STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

RDMA on MANA

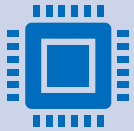
Microsoft Azure Network Adapter

Ajay Sharma
Long Li

Agenda

- RDMA Background
- MANA Overview
- DPDK over MANA
- RDMA over MANA
- Current State

RDMA Motivation



Modern datacenter applications demand high throughput and low latency.

Standard TCP/IP stacks cannot meet these requirements

Remote Direct Memory Access (RDMA) saves network stack overhead

RDMA Benefits

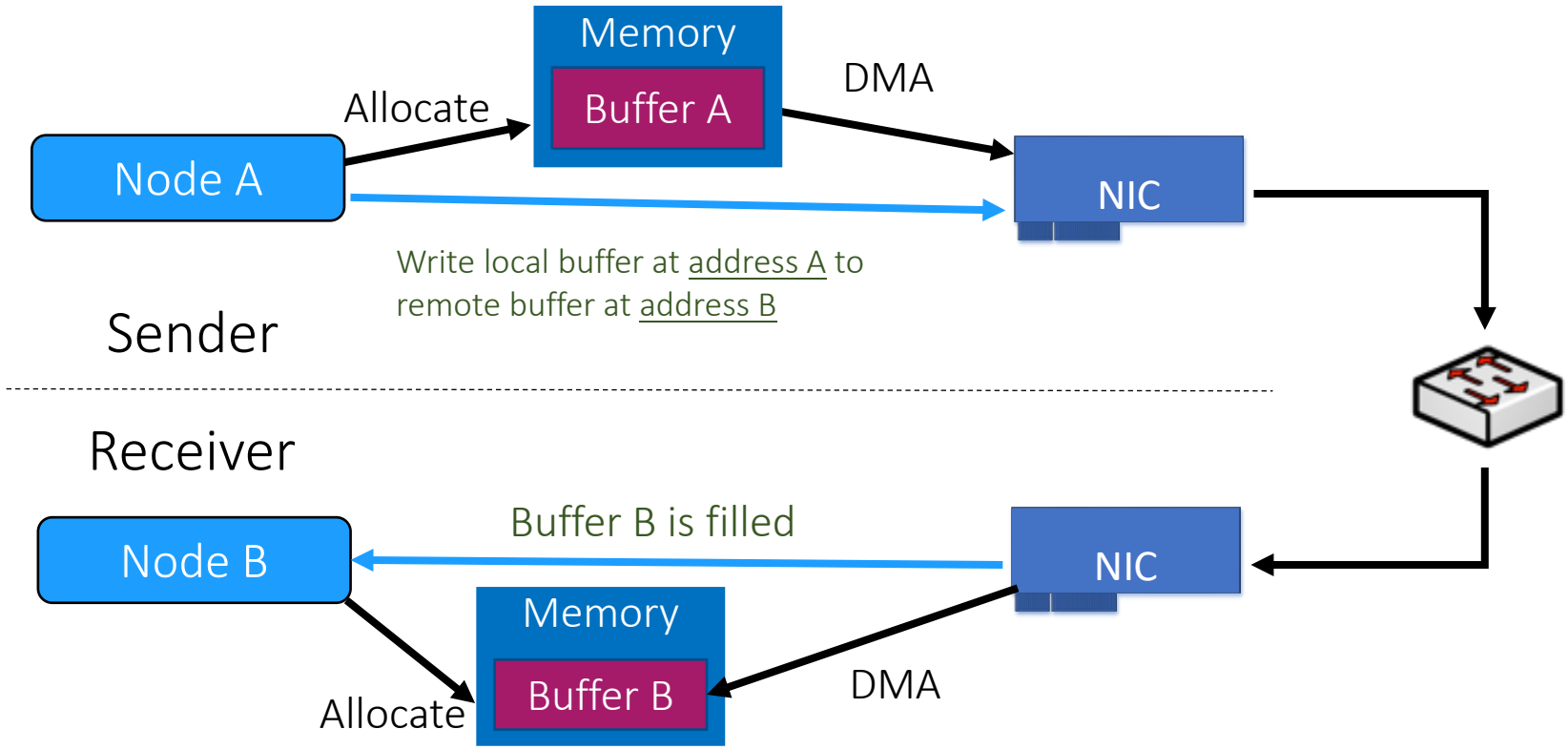
Improved tail latency for storage IO

- Eliminate Network latency
  - Remove TCP overhead, data copies, HTTP formatting
- Switches and NICs form a lossless Fabric
  - Hop-by-Hop backpressure-based flow control
  - Eliminates retransmission due to congestion drops
- DCQCN prevents queue buildup in the switches
  - ECN + E2E congestion control implemented in the NIC

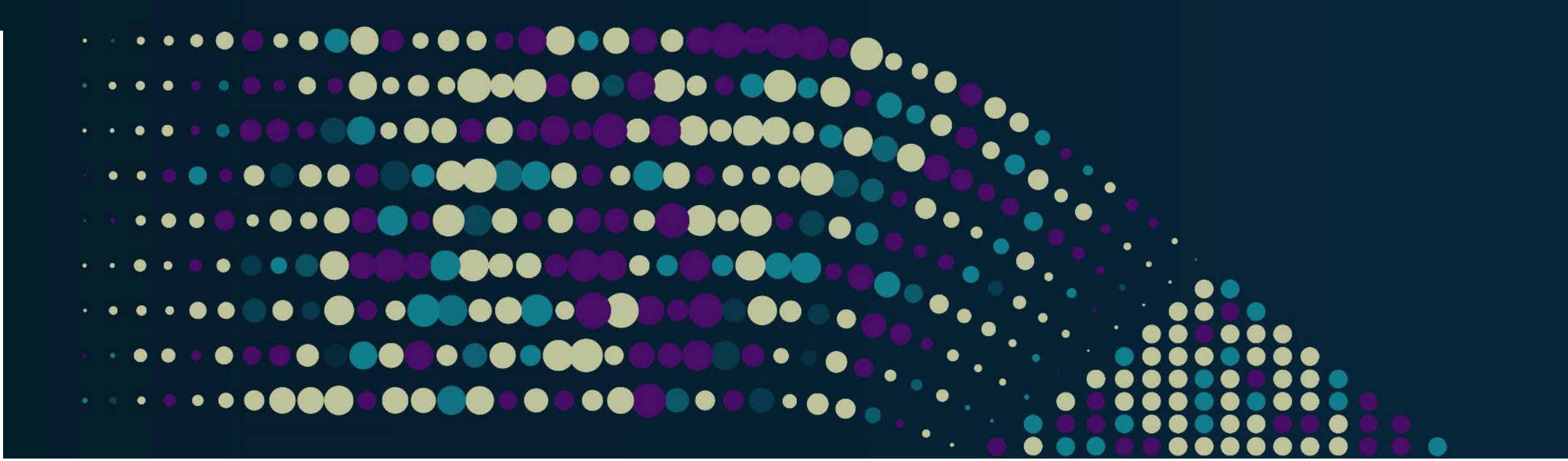
CPU savings

- No CPU involvement in the READ and WRITE ops
  - Reduces CPU load on storage nodes

RDMA: Remote Direct Memory Access



RDMA bypasses host OS stack → frees host CPU, lowers latency



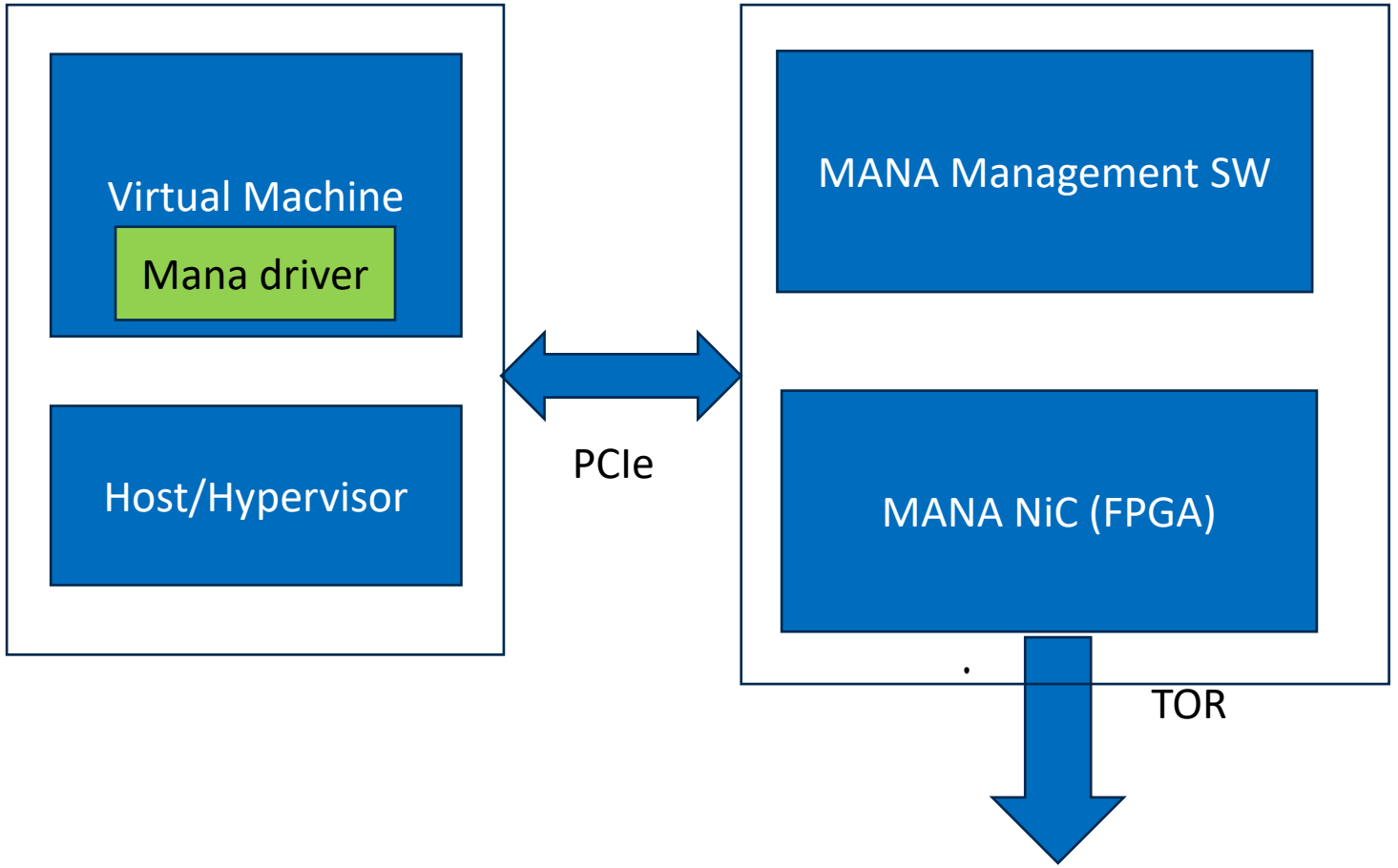
Introducing MANA

Microsoft Azure Network Adapter (MANA)

- **M**icrosoft **A**zure **N**etwork **A**dapter (MANA) leverages both the latest and future hardware acceleration features in Azure and provide competitive performance.
- MANA provides performance, availability, extensibility, and servicing features critical to the ever-evolving cloud landscape.
- MANA is designed with RDMA performance and quality in mind - enabling customers to achieve low latency and high throughput required for their workloads.
- MANA is implemented in FPGA RTL - furthers Microsoft's investments into FPGA technology

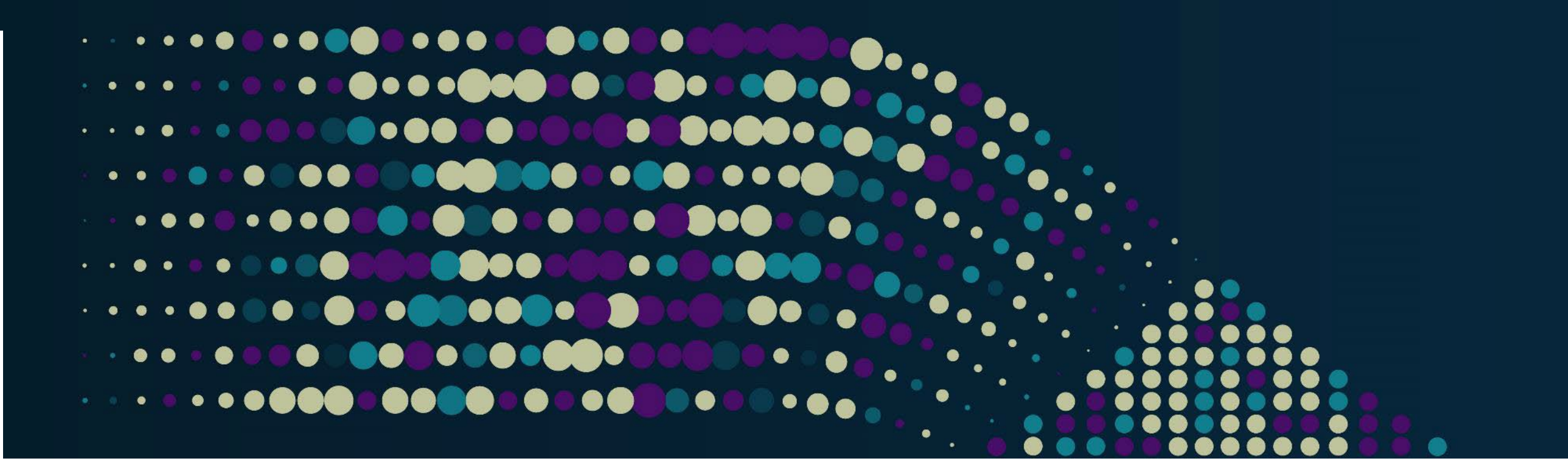
- P.S. MANA is now available for customer to test-drive as part of the Azure Boost Preview

MANA Components



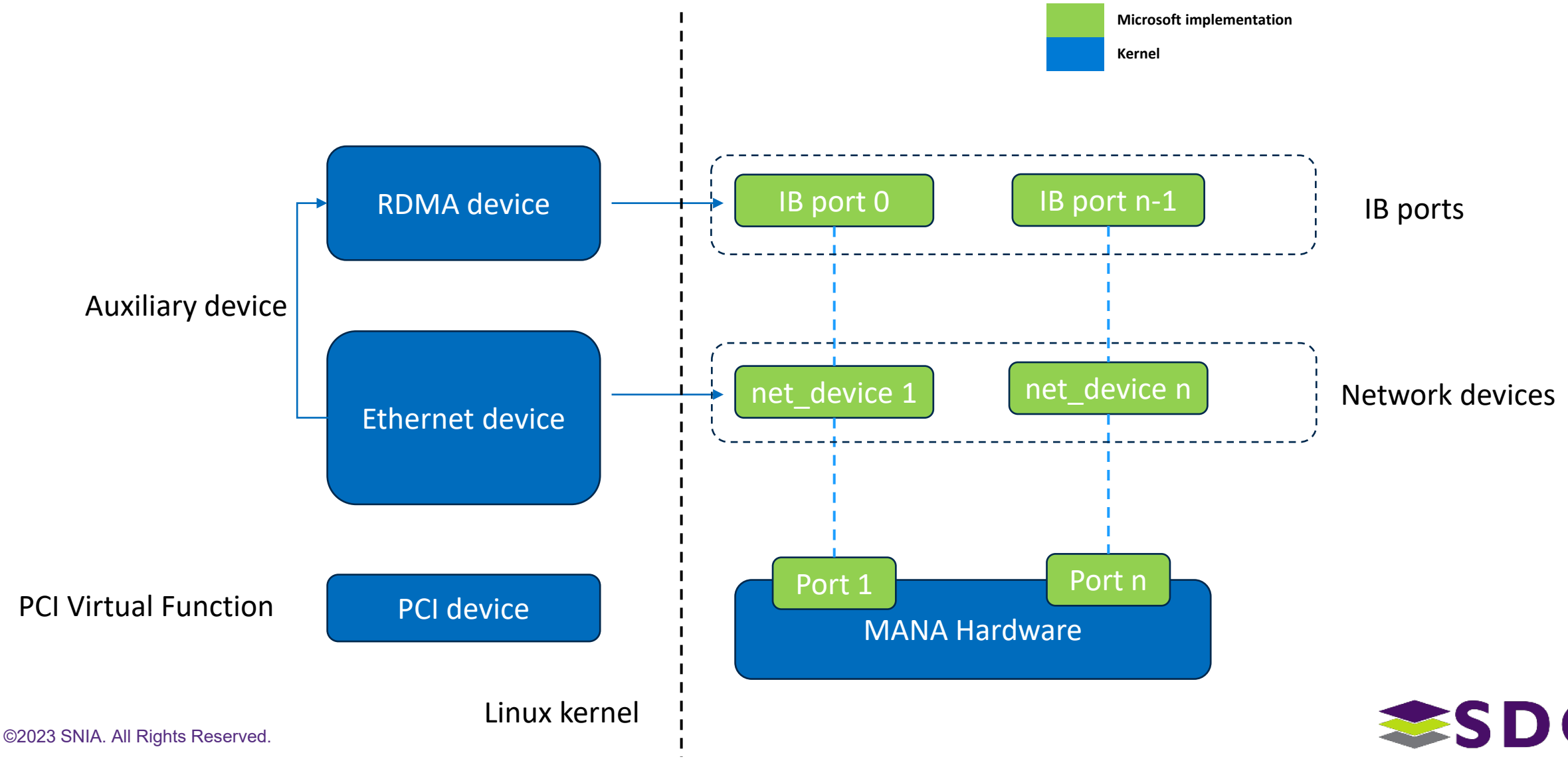
Linux MANA RDMA Driver

- Modeled as an auxiliary device to Linux MANA Ethernet driver
  - Hardware is exposed as a PCI device
  - Support multiple network devices over one PCI function
  - Each network device can optionally expose an RDMA port
- RoCE v2 RDMA only support
- Support two types of queue pairs
  - RAW – used to expose native device queue to user-mode
    - Used by DPDK
  - RC – reliable connection
    - Support CM verbs
    - RC queue pair

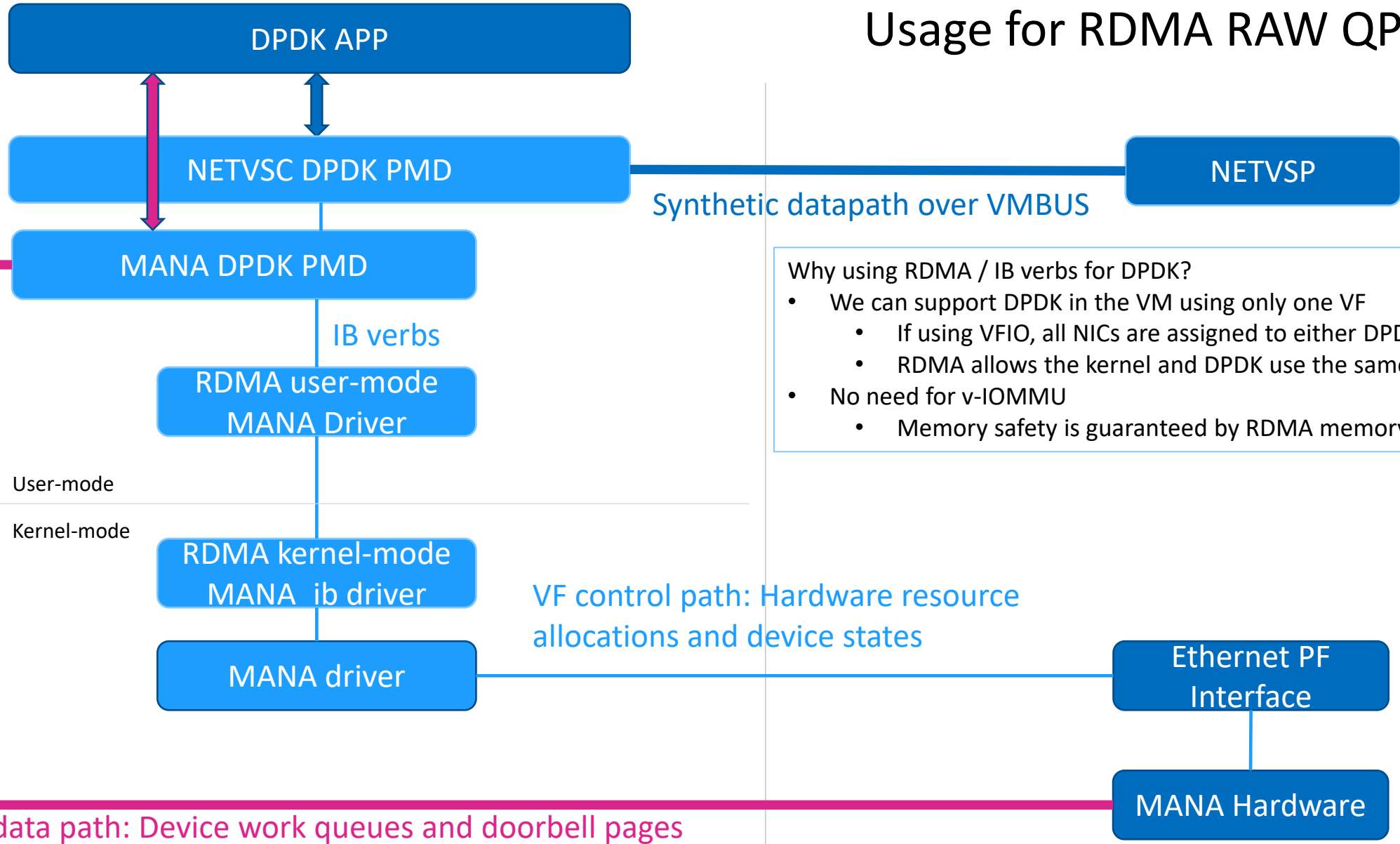


DPDK over MANA

Linux MANA RDMA Device Model



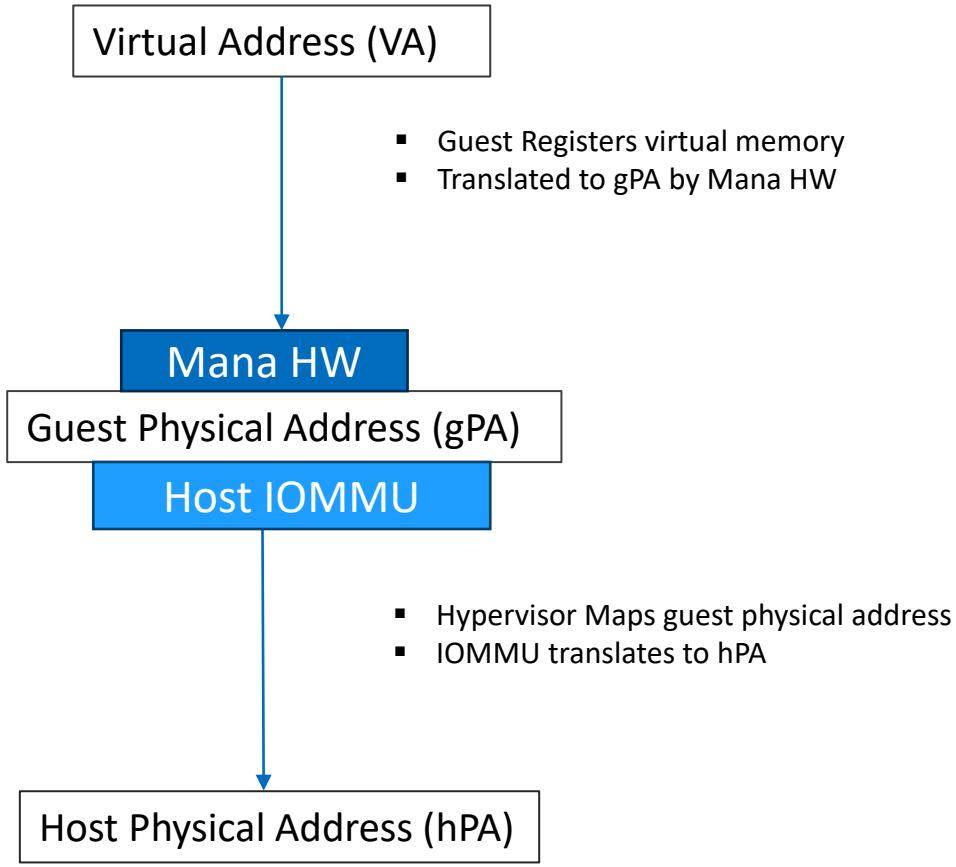
Usage for RDMA RAW QP for DPDK



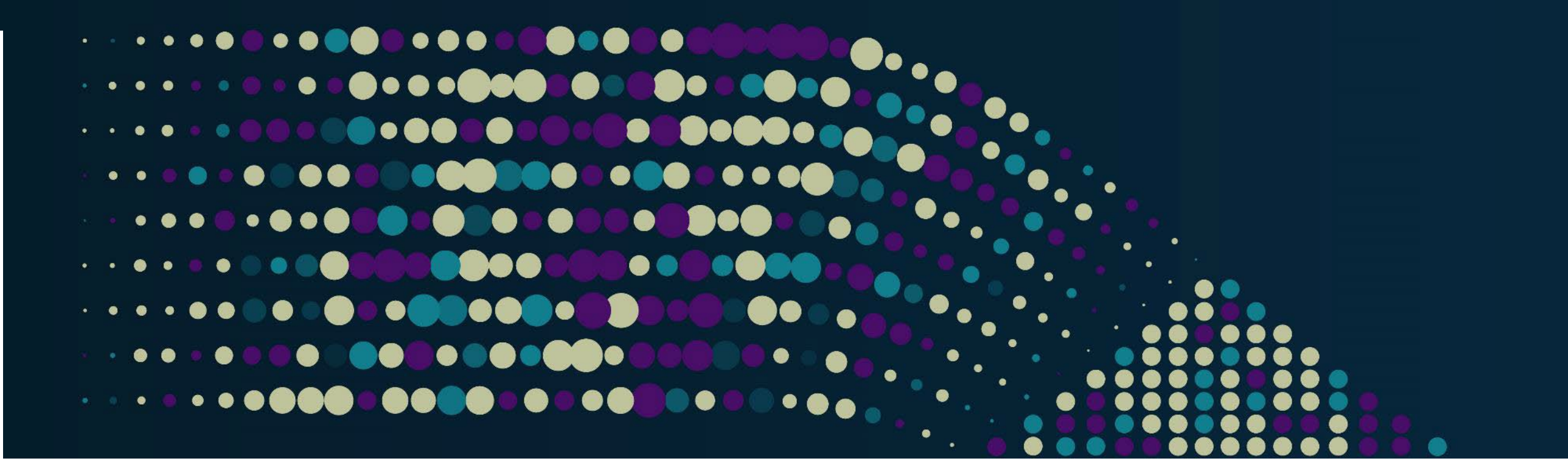
Why using RDMA / IB verbs for DPDK?

- We can support DPDK in the VM using only one VF
  - If using VFIO, all NICs are assigned to either DPDK or kernel but not both
  - RDMA allows the kernel and DPDK use the same VF
- No need for v-IOMMU
  - Memory safety is guaranteed by RDMA memory registration

Memory Registration



- Virtual IOMMU for the guest is not required.
- viommu very expensive to implement in the Hypervisor
- Inter process memory safety is guaranteed by the memory registration.
- It's essential for the container workload.

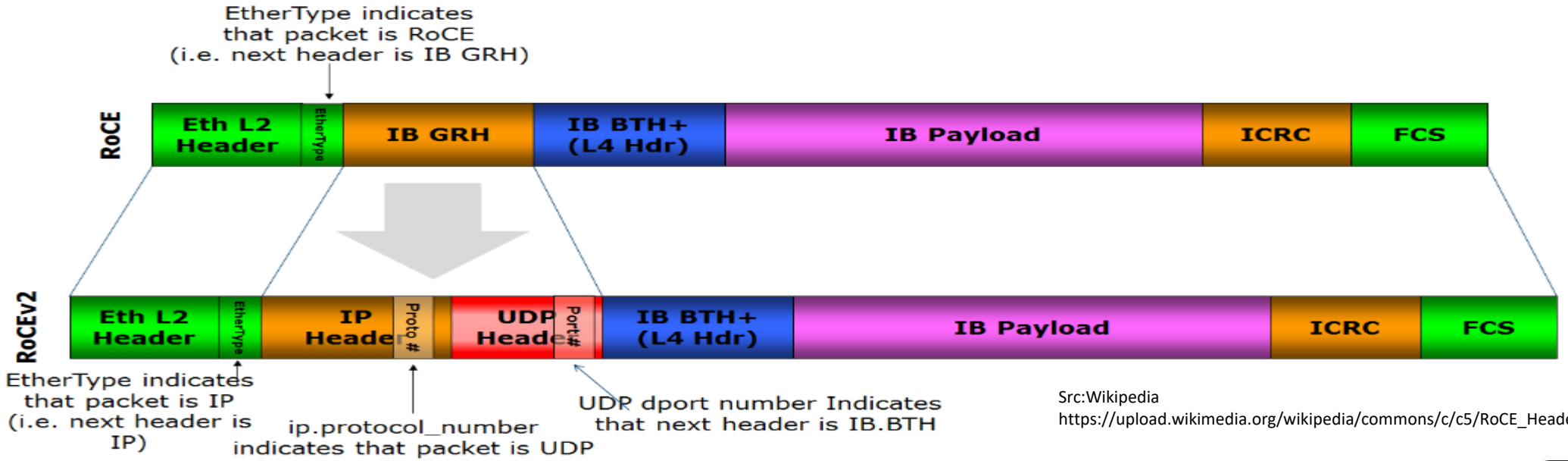


RDMA over MANA

RoCEv2 Packet Format

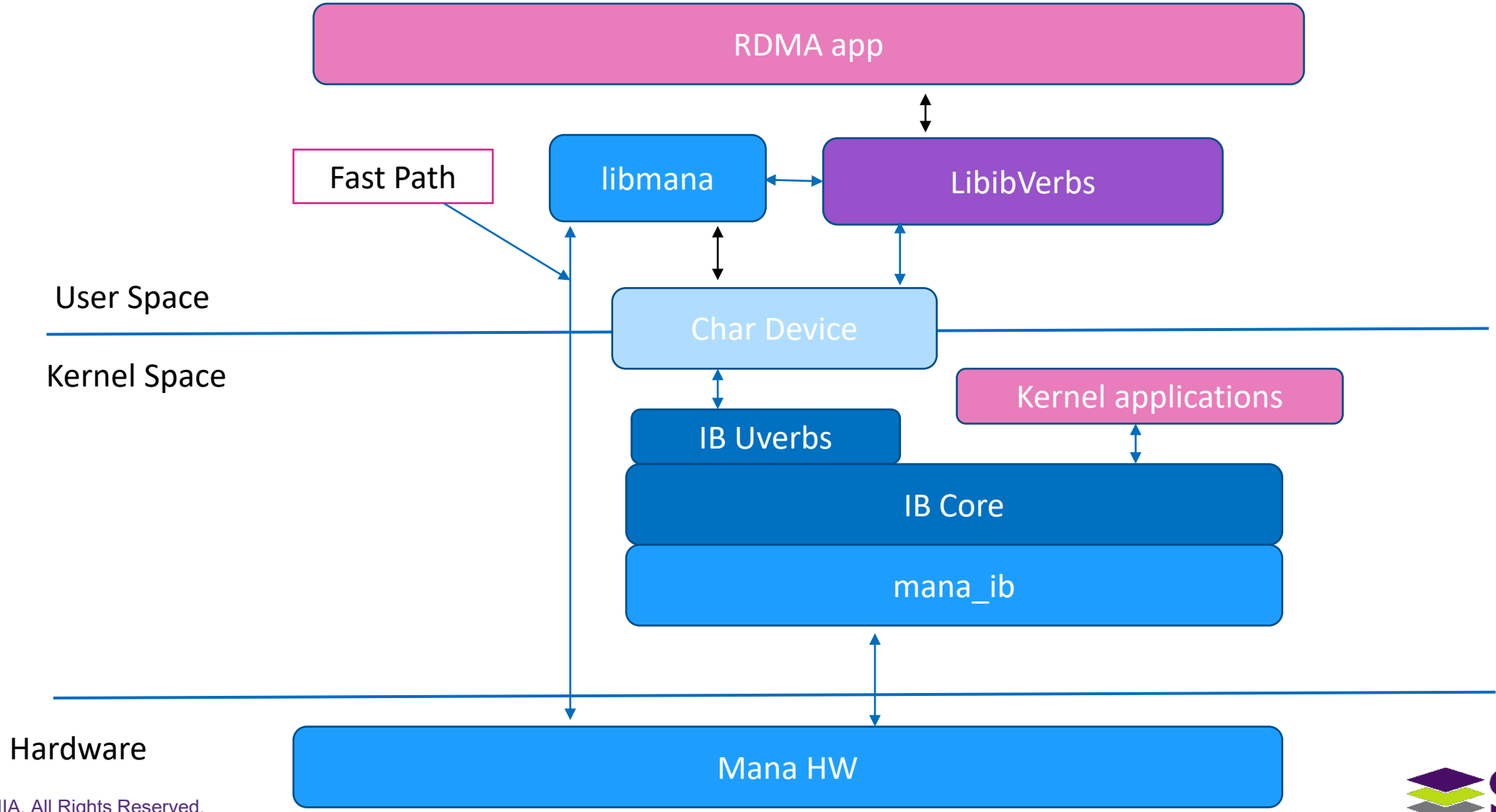
The IP/UDP header are used to:

1. Route the packet to the correct node
2. Indicate this is an RDMA packet using UDP protocol number 4791
3. Indicate RDMA packet length (BTH don't have packet length field)



Src:Wikipedia
https://upload.wikimedia.org/wikipedia/commons/c/c5/RoCE_Header_format.png

Microsoft ibverbs implementation



Acknowledgments

Marina Lipshteyn

Mahmoud Elhaddad

Matt Reat

Murtuza Naguthanwala

Siri Velauthapillai

