OFA IWG meeting – 3/12/2013

Attendees

First	Last	Company	01/29/13	02/12/13	02/26/13	03/12/13
Samantha	Jian-Pielak	Canonical	X	X		X
Tom	Reu	Chelsio		X	X	X
Martin	Schlining	DataDirect Networks	X	X	X	X
Alex	Nicholson	Emulex		X	X	X
Brad	Benton	IBM		X	X	X
Pradeep	Satyanarayana	IBM		X	X	X
Harry	Cropper	Intel	X	X	X	X
Mitko	Haralanov	Intel	X	X	X	X
Jess	Robel	Intel	X	X	X	X
Guy	Ergas	Mellanox			X	
Yevgeny	Petrilin	Mellanox				
Rupert	Dance	Software Forge	X	X	X	X
Glenn	Martin	UNH-IOL	X			X
Edward	Mossman	UNH-IOL	X	X	X	X
Bob	Noseworthy	UNH-IOL	X	X		
Nate	Rubin	UNH-IOL	X	X	X	

2013 Events

April 2013 dates

- IBTA **PF23** 04/01 ➔ 04/12/2013
- OFA April Interop Debug Event 04/08 ➔ 04/12/2013

October 2013

- IBTA **PF24** 10/07 ➔ 10/18/2013
- OFA Interop Debug Event 10/14 ➔ 10/18/2013

OFA Conferences – Monterey California

- OFA User Conference 04/18 ➔ 04/19/2013
- OFA Developer Workshop 04/21 ➔ 04/23/2013
- **Early Bird Deadline** **Extended**: 03/29/2013

IBTA Plugfest 22 results are available here:
http://www.infinibandta.org/content/pages.php?pg=integrators_list_overview

OFA EWG Status

- OFED 3.5 GA Released 02/14/2013
- OFED 3.5.1 Planned Support for RHEL 6.4
- CentOS Current Distribution 6.4
- RHEL Current Distribution 6.4
- Scientific Linux Current Distribution 6.3

Next OFED discussion

1) OFED 3.5.1will support RHEL 6.4
2) The EWG is proposing to do one or more Technology Previews based on OFED 3.5.1. One of these will be from Intel and will provide XEON Phi support. Another could include the Emulex and IBM drivers.
3) OFED 3.5 does include support for RSockets

RSockets – Sean Hefty

1) RSockets is a user space implementation, so there's nothing that needs to be pushed upstream. An initial release of RSockets was included in OFED 3.5 as part of the librdmacm 1.0.16 release. The 1.0.17 release extended RSockets capabilities with more features.
2) There should be an rsocket.7 man page installed as part of the librdmacm, along with a sample application (rstream) and a socket preload library (librspreload.so). The preload library is installed in %lib%/rdma by default to avoid potential compatibility issues.
3) RSockets should run over InfiniBand or RoCE. iWARP support is not yet available.

OFA Cluster Cable Needs

1) The OFA Cluster is in dire need of new QSFP Cables that have been listed on the IBTA Integrators' List for PF22. Here is a list of devices in the cluster that need to have FDR capable cables
   a) 4 FDR HCA to be tested
   b) 2 FDR Switches
   c) 1 FDR SRP
   d) Many FDR ➔ QDR links
2) There are also many other QDR links that need to be upgraded with the latest PF22 cables.
3) Here is the list of cables needed to fully upgrade the cluster
   a) **FDR** cables needed:
      i) Quantity: 35
      ii) Length: 3-5 meters
      iii) Type: Copper or AOC
   b) **QDR** cables needed:
      i) Quantity: 25
      ii) Length: 3-5 meters
      iii) Type: Copper or AOC

Event updates – Interop GA Event – IB – Edward

1) Cable status – SWF loaned PF22 cables which enabled Logo GA Testing
2) Link Init – this is done and most of the issues discovered during the October Interop Debug event have been resolved.
3) Fabric Init – this is done and Edward believes everything passed
4) SRP done – results will be comparable to what we saw during the Interop Debug event. Not all products support SRP.
5) RDMA Part B is complete – this is the stress portion
6) IpoIB in both connected and datagram mode has been completed.
7) Testing that has not been completed
   a) uDAPL
   b) SM
   c) MPI
8) Test that have been deprecated or no longer run because of new OFA Logo Policy
   a) NFSoRDMA

b) RDS – this has not been back ported to 2.x and OFED 3.5 only supports 3.x kernels

9) Overall results are considerably better than what they saw in the October Debug

Event updates – Interop GA Event – iWARP

1) UNH-IOL will start testing later this week

April Interop Debug

1) UNH-**IOL r**eceived IBM devices - 2U server arrived but they are still waiting for the RDMA Channel Adapters.
2) UNH-IOL is also waiting for firmware updates for April from Mellanox.
3) Emulex Switch Partner
   a) One of their partners is willing to have their unannounced switch being a part of the UNH testing but there can be no performance testing (we don't do this as a matter of principle in the OFILP) and the results must be kept under NDA.

Suggested Modifications for the RoCE section of the test plan – Brad Benton

1) **Page 19: RDMA stress**: I think that we need to simultaneously stress RoCE/IB traffic and IP level Ethernet traffic. This will ensure that the Ethernet and IB/RDMA portions of the hardware, drivers & libraries work properly together. The IBM Adapter is both an Ethernet NIC and an RNIC and they want to see this functionality tested simultaneously
   a) Brad says that uperf works well to generate Ethernet traffic.
   b) IB can use dapltest or other utilities. MPI would be great but there are issues with little and big endian this is a stretch too far for April 2013.
   c) He would like to see us exercise a rough test plan algorithm and then refine it during Interop Debug event
2) **Page 24, Table 21**: Brad wants to understand how IPoCE is different from just standard IP over the RoCE CA? Or is this referring specifically to IP traffic over a CEE fabric?
   a) There is no IPoIB with RoCE. This is standard IP traffic over the CE interface.
   b) The group agrees that this part of the test spec needs fleshing out.
3) **Page 26, 1.12** – just want to note that they to do x86_64 and ppc64 interoperability.
   a) We need to add and allow mixed system architectures for x86_64 and ppc64 interoperability. This also implies mixed endianness between those systems.
   b) It works at the wire level but they want to test handshakes during RDMA-CM
   c) They expects problems in certain areas
4) **Section 9.2,** Operating System –
   a) The IBM implementation cannot use CentOS or SL – there is no support for the IBM RCA in these Distros. So they will probably use RHEL 6.2 running with ConnectIB. We will still use SL 6.3 on the rest of the cluster
   b) For the IBM Power system, RHEL6.x will have to be used. There are no CentOS, Scientific Linux or Ubuntu distributions for Power platforms.
   c) OFED Version updates
5) **Section 13.1.3**, TI testing (iSER)
   a) At this point IBM does not care if it is included in the Interop Testing for April 2013. They agree that we should wait for demand
6) **General Enhancements**
   a) RDMA_CM
   b) IBM would like to see explicit rdma_cm tests, particularly for processor-heterogeneous (x86_64/ppc64) setups.
      i) The perf tests can do this when setting up queue pairs.

ii) Look for any other rdma_cm scenarios that will test this.

c) For heterogeneous testing, it would be good to use tests that are network byte order aware and the results of data transfers are subsequently validated. He wants to enhance existing perf test to validate that data received was data sent. This could be in uDAPL
   i) Mitko says that if network byte order is wrong you won't get to data validation
   ii) For heterogeneous testing, it would be good to use tests that are network byte order aware and the results of data transfers are subsequently validated.

d) They would like to see us test RSockets over RoCE
   i) This should cover both IB and RoCE

e) Test IPv6

f) Test bonding over RoCE Ethernet interfaces
   i) General Bonding driver – wants to make sure Link Aggregation works. If they have two devices, they would like to test fail over. Bonding from IB is limited and they want to make sure the basic RDMA_CM tests work.
   ii) **Pradeep** – wants to make sure that if a bonded link fails then the other link takes over. This is a not APM.
   iii) Brad agrees with Rupert that currently there is no APM from one HCA to another. People want RoCE to support link aggregation and failover with multiple Channel Adapters.

7) Brad says that his priority is based on the order he presented.
   a) Pradeep feels that IPv6 is very important and should be above bonding

Future Considerations

1) RoCE & VLANs. In the Mellanox implementation, a VLAN ID is stored as the 12th and 13th bytes of the GID. Is that compatible across vendors?
   a) Brad asks how do we manage RoCE and VLAN
   b) Alex feels this may not be at the test site.
2) Validating Priority Flow Control (802.1Qbb)
3) Routable RoCE is important.

Action Requests (ARs)

Rupert Dance

2/26/2013

1) Secure 50-60 new cables to get the OFA Cluster up and ready for GA testing – done 2/28/2013
2) OpenFabrics Interoperability Logo Program Changes – in process 3/24/2013
3) Contact Sean Hefty about RSockets and SDP – done 3/25/2013

1/29/2013

1) Check with Mellanox regarding the state of their firmware updates and the readiness to do GA testing
   a) Received firmware updates on 2/18/2013
2) Discuss with the EWG and determine what are next OFED release – done 3/4/2013

12/04/2012

1) Log SRP bug with RedHat

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
4) Incorporate ICR into Logo Program

UNH-IOL

2/26/2013

Edward: OFA April Interop Debug Event - need updated link on UNH-IOL website – done 2/27/2013

2/14/2013

Bob Noseworthy: do you have contacts for IBM within DCB. Bob will check with Blade Networks as well.

10/9/2012

1) **Nate**: verify that rping works for RoCE as well as iWARP
2) IOL Check the OpenMPI commands for RoCE