
Switch 1	Switch 2 Switch 3	Switch 4	34
Switch	Switch 2 Switch 3	Switch 4	35
			37
			38
			39
			40
	Switch 5		41
	Owneed		42

7 IWA	RP HW DES		NNECTIVITY				1
		Th	e Test contains 2	major parts - this o	lescription is for ea	ch of those parts.	2
7.1 IW	ARP BASIC		(P1P1)				3
	7.1.1 RNIC	1 ON ONE HOST S			TO RNIC 2 ON AN	OTHER HOST OR TO A	5
10GвE	зwiтсн.						6
	7.1.2 WITH	10GBE CABLES					7
7.2 Sv	VITCHES ANI	D SOFTWARE NE	EDED				8
	7.2.1 Switc	HES PROVIDED BY	r OEMs				9
		It i ve is i on	s necessary that S rsions supported recommended tha the Switch.	Switches provided by the Switch OEM to EMs provide a s	by OEMs cover the ls. Port count is not witch per variations	full breadth of software critical for the tests. It of software supported	10 11 12 13
	7.2.2 OPEN	FABRICS SOFTWA	RE RUNNING ON	RNICs			14
		Wi nir tes	here there are dep ng on RNICs, such sting, and any kno	pendencies of OEM n software should b own dependencies	I provided with Ope e provided to UNH- should be commur	nFabrics software run- IOL for interoperability icated to UNH-IOL.	16 17
7.3 CL		INECTIVITY					10
	7.3.1 Hosts	S AND TARGETS 1	-6 SHOULD BE CO	ONNECTED TO SWI	TCHES USING 10G	BE CABLES.	20
							21
		Figur	e 2 Template fo	or iWARP Intero	p Setup		22
-	Host or Farget 1	Host or Target 2	Host or Target 3	Host or Target 4	Host or Target 5	Host or Target 6	23 24 25
							26
		_	2		2		27
		1					28
	1		1	1		1	30
		↓ 1	<b>v</b>	↓ ↓.	<b>V</b>		31
			• • •		<b>,</b>		32
	Swite	:h 1 S	Switch 2	Switch 3	Switch 4		33
							34
							36
							37
			V V	¥ ¥			38
							39

41 42

Switch 5

	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 SW & HW INSTALLATION				1
8.1 BURNING THE FW				2
8.1.1 PLEASE REFER TO FIRMWAR	E BU	JRNI	NG TOOLS AND PROCEDURES DOCUMENTATION FROM HCA IB VENDOR	3
8 1 2 NO FIRMWARE BURNING REC				4
				5
				6
8.2.1 PLEASE REFER TO SOFTWAR	RE IN	ISTA	LLATION MANUAL FROM HCA IB VENDOR.	7
8.2.2 PLEASE REFER TO SOFTWAR	RE IN	ISTA	LLATION MANUAL FROM RNIC VENDOR.	8
9 GENERAL INSTRUCTIONS				9
				11
3.1 TIKST STEP INSTRUCTIONS	1)	D	the FW release XXX on all UCAs and DNICs using the shows proce	12
	1)	dur	e as required by vendor.	13
	2)	Но	st and Target Configuration	14
	_,	a)	Install OFED software on host systems (using a 64 bit OS) configured to	15
		u)	run OFED.	16
		b)	Install WinOF software on host systems (using a 64 bit OS) configured	17
			to run WinOF.	18
		c)	Configure non-OFED systems for use in the cluster as per the vendors	19
				20
		d)	Configure iSER/SRP targets for use in the cluster as per the vendors in- structions.	21 22
	3)	lns ver	tall the switch or gateway with the candidate SW stack as required by ndor.	23 24
	4)	Bui	n the switch or gateway with the released FW as required by vendor.	25
	5)	Со	nnect the Hosts and Targets to an appropriate switch following the basic	26
	,	cor	inectivity.	27
				28
				29
				30
				31
				32
				33
				34
				30
				37
				38
				39
				40
				41

10 INFINIBAND SPECIFIC INTEROP PROCEDURES USING OFED			1
No tes	o <b>te</b> : L sts. S	INH-IOL has created automated scripts to run many of the OFED based . ee <u>http://www.iol.unh.edu/downloads/OFA/scripts.rar</u>	2 3
10.1 IB LINK INITIALIZE USING OF	ED		4
10.1.1 Disconnect the full topology and select a cable whose length should be a maximum of 15 meters for SDR and 10 meters for DDR when using copper cables.		6	
1)	Ver	ify that no SM is running	0
2)	Co	nnect two devices back to back	9
3)	ssh	to one of the two devices	10
	a)	Run "ibdiagnet -lw 4x" to verify portwidth	11
	b)	Run "ibdiagnet -ls 2.5" to check link speed. Interpret output and com- pare to advertised speed.	12 13
	<b>No</b> oth	te: This command will only produce output if the link speed is anything er than SDR. Keep this in mind during your interpretation of the output.	14
4)	Re	peat steps 1-3 with a different device pairing.	16
	a)	All device pairs must be tested except target to target: HCA to HCA,	17
		HCA to Switch, HCA to Target, Switch to Switch, and Switch to Target.	18
	b)	Each device must link to all other devices in order for the device to pass link init over all.	19
10.1.2 Recommendations			20
In	orde	to determine Switch to Target and Switch to Switch link parameters, run	21
CO	commands from an HCA linked to the switch under test. This does require more		23
	erpre		24
			25
			26
			27
			28
			29
			30
			31
			3Z
			34
			35
			36
			37
			38
			39
			40

10.2 IB FABRIC INITIALIZATION USING OFED		
10.2.1 Architect the Network we wa	ant to build.	2
1	Create a table of IP addresses to assign.	3
2	Create topology file - this makes sure that the subnet is configured as expected - i.e. SDR and DDR links. This inserts name of devices as	4
	well as the GUID.	6
3	See Figure 3- Sample Network Configuration below.	7
		8
10.2.2 Connect the HCAs and switches as per the Architected Network and make sure that no SM/SA is		
running on the Fabric.		10
		12
10.2.3 Procedure		13
1	Start an SM on a device and let it initialize (all SM's will need to be	14
2	Visually verify that all devices are in the active state. Orange led will	15
	be on if the port is active.	16
3	Run "ibdiagnet -pc" to clear all port counters	17
5	Run "ibdiagnet -c 1000" to send 1000 node descriptions.	10
6	Run "ibdiagnet -r" to generate fabric report.	20
7	Run "ibchecknet" to build guid list.	21
10.2.4. Varification Dressdures		22
10.2.4 Vernication Procedures	Boview "DM Counters" costion of the febrie report. There should be	23
I.	no illegal PM counters. The Specification says there should be no er-	24
	rors in 17 seconds.	26
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	27
	switches in the fabric.	28
3	Review the ibchecknet report and verify that there are no duplicate	29
	GUIDS IN the fabric	30
		32
R	estart all devices in the fabric and follow Sections 10.2.3 and 10.2.4.	33
к b	been used. All SMs on managed switches and one instance of <b>opensm</b>	
r	ust be used.	35
-	ash daviaa muat paga all varification procedures with every SM to page	36
E F	ach device must pass an venneation procedures with every SM to pass abric Initialization test.	37
		30
		40
		41
		42

## Table 30 - ibdiagnet commands

Commands	Description
ibdiagnet -h	Help
Ibdiagnet - pc	Clear Counter
Ibdiagnet -lw 4x - ls 2.5	Specify link width and speed
Ibdiagnet -c 1000	send 1000 Node Descriptions



#### Figure 3 - Sample Network Configuration

Last Modified: 3/29/09 8:19 am

10.3 IB IPOIB CONNECT MODE	(C	M) USING OFED	1
10.3.1 Setup			2
	Co	nnect the HCAs and switches as per the Architected Network and make sure	3
	tha	t no SM is running on the Fabric.	4
	Thi	s 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	7
	WO	uld only run SM/SA for the partner pair (with a switch in the middle). This pro-	8
	cec	edure has been developed for Linux and may be ported to Windows if there	
	suf	ficient vendor support.	10
	<b>Optional</b> : In the procedures below, an IB analyzer can be inserted in the apr		11
	pria	priate link to obtain traces and validate the aspects of the procedures specifically	
	det	ailed below in subsequent sections.	13
10.3.2 IPoIB Interface Creation and IPoIB Subnet Creation		14	
	1)	Configure IPoIB address. All addresses must reside on the same subnet.	15
	,	a) Set interfaces to 10.0.0.x/24 (10.0.0.x/netmask 255.255.255.0) using	10
	the command <i>ifconfig ib0 10.0.0.x netmask 255.255.255.0</i>		18
10.3.3 .Bringing the IPoIB in Connected Mode		19	
	1)	Set "SET_IPOIB_CM=yes" in file /etc/infiniband/openib.conf	20
	2)	Restart driver "/etc/init.d/openibd restart"	21
	3)	Validate CM mode by checking that "/sys/class/net/ <i f="" name="">/mode" equal</i>	22
		to 'connected'	23
	4)	Repeat steps 1-3 in section 10.3.3 on all nodes being tested.	24
			25
10.3.4 Ping Procedures			26
Sten A	1)	Stop all SM's and verify that none are running	28
	י) 2)	Bower cycle all switches in the fabric (this insures that the new SM will con	20
	2)	figure all the links and create the multi-cast join).	30
	3)	Start an SM (All SM's will need to be tested) and let it initialize	31
		a) Visually verify that all devices are in the active state. Orange led will be	32
	on if the po	on if the port is active.	33
		b) Run "ibdiagnet -r" and verify that the SM you started is the one that is	34
	running and and that it is the master. You will need to know the	running and and that it is the master. You will need to know the GUID of the device since the SM will be reassigned on each report	35
		c) Verify that all nodes and switches were discovered	36
		<ul> <li>Verify that all houses and switches were discovered.</li> </ul>	37
		number of switch platforms present. This is because some switches have	30 20
		multiple switch chips.	40
		Examine the arp table (via arp -a) and remove the destination node's ib0 a	
		dress from the sending node's arp table (via arp -d).	42
OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		IB IPoIB Connect Mode (CM) using OFED March 27, 2009 RELEASE 1.28 DRAFT	_
--	------------------	---	----------------
	5)	Ping every HCA except localhost with packet sizes of 64, 256, 511, 512, 1024, 1025, 2044, 4096, 8192, 16384, 32768, and 65507.	1 2
		a) 100 packets of each size will be sent	3
		b) Repeat step #4 before issuing each ping command. Every packet size is a new ping command.	4 5
	6)	In order to pass Step A, a reply must be received for every ping sent (without losing a single packet) while using each one of the SMs available in the cluster.	6 7 8
Step B	1)	Bring up all HCAs but one.	9
	2)	Start an SM (all SMs will need to be tested).	10
	3)	Check for ping response between all node (All to All).	11
		a) A response from the disconnected HCA should not be returned.	12
	4)	Disconnect one more HCA from the cluster.	13
	5)	Ping to the newly disconnected HCA from all nodes (No response should be returned).	14 15
	6)	Connect the first machine (the one that was not connected) and check for ping response from all nodes that are still connected.	16 17
	7)	Connect the disconnected HCA to a different switch on the subnet which will change the topology.	18 19
	8)	Ping again from all nodes (this time we should get a response).	20
	9)	Follow Step B, this time bring the interface down and then back up using if- config ibX down and ifconfig ibX up commands instead of physically discon- necting the HCAs.	21 22 23
		<b>Note</b> : Each step must exhibit the expected behavior while using each SM in order for the device to pass Step B overall.	24 25
Step C	Fo ins for	low Step A and B using a different SM until all SM's have been used. Only one tance of each available SM is required. Steps A, B, and C must pass in order the device to pass 10.3.4 overall.	26 27 28
10.3.5 SFTP PROCEDURE			29
	SF	TP procedures require an SFTP server to be configured on each machine in	30
	the	partner pair. An SFTP client needs to be available on each machine as well.	31
	١h	e default RHEL install includes both.	32
	A 4 par	MB file will be SFTP'd to the partner and then SFTP'd back and binary com- red to the original file, this will be done in each direction and then bidirectional	33 34 35
	usi	ng every Sivi available.	36
10.3.5.1 SETUP			37
	1)	Make sure vsftpd is installed on each node for SFTP application.	38
	, 2)	A special account for this should be created as follows:	39
	_,	b) Username: Interop	40
		c) Password openfabrics	41
			42

10.3.5.2 PROCEDURE		1
1)	Run SFTP server on all nodes.	2
2)	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 you started.	4
3)	SFTP:	6
	a) Connect an HCA pair via SFTP on IPoIB using the specified user name and password.	7
	b) Put the 4MB file to the /tmp dir on the remote host.	9
	c) Get the same file to your local dir again.	10
	d) Compare the file using the command <i>cmp tfile tfile.orig</i> .	11
	i) The two must be identical	12
4)	Repeat the procedure with a different SM.	13
	<b>Note</b> : Every node must SFTP the 4MB file to all others using all SM's and the	14
	files must be identical as determined by the binary compare in order for the	15
	device to pass 10.3.5 overall.	16
	<b>Note</b> : Sections 10.3.4 and 10.3.5 must pass using the configuration determined by sections 10.3.1 10.3.2 and 10.3.3 for the device to pass IPoIB	17
	Connected mode overall.	19
		20
		21
		22
		23
		24
		25
		26
		21
		29
		30
		31
		32
		33
		34
		35
		00
		36

- 39 40
- 41
- 42

10.4 IB IPOIB DATAGRAM MOD	DE (	DM) USING OFED	1
10.4.1 Setup			2
	Coi tha	nnect the HCAs and switches as per the Architected Network and make sure t no SM is running on the Fabric.	3 4
			5
	Thi An	s procedure, as the previous ones, will be based on the cluster connectivity. SM/SA which supports IPoIB (sufficient IB multicast support) will be running	6
	on	the HCAs, or on a switch with an embedded SM/SA or a third HCA which	0
	cec	luid only run SM/SA for the partner pair (with a switch in the middle). This pro-	0
	suf	ficient vendor support.	10
	_		11
	Ор	<b>itional</b> : In the procedures below, an IB analyzer can be inserted in the appro-	12
	pria	ate link to obtain traces and validate the aspects of the procedures specifically alled below in subsequent sections	12
	uct		14
10.4.2 IPoIB Interface Creation a	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
10.4.3 .Bringing the IPoIB in Data	agra	am Mode	19
	1)	Set "SET_IPOIB_CM=no" in file /etc/infiniband/openib.conf	20
	2)	Restart driver "/etc/init.d/openibd restart"	21
	3)	Validate DM mode by checking that "/sys/class/net/ <i f="" name="">/mode" equal</i>	22
	-,	to 'datagram'	23
	4)	Repeat steps 1-3 in section 10.4.3 on all nodes being tested.	24
			25
			26
10.4.4 Ping Procedures			27
Step A	1)	Stop all SM's and verify that none are running	28
	2)	Power cycle all switches in the fabric (this insures that the new SM will con-	29
		figure all the links and create the multi-cast join).	30
	3)	Start an SM (All SM's will need to be tested) and let it initialize	31
		a) Visually verify that all devices are in the active state. Orange led will be	32
		on if the port is active.	33
		b) Run "ibdiagnet -r" and verify that the SM you started is the one that is	34
		the device since the SM will be reassigned on each reboot.	35
		c) Verify that all nodes and switches were discovered.	37
		<b>Note:</b> Ibdiagnet may show more switches than indicated by the physical	38
		number of switch platforms present. This is because some switches have	39
		multiple switch chips.	40
	4)	Examine the arp table (via arp -a) and remove the destination node's ib0 ad-	41
		dress from the sending node's arp table (via arp -d).	42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		B IPolB Datagram Mode (DM) using OFED March 27, 2009 RELEASE 1.28 DRAFT	_
	5)	Ping every HCA except localhost with packet sizes of 64, 256, 511, 512, 1024, 1025, 2044, 4096, 8192, 16384, 32768, and 65507.	1 2
		a) 100 packets of each size will be sent	3
		b) Repeat step #4 before issuing each ping command. Every packet size is a new ping command.	4 5
	6)	In order to pass Step A, a reply must be received for every ping sent (without losing a single packet) while using each one of the SMs available in the cluster.	6 7 8
Step B	1)	Bring up all HCAs but one.	9
	2)	Start an SM (all SMs will need to be tested).	10
	3)	Check for ping response between all node (All to All).	11
		a) A response from the disconnected HCA should not be returned.	12
	4)	Disconnect one more HCA from the cluster.	13
	5)	Ping to the newly disconnected HCA from all nodes (No response should be returned).	14 15
	6)	Connect the first machine (the one that was not connected) and check for ping response from all nodes that are still connected.	16 17
	7)	Connect the disconnected HCA to a different switch on the subnet which will change the topology.	18 19
	8)	Ping again from all nodes (this time we should get a response).	20
	9)	Follow Step B, this time bring the interface down and then back up using if- config ibX down and ifconfig ibX up commands instead of physically discon- necting the HCAs.	21 22 23
		<b>Note</b> : Each step must exhibit the expected behavior while using each SM in order for the device to pass Step B overall.	24 25
Step C	Fo ins for	low Step A and B using a different SM until all SM's have been used. Only one tance of each available SM is required. Steps A, B, and C must pass in order the device to pass 10.4.4 overall.	26 27 28
10.4.5 SFTP PROCEDURE			29
	SF	TP procedures require an SFTP server to be configured on each machine in	30
	the	partner pair. An SFTP client needs to be available on each machine as well.	31
	In	e default RHEL Install includes both.	32
	A 4 pai	MB file will be SFTP'd to the partner and then SFTP'd back and binary com- red to the original file, this will be done in each direction and then bidirectional ng every SM available	33 34 35
	031		36
10.4.5.1 SETUP			37
	1)	Make sure vsftpd is installed on each node for SFTP application.	38
	2)	A special account for this should be created as follows:	39
		b) Username: Interop	40
		c) Password: openfabrics	41
		· ·	42

10.4.5.2 PROCEDURE			1
	Run	SFTP server on all nodes.	2
	1)	Start an SM (all SM's will need to be tested) and let it initialize	3
	1)	Norify that the supping SM is the one you started	4
	0)		6
	2)		7
		a) Connect an HCA pair via SFTP on IPoIB using the specified user name and password.	8
		p) Put the 4MB file to the /tmp dir on the remote host.	9
		c) Get the same file to your local dir again.	10
		d) Compare the file using the command <i>cmp tfile tfile.orig</i> .	12
		i) The two must be identical	13
	3)	Repeat the procedure with a different SM.	14
		Note: Every node must SFTP the 4MB file to all others using all SM's and the	15
		iles must be identical as determined by the binary compare in order for the device to pass 10.4.5 overall.	16 17
		<b>Note</b> : Sections 10.4.4 and 10.4.5 must pass using the configuration deter-	18
		The doy sections 10.4.1, 10.4.2, and 10.4.3 for the device to pass IPOIB	19
			20
			21
			22
			23
			24
			20 26
			20
			28
			29
			30
			31
			32
			33
			34
			35
			36
			১/ २२
			39
			40

10.5 IB SM FAILOVER AND HANDO	OVE	R PROCEDURE USING OFED	1
10.5.1 SETUP			2
1)	С	onnect 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	e from the same vendor. They will be tested pairwise; two at a time.	5
10.5.2 PROCEDURE			6
1)	Di pa	sable all SMs in the cluster then start a SM on either machine in a chosen ir.	7 8
2)	R	un "saquery" on a node in the fabric.	9
	a)	Verify that all nodes in the cluster are present in the output	10
3)	U: m	sing the ibdiagnet tool with the -r option, verify that the running SM is the aster.	12
4)	St	art a SM on the second machine in the current pair.	13
5)	Ve	erify that the SMs behave according to the SM priority rules. Use "ibdi-	14
3)	ag	gnet -r" again.	15
	a)	SM with highest numerical priority value is master and the other is in	10
	,	standby.	18
	a)	If both SMs have the same priority value then the SM with the smallest guid is master and the other is in standby.	19
6)	R	un "saquery" on either machine in the current pair.	20
	a)	Verify that all nodes in the cluster are present in the output.	21
7)	SI	nutdown the master SM.	23
8)	Ve	erify the other active SM goes into the master state using "ibdiagnet -r"	24
,	ag	jain.	25
9)	R	un "saquery" on either machine in the current pair.	26
	a)	Verify that all nodes in the cluster are present in the output.	27
10	) St	art the SM you just shutdown.	28
11)	, ) Ve	erify that the newly started SM resumes it's position as master while the	29
,	ot	her goes into standby again.	30
12)	) R	un "saquery" on either machine in the current pair.	31
	a)	Verify that all nodes in the cluster are present in the output.	32
13	) SI	nutdown the standby SM.	33
14	, ) Ve	erify that the previous master SM is still the master.	34
15	) R	un "saguery" on either machine in the current pair.	35
	, ຊ)	Verify that all nodes in the cluster are present in the output	30
16	ч) \ D	verify that an nodes in the statistic present in the supprise	38
IO,	m m	et (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

OFA Interoperability Working Group	IB SM Failover and Handover Procedure using OFED	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLA	N RELEASE 1.28	DRAFT

c) Both SMs having equal num	erical priority values.
17) Repeat steps 1-16 until all possib	ble SM pairs from identical vendors in the 2
cluster have been tested.	3
18) All of the "saquery" commands m	nust return the expected list of nodes in 4
	eceive a passing grade. 5
	0
	/
	9
	10
	11
	12
	13
	14
	15
	16
	17
	18
	19
	20
	22
	23
	24
	25
	26
	27
	28
	29
	31
	32
	33
	34
	35
	36
	37
	38
	39
	40
	41
	42

10.6 IB SRP USING OFED			1
10.6.1 SETUP			2
	Co	nnect the HCAs and switches as per the Architected Network and make sure	3
	tha	t no SM is running on the Fabric.	4
			5
	1)	Start an SM (all SM's will need to be tested) and let it initialize	6
	1)	a) Verify that the running SM is the one that you started	/ 8
	2)	Choose a node to work with	9
	2) 3)		10
	3) 4)	Load ern module with ern, eg. tablesize=255	11
	4)	a) <b>Example:</b> module with sip_sg_tablesize=235	12
		a) Example. mouprobe ib_sip sip_sg_tablesize=255	13
	5)	b) Let it initialize	14
	5)		15
	$\sim$	a) <b>Example</b> : ismod   grep ib_srp	16
	6)	Load srp_daemon with -e -o -n options	18
		a) Example: srp_daemon -e -o -n	19
		b) Let it initialize	20
	7)	Find all volumes from all targets	21
		a) Use Isscsi	22
	8)	Perform 6GB read from srp volume to null	23
		a) <b>Example</b> : dd if=\$drive of=/dev/null count=600 bs=10M	24
	9)	Perform 6GB write from zero to srp volume	25
		a) <b>Example</b> : dd if=/dev/zero of=\$drive count=600 bs=10M	26
	10)	Repeat step #8 and #9 for all volumes found for all targets as determined by step #7	27
	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.	3Z
		Note: An HCA must successfully complete all DD operations to and from all	34
		One volume per target is all that is required.	35
			36
			37
			38
			39
			40
			41

## **10.7 IB ETHERNET GATEWAY USING OFED**

### 10.7.1 Procedure

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-	2
	work or Ethernet device. Start the SM to be used in this test.	Ę

- 2) Determine which ULP your ethernet gateway uses and be sure that ULP is running on the host (VNIC or IPoIB).
- Restart the ULP or using the tool provided by the ULP, make sure that the host "discovers" the Ethernet Gateway. Configure the interfaces and make sure they are up.
- 4) Run ping from the host to the Ethernet device. While the ping is running, kill the master SM. Verify that the ping data transfer is unaffected.
- 5) Reboot the Ethernet Gateway. After the Ethernet Gateway comes up, verify that the host can discover the Ethernet Gateway as it did before and we are able to configure the interfaces.
- 6) 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.
- Unload the ULP used by Ethernet Gateway and check that the Ethernet Gateway shows it disconnected. Load the ULP and verify that the Ethernet gateway shows the connection.
- 8) Repeat step 4 by using ssh and scp instead of ping.

# **10.8 IB FIBRECHANNEL GATEWAY USING OFED**

# 10.8.1 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
•		6
2)	Configure the FC Gateway appropriately (now to do this is vendor specific).	7
3)	Use ibsrpdm tool in order to have the host "see" the FC storage device. Add the storage device as target.	8 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. While the test is running, kill the master SM. Verify that the test completes properly.	11 12 13
6)	Unload the SRP host / SRP Target (target first/host first) and check that the SRP connection is properly disconnected.	14
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 target.	19 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 be tested, until each HCA and each SM has been tested with the FC	22
	Gateway.	20
		25
		26
		27
		28
		29
		30
		31
		32
		33
		34
		35
		36
		37
		38
		39
		40
		41

11 ETHERNET SPECIFIC INTEROP PROCEDURES USING OFED		1	
<b>11.1 ETHERNET LINK INITIALIZE</b>	11.1 ETHERNET LINK INITIALIZE USING OFED		
11.1.1 PURPOSE			3
	Th	e Ethernet Link Initialize test is a validation that all Ethernet devices receiving	4
	the	OFA Logo can link and pass traffic under nominal (unstressed) conditions.	5
11 1 2 Resource Requirements			0
	1)	Gigabit or 10 Gigabit Ethernet RNIC	8
	·) 2)	Gigabit or 10 Gigabit Ethernet Switch	9
	2) 3)		10
	5)	Compliant Cables	11
	ть	a validation of the underlying transport infractructure is accordial to the and	12
	USe	ers experience of the operation of the OFED software stack. To this end, this	13
	tes	t confirms that Ethernet devices receiving the OFA Logo can suitably link and	14
	pas	ss traffic in any configuration. Exhaustive compliance testing of BER perfor-	15
	eve	er, successful completion of this test provides further evidence of the	16
	rob	ustness of the OFA logo bearing device.	18
11.1.4 Procedure			19
	1)	Connect the two link partners together utilizing compliant 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
	1)	Beneat station is required).	25
	4)	switch, and RNIC to RNIC link partner combinations. Previously tested com-	26
		binations resident in the OFILG cluster may be omitted.	28
11.1.5 OBSERVABLE RESULTS			29
	1)	Link should be established on both ends of the channel.	30
	2)	Traffic should pass in both directions. Error rates of 10e-5 or better should	31
		be readily confirmed (no lost frames in 10,000).	32
11.1.6 POSSIBLE PROBLEMS			33
	1)	Traffic directed to a switches IP management address may not be pro-	34
		switch to a remote responder.	35
			36
			31 20
			30
			40

11.2 ETHERNET FABRIC INITIALIZE USING OFED							
11.2.1 PURPOSE			2				
	The E	thernet Fabric Initialization test is a validation that all Ethernet devices re-	3				
	ceiving the OFA Logo properly interoperate with common OSI Layer 2 p						
	incluc	uding Link Aggregation, RSTP, and MSTP under nominal (unstressed) con-					
	10115.		6				
11.2.2 Resource Requirements			7				
	1) G	igabit or 10Gigabit Ethernet RNIC,	8				
	2) G	igabit or 10Gigabit Ethernet Switch	9				
	3) C	ompliant Cables	11				
11.2.3 Discussion			12				
	The v	alidation of the underlying transport infrastructure is essential to the end-	13				
	users	experience of the operation of the OFED software stack. To this end, this	14				
	test c	onfirms that Ethernet devices receiving the OFA Logo can suitably form link	15				
	MSTP in the selected plugfest or cluster Network Architecture configuration. Ne						
	ther exhaustive interoperability configuration permutations nor IEEE 802.1 com-						
	pliance of this	e testing is performed as part of this test; however, successful completion	18				
	device.						
			20				
	Note: lvina	IP Connectivity is desired to ensure connectivity is stable and that under-	21				
	RDM/	A traffic. RDMA traffic is desired to observe the effects of topology changes	22				
	on the	iWARP protocol.	23				
11 2 4 Procedure							
	1) A	rchitect the desired network from available cluster and plugfest partici-	26				
	р, р	ants, similar to that shown in the Cluster Connectivity Section 7.3. All ca-	27				
	b	ing must be compliant cables. Most RNIC-to-RNIC paths should traverse 2	28				
	0	more switches.	29				
	а	Create a table of IP addresses to assign to RNICs and switch management entities	30				
	h	When MSTP is supported, create a VI AN topology with at least 2	31				
	b	VLANs (high and normal priority) Create 802.1q VLAN trunk links be-	32				
		tween supporting switches.	33				
	C	When Link Aggregation is supported by both link partners, create a 2-4 channel link aggregate between the link partners.	34 35				
		<b>Note</b> : This includes RNICs supporting Link Aggregation, as well as switch to switch links.	36 37				
	d	Set spanning tree priorities such that desired bridge(s) becomes root bridge(s).	38 39				
	е	See <u>Cluster Connectivity</u> .	40				

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		Ethernet Fabric Initialize using OFED RELEASE 1.28	March 27, 2009 DRAFT
	2)	Connect the RNICs and switches as per the Arch sure that desired bridge is the root bridge (in the priate bridge per VLAN) is running on the Fabric.	nitected Network and make case of MSTP, the appro-
	3)	Verify IP connectivity to all IP attached stations in minimum size ICMP echo requests from all RNIC verify cluster connectivity.	the Cluster. Source 1000 to all other IP entities to
	4)	Verify RDMA connectivity to all RDMA attached s Source 100 2k RDMA reads from each RNIC to a cluster RDMA connectivity.	stations in the Cluster. all other RNICs to verify
11.2.5 OBSERVABLE RESULTS			
	1)	In all cases, the desired root bridge (or in the cas desired root bridges) should always become the	e of MSTP topologies, the root bridge.
	2)	IP connectivity should occur to all stations without	it loss of responses.
	3)	RDMA connectivity should occur to all stations w	ithout loss of responses.

Vlan 2

#### - Ethernet RDMA Link / or Link Aggregate Legend: Ethernet non-RDMA Addressing (CX4, SR, LRM, etc) IP Address 132.177.125.192 /26 - Spanning Tree for VLAN1, tagged or untagged **Ethernet RDMA Addressing** IP Address 10.10.0.xx /16 - Spanning Tree for VLAN2, tagged or untagged 125.210 125.211 125.212 125.213 125.214 125.215 125.216 125.217 125.218 10.10.0.1 10.10.0.2 10.10.0.3 10.10.0.4 10.10.0.5 10.10.0.6 10.10.0.7 10.10.0.8 10.10.0.9 Vendor Vendor Vendor Vendor Vendor Vendor Vendor Vendor Vendor RNIC F RNIC A RNIC B RNIC C RNIC D RNIC E RNIC G RNIC H iSER Rhea Atlas Telesto Titan Miranda Phoebe Hyperion Calypso Target Untagged Tagged Untagged Vlan 1 Untaggeo Vlan 2 Untagged Untagged Vlan 2 Untagged Tagged Untagged Untagged Vlan 2 Vlan 1 Vlan 1 & 2 Vlan 1 & 2 Vlan 1 Vlan 1 3 2 2 4 5 2 1 2 3 3 1 1 4 Switch C Switch D Switch E Switch F Switch G Switch H 125.202 125.203 125.204 125.205 125.206 125.207 5 6 5 3 2 6 4 1 2 3 1 Λ Simulated Cable Fault Tagged Vlan links 65 32 6 5 4 3 2 1 4 1 Switch Switch А В 125.200 125.201 Vlan 1 MSTP Vlan 2 MSTP Root Brige Root Brige **Note on final Network Architecture** Dependent on: RNIC VLAN tag support, Link Agg support, MSTP support (RSTP support is assumed) and Port Type. Layer-3 Routing will not be utilized.

11.3.1 PURPOSE       2         The Ethermet Fabric Reconvergence test is a validation that all Ethermet devices receiving the OFA Logo properfy converge in the event of a topology change in a timely manner, minimally impacting the fabric.       3         11.3.2 Resource Requirements       1       Gigabit or 10Gigabit Ethermet RNICs       7         2)       Gigabit or 10Gigabit Ethermet RNICs       7       7         3)       CX4 Cables       7         11.3.3 Discussion       7       7       7         11.3.4 Procedure       7       7       7         Note: IP Connectivity is desired to ensure connectivity is stable and that underlying frampoint infrastructure is especiate, and/or an RSTP or MSTP per the selected pulgets or culser Network Architecture configuration. Neither exhaustive topology change effecting a link aggregate, and/or an plance testing is performed as part of this test; however, successful completion of the OFA logo bearing device.       7         11.3.4 Procedure       1       7       7         11.3.4 Procedure       1       9       7	11.3 ETHERNET FABRIC RECONVERGENCE USING OFED					
The Ethernet Fabric Reconvergence test is a validation that all Ethernet devices receiving the OFA Logo properly converge in the event of a topology change in a timely manner, minimally impacting the fabric.       3         11.3.2 Resource Requirements       •         •       •       0 Gigabit or 10Gigabit Ethernet RNICs       7         •       •       0 Gigabit or 10Gigabit Ethernet Switches       7         •       •       0 Gigabit or 10Gigabit Ethernet Switches       7         •       •       0 Gigabit or 10Gigabit Ethernet Switches       7         •       •       0 Gigabit or 10Gigabit Ethernet Switches       7         •       •       0 CX4 Cables       7         11.3.3 Discussion       •       •       7         11.3.4 Procedure       The validation of the underlying transport infrastructure is essential to the endustrustive topology change permutations nor IEEE 802.1 compliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.       7         •       Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of the fabric.       7         11.3.4 Procedure       •       •       •       7         11.3.4 Procedure       •       •       •       7 <th colspan="6">11.3.1 PURPOSE</th>	11.3.1 PURPOSE					
<ul> <li>receiving the OFA Logo properly converge in the event of a topology change in a timely manner, minimally impacting the fabric.</li> <li>11.3.2 Resource Requirements         <ol> <li>Gigabit or 10Gigabit Ethernet RNICs</li> <li>Gigabit or 10Gigabit Ethernet Switches</li> <li>CX4 Cables</li> </ol> </li> <li>11.3.3 Discussion         <ol> <li>The validation of the underlying transport infrastructure is essential to the endusers experience of the operation of the OFED software stack. To this end, this test confirms that Ethernet devices receiving the OFA Logo can suitably reconverge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture configuration. Netther exhaustive topology change effecting a link aggregate, and/or an estimate the start of the robustness of the OFA logo bearing device.</li> </ol> </li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.</li> <li>Note: IP Connectivity remains that of the selected Network Architecture configuration.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restor et pervolusly removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restor et he fabric and flows Steps 1-8 each time with adifferent switch in the fabric and flow</li></ul>		The	e Ethernet Fabric Reconvergence test is a validation that all Ethernet devices	3		
11.3.2 Resource Requirements       1         1) Gigabit or 10Gigabit Ethernet RNICs       7         2) Gigabit or 10Gigabit Ethernet Switches       9         3) CX4 Cables       10         11.3.3 Discussion       11         11.3.4 Procedure       11         11.3.5 Discussion       11         11.3.4 Procedure       11         11.3.5 Discussion       11         11.3.4 Procedure       11         11.3.4 Procedure       12         11.3.4 Procedure       11         11.3.4 Procedure       11         11.3.4 Procedure       12         11.3.4 Procedure       12         11.3.4 Procedure       11         11.3.4 Procedure       11         11.3.4 Procedure       11         11.3.4 Procedure       12         11.3.4 Procedure		rec	eiving the OFA Logo properly converge in the event of a topology change in a	4		
11.3.2 Resource Requirements <ul> <li>Gigabit or 10Gigabit Ethernet RNICs</li> <li>Cigabit or 10Gigabit Ethernet Switches</li> <li>CX4 Cables</li> </ul> 11.3.3 Discussion <ul> <li>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 test confirms that Ethernet devices receiving the OFA Logo can suitably recomparison of the operation of the OFED software stack. To this end, this test confirms that Ethernet devices receiving the OFA Logo can suitably recomparison of this test, however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.         Note: IP Connectivity is desired to ensure connectivity is stable and that underly fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the WARP protocol.          <ul> <li>Power off all switches. Connectivity remains that of the selected Network Architect configuration.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Create a new redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is m</li></ul></li></ul>		tim	ely manner, minimally impacting the fabric.	5		
1) Gigabit or 10Gigabit Ethernet RNICs       8         2) Gigabit or 10Gigabit Ethernet Switches       9         3) CX4 Cables       10         11.3.3 Discussion       11         11.3.4 Discussion       11         11.3.4 Procedure       1         11.3.4 Procedure       1 </th <th>11.3.2 Resource Requirements</th> <th></th> <th></th> <th>6</th>	11.3.2 Resource Requirements			6		
<ul> <li>2) Gigabit or 10Gigabit Ethernet Switches</li> <li>3) CX4 Cables</li> <li>11.3.3 Discussion</li> <li>11.3.4 Procedure</li> <li>11.3.</li></ul>	•	1)	Gigabit or 10Gigabit Ethernet RNICs	7		
<ul> <li>a) CX4 Cables</li> <li>b) evaluation of the underlying transport infrastructure is essential to the end- users experience of the operation of the OFED software stack. To this end, this test confirms that Ethernet devices receiving the OFA Logo can suitably reconfigu- tation. Neither exhaustive topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster. Network Architecture config- uration. Neither exhaustive topology change permutations nor IEEE 802.1 com- pliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.</li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.</li> <li>11.3.4 Procedure</li> <li>1) Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.</li> <li>2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>3) Power up all switches. Verify IP connectivity of connected nodes.</li> <li>4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>3) Remove a redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming of top alsone of the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity i</li></ul>		, 2)	Gigabit or 10Gigabit Ethernet Switches	8		
11.3.3 Discussion       10         11.3.3 Discussion       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 test confirms that Ethernet devices receiving the OFA Logo can suitably recon- verge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture config- uration. Neither exhaustive topology change permutations nor IEEE 802.1 com- pliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.         Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.         11.3.4 Procedure       1         1       Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.         2)       Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.         3)       Power up all switches. Verify IP connectivity of connected nodes.         4)       Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.         5)       Remove a redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes.         6)       Create a new redundain switch-to-switch interconne		-/	CX4 Cables	9		
<ul> <li>The validation of the underlying transport infrastructure is essential to the endusers experience of the operation of the OFED software stack. To this end, this test confirms that Ethernet devices receiving the OFA Logo can suitably reconverge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture configuration. Neither exhaustive topology change permutations nor IEEE 802.1 compliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.</li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.</li> <li>11.3.4 Procedure         <ol> <li>Power off all switches. Connectivity remains that of the selected Network Architect configuration.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Power up all switches. Verify IP connect. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintaine</li></ol></li></ul>	11 3 3 Discussion	0)		11		
<ul> <li>Intervalication of the underlying transport infrastructure is escarching to the endo- users experience of the operation of the OFED softwars stack. To this end, this test confirms that Ethernet devices receiving the OFA Logo can suitably recon- verge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture config- uration. Neither exhaustive topology change permutations nor IEEE 802.1 com- pliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.</li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.</li> <li>Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all</li></ul>	TI:3.3 DISCUSSION	Th	a validation of the underlying transport infractructure is accordial to the and	12		
test confirms that Ethernet devices receiving the OFA Logo can suitably reconverge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture configuration. Neither exhaustive topology change permutations nor IEEE 802.1 compliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device. Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.  11.3.4 Procedure 1) Power off all switches. Connectivity remains that of the selected Network Architect configuration. 2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric. 3) Power up all switches. Verify IP connectivity of connected nodes. 4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is maintained to all nodes. 5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes. 7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 9) Restart all devices in the fabric as the desired root bridge. Repeat, as time allows, application and removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 9) Restart all devices in the fabric can be desired root bridge. Repeat, as time allows, application and removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 9) Restart all devices in the fab			ers experience of the operation of the OFED software stack. To this end, this	13		
<ul> <li>verge in the event of a topology change effecting a link aggregate, and/or an RSTP or MSTP per the selected plugfest or cluster Network Architecture configuration. Neither exhaustive topology change permutations nor IEEE 802.1 compliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.</li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the IWARP protocol.</li> <li>11.3.4 Procedure</li> <li>11.3.4 Procedure</li> <li>11 Power off all switches. Connectivity remains that of the selected Network Architect configuration.</li> <li>2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>3) Power up all switches. Verify IP connectivity of connected nodes.</li> <li>4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, and and and and and and and and and and</li></ul>		tes	t confirms that Ethernet devices receiving the OFA Logo can suitably recon-	14		
<ul> <li>Ites if of WS IP per the selected pluglest of cluster Network Architecture conliguration. Net lifet exhaustive topology change permutations nor IEEE 802.1 completion of this test provides further evidence of the robustness of the OFA logo bearing device.</li> <li>Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.</li> <li>11.3.4 Procedure</li> <li>1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.</li> <li>2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>3) Power up all switches. Verify IP connectivity of connected nodes.</li> <li>4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>5) Remove a redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired rool bridge. Repeat, as time allows, and a stop of the absence of RSTP/MSTP).</li> </ul>		ver	ge in the event of a topology change effecting a link aggregate, and/or an	15		
pliance testing is performed as part of this test; however, successful completion of this test provides further evidence of the robustness of the OFA logo bearing device.       17         Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic. BMA traffic is desired to observe the effects of topology changes on the iWARP protocol.       20         11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.       25         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       31         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is maintained to all nodes.       31         5) Remove a redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       37         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         9) Restart all devices i		ura	tion. Neither exhaustive topology change permutations nor IEEE 802.1 com-	16		
of this test provides further evidence of the robustness of the OFA logo bearing device. 19 Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the IWARP protocol. 23 11.3.4 Procedure 11.3.4 Procedure 24 1) Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration. 26 27 the fabric. 28 3) Power up all switches. Verify IP connectivity of connected nodes. 29 4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes. 31 5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes. 36 6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes. 36 7) Remove a channel from a link aggregate. Verify IP and RDMA con- nectivity is maintained to all nodes. 36 80 Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 36 80 Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 36 80 Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 37 80 Restart all devices in the fabric and follow Steps 1-8 each time with a dif- ferent switch in the fabric as the desired root bridge. Repeat, as time allows, 40 81 82 82 83 83 84 84 85 84 85 85 85 85 85 85 85 85 85 85 85 85 85		plia	ince testing is performed as part of this test; however, successful completion	17		
device.       19         Note: IP Connectivity is desired to ensure connectivity is stable and that under- lying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.       22         11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.       26         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       29         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.       31         5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.       33         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA con- nectivity is maintained to all nodes.       36         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         9) Restart all devices in the fabric as the desired root bridge. Repeat, as t		of t	his test provides further evidence of the robustness of the OFA logo bearing	18		
Note: IP Connectivity is desired to ensure connectivity is stable and that underlying fabric issues (such as link flapping) are not masked by TCP transport of RDMA traffic. RDMA traffic is desired to observe the effects of topology changes on the iWARP protocol.       23         11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.       26         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       29         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.       30         5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.       31         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       37         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, and and point the fabric as the desired root bridge. Repeat, as time allows, and and point the fabric as the desired root bridge. Repeat, as time allows, and and po		device.				
lying fabric issues (such as link flapping) are not masked by TCP transport of       21         RDMA traffic. RDMA traffic is desired to observe the effects of topology changes       22         on the iWARP protocol.       23         11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.       26         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       29         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.       31         5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.       31         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       36         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       37         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         9) Restart all devices in the fabric as the desired root bridge. Repeat, as time allows, the fabric as the desired root b		Note: IP Connectivity is desired to ensure connectivity is stable and that under				
<ul> <li>11.3.4 Procedure</li> <li>11.3.4 Procedure</li> <li>1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.</li> <li>2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>2) Disconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>3) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows,</li> </ul>		lyin	g fabric issues (such as link flapping) are not masked by TCP transport of	21		
11.3.4 Procedure       24         11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.       25         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       29         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.       31         5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.       33         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       37         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       38         9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, for the fabric as the desired root bridge. Repeat, as time allows, for the fabric as the desired root bridge. Repeat, as time allows, for the fabric as the desired root bridge.		RD on	MA traffic. RDMA traffic is desired to observe the effects of topology changes the iWARP protocol	22		
11.3.4 Procedure       24         1) Power off all switches. Connectivity remains that of the selected Network Architect configuration.       26         2) Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.       27         3) Power up all switches. Verify IP connectivity of connected nodes.       29         4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.       30         5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.       33         6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.       36         7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       37         8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.       39         9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 40		on		23		
<ol> <li>Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.</li> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restart all devices in the fabric and follow Steps 1-8 each time with a dif- ferent switch in the fabric as the desired root bridge. Repeat, as time allows,</li> </ol>	11.3.4 Procedure			24		
<ol> <li>Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.</li> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, the start all devices in the fabric as the desired root bridge. Repeat, as time allows, the start all devices in the fabric as the desired root bridge. Repeat, as time allows, the start all devices in the fabric as the desired root bridge.</li> </ol>		1)	Power off all switches. Connectivity remains that of the selected Network Ar- chitect configuration.	25 26		
<ol> <li>Power up all switches. Verify IP connectivity of connected nodes.</li> <li>Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, the desired root bridge. Repeat, as time allows, the desired root bridge.</li> </ol>		2)	Disconnect one switch, leaving its attached RNICs isolated from the rest of the fabric.	27 28		
<ul> <li>4) Reconnect disconnected switch to original location. Verify IP and RDMA connectivity is restored to all nodes.</li> <li>5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 40</li> </ul>		3)	Power up all switches. Verify IP connectivity of connected nodes.	29		
<ul> <li>(a) Rebolined a under the original reduction verify in and rebuilt an</li></ul>		4)	Reconnect disconnected switch to original location. Verify IP and RDMA	30		
<ul> <li>5) Remove a redundant switch-to-switch interconnect. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 40</li> </ul>			connectivity is restored to all nodes.	31		
<ul> <li>connectivity is maintained to all nodes.</li> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 41</li> </ul>		5)	Remove a redundant switch-to-switch interconnect. Verify IP and RDMA	32		
<ul> <li>6) Create a new redundant switch-to-switch interconnection (potentially forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 41</li> </ul>		,	connectivity is maintained to all nodes.	33		
forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con- nectivity is maintained to all nodes. 36 7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 38 8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes. 40 9) Restart all devices in the fabric and follow Steps 1-8 each time with a dif- ferent switch in the fabric as the desired root bridge. Repeat, as time allows, 42		6)	Create a new redundant switch-to-switch interconnection (potentially	34		
<ul> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 41</li> </ul>			forming a loop in the absence of RSTP/MSTP). Verify IP and RDMA con-	35		
<ul> <li>7) Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 42</li> </ul>		- `		36		
<ul> <li>8) Restore the previously removed channel to the link aggregate. Verify IP and RDMA connectivity is maintained to all nodes.</li> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 41</li> </ul>		()	Remove a channel from a link aggregate. Verify IP and RDMA connectivity is maintained to all nodes	37		
<ul> <li>RDMA connectivity is maintained to all nodes.</li> <li>Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows,</li> </ul>		8)	Restore the previously removed channel to the link aggregate. Varify ID and	38		
<ul> <li>9) Restart all devices in the fabric and follow Steps 1-8 each time with a different switch in the fabric as the desired root bridge. Repeat, as time allows, 42</li> </ul>		0)	RDMA connectivity is maintained to all nodes.	39		
ferent switch in the fabric as the desired root bridge. Repeat, as time allows,		9)	Restart all devices in the fabric and follow Steps 1-8 each time with a dif-	40 74		
	fere		ferent switch in the fabric as the desired root bridge. Repeat, as time allows,	+ı ⊿2		

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	Ethernet Fabric Reconvergence using OFED RELEASE 1.28	March 27, 2009 DRAFT	
	until each switch has been the root of a spanning tra isolated from the fabric at least once, each switch h pology change (removal or addition of a link), and a seen a removal and restoration of a link.	ee, each switch has been as seen at least one to- all link aggregates have	1 2 3
	<b>Note</b> : In the presence of no hardware/ firmware/ so ously tested combinations resident in the OFILG cl	oftware changes, previ- uster may be omitted.	4 5 6 7
11.3.5 OBSERVABLE RESULTS			8
	<ol> <li>In all cases, the desired root bridge (or in the case desired root bridges) should always become the ro</li> </ol>	of MSTP topologies, the ot bridge.	9 10
	2) IP and RDMA connectivity should be restored to al note: this could be further clarified (some topology impact traffic, some will), RSTP likely would conver and thus 2s could form an 'extreme' upper-bound for the cases of traffic interruption.	l stations rapidly. Editors changes should not rge in well under 2sec r reconvergence times in	11 12 13 14
11.3.6 POSSIBLE PROBLEMS	Time limitations of the plugfest may prevent full evaluat RNICs. In this case, 'switch to switch' links and 'switch lected in random order to provide as much coverage as	tion of all switches and to RNIC' links will be se- s time allows.	15 16 17 18 19

11.4 ETHERNET FABRIC FAILON	VER		1		
11.4.1 PURPOSE			2		
	The	Ethernet Fabric Failover test is a validation that the Ethernet switch fabric de-	3		
	vice	es receiving the OFA Logo properly recovers in the event of the loss of the root tch and a new topology converges in a timely manner, minimally impacting the	4		
	fab	ric.	6		
11 4 2 Resource Requirements			7		
	1)	Gigabit or 10Gigabit Ethernet RNICs	8		
	2)	Gigabit or 10Gigabit Ethernet Switches	9		
	3)	Compliant Cables	10		
11.4.3 Discussion	- /		12		
	The	e validation of the underlying transport infrastructure is essential to the end-	13		
	use	ers experience of the operation of the OFED software stack. To this end, this	14		
	tes	ge in the event of a failure of the root bridge of an RSTP or MSTP topology	15		
	per the selected plugfest or cluster Network Architecture configuration. A root				
	connects in its given VLAN domain. Loss of the current root bridge requires re- discovery and re-election of a new root bridge, possibly further delaying network				
	re-o	convergence.	20		
	It is	assumed that the Network Architecture will be selected such that the core	21		
	of any one switch will not interrupt the flow of cluster traffic. Additionally, it sumed that the core switches will have no direct connection to any RNIC, ar				
	these core switches will be selected to serve as the root of the spanning trees is presumed that every switch under test can serve as one of the cluster's co				
	swi	switches.			
	Nei	ther exhaustive topology change permutations nor IEEE 802.1 compliance	27		
	tes	ting is performed as part of this test; however, successful completion of this	28		
	tes	t provides further evidence of the robustness of the OFA logo bearing device.	29		
	No	te: IP Connectivity is desired to ensure connectivity is stable and that under-	30		
	lyin na	g fabric issues (such as link flapping) are not masked by TCP transport of	32		
	on	the iWARP protocol.	33		
11.4.4 Dropoduro			34		
11.4.4 FIOCEddie	1)	Power off all switches. Connectivity remains that of the selected Network Ar-	35		
	')	chitect configuration.	36		
		<b>Note</b> : Selected Architecture should allow redundant paths to all RNICs in the event of the loss of the desired root switch.	38		
	2)	Power up all switches. Verify IP and RDMA connectivity of connected nodes.	39 40		

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		Ethernet Fabric Failover using OFED RELEASE 1.28	March 27, 2009 DRAFT	
	3)	In a single RSTP environment, remove power from or In an MSTP environment, remove power from or the root of a selected VLAN. Verify IP connectivi all nodes.	om the current root switch. nly one switch serving as ty is eventually restored to	1 2 3
	4)	Restart all devices in the fabric and follow Section each time with a different switch in the fabric as NOTE: This may require the location of the switch within the selected Network Architecture to ensur- all RNICs. Repeat as time allows until each switch while serving as the root of a spanning tree. Note hardware/ firmware/ software changes, previous ident in the OFILG cluster may be omitted.	ons 10.A.1 through 10.A.3, the desired root bridge. ch-under-test to be moved are redundant paths exist to ch has been powered off e: In the presence of no ly tested combinations res-	4 5 6 7 8 9 1( 1 <sup>-</sup>
11.4.5 OBSERVABLE RESULTS				12
	1)	In all cases, the desired root bridge (or in the cas desired root bridges) should always become the	se of MSTP topologies, the root bridge.	14 1{
	2)	IP and RDMA connectivity should be restored to note: the term 'rapidly' could be further clarified, I switch will significantly increase convergence tin	all stations rapidly. Editors however the loss of the root nes.	16
11.4.6 POSSIBLE PROBLEMS				18
	Av wit RN ne	ailable switch ports may restrict a given switch's at hin the Network Architecture. In such cases, if a r IICs is not possible in the event the root switch is ctivity to the effected RNICs will naturally not be re	bility to serve in any location edundant path to a set of lost, then reconverged con- equired.	20 21 22 23
	Ad sw mu	ditionally, time limitations of the plugfest may prev itches. In this case, each switch will be selected in ich coverage as time allows.	rent full evaluation of all random order to provide as	24 25 26
				27 28 28
				30
				32
				35
				37
				38 39 4(

- 41
- 42

11.5 IWARP CONNECTIVITY USING OFED 11.5.1 UNH-IOL INTEROP SUITE					
	See UNH-IOL iWARP Interoperability Test Suite for full details	3			
		4			
11.5.2 IWARP SETUP		5			
	The interoperability tests can be run in point to point mode or switched. Connect 2 iWARP hosts RNICs together or to a 10GbE switch.	6 7			
11.5.3 Test Procedure		8			
Step A:	Group 1: Single RDMA Operations Over A Single Connection:	9 10			
	TEST 1.1: RDMA WRITE	11			
	TEST 1.2: RDMA READ	12			
	TEST 1.3: RDMA SEND	13			
	TEST 1.4: RDMA SENDINV	15			
	TEST 1.5: RDMA SENDSE	16			
	TEST 1.6: RDMA SENDSEINV	17			
	TEST 1.7: RDMA TERMINATE	18			
	TEST 1.8: LARGE RDMA WRITE	19			
	TEST 1.9: LARGE RDMA READ	21			
		22			
Step B	Group 2: Multiple RDMA Operations Over A Single Connection:	23			
	Test 2.1: Sequence of 10 RDMA Write Commands	24			
	Test 2.2: Sequence of 10 RDMA Read Commands	26			
	Test 2.3: Sequence of 10 RDMA Send Commands	27			
	<ul> <li>Test 2.4: Sequence of 10 RDMA Sendinv Commands</li> </ul>	28			
	<ul> <li>Test 2.5: Sequence of 10 RDMA Sendse Commands</li> </ul>	29			
	Test 2.6: Sequence of 10 RDMA Sendseinv Commands	31			
	Test 2.7: Sequence of 10 RDMA Terminate Commands	32			
	<ul> <li>Test 2.8: Sequence of Interleaved RDMA Write And Read Com- mands</li> </ul>	33 34			
	Test 2.9: Sequence of Interleaved RDMA Write And Terminate Com- mands	35 36			
	<ul> <li>Test 2.10: Sequence of Interleaved RDMA Read And Terminate Commands</li> </ul>	37 38			
	Test 2.11: Sequence of Interleaved RDMA Send And Terminate Com- mands	39 40 41			

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		iWARP Connectivity using OFED RELEASE 1.28	March 27, 2009 DRAFT	-
	•	Test 2.12: Sequence of Interleaved RDMA Ser Commands	ndinv And Terminate	1 2
	•	Test 2.13: Sequence of Interleaved RDMA Ser Commands	ndse And Terminate	3 4
	•	Test 2.14: Sequence of Interleaved RDMA Sen Commands	dseinv And Terminate	5 6
	•	Test 2.15: Sequence of Interleaved RDMA Wri RDMA Commands	te With All Other	7 8
	•	Test 2.16: Sequence of Interleaved RDMA Rea RDMA Commands	ad With All Other	9 1
	•	Test 2.17: Sequence of Interleaved RDMA Ser RDMA Commands	nd With All Other	1 1
	•	Test 2.18: Sequence of Interleaved RDMA Ser RDMA Commands	ndinv With All Other	1
	•	Test 2.19: Sequence of Interleaved RDMA Ser RDMA Commands	ndse With All Other	1
	•	Test 2.20: Sequence of Interleaved RDMA Ser RDMA Commands	ndseinv With All Other	1
Step C	Gro	<b>oup 3</b> : Multiple Connections:		2
	•	Test 3.1: Single RDMA Operations Over Multi	ple Connections	2
	•	Test 3.2: Multiple RDMA Operations Over Mul	tiple Connections	2
	•	Test 3.3: RDMA Operations Over 25 Connecti	ons	2
	•	Test 3.4: Simultaneous Operations Over 25 C	onnections	2
		·		2
Step D	Gro	up 4: Disconnect/Reconnect Physical Connect	ions:	2
				2
	•	Test 4.1: Termination Followed By A WRITE		3
	•	Test 4.2: Termination Followed By A READ		3
Stop F	<b>C</b>	un F. Chood Narotistian.		3
Step E	Gro	up 5: Speed Negotiation.		3
	•	Test 5.1: RNICs Operating At 10g And 1g Spe	ed	3
				3
Step F	Gro	<b>up 6</b> : RDMA Error Ratio:		3
		Toot 6 1: Soquence of All Zaras		3
	•			3
	•	lest 6.2: Sequence of All Ones		4
	•	Test 6.3: Sequence of Ones Followed By Zero	DS	4
				- 4

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		iWARP Connectivity using OFED RELEASE 1.28	March 27, 2009 DRAFT
	•	Test 6.4: Sequence of Interleaved Ones Ar	nd Zeros
Step G	Gr	oup 7: Stress Patterns Over RDMA:	
	•	Test 7.1: RDMA Read After Prolonged RDI	MA Write Operations
	•	Test 7.2: RDMA Read After Prolonged RDI	MA Read Operations
	•	Test 7.3: RDMA Read After Prolonged RDI	MA Send Operations
	•	Test 7.4: RDMA Read After Prolonged RDI	MA Sendinv Operations
	•	Test 7.5: RDMA Read After Prolonged RDI	MA Sendse Operations
	•	Test 7.6: RDMA Read After Prolonged RDI	MA Sendseinv Operations
Step H	Gr	<b>coup 8</b> : Parameters:	
	•	Test 8 1 <sup>.</sup> Markers Support	
	•	Test 8 2: CRC Support	

TI iSER using OFED RELEASE 1.28

12 TRANS	SPORT INDEPENDENT IN	TE	ROP PROCEDURES USING OFED	1			
12.1 TI IS				2			
12.1.1 IB	Setup			3			
· _ · · · · _ • • • • • •			nnect initiator/target to switch as well as run one or more SMs (embedded in	4			
	ſ	the	switch or host based). If more than one SM, let the SMs split into master and	5			
			/e.	6			
	,	<b>Ontional:</b> In the procedures below, an IR analyzer can be inserted in the appro-					
		pria	the link to obtain traces and validate the aspects of the procedures specifically	8			
detailed below in subsequent sections.				9			
				11			
		Co	anect iSER bost initiator and target RNICs to an 10GbE switch	12			
		00		13			
12.1.3 Pro	ocedure			14			
		1)	Load iSER target and iSER initiator to hosts from OpenFabrics tree, check	15			
			ISER connection.	16			
		2)	Run basic dd application from iSER initiator host connected to target.	17			
		3)	[IB Specific Test] Run basic dd application from iSER initiator host con-	18			
		completes properly.	19				
	,	4)	Unload iSER initiator from a Host and check iSER connection properly dis-	20			
		,	connected on a target host.	21			
	!	5)	Unload iSER target from a Host and check iSER connection properly dis-	22			
			connected on an initiator host.	23			
		6)	[IB Specific Test] Repeat steps 2-5 now with the previous slave SM (we did	25			
			not actually stop the target).	26			
				27			
				28			
				29			
				30			
				31			
				32			
				33			
				34			
				30			
				30			
				01			

39 40

38

TI NFS over RDMA using OFED RELEASE 1.28

12.2 TI NFS OVER RDMA USING (	<b>DFED</b> 1	
Νο	<b>te</b> : This procedure was written by NetApp and Open Grid Computing. For ad- $^2$	
diti	onal help, please use the following links: 3	
1)	http://lxr.linux.no/linux+v2.6.28/Documentation/filesvstems/nfs-rdma.txt	
2)	5 http://www.connectathon.org/nfstests.html	
-/	nfs rdma devel@lists sourceforge net	
12.2.1 Installation		
12.2.1 IIIStallation	Verify that of sutils 1.1.2 or greater is installed on all the clients	
')	2) /shin/mount of a //	Ο
	a) /sbit/mount.ms -v	1
	b) If the version is less than 1.1.2 or the command does not exist, you should install the latest version of nfs-utils.	2
	i) <u>http://www.kernel.org/pub/linux/utils/nfs</u> 13	3
2)	After building the nfs-utils package, there will be a mount.nfs binary in the utils/mount directory. This binary can be used to initiate NFS v2, v3, or v4 mounts. To initiate a v4 mount, the binary must be called mount.nfs4. The standard technique is to create a symlink called mount.nfs4 to mount.nfs. This mount.nfs binary should be installed at /sbin/mount.nfs as follows:	4 5 6 7
	a) sudo cp utils/mount/mount.nfs /sbin/mount.nfs	q
	<b>Note</b> : mount.nfs and therefore nfs-utils-1.1.2 or greater is only needed on the NFS client machine. You do not need this specific version of nfs- utils on the server. Furthermore, only the mount.nfs command from nfs- utils-1.1.2 is needed on the client.	0 1 2
3)	Verify that you are using a Linux kernel with NFS/RDMA	3
	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 Linux kernel can be found at:	4 5 6
	i) <u>ftp://ftp.kernel.org/pub/linux/kernel/v2.6/</u> 27	7
	b) Download the sources as needed and place them in an appropriate lo-	8 9
4)	Configure the RDMA stack 30	0
(ד	a) Make sure your kernel configuration has PDMA support enabled Under 3	1
	Device Drivers -> InfiniBand support, update the kernel configuration to 32	2
	enable InfiniBand support [ <b>Note</b> : the option name is misleading. En- abling InfiniBand support is required for all RDMA devices (IB, iWARP, etc.)].	3 4
	b) Enable the appropriate IB HCA support (mlx4, mthca, ehca, ipath, etc.)	5 6
	a) If you are using latiniPand, be sure to enable ID over latiniPand over 3	7
	port.	8
5)	Configure the NFS client and server	9
3)	4(	0
	41	1
	42	2

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN			TI NFS over RDMA using OFED March 27, 2009 RELEASE 1.28 DRAFT	
	ä	a)	Your kernel configuration must also have NFS file system support and/or NFS server support enabled. These and other NFS related con- figuration options can be found under File Systems -> Network File Sys- tems.	1 2 3
	6) I	Buil	d, install, reboot	4
	ä	a)	The NFS/RDMA code will be enabled automatically if NFS and RDMA are turned on. The NFS/RDMA client and server are configured via the hidden SUNRPC_XPRT_RDMA config option that depends on SUNR-PC and INFINIBAND. The value of SUNRPC_XPRT_RDMA will be:	5 6 7 8
			<ul> <li>N if either SUNRPC or INFINIBAND are N, in this case the NFS/RDMA client and server will not be built</li> </ul>	9 1(
			<ul> <li>ii) - M if both SUNRPC and INFINIBAND are on (M or Y) and at least one is M, in this case the NFS/RDMA client and server will be built as modules</li> </ul>	11 12 13
			<ul> <li>iii) - Y if both SUNRPC and INFINIBAND are Y, in this case the NFS/RDMA client and server will be built into the kernel</li> </ul>	14
	I	<b>b</b> )	Therefore, if you have followed the steps above and turned on NFS and RDMA, the NFS/RDMA client and server will be built.	16
	(	C)	Build a new kernel, install it, boot it.	1/
	7) (	Che	eck RDMA Setup	19
	ä	a)	If you are using InfiniBand, make sure there is a Subnet Manager (SM) running on the network.	2(
	I	o)	To further test the InfiniBand software stack, use IPoIB to ping two hosts	22
8	8) (	Che	eck NFS Setup	2
	i	a)	For the NFS components enabled above (client and/or server), test their functionality over standard Ethernet using TCP/IP or UDP/IP.	24 2!
	9) I	NFS	S/RDMA Setup	26
	i	a)	Use two machines, one to act as the client and one to act as the server.	27
	I	<b>b</b> )	On the server system, configure the /etc/exports file and start the NFS/RDMA server. Exports entries with the following formats have been tested:	28
			i) /vol0 192 168 0 47/fsid=0 pv async insecure no root squash)	30
			ii) /vol0 192.168.0.0/255.255.255.0(fsid=0.rw.async.insecure, no_root_squash)	32
			cure,no_root_squash)	33
	(	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.	34
			<b>Note</b> : The "insecure" option must be used because the NFS/RDMA client does not use a reserved port.	36
	(	d)	Start the NFS server	39
			<ul> <li>If the NFS/RDMA server was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the</li> </ul>	39
			RDMA transport module:	41
			1. \$ modprobe svcrdma	42

<ul> <li>ii) Regardless of how the server was built (module or built-in), start the server: <ol> <li>Server:</li> <li>Server:</li> <li>Server to listen on the RDMA transport:</li> <li>Secho rdma 20049 &gt; /proofs/nfds/portlist</li> </ol> </li> <li>(Do the client system</li> <li>If the NFS/RDMA client was built as a module (CONFIG_SUNRPC_XFRT_RDMA=m in kernel config), load the RDMA client module:</li> <li>Smodprobe xprtrdma.ko</li> <li>Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:</li> <li>\$modprobe xprtrdma.ko</li> <li>Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:</li> <li>\$modprobe xprtrdma.ko</li> <li>Please see <u>Connectation</u> for instructions on how to run the available tests:</li> <li>Run the following tests</li> <li>Test 1 - File and directory creation</li> <li>Test 2 - File and directory creation</li> <li>Test 4 - Stettr and lookup</li> <li>Test 5 - Read and write</li> <li>Test 5 - Read in</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Stafts</li> </ul>	OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		TI NFS over RDMA using OFED RELEASE 1.28	March 27, 2009 DRAFT
1. \$ /etc/init.d/nfs start or \$ service nfs start         iii) Instruct the server to listen on the RDMA transport:         1. \$ echo rdma 20049 > /proc/fs/nfsd/portlist         e) On the client system         i) If the NFS/RDMA client was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module:         1. \$ modprobe xprtrdma.ko         ii) Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:         1. \$ modprobe xprtrdma.ko         iii) To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.         12.2.2 NFS/RDMA Test Procedure         1) Please see Connectation for instructions on how to run the available tests:         2) Run the following tests         a) Test 1 - File and directory creation         b) Test 2 - File and directory creation         c) Test 3 - Lookups across mount point         c) Test 4 - Setattr, gatattr, and lookup         e) Test 5 - Read and write         g) Test 5 - Read and write         g) Test 7 - Link and rename         k) Test 7 - Link and rename         k) Test 8 - Symlink and readlink         n) Test 9 - Statfs			ii) Regardless of how the server was built (mo	odule or built-in), start the 1
<ul> <li>iii) Instruct the server to listen on the RDMA transport: <ol> <li>\$ echo rdma 20049 &gt; /proc/fs/nfsd/portlist</li> <li>On the client system</li> <li>ii) If the NFS/RDMA client was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module: <ol> <li>S modyrobe xprtdma.ko</li> <li>Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:</li> <li>\$ mount - ordma port-20049 &lt; PolB-server-name-or-ad- dress&gt;:/<export>/mnt</export></li> </ol> </li> <li>10) To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.</li> </ol></li></ul> <li>12.2.2 NFS/RDMA Test Procedure <ol> <li>Please see Connectation for instructions on how to run the available tests:</li> <li>Run the following tests</li> <li>Test 1 - File and directory creation</li> <li>Test 2 - File and directory removal</li> <li>Test 3 - Lookups across mount point</li> <li>Test 4 - Setatr, getatr, and lookup</li> <li>Test 5 - Read and write</li> <li>Test 5 - Read and write</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Link and rename</li> <li>Test 8 - Symlink and readlink</li> <li>Test 9 - Statfs</li> </ol></li>			1. \$ /etc/init.d/nfs start <b>or</b> \$ service nfs st	art 2
<ul> <li>1. \$ echo rdma 20049 &gt; /proc/fs/nfsd/portlist</li> <li>e) On the client system <ol> <li>i) If the NFS/RDMA client was built as a module</li> <li>(CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module: <ol> <li>1. \$ modprobe xprtrdma.ko</li> <li>ii) Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:</li> <li>1. \$ mourt or dma.port=20049 <ipoib-server-name-or-ad-dress- rexport="">/mnt</ipoib-server-name-or-ad-dress-></li> </ol> </li> <li>12.2.2 NFS/RDMA Test Procedure <ol> <li>Please see <u>Connectation</u> for instructions on how to run the available tests:</li> <li>2) Run the following tests <ol> <li>a) Test 1 - File and directory creation</li> <li>b) Test 2 - File and directory removal</li> <li>c) Test 3 - Lookups across mount point</li> <li>d) Test 4 - Setattr, and lookup</li> <li>f) Test 5 - Read and write</li> <li>g) Test 5 - Read and write</li> <li>g) Test 5 - Read</li> <li>j) Test 7 - Link and rename</li> <li>k) Test 7 a - Rename</li> <li>i) Test 8 - Symlink and readlink</li> <li>m) Test 9 - Statfs</li> </ol> </li> </ol></li></ol></li></ul>			iii) Instruct the server to listen on the RDMA tr	ransport: 4
<ul> <li>e) On the client system <ol> <li>i) If the NFS/RDMA client was built as a module</li> <li>(CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module: <ol> <li>1. \$ modprobe xprtrdma.ko</li> </ol> </li> <li>ii) Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server: <ol> <li>1. \$ mount - ordma.port=20049 <ipoib-server-name-or-ad-dress>-/<export>/mnt</export></ipoib-server-name-or-ad-dress></li> </ol> </li> <li>12.2.2 NFS/RDMA Test Procedure <ol> <li>Please see <u>Connectathon</u> for instructions on how to run the available tests:</li> <li>2) Run the following tests <ol> <li>a) Test 1 - File and directory removal</li> <li>b) Test 2 - File and directory removal</li> <li>c) Test 3 - Lookups across mount point</li> <li>c) Test 4 - Getattr and lookup</li> <li>f) Test 5 - Read and write</li> <li>g) Test 5 - Read and write</li> <li>g) Test 5 - Read and write</li> <li>g) Test 5 - Read</li> <li>j) Test 7 - Link and rename</li> <li>k) Test 7 - Rename</li> <li>j) Test 7 - Link and rename</li> <li>k) Test 7 - Link and rename</li> <li>k) Test 7 - Link and rename</li> <li>k) Test 9 - Statfs</li> </ol> </li> </ol></li></ol></li></ul>			1. \$ echo rdma 20049 > /proc/fs/nfsd/por	tlist 5
<ul> <li>i) If the NFS/RDMA client was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module: <ol> <li>\$ modprobe xptrdma.ko</li> <li>Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:</li> <li>\$ mount - or fman, port=20049 </li> </ol> </li> <li>12.2.2 NFS/RDMA Test Procedure <ol> <li>Please see <u>Connectation</u> for instructions on how to run the available tests:</li> <li>Rest 1 - File and directory creation</li> <li>Test 1 - File and directory removal</li> <li>Test 4 - Setattr, getattr, and lookup</li> <li>Test 5 - Read and write</li> <li>Test 5 - Read and write</li> <li>Test 5 - Read and write</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Link and rename</li> <li>Test 9 - Statfs</li> </ol></li></ul>		e)	On the client system	6
1. \$ modprobe xptrdma.ko       iii         ii)       Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:         1. \$ mount - ordma.port=20049 <ipoib-server-name-or-address>:/<export>/mmt         iiii)       To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.         12.2.2 NFS/RDMA Test Procedure       1         1)       Please see Connectathon for instructions on how to run the available tests:         2)       Run the following tests         a)       Test 1 - File and directory creation         b)       Test 2 - File and directory removal         c)       Test 3 - Lookups across mount point         d)       Test 4 - Getattr and lookup         e)       Test 5 - Read and write         g)       Test 5 - Read and write         g)       Test 5 - Read         i)       Test 7 - Link and rename         ii)       Test 7 - Link and rename         iii)       Test 8 - Symlink and readlink         n)       Test 9 - Statfs</export></ipoib-server-name-or-address>		,	<ul> <li>i) If the NFS/RDMA client was built as a mod (CONFIG_SUNRPC_XPRT_RDMA=m in k RDMA client module:</li> </ul>	lule 7 kernel config), load the 8
<ul> <li>ii) Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server: <ol> <li>\$ mount - or dma,port=20049 <ipoib-server-name-or-address>-/keyport&gt;/mnt</ipoib-server-name-or-address></li> <li>iii) To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.</li> </ol> </li> <li>12.2.2 NFS/RDMA Test Procedure <ol> <li>Please see <u>Connectathon</u> for instructions on how to run the available tests:</li> <li>Run the following tests <ol> <li>Test 1 - File and directory creation</li> <li>Test 2 - File and directory removal</li> <li>Test 3 - Lookups across mount point</li> <li>Test 4 - Getattr and lookup</li> <li>Test 5 - Read and write</li> <li>Test 5 - Read</li> <li>Test 6 - Readdir</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Link and readlink</li> <li>Test 9 - Statfs</li> </ol> </li> </ol></li></ul>			1. \$ modprobe xprtrdma.ko	1
1. \$ mount -o rdma,port=20049 <ipoib-server-name-or-address>:/<export> /mnt       14         iii) To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.       15         12.2.2 NFS/RDMA Test Procedure       1) Please see Connectathon for instructions on how to run the available tests:       17         2) Run the following tests       a) Test 1 - File and directory creation       10         b) Test 2 - File and directory removal       11         c) Test 3 - Lookups across mount point       12         d) Test 4 - Setattr, getattr, and lookup       12         e) Test 5 - Read and write       12         g) Test 5 - Read       12         i) Test 6 - Readdir       13         j) Test 7 - Link and rename       13         k) Test 7 a - Rename       14         i) Test 8 - Symlink and readlink       13         m) Test 9 - Statfs       14</export></ipoib-server-name-or-address>			ii) Regardless of how the client was built (mo command to mount the NFS/RDMA server	dule or built-in), use this 11
<ul> <li>iii) To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.</li> <li>1) Please see <u>Connectathon</u> for instructions on how to run the available tests: <ol> <li>Run the following tests</li> <li>Test 1 - File and directory creation</li> <li>Test 2 - File and directory removal</li> <li>Test 3 - Lookups across mount point</li> <li>Test 4 - Getattr getattr, and lookup</li> <li>Test 5 - Read and write</li> <li>Test 6 - Readdir</li> <li>Test 7 - Link and rename</li> <li>Test 7 - Link</li> <li>Test 8 - Symlink and readlink</li> <li>Test 9 - Statfs</li> </ol> </li> </ul>			<ol> <li>\$ mount -o rdma,port=20049 <ipoib-se dress&gt;:/<export> /mnt</export></ipoib-se </li> </ol>	erver-name-or-ad-
12.2.2 NFS/RDMA Test Procedure       1         1) Please see Connectation for instructions on how to run the available tests:       17         2) Run the following tests       19         a) Test 1 - File and directory creation       20         b) Test 2 - File and directory removal       21         c) Test 3 - Lookups across mount point       22         d) Test 4 - Setattr, getattr, and lookup       23         e) Test 5 - Read and write       26         g) Test 5 - Read       28         i) Test 5 - Read       28         ii) Test 6 - Readdir       29         ji) Test 7 - Link and rename       30         k) Test 7a - Rename       31         li) Test 9 - Statfs       36         an Test 9 - Statfs       36			<li>iii) To verify that the mount is using RDMA, ru check the "proto" field for the given mount.</li>	n "cat /proc/mounts" and
1) Please see Connectation for instructions on how to run the available tests:2) Run the following testsa) Test 1 - File and directory creationb) Test 2 - File and directory removalc) Test 3 - Lookups across mount pointd) Test 4 - Setattr, getattr, and lookupe) Test 4a - Getattr and lookupf) Test 5 - Read and writeg) Test 5a - Writeh) Test 5b - Readi) Test 6 - Readdirj) Test 7 - Link and renamek) Test 7a - Renamej) Test 8 - Symlink and readlinkn) Test 9 - Statfs	12.2.2 NFS/RDMA Test Procedure			11
2)Run the following tests19a)Test 1 - File and directory creation20b)Test 2 - File and directory removal21c)Test 3 - Lookups across mount point22d)Test 4 - Setattr, getattr, and lookup23e)Test 4 - Setattr, getattr, and lookup24f)Test 5 - Read and write25g)Test 5 - Read and write26g)Test 5a - Write26g)Test 5b - Read28i)Test 6 - Readdir29j)Test 7a - Rename30k)Test 7b - Link32m)Test 9 - Statfs363738394041	1)	Ple	ase see <u>Connectathon</u> for instructions on how t	o run the available tests:
a) Test 1 - File and directory creation20b) Test 2 - File and directory removal21c) Test 3 - Lookups across mount point22d) Test 4 - Setattr, getattr, and lookup23e) Test 4 - Getattr and lookup24f) Test 5 - Read and write26g) Test 5a - Write27h) Test 5b - Read28i) Test 6 - Readdlir29j) Test 7 - Link and rename30k) Test 7a - Rename31l) Test 9 - Statfs36a) Test 9 - Statfs36a) Test 9 - Statfs36	2)	Ru	n the following tests	1
b) Test 2 - File and directory removal 21 c) Test 3 - Lookups across mount point 22 d) Test 4 - Setattr, getattr, and lookup 23 e) Test 4a - Getattr and lookup 24 f) Test 5 - Read and write 26 g) Test 5a - Write 27 h) Test 5b - Read 28 i) Test 6 - Readdir 29 j) Test 7 - Link and rename 30 k) Test 7a - Rename 31 l) Test 7b - Link 32 m) Test 8 - Symlink and readlink 33 m) Test 9 - Statfs 35		a)	Test 1 - File and directory creation	20
c) Test 3 - Lookups across mount point       22         d) Test 4 - Setattr, getattr, and lookup       23         e) Test 4a - Getattr and lookup       24         f) Test 5 - Read and write       26         g) Test 5a - Write       27         h) Test 5b - Read       28         i) Test 6 - Readdir       29         j) Test 7 - Link and rename       30         k) Test 7a - Rename       31         l) Test 8 - Symlink and readlink       34         n) Test 9 - Statfs       36         37       38         38       39         40       41		b)	Test 2 - File and directory removal	2
d) Test 4 - Setattr, getattr, and lookup       23         e) Test 4a - Getattr and lookup       24         f) Test 5 - Read and write       26         g) Test 5a - Write       27         h) Test 5b - Read       28         i) Test 6 - Readdir       29         j) Test 7 - Link and rename       30         k) Test 7a - Rename       31         l) Test 7b - Link       32         m) Test 8 - Symlink and readlink       33         n) Test 9 - Statfs       36         37       38         38       39         40       41		c)	Test 3 - Lookups across mount point	2:
e) Test 4a - Getattr and lookup       24         f) Test 5 - Read and write       26         g) Test 5a - Write       27         h) Test 5b - Read       28         i) Test 6 - Readdir       29         j) Test 7 - Link and rename       30         k) Test 7a - Rename       31         l) Test 7b - Link       32         m) Test 8 - Symlink and readlink       33         n) Test 9 - Statfs       36         37       38         39       40         41       41		d)	Test 4 - Setattr, getattr, and lookup	2:
f)       Test 5 - Read and write       25         g)       Test 5a - Write       27         h)       Test 5b - Read       28         i)       Test 5b - Read       28         i)       Test 6 - Readdir       29         j)       Test 7 - Link and rename       30         k)       Test 7a - Rename       31         l)       Test 7b - Link       32         m)       Test 8 - Symlink and readlink       33         n)       Test 9 - Statfs       36         37       38       39         40       40       41		e)	Test 4a - Getattr and lookup	24
g)       Test 5a - Write       20         n)       Test 5b - Read       28         i)       Test 6 - Readdir       29         j)       Test 7 - Link and rename       30         k)       Test 7a - Rename       31         l)       Test 7b - Link       32         m)       Test 8 - Symlink and readlink       33         n)       Test 9 - Statfs       36         37       38       39         40       41       41		f)	Test 5 - Read and write	28
h) Test 5b - Read       28         i) Test 6 - Readdir       29         j) Test 7 - Link and rename       30         k) Test 7a - Rename       31         l) Test 7b - Link       32         m) Test 8 - Symlink and readlink       34         n) Test 9 - Statfs       36         37       38         39       40         41       41		g)	Test 5a - Write	2
i) Test 6 - Readdir 29 j) Test 7 - Link and rename 30 k) Test 7a - Rename 31 l) Test 7b - Link 32 m) Test 8 - Symlink and readlink 33 n) Test 9 - Statfs 35 36 37 38 39 40 41		h)	Test 5b - Read	2
j)       Test 7 - Link and rename       30         k)       Test 7a - Rename       31         l)       Test 7b - Link       32         m)       Test 8 - Symlink and readlink       33         n)       Test 9 - Statfs       35         36       37       38         39       40       41		i)	Test 6 - Readdir	29
k)       Test 7a - Rename       31         l)       Test 7b - Link       32         m)       Test 8 - Symlink and readlink       33         n)       Test 9 - Statfs       35         36       37       38         39       40       41		j)	Test 7 - Link and rename	3
I)       Test 7b - Link       32         m)       Test 8 - Symlink and readlink       33         n)       Test 9 - Statfs       35         36       37       38         39       40       41		k)	Test 7a - Rename	3
m) Test 8 - Symlink and readlink n) Test 9 - Statfs 36 37 38 39 40 41		I)	Test 7b - Link	3:
34         n) Test 9 - Statfs         36         37         38         39         40         41		m)	Test 8 - Symlink and readlink	33
35 36 37 38 39 40 41		n)	Test 9 - Statfs	34
36 37 38 39 40 41		,		3
37 38 39 40 41				30
30 39 40 41				చ 2
40 41				्र २
41				4
				4

12.3 TI RELIABLE DATAGRAM SE	ER\	ICE (RDS) USING OFED	1
12.3.1 RDS-Ping Procedure	12.3.1 RDS-Ping Procedure		
- 1	1)	Use the command <i>modprobe rds</i> to add RDS support to the kernel	3
2	2)	Verify that the kernel supports RDS by issuing the <i>rds-info</i> command.	4
		a) The rds-info utility presents various sources of information that the RDS kernel module maintains. When run without any optional arguments rds-info will output all the information it knows of.	5 6 7
3	3)	Start one of the Subnet Managers in the cluster	8
		<b>Note</b> : RDS is IP based so you need to provide a host address either through an out of band Ethernet connection or through IPoIB. RDS also requires the LIDs to be set in an InfiniBand Fabric and therefore an SM must be run.	9 10 11
		Note: All SMs in the fabric should be tested.	12
4	4)	Chose a host and use <i>rds-ping host</i> to communicate with every other end point in the fabric.	13 14
		<ul> <li>rds-ping is used to test whether a remote node is reachable over RDS.</li> <li>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.</li> </ul>	15 16 17
		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 bound; instead, the kernel processes incoming packets and responds to them.	18 19 20
5	5)	Verify that all nodes respond without error.	21
12.3.2 RDS-STRESS PROCEDURE			23
1	1)	Chose a host and start a passive receiving session for the RDS Stress test. It only needs to be told what port to listen on.	24 25
		a) \$ rds-stress -p 4000	26
2	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 s given configuration options which both instances will use.	27 28
		a) \$ rds-stress -T 5 -s recvhost -p 4000 -t 1 -d 1	29
3	3)	Every second, the parent process will display statistics of the ongoing stress test. If the -T option is given, the test will terminate after the specified time	31 32
	4 \	and a summary is printed.	33
4	+) -\	Perset store 1.4 until all and paints in the eluster have been tested	34
5	<b>)</b> )	Repeat steps 1-4 until all end points in the cluster have been tested.	35
			36
			37
			38
			39
			4U ⊿1
			- T I

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	TI SDP using OFED RELEASE 1.28	March 27, 2009 DRAFT
12.4 TI SDP USING OFED		1
12.4.1 IB SETUP		2
( t	connect the HCAs and switches as per the Architected Net no SM is running on the Fabric.	etwork and make sure 3 4
	his procedure, as the previous ones, will be based on th in SM/SA which supports IPoIB (sufficient IB multicast su n the HCAs, or on a switch with an embedded SM/SA or yould only run SM/SA for the partner pair (with a switch ir edure has been developed for Linux and maybe ported t ufficient vendor support.	e cluster connectivity. upport) will be running a third HCA which the middle). This pro- o Windows if there is 10
(   	<b>Optional</b> : In the procedures below, an IB analyzer can be riate link to obtain traces and validate the aspects of the p etailed below in subsequent sections.	<ul> <li>inserted in the appro-</li> <li>inserted in the a</li></ul>
12.4.2 IWARP SETUP		14
0	connect SDP host client and server RNICs to an 10GbE	switch. 16
12.4.3 INSTALLATION REQUIREMENT	s	17
۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰	<ul> <li>Iake sure the following are installed on all nodes:</li> </ul>	18
		19
	) vsttpd - for SFTP application.	20
	) ssnd - for SCP application.	22
12.4.4 CREATING A USER NAME	pecial account for this should be created as follows:	23 24
1	) Username: interop.	25
2	) Password: openfabrics.	26
12.4.5 Environment Variables		27
1	) Set LD_PRELOAD to:	20
	a) On 64bit machines - /DEFAULT_INSTALL_LOCA	TION/lib64/libsdp.so 30
	b) On 32bit machines - /DEFAULT_INSTALL_LOCA	TION /lib/libsdp.so 31
	c) <b>Example</b> : export LD_PRELOAD=/usr/local//lib64/	libsdp.so 32
2	) Set SIMPLE_LIBSDP to 1 - this says to use SDP	33
	a) Example: export SIMPLE_LIBSDP=1	34
3	) After setting the environment variables restart the xine	etd. 36
40.4.0	a) <b>Example</b> : /etc/init.d/xinetd restart	37
12.4.6 NETPERF PROCEDURE		38
	) Start an Sivi (all Sivi's will need to be tested) and let it	iniualize 39
	<ul> <li>a) verify that the running SW is the one you started.</li> <li>Start a petperf approx on one pade</li> </ul>	40
2		41

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		TI SDP using OFED RELEASE 1.28	March 27, 2009 DRAFT
		a) <b>Example</b> : /netnerf -n {nort number}	
	3)	From all the other nodes run:	
	5)	<ul> <li>a) [For IB] . /net perf -p {port number} -H {server n TCP_STREAMm {message size} -s {local b</li> </ul>	ode's IPoIB} -l 1 -t uffer size}
		<ul> <li>b) [For iWARP] . /net perf -p {port number} -H {ser TCP_STREAMm {message size} -s {local b</li> </ul>	ver node's IP} -I 1 -t uffer size}
		c) <b>Example</b> : /net perf -p 2006 -H 11.4.10.36 -I 1 -t 1000 -s 1024	TCP_STREAMm
		<ul> <li>d) Where message size is 10, 100, 1000, 10000 a 1024, 6000.</li> </ul>	nd local buffer size is
		e) Repeat these steps until all message sizes and been used from all nodes	all buffer sizes have
		f) Kill the netperf server	
	4)	Repeat step #2 and #3 with a different node acting until all nodes have done so.	as the netperf server
	5)	Repeat the netperf procedure with a different SM ru SMs have been used.	inning until all available
		<b>Note</b> : All nodes are expected to act as a server and SM. All operations must finish successfully for the coverall.	as a client using every levice to pass 10.11.6
12.4.7 SFTP PROCEDURE			
	SF the The	TP procedures require an SFTP server to be configure partner pair. An SFTP client needs to be available or default RHEL install includes both.	red on each machine in n each machine as well.
	A 4 par usii	MB file will be SFTP'd to the partner and then SFTF ed to the original file; this will be done in each direction ng every SM available.	P'd back and binary com- on and then bidirectional
12.4.7.1 SETUP			
	1)	Make auro votted is installed on each pade for SET	Dapplication
	י) 2)	A appendiate appoint for this should be prosted as follo	
	2)	A special account for this should be created as folic	JWS.
		a) Osemane. Interop	
		b) Password: openiabrics	
12.4.7.2 Procedure			
	1)	Run SFTP server on all nodes.	
		a) <b>Example</b> : /etc/init.d/vsftpd start	
	2)	Verify SDP is running.	
		a) Ismod   grep sdp	
		<ul> <li>b) ib_sdp should be greater than 0 - reference cou than 0. Each connection opens three reference</li> </ul>	unt should be greater counts.
		-	

OFA Interoperability Working Group		TI SDP using OFED	March 27, 2009	_
		NEELAGE 1.20		_
	C	<ul> <li>During these transactions double check that sdp con established, you can see it in /proc/net/sdp/conn_ma</li> </ul>	nection has been iin.	1 2
:	3) 5	Start an SM (all SM's will need to be tested) and let it init	ialize	3
	a	) Verify that the running SM is the one you started.		4
	4) 8	SFTP:		5
	â	<ul> <li>Connect an HCA pair via SFTP on IPoIB using the sp and password.</li> </ul>	pecified user name	6 7
	t	) Put the 4MB file to the /tmp dir on the remote host.		8
	C	) Get the same file to your local dir again.		9
	C	I) Compare the file using the command <i>cmp tfile tfile.or</i>	rig.	10
		i) The two must be identical		12
	e	e) Repeat the procedure with a different SM.		13
	5) F	Repeat the procedure with a different SM.		14
	1	lote: Every node must SFTP the 4MB file to all others usi	ing all SM's and the	15
	f	iles must be identical as determined by the binary comp	are in order for the	16
	C	levice to pass 10.11.7 overall.		17
12.4.8 SCP PROCEDURE	A 4 B 4	D file will be CODId to the postner and then CODId back or		18
1	4 4 IV	e original file, this will be done in each direction and then	bidirectional using	19
	ever	/ SM available.	5	20
12 4 8 1 SETUP				21
				23
	1) A	A special account for this should be created as follows:		24
	6	) Username: Interop		25
	t	<ul> <li>Password: openfabrics</li> </ul>		26
12.4.8.2 Procedure				27
	1) 8	Start an SM (all SM's will need to be tested) and let it init	ialize	28
	â	) Verify that the running SM is the one you started.		29
:	2) 8	SCP:		30
	â	) Put the 4MB file to the /tmp dir on the remote host via	a SCP.	32
	t	) Get the same file to your local dir again via SCP.		33
	C	c) Compare the file using the command <i>cmp tfile tfile.or</i>	rig.	34
		i) The two must be identical		35
	(	<ol> <li>Repeat step #2 with a different HCA pair until all HCA ed with all others (All to All).</li> </ol>	As have been test-	36 37
:	3) F	Repeat the procedure with a different SM.		38
	ľ	lote: Every node must SCP the 4MB file to all others usi	ng all SM's and the	39
	f	iles must be identical as determined by the binary comp	are in order for the	40
	C	ievice io pass 10.11.0 overall.		41
				42

OFA Interoperability Working Group	TI SDP using OFED	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLAN	RELEASE 1.28	DRAFT

Note:Sections 10.11.6, 10.11.7 and 10.11.8 must pass using the configura-<br/>tion determined by sections 10.11.1/10.11.2, 10.11.3, 10.11.4 and 10.11.5<br/>for the device to pass SDP overall.13

12.5 TI UDAPLTEST COMMANDS		1
Se	ver Command: dapItest -T S -D <ia_name></ia_name>	2
		3
12.5.1 IB SETOP	The late date apply people to be verified to be sure that the correct interface is	4
·	used. By default the dapl interface for IB is ib0 and for iWARP is eth2. If these are not correct for the current cluster then errors will occur.	5 6 7
•	It is also important to verify that the desired dapl library is being used.	8
•	To run dapltest on IB, an SM needs to be running.	9
12.5.2 GROUP 1: POINT-TO-POINT TO	POLOGY	10
[1.7	1 1 connection and simple send/recy:	11
Ľ	<ul> <li>dapltest -T T -s <server name=""> -D <ia name=""> -i 100 -t 1 -w 1 -R BE</ia></server></li> </ul>	12
	<ul> <li>client SR 256 1 server SR 256 1</li> </ul>	13
[1.2	2] Verification, polling, and scatter gather list:	14
Ľ	<ul> <li>dapltest -T T -s <sever name=""> -D <ia name=""> -i 100 -t 1 -w 1 -V -P -R BE</ia></sever></li> </ul>	16
	• client SR 1024 3 -f \	17
	• server SR 1536 2 -f	18
12.5.3 GROUP 2: SWITCHED TOPOLOG	SY	19
Infi	niBand Switch: Any InfiniBand switch	20
		21
IW/	ARP Switch: 10 GbE Switch	22
[2.7	<ol> <li>Verification and private data:</li> </ol>	23
	• dapltest -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE</ia_name></server_name>	24
	• client SR 1024 1 \	20
	• server SR 1024 1	20
[2.2	2] Add multiple endpoints, polling, and scatter gather list:	28
	• dapltest -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 10 -V -P -R</ia_name></server_name>	29
	BE client SR 1024 3 \	30
	• server SR 1536 2	31
[2.3	3] Add RDMA Write :	32
	<ul> <li>dapItest -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE</ia_name></server_name></li> </ul>	33
	• client SR 256 1 \	34
	server RW 4096 1 server SR 256 1	35
[2.4	I] Add RDMA Read:	36
	<ul> <li>dapItest -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE</ia_name></server_name></li> </ul>	37
	• client SR 256 1 \	30
	server RR 4096 1 server SR 256 1	29 20
12.5.4 GROUP 3: SWITCHED TOPOLOG	BY WITH MULTIPLE SWITCHES	41
[3.7	] Multiple threads, RDMA Read, and RDMA Write:	42

<ul> <li>dapItest -T T -s <server_name> -D <ia_name> -i 100 -t 4 -w 8 -V -P -R E</ia_name></server_name></li> </ul>	3E 1
<ul> <li>client SR 256 1 \</li> </ul>	2
<ul> <li>server RR 4096 1 server SR 256 1 client SR 256 1 server RR 4096 1</li> </ul>	\ 3
server SR 256 1	4
[3.2] Pipeline test with RDMA Write and scatter gather list:	5
<ul> <li>dapltest -T P -s <server_name> -D <ia_name> -i 1024 -p 64 -m p RW 8192 2</ia_name></server_name></li> </ul>	7
[3.3] Pipeline with RDMA Read:	8
<ul> <li>InfiniBand: dapItest -T P -s <server_name> -D <ia_name> -i 1024 -p 6</ia_name></server_name></li> <li>-m p RR 4096 2</li> </ul>	34 9 10
<ul> <li>iWARP: dapItest -T P -s <server_name> -D <ia_name> -i 1024 -p 64 - p RR 4096 1</ia_name></server_name></li> </ul>	m 11 12
[3.4] Multiple switches:	13
<ul> <li>dapItest -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 10 -V -P</ia_name></server_name></li> </ul>	-R 14
BE client SR 1024 3 \	15
server SR 1536 2	16
	17
	18
	20
	21
	22
	23
	24
	25
	26
	27
	28
	29 30
	31
	32
	33
	34
	35
	36
	37
	38
	39
	40 /1
	42

12.6 TI BASIC RDMA INTERO	P USING OFED - COMMAND LINE	1
12.6.1 Purpose		2
	To demonstrate the ability of endpoints to exchange core RDMA operations	3
	points at the RDMA level, in a simple network configuration.	- 4
	The Basic RDMA interop test identifies interoperability issues in one of four ways	3: 7
	<ul> <li>The inability to establish connections between endpoints</li> <li>The failure of RDMA operations to complete</li> <li>Incorrect data after the completion of RDMA exchanges</li> </ul>	8 9 10
12.6.2 General Setup	• Inconsistent performance levels.	11
	The RDMA interop procedure can be carried out using the OFA Verbs API to create RDMA Connections and send RDMA operation or by using a 3rd party traffic generation tool such as <u>XANStorm</u> .	12 13 14
12.6.3 Topology		16
	The topology of the network that interconnects the switches can be changed to validate operation of the endpoints over different networks paths. It is recommended that this procedure first be executed between endpoints connected by a single switch, and then the process repeated for more complex network configurations.	17 18 18 1- 19 20
12.6.4 IB Setup		21
	Connect endpoints to switch and run one or more SMs (embedded in the switch or host based).	1 23 24
	Optional: Insert analyzer in the appropriate link to obtain traces as needed.	25
12.6.5 iWARP Setup		27
	Connect iWARP RDMA endpoints to an 10GbE switch.	28
	Optional: insert analyzer to capture traces as needed.	29 30
12.6.6 RDMA Connectivity Setup		31
	Create two IP connections between each unique pair of RDMA interfaces in-	32
	volved in the testing process. The creation of two IP connections per interface pair ensures that the different semantics associated with active and passive sides of the connection are exercised	33 5 34
		35
	<ol> <li>For each unique pair of RDMA interfaces create two RDMA streams where each RDMA interface is sending RDMA data (Requestor) on a connection</li> </ol>	30
	and receiving RDMA data (Target) on the other connection.	38
12.6.7 Small RDMA READ Proce	edure	39
	1) Select the two devices that will be tested:	40
	2) On the server device issue the following command on command line:	41 42

		a) ib_read_bw -d <dev_name> -i <port></port></dev_name>	1
3	3)	On the client device issue the following command on command line:	2
		a) ib_read_bw -d <dev_name> -i <port> -s 1 -n 100000</port></dev_name>	3
4	1)	Verify that the operation completed without error and the level of perfor- mance achieved is reasonable and as expected.	4 5
12.6.8 Large RDMA READ Proced	dur	e	6
1	I)	Select the two devices that will be tested:	7
2	2)	On the server device issue the following command on command line:	ð 0
		a) ib_read_bw -d <dev_name> -i <port></port></dev_name>	10
3	3)	On the client device issue the following command on command line:	11
		a)	12
4	1)	Verify that the operation completed without error and the level of perfor- mance achieved is reasonable and as expected.	13 14
12.6.9 Small RDMA Write Procedu	ure	•	15
1	I)	Select the two devices that will be tested:	16
2	2)	On the server device issue the following command on command line:	17
		a) ib_write_bw -d <dev_name> -i <port></port></dev_name>	18
3	3)	On the client device issue the following command on command line:	20
		a) ib_write_bw -d <dev_name> -i <port> -s 1 -n 100000</port></dev_name>	20
4	4)	Verify that the operation completed without error and the level of perfor- mance achieved is reasonable and as expected.	22
12.6.10 Large RDMA Write Procee	dui	re	24
1	I)	Select the two devices that will be tested:	25
2	2)	On the server device issue the following command on command line:	26
		a) ib_write_bw -d <dev_name> -i <port></port></dev_name>	27
3	3)	On the client device issue the following command on command line:	28
		a)	29
4	1)	Verify that the operation completed without error and the level of perfor- mance achieved is reasonable and as expected.	30
12.6.11 Small RDMA SEND Proce	edι	ire	32
T d	Thi: buf	s procedure may fail due to the inability of a endpoint to repost the consumed fers.	34 35
1	n	Select the two devices that will be tested.	36
2	·, ?)	On the server device issue the following command on command line:	37
2	-,	a) ib send by -d <dev name=""> -i <nort></nort></dev>	38
3	3)	On the client device issue the following command on command line:	39
5	')	a) ib writesend by $d \leq dev name > -i \leq nort > -s = 1 - n = 100000$	40
			41
			42

OFA Interoperability Working Group	TI BASIC RDMA Interop using OFED - Command Line	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PL	AN RELEASE 1.28	DRAFT

	4)	Verify that the operation completed without error and the level of perfor- mance achieved is reasonable and as expected.	1
12.6.12 Large RDMA SEND Pro	ocedu	Ire	3
	This	procedure may fail due to the inability of a endpoint to repost the consumed	4
	buff	ers.	5
	1)	Select the two devices that will be tested:	6
	2)	On the server device issue the following command on command line:	7
	,	a) ib send bw -d <dev name=""> -i <port></port></dev>	8
	3)	On the client device issue the following command on command line:	9 10
	-,	a) ib send bw-d <dev name="">-i <pre>sport&gt;-s 1000000 -n 100</pre></dev>	11
	4)	Verify that the operation completed without error and the level of perfor-	12
	•)	mance achieved is reasonable and as expected.	13
			14
Example:			15
Devinto - Server			16
hca_id: mthca0			17
fw_ver: 1.2.0		-0000-b 4-b	18
sys image guid: 0002:	:c902 02:c9	:0020:b4dc 02:0020:b4df	19
vendor_id: 0x02c	;9		20
vendor_part_id: 2520	04		21
board id: MT 02	23000	00001	22
phys_port_cnt: 1			24
port: 1			25
max mtu:	2048	(4)	26
active_mtu: 2	2048	(4)	27
sm_lid: 1			28
port_inc: 0	x00		29
	<b>b</b> 0		30
Command Line: Ib_read_bw -d mti	ncau	-1 1	31
			32
DevInfo - Client			33 24
fw ver: 2.2.238	3		34
node_guid: 0002:	:c903	:0000:1894	36
sys_image_guid: 000	02:c9	03:0000:1897	37
vendor part id: 254	.ອ 18		38
hw_ver: 0xA0			39
board_id: MT_04	4A011	0002	40
priys_port_crit. 2 port: 1			41
1			42

OFA Interoperability Working Group	TI BASIC RDMA Interop using OFED - Command Line	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLA	N RELEASE 1.28	DRAFT

si m a si p	rate: hax_mtu: ctive_mtu: m_lid: prt_lid:	PORT_ACTIVE (4) 2048 (4) 2048 (4) 1 1		1 2 3 4
p	ort_lmc:	0x00		5
Command Line	ib_send_bw -	d mlx4_0 -i 1 10.0.0.1 -s 1	-n 100	6
		_		/ 8
				9
				10
				11
				12
				13
				14
				15
				16
				17
				18
				19
				20
				21
				22
				20
				25
				26
				27
				28
				29
				30
				31
				32
				33
				34
				35
				36
				37
				30 20
				39

- 40 41
- 42
| 12.7 TI BASIC RDMA INTEROP USING OFED - USING XANSTORM   | 1  |
|--|----|
| 12 7 1 Load XANStorm Test Configuration file   | 2  |
|  | 3  |
| <2 vml version="2.0" encoding="1.ITE_0" standolong="vec" 2   | 4  |
| <pre></pre> ///////////////////////////////////  | -  |
| <pre><!--DOCTIFE Xall.iviTTConfiguration--></pre>  | C  |
| <i <="" configuration="" iviti="" td=""><td>6</td></i>   | 6  |
| <rdmastreamlist></rdmastreamlist>  | 7  |
| <rdmastream id="1"></rdmastream>   | 8  |
| <requester></requester>  | 0  |
| <target></target>  | 9  |
|  | 10 |
| <rdmastream id="2"></rdmastream>   | 11 |
| <requester></requester>  | 12 |
| <target></target>  | 13 |
|  | 1/ |
| <rdmastream id="3"></rdmastream>   | 17 |
| <requester></requester>  | 15 |
| <target></target>  | 16 |
|  | 17 |
| <rdmastream id="4"></rdmastream>   | 18 |
| <requester></requester>  | 10 |
| <target></target>  | 10 |
|  | 20 |
| <rdmastream id="5"></rdmastream>   | 21 |
| <requester></requester>  | 22 |
| < Talget>  | 23 |
|  | 24 |
| <requester></requester>  | 2  |
| <taraet></taraet>  | 20 |
|  | 26 |
| <rdmastream id="7"></rdmastream>   | 27 |
| <requester></requester>  | 28 |
| <target></target>  | 29 |
|  | 20 |
| <rdmastream id="8"></rdmastream>   | 50 |
| <requester></requester>  | 31 |
| <target></target>  | 32 |
|  | 33 |
|  | 34 |
| <commandsequencelist></commandsequencelist>  | 35 |
| <commandsequence id="Small RDMA READ Procedure"></commandsequence>                                 | 55 |
| <rdmaoperation count="1000000" delay="0" size=" 1 B" type="Read"></rdmaoperation>                  | 36 |
|  | 37 |
| <commandsequence id="Large RDMA READ Procedure"></commandsequence>                                 | 38 |
| <rumaoperation count="10000" delay="0" size=" 64 MB" type="Read"></rumaoperation>                  | 39 |
|  | 40 |
| < CommandSequence ID="Small RDIVIA WITE Procedure" >   | 40 |
| <rdiviaoperation "="" b"="" count="1000000" delay="0" i="" size="" type="write"></rdiviaoperation> | 41 |
|  | 42 |

OFA Interoperability Working Group T	I BASIC RDMA Interop using OFED - Using XANStorm	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLAN	RELEASE 1.28	DRAFT

<commandsequence ddma="" dreadure"="" id="" orga="" rite="" w=""></commandsequence>	4
< CommandSequence ID= Large RDIVIA White Procedure > < RDMAOneration size=" 64 MB" count="10000" type="Write" delay="0" />	1
	.2
<commandsequence id="Small RDMA SEND Procedure"></commandsequence>	3
<pre><rdmaoperation count="1000000" delay="0" size=" 1 B" type="Send"></rdmaoperation></pre>	4
	5
<commandsequence id="Large RDMA SEND Procedure"></commandsequence>	6
<rdmaoperation count="10000" delay="0" size=" 64 MB" type="Send"></rdmaoperation>	0
	1
<commandsequence id="Small RDMA Verify Data Procedure"></commandsequence>	8
<rdmaoperation count="1000000" delay="0" size=" 1 B" type="Verify"></rdmaoperation>	9
	10
<commandsequence id="Large RDMA Verify Data Procedure"></commandsequence>	11
<rdmaoperation count="10000" delay="0" size=" 64 MB" type="Verify"></rdmaoperation>	10
	12
	13
<executionstream block="ON" checked="true" count="1" dolor="0"></executionstream>	14
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	15
<commandsequence>Small RDMA RFAD Procedure</commandsequence>	16
	17
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	17
<pre><rdmastream>2</rdmastream></pre>	18
<commandsequence>Large RDMA READ Procedure</commandsequence>	19
	20
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	21
<rdmastream>3</rdmastream>	22
<commandsequence>Small RDMA Write Procedure</commandsequence>	22
	23
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	24
<rdmastream>4</rdmastream>	25
<commandsequence>Large RDMA Write Procedure</commandsequence>	26
	27
<executionstream block="UN" checked="true" count="1" delay="0"></executionstream>	28
<rdmasheah>3</rdmasheah> <commandsaguanca>Small RDMA SEND Bracadura</commandsaguanca>	20
	29
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	30
<pre><rdmastream>6</rdmastream></pre>	31
<commandsequence>Large RDMA SEND Procedure</commandsequence>	32
	33
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	34
<rdmastream>7</rdmastream>	25
<commandsequence>Small RDMA Verify Data Procedure</commandsequence>	30
	36
<executionstream block="ON" checked="true" count="1" delay="0"></executionstream>	37
<rdmastream>8</rdmastream>	38
<commandsequence>Large RDMA Verify Data Procedure</commandsequence>	39
	10
	40
	41
	42

OFA Interoperability Working Group	TI BASIC RDMA Interop using OFED - Using XANStorm
OFA-IWG INTEROPERABILITY TEST PLA	AN RELEASE 1.28

## 12.7.2 Run XANStorm Application

XANStorm											
le Edit View Execute Help	75		15								
00000000	🚔 S, R,	w v	* 🔕 🤅	3 1							
Connections Command Sequences Res	ults										
	Command Sequences							RDMA Connec	tions		
Sequence ID	RDMA Operation	Size	Count	Delay	1 ID		Requester	Target	Status		
- Small RDMA READ Procedure		F			1		1		Disconnected		
E	Read	18	· 100000	0	2				Disconnected		
- Large RDMA READ Procedure	Cherch and a state	1			3				Disconnected		
	Read	1 MB	· 100	0	4				Disconnected		
G- Small RDMA Write Procedure					5				Disconnected		
	Write	18	▼ 100000	0	6				Disconnected		
🗈 📕 Large RDMA Write Procedure					7				Disconnected		
	Write	1 MB	▼ 100	0	8				Disconnected		
Small RDMA SEND Procedure								Even dan Chu			
	Send •	18	▼ 100	0	1 m		-	Execution Site	Jams	1	-
Large RDMA SEND Procedure	-	-				Connection ID	Command Seque	ence	Count	Delay	Block
	Send •	1 MB	▼ 100	0	×	1	Small RDMA	READ Procedure	1	0	ON
Small RDMA Verify Data Procedure	6. A	1		1.	×	2	Large RDM/	A READ Procedure	1	0	ON
	Verify	18	▼ 100000	0	× :	3	Small RDMA	Write Procedure	1	0	ON
E Large RDMA Venity Data Procedure	in at	T	Lun	1.		4	Large RDM/	A Write Procedure	1	0	ON
	venty	1 MB	♥ 100	0		5	Small RDMA	SEND Procedure	1	0	ON
					*	7	Large RDM/	A SEND Procedure	1	0	ON
						0	Large DDM	A Verify Data Procedure	1	0	ON
						P	Large RUM	A verity Data Procedure	1	0	UN

~~

#### 12.8 TI RDMA OPERATIONS USING OFED - COMMAND LINE 2 12.8.1 Purpose 3 This test is designed to identify problems that arise when RDMA operations are 4 performed over interconnection devices in the fabric. The test is not designed to measure the forwarding rate or switching capacity of a device, but does use per-5 formance measures to identify failures. 6 7 Test failures are identified by the following events: 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 11 Inconsistent performance levels. 12 12.8.2 General Setup 13 The RDMA interop procedure can be carried out using the OFA Verbs API to 14 create RDMA Connections and send RDMA operation or by using a 3rd party traffic generation tool such as XANStorm. 15 16 12.8.3 Topology 17 This test does not define a detailed topology and can be used either on a single 18 switch or across a RDMA fabric that may include gateways to and from other 19 technologies. The test configuration depends on the number of endpoints available to perform the testing. 20 21 12.8.4 Switch Load 22 The switch load test validates proper operation of a switch when processing a 23 large number of small RDMA frames. This test is analogous to normal switch 24 testing. 25 1) Attach a device to each port on the switch. 26 27 2) Select two ports on the switch to test (This will be your control stream) 28 3) Generate RDMA WRITE Operations of size 1024 bytes 100, 000 times on each device by issuing the following commands 29 30 a) On the server device issue the following command on command line: 31 ib\_read\_bw -d <dev\_name> -i <port> 32 b) On the client device issue the following command on command line: 33 i) ib read bw -d <dev name> -i <port> -s 1024 -n 100000 34 This must be done on both devices at the same time. 35 5) On all other pairs generate RDMA WRITE Operations of size 1 byte continu-36 ously until the control stream completes. 37 6) Repeat above steps until all port pairs are tested. 38 7) Repeat the above steps with all endpoint pairs, except the control stream 39 changed such that the size of the RDMA WRITE operation is 1,000,000 40 bytes (~1 MB) 41

12.8.5 Switch FAN in			1				
	The	e switch fan in test attempts to validate proper operation of RDMA exchanges	2				
	in t	ne presence of traffic loads that exceed the forwarding capacity of the switch.					
	pai	r.	4				
	•		5				
	1)	Connect all possible endpoint pairs such that data exchanges between pairs must traverse the pair of ports interconnecting the switch. The control con- nections must be across the interconnect network.	6 7 8				
	2)	Select two ports such that it has to cross both switches. (This will be your control stream)	9 10				
	3)	Generate RDMA WRITE Operations of size 1024 bytes 100, 000 times on each device by issuing the following commands	11 12				
		a) On the server device issue the following command on command line:	13				
		i) ib_read_bw -d <dev_name> -i <port></port></dev_name>	14				
		b) On the client device issue the following command on command line:	15				
		i) ib_read_bw -d <dev_name> -i <port> -s 1024 -n 100000</port></dev_name>	16				
	4)	This must be done on both devices at the same time.	17				
	5)	On all other pairs generate RDMA WRITE Operations of size 1 byte continu- ously until the control stream completes.	18 19				
	6)	Repeat above steps until all port pairs are tested.	20				
	7)	changed such that the size of the RDMA WRITE operation is 1,000,000 bytes (~1 MB)	22 23 24 25 26 27 28				
			29				
			30				
			31				
			33				
			34				
			35				
			36				
			37				
			38				
			39				
			40				
			41				

12.9 TI RDMA OPERATIONS USING OFED - USING XANSTORM	1
12.9.1 Load XANStorm Test Configuration file	2
	3
<2 vml version="2.0" encoding="LITE-8" standalone="ves" 2>	4
xan:iWITTConfiguration	5
<pre><xan:iwittconfiguration></xan:iwittconfiguration></pre>	0
<iwarpagentlist></iwarpagentlist>	0
<rdmastreamlist></rdmastreamlist>	7
<rdmastream id="1"></rdmastream>	8
<requester></requester>	9
<target></target>	10
	11
<rdmastream id="2"></rdmastream>	10
<requester></requester>	12
< Iarget>	13
<td>14</td>	14
<requester></requester>	15
<target></target>	16
	17
<rdmastream id="4"></rdmastream>	10
<requester></requester>	10
<target></target>	19
	20
<rdmastream id="5"></rdmastream>	21
<requester></requester>	22
< larget>	23
	24
	27
<commandsequence id="Control Stream Sequence"></commandsequence>	25
<pre><bdmaoperation count="100000" delay="0" size=" 1 kB" type="Write"></bdmaoperation></pre>	26
	27
<commandsequence id="Other Pairs Sequence"></commandsequence>	28
<rdmaoperation count="100000000" delay="0" size=" 1 B" type="Write"></rdmaoperation>	29
	30
	24
<executionstreamlist></executionstreamlist>	31
<executionstream block="OFF" checked="true" count="1" delay="0"></executionstream>	32
<rdmastream>1</rdmastream>	33
<commandsequence>Control Stream Sequence</commandsequence>	34
	35
<pdmastroom>2</pdmastroom>	36
<commandsequence>Control Stream Sequence</commandsequence>	27
	57
<executionstream block="OFF" checked="true" count="10000" delav="0"></executionstream>	38
<rdmastream>3</rdmastream>	39
<commandsequence>Other Pairs Sequence</commandsequence>	40
	41
<executionstream block="OFF" checked="true" count="10000" delay="0"></executionstream>	42
	r 🗠

OFA Interoperability Working Group	TI RDMA Operations using OFED - Using XANStorm	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLA	N RELEASE 1.28	DRAFT

<rdmastream>4</rdmastream>	1
<commandsequence>Other Pairs Sequence</commandsequence>	2
	3
<pre><rdmastream>5</rdmastream></pre>	4
<commandsequence>Other Pairs Sequence</commandsequence>	5
	6
	7
	1
	8
	9
	10

OFA Interoperability Working Group	TI RDMA Operations using OFED - Using XANStorm	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLAN	RELEASE 1.28	DRAFT

## 12.9.2 Run XANStorm Application

XANStorm														
ile Edit View Execute H	lelp													
		<b>-</b> !	5, R	V	K V, 🗶		2	Image: A state of the state						
Connections Command Sequ	ences Results	1												
	Command S	equenc	es							RDMA Cor	nections			
Sequence ID	RDMA Ope	ration	Size		Count	Delay	I	)	Reque	ster Target		Statu	s	
😑 📕 Control Stream Seque	nce						1					Disco	nnected	
	Write	-	1 kB	-	100000	0	2					Disco	nnected	
Other Pairs Sequence							3					Disco	nnected	
	Write	-	1B	-	100000000	0	4					Disco	nnected	
							5					Disco	nnected	
										Execution	Streams			
								Connection ID	Co	ommand Sequence		Count	Delay	Block
							1	<b>c</b> 1		Control Stream Se	quence	1	0	OFF
							1	<b>¢</b> 2		Control Stream Se	quence	1	0	OFF
							3	6 3		Other Pairs Seque	nce	10000	0	OFF
								<b>c</b> 4		Other Pairs Seque	nce	10000	0	OFF
								6 5		Other Pairs Seque	100	1000	0	OFF

12.10 TI MPI - HP-MPI USING OF	ED	1					
12.10.1 CLUSTER SETUP		2					
No	Note: The tests referenced below are in the following location and contain both $^{3}$						
32	and 64 bit versions:	4					
	http://www.iol.unh.edu/downloads/OFA/HP/ofatests_v3.tar.gz	5					
Να	te: HP-MPI is not part of the OFA Stack	7					
	·	8					
1)	Ethernet or some form of TCP/IP must be installed and configured on all	9					
	systems.	10					
2)	OFED library path must be configured on all systems (Idconfig should be executed after OFED installation).	11 12					
3)	OFED uDAPL /etc/dat.conf must match /sbin/ifconfig setup. (Modify	13					
	/etc/dat.conf and change the netdev reference to the appropriate interface	14					
		15					
4)	Suse Linux Enterprise Server 9 and 10.	16 17					
5)	All systems must be setup with identical user accounts (SSH access with no password prompts (key's setup) or rsh with .rhosts setup).	18 19					
6)	HP-MPI must be installed (in the same location) on all the machines in the cluster (or copying the HP-MPI tree to a shared directory also works).	20 21					
7)	Increase the max lockable memory limits on all the machines in the cluster:	22					
	a) edit /etc/security/limits.conf and add the following:	23					
	i) * hard memlock 500000	24					
	ii) * soft memlock 500000	25					
8)	A shared directory is very much recommended for ease of use in running the below tests.	26 27					
9)	Perl should be installed and located at /usr/bin/perl (or else the ex-	28					
	itpath/kill.pl script needs to be edited to point at an appropriate installation of	29					
	pen.	30					
12.10.2 REQUIRED FILES		31					
1)	HP-MPI is packaged as a binary .rpm. Version 2.3.1 has OFED 1.4.1 sup-	32					
	a) hpmpi-2 03 01 00-20090402r v86 64 rpm	33					
	a) <u>hpmpi 2.02.01.00-20030402r.206_04.pm</u> .	34					
	b) <u>hphph-2.03.01.00-200904021.1360.1phh</u> .	30					
2)	The version of an installed HP-MPI can be checked using mpiruh -version	30					
12.10.3 IEST SUITE INSTRUCTIONS		38					
1)	Although not absolutely required, these tests are easier to run from a shared directory, and the below instructions assume the use of one.	39					
2)	Download and unpack tests:	40					
2)	a) ofatests v2 tar dz	41					
		42					

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		TI MPI - HP-MPI using OFED RELEASE 1.28	March 27, 2009 DRAFT
:	3)	Unpack this preferably into a shared directory and cd int packs into:	o the directory it un-
		b) tar zxvf ofatests_v2.tar.gz	
		c) cd ofatests/	
	4)	Construct a file "hosts" on the machine you'll be running format should be:	"mpirun" from. The
		a) first_machine_name 2	
		b) second_machine_name 2	
		c) etc	
	5)	Later when this file is given with the "-hostfile" option to ' will launch two ranks on the first machine, two on the se	'mpirun", HP-MPI cond, etc.
12.10.4 BUILDING THE TESTS			
	1)	All the HP-MPI tests are shipped as a single binary hpmp permanent unrestricted license built in. It will run any of tests based on the first command line argument:	oitest.x which has a the following five
		a) IMB (command line "IMB")	
		b) rings2 (command line "rings2")	
		c) hello world (command line "hw")	
		d) fork (command line "fork")	
		e) ping pong ring (command line "ppr")	
		f) alltoone (command line "alltoone")	
		<b>Note</b> : For reference, these tests are included individ ests/src directory (except IMB which is available fror	ually in the ofat- n Intel).
12.10.5 RUNNING THE TESTS			
	1)	The test directory contains two scripts: "runit.ib" and "run runs the test in several different modes:	hit.iwarp", which
		a) for runit.ib:	
		i) IBV in RDMA mode with IBV intra-host	
		ii) IBV in SRQ mode with IBV intra-host	
		iii) UDAPL in RDMA mode with UDAPL intra-host	
		b) for runit.iwarp:	
		i) UDAPL in RDMA mode with UDAPL intra-host	
:	2)	Use runit.ib on a cluster with InfiniBand cards, or runit.iwa iWARP cards.	arp on a cluster with
12.10.6 CHECKING THE TEST STATU	JSE	S	
	1)	The "runit.*" scripts run all the tests one after the other, r or "FAILED" to stdout for each. If failures occur, they are	eporting "passed" logged in

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		ΤI	MPI - HP-MPI using OFED RELEASE 1.28	March 27, 2009 DRAFT	
	LO fail	G.ib ed jo	failures or LOG.iwarp.failures along with the ob.	full stdout/stderr for the	1 2
12.10.7 TEST DESCRIPTIONS					3
	1) Th	e HF	P-MPI test suite includes five tests:	4	4
	a)	det	ection	Ę	5
		i)	This is a simple ping pong application that i default interconnect selection. The other test terconnect selection.	s used to test HP-MPI's st cases use explicit in-	6 7 8
	b)	IMI	3	(	9
		i)	This is the Intel MPI Benchmark. If this pase teroperability of HP-MPI with the installed C	ses, then the basic in-	10 11
	c)	ring	gs2		12
		i)	This is a proprietary HP test which has a go interconnects to the point of failure. It also in tions	ood history of stressing ncludes 1sided opera-	13 14
	d)	for	1011S.		15
	u)	1011	Now PDMA implementations often have for	kissues and as now	16
		1)	OS kernels come out the fork problems son test makes a point of stressing that code pa	netimes re-appear. This	18
	e)	exi	tpath		19
	- ,	i)	The purpose of this test is simply to make su drivers etc remain stable when applications	re machines and OFED repeatedly terminate	20 21 22
	0		abnormally.		23
	T)	allt			24
		I)	This test has all the ranks send a flood of m make sure the interconnect can handle hea	vy load in that message	25 26
			pattern.		27
				2	28
					29
					30
					31
					32
					33
					34
					35 20
					30
					38
					39
					40

12.11 TI MPI - INTEL MPI USING O	OFE	D	1	
12.11.1 GENERAL ISSUES				
Note: Intel MPI is not part of the OFA Stack				
1)	Ne	twork configuration requirements	4	
1)		Ethernet must be installed and configured on all systems	5	
	a) b)	Ethernet must be installed and conligured on all systems.	7	
	(U	Lins hames must match hostnames.	8	
	C)	/etc/nosts should be setup with static IB nostnames and addresses.	9	
2)	OF		10	
	a)	OFED library path must be configured on all systems (Idconfig should be executed after OFED installation).	11 12	
	b)	OFED uDAPL /etc/dat.conf must match /sbin/ifconfig setup.	12	
3)	Se	tup Requirements	14	
	a)	All systems must be setup with identical user accounts on all nodes	15	
		(SSH access with no password prompts (key's setup) or rsh with .rhosts setup).	16	
	b)	Requires NFS setup from headnode and mount points (/home/test/ex- port) on user accounts.	17	
		<b>Note</b> : any node on the cluster can be setup as the headnode.	19	
	c)	MPI testing requires a reliable IB fabric without other fabric interop test- ing occurring.	20 21	
4)	He	re is the location for the free Intel MPI runtime environment kit	22	
	a)	http://www.intel.com/cd/software/products/asmo-na/eng/222346.htm	23	
5)	́Не	re is the location for the Intel MPI Benchmarks	24	
- /	a)	http://www.intel.com/cd/software/products/asmo-na/eng/clus-	20	
	- /	ter/mpi/219848.htm	27	
12.11.2 SETUP FOR THE CLUSTER			28	
1)	Ins	tall same O/S version on homogenous x86_64 systems. (Recommend	29	
	RH	IEL 5.2, EM64T). See the Figure 5- Intel Requirements for Homogeneous	30	
		with the second se	31	
	a)	nttp://www3.intel.com/cd/software/products/asmo-na/eng/308295.ntm	32	
2)	ins na	mes registered with DNS.	33 34	
3)	Ve	rify "hostname" on each system returns the hostname that DNS reports.	35	
			36	
			37	
			38	
			39	

- 40
- 41
- 42

		1
Hardware		2
Minimum	IA-32. Intel® 64 or IA-64 (formerly Itanium) architecture-based system.	3
Requirements	Examples of such Intel processors are:	4
	Intel® Pentium® 4 processor, or	5
	Intel® Xeon® processor, or	6
	Intel® Itanium® processor, or Intel® Core™2 Duo processor (example of Intel® 64 architecture)	7
		8
	Note that it is assumed that the processors listed above are configured into	9
	4 GB of RAM (8 GB of RAM recommended)	10
	1 GB of hard disk space (10 GB of space recommended)	10
Operation Custom		12
Operating System	Support	13
All three	Red Hat* Enterprise Linux* 4.0, 5.0	14
architectures	SUSE* Linux Enterprise Server* (SLES) 9, 10	10
IA-32 and Intel	Microsoft* Windows Vista*	10
64 architectures	The osoit windows visita	1/
Intel® 64 and	SGI ProPack* 5	10
architectures		19
IA-32 architecture only	Microsoft Windows* XP	20
Intel® 64	Red Hat Fedora 7 through 8	22
architecture only	cAos* 2	23
	CentOS* 4.6, 5.1	24
	openSuSE* Linux* 10.3	25
	Microsoft* Windows Compute Cluster Server 2003*	26
	Microsoft* Windows Server 2003*	27
	Microsoft* Windows XP Professional x64 Edition*	28
	Microsoft* Windows HPC Server 2008*	29
	Microsoft* Windows Server 2008*	30
Other Supported S	Software	31
Intel® MPI Bench	marks	32
Intel® Math Kern	el Library	33
Intel® Trace Anal	yzer and Collector	34
Intel® C++ Com	piler	35
Intel® Fortran Co	mpiler	36
Microsoft* Visual	Studio and Visual C++ Compilers	37
GNU C, C++, and	FORTRAN Compilers	38
OpenFabrics* Ent	erprise Distribution (OFED*)	. 39
		- 40

## Figure 5 - Intel Requirements for Homogeneous Environment

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	TI MPI - Intel MPI using OFED RELEASE 1.28	March 27, 2009 DRAFT
12.11.3 Setup information for OFED		1
1)	Install the current version of OFED on all systems.	2
2)	Bump up the max locked memory limits on the system.	3
	edit /etc/security/limits.conf and add the following:	4
	* hard memlock 500000	5
	* soft memlock 500000	7
3)	Run /sbin/ldconfig to pick up new OFED library path	8
4)	Modify /etc/hosts and add IB hostnames and addresses interfaces	s for the IB network 9
5)	Modify /etc/dat.conf and change the netdev reference to terface (ib0 or ib1) being used	o the appropriate in-
6)	Run OpenSM either on the headnode OR from one of t by pinging IB addresses on all systems.	he switches. Verify
12.11.4 Setup information for Intel M	א	1-
1)	Install Intel MPI in /opt/intel/mpi/3.1 on every system.	10
2)	Add identical user account (/home/test) on every syster "useradd –m test –u 555 –g users	n. For example
3)	Update the .bashrc for /home/test on every system:	1
	export PATH=\$PATH:./	21
	source /opt/intel/mpi/3.1/bin64/mpivars.sh	2
	# for IB, (mpi will default to rdssm if nothing defined)	22
	export I_MPI_DEVICE=rdssm	23
	# for ethernet	24
	export I MPI DEVICE=sock	2
	export MPIEXEC TIMEOUT=180	20
	ulimit -c unlimited	2
4)	Add .mpd.conf file in /home/test on every system.	20
•,	add single line "MPD_SECRETWORD=testing" to mpc	L conf 3
	chmod 600 /home/test/ mpd conf	3
5)	Add 2 mpd hosts files in /home/test on the headnode	one for ethernet and $3^{\circ}$
	one for IB	33
	Create mpd.hosts.ethernet and add a line for every sys using ethernet addresses or hostnames	tem on the cluster 34
	Create mpd.hosts.ib and add a line for every system or IPoIB addresses	the cluster using 30
6)	Add nfs export /home/test/export on headnode and cha mount points:	inge /etc/fstab for 38
	edit /etc/exports and add "/home/test/exports *(rw)" on	headnode 40
	edit /etc/fstab and add "hostname:/home/test/exports /ho on all other nodes	ome/test/exports nfs" 4 42

OFA Interoperability Working Group		TI MPI - Intel MF RELEASE	PI using OFED = 1.28	March 27, 2009 DRAFT
	7)	Untar the Intel Test	Suites on the headnode in /ho	me/test/exports
	8)	run mpdboot on the cluster and want to	e head node. For example: if yo run over ethernet:	ou have 6 nodes on the
		From the /home/tes ernet"	st directory run: "mpdboot –n 6	-r ssh -f ./mpd.host.eth-
	9)	Run test suite over	Ethernet to validate your instal	lation:
		"export I_MPI_DE∨	/ICE = sock"	
		run tests(refer to	test plan)	
		"mpdallexit"		
	10)	Run test suite over	IB	
	,	export I MPI DEVI	ICE = rdssm	
		mpdboot –n 6 –r ss	sh –f./mpd.host.ib	
		run tests (refer t	o test plan)	
		"mndallevit"		
12 11 5 ADDITIONAL INFORMATION		mpdallexit		
12.11.5 ADDITIONAL INFORMATION	N 1)	Co to the individual	I toot directories and follow the	atona in the respective
	1)	README-*.txt files	. The recommended order for r	unning the test suites in
		the order of increas	sing execution time:	0
		a) mpich2-test: se	e README-mpich2-test.txt file	
	2)	For Intel MPI Supp	ort Services go to:	
		http://www.intel.com	n/support/performancetools/clu	ster/index.htm
		See the Intel MPI R	Reference Manual for Additiona	l information
12.11.6 INTEL MPI BENCHMARK S	ETU	P		
	Th	e IMB tests must be	compiled with the -DCHECK co	ompiler flag set, to enable
	aut	omatic self-checking	g of the results. Modify the appr	opriate make_arch file as
	foll	OW:		
		MPI_HOME =		
		MPI_INCLUDE = .		
		LIB PATH =		
		LIBS =		
		CC =	mpicc	
		OPTELAGS = -	0	
		CPPFLAG5 =		
12.11.7 INTEL IHV TEST SUITE SET	ΓUΡ			
	All	test suites are config	gured, built, and run in a uniforr	n way.

<ul> <li>Configure for mpich2-test: ./configure –with-mpich2=/opt/intel/mpi/3.1 – cc=mpicc –f77=mpif77 –cxx=mpicxx</li> <li>Configure for IntelMPITEST: ./configure –with-mpich2=/opt/intel/mpi/3.1</li> <li>If you installed the library to another location, then replace the default Intel(R) MPI Library installation path "/opt/intel/mpi/2.0".</li> <li>A detailed description of the extra configuration options is contained in the</li> </ul>	1 2 3 4
<ul> <li>Configure for IntelMPITEST: ./configure –with-mpich2=/opt/intel/mpi/3.1</li> <li>If you installed the library to another location, then replace the default Intel(R) MPI Library installation path "/opt/intel/mpi/2.0".</li> <li>A detailed description of the extra configuration options is contained in the</li> </ul>	2 3 4
<ol> <li>If you installed the library to another location, then replace the default Intel(R) MPI Library installation path "/opt/intel/mpi/2.0".</li> <li>A detailed description of the extra configuration options is contained in the</li> </ol>	4
A detailed description of the extra configuration options is contained in the	5
spective README-*.txt file.	·e- 6
2) Run the tests:	8
If you use a Bourne-compatible shell (sh, bash, ksh, etc.), do:	9
export MPIEXEC_TIMEOUT=180	10
nohup make testing > xlog 2>&1 &	11
If you use a Csh-compatible shell (csh, tcsh, etc.), do:	12
seteny MPIEXEC_TIMEOUT 180	13
nohup make testing >&! xlog &	14
The expected duration of the test run is detailed in the respective README-*.	15 xt ₄c
file.	10
3) Check the results:	18
aron "Space" summany xml Lwo L	19
grop ">foil" ourmon(xml   wo	20
grep <rail -i<="" summary.xml="" td="" wc=""  =""><td>21</td></rail>	21
ADME-*.txt file.	22
12.11.8 TEST PROCEDURE	24
These sets of tests should be run for both Intel mpich2-test and the IntelMPITE suite:	эт 25 26
<b>Note:</b> "Set ulimit –c unlimited" to capture core files in case of abnormal termin tions.	a- 27 28
Test suite mpich2-test: use default settings with no environment variables.	28
Test suite IntelMPITEST: use default settings with no environment variables	31
12 11 9 INTERPRETING THE RESULTS	33
<ol> <li>For mpich2-test test suites: See Table <u>- TI - Intel MPICH2 Suite - (Not part</u> OFA Stack)</li> </ol>	of 34 35
The <b>summary.xml</b> file produced by the test suites has the following unifo format:	m 36
<ul> <li>The file header contains information on the test suite and testing env ronment.</li> </ul>	38
The rest of the file represents the results of the test suite run.	40
<ol> <li>For IntelMPITEST test suite: See Table - TI - Intel MPI Test Suite De- scription - (Not part of OFA Stack)</li> </ol>	41

TI MPI - Intel MPI using OFED	March 27, 2009
RELEASE 1.28	DRAFT

The <b>Tests/summary.xml</b> file produced by the test suites has the following uniform format:	-
• The file header contains information on the test suite and testing environment	
The rest of the file represents the results of the test suite run.	1
	é
	-
	8
	Q
	,
	,
	,
	,
	,
	,
	4
	4
	4
	4
	4
	4
	4
	4
	4
	-

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN

12.12 TI MPI - OPEN MPI USING C	FED	1
12.12.1 Cluster setup		2
1)	Network configuration requirements	3
	a) All systems must be reachable by each other a common network that supports TCP (Ethernet, IPoIB, etc.).	4 5
	<ul> <li>All nodes must agree on the IP addresses for all TCP networks on all systems (e.g., via /etc/hosts, DNS, or some other mechanism).</li> </ul>	6 7
	c) If multiple, physically separate OpenFabrics networks are used in the testing, then all the devices on each network must report a subnet ID through the verbs stack that is both the same as all other ports on the same physical network and unique among all other ports on other physical networks.	8 9 10 11 12
	<b>Note</b> : this is a new requirement among all the MPI's. This likely means that IB vendors will need to change the default subnet ID reported by their systems. It is only necessary for testing scenarios when multiple physically separate OpenFabrics networks are available, such as (but not limited to):	13 14 15 16
	<ul> <li>so-called "multi-rail" scenarios, where one or more systems in the test have multiple OpenFabrics devices, each connected to physi- cally separate networks</li> </ul>	17 18 19
	<ul> <li>when some systems are connected to IB network A, and other sys- tems are connected to IB network B</li> </ul>	20
2)	The same version of OFED must be installed in the same filesystem location on all systems under test.	22
3)	The same version of Open MPI must be available in the same filesystem lo- cation on all systems under test.	23 24
	<ul> <li>a) Open MPI can be installed once on a shared network filesystem that is available on all nodes, or can be individually installed on all systems. The main requirement is that Open MPI's filesystem location is the same on all systems under test.</li> </ul>	25 26 27 28
	b) If Open MPI is built from source, theprefix value given to configure should be the filesystem location that is common on all systems under test. For example, if installing to a network filesystem on the filesystem server, be sure to specify the filesystem location under the common mount point, not the "native" disk location that is only valid on the file server.	29 30 31 32 33
	c) Note that Open MPI only started supporting iWARP as of the v1.3 se- ries. As of this writing, Open MPI v1.3 has not yet been released, but nightly "snapshot" tarballs are available on the Open MPI web site.	34 35 36
	<ul> <li>d) The version of Open MPI can be obtained by running "ompi_info   head".</li> </ul>	37 32
4)	All systems must be setup with at least one identical user account. This user must be able to SSH or RSH to all systems under test from the system that will launch the Open MPI tests with no additional output to stdout or stderr (e.g., all SSH host keys should already be cached, no password/passphrase prompts should be emitted, etc.).	39 40 41 42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		TI MPI - Open MPI using OFED RELEASE 1.28	March 27, 2009 DRAFT
	5)	The lockable memory limits on each machine shou limited locked memory per process.	ld be set to allow un-
	6)	The underlying OpenFabrics network(s) used in the and reliable.	e test should be stable
	7)	No other fabric interoperability tests should be runn tests.	ing during the Open MPI
	8)	Note that Open MPI is included in some Linux distr ating systems. Multiple versions of Open MPI can p system as long as they are installed into separate f configured with a differentprefix argument). All MI run with a single installation of Open MPI.	ibutions and other oper- beacefully co-exist on a ilesystem locations (i.e., PI tests must be built and
	9)	MPI tests should be run across at least 2 separate of the OpenFabrics networks (vs. using just shared communication).	systems to force the use I memory for in-system
	10)	Open MPI currently assumes a "mostly homogener software and hardware of all systems under test are They should be the same general server vendor m hardware accessories, running the same operating the same features enabled, etc. Some "drift" betwee test is acceptable (in both software and hardware), interoperability testing, it is best to test with system neous as possible. Specifically: this interoperability mogeneous and heterogeneous network hardware server hardware or software. Open MPI v1.3.x doe neous HCA/ RNICs in a single MPI job. Such heter hardware scenarios may be supported via manual "out of the box" in a future version.	ous" scenario, where the e more-or-less the same. odel with similar operating system with en the systems under but for the scope of this s that are as homoge- testing focuses on ho- , not heterogeneous s not support heteroge- ogeneous network parameter tweaking, or
12.12.2 TEST SETUP			
	1)	The following values are used in examples below:	
		code resides.	the Open MPI source
		<ul> <li>\$MPIHOME: The absolute directory location of tion that is common to all systems under test.</li> </ul>	the Open MPI installa-
	2)	Open MPI can be used from the OFED installation, quired, can be downloaded and installed from the r	or, if a later version is re- nain Open MPI web site:
		http://www.open-mpi.org/	
		<ul> <li>a) If building Open MPI from source, and if the Op headers are installed in a non-default location, openib=<dir> option to configure to specify the location.</dir></li> </ul>	penFabrics libraries and be sure to use thewith- OpenFabrics filesystem
	3)	Create a hostfile listing the hostname of each syste test. If a system under test can run more than one multiprocessor or multicore systems), either add a each hostname indicating how many processes to r the hostname as many times as MPI processes are	m that will be used in the MPI process (such as "slots" parameter after run on that system, or list e desired. For example,

4)

5)

for am	two 4 processor systems named node1.example.com and node2.ex- ple.com:	1 2
	shell\$ cat hostfile.txt	3
	node1.example.com slots=4	4
	node2.example.com slots=4	5
	shell\$ cat equivalent-hostfile.txt	6
	node1.example.com	7
	node1.example.com	8
	node1.example.com	9
	node1.example.com	11
	node2.example.com	12
	node2.example.com	13
	node2 example com	14
	node2 example com	15
	shells	16
On	en MPI defaults to probing all available networks at run time to determine	17
whi <b>*oi</b> to∃ fore	ich to use. OpenFabrics testing should specifically force Open MPI to <b>nly</b> <sup>*</sup> use its OpenFabrics stack for testing purposes (e.g., do not fail over ICP if the OpenFabrics stack is unavailable). There are three ways to ce Open MPI to use the OpenFabrics stack by default:	18 19 20 21
a)	Set a per-user file that is visible on all nodes (either if the \$HOME is a networked filesystem that is common to all systems under test, or this process is invoked on all systems):	22 23
	shell\$ mkdir \$HOME/.openmpi	24
	shell\$ cat > \$HOME/.openmpi/mca-params.conf < <eof< td=""><td>26</td></eof<>	26
	btl = openib,self,sm	27
	EOF	28
b)	Set an environment variable on the node/shell where mpirun is invoked:	29
	# sh-flavored shells	30
	shell\$ export OMPI_MCA_btl=openib,self,sm	31
	# csh-flavored shells	32
	shell% setenv OMPI_MCA_btl openib,self,sm	33
c)	Add an extra command line parameter to mpirun (not shown in all the examples below):	34 35
	shell\$ mpirunmca btl openib,self,smrest of command line	36
Op ality with	en MPI includes several trivial test programs to verify basic MPI function- y. Assuming the Open MPI source tree is available, the tests can be built	37 38 30
**11	shells of SOMPL SOURCE TREE/examples	40
	shell\$ make	41
		42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		RELEASE 1.28 March 27, 2009 DRAFT
	6)	NetPIPE should be obtained from its main web site:
		http://www.scl.ameslab.gov/netpipe/
		a) Open MPI should be in the \$PATH so that "mpicc" can be found. The test suite can then be built with:
		shell\$ cd NetPIPE-3.7.1
		shell\$ make mpi
	7)	The Intel MPI Benchmarks should be obtained from the same URL provided in the Intel MPI test section of this document.
		a) The test suite can be built with:
		shell\$ cd IMB_3.1/src
		shell\$ make -f make_mpich MPI_HOME=\$MPIHOME
	8)	It may be desirable to set the shell to unlimit the size of corefiles for analysis
		of aborted tests. This limit should be set in the shell startup files of the test
		user on every hode.
12.12.3 IEST PROCEDURE	4)	The following values are used in everylas helow
	1)	a) COMPLE COUPCE, TREE The directory where the Open MPL sources
		a) \$0MPI_SOURCE_TREE: The directory where the Open MPI source code resides.
		<ul> <li>\$MPIHOME: The absolute directory location of the Open MPI installa- tion that is common to all systems under test.</li> </ul>
		c) \$NP: The number of MPI processes to use in the test. Unless otherwise specified, it is usually the sum of the number of processors on all sys- tems under test.
		d) \$HOSTFILE: The absolute filename location of the hostfile.
	2)	Ensure that the Open MPI installation includes OpenFabrics support:
		shell\$ \$MPIHOME/bin/ompi_info   grep openib
		MCA btl: openib (MCA v1.0, API v1.0.1, Component v1.4)
		The exact version numbers displayed will vary depending on your version of Open MPI. The important part is that a single "btl" line appears showing the openib component.
	3)	Basic Open MPI run-time functionality can first be verified by running simple non-MPI applications. This ensures that the test user's rsh and/or ssh set- tings are correct, etc.
		shell\$ \$MPIHOME/bin/mpirun -np \$NPhostfile \$HOSTFILE hostname
		The output should show the hostname of each host listed in the hostfile. If a host was listed with "slots=X", the hostname should appear X times. The list
		of hostnames may appear in random order; this is normal. Note that any se- rial application can be run; "hostname" is a good, short test that clearly iden- tifies that specific hosts were used, etc.
	4)	Basic Open MPI functionality can be verified with several trivial test pro- grams that are included in Open MPI. Run them with:
		shell\$ cd \$OMPI_SOURCE_TREE/examples

		_
	shell\$ \$MPIHOME/bin/mpirun -np \$NPhostfile \$HOSTFILE hello_c	1
	shell\$ \$MPIHOME/bin/mpirun -np \$NPhostfile \$HOSTFILE ring_c	2
	The first program prints a simple "hello world" message from each MPI pro-	3
	that trivial MPI applications are able to start, properly initialize, properly fi-	4
	nalize, and exit successfully. The lines may output out of order; this is normal.	6
	The second program sends a message around in a ring. In addition to testing	7
	the same functionality as "hello world", it exercises basic message passing (using the OpenEabrics verbs stack in this case). The output should indicate	8
	that a message was sent around a ring 10 times, and then that each process	9
	exited successfully. Some lines may be output out of order; this is normal.	10
	The same two test programs are also available in C++, Fortran 77, and For- tran 90, but they are not relevant to this test.	11 12
5)	NetPIPE can only be run with 2 MPI processes. It can be invoked:	13
	shell\$ cd NetPIPE-3.7.1	14
	shell\$ \$MPIHOME/bin/mpirun -np \$NPbynodehostfile \$HOSTFILE \	15
		10
	nodes (to force testing of the network, as opposed to shared memory).	18
	NetPIPE will run through ping-pong benchmarks of a variety of message	19
	sizes. It is fairly obvious if NetPIPE hangs or fails to complete successfully.	20
6)	The Intel MPI benchmarks can be invoked with the following:	21
	shell\$ cd IMB_3.1/src	22
	shell\$ \$MPIHOME/bin/mpirun -np \$NPbynodehostfile \$HOSTFILE \ IMB-MPI1 -multi 0 PingPong PingPing	23 24
	shell\$ \$MPIHOME/bin/mpirun -np \$NPhostfile \$HOSTFILE IMB-MPI1	25
	shell\$ \$MPIHOME/bin/mpirun -np \$NPhostfile \$HOSTFILE IMB-IO	26
	The first command runs just the PingPong and PingPing point-to-point	21
	benchmarks, but makes all the MPI processes active in a pairwise fashion. The "bynode" option forces Open MPI to place successive MPI processes	29
	on separate nodes (to force testing of the network, as opposed to shared	30
	memory).	31
	The second command runs all the benchmarks in the suite. Depending on the number of processes in the test, it may take a while to run.	32 33
	The third command runs a variety of MPI file tests, each of which involve MPI	34
	these warnings are a known issue and are safe to ignore. Depending on the	35
	number of processes in the test and the back-end filesystem used, it may	36
	take a long time to run. Periodic "hang"-like behavior is also not uncommon (largely caused by filesystem issues). For small node/process counts, hangs	37
	shouldn't last for more than 1-2 minutes each. For larger node/process	30
	counts, the hangs may be longer.	40
		41

12.13 TI MPI - Оню STATE UNIVE 12.13.1 MVAPICH 1 - Setup	RSI	TY USING OFED	1 2
1)	Ne	twork configuration requirements	3
	a)	All systems must be reachable by each other a common network that supports TCP (Ethernet, IPoIB, etc.)	4 5
	b)	All nodes must agree on the IP addresses for all TCP networks on all systems (e.g., via /etc/hosts, DNS, or some other mechanism).	6 7
2)	The on	e same version of OFED must be installed in the same filesystem location all systems under test.	8 9
3)	M∨ MV	APICH1 is included in OFED distributions. The updated versions of APICH1 can be obtained from OpenFabrics website.	10 11
4)	The cat	e same version of MVAPICH must be available in the same filesystem lo- ion on all systems under test.	12 13
	a)	MVAPICH1 can be installed once on a shared network filesystem that is available on all nodes, or can be individually installed on all systems. The main requirement is that MVAPICH1 filesystem location is the same on all systems under test.	14 15 16 17
5)	All mu will (e.g	systems must be setup with at least one identical user account. This user st be able to SSH or RSH to all systems under test from the system that launch the MVAPICH1 tests with no additional output to stdout or stderr g., all SSH host keys should already be cached, no password/passphrase mpts should be emitted, etc.).	18 19 20 21
6)	The lim cor	e lockable memory limits on each machine should be set to allow un- ited locked memory per process. This can be achieved by using ulimit nmand.	22 23 24
7)	The No MV	e underlying IB network(s) used in the test should be stable and reliable. other fabric interoperability tests should be running during the APICH1 tests.	25 26
8)	Mu as a d sta	Itiple versions of MVAPICH can peacefully co-exist on a system as long they are installed into separate filesystem locations (i.e., configured with ifferentprefix argument). All tests must be built and run with a single in- llation of MVAPICH.	27 28 29 30
9)	M∨ the mu	APICH tests should be run across at least 2 separate systems to force use of the IB networks (vs. using just shared memory for in-system com- nication).	31 32 33
12.13.2 MVAPICH 1 - TEST SETUP A	ND	PROCEDURE	34
1)	Tes	st Setup	35
	a)	Create a hostfile listing the hostname of each system that will be used in the test. If a system under test can run more than one MPI process (such as multiprocessor or multicore systems) list the hostname as many times as MPI processes are desired. For example, for two 2 pro- cessor systems named host1 and host2	36 37 38 39
		\$ cat hostfile.txt host1 host1	40 41 42

#### host2 host2 2 b) Download and install Intel® MPI Benchmarks on all nodes from: 3 http://www.intel.com/cd/software/products/asmo-4 na/eng/cluster/mpi/219848.htm 5 Follow the instructions below to install: 6 7 untar downloaded archive i) 8 ii) open <natured directory>/src/make mpich and fill in the following variables: 9 MPI HOME=<path to mvapich1 directory> #mine was 10 /usr/mpi/gcc/mvapich-1.0.1 11 **CPPFLAGS= -DCHECK** 12 iii) gmake -f make mpich 13 This will install the benchmarks inside the MPI HOME/tests directory 14 15 Note: Intel® MPI Benchmarks are installed with OFED installation by default 16 17 c) Enter all nodes and run the following commands: 18 echo "PATH=\\$PATH:<path to mvapich1 directory>/bin:<path to i) mvapich1 directory>/tests/IMB-3.0" >> /<username>/.bashrc # or 19 .cshrc 20 ii) echo "ulimit -l unlimited" >> /<username>/.bashrc # or .cshrc 21 iii) source /<username>/.bashrc # or .cshrc 22 Note: these commands may fail or produce unexpected results with a 23 shared \$HOME 24 2) Testing Procedure 25 a) The following values are used in the examples below 26 27 \$MPIHOME - The absolute directory location of the MVAPICH ini) stallation that is common to all systems under test 28 ii) \$NP - The number of MPI processes to use in the tests. Unless oth-29 erwise specified, it is usually the sum of the number of cores on all 30 systems under test 31 iii) \$HOSTFILE - The absolute location of the hostfile 32 b) Run Intel® MPI Benchmarks: 33 Run the PingPong and PingPing point-to-point tests 34 \$MPIHOME/bin/mpirun -np \$NP IMB-MPI1 -multi 0 PingPong Ping-35 Ping -hostfile \$HOSTFILE 36 ii) Run all the tests (PingPong, PingPing, Sendrecv, Exchange, Bcast, 37 Allgather, Allgatherv, Alltoall, Reduce, Reduce\_scatter, Allreduce, 38 Barrier), in non-multi mode. 39 \$MPIHOME/bin/mpirun -np \$NP IMB-MPI1 -multi 0 -hostfile \$HOST-40 FILE 41

12.13.3 MVAPICH 2 - SETUP			1
1)	Do	wnload and install OFED on all nodes from:	2
	<u>htt</u>	p://www.openfabrics.org/downloads/OFED	3
2)	Do	wnload and install Intel® MPI Benchmarks on all nodes from:	4
	htt	p://www.intel.com/cd/software/products/asmo-	5
	na	/eng/cluster/mpi/219848.htm	6
	Yo	u will have to accept a license. Follow the instructions below to install.	/ 0
	a)	untar downloaded archive	0
	b)	open <untarred directory="">/src/make_mpich and fill in the following vari- ables:</untarred>	10
		<ul> <li>MPI_HOME=<path directory="" mvapich2="" to=""> #mine was /usr/mpi/gcc/mvapich2-1.0.3</path></li> </ul>	12
		ii) CPPFLAGS= -DCHECK	13
	c)	gmake -f make mpich	14
	Ť	his will install the benchmarks inside the MPI HOME/tests directory	15
3)	All	nodes should be physically connected.	17
4)	Er	ter all nodes and run the following cmds:	18
-,	a)	echo "PATH=\\$PATH: <pre>path to myapich2 directory&gt;/bin:<pre>coath to</pre></pre>	19
	.,	mvapich2 directory>/tests/IMB-3.0" >> / <username>/.bashrc # or .cshrc</username>	20
	b)	echo "ulimit -l unlimited" >> / <username>/.bashrc;</username>	21
	c)	source / <username>/.bashrc # or .cshrc</username>	22
5)	Cr	eate an mpi ring:	23
	a)	Construct a file called hosts that has the following format. Include as many lines as you have hosts. Be sure to leave a blank line at the end of the file:	24 25 26
		i) <host>ifhn=<infiniband address="" ip=""></infiniband></host>	27
	b)	Run the following commands	28
	,	<ul> <li>i) mpdboot -n `cat hosts wc -l` -f hostsifhn=<localhost infiniband="" ip<br="">address&gt;</localhost></li> </ul>	29 30
		ii) mpdtrace -I #OPTIONAL shows current ring members	31
12.13.4 MVAPICH 2 - TEST PROCE	URF	,	32
Sten A:	un a	subnet manager from one node only	33
	ana		34
Step B R	un In	tel® MPI Benchmarks:	35
1)	Ти	o sets of tests should be run, with these command lines	36
,	a)	mpirun -np <number node="" nodes="" number="" of="" processors="" x=""> IMB-MPI1</number>	31
	.,	-multi 0 PingPong PingPing	30
	a)	mpirun -np <number node="" nodes="" number="" of="" processors="" x=""> IMB-MPI1</number>	40
	Th bu	e first command runs just the PingPong and PingPing point-to-point tests, t makes all tasks active (pairwise).	41 42

The second command runs all the tests (PingPong, PingPing, Sendrecv, Ex- 1 change, Bcast, Allgather, Allgatherv, Alltoall, Reduce, Reduce\_scatter, Allre-2 duce, Barrier), in non-multi mode. 2) If the test passes shutdown current subnet manager and start another one on a different node; run both tests again. 3) Repeat until all nodes have run a subnet manager and passed all tests. 

13 INFINIBAND SPECIFIC INTEROP PROCEDURES USING WINOF			1
13.1 IB LINK INITIALIZE USING WINOF			2
13.1.1 Disconnect the full topology and select a cable whose length should be a maximum of 15 meters for SDR and 10 meters for DDR when using copper cables.			3 4 5
<b>No</b> pai Wi	red t nOF	The WinOF Subnet Manager and diagnostics are still evolving as com- o OFED. Therefore, you must include an OFED Linux node along with the node to run diagnostics for this test.	6 7 8
1)	Ve	rify that no SM is running	9 10
2)	Со	nnect two devices back to back	10
3)	ssł	n to the OFED node.	12
	a)	Run "ibdiagnet -lw 4x" to verify portwidth	13
	b)	Run "ibdiagnet -ls 2.5" to check link speed. Interpret output and com- pare to advertised speed.	14 15
	<b>No</b> oth	<b>te</b> : This command will only produce output if the link speed is anything her than SDR. Keep this in mind during your interpretation of the output.	16 17 18
4)	Re	peat steps 1-3 with a different device pairing.	19
	a)	All device pairs must be tested except target to target: HCA to HCA, HCA to Switch, HCA to Target, Switch to Switch, and Switch to Target.	20 21
	b)	Each device must link to all other devices in order for the device to pass link init over all.	22 23
13.1.2 Recommendations			24
In col inte	orde mma erpre	r to determine Switch to Target and Switch to Switch link parameters, run ands from an HCA linked to the switch under test. This does require more etation of the output to differentiate the reported parameters.	25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 28
			29

13.2 IB FABRIC INITIALIZATION U	SING WINOF	1
13.2.1 Architect the Network we w	ant to build.	2
р р	<b>lote</b> : The WinOF Subnet Manager and diagnostics are still evolving as com- ared to OFED. Therefore, you must include an OFED Linux node along with the	3 4 -
v H	ICA to HCA pairs.	5 6
1	) Create a table of IP addresses to accign	7
2	) Create topology file - this makes sure that the subnet is configured as ex-	8
	pected - i.e. SDR and DDR links. This inserts the name of devices as well as	9
	the GUID.	10
3	5) See <u>Figure 6- Sample Network Configuration</u> below.	11
		12
13.2.2 Connect the HCAs and swit	ches as per the Architected Network and make sure that no SM/SA is	13
furning on the Fabric.		14
		15
13.2.3 Procedure		16
1	) Start an SM on a device and let it initialize (all SMs will need to be tested)	17
2	) Visually verify that all devices are in the active state. Orange led will be on if	18
	the port is active.	19
3 4	<ul> <li>Wait 17 seconds as per the specifications requirements</li> </ul>	20
5	<ul> <li>Run "ibdiagnet -c 1000" to send 1000 node descriptions.</li> </ul>	21
6	) Run "ibdiagnet -r" to generate fabric report.	22
7	) Run "ibchecknet" to build guid list.	23
		24
13.2.4 Verification Procedures		25
1	) Review "PM Counters" section of the fabric report. There should be no il- legal PM counters. The Specification says there should be no errors in 17	26 27
2	Seconds.	28
2	running SM is the one you started and verify number of nodes and switches	29
	in the fabric.	30
3	Review the ibchecknet report and verify that there are no duplicate GUIDs in the febric	31
	the fabric	32
		33
F	Restart all devices in the fabric and follow Sections 13.2.3 and 13.2.4. Run the	34
S	M from a different device in the fabric until all SMs present have been used. All	35
	SMs on managed switches and one instance of <b>opensm</b> must be used.	36
E	ach device must pass all verification procedures with every SM to pass Fabric	37
lı lı	nitialization test.	38
		39
		40
		41

# Table 31 - ibdiagnet commands

Commands	Description
ibdiagnet -h	Help
Ibdiagnet - pc	Clear Counter
Ibdiagnet -lw 4x - ls 2.5	Specify link width and speed
Ibdiagnet -c 1000	send 1000 Node Descriptions



13.3 IB IPOIB DATAGRAM MOI	DE (	<b>DM</b> ) ເ	JSING WINOF	1	
13.3.1 Setup				2	
	<b>No</b> t will	t <b>e</b> : Wir suppc	OF 2.0.2 only supports IPoIB Datagram Mode. Future WinOF releases ort IPoIB Connected-Mode.	4	
	0			5	
	tha	nnect t t no Sl	M is running on the Fabric.	6 7	
	Thi	s nroc	edure, as the previous ones, will be based on the cluster connectivity	8	
	An	SM/SA	A 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 would only run SM/SA for the partner pair (with a switch in the middle). This pr		10		
			y run SM/SA for the partner pair (with a switch in the middle). This pro-	11	
	ceo	lure ha	as been developed for the Windows environment.	12	
	Op	tional	: 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	
	det	detailed below in subsequent sections.			
13.3.2 IPoIR Interface Creation a	bnd		Subnet Creation	16	
15.5.2 IFOID Interface Creation and IFO			Subiel Creation	17	
	1)	Conii	gure iPoiB address. All addresses must reside on the same subnet.	18	
	2)	Verify	which 'Local Area Connection' the IPoIB interfaces are bound to:	19	
		a) S	tart   Server Manager   View Network Connections.	20	
		b) F	ind the OpenFabrics IPoIB interfaces (one per HCA port). If your plat-	21	
		fo	orm has two Ethernet ports, then IPoIB interfaces likely will be assigned	22	
		n n	et ports are assigned 'Local Area Connection' and 'Local Area Con-	23	
		n	ection 2'.	24	
	3)	Set in	terfaces to 10.0.0.x/24 (10.0.0.x/netmask 255.255.255.0) using the fol-	25	
	,	lowing	g commands:	26	
		a) n	etsh interface ip set address "Local Area Connection 3" static	27	
		1	0.10.4.x 255.255.255.0	28	
		b) n	etsh interface ip set address "Local Area Connection 4" static	29	
		1	0.10.4.y 255.255.255.0	30	
	4)	View	the IPoIB IP address using the following command	31	
		a) n	etsh interface ip show address "Local Area Connection 3"	32	
				33	
				34	
13.3.3 Ping Procedures				35	
Step A	1)	Stop	all SM's and verify that none are running	36	
	2)	Powe	er cycle all switches in the fabric (this insures that the new SM will con-	37	
		figure all the links and create the multi-cast join).			
	3)	Start	an SM (All SM's will need to be tested) and let it initialize	39	
		Note	For link testing it is recommended to use an OFED Linux OpenSM as	40	
		the W	/indows version of OpenSM does not support all SA queries and func-	41	
		tional	ity of the UFED 1.4 OpenSM.	42	

	<b>Note</b> : All WinOF installed systems contain a disabled OpenSM windows service. A WinOF installation option/feature is to automatically 'start/enable' the OpenSM service on the local node.	1 2
	<ul> <li>Start   Server Manager   Configuration   Services   InfiniBand Subnet Manager   Automatic   apply</li> </ul>	4
	<ul> <li>Start   Apply will enable the local OpenSM to start and be started upon system boot.</li> </ul>	5
	a) Visually verify that all devices are in the active state. Orange led will be on if the port is active.	/ 8
	b) From a Linux system, Run "ibdiagnet -r" and verify that the SM you started is the one that is running and and that it is the master. You will need to know the GUID of the device since the SM will be reassigned on each reboot; the Windows 'vstat' command displays HCA info.	9 10 11 12
	c) Verify that all nodes and switches were discovered.	13
	d) WinOF 2.0.2 does not provide a ibdiagnet utility.	14
	<b>Note</b> : Ibdiagnet may show more switches than indicated by the physical number of switch platforms present. This is because some switches have multiple switch chips.	15 16 17
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).	18 19
5)	Ping every IPolB interface IPv4 address except localhost with packet sizes of 64, 256, 511, 512, 1024, 1025, 2044, 4096, 8192, 16384, 32768, and 65507. 'ping /?' displays ping help.	20 21
	a) 100 packets of each size will be sent	22
	b) Every packet size is a new ping command.	23
6)	In order to pass Step A, a reply must be received for every ping sent (without losing a single packet) while using each one of the SMs available in the cluster.	25 26 27
Step B 1)	Bring up all HCAs but one.	20 29
2)	Start an SM (all SMs will need to be tested).	30
3)	Check for ping response between all node (All to All).	31
	a) A response from the disconnected HCA should not be returned.	32
4)	Disconnect one more HCA from the cluster.	33
5)	Ping to the newly disconnected HCA from all nodes (No response should be returned).	34 35
6)	Connect the first machine (the one that was not connected) and check for ping response from all nodes that are still connected.	36 37
7)	Connect the disconnected HCA to a different switch on the subnet which will change the topology.	38 39
8)	Ping again from all nodes (this time we should get a response).	40 41 42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	IB IF	OIB Datagram Mode (DM) using WinOF RELEASE 1.28	March 27, 2009 DRAFT	_
	9) Fo Se ne th	ollow Step B, this time bring the interface down a erver Manager   View Network Connections   IP ection) disable and enable commands instead o e HCAs.	and then back up: Start   oIB(Local Area con- f physically disconnecting	1 2 3
	No or	ote: Each step must exhibit the expected behavior der for the device to pass Step B overall.	ior while using each SM in	4 5 6
Step C	Follow instan for the	v Step A and B using a different SM until all SM's ce of each available SM is required. Steps A, B, e device to pass 13.3.3 overall.	have been used. Only one and C must pass in order	7 8 9
13.3.4 FTP PROCEDURE				11
	FTP p partne FTP c	rocedures requires an FTP server to be configur r pair. An FTP client needs to be available on ea lient is a standard Windows component.	ed on each machine in the ach machine as well; an	12 13 14
	An FT which	P server is a component of the IIS ' <b>Internet Infor not</b> a part of a standard Windows installation:	mation Services' manger	15 16
	See S ager.	tart   Server Manager   Roles   Add IIS. Configu	re FTP server via IIS man-	17 18 19 20
13.3.4.1 SETUP				21 22
	1) M	ake sure ftpd is installed on each node for the F	TP application.	23
	2) A	special account for this should be created as fo	llows:	24
	b)	Username: Interop		25
	c)	Password: openfabrics		26
13.3.4.2 PROCEDURE				27
	Run F	TP server on all nodes		28
				30
	1) St	art an SM (all SMs will need to be tested) and le ork utilities docs)	et it initialize (ref MS Net-	31
	a)	Verify that the running SM is the one you star	ted.	32
	2) F	······································		33
	a)	Connect an HCA pair via FTP on IPoIB using and password.	the specified user name	34 35
	b)	Put the 4MB file to the %windir%\temp folder dows\Temp) on the remote host.	(generally C:\Win-	36
	c)	Get the same file to your local dir again.		30 20
	d)	Binary compare the file using the Windows co tfile.orig'.	ommand 'fc /B tfile	40
		i) The two must be identical		42

3)	Repeat the procedure with a different SM.	1
	Note: Every node must FTP the 4MB file to all others using all SMs and the	2
	files must be identical as determined by the binary compare in order for the	3
	device to pass 13.3.4 overall.	4
	<b>Note</b> : Sections 13.3.3 and 13.3.4 must pass using the configuration determined by sections 13.3.1 and 13.3.2 for the device to pass IPoIP Datagram	5
	mode overall.	6
		7
		8
		9
		10
		11
		12
		13
		14
		10
		17
		18
		10
		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 IB SM FAILOVER AND HANDO	VER PROCEDURE USING WINOF	1
13.4.1 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
13.4.2 PROCEDURE		6
1)	Disable all SMs in the cluster.	0
2)	Start a SM on either machine in a chosen pair.	9
	<ul> <li>a) Start   Server Manager   Configuration   Services   InfiniBand Subnet Manager   start   apply</li> </ul>	10 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
	<ul> <li>Start   Control Panel   Network and Sharing Center   Manage Net- work Connections.</li> </ul>	16 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
	<ul> <li>c) Verify the IPoIB devices (one per cabled connected HCA port) are visible &amp; operational from a device driver perspective using Device Manager</li> </ul>	21 22 23
	i) Start   Run   devmamt.msc	24
	<ul> <li>d) Ping the IPoIB interface IPv4 address local and remote, verify traffic is actually going in/out over IPoIB 'local area connection x'.</li> </ul>	25 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
	<b>Note</b> : 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 with the smallest guid is master and the other is in standby.	30 31 32
7)	Verify that all nodes in the cluster are present - ping all IPoIB interfaces	33
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 other goes into standby again; see '%windir%\temp\osm.log'.	39 40
13)	Verify that all nodes in the cluster are present - ping all IPoIB interfaces	41 42

<ul> <li>15) Verify that the previous master SM is still the master; view "%windir%\templosm.log'.</li> <li>16) Verify that all nodes in the cluster are present - ping all IPolB interfaces</li> <li>17) Repeat proceeding steps [1-16] 2 more times with the same node pair, ensuring that the below criteria is met (total of 3 tests per pair which can be run in any order): <ul> <li>a) First SM to be started having highest numerical priority value.</li> <li>b) Second SM to be started having highest numerical priority value.</li> <li>c) Both SMs having equal numerical priority values.</li> </ul> </li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>	14)	Shutdown the standby SM.	-
<ul> <li>16) Verify that all nodes in the cluster are present - ping all IPoIB interfaces</li> <li>17) Repeat proceeding steps [1-16] 2 more times with the same node pair, ensuing that the below criteria is met (total of 3 tests per pair which can be run in any order): <ul> <li>a) First SM to be started having highest numerical priority value.</li> <li>b) Second SM to be started having highest numerical priority value.</li> <li>c) Both SMs having equal numerical priority values.</li> </ul> </li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>	15)	Verify that the previous master SM is still the master; view '%windir%\temp\osm.log'.	2
<ul> <li>17) Repeat proceeding steps [1-16] 2 more times with the same node pair, ensuring that the below criteria is met (total of 3 tests per pair which can be run in any order): <ul> <li>a) First SM to be started having highest numerical priority value.</li> <li>b) Second SM to be started having highest numerical priority value.</li> <li>c) Both SMs having equal numerical priority values.</li> </ul> </li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>	16)	Verify that all nodes in the cluster are present - ping all IPoIB interfaces	2
<ul> <li>a) First SM to be started having highest numerical priority value.</li> <li>b) Second SM to be started having highest numerical priority value.</li> <li>c) Both SMs having equal numerical priority values.</li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>	17)	Repeat proceeding steps [1-16] 2 more times with the same node pair, en- suring that the below criteria is met (total of 3 tests per pair which can be run in any order):	6
<ul> <li>b) Second SM to be started having highest numerical priority value.</li> <li>c) Both SMs having equal numerical priority values.</li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>		a) First SM to be started having highest numerical priority value.	8
<ul> <li>c) Both SMs having equal numerical priority values.</li> <li>18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.</li> </ul>		b) Second SM to be started having highest numerical priority value.	(
18) Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.		c) Both SMs having equal numerical priority values.	,
	18)	Repeat steps 1-17 until all possible SM pairs from identical vendors in the cluster have been tested.	
			,
			,
			,
			4
			4
			4
			4
			4
			1
			4
			,
			4
13.5 IB SRP USING WINOF			1
--------------------------	----	---	----------------
13.5.1 SETUP			2
	1)	Connect the HCAs and switches as per the Architected Network and make sure that no SM is running on the Fabric.	3 4
	2)	Configure and Start a Linux OFED SRP target - VDISK BLOCKIO mode; (some assembly required) - <u>https://wiki.openfabrics.org/tiki-</u> index.php?page=SRPT+Installation	5 6 7
		a) assume /dev/sdb1 & /dev/sdc1 are formatted with /sbin/mkfs.msdos	8
		<ul> <li>b) Setting SRPT_LOAD=yes in /etc/infiniband/openib.conf is not good enough. It only loads ib_srpt module and does not load scst and its dev_handlers.</li> </ul>	9 10 11
		c) modprobe scst	12
		d) modprobe scst_vdisk	13
		e) echo "open vdisk0 /dev/sdb BLOCKIO" > /proc/scsi_tgt/vdisk/vdisk	14
		<ul> <li>f) echo "open vdisk1 /dev/sdc BLOCKIO" &gt; /proc/scsi tgt/vdisk/vdisk</li> </ul>	15
		g) echo "add vdisk0 0" >/proc/scsi tgt/groups/Default/devices	16
		h) echo "add vdisk1 1" >/proc/scsi_tgt/groups/Default/devices	17 18 10
13.5.2 WINDOWS PROCEDURE			20
	1)	Start an SM (all SM's will need to be tested) and let it initialize	21
		a) Verify that the running SM is the one that you started	22
	2)	Choose a node to work with	20
	3)	Verify the SRP driver loaded correctly; locate the SRP Miniport.	25
		a) Start  Control Panel   Device Manager   Storage Controllers [InfiniBand SRP Miniport]	26 27
	4)	Discover + Enable (bring online) the SRP drive(s)	28
		a) Start   Server Manager   Storage   Disk Management	29
	5)	You will find a basic 'unknown' and 'offline' disk; this one of your SRP volume(s).	30 31
	6)	Right-click the offline disk and select 'online'.	32
	7)	Right-click the volume space, assign the drive letter 'T'.	33
	8)	Right-click the volume space, format the volume.	34
	9)	Access the SRP drive via assigned drive letter. From a Windows/DOS command prompt window, execute the following commands.	35
		a) vol T:	37
		b) dir T:\ (should be empty)	30
		c) mkdir T:\tmp	35 21
		d) copy /B WinOF_wlh_x64.msi T:\tmp	41
		e) fc /B WinOF_wlh_x64.msi T:\tmp\WinOF_wlh_x64.msi	42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN		IB SRP using WinOF RELEASE 1.28	March 27, 2009 DRAFT	-
				_
	f)	copy /B T:\tmp\WinOF_wlh_x64.msi T:\tmp\WOF2.m	ısi	1
	g)	fc /B T:\tmp\WinOF_wlh_x64.msi T:\tmp\WOF2.msi		2
	h)	fc /B WinOF_wlh_x64.msi T:\tmp\WOF2.msi		3
	i)	copy /B T:\tmp\WOF2.msi WOF3.msi		4
	j)	fc /B WinOF_wlh_x64.msi WOF3.msi		5
	k)	del T:\tmp\WOF2.msi		7
	I)	del T:\tmp\WinOF_wlh_x64.msi		8
	m)	dir T:\tmp (should be empty)		9
	n)	rmdir T:\tmp		10
	o)	dir T:\ (should be empty)		11
	p)	del WOF3.msi		12
	10) Foi	each SRP target located in Procedure #4		13
	a)	Repeat step 9 for all volumes found for all targets as	s determined by	14
		Windows Procedure step #4 - see <u>Discover + Enabl</u>	e (bring online) the	16
	11) Tal	e SRP drive offline		17
	a)	Start   Server Manager   Storage   Disk Managemen	ht .	18
	a) b)	Right-click the online disk and select 'offline'		19
	c)	dir T·\ (should fail)		20
	0) 12) Re	boot all devices in the fabric and repeat the procedur	e using a different	21
	SN		e using a unerent	22
	Note: A	n HCA must successfully complete all operations to a	and from all volumes	24
	on all ta	argets using all available SM's in order to pass SRP t	esting. One volume	25
	pertar			26
				27
				28
				29
				30
				32
				33
				34
				35
				36
				37
				38 20
				39 40
				41
				42

13.6 IB UDAPLTEST COMMANDS U	SING WINOF	1
Serve	er Command: dapl2test -T S -D <ia_name></ia_name>	2
		3
13.6.1 IB SETUP		4
•	orrect interface is used. The DAPL interface for IB is ibnic0v2.	5 6
• 1	t is also important to verify that the desired dat/dapl libraries are available	7
	%windir%\dat2.dll	8
	%windir%\dapl2.dll	9
• 1	Fo run dapl2test on IB, an SM needs to be running.	10
13 6 2 GROUP 1. POINT-TO-POINT TOP		11
	1 connection and simple send/recv:	12
[1.0]	danl2test TT s centrer names D cia names i 100 t 1 w 1 B BE	14
•	client SR 256 1 server SR 256 1	15
[1 /]	Verification polling and scatter gather list:	16
['] •	danl2test TT s ceaver names D cia names i 100 t1 w 1 V P P	17
	BE	18
	client SR 1024 3 -f \	19
	server SR 1536 2 -f	20
13.6 IB UDAPLTEST COMMANDS USING WINOF         Server Command: dapl2test -T S -D <ia_name>         13.6.1 IB SETUP         • The % SystemDrive%/UDAT/dat.conf needs to be verified to be sure that the correct interface is used. The DAPL interface for IB is ibnic0v2.         • It is also important to verify that the desired dat/dapl libraries are available         • %windir%/dat2.dll         • %windir%/dat2.dll         • Windir%/dat2.dll         • To run dapl2test on IB, an SM needs to be running.         13.6.2 GROUP 1: POINT-TO-POINT TOPOLOGY         [1.3] 1 connection and simple send/recv:         • dapl2test -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -R BE         • client SR 256 1 server SR 256 1         [1.4] Verification, poling, and scatter gather list:         • dapl2test -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE         • client SR 1024 3 -f\         • server SR 1536 2 -f         13.6.3 GROUP 2: SWITCHED TOPOLOGY         InfiniBand Switch: Any InfiniBand switch         [2.5] Verification and private data:         • dapl2test -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE         • client SR 1024 1\         • server SR 1024 1         [2.6] Add multiple endpoints, polling, and scatter gather list:         • dapl2test -T T -s <server_name> -D <ia_name> -i 100 -t 1 -w 1 -V -P -R BE         • g</ia_name></server_name></ia_name></server_name></ia_name></server_name></ia_name></server_name></ia_name>		21
Infini	Band 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 BE</ia_name></server_name>	25 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</ia_name></server_name>	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 BE</ia_name></server_name>	34 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 BE</ia_name></server_name>	39 40
	client SR 256 1 \	41
		42

OFA Interoperability Working Group OFA-IWG INTEROPERABILITY TEST PLAN	IB uDAPLTEST Commands using WinOF RELEASE 1.28	March 27, 2009 DRAFT
	• server RR 4096 1 server SR 256 1	
13.6.4 GROUP 3: SWITCHED TOPO	LOGY WITH MULTIPLE SWITCHES	
	[3.5] Multiple threads, RDMA Read, and RDMA Write:	
	<ul> <li>dapl2test -T T -s <server_name> -D <ia_name></ia_name></server_name></li> <li>BE</li> </ul>	> -i 100 -t 4 -w 8 -V -P -R
	client SR 256 1 \	
	• server RR 4096 1 server SR 256 1 client SR 25	56 1 server RR 4096 1 \
	server SR 256 1	
	[3.6] Pipeline test with RDMA Write and scatter gather	list:
	<ul> <li>dapl2test -T P -s <server_name> -D <ia_name< li=""> <li>8192 2</li> </ia_name<></server_name></li></ul>	> -i 1024 -p 64 -m p RW
	[3.7] Pipeline with RDMA Read:	
	<ul> <li>dapl2test -T P -s <server_name> -D <ia_name: 4096 2</ia_name: </server_name></li> </ul>	> -i 1024 -p 64 -m p RR
	[3.8] Multiple switches:	
	<ul> <li>dapl2test -T T -s <server_name> -D <ia_name></ia_name></server_name></li> </ul>	> -i 100 -t 1 -w 10 -V -P -R
	• BE client SR 1024 3 \	
	server SR 1536 2	
3.6.5 WINOF DAPL2TEST WRAPPE	ER SCRIPTS	
	All the specified DAPL tests are conveniently located in DAPL test server & client scripts.	the WinOF distributed
	<ul> <li>%ProgramFiles(x86)%\WinOF\dt-svr.bat</li> </ul>	
	<ul> <li>To run the dapl2test Server, to a Windows of type 'dt-svr'. Only one server is necessary communicate with a single dapl2test server different nodes can exist. A single dapl2test with only one dapl2test server at a time.</li> </ul>	cmd-prompt window – multiple clients can r; multiple servers on t client communicates
	<ul> <li>No further server action is required as the or sistent; looping waiting for dapltest client re</li> </ul>	dapl2test server is per- equests.
	<ul> <li>%ProgramFiles(x86)%\WinOF\dt-cli.bat</li> </ul>	
	'dt-cli' no arguments, will display dt-cli com	mand args & options.
	<ul> <li>Dapl2test client invocation: 'dt-cli IPoIB_IPv cmd'</li> </ul>	v4_server_address
	<ul> <li>If the dt-svr command was executed on a sy interface address is 10.10.4.200 then</li> </ul>	vstem where the IPoIB
	<ul> <li>'dt-cli 10.10.4.200 interop' would run the at tween the client and server.</li> </ul>	oove dap2tests be-
	<ul> <li>'dt-cli 10.10.4.200 conn' is a simple, quick t client   server connection is operational.</li> </ul>	est to verity dapl2test

13.7 IB MPI - INTEL MPI USING W	NО	<b>DF</b>	1
13.7.1 Requirements			2
1)	Inte froi	el MPI is not part of the WinOF installation; acquire Intel MPI installer file m Intel.	3 4
2)	Ins x86	stall same O/S version (Windows Server 2008-HPC) on homogenous 6_64 systems.	5
3)	MF occ	PI testing requires a reliable IB fabric without other fabric interop testing curring.	7
4)	Pri	ivate Ethernet Network configuration	)
	a)	DNS names must match hostnames in hosts file.	10
5)	Ŵi	nOF Installation requirements	11
,	a)	Install the latest version of WinOF on all systems (double-click WinOF_wlh_x64.msi); see	12 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 switches.	16 17
		iv) If OpenSM on the headnode, select WinOF install feature 'OpenSM Started'.	18 19
	b)	Once WinOF installation on all nodes has completed, configure IPoIB interfaces.	20 21
		<ul> <li>i) %windir%\system32\Drivers\etc\hosts should be setup with IB host- names and static IP addresses.</li> </ul>	22 23
		ii) Assign IPv4 address, from hosts file, to each IPoIB interface; Exam- ple: Local Area Connection 3 is the 1st IPoIB interface.	24 25
		<ul> <li>netsh interface ip set address "Local Area Connection 4" static 10.10.4.y 255.255.255.0</li> </ul>	26
		This allows you to <b>set</b> the IPoIB IP address.	11 20
		netsh interface ip show address "Local Area Connection 3"	29
		This allows you to <b>view</b> the IPoIB IP address.	30
		iii) Verify by pinging IPoIB interface addresses on all nodes.	31
		3	32
42.7.2 Setup information for Intel MDI	I		33
13.7.2 Setup information for inter MPI	tall I	Intel MPI on every cluster node:	34
113	tan i		35
1)	Inte	el MPI runtime environment kit	36
	a)	http://www.intel.com/cd/software/products/asmo-na/eng/308295.htm	37
2)	Inte	el MPI Benchmarks,	20 20
	a)	http://www.intel.com/cd/software/products/asmo-na/eng/clus-	10
3)	Ad	ld identical user account (%SystemDrive%\users\test) on every node.	∔1 42

4) Headnode mount points (%SystemDrive%\test\export) on user accounts.

				2
13.7.3 Additional Information				3
	1)	Go t	o the individual test directories and follow the steps in the respective	4
	•		DMEtxt mes.	6
	2)	For	ntel MPI Support Services go to:	7
		a) [	<u> http://software.intel.com/en-us/articles/intel-mpi-library-for-win-</u> <u>dows/all/1/</u>	8
		b) 3	See Intel MPI Reference Manual for Additional information	9
				10
13.7.4 Intel MPI (MVAPICH 2) - Te	st P	roced	ure	11
, , ,	1)	Run	a subnet manager from one node only.	12
	2)	Run	Intel® MPI Benchmarks from the HPC head-node	13
	_)	a) -	Two sets of tests should be run, with these command lines	14
		a)	wo sets of tests should be full, with these command lines	15
			IMB-MPI1 -multi 0 PingPong PingPing	16 17
			mpirun -np <number node="" nodes="" number="" of="" processors="" x=""></number>	18
			IMB-MPI1	19
			The first command runs just the PingPong and PingPing point- to-point tests, but makes all tasks active (pairwise).	20 21
			The second command runs all the tests (PingPong, PingPing, Sendrecv, Exchange, Bcast, Allgather, Allgatherv, Alltoall, Re- duce, Reduce_scatter, Allreduce, Barrier), in non-multi mode.	22 23 24
		b) (	f the test passes shutdown current subnet manager and start another one on a different node; run both tests again.	25 26
	3)	Rep	eat until all nodes have run a subnet manager and passed all tests.	27
				28
				29
13.7.5 Interpreting the results				30
	1)	TBA		31
				32
				33
				34
				35
				36
				37

14 BUG REPORTING METHODOLOGY	DURING	PRE-TESTING	1
The inte	following roperabilit	bug reporting methodology will be followed during the execution of y pre-testing at UNH-IOL.	23
1)	UNH-IOL Logic, Me focal poin	and the OEMs (e.g. Chelsio, Data Direct, Flextronics, Intel, LSI Ilanox, NetEffect, Obsidian, QLogic and Voltaire) will assign a t of contact to enable fast resolution of problems.	4 5 6
2)	Bug repo	ts will include:	7
	a) Detai	ed fail report with all relevant detail (Test/Application, Topology.).	8
	b) [For l	B] IB trace if needed.	9
	c) [For i'	WARP] iWARP, TCP and SCTP traces if needed.	11
3)	Bug report the switch	ts will be sent via mail by UNH-IOL to the focal point assigned by OEM	12
4)	Bug repo	ts and suggested fixes will be sent to the OpenFabrics devel-	14
	opment c UNH-IOL be mainta	ommunity - <u>OFA Bugzilla</u> . When such reports are communicated, will ensure that confidentiality between UNH-IOL and the OEM will ined. Bug reports will be generalized and not include any company	15 16
	specific p	roprietary information such as product name, software name,	17
5)		C.	18
5)	OpenFab	rics repository. Documentation related to fixes will not mention any	20
Na	Company	specific prophetary information.	21
IET	e: This les F iWARP	during or after interoperability testing at plugfests.	22
			23
			24
			26
			27
			28
			29
			30
			31
			১∠ 33
			34
			35
			36
			37
			38
			39
			4U ⊿1
			- <b>T</b> I

#### **15 RESULTS SUMMARY**

#### **15.1 INFINIBAND SPECIFIC TEST RESULTS**

Please add a check mark whenever a test case passes and when the system is behaving according to the criteria mentioned below. Otherwise indicate a failure along with a comment explaining the nature of the failure.

#### **Results Table 1 - IB Link Initialize**

Test #	Test	Pass	Fail	Comment	8
1	Phy link up all ports				9
2	Logical link up all ports switch SM				1
3	Logical link up all ports HCA SM				1:
					1

#### **Results Table 2 - IB Fabric Initialization**

Test #	Test	Pass	Fail	Comment	17
1	Verify that all ports are in Armed or Active state				18

#### **Results Table 3 - IB IPoIB - Connected Mode (CM)**

Test #	Test	Pass	Fail	Comment	2
1	Ping all to all - Ping using SM 1				2
2	Ping all to all - Ping using SM 2				2
3	Ping all to all - Ping using SM 3				2
4	Ping all to all - Ping using SM 4				2
5	Ping all to all - Ping using SM 5				3
6	Ping all to all - Ping using SM 6				3
7	Ping all to all - Ping using SM x				3
8	Connect/Disconnect Host				3
9	FTP Procedure				3

1 2

3

4

5 6

7

14 15

16

21 22

- 38 39
- 40
- 41
- 42

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			

#### **Results Table 4 - IB IPoIB - Datagram Mode (DM)**

#### Table 5 - IB SM Failover/Handover

Test #	Test	Pass	Fail	Comment	
1	Basic sweep test				
2	SM Priority test				
3	Failover test - Disable SM1				
4	Failover test - Disable SM2				

#### **Results Table 6 - IB SRP**

Test #	Test	Pass	Fail	Comment	4
1	Basic dd application				0.00
2	IB SM kill				3
3	Disconnect Initiator				3
4	Disconnect Target				3

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				

### **Results Table 8 - Ethernet Gateway - (IB Specific)**

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			

#### **15.2 ETHERNET SPECIFIC TEST RESULTS** 2 **Results Table 9 - Ethernet Link Initialize** 3 Test # Test Fail Pass Comment 4 5 1 Phy link up all ports 6 2 Verify basic IP connectivity 7 8 9 10 **Results Table 10 - Ethernet Fabric Initialize** 11 Test # Test Pass Fail Comment 12 1 Fabric Initialization 13 14 15 16 **Results Table 11 - Ethernet Fabric Reconvergence** 17 18 Test # Test Pass Fail Comment 19 1 Fabric Reconvergence 20 21 22 **Results Table 12 - Ethernet Fabric Failover** 23 24 Test # Test Pass Fail Comment 25 Fabric Failover 1 26 27 28 29 **Results Table 13 - iWARP Connectivity** 30 Test # Pass Fail 31 Test Comment 32 1 Group 1 - Verify that each single iWARP operation over 33 single connection works 34 2 Group 2 - Verify that multiple iWARP operations over a 35 single connection work 36 3 Group 3 - Verify that multiple iWARP connections work 37 4 Group 4 - Verify that disconnect/reconnect physical con-38 nections work 39 5 Group 5 - Verify that IP Speed negotiation work 40

41 42

1

Group 6 - Verify that iWARP error ratio work

OFA Interoperability Working Group	Ethernet Specific Test Results	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLAN	RELEASE 1.28	DRAFT

## **Results Table 13 - iWARP Connectivity**

Test #	Test	Pass	Fail	Comment	
7	Group 7 - Verify that stress pattern over iWARP work				
8	Group 8 - Verify that iWARP parameter negotiation work				

#### **15.3 TRANSPORT INDEPENDENT TEST RESULTS**

#### **Results Table 14 - TI iSER**

Test #	Test	Pass	Fail	Comment	6
1	Basic dd application				7
2	IB SM kill				8
3	Disconnect Initiator				1
4	Disconnect Target				1
5	Repeat with previous SM Slave				1

#### **Results Table 15 - TI NFS Over RDMA**

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			

#### **Results Table 16 - TI RDS**

Test #	Test	Pass	Fail	Comment	34
1	rds-ping procedure				3
2	rds-stress procedure				36

#### **Results Table 17 - TI SDP**

Test #	Test	Pass	Fail	Comment
1	netperf procedure			
2	FTP Procedure			
3	IB SCP Procedure			
4	iWARP SCP Procedure			

#### **Results Table 18 - TI uDAPL**

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			

## **Results Table 19 - TI Basic RDMA Interop**

Test #	Test	Pass	Fail	Comment	35
1	Small RDMA READ				36
					37
2	Large RDMA READ				38
3	Small RDMA Write				39
4	Large RDMA Write				41

42

#### **Results Table 19 - TI Basic RDMA Interop**

http://www.openfabrics.org/

Test #	Test	Pass	Fail	Comment	3
5	Small RDMA SEND				4
6	Large RDMA SEND				5
7	Small RDMA Verify				6
8	Large RDMA Verify				8

#### **Results Table 20 - TI RDMA operations over Interconnect Components**

			I	1	12
Test #	Test	Pass	Fail	Comment	13
1	Switch Load				14
2	Switch Fan In				15

HP-MPI Test Results RELEASE 1.28

\_

## 15.4 HP-MPI TEST RESULTS

#### Results Table 21 - TI MPI - HP-MPI - (Not part of OFA Stack)

Test #	Test Suite	Pass	Comment	6
1	IMB			7
2	rings2			8
3	fork			9
4	exitpath			11
5	alltoone			12
				13

Intel MPI Test Results RELEASE 1.28

1 2 3

4

24 25

26

#### **15.5 INTEL MPI TEST RESULTS**

#### Results Table 22a - Intel MPI Benchmark Summary

Test #	Test Suite	Pass	Fail	Comment	(
1	Test 1: PingPong				1
2	Test 1: PingPing				8
3	Test 1: Sendrecv				
4	Test 1: Exchange				
5	Test 1: Allreduce				
6	Test 1: Reduce				
7	Test 1: Reduce_scatter				
8	Test 1: Allgather				
9	Test 1: Allgatherv				
10	Test 1: Alltoall				
11	Test 1: Alltoallv				*
12	Test 1: Bcast				
13	Test 1: Barrier				

#### Results Table 22b - TI MPI - Intel MPICH2 (Not part of OFA stack) Pass/Fail Summary

Test #	Test Suite	Pass	Fail	Comment	2
1	attr				29
2	coll				3(
3	comm				31
4	datatype				32
5	errhan				34
6	group				35
7	info				36
8	init				31
9	pt2pt				39
10	rma				4(
11	spawn				41

### Results Table 22b - TI MPI - Intel MPICH2 (Not part of OFA stack) Pass/Fail Summary

Test #	Test Suite	Pass	Fail	Comment	2
12	topo				4
13	io				5
14	F77				6
15	схх				8
16	threads				9

#### Results Table 22c TI MPI - Intel MPI (Not part of OFA stack) Test Failure Details

Test #	Test Suite	Pass	Comment	
1	testlist2l (1085 tests)			
2	testlist2-21 (23 tests)			-
3	testlist4 (216 tests)			1
4	testlist4lg (1 test)			1
5	testlist6 (46 tests)			2

## **15.6 OPEN MPI TEST RESULTS**

	Results Table 23 -	TI MPI	- Open	MPI	2 3 4
Test #	Test Suite	Pass	Fail	Comment	5
	Phase 1: "	Short"	tests	·	7
2	OMPI built with OpenFabrics support				8
3	OMPI basic functionality (hostname)				10
4.1	Simple MPI functionality (hello_c)				11
4.2	Simple MPI functionality (ring_c)				12
5	Point-to-point benchmark (NetPIPE)				13
6.1.1	Point-to-point benchmark (IMB PingPong multi)				15
6.1.2	Point-to-point benchmark (IMB PingPing multi)				16
	Phase 2: "	'Long"	tests		17 18
6.2.1	Point-to-point benchmark (IMB PingPong)				19
6.2.2	Point-to-point benchmark (IMB PingPing)				20
6.2.3	Point-to-point benchmark (IMB Sendrecv)				21
6.2.4	Point-to-point benchmark (IMB Exchange)				23
6.2.5	Collective benchmark (IMB Bcast)				24
6.2.6	Collective benchmark (IMB Allgather)				25
6.2.7	Collective benchmark (IMB Allgatherv)				20
6.2.8	Collective benchmark (IMB Alltoall)				28
6.2.9	Collective benchmark (IMB Reduce)				29
6.2.10	Collective benchmark (IMB Reduce_scatter)				30
6.2.11	Collective benchmark (IMB Allreduce)				32
6.2.12	Collective benchmark (IMB Barrier)				33
6.3.1	I/O benchmark (IMB S_Write_Indv)				34
6.3.2	I/O benchmark (IMB S_IWrite_Indv)				35
6.3.3	I/O benchmark (IMB S_Write_Expl)				37
6.3.4	I/O benchmark (IMB S_IWrite_Expl)				38
6.3.5	I/O benchmark (IMB P_Write_Indv)				39
6.3.6	I/O benchmark (IMB P_IWrite_Indv)				40 41

## Results Table 23 - TI MPI - Open MPI

Test #	Test Suite	Pass	Fail	Comment 2
6.3.7	I/O benchmark (IMB P_Write_Shared)			4
6.3.8	I/O benchmark (IMB P_IWrite_Shared)			5
6.3.9	I/O benchmark (IMB P_Write_Priv)			6
6.3.10	I/O benchmark (IMB P_IWrite_Priv)			8
6.3.11	I/O benchmark (IMB P_Write_Expl)			9
6.3.12	I/O benchmark (IMB P_IWrite_Expl)			1
6.3.13	I/O benchmark (IMB C_Write_Indv)			1
6.3.14	I/O benchmark (IMB C_IWrite_Indv)			1
6.3.15	I/O benchmark (IMB C_Write_Shared)			1
6.3.16	I/O benchmark (IMB C_IWrite_Shared)			1
6.3.17	I/O benchmark (IMB C_Write_Expl)			1
6.3.18	I/O benchmark (IMB C_IWrite_Expl)			1
6.3.19	I/O benchmark (IMB S_Read_Indv)			1
6.3.20	I/O benchmark (IMB S_IRead_Indv)			2
6.3.21	I/O benchmark (IMB S_Read_Expl)			2
6.3.22	I/O benchmark (IMB S_IRead_Expl)			2
6.3.23	I/O benchmark (IMB P_Read_Indv)			2
6.3.24	I/O benchmark (IMB P_IRead_Indv)			2
6.3.25	I/O benchmark (IMB P_Read_Shared)			2
6.3.26	I/O benchmark (IMB P_IRead_Shared)			2
6.3.27	I/O benchmark (IMB P_Read_Priv)			2
6.3.28	I/O benchmark (IMB P_IRead_Priv)			3
6.3.29	I/O benchmark (IMB P_Read_Expl)			3
6.3.30	I/O benchmark (IMB P_IRead_Expl)			3
6.3.31	I/O benchmark (IMB C_Read_Indv)			3
6.3.32	I/O benchmark (IMB C_IRead_Indv)			3
6.3.33	I/O benchmark (IMB C_Read_Shared)			3
6.3.34	I/O benchmark (IMB C_IRead_Shared)			3
6.3.35	I/O benchmark (IMB C_Read_Expl)			3
6.3.36	I/O benchmark (IMB C_IRead_Expl)			4

OFA Interoperability Working Group	Open MPI Test Results	March 27, 2009
OFA-IWG INTEROPERABILITY TEST PLAN	RELEASE 1.28	DRAFT

# Results Table 23 - TI MPI - Open MPI

Test #	Test Suite	Pass	Fail	Comment	3
6.3.37	I/O benchmark (IMB Open_Close)				4
					5
					6
					7
					õ
					9 10
					11
					12
					13
					14
					15
					16
					17
					18
					20
					21
					22
					23
					24
					25
					26
					27
					20
					30
					31
					32
					33
					34
					35
					36
					37
					30 30
					40
					41
					42

OSU MPI Test Results RELEASE 1.28

#### **15.7 OSU MPI TEST RESULTS**

#### **Results Table 24 - TI MPI - OSU**

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			

#### **Results Table 25 Remarks**

**General Remarks:** Comments about the set-up, required updates to the TD, and any other issues that came up during the testing.