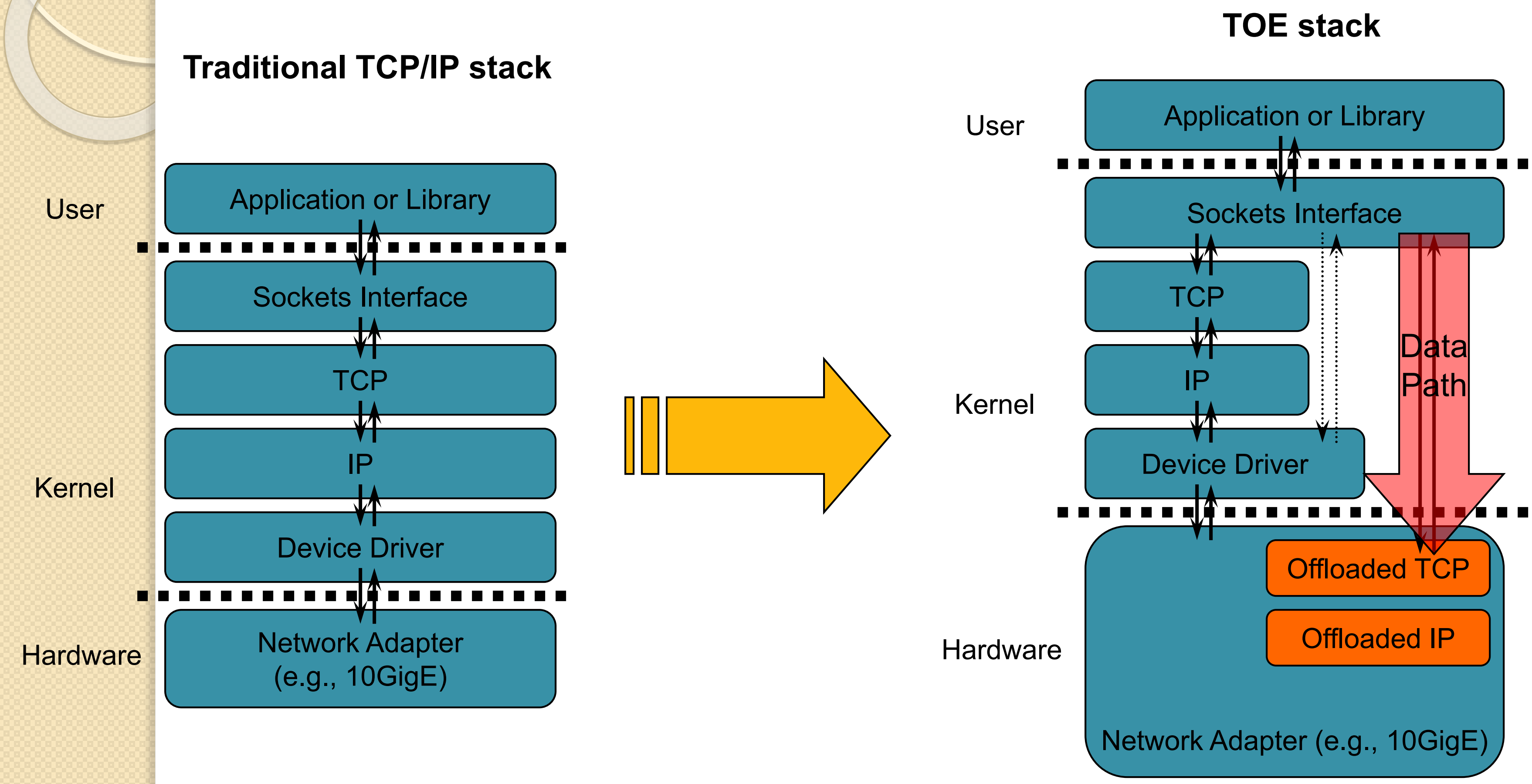
EE542
Lecture 12: Network with RDMA

Internet and Cloud Computing
Young Cho
Department of Electrical Engineering
University of Southern California

Ethernet: Technology Trends

- Regular Ethernet adapters
  - Layer-2 adapters
  - Rely on host-based TCP/IP to provide network/transport functionality
  - Could achieve a high performance with optimizations
- TCP Offload Engines (TOEs)
  - Layer-4 adapters
  - Have the entire TCP/IP stack offloaded on to hardware
  - Sockets layer retained in the host space
- RDMA-aware adapters
  - Layer-4 adapters
  - Entire TCP/IP stack offloaded on to hardware
  - Support more features than TCP Offload Engines
    - No sockets ! Richer RDMA interface !
    - E.g., Out-of-order placement of data, RDMA semantics

What is a TCP Offload Engine (TOE)?



RDMA

- A method for interconnecting platforms in high-speed networks that overcomes many of the difficulties encountered with traditional networks such as TCP/IP over Ethernet.
  - new standards
  - new protocols
  - new hardware interface cards and switches
  - new software

Remote Direct Memory Access

- Remote
  - data transfers between nodes in a network
- Direct
  - no Operating System Kernel involvement in transfers
  - everything about a transfer offloaded onto Interface Card
- Memory
  - transfers between user space application virtual memory
  - no extra copying or buffering
- Access
  - send, receive, read, write, atomic operations

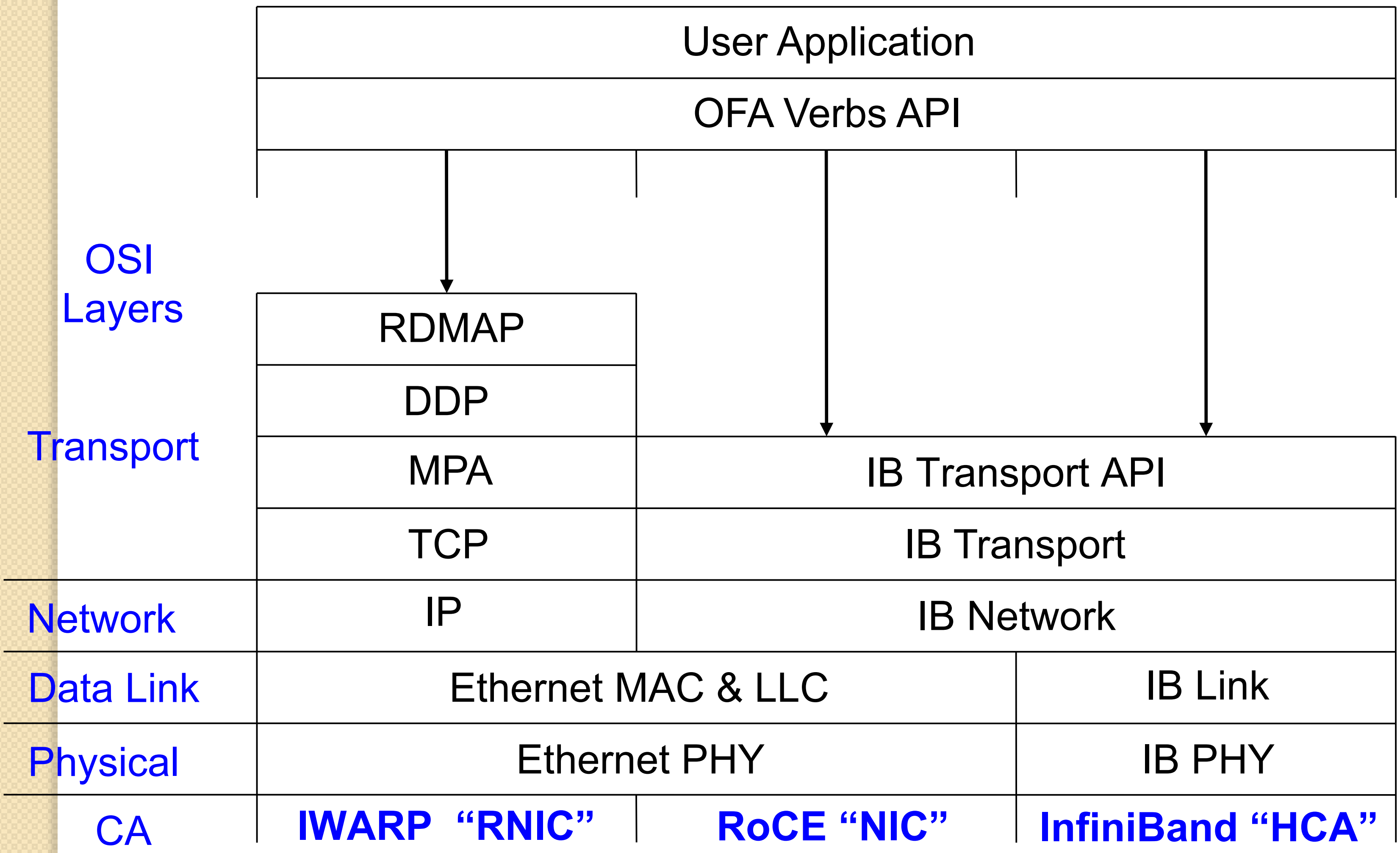
RDMA Benefits

- High throughput
- Low latency
- High messaging rate
- Low CPU utilization
- Low memory bus contention
- Message boundaries preserved
- Asynchronous operation

RDMA Technologies

- InfiniBand – (41.8% of top 500 supercomputers)
  - SDR 4x – 8 Gbps
  - DDR 4x – 16 Gbps
  - QDR 4x – 32 Gbps
  - FDR 4x – 54 Gbps
- iWarp – internet Wide Area RDMA Protocol
  - 10 Gbps
- RoCE – RDMA over Converged Ethernet
  - 10 Gbps
  - 40 Gbps

RDMA Architecture Layering



OSI Layers			
	User Application		
	OFA Verbs API		
Transport	RDMAP		
	DDP		
	MPA	IB Transport API	
	TCP	IB Transport	
Network	IP	IB Network	
Data Link	Ethernet MAC & LLC		IB Link
Physical	Ethernet PHY		IB PHY
CA	**IWARP “RNIC”**	**RoCE “NIC”**	**InfiniBand “HCA”**

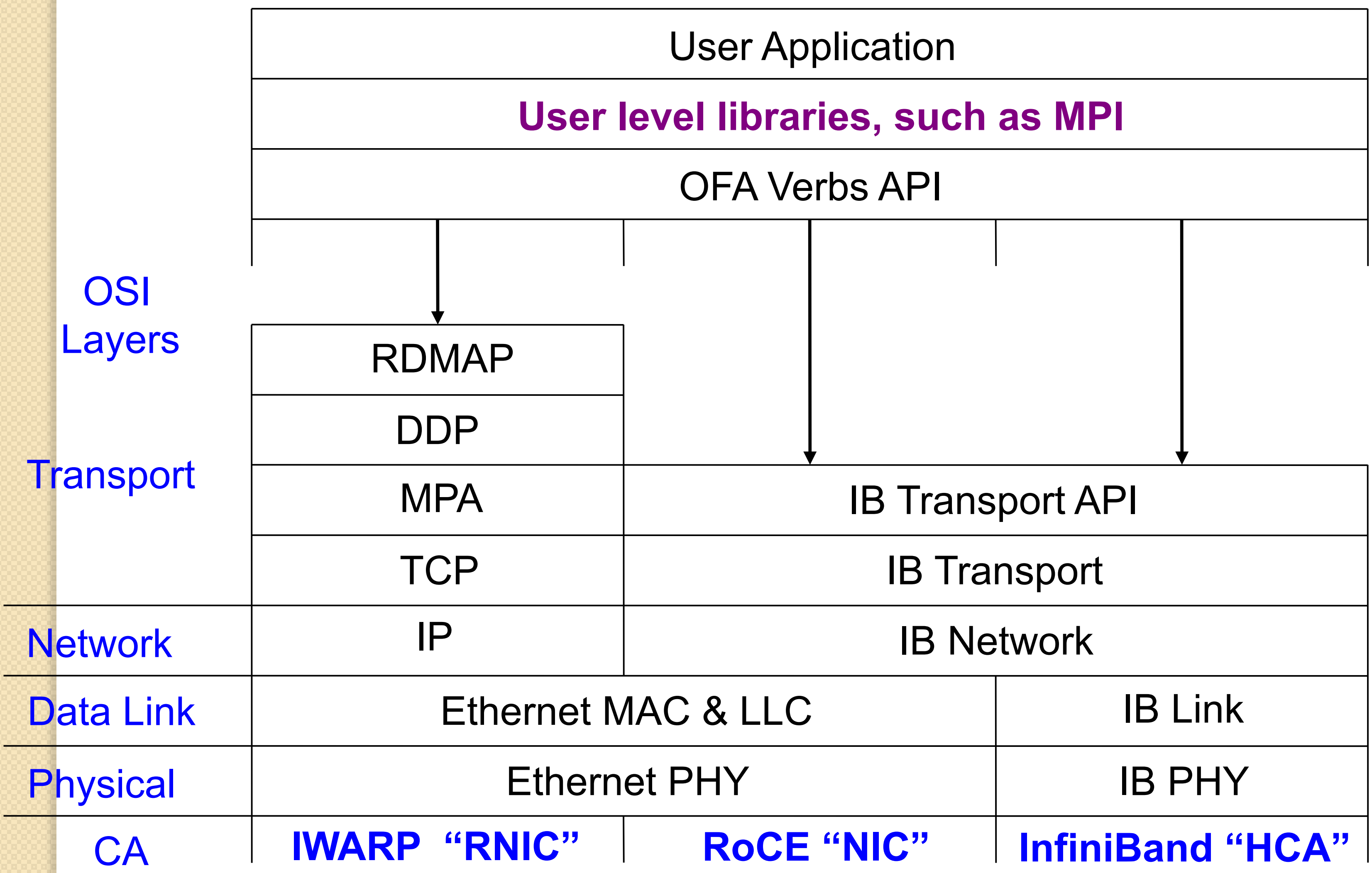
Specification

- InfiniBand specification
  - semantic description of required behavior
  - no syntactic or operating system specific details
  - implementations free to define their own API
    - syntax for functions, structures, types, etc.
- OpenFabrics Alliance (OFA)
  - one possible syntactic definition of an API
  - in syntax, each “verb” becomes an equivalent “function”
  - done to prevent proliferation of incompatible definitions
  - was an OFA strategy to unify InfiniBand market

Libraries that access RDMA

- MPI – Message Passing Interface
  - Main tool for High Performance Computing (HPC)
    - Physics, fluid dynamics, modeling and simulations
  - Many versions available
    - OpenMPI
    - MVAPICH
    - Intel MPI

Layering with user level libraries



User Application

User level libraries, such as MPI

OFA Verbs API

OSI Layers

RDMAP

DDP

Transport

MPA

TCP

IB Transport API

IB Transport

Network

IP

IB Network

Data Link

Ethernet MAC & LLC

IB Link

Physical

Ethernet PHY

IB PHY

CA

IWARP “RNIC”

RoCE “NIC”

InfiniBand “HCA”

Additional ways to access RDMA

- File systems
  - Lustre – parallel distributed file system for Linux
  - NFS_RDMA – Network File System over RDMA
- Storage appliances by DDN and NetApp
  - SRP – SCSI RDMA (Remote) Protocol – Linux kernel
  - iSER – iSCSI Extensions for RDMA – Linux kernel

Additional ways to access RDMA

Pseudo sockets libraries

- SDP – Sockets Direct Protocol – supported by Oracle
- rsockets – RDMA Sockets – supported by Intel
- mva – Mellanox Messaging Accelerator
- SMC-R – proposed by IBM

Similarities between TCP and RDMA

- Both utilize the client-server model
- Both require a connection for reliable transport
- Both provide a reliable transport mode
  - TCP provides a reliable in-order sequence of **bytes**
  - RDMA provides a reliable in-order sequence of **messages**

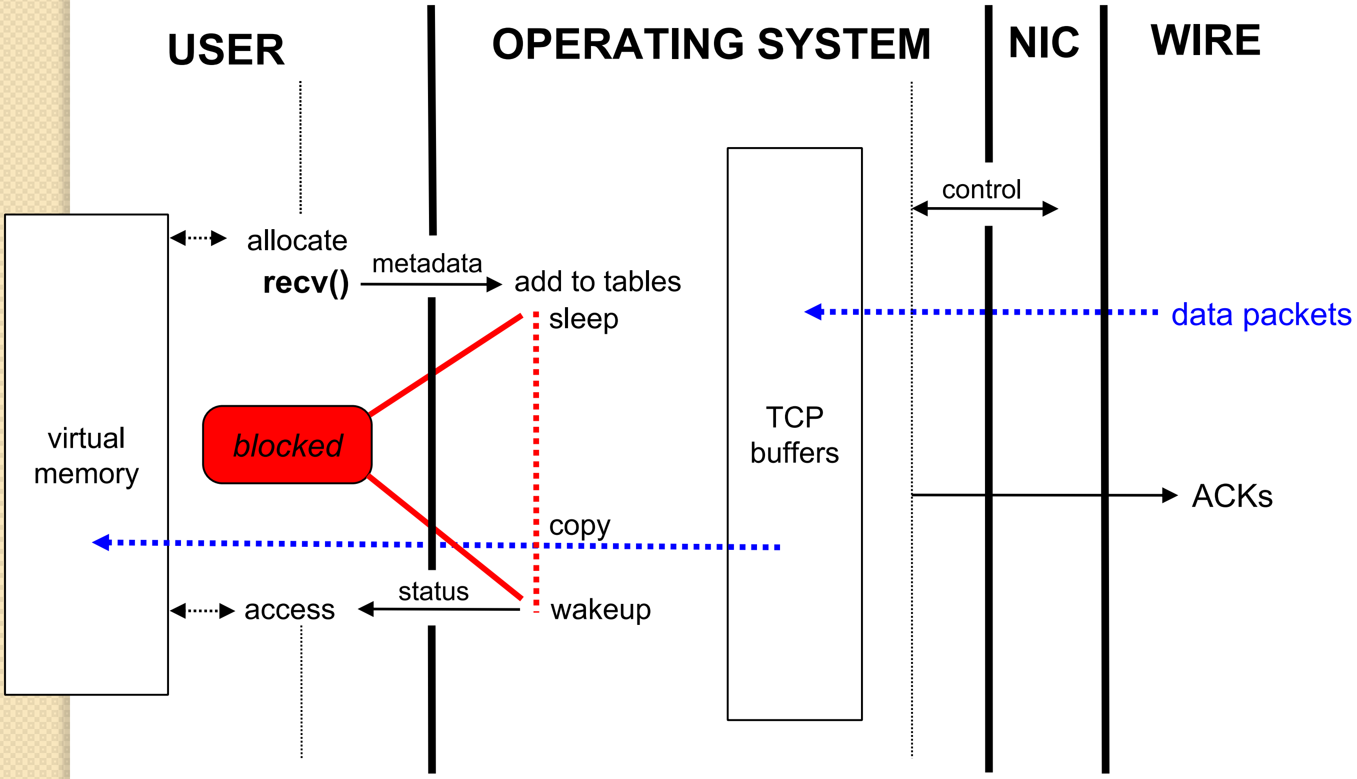
How RDMA differs from TCP/IP

- “zero copy” – data transferred directly from virtual memory on one node to virtual memory on another node

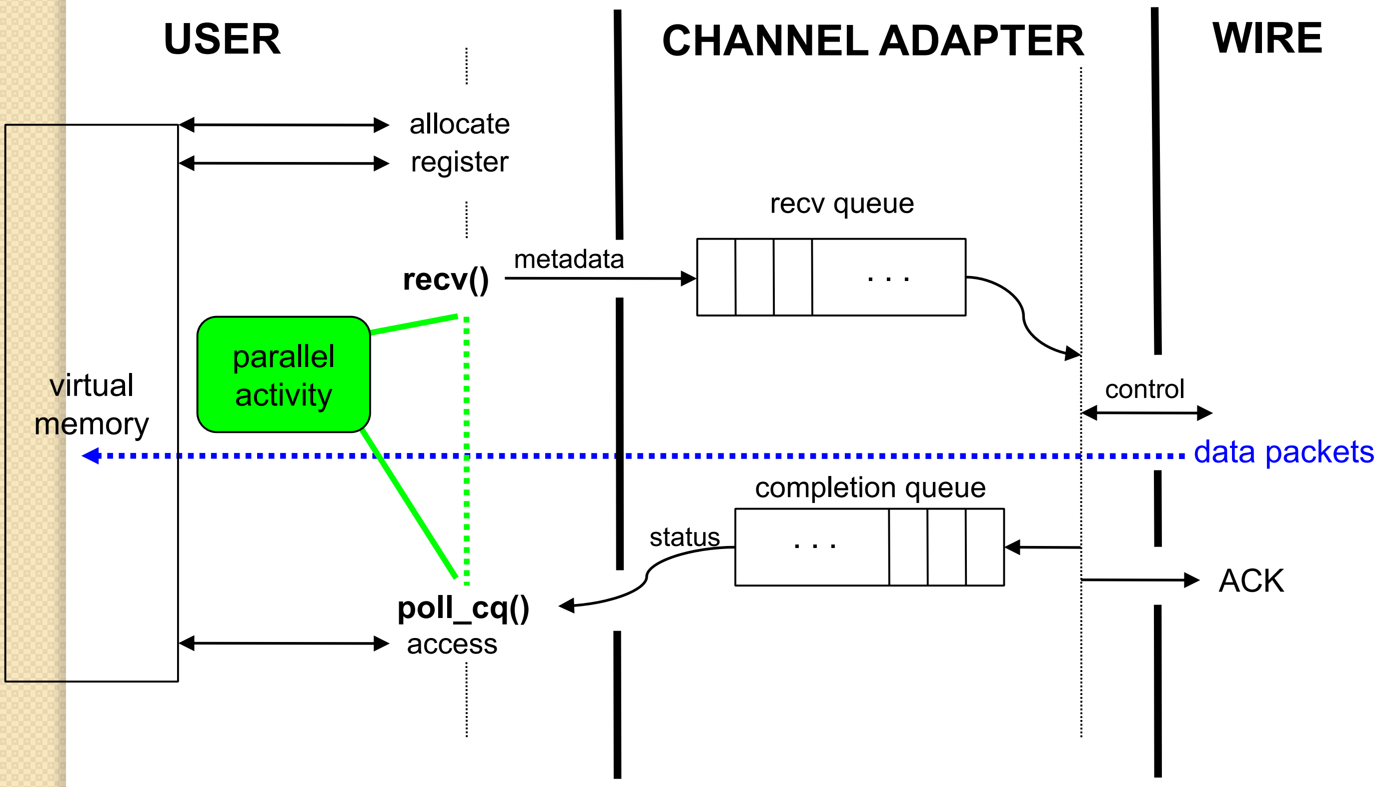
- “kernel bypass” – no operating system involvement during data transfers

- asynchronous operation – threads not blocked during I/O transfers

TCP RECV()


USER
OPERATING SYSTEM
NIC
WIRE
control
allocate
metadata
recv()
add to tables
sleep
data packets
virtual
memory
blocked
TCP
buffers
ACKs
copy
status
access
wakeup

RDMA RECV()


USER
CHANNEL ADAPTER
WIRE
allocate
register
recv queue
metadata
recv()
. . .
virtual
memory
parallel
activity
control
data packets
completion queue
status
. . .
ACK
poll_cq()
access

RDMA access model

- ❖Messages – preserves user's message boundaries
- ❖Asynchronous – no blocking during a transfer, which
  - –starts when metadata added to work queue
  - –finishes when status available in completion queue
- ❖ 1-sided (unpaired) and 2-sided (paired) transfers
- ❖ No data copying into system buffers
  - –order and timing of send() and recv() are **relevant**
    - •recv() must be waiting before issuing send()
  - –memory involved in transfer is **untouchable** between start and completion of transfer

Kernel Bypass

- User interacts directly with CA queues
- Queue Pair from program to CA
  - work request – data structure describing data transfer
  - send queue – post work requests to CA that send data
  - secv queue – post work requests to CA that receive data
- Completion queues from CA to program
  - work completion – data structure describing transfer status
  - Can have separate send and receive completion queues
  - Can have one queue for both send and receive completions

iWARP

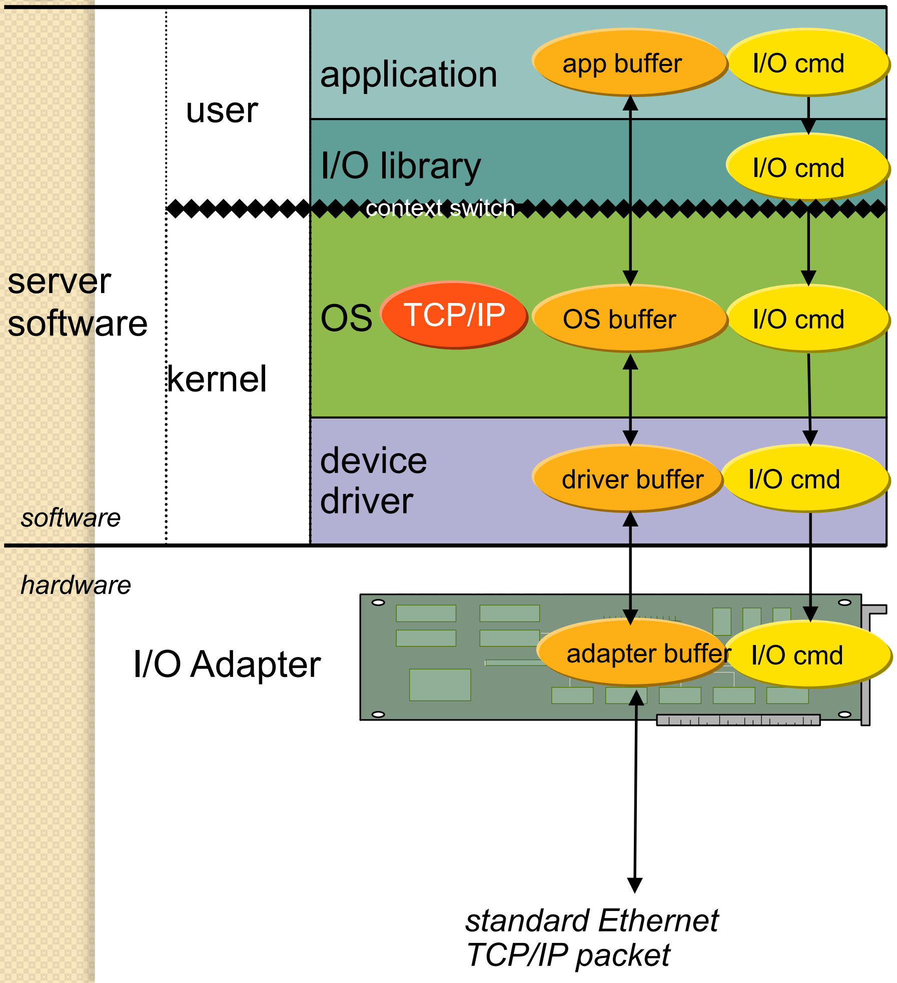
- Internet Wide Area RDMA Protocol
- RDMA over TCP/IP
  - compatible with the existing Internet infrastructure
- Uses RDMA and OS bypass to move data without the CPU or OS being involved, greatly increasing performance.
- Protocol offload – RDMA-enabled Network Interface Card (RNIC)

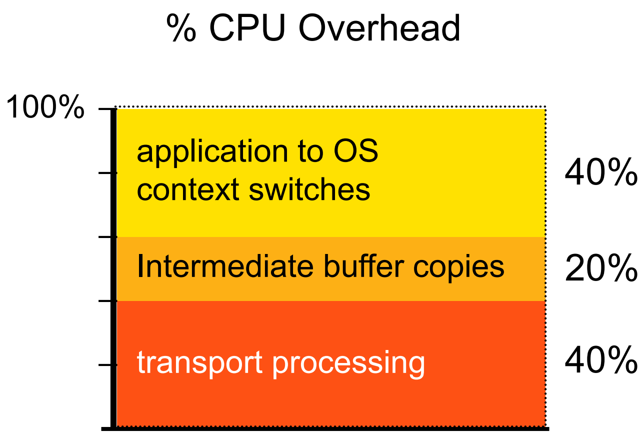
Networking Performance Barriers

Packet Processing

Intermediate Buffer Copies

Command Context Switches



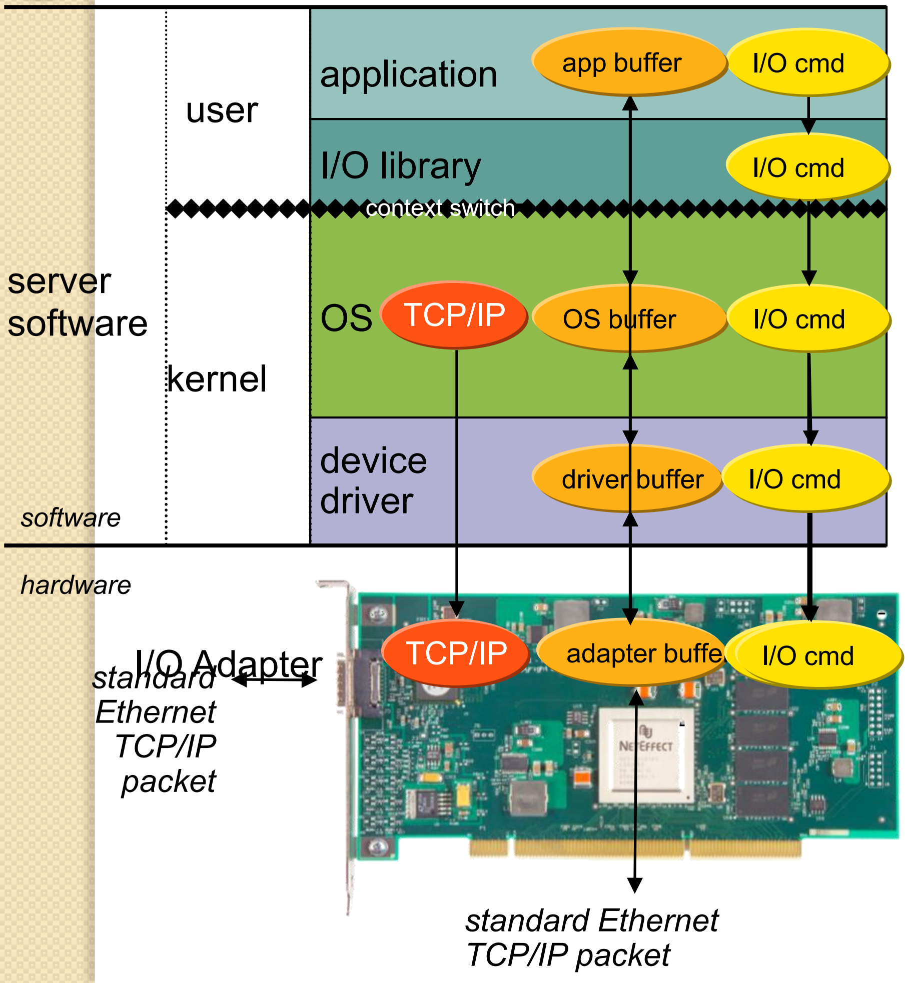


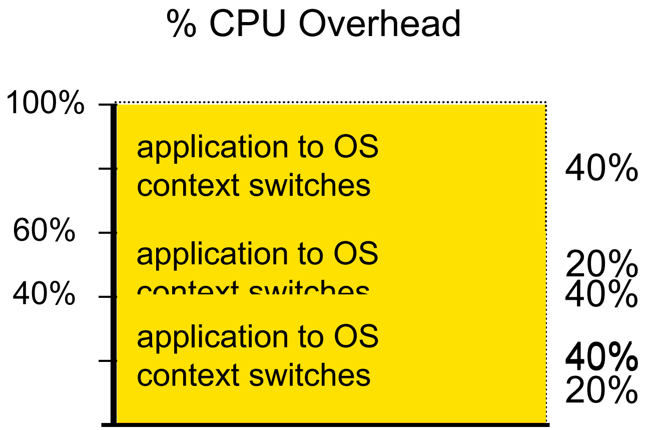
Eliminate Networking Performance Barriers With iWARP

Packet Processing

Intermediate Buffer Copies

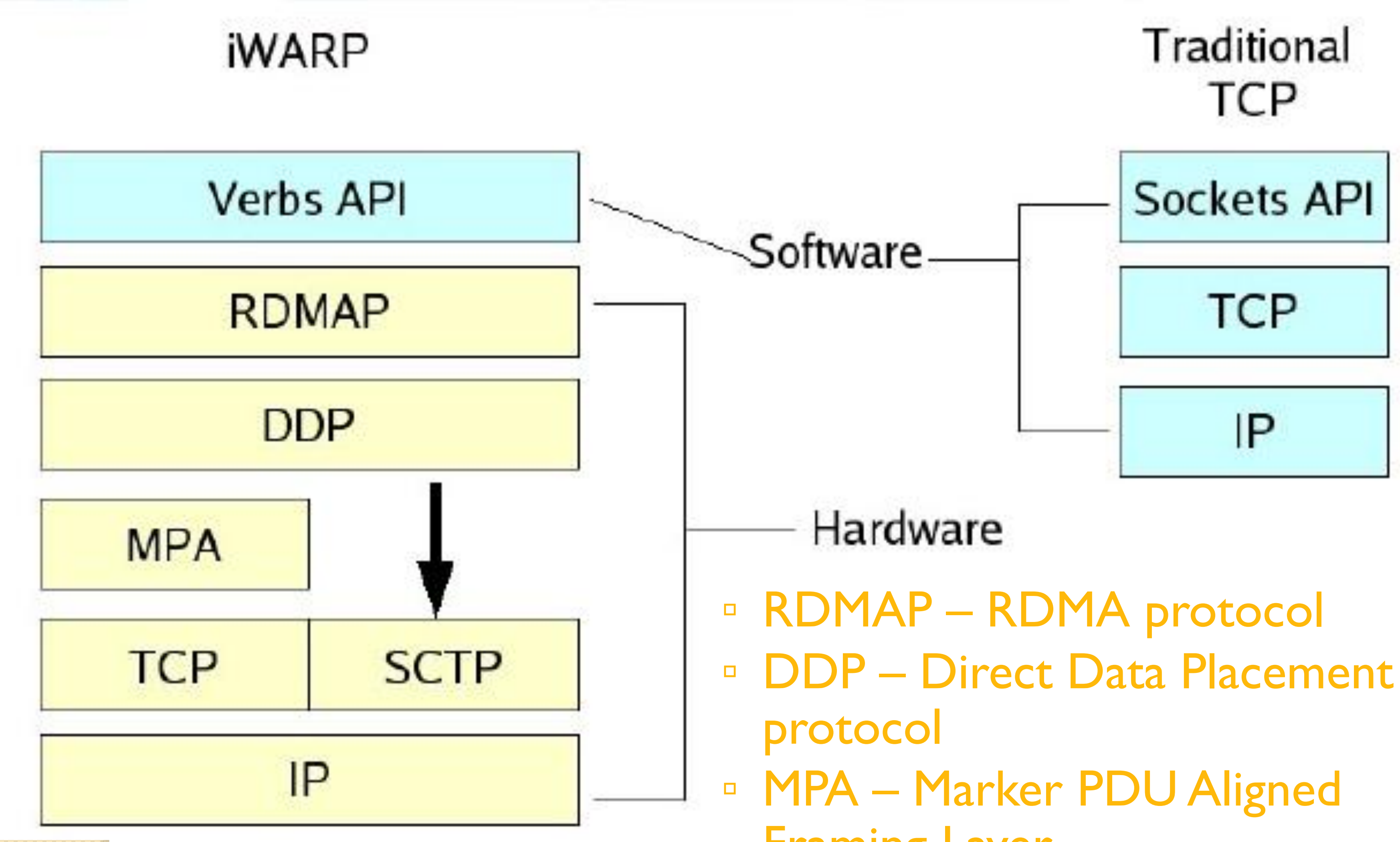
Command Context Switches





- Transport (TCP) offload
- RDMA / DDP
- User-Level Direct Access/ OS Bypass

iWARP Protocol Stack

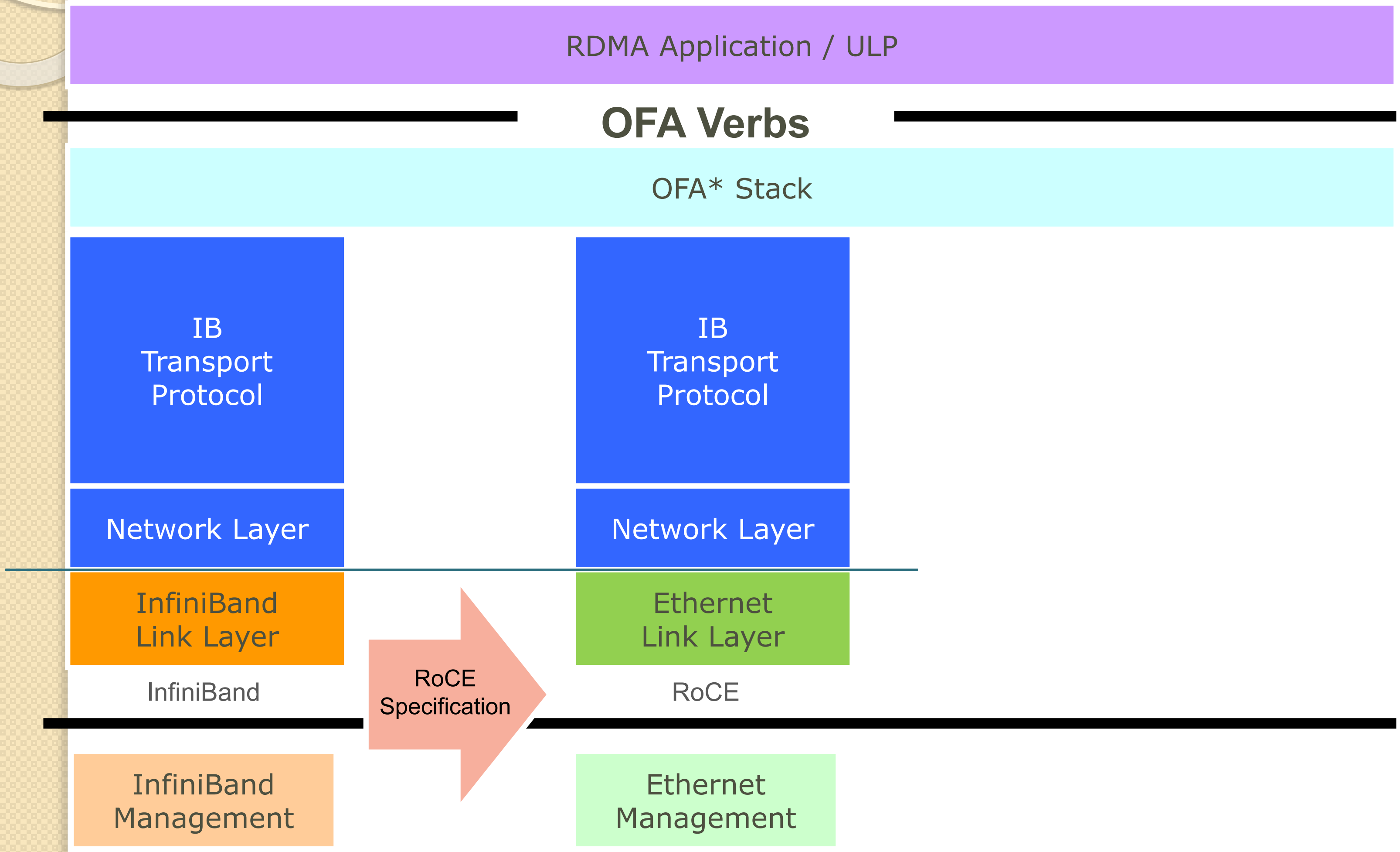


- RDMAP – RDMA protocol
- DDP – Direct Data Placement protocol
- MPA – Marker PDU Aligned Framing Layer

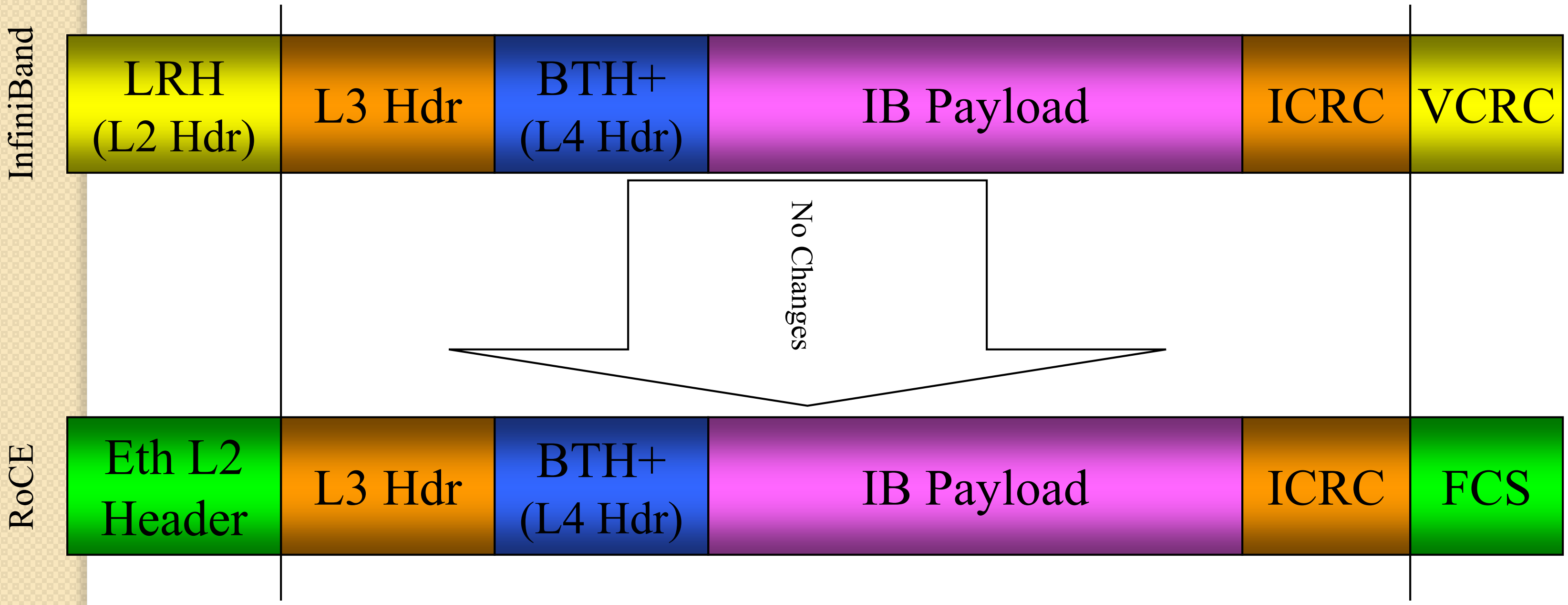
iWARP Protocol Stack

- Verbs layer is the user-level interface to the RDMA-enabled NIC.
- RDMAP layer is responsible for RDMA operations, joint buffer management with DDP.
- DDP layer is used for direct zero-copy data placement, as well as segmentation and reassembly.
- MPA layer assigns boundaries to DDP messages

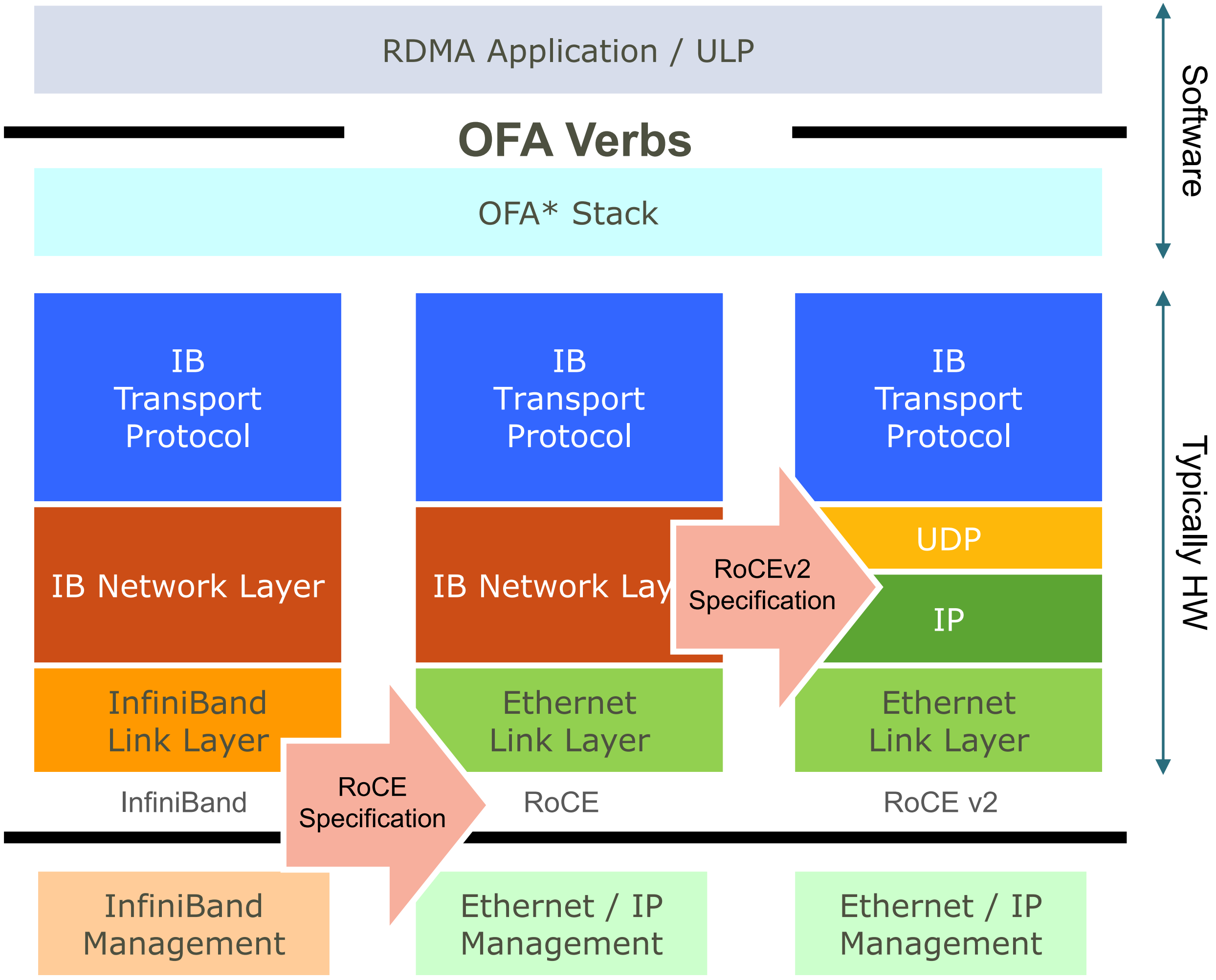
RDMA over Converged Ethernet



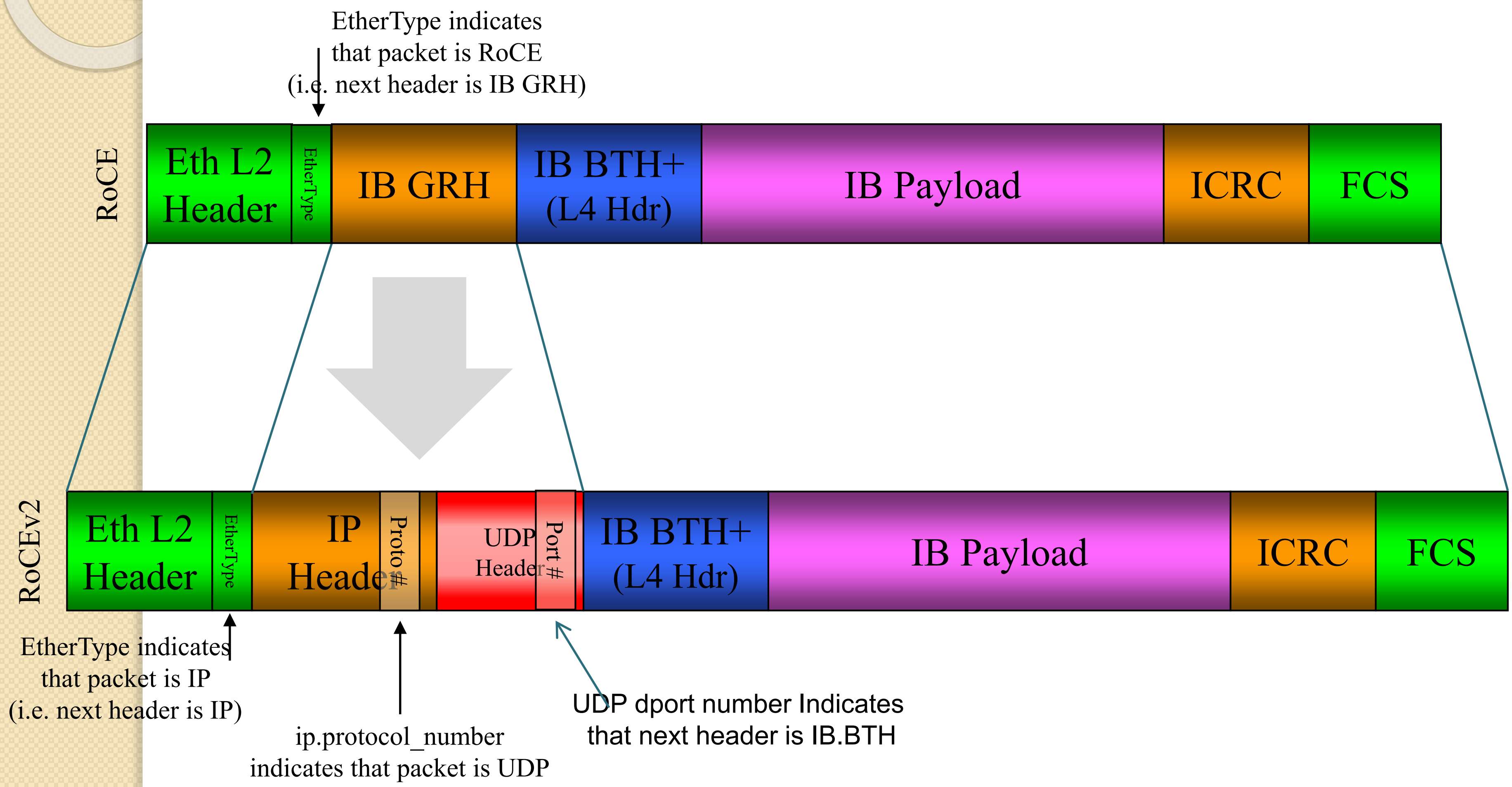
The RoCE Packet Format



RoCEv2 – Extension



RoCEv2 - IP Routable Packet Format



Modern RDMA

- Several major vendors: Qlogic (Infiniband), Mellanox, Intel, Chelsio, others

- RDMA has evolved from the U/Net approach to have three "modes"
  - Infiniband (Qlogic PSM API): one-sided, no "connection setup"
  - More standard: "qpair" on each side, plus a binding mechanism (one queue is for the sends, or receives, and the other is for sensing completions)
  - One-sided RDMA: after some setup, allows one side to read or write to the memory managed by the other side, but pre-permission is required
  - RDMA + VLAN: needed in data centers with multitenancy

Software RDMA Drivers

- Softiwarp
  - www.zurich.ibm.com/sys/rdma
  - open source kernel module that implements iWARP protocols on top of ordinary kernel TCP sockets
  - interoperates with hardware iWARP at other end of wire

- Soft RoCE
  - www.systemfabricworks.com/downloads/roce
  - open source IB transport and network layers in software over ordinary Ethernet
  - interoperates with hardware RoCE at other end of wire