OFA IWG meeting – 11/05/2013

Attendees

First	Last	Company	09/10/13	09/24/13	10/22/13	11/05/13
Tom	Reu	Chelsio	X	X	X	X
Martin	Schlining	DataDirect Networks		X	X	
Jeff	Kopko	Emulex	X	X	X	
Pradeep	Satyanarayana	IBM	X	X	X	X
Harry	Cropper	Intel	X		X	X
Don	Wood	Intel				X
Mitko	Haralanov	Intel			X	
Jess	Robel	Intel	X	X		X
Jim	Ryan	Intel				X
Rupert	Dance	Software Forge	X	X	X	X
James	English	UNH-IOL	X	X	X	X
Glenn	Martin	UNH-IOL		X		X
Edward	Mossman	UNH-IOL	X	X		X
Bob	Noseworthy	UNH-IOL	X			X
Nate	Rubin	UNH-IOL	X	X	X	X

Agenda

2013-2014 Events

Meeting Minutes from 10/22/2013

Interop and Plugfest Events

- **November/December 2013**
  - OFA Interop Logo GA Event — Waiting for developers to resolve **Critical Bugs** – see below
- **April 2014**
  - IBTA PF25 — 03/31 ➔ 04/11/2014
  - OFA Interop Debug Event — 04/07 ➔ 04/11/2014

OFA EWG Status

- OFED 3.5-1
  - OFED 3.5-1 GA
- OFED 3.5-2
  - OFED 3.5-2 RC2
- OFED 3.5-2 **GA**
  - Waiting for developers to resolve **Critical Bugs** – see below
- OFED 3.12
  - Daily Build released 11/11/2013
  - RC in December
- Kernel 3.12
  - 3.12 has been officially released

October 2013 Interop Debug Event - Update from Edward at **UNH-IOL**

1) IB all done except
   a) Fabric Init – everything passed
   b) IPoIB is looking very good in both modes. Minor issue with newer products but resolved.
   c) MPI – full pass on all adapters
   d) NFSoRDMA it is good except for 2 of the newer HCAs
   e) RDMA Interop and Stress all passed.
   f) SRP – one known issue with a new product – may well not be resolved for OFED 3.5-2
      i) Possible to add to list but with a known Issue
   g) uDAPL – success for all three groups.
   h) SM Failover is still pending - **Open**
   i) Link Initialization – still needs to be done - **Open**

2) RoCE - very exciting in RoCE department
   a) Link Init is 100% passing for Emulex, IBM and Mellanox
   b) MPI was very successful
   c) RDMA Interop and Stress – still **Open**
   d) NFSoRDMA – still **Open**
      i) **James**: Some Interop issues causing hangs and he is working with vendors to resolve this.
   e) uDAPL – this passed 100%
   f) RSockets – this passed 100%

3) iWARP
   a) Still having configuration issues with new Chelsio iWARP cards
      i) T5 40 GbE uses QSFP+ whereas the 10 GbE adapters use CX4 and SFP+. They are trying to bridge cap between the two. They are going through a40 GbE switch and then a 4-way breakout cable to the other nodes.
      ii) Cards link to each other but not to the switch and the rest of the cluster.
      iii) Bob says they have borrowed things. Problem is on the 40 GbE to 40 GbE side: the 10 GbE switch can disable the white list and support any transceiver. On the 40 GbE switch, they don't know how to disable the white list. The problem is with both the IBM switch and the Emulex partner switch.
      iv) **Tom** - they are working with IBM and that there is a hardware rev on the RNIC. Please send him that information.
      v) **Bob**: Please share the transceivers and cables that work with this switch and send us some compatible hardware.
      vi) Edward will take lead to contact
   b) MPI – some configuration issues ae preventing the completion of these test. He feels that this is not an OFED issue but something on their end.
   c) RDMA Interop – some issues with new T5 adapters.
      i) Used Dr Russell's tools for the last few events
   d) uDAPL – good results overall except for T5.

Review of the problem with the use of ib_send_bw on iWARP adapters

The Test Plan for iWARP used to use rdma_bw but this tool was deprecated in OFED 3.5. There is a conflict with the use of ib_send_bw in iWARP for the following reasons:

Harry Cropper – from email

This test is streaming RDMA Sends with no pacing/credits which means the receiver can run out of RDMA Receive buffers. In IB, this works because the IB protocol has a credit mechanism. However, the iWARP protocol does not have a credit mechanism. The iWARP designers felt that most applications are doing credit mechanisms anyway and is not a good fit for IP networks.

So what does the iWARP specification say about an incoming RDMA Send when no RDMA Receive buffers are posted? It says two things:

- The connection is torn down.
- An implementation MAY choose to handle the no buffer situation gracefully in a way that does not tear down the connection.

The NE020 does the first and Chelsio does the second. Both behaviors are correct.

So the objections to this test are:

1) The test is acting in a way that is outside the intent of how iWARP applications are intended to work.
2) Both behaviors are valid so whether the test fails or succeeds, it must be considered to be passing.

Don Wood – Intel – from email

The following quotes are from "RDMA Protocol Verbs Specification (Version 1.0)" available at http://www.rdmaconsortium.org/home/draft-hilland-iwarp-verbs-v1.0-RDMAC.pdf.

From Section 8.1.2.1 Send/Receive - At the end of the first paragraph:
"If a WQE is not available on the RQ to describe the Untagged Buffer for the incoming Send Message Type, then the LLP Stream MAY be terminated.
If the LLP Stream is not terminated, the reader should see Section 13.2 - Graceful Receive Overflow Handling for one implementation option."

From Section 13.2

"A valid implementation option is to gracefully handle Receive Queue or Shared-Receive Queue overflow. In a strictly layered model, this may be difficult but in an RNIC implementation, this should be feasible.

In the current architecture, if there are no Receive Queue Work Queue Elements available when an Untagged Message arrives then the connection is dropped. This is true if there is a Shared Receive Queue or a dedicated receive queue.

In this case, the implementation (RI/RNIC), which is not relying on an external LLP, may choose to handle this gracefully through LLP mechanisms. In this case, the RI will choose to not drop the connection and instead appear to pause receive queue processing until more WQEs have been posted to the RQ or S-RQ."

Dr Russell UNH-IOL – from email
I have looked carefully at the code in the latest perftest-2.0 for ib_send_bw (the source file is send_bw.c). Harry is correct that this code does NOT use a credit scheme of any type between sender and receiver, so it is VERY POSSIBLE that the sender will "over run" the receiver (i.e., send a message when a recv has not yet been posted).

The "chapter and verse" cited by Don Wood also confirms that there are 2 possible ways iWARP can deal with this situation and still be considered conforming to the "standard" -- terminate the connection, or try to deal with it by NOT terminating the connection. So if Chelsio does it one way, and Intel does it another, both have to be considered "standard". And the ib_send_bw will definitely provoke this situation.

Harry Cropper – live commentary
The iWARP specs are old but Bob Noseworthy says that the IETF typically does not take RFC to a full standard.

Don Wood – live commentary
1) The issue is that the iWARP designers wanted to be Hi-Performance which means pre-posted buffers and credit exchanges should be done at App level instead of doing in hardware. If there are no buffers then it's an application error. There is an option to not make it terminate. Both are valid options.
2) Ib_send_bw sends non- stop and there is neither SW Credit Flow nor Hardware Credit flow. Intel terminates and Chelsio takes the other spec options and handles the error.
3) **Pradeep**: somewhere in OFED 1.4.1 introduced rdma_bw tools. Mellanox is maintainer. Difference was that rdma_bw invoked RDMA_CM. Now the ib tools have option to invoke RDMA-CM with a flag and so that is why rdma_bw was deprecated after OFED 1.5.4.1.
4) **Rupert** – what is status of ib_read_bw and ib_write_bw
   a) Don will complete his testing and let us know.
5) **Pradeep**
   a) Suggests we look and the –n option for b_send_bw and don't flood the partner. Perhps that will solve the lack of credits issue.
6) **Don Wood**
   a) He suggests that we get the user space app "rdma_bw" and install it with OFED 3.5-2. It should install and work fine.
   b) Intel will not change hardware or driver.
7) **Rupert' suggestions**
   a) He asks that Intel validate that the perfest read and write utilities work on the Intel adapter.
   b) We have several options regarding the send issue
      i) Continue to use Dr Russell's tool which was developed as part of the OFA training materials
         (1) Intel does not like this option because this is not open source code and they cannot see what this tool is doing. They are concerned about IP and would want to know those things before sanctioning the use of that tool. .
      ii) Eliminate the send command for iWARP
      iii) Put the ib_send_bw operation into Beta status for iWARP until the April 2014 event and give more time to resolve the problem.
8) Conclusion
   a) Rupert called for a Vote by acclamation to put the ib_send_bw operation into Beta status for iWARP until the April 2014 Interop Debug event. There were no objections.

Open Issues

1) **Rupert**: Just a reminder to the group that the OFA Debug Events do not generate official reports. UNH-IOL should send reminders to all vendors regarding problems at this event
2) Updates to the OFA-IWG Interoperability Test Plan-v1.49-v6
   a) Add the "-s" flag to both server and client for RDMA Interop
      i) Pradeep has found that without the “-s” flag on the server that the command can fail with large sizes even though it works fine on small sizes without the flag.
   b) Eliminates the special "-t 126" flag that was used for Emulex. Emulex told us that this was no longer needed.

Action Requests (ARs)

Chelsio

11/5/2013

1) Send updated firmware for T5 RNIC.
2) Send UNH-IOL transceivers and cables that will work with the T5 adapter and the IBM switch

Emulex – Jeff

iWARP vendors

10/22/2013

1) Solve bugs in OFED 3.5 and 3.5-2

IBM - Pradeep

Intel Don Woods

11/5/2013

1) Validate that the OFED 3.5-2 perfest utilities ib_read_bw and ib_write_bw work on the Intel adapter.

Software Forge - Rupert Dance

10/22/2013

1) Add the –s option to both client and server in the RDMA Interop tests as per Pradeep's recommendation – done 11/5/2013

2) remove the "-t 126" option for Emulex RCA – done 11/5/2013

Older

1) Update Logo Program to include OFA Policy regarding OFILG Membership and granting of the Logo
2) Check with EWG and XWG about the distribution of OFED Binaries
3) Create Logo with version or change Logo Guidelines – see 4/3/2012 minutes

UNH-IOL

11/5/2013

1) Send all vendors a list of known issues from the October Interop Debug event. This will not be a full blown logo report.
2) Send Chelsio information about problems linking with 40 GbE IBM switch

9/10/2013

1) Nate and James will work with Jeff and Pradeep to resolve issues they have found with NFS0RDMA over RoCE – in process 9/24/2013