Agenda

1. What is RDMA
2. RDMA VS TCP/IP
3. IB (InfiniBand) Topology and Addressing
4. OFED Software Architecture

What is RDMA?

Current Situation and Prospect of RDMA

TERADATA

Performance acceleration vs. Ethernet

- 2X faster SQL rack to rack query
- 4X faster data load

IBM

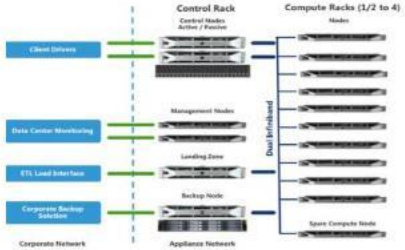
InfiniBand and RoCE

- Mellanox VPI adapter card
- Mellanox Switch (36 ports)
- Mellanox Fabric Management

ORACLE

Application Performance Improved up to 10X

Microsoft



"InfiniBand is by far the fastest and most efficient switch fabric for running enterprise data centers. " , Larry Ellison said.

What is RDMA?

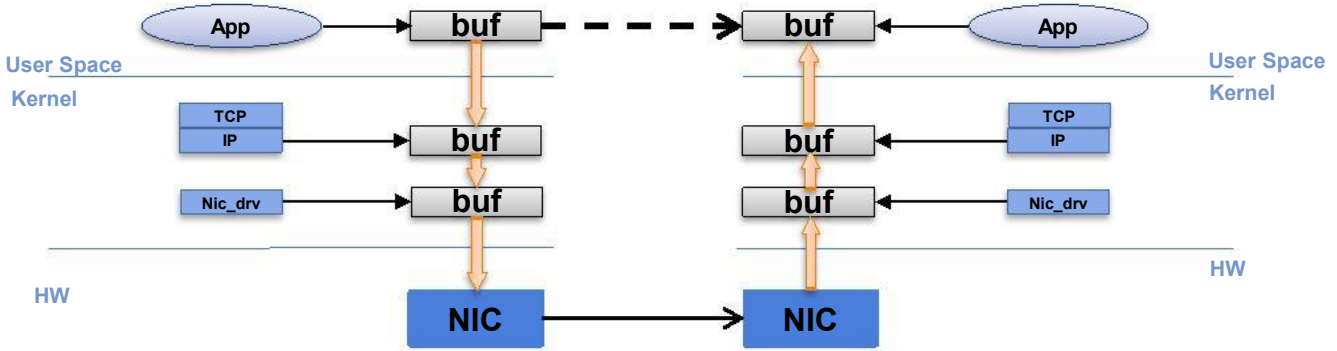
Direct memory access (DMA) is an ability of a device to access host memory directly, without the intervention of the CPU(s).

RDMA (Remote DMA) is the ability of accessing (i.e. reading from or writing to) memory on a remote machine without interrupting the processing of the CPU(s) on that system.

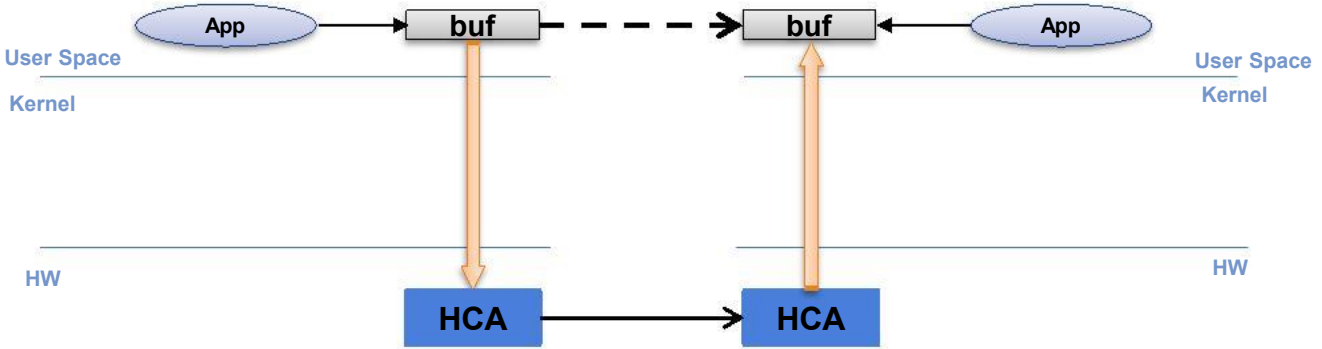
- RDMA advantages:
  - Direct user-level access to HW
    - Zero-copy
    - Kernel bypass in the fast path
    - User buffers are accessed directly
    - No CPU involvement
  - Asynchronous communication
    - Computation and communication overlap
  - HW managed transport
    - SW deals with buffers, not packets
    - Per “socket” context maintained in HW
      - No need for OS to multiplex HW
  - Explicit memory management
  - Improve small packet throughput

RDMA Key Feature – Zero Copy

Socket Based

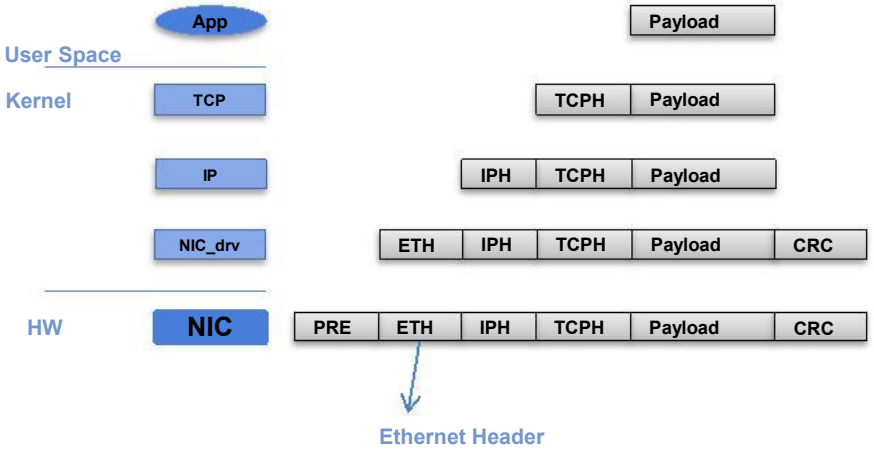


RDMA Based



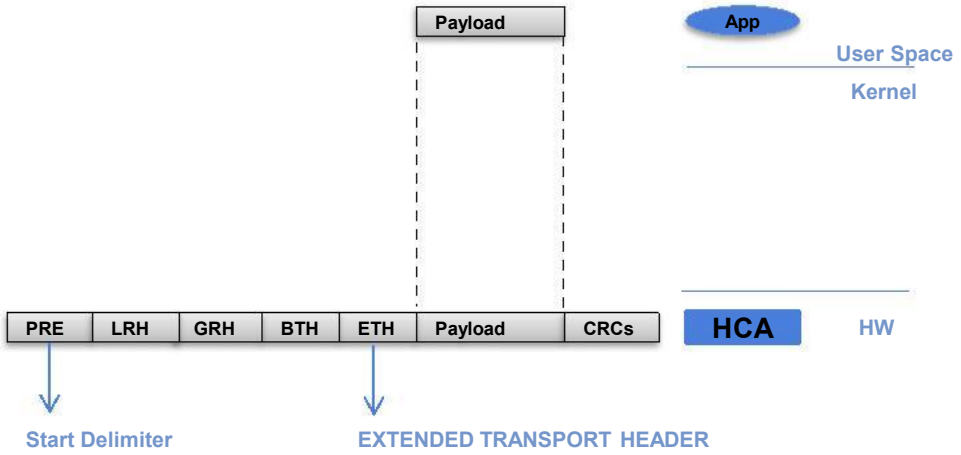
RDMA Key Feature – Full Protocol Offloading

Socket Based



- **Benefits**
  - **Socket API with long history. Many SW engineers familiar with it.**
  - **TCP/IP for both intranet and internet; RDMA currently only for intranet**

RDMA Based

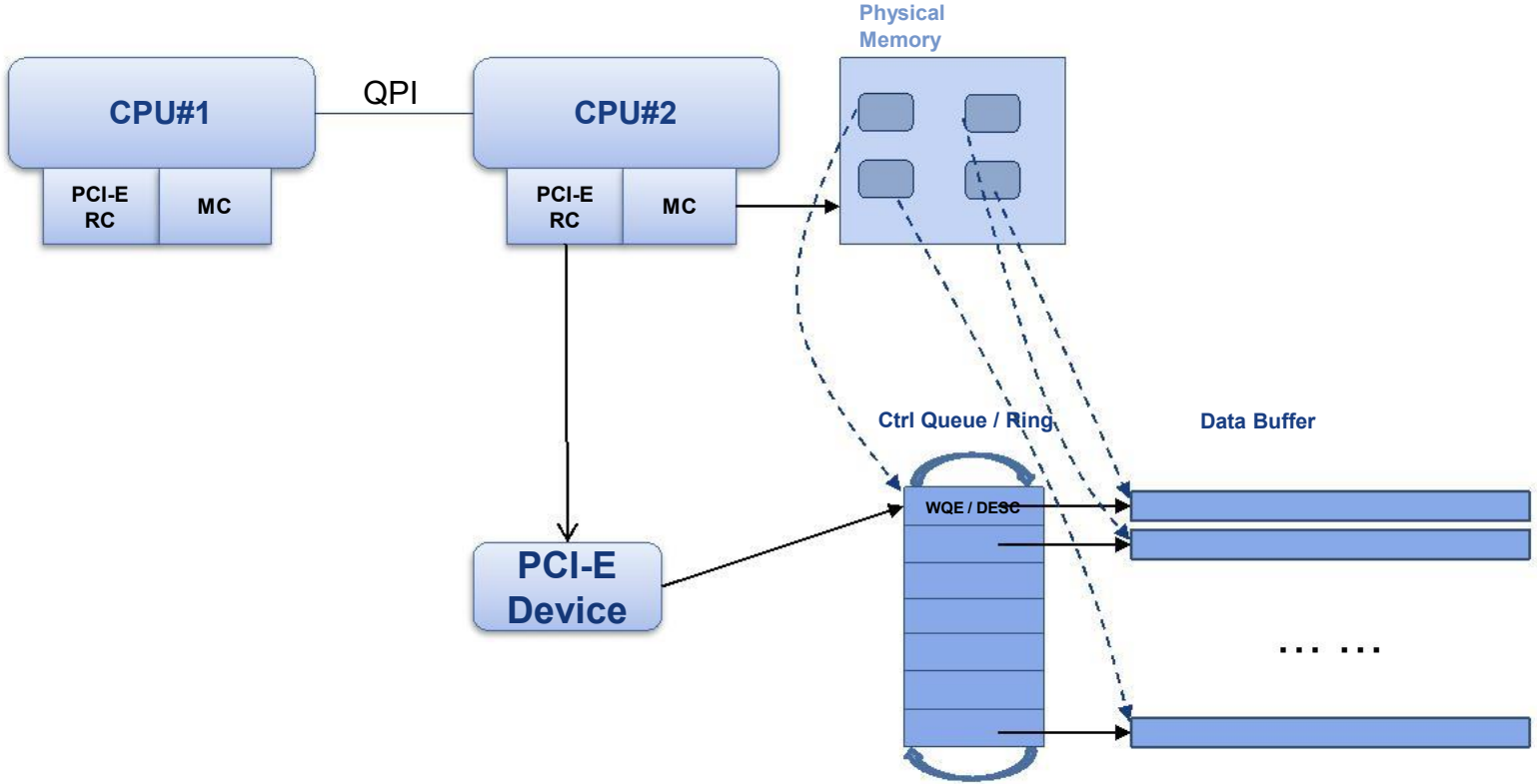


- Benefits
  - Low latency – stack bypass and copy avoidance
  - Kernel bypass – reduces CPU utilization
  - Reduces memory bandwidth bottlenecks
  - High bandwidth utilization
  - Available for both IB network and Ethernet fabric

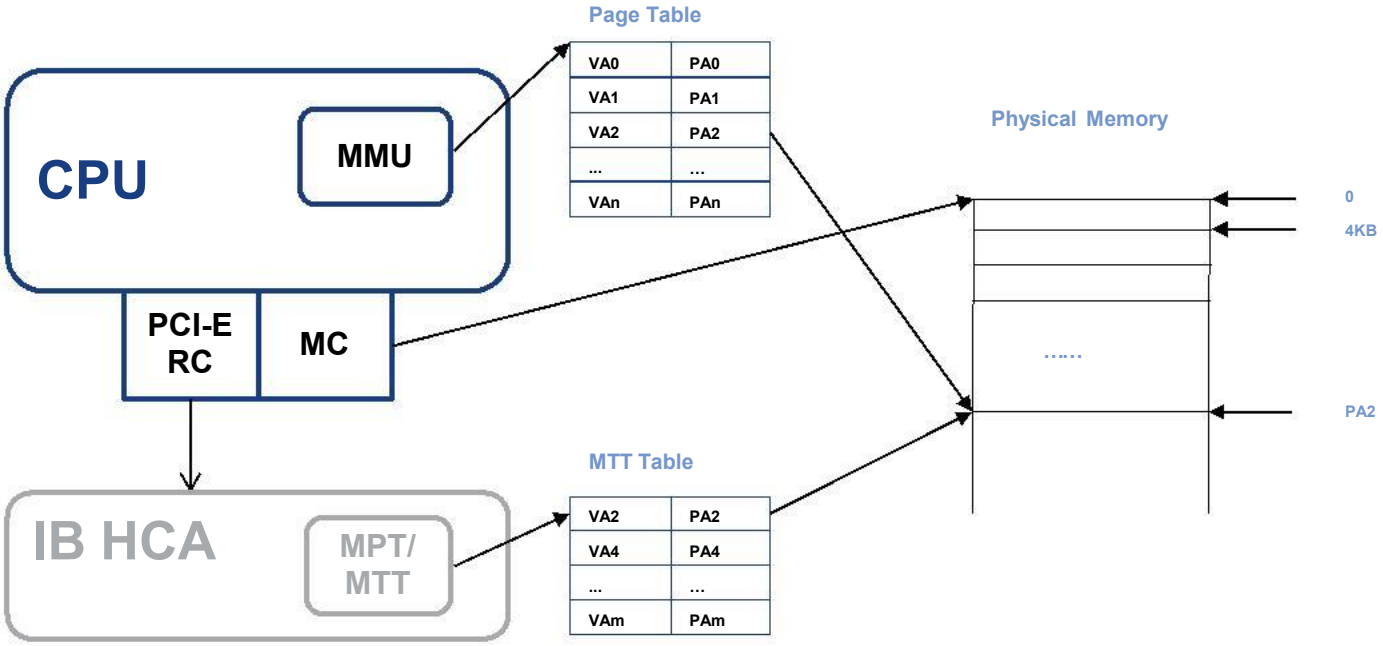
Terms and Related Technology

- PCIe
  - DMA
  - Interrupt / MSI /MSI-x
- Infiniband
  - IBTA
- OFED (Open Fabric Enterprise Distribution)
  - Verbs interface
- Memory registration
  - User space memory access
- RoCE (RDMA over Converged Ethernet)
  - DCB/DCBx/ETS/PFC/

How PCI-E Networking Device works on X86



Memory Registration



Infiniband Foundation

What is InfiniBand?

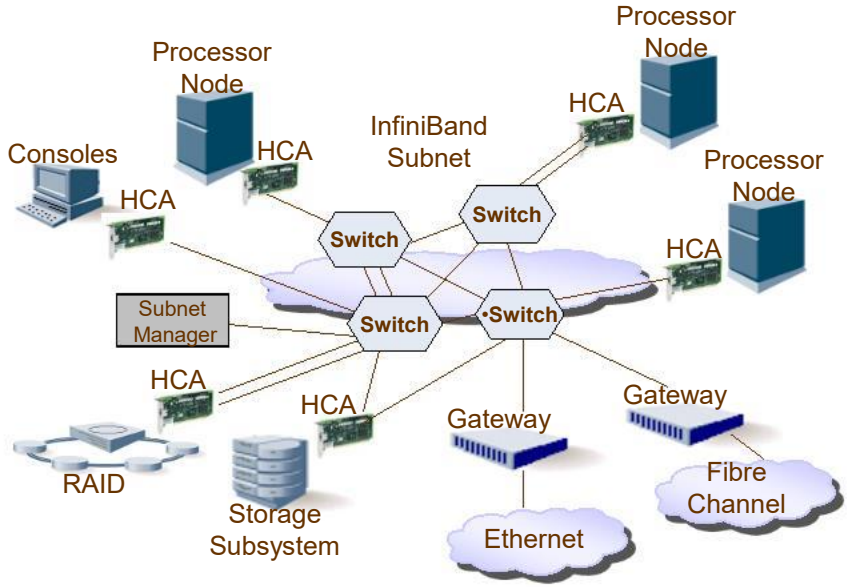
- Industry standard defined by the InfiniBand Trade Association
  - Originated in 1999
  - InfiniBand® Trade Association (IBTA) www.infinibandta.org
- InfiniBand™ specification defines an input/output architecture used to interconnect servers, communications infrastructure equipment, storage and embedded systems
- InfiniBand is a pervasive, low-latency, high-bandwidth interconnect which requires low processing overhead and is ideal to carry multiple traffic types (clustering, communications, storage, management) over a single connection.
- As a mature and field-proven technology, InfiniBand is used in thousands of data centers, high-performance compute clusters and embedded applications that scale from small scale to large scale

InfiniBand Feature Highlights

- Serial High Bandwidth Links
  - 10Gb/s to 200Gb/s HCA links
  - Up to 6.4Tb/s switch-switch
- Ultra low latency
  - Under 1 us
- Reliable, lossless, self-managing fabric
  - Link level flow control
  - Congestion control
- Full CPU Offload
  - Hardware Based Transport Protocol
  - Reliable Transport
  - Kernel Bypass
- Memory exposed to remote node
  - RDMA-read and RDMA-write
- Quality Of Service
  - I/O channels at the adapter level
  - Virtual Lanes at the link level
- Scalability/flexibility
  - Up to 48K nodes in subnet, up to 2^{128} in network

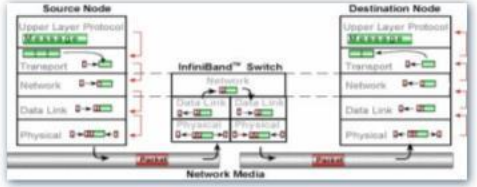
The InfiniBand Architecture

- Industry standard defined by the InfiniBand Trade Association(IBTA)
- Defines System Area Network architecture
  - Comprehensive specification: from physical to applications
- Architecture supports
  - Host Channel Adapters (HCA)
  - Switches
  - SM
  - Gateway
  - Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload



IB Architecture Layers

- **Software Transport Verbs and Upper Layer Protocols:**
  - Interface between application programs and hardware.
  - Allows support of legacy protocols such as TCP/IP
  - Defines methodology for management functions



- **Transport:**
  - Delivers packets to the appropriate Queue Pair; Message Assembly/De-assembly, access rights,..

- **Network:**
  - How packets are routed between Different Partitions /subnets

- **Data Link (Symbols and framing):**
  - Flow control (credit-based); How packets are routed , from Source to Destination on the same Partition Subnet

- **Physical:**
  - Signal levels and Frequency; Media; Connectors

Link Layer: Addressing

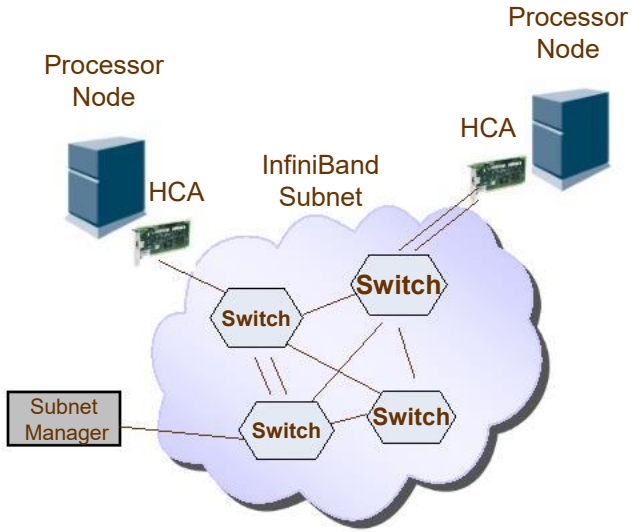
- Local ID (LID)
  - 16 bit field in the Local Routing Header (LRH) of all IB packets
  - Used to rout packet in an InfiniBand subnet
  - Each subnet may contain up to:
    - 48K unicast addresses
    - 16K multicast addresses
- Assigned by Subnet Manager at initialization and topology changes
- Not Persistent through reboots
- Address ranges

0x0000 = reserved

0x0001 = 0xBFFF = Unicast

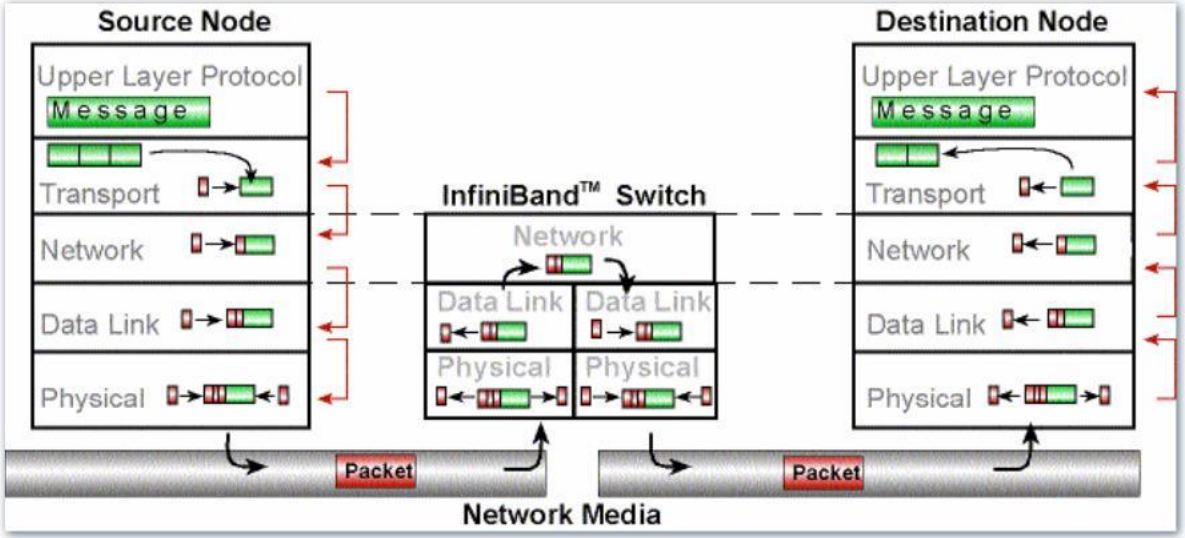
0xc000 = 0xFFFE = Multicast

0xFFFF = Reserved for special use



Transport Layer – Responsibilities

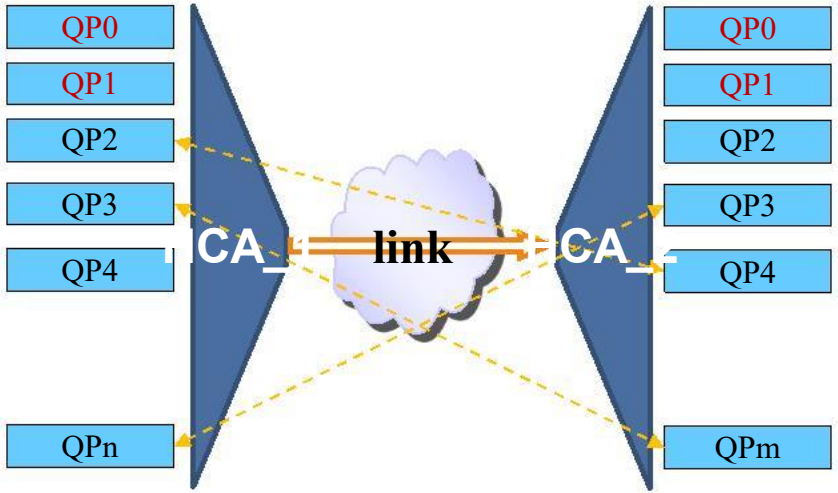
- The network and link protocols deliver a packet to the desired destination.
- The transport Layer
  - Segmenting Messages data payload coming from the Upper Layer into multiple packets that will suit Valid MTU size
  - Delivers the packet to the proper **Q**ueue **P**air (assigned to a specific session)
  - Instructs the QP how to process the packet's data. (Work Request Eelement)
  - *R*eassembles the Packets arriving from the Other side into Messages



Packet Format

InfiniBand VS TPC/IP

TCP/IP	InfiniBand
TCP/UDP port	QP number
IP address/MAC address	LID/GID



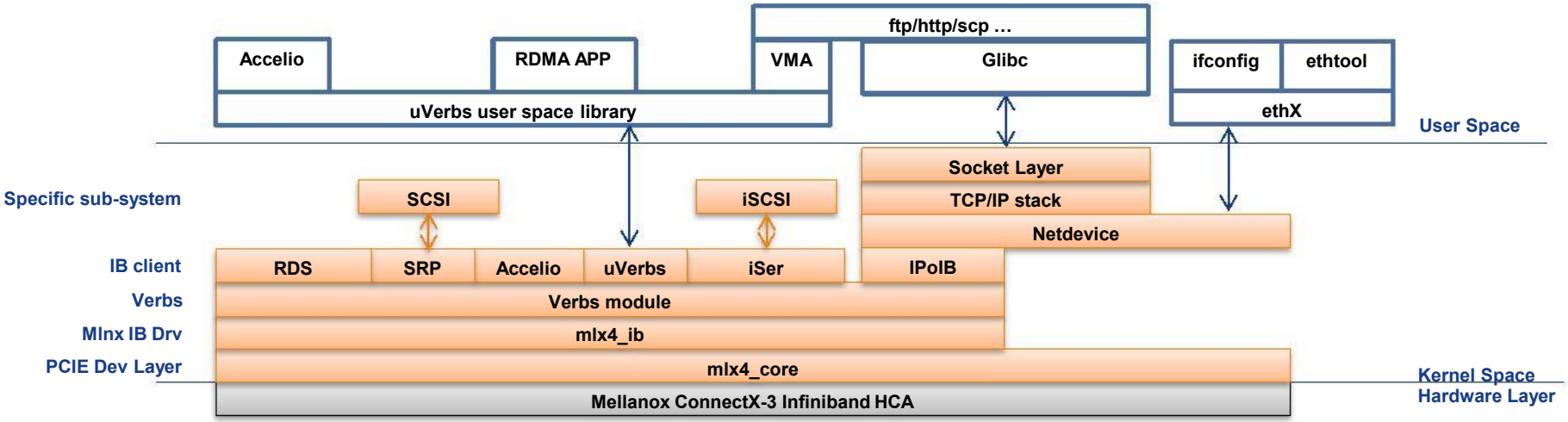
OFED

OFED

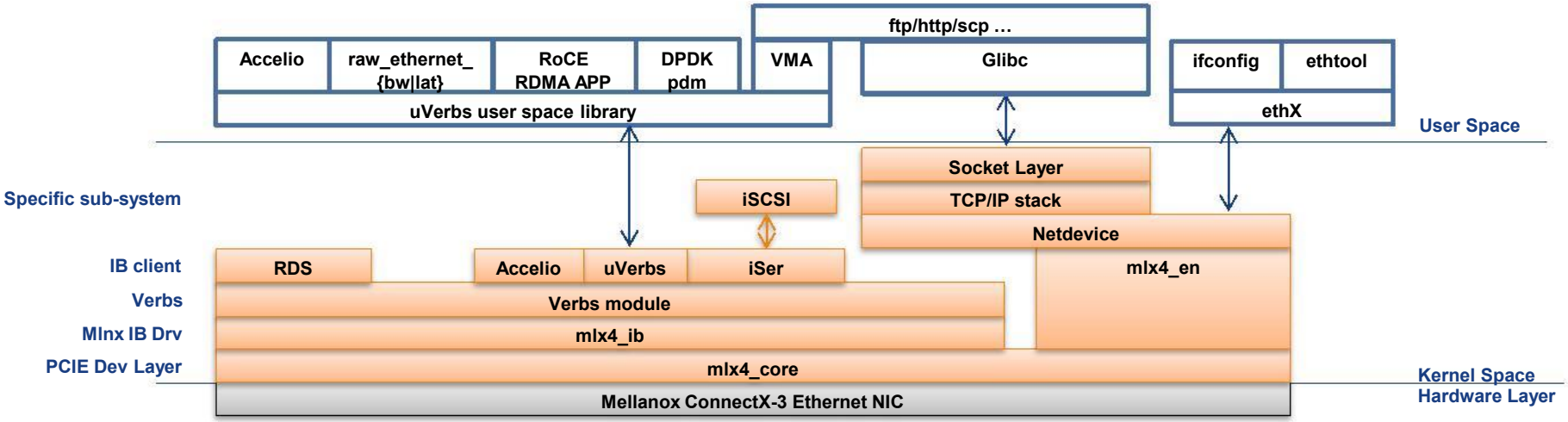
- Openfabrics Aliliance
  - https://www.openfabrics.org/index.php

- The OpenFabrics Enterprise Distribution (OFED™) is open-source software for RDMA and kernel bypass applications. OFED is used in business, research and scientific environments that require highly efficient networks, storage connectivity and parallel computing. The software provides high performance computing sites and enterprise data centers with flexibility and investment protection as computing evolves towards applications that require extreme speeds, massive scalability and utility-class reliability.

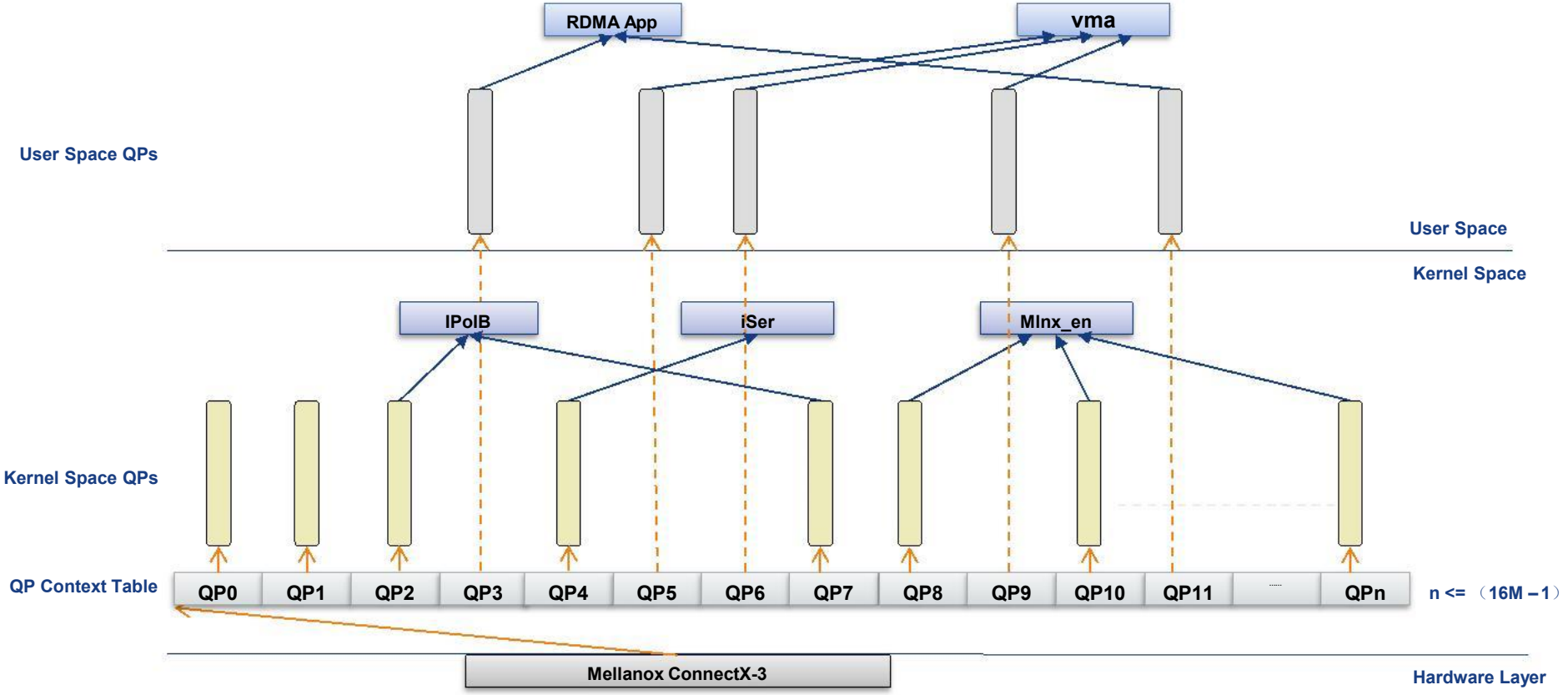
OFED SW (VPI) Arch on Infiniband Fabric



OFED SW (VPI) Arch on Ethernet Fabric



How QPs work with OFED



Q&A

THANK YOU