Support Infiniband Link Layer

Shahaf Shuler

DPDK Summit Userspace - Dublin- 2017

Agenda

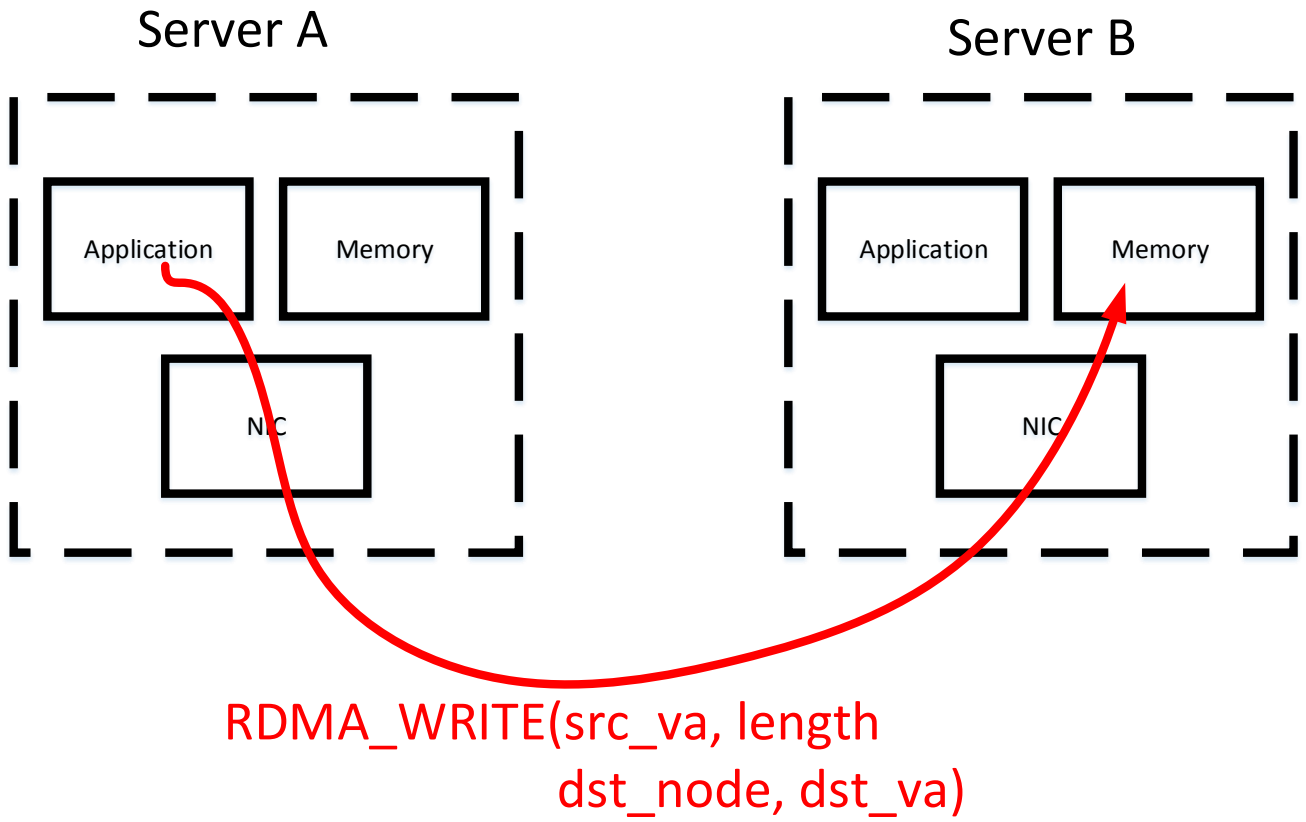
- Why Infiniband in DPDK?
- Infiniband Intro
- Infiniband network addressing
- IPoIB
- POC results
- Upstream Infiniband to DPDK

Why Infiniband?

- Many large scale Infiniband clusters in the HPC market
- Fast packet processing is required also there
  - Parallel Distributed storage applications
- Infiniband was defined with user space direct access from the start
  - DPDK is better optimized for high packet rate than the original verbs API
    - Max rate with verbs : ~30Mpps
    - Max rate with DPDK: ~50Mpps

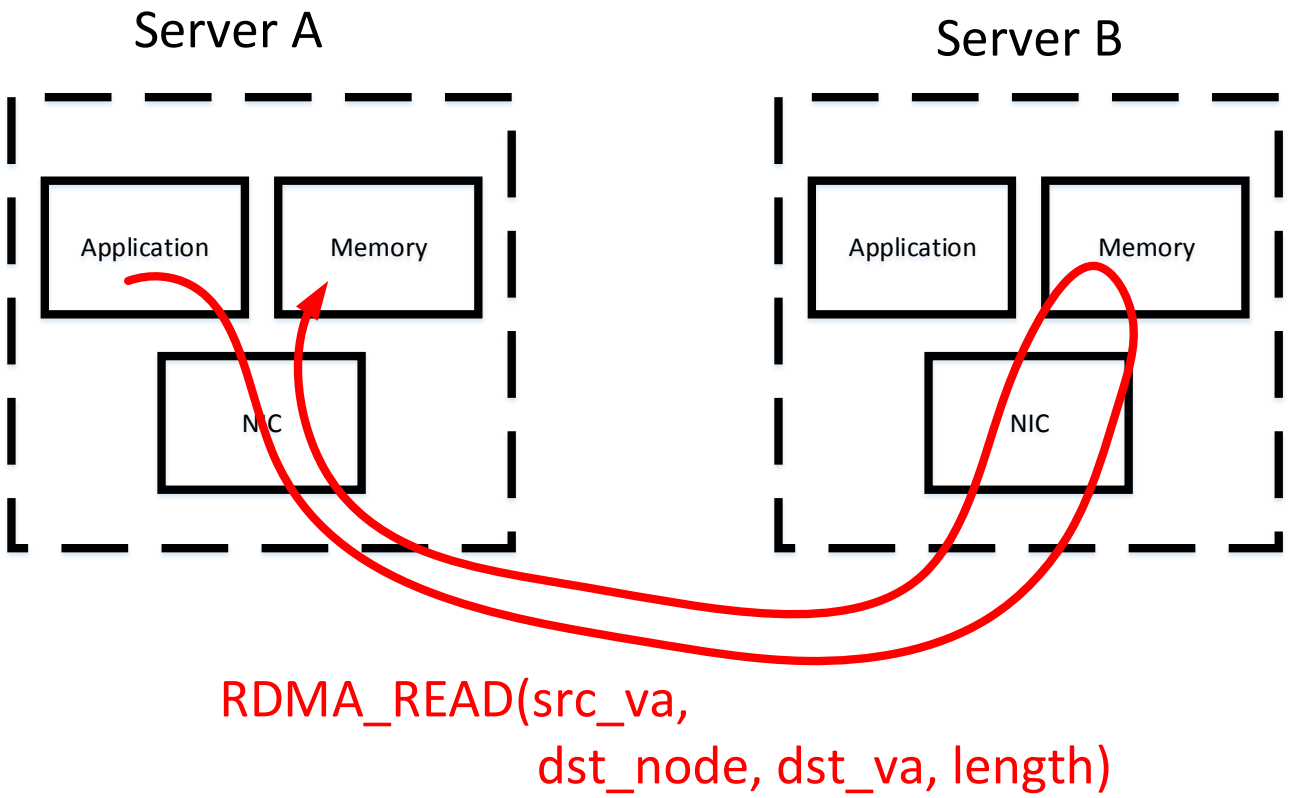
Infiniband Intro

- Infiniband (IB) computer network standard
- Centralized subnet management using the SM
- Benefits
  - Flatter topology
  - Low latency – 0.7 usec
  - L4 queues
  - RDMA and remote atomic operation



Infiniband Intro

- Infiniband (IB) computer network standard.
- Centralized subnet management using the SM
- Benefits
  - Flatter topology
  - Low latency – 0.6 usec
  - L4 queues
  - RDMA and remote atomic operation



Infiniband Network Addressing

Ethernet	Infiniband	Comments
MAC (6 Bytes)	LID (2 Bytes)	LIDs are configured by the SM. LIDs are statically configured in the switch.
VLAN (4 Bytes)	PKey	No VLAN offloads
IPv4 (4 Bytes) IPv6(16 Bytes)	GID (16 Bytes)	GIDs are fixed. cannot be changed.
UDP/TCP port (2 Bytes)	QP (3 Bytes)	Apart from MC, packet targets single QP. No IB RSS yet. No promiscuous No all multi
ARP	Path query	{LID,MTU} = PQ(GID)

bits bytes	31-24		23-16			15-8	7-0
0-3	VL	LVer	SL	Rsv2	LNH	Destination Local Identifier	
4-7	Reserve 5	Packet Length (11 bits)				Source Local Identifier	

bits bytes	31-24		23-16	15-8	7-0
0-3	IPVer	TClass	FlowLabel		
4-7	PayLen			NxtHdr	HopLmt
8-11	SGID[127-96]				
12-15	SGID[95-64]				
16-19	SGID[63-32]				
20-23	SGID[31-0]				
24-27	DGID[127-96]				
28-31	DGID[95-64]				
32-35	DGID[63-32]				
36-39	DGID[31-0]				

bits bytes	31-24		23-16				15-8	7-0
0-3	OpCode		SE	M	Pad	TVer	Partition Key	
4-7	Reserved 8 (masked in ICRC)		Destination QP					
8-11	A	Reserved 7	PSN - Packet Sequence Number					

IPoIB

- IPoIB – encapsulation of IP packet in IB message.
- IPoIB RFC - https://tools.ietf.org/html/rfc4391
- Linux generic netdev for InfiniBand
- It is possible to do RSS and promiscuous (all IPs)
- It is possible to use TSO and checksum offloads

```
$ ip address show
2: ib0: <BROADCAST,MULTICAST,UP,LOWER UP> mtu 4092 qdisc mq state UP qlen 1024
    link/infiniband a0:00:03:00:fe:80:00:00:00:00:00:00:00:02:c9:03:00:2f:ff:d1 brd
00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff
    inet 50.50.50.12/24 brd 50.50.50.255 scope global ib0
       valid lft forever preferred lft forever
    inet6 fe80::202:c903:2f:ffd1/64 scope link
       valid lft forever preferred lft forever
3: eth3: <BROADCAST,MULTICAST,UP,LOWER UP> mtu 1500 qdisc mq state UP qlen 1000
    link/ether e4:1d:2d:e8:34:ac brd ff:ff:ff:ff:ff:ff
    inet 40.40.40.12/24 brd 40.40.40.255 scope global ens3
       valid lft forever preferred lft forever
    inet6 fe80::e61d:2dff:fee8:34ac/64 scope link
       valid lft forever preferred lft forever
```

POC Results

- POC
  - Done with Weka.io (Parallel distributed Storage App based on DPDK)
  - DPDK v17.05 patched with IPoIB support
  - Replacing SM query with udp socket IB address exchange inside the application with the PMD help
  - rte_flow rules based on well knows udp port to steer traffic to the PMD queues
- HW
  - ConnectX-4, single port, speed 56Gb/sec
  - Intel(R) Xeon(R) CPU E5-2643 v4 @ 3.40GHz. Cluster size – 20 nodes.

	Single core	2 cores
Uni-dir (262x4KB from one to other)	45.3 Gbps	50Gbps
Bidir (262x4KB both ways)	28.36Gbps (per direction)	39.2Gbps (per direction)

Discussion: Upstream Infiniband to DPDK

- rte_ib_dev ?
- Sub section inside ethdev for IPoIB ?
- Helper libraries for IB address resolution ?
- Mbuf fields change
  - Application sets headers starting from IPoIB header
  - PMD needs to add the IB headers

Questions?

Shahaf Shuler

shahafs@mellanox.com

Backup: Infiniband Address resolution

- GUID = IPoIB_ARP (IP)
  - Kernel service
- {LID, MTU} = PATH_RECORD(GUID)
- Infiniband CM has Socket semantic API (listen, connect)
- User-space application listen and respond with the QP parameters