OPENFABRICS ALLIANCE

Multi-Path RDMA

ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP

Elad Raz

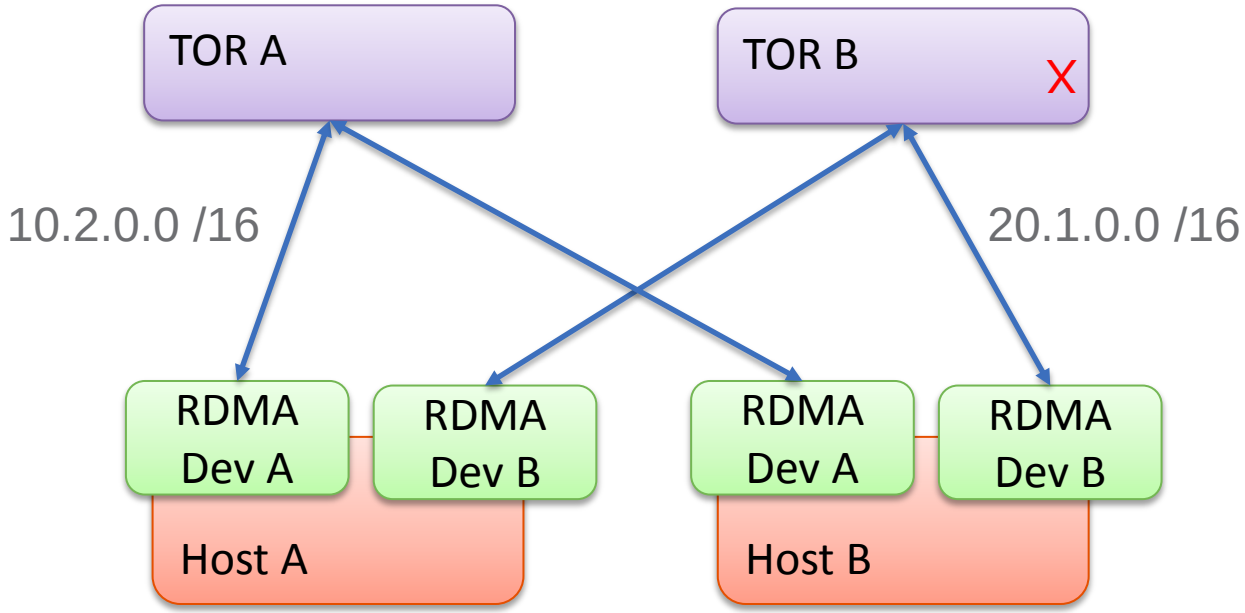
Mellanox Technologies

Agenda

- Motivation
- Introducing Multi-Path RDMA
- Design
- Status and initial results
- Next steps
- Conclusions

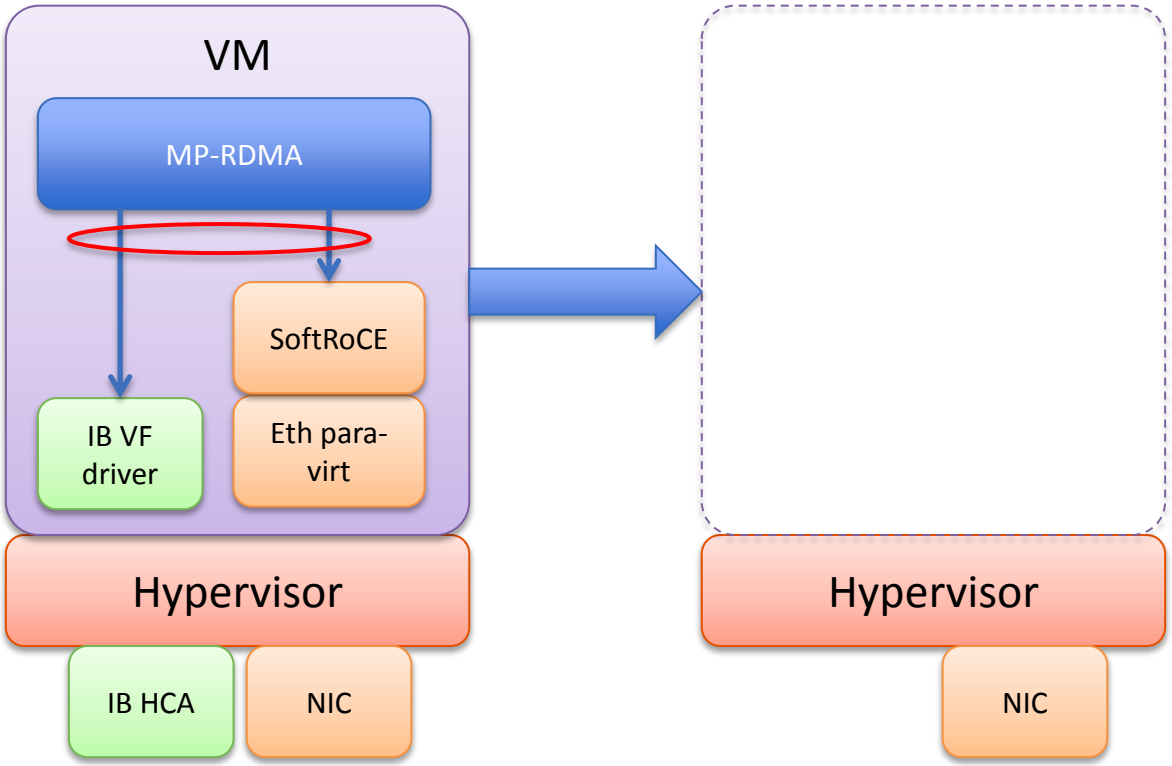
MP-RDMA Motivation (1)

- Failovers and High Availability Support
- Bandwidth Aggregation
- L3 datacenter support

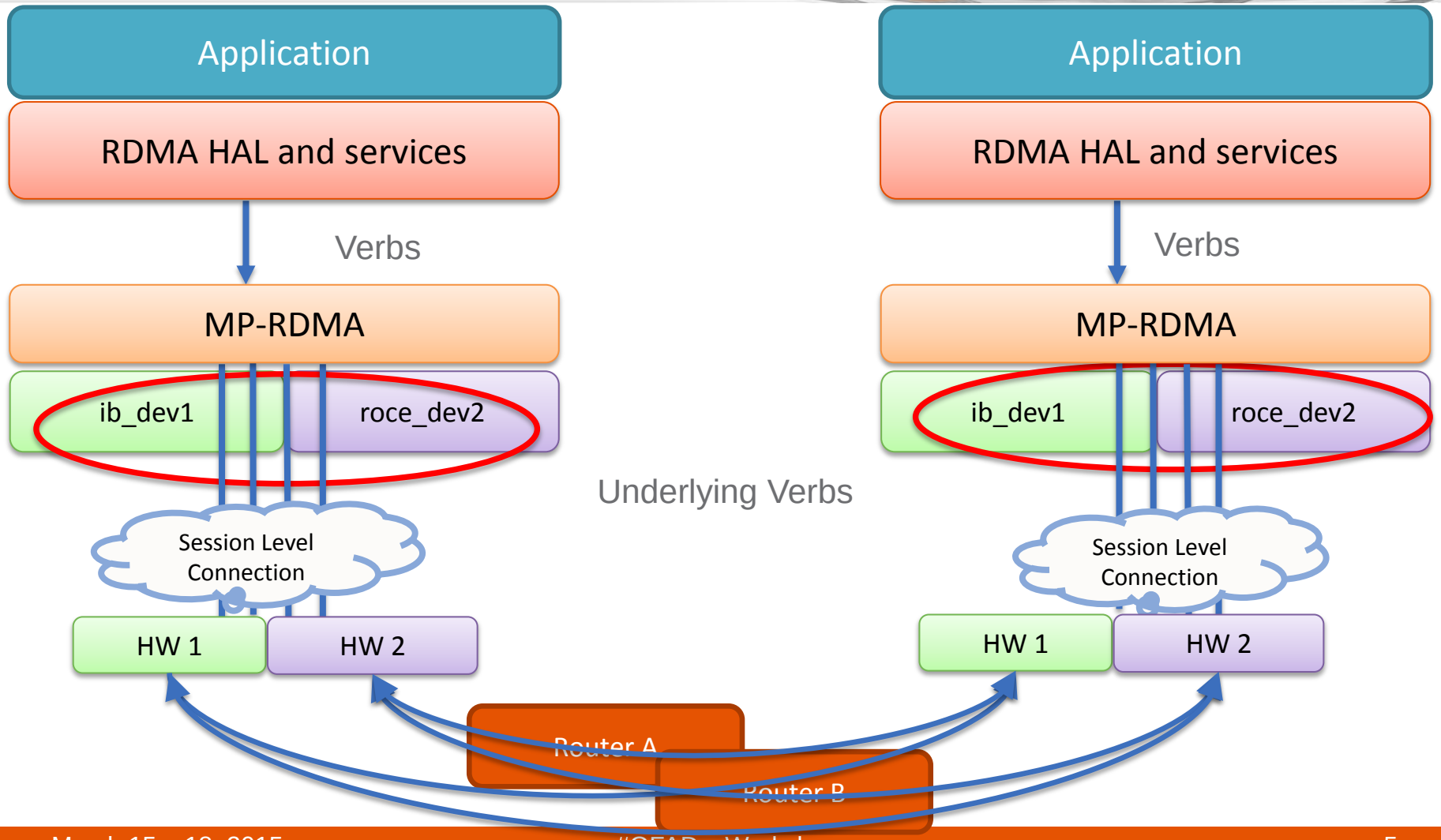


MP-RDMA Motivation (2)

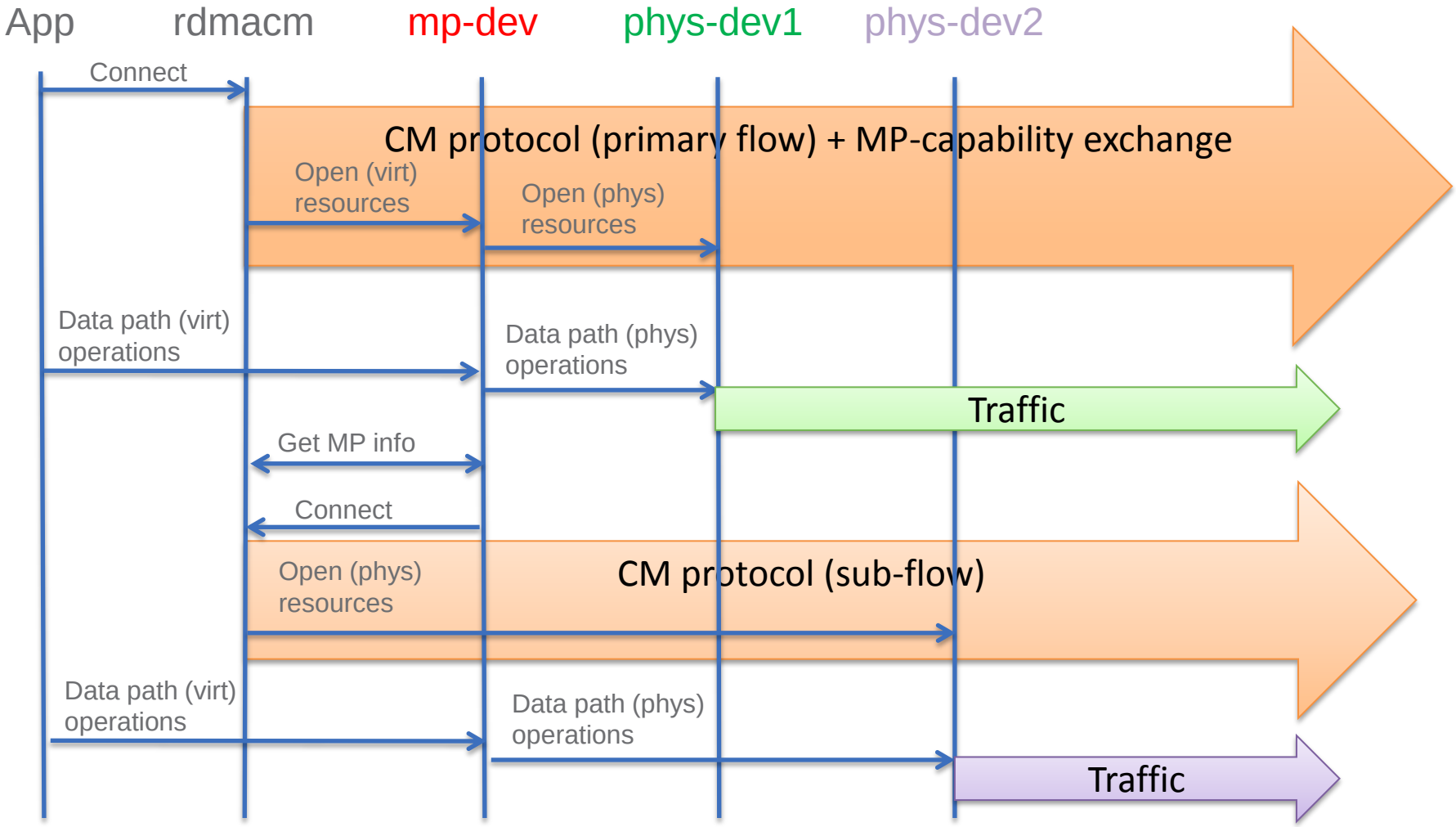
- Transparent migration



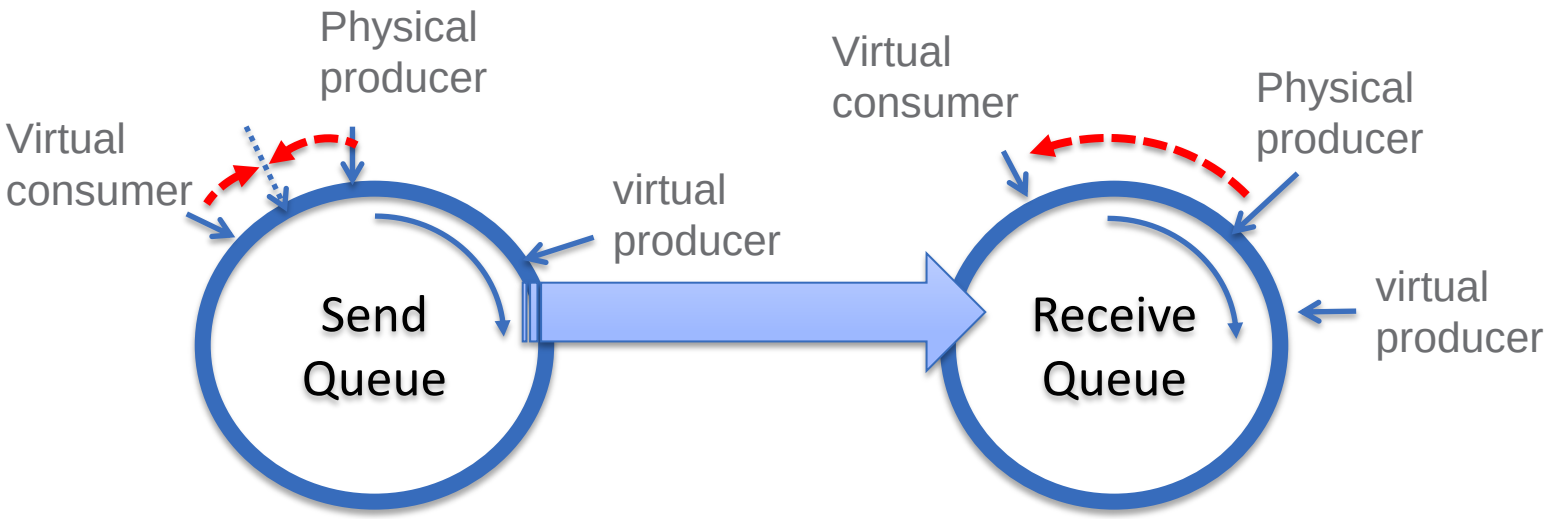
What is Multi-Path RDMA?



MP-RDMA Operation



Connection Migration



MP-RDMA Comparison

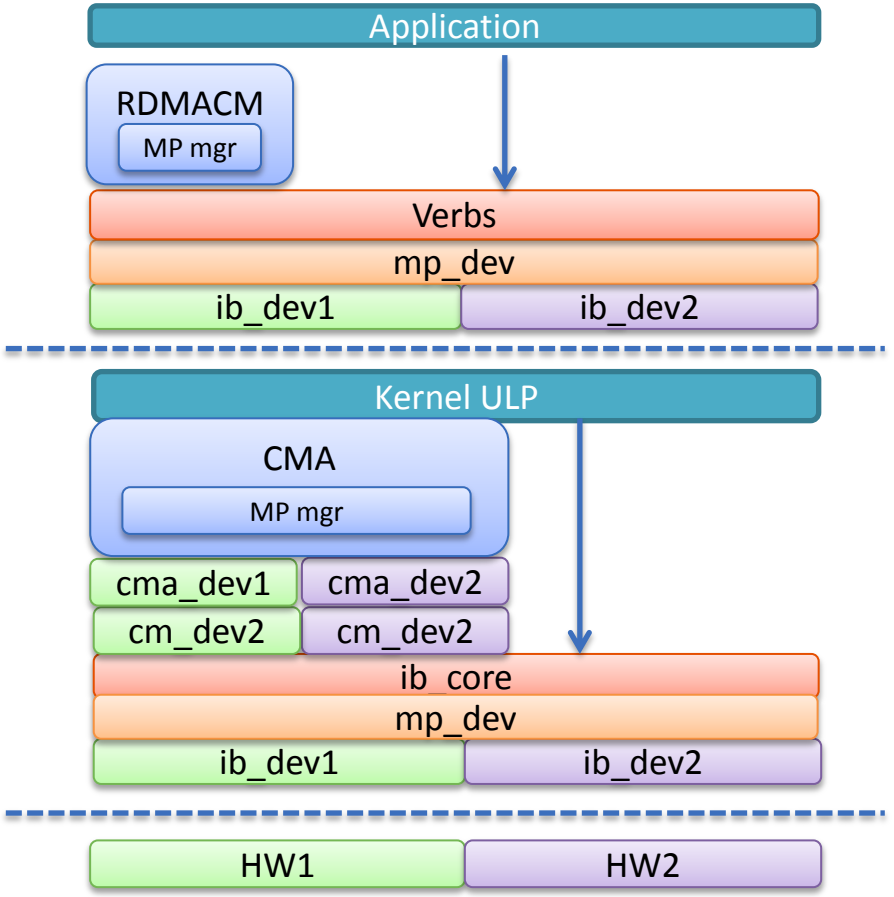
	Automatic Path Migration	RoCE-LAG	MP-RDMA
Multi-Port failover	✔	✔	✔
Bandwidth aggregation	✔	✔	✔
Application agnostic		✔	✔
L3 session			✔
Multi-device failover			✔
Migration support			✔

MP-RDMA and MP-TCP

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

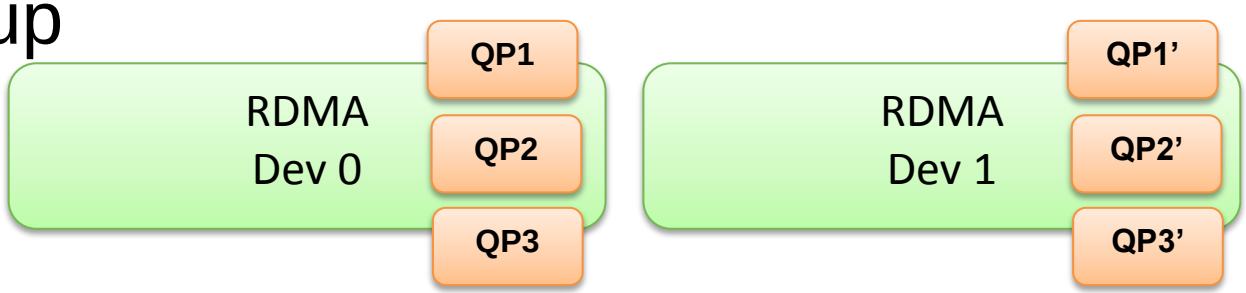
MP-RDMA Design

- 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

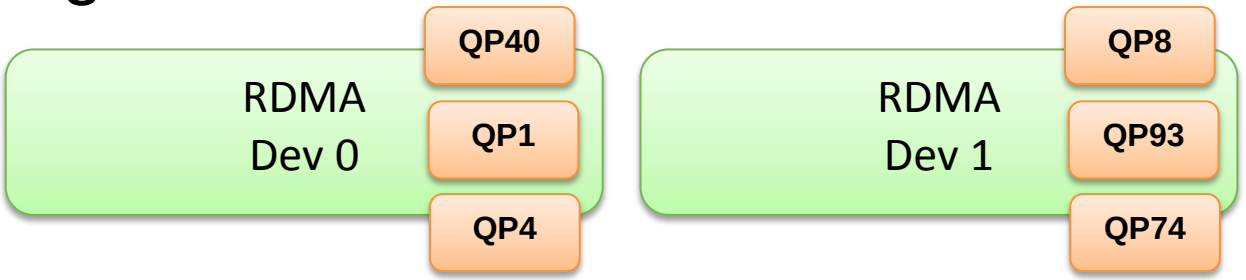


Policies

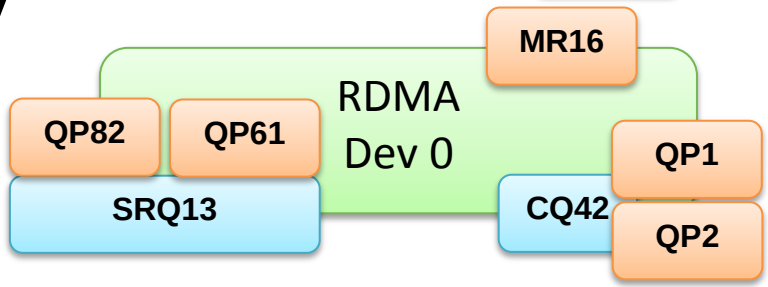
- Active backup



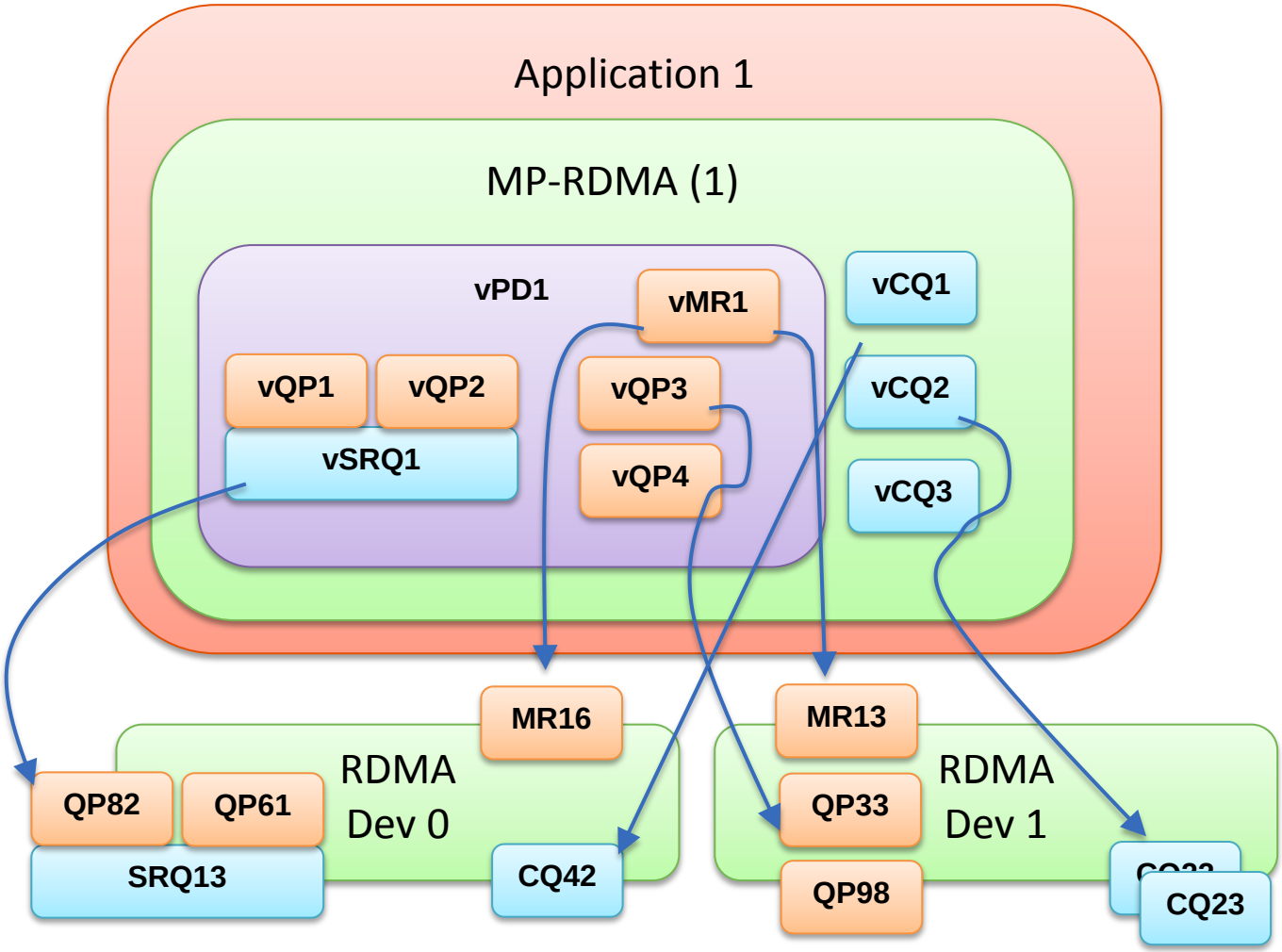
- Load Balancing



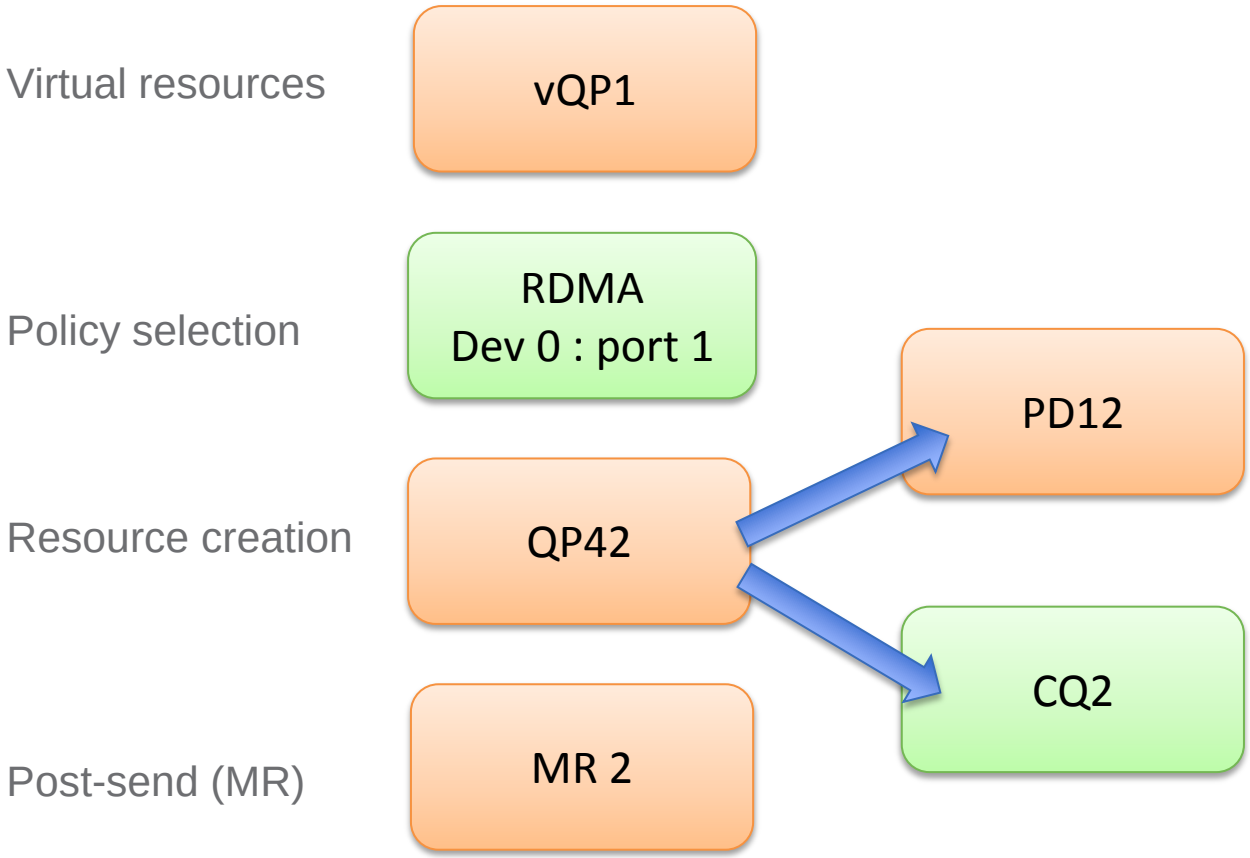
- Efficiency



Virtual Namespaces

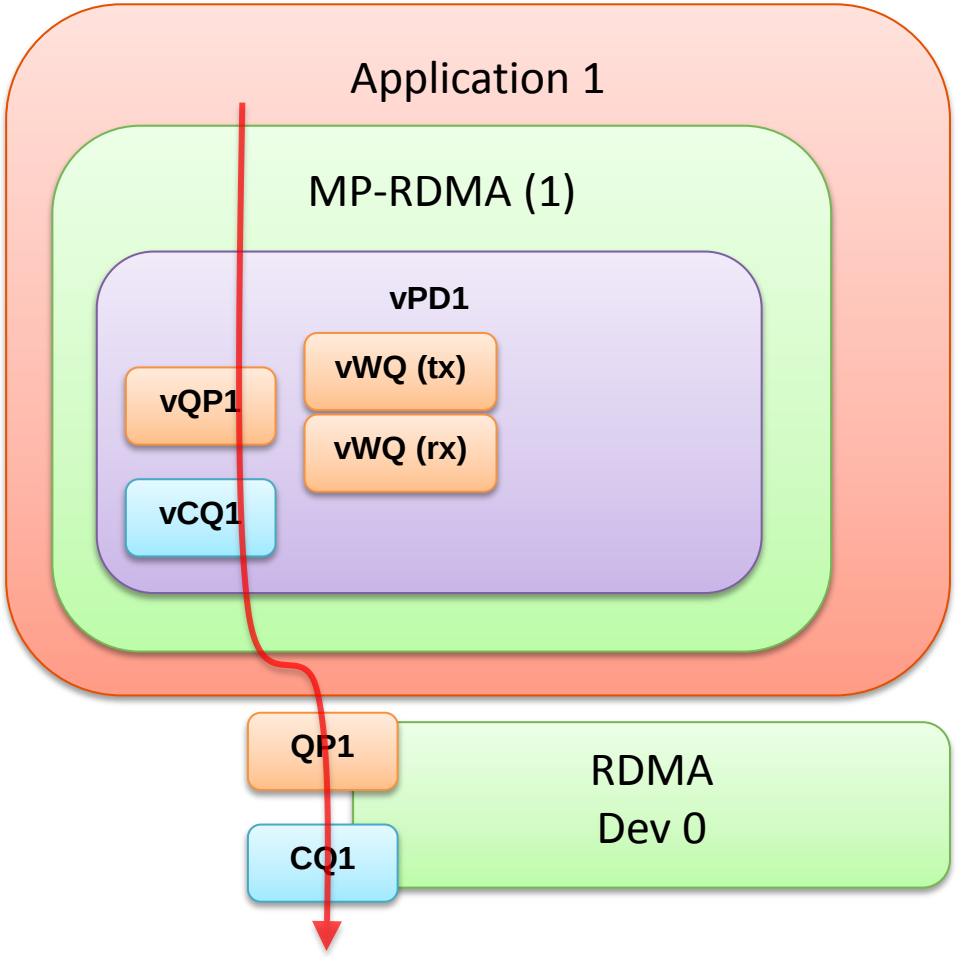


Resource Creation



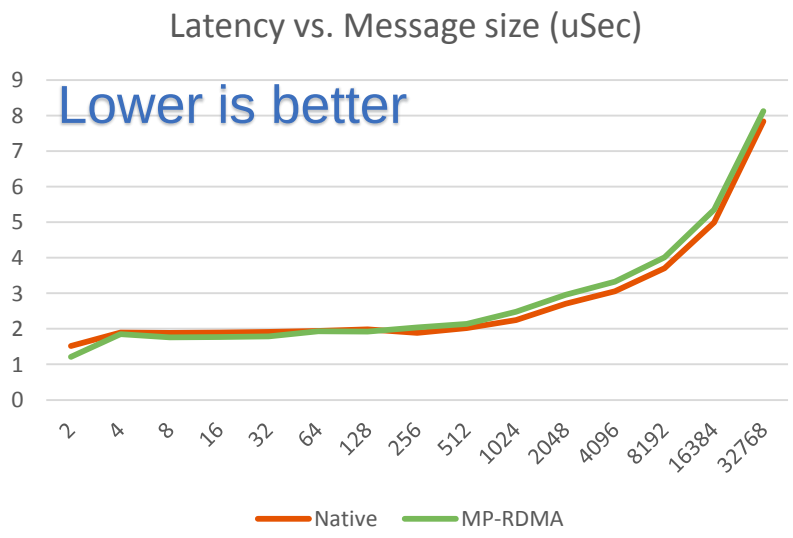
Data Path

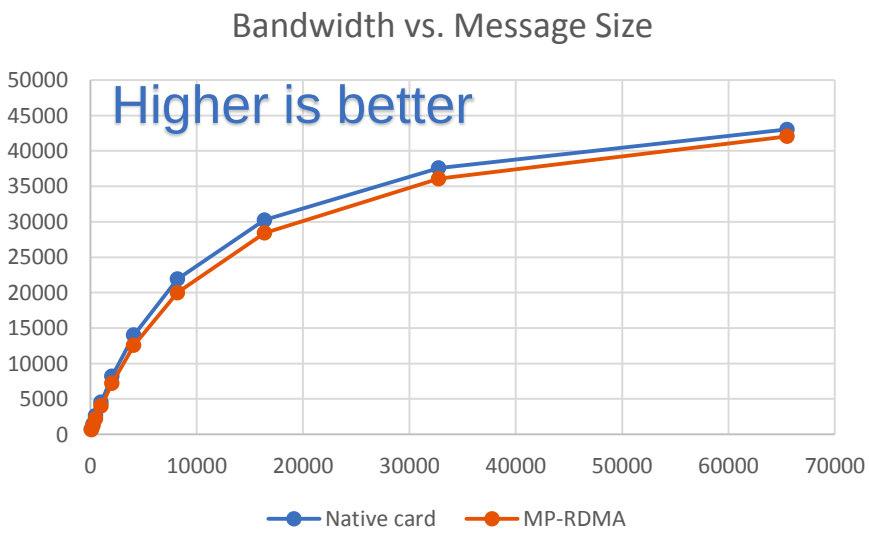
- Translate MRs:
  - ibv_post_send
  - ibv_post_recv
- Translate QPs:
  - ibv_poll_cq
- Monitor WQs:
  - PSN
  - Completed



Status and Initial Results

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





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

OPENFABRICS ALLIANCE

ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP