



OPENFABRICS
ALLIANCE

13th ANNUAL WORKSHOP 2017

THE LINUX SOFTROCE DRIVER

Liran Liss

Mellanox Technologies

March, 2017

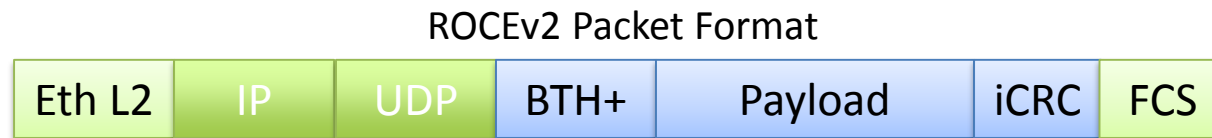
AGENDA

- **Introduction**
- **The RXE driver**
- **Development status and availability**
- **Configuring and using SoftRoCE**
- **Future work**

INTRODUCTION

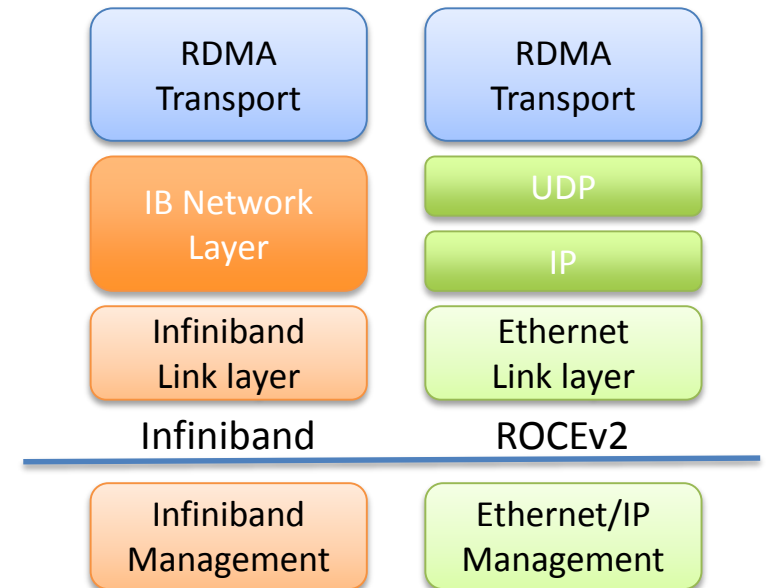
▪ SoftRoCE is a software implementation of the IBTA RoCEv2 specification

- RDMA transport services over Ethernet network



▪ Use cases

- Development
- Testing
- Asymmetric deployments
 - E.g., software clients connecting to hardware high-performance servers



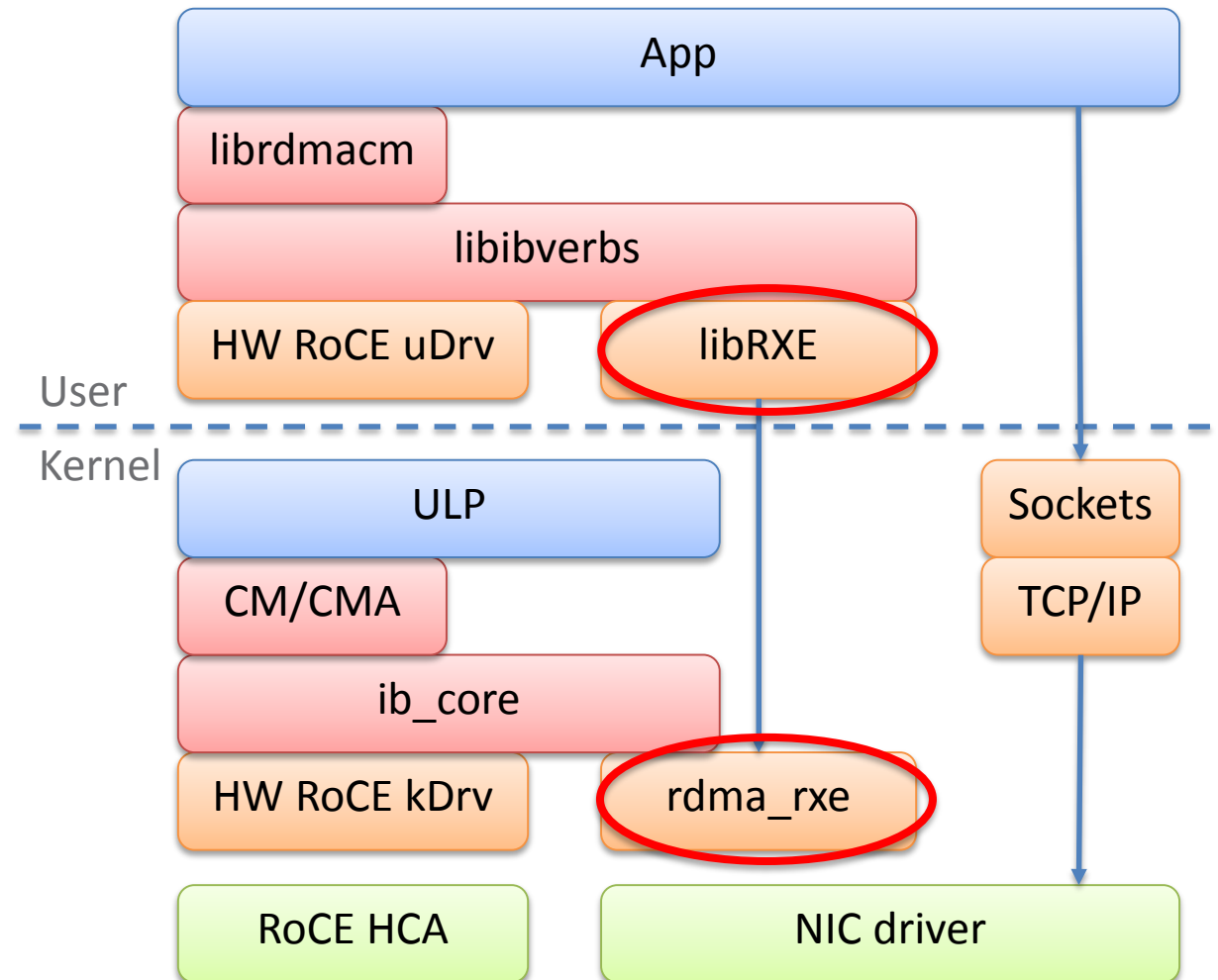
RXE DRIVER

■ Standard RDMA provider

- Kernel rdma_rxe module
- User-space librxe library

■ Kernel transport

- Supports both kernel ULPs and applications
- Asynchronous progress



RXE DRIVER (CONT.)

■ Kernel-allocated / user-mapped queues

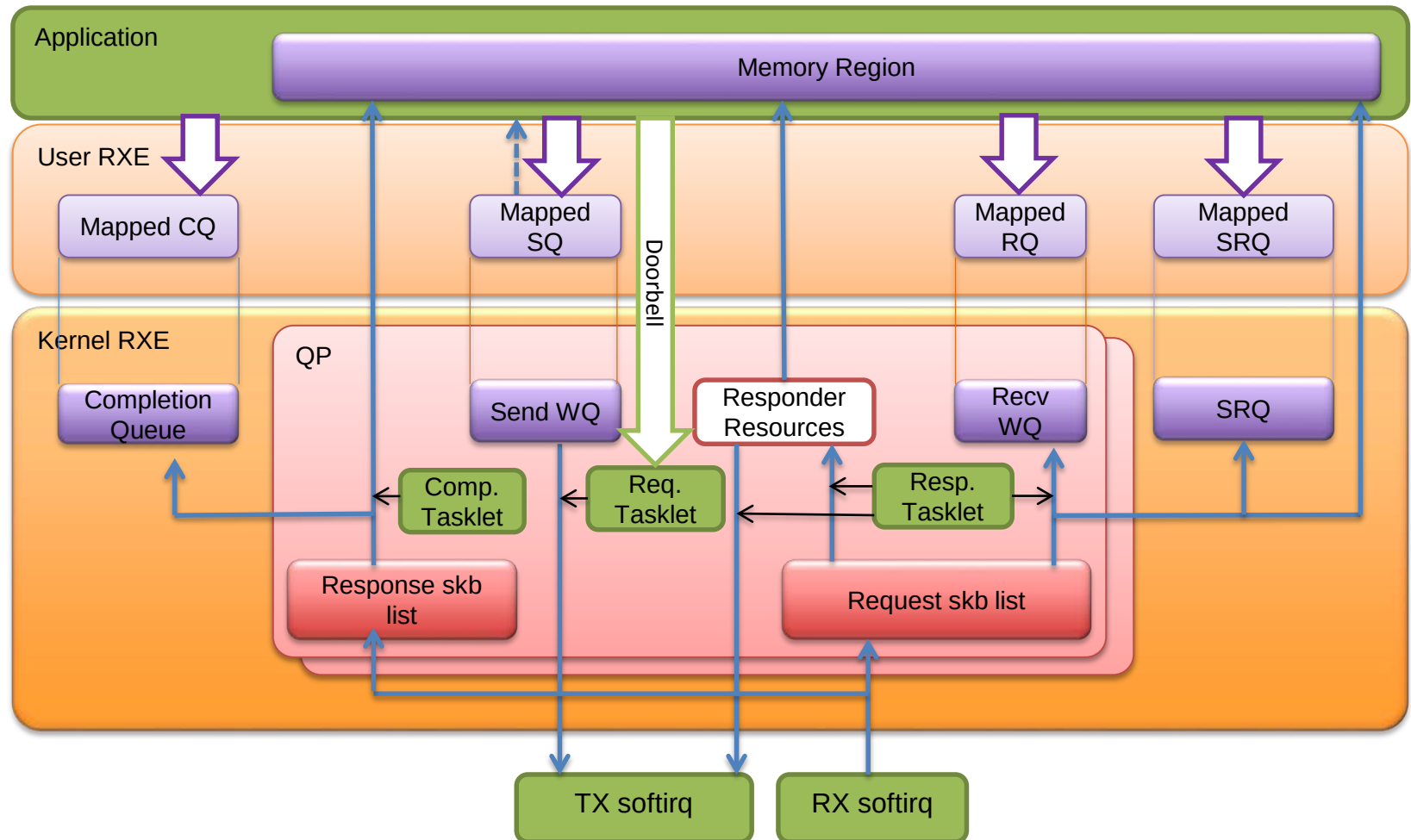
- User-space prepares WQEs
- System calls used only for send doorbells
- No system calls for
 - Polling completions
 - Posting receive WQEs

■ Pinned memory regions

- Data copied in/out of buffers from any context

■ Per-QP tasklets

- Requester
- Responder
- Completer



DEVELOPMENT

- **Initial driver developed by System Fabrics Works**
- **Restarted as Open-Source Github project in 2014**
- **Driver submitted upstream and accepted in 2016**
 - Available from Kernel 4.8
 - config RDMA_RXE
- **Library included in rdma-core repo**
- **Bug fixes**
 - Logic and Semantics
 - Support 0-byte operations
 - Update work queue state before generating completion
 - Response PSN should increase monotonically
 - Correctly handle erroneous WQEs
 - Correctly handle duplicate atomic requests
 - Memory leaks and reference counts
 - Bad socket ref-count
 - QP ref-count while running tasklets
 - Freeing packets in some error flows
 - Races
 - QP state update vs. sending out skbs
 - Packet processing vs. QP teardown
- **Performance improvements (ongoing)**

STATUS

■ Features

- UD, RC, and UC transports
- SRQ
- Atomics support
- Multicast
- Fast registration, local/remote invalidate

■ Test configurations

- Bare-metal NIC
- SRIOV VF

■ ULPs and applications

- Low-level Verbs testing
- RDMACM tests
- Perftest
- iSER initiator + target

■ Interoperability

- Validated successfully with HW RoCE
 - ConnectX-3
 - ConnectX-4

AVAILABILITY

- **Linux distributions (planned)**

- Redhat / CentOS 7.4
- Ubuntu 17.10

- **Mellanox OFED**

- Included in MOFED 4.0
- Based on Kernel 4.9
- Backports for Kernel 4.8
 - May be installed on Ubuntu 16.10
- Forward-ports for Kernel 4.10
 - May be installed on Ubuntu 17.04

CONFIGURATION

▪ Once installed, RXE is managed via 'rxe_cfg' tool

- Start / stop RXE driver stack
- Add / remove / show configured devices
- Configured interfaces are persisted at /var/rxe/rxe

NAME

rxe_cfg - rxe configuration tool for RXE (Soft RoCE)

SYNOPSIS

```
rxe_cfg [status]
rxe_cfg start [-p proto]
rxe_cfg stop
rxe_cfg persistent
rxe_cfg add [-n] ethN
rxe_cfg remove [-n] ethN|rxen
rxe_cfg crc enable|disable
rxe_cfg mtu [-f] [rxen] mtu_size
```

EXAMPLE

```
dev-l-vrt-097:/kernel=> sudo rxe_cfg start
Name      Link  Driver      Speed  NMTU  IPv4_addr  RDEV  RMTU
eno2       yes   igb
enp17s0    yes   mlx5_core
eth0       no    mlx4_en

dev-l-vrt-097:/kernel=> #sudo rxe_cfg add enp17s0

dev-l-vrt-097:/kernel=> #sudo rxe_cfg status

dev-l-vrt-097:/kernel=> sudo rxe_cfg start
Name      Link  Driver      Speed  NMTU  IPv4_addr  RDEV  RMTU
eno2       yes   igb
enp17s0    yes   mlx5_core          rxe0  1024  (3)
eth0       no    mlx4_en
```

EXAMPLE (CONT.)

```
dev-l-vrt-097:/kernel=> ip addr show dev enp17s0
8: enp17s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP qlen 1000
    link/ether 7c:fe:90:75:3e:20 brd ff:ff:ff:ff:ff:ff
    inet 11.134.97.1/16 brd 11.134.255.255 scope global enp17s0
    ...

dev-l-vrt-097:/kernel=> ibv_devinfo -v -d rxe0

hca_id: rxe0
  transport:                InfiniBand (0)
  fw_ver:                   0.0.0
  node_guid:                 7efe:90ff:fe75:3e20
  sys_image_guid:            0000:0000:0000:0000
  vendor_id:                 0x0000
  vendor_part_id:            0
  hw_ver:                   0x0
  phys_port_cnt:             1
  ...
  max_wq_type_rq:           0
    port: 1
      state:                 PORT_ACTIVE (4)
      max_mtu:               4096 (5)
      active_mtu:            1024 (3)
      ...
      phys_state:            LINK_UP (5)
      GID[ 0]:               fe80:0000:0000:0000:7efe:90ff:fe75:3e20
      GID[ 1]:               0000:0000:0000:0000:0000:ffff:0b86:6101
```


IT'S JUST ROCE...

No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	11.134.146.5	11.134.145.5	RROCE	570	RC Send only QP=0x0000ea
2	0.000275	11.134.145.5	11.134.146.5	RROCE	62	RC Acknowledge QP=0x000011
3	0.000307	11.134.145.5	11.134.146.5	RROCE	570	RC Send only QP=0x000011
4	0.000345	11.134.146.5	11.134.145.5	RROCE	62	RC Acknowledge QP=0x0000ea

```
Frame 1: 570 bytes on wire (4560 bits), 570 bytes captured (4560 bits) on interface 0
Ethernet II, Src: Mellanox_ae:86:f9 (e4:1d:2d:ae:86:f9), Dst: Mellanox_ae:87:0c (e4:1d:2d:ae:87:0c)
Internet Protocol Version 4, Src: 11.134.146.5, Dst: 11.134.145.5
  0100 .... = Version: 4
  .... 0101 = Header Length: 20 bytes
  Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
  Total Length: 556
  Identification: 0xdc25 (56357)
  Flags: 0x02
  Fragment offset: 0
  Time to live: 64
  Protocol: UDP
  Header checksum: 0x0000 (none)
  Source: 11.134.146.5
  Destination: 11.134.145.5
  [Source Geo: ...]
  [Destination Geo: ...]
User Datagram Protocol, Src Port: 5555, Dst Port: 5555
  Length: 536
  Checksum: 0x0000 (none)
  [Stream index: 0]
InfiniBand
  Base Transport Header
    Opcode: 4
    0... .... = Solicited Event: False
    .0.. .... = MigReq: False
    ..00 .... = Pad Count: 0
    .... 0000 = Header Version: 0
    Partition Key: 65535
    Reserved (8 bits): 0
    Destination Queue Pair: 0x0000ea
    1... .... = Acknowledge Request: True
    .000 0000 = Reserved (7 bits): 0
    Packet Sequence Number: 12648430
    Invariant CRC: 0x33562b44
Data (512 bytes)
```

dev-l-vrt-146-005:/kernel=> ibv_rc_pingpong -d rxe0 -g 1 -n 1 -s 512 dev-l-vrt-145-005
local address: LID 0x0000, QPN 0x000011, PSN 0xc0ffee, GID ::ffff:11.134.146.5
remote address: LID 0x0000, QPN 0x0000ea, PSN 0xc0ffee, GID ::ffff:11.134.145.5
1024 bytes in 0.00 seconds = 26.60 Mbit/sec
1 iters in 0.00 seconds = 308.00 usec/iter

FUTURE WORK

Features

- **RDMA features**

- Memory windows
- On-demand-paging
- SRQ resizing
- XRC

- **Congestion control**

- Flow fairness
- Lossy networks

- **Container support**

- Should come for free once RoCE network namespace support is completed

FUTURE WORK

Optimizations

■ 0-copy for *all* transmitted packets

- Leverage the fact that MRs are pinned
- Applies to
 - Send / RDMA-W packets
 - RDMA-R response packets
- For UD, generate completions when skbs are destroyed
 - Rather than when enqueueing copied skbs
- Note: still need to go over all the data for CRC calculation

■ Efficient 1-copy for *all* received packets

- Remove ***completely*** any skb re-queuing
- Scatter data to MR directly from NAPI context
 - Combine with CRC calculation

FUTURE WORK

Optimizations

- **Per-CPU execution context**
 - Replace existing per-QP tasklets for sender/responder/completer tasks
- **Transmit packet batching**
 - Output a batch of skbs as a list
 - Results in skb->xmit_more in netdev
- **Cache route for connected QPs**
 - Eliminates FIB lookup on each packet
 - Similar to TCP sockets
- **Use synchronous hash crypto driver for CRC calculations**
 - E.g., use crc32-pclmul driver on x86
- **Select different UDP source ports for each QP**
 - Utilize all NIC transmission queues
 - Leverage ECMP in the network



OPENFABRICS
ALLIANCE

13th ANNUAL WORKSHOP 2017

THANK YOU

Liran Liss

Mellanox Technologies