Multi-Path RDMA

Elad Raz
Mellanox Technologies
Agenda

• Motivation

• Introducing Multi-Path RDMA

• Design

• Status and initial results

• Next steps

• Conclusions
MP-RDMA Motivation (1)

- Failovers and High Availability Support
- Bandwidth Aggregation
- L3 datacenter support
MP-RDMA Motivation (2)

- Transparent migration
What is Multi-Path RDMA?

RDMA HAL and services

Verbs

MP-RDMA

ib_dev1  roce_dev2

Session Level Connection

HW 1  HW 2

Router A  Router B

Verbs

RDMA HAL and services

MP-RDMA

ib_dev1  roce_dev2

Session Level Connection

HW 1  HW 2
MP-RDMA Operation

CM protocol (primary flow) + MP-capability exchange

CM protocol (sub-flow)

March 15 – 18, 2015 #OFADevWorkshop
Connection Migration

Virtual consumer

Send Queue

Physical producer

virtual producer

Receive Queue

Virtual consumer

Physical producer

virtual producer

March 15 – 18, 2015
#OFADevWorkshop
## MP-RDMA Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>Automatic Path Migration</th>
<th>RoCE-LAG</th>
<th>MP-RDMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Port failover</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Bandwidth aggregation</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Application agnostic</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>L3 session</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Multi-device failover</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Migration support</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>
## MP-RDMA and MP-TCP

<table>
<thead>
<tr>
<th></th>
<th><strong>MP-TCP</strong></th>
<th><strong>MP-RDMA</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>MP Messages</td>
<td>TCP options</td>
<td>CMA MADs (private data)</td>
</tr>
<tr>
<td>Address management</td>
<td>Add/remove addresses</td>
<td>Add/remove CMA address</td>
</tr>
<tr>
<td>Flow management</td>
<td>Add/remove TCP sub-flows</td>
<td>Establish/migrate/teardown QPs</td>
</tr>
<tr>
<td>Communication endpoint</td>
<td>TCP socket</td>
<td>MP RDMA device</td>
</tr>
<tr>
<td>Data sequencing</td>
<td>Byte-stream divided between sub-flows Flow + session based sequencing</td>
<td>QP and actual HW WQEs (performance)</td>
</tr>
<tr>
<td>Sub-flow address combinations</td>
<td>Any IP interface to any peer IP interface, subject to middle-boxes (e.g., firewalls, NAT)</td>
<td>Any RDMA addressing to any peer RDMA addressing, subject to the same Technology (IB, RoCE, iWARP)</td>
</tr>
</tbody>
</table>
MP-RDMA Design

• User/kernel mp-rdma driver
  – Device instance hosting MP-capable resources
  – Implements resource virtualization and connection failover
  – Uses underlying physical devices transparently

• RDMACM/CMA support
  – MP capability negotiation
Policies

• Active backup

• Load Balancing

• Efficiency
Virtual Namespaces

Application 1

MP-RDMA (1)

vPD1
vMR1
vQP1
vQP2
vQP3
vQP4
vCQ1
vCQ2
vCQ3
vSRQ1

RDMA Dev 0

MR16
QP82
QP61
SRQ13
CQ42

RDMA Dev 1

MR13
QP33
QP98
CQ22
CQ23
Resource Creation

Virtual resources
- vQP1

Policy selection
- RDMA Dev 0: port 1
- PD12

Resource creation
- QP42
- CQ2

Post-send (MR)
- MR 2
Data Path

- Translate MRs:
  - `ibv_post_send`
  - `ibv_post_recv`

- Translate QPs:
  - `ibv_poll_cq`

- Monitor WQs:
  - PSN
  - Completed
Status and Initial Results

- User-space driver progressing nicely
  - Resource management
  - Connection management
  - Failover
  - Data path for RC send-receive operations

- Encouraging initial results

**Latency vs. Message size (uSec)**

**Bandwidth vs. Message Size**

**Lower is better**

**Higher is better**
Next Steps

- Kernel MP-RDMA driver and connection model
- Dynamic device removal notifications
- RDMA and Atomic support
- Datagram and Multicast support
- Consider future standardization
  - IBTA CM extensions
  - RFC
- Open-source the code
Conclusions

• MP-RDMA solves multiple requirements
  – Multi-devices failovers
  – Transparent BW aggregation
  – Transparent RDMA migration
  – Multi-homed hosts

• Modeled over MP-TCP

• Promising initial results
Thank You