OFA EWG Meeting – 11/11/2013

Attendees

First	Last	Company	09/30/13	10/14/13	10/28/13	11/11/13
Steve	Wise	Chelsio	X			
Pradeep	Kankipati	Emulex	X		X	X
Pradeep	Satyanarayana	IBM	X		X	X
Tom	Elken	Intel	X	X	X	X
Tatyana	Nikolova	Intel	X	X	X	X
Robert	Woodruff	Intel	X	X	X	X
Vladimir	Sokolovsky	Mellanox	X		X	X
Bill	Snapko	SGI		X	X	X
Rupert	Dance	Software Forge	X	X	X	X
John	Jolly	SUSE		X		

OFED status

1) OFED 3.5-1GA released
2) OFED 3.5-2 RC2 released
3) OFED 3.12 with RHEL 6.5 – **Vlad** leading
   a) Daily build: http://www.openfabrics.org/downloads/OFED/ofed-3.12-daily/OFED-3.12-20131111-1359.tgz
   b) libocrdma user space and kernel packages have been included in build
      i) Vlad found Rupert's email from 10/31/2013 and he added the package by himself and did not wait for Pradeep. There is a small issue with library and he has field Bug 2454 assigned it to himself. He asks Rupert to setup Pradeep and re-assign the bug to Pradeep to resolve it.
   c) OFED 3.12 supports RHEL 6.5 and he is working on a backport to 6.4
   d) Vlad has not enabled all compilation because he has the following packages ready:
      i) ibcore, IPoIB, mlx4
      ii) mlx5 is included but he has not checked compilation yet.
      iii) Compat, compat-rdma trees have been updated and so we can begin doing backporting.
   e) If you want to install, you can only can select subset
4) OFED 3.13
   a) He will provide updates for this after we have 3.12 stable

OFED Bugs

Bugs 2441 and 2445 - dapltest bugs with Chelsio products

1) **Woody** feels that if they do not affect general operations and Chelsio is OK with it, then we should go ahead and release without fixing them.
2) **Rupert** should check with Chelsio

Bug 2452

1) Memory registration takes a very long time when using Intel MPI. The problem does not exist in 1.5.4.1 but the problem exists on all higher releases of OFED and in Mellanox OFED as well.
2) This is causing a problem on the Intel Endeavor cluster and it is slowing performance.
3) **Vlad** is putting additional resources on it because Jack is sick
4) **Vlad** asks that they try and install with RHEL 6.5 and OFED 3.12. If problem is fixed upstream then it will help Vlad to determine what the problem is. Mellanox OFED and OFED 1.5.4.1 all use different drivers.
5) **Vlad** – does it just happen after running Intel MPI
6) **Woody**
   a) Verify that it happens with dapltest and not just Intel PMI.
   b) He is not sure it will run on 3.12 with just ibcore.
   c) We should not release 3.5-2 until this is resolved.

Bug 2453
1) Stan was trying to run on newer patch updated kernel - kernel 2.6.32-358.23.2
2) Vlad says compat issue and not hard to fix – he can do by tomorrow.
3) Which kernels do we try to support – this is not the default kernel. This could happen again with other updated kernels that can cause backport issues.
4) Tom Elkin – if easy to fix we should do it. But if it is difficult, then we might choose to delay.
   a) If OFED is already GA, then we point to the next release.

Bugs 2419-2423 - ib_read_bw and ib_write_bw issues
1) Vlad – Gil is new to this problem but he is working on it.

Bug 2449 – NFSoRDMA on PPC
1) Jeff Becker and Pradeep are working on this but they need access to OFA Cluster so they can test against PPC and RoCE adapters
2) Rupert will arrange

Mellanox: proposed solution to ibutils2 – binary
Rupert explains that Mellanox does not want to include the source in OFED because there is proprietary code. However without this we have no way to use ibdiagnet to check for FDR links. But Mellanox is willing to provide the binary for ibdiagnet.
Woody
1) Recommending 3rd party binaries is not customary but Woody agrees that we should make the binary available.
2) **AR**: Follow up with Ira and see if there are other alternatives

Bug 2454 **-** Pradeep Kankipati - Emulex
1) Send kernel and user space drivers to Vlad and "cc" the EWG
2) There is a problem with the packages that were submitted and they do not compile.

Opens

Problem with ib_send_bw and Intel iWARP adapters – Tatyana
1) She is uncomfortable doing Logo Testing with Dr Russell's tool because it is not open source.
2) Woody – we should not use something in Interop that you do not have source for.
3) Tatyana – suggestions and comments
   a) Put ib_send_bw into Beta
   b) She is testing ib_read_bw and ib_write_bw
   c) Different adapters seem to cause problems – T4
   d) She can debug in my own environment. She wants these OFED tools to work.

History of the ib_send_bw problem – Notes from the IWG meeting
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.

ARs

Emulex – Pradeep Kankipati

11/11/2013

1) Resolve the compilation issue in Bug 2454

10/28/2013

2) Send kernel and user space drivers to Vlad and "cc" the EWG – done - Vlad pulled this on 11/11/2013

Rupert

11/11/2013

1) Add Pradeep as maintainer – done 11/11/13
2) Check with Chelsio to see when they are going to work on Bugs 2441 and 2445 – done 11/21/2013
3) Provide access to OFA Cluster an the iWARP and RoCE hardware for those EWG members who need it to resolve bugs – done 11/13/2013
4) Follow up with Ira and see if there are other alternatives to ibdiagnet and ibutils2

10/28/2013

1) Send document to Emulex describing the procedure for vendors to create drivers and packages for their products an upload to the OFA Download site – done 10/31/2013
2) Update release notes to explain the uninstall procedure and the problem with packages installed by Distros – see Bug 2403

10/14/2013

1) Add a note indicating a known limitation in the ibutils package that it does not support FDR, FDR 10 or EDR speeds when using ibdiagnet.

8/19/2013

1) Update the Bugzilla list of Maintainers using OFA Maintainers list because it is more accurate

Steve Wise

9/30/2013

1) Steve Wise agreed to review bugs 2402 and 2445 and try to get them resolved – done 11/21/2013 - Chelsio has indicated that they do not have the resources to fix these now and will target OFED 3.13

Tatyana

10/14/2013

1) Review Bug 2441 and other iWARP bugs – done 10/31/2013 – this is a Chelsio bug

UNH-IOL

10/28/2013

1) Verify suggested fix to Bug 2450 – done 11/12/2013
   a) The problem still exists but Sean Hefty has found a solution

Vlad

10/28/2013

1) Move Bug 2416 to later and put in a target release date

9/30/2013

1) Review all of the bugs assigned to Mellanox developers with his team – done 11/14/2013

Woody

11/11/2013

1) Verify that the Memory Registration bug (2452) happens with dapltest and not just Intel PMI.

9/30/2013

1) Work with Arlin regarding the status of the bugs in OFED 3.5 and 3.5-x