

ASAP2
Accelerated Switching & Packet Processing

Mellanox
DPDK



Rivermax VMA Tech
Broader Flow
Narrow CPU Utilization



Hardware Offloading To RDMA and Beyond

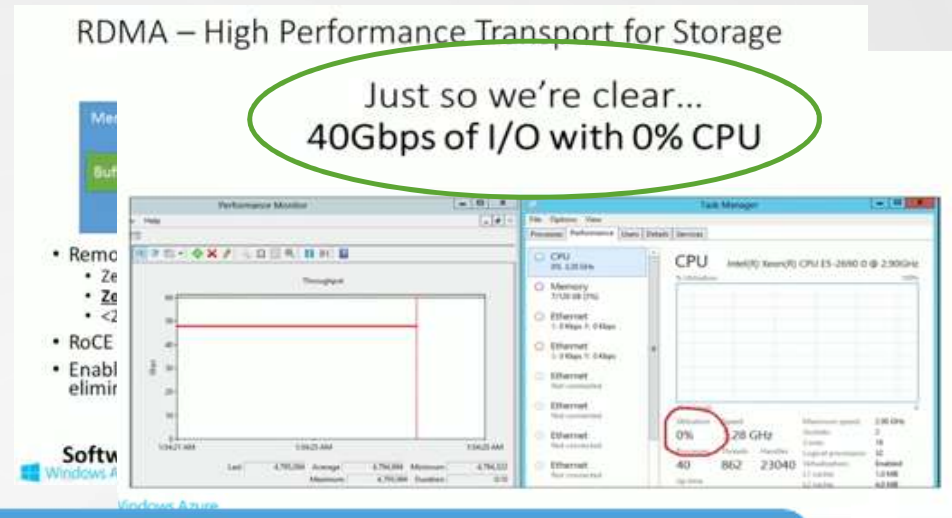
KBNets Key Note 2018

Aug 20th 2018, Eitan Zahavi

RDMA Enables Hyperscale Clouds Storage

“To make storage cheaper we use lots more network!
 How do we make Azure Storage scale?
 RoCE enabled at 40GbE for Windows Azure Storage,
 achieving massive COGS savings”

Microsoft Keynote, Albert Greenberg, SDN in Azure Infrastructure @ ONS14



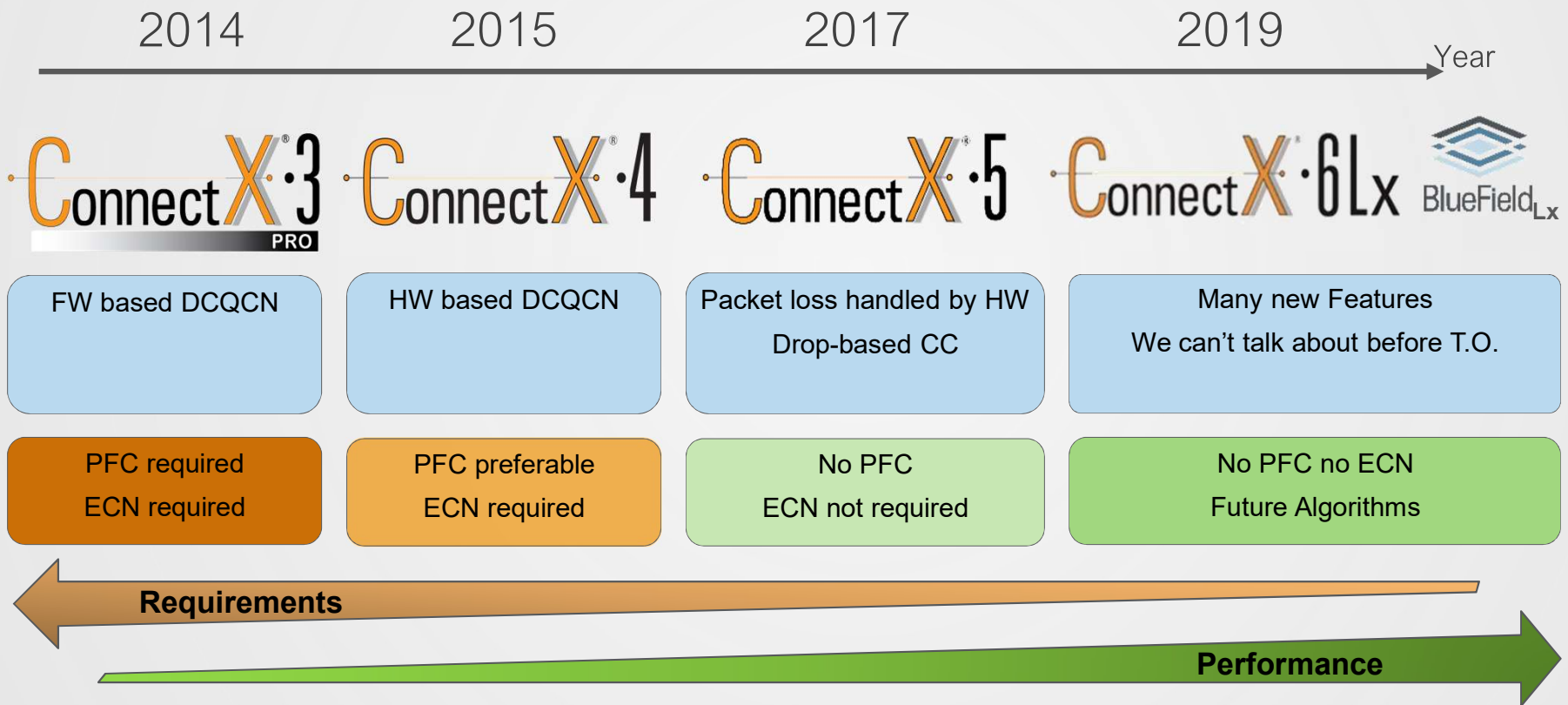
“网络比同行高出26倍，网络转发能力高出8倍。最大内网带宽为3倍。存储方面以I1为例，随机IOPS达到同行的2倍，存储性能高出6倍”



The Network is 26 times more powerful than competitors, and its forwarding capacity is 8 times higher. The maximum internal network bandwidth is 3 times. In terms of storage, for example, I1, random IOPS are twice as high as competitors, and storage performance is 6 times higher



RoCE Congestion Control Roadmap



Network Acceleration Technologies Beyond RDMA



Network Acceleration Technologies Beyond RDMA



What is Mellanox VMA?

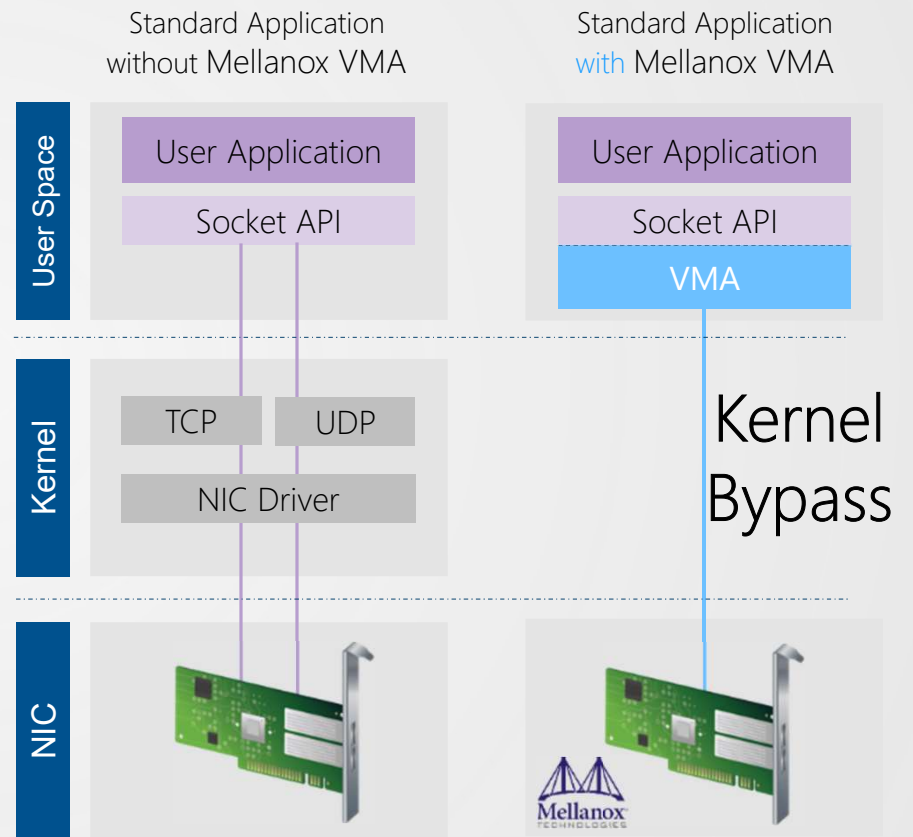
KERNEL BYPASS

Reduce Kernel overhead with direct network adapter access



SINGLE SIDED

Requires no application changes – Standard sockets TCP, UDP (Unicast, Multicast)



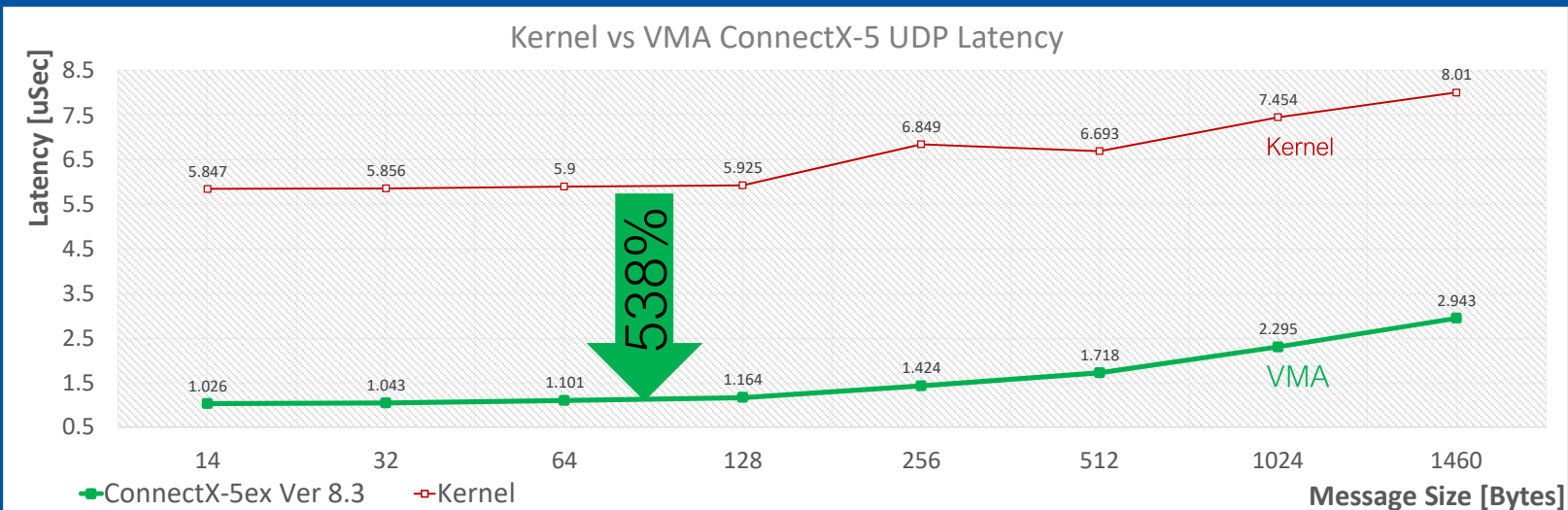
VMA the Lowest Latency Networking Stack



Mark Russinovich "Inside Azure Datacenter Architecture" @ Microsoft Build 2018 :

PTP UDP Ping Latency

VM results: 190uSec → Azure FPGA: 28uSec → VMA on Azure: 6uSec

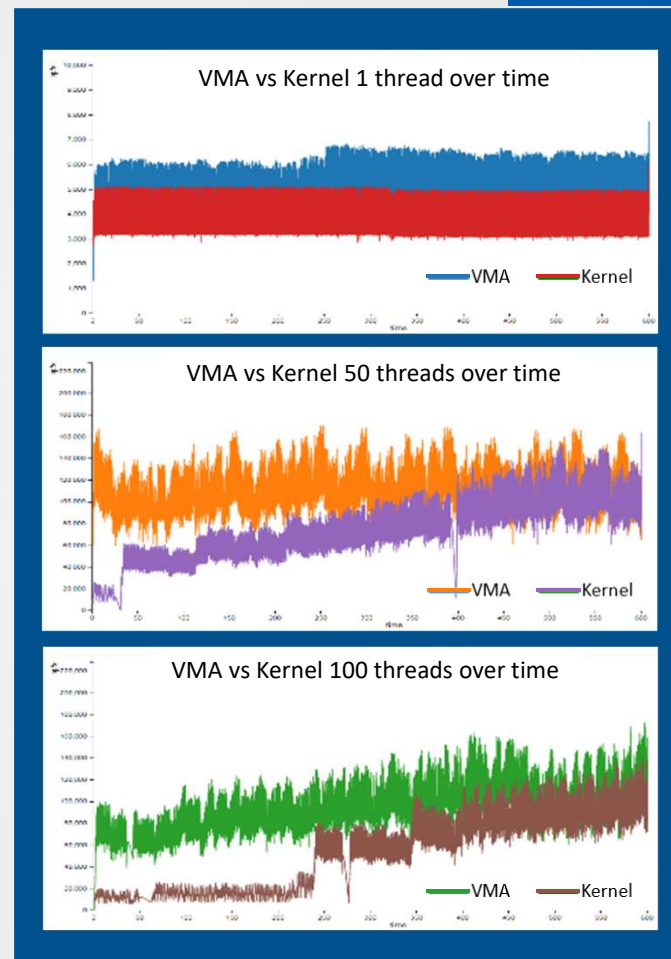
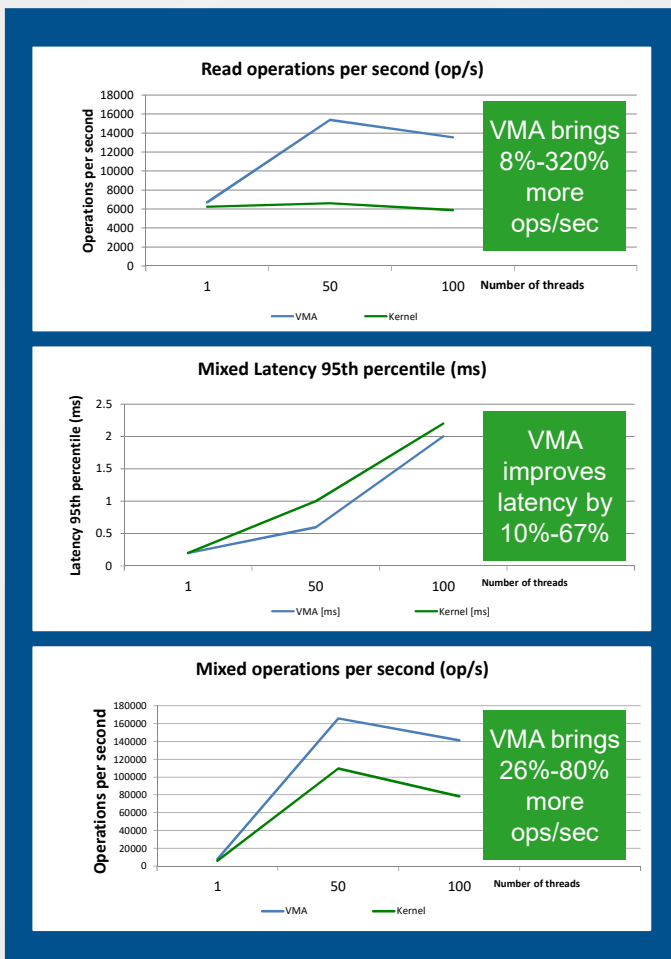


VMA Speeds Up Apache Cassandra



- ConnectX-4 VMA @ 10GbE
- Read:
 - 320% op/sec @ 50 threads
- Latency 95th percentile
 - 67% less @ 50 threads

Cassandra with Mellanox ConnectX-4 10GbE running with VMA brings 8%-80% better performance

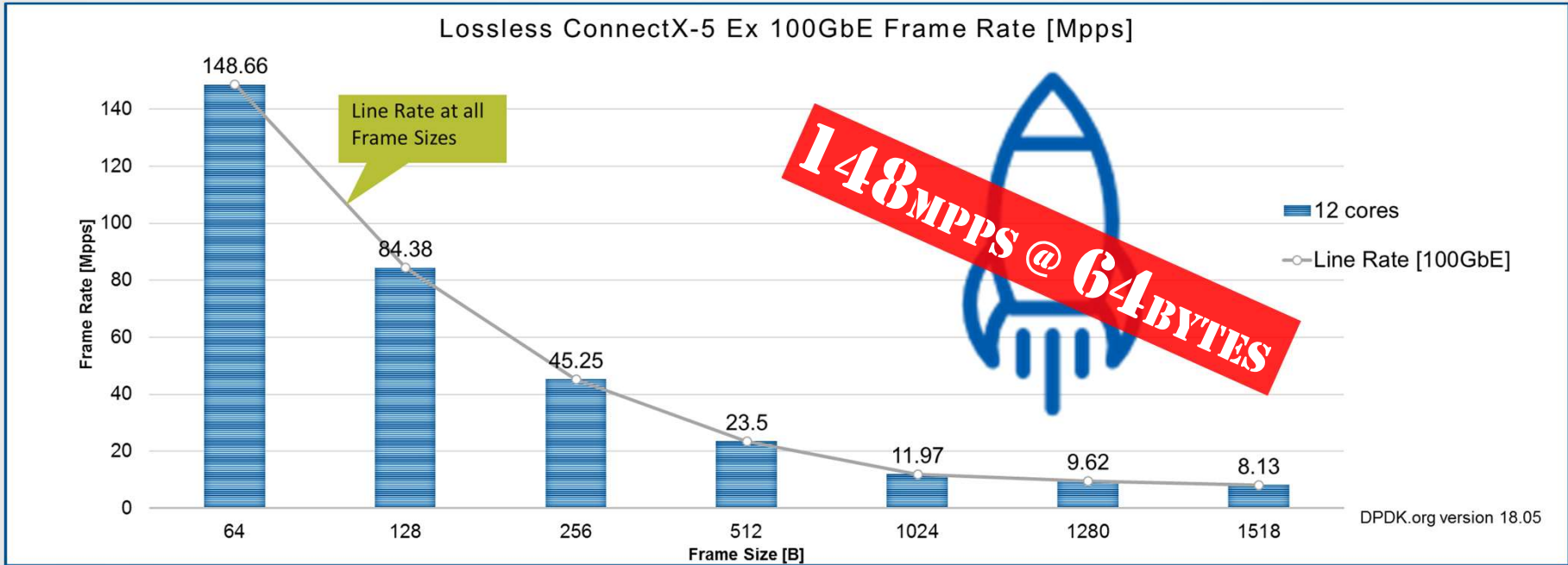


DPDK with Mellanox - Industry Leading Performance

Highest Performance and Message Rate in the Market!!!

66% lower latency compared to competition

DPDK with Mellanox



DPDK Single Core Performance - dpdk.org report

- Mellanox ConnectX-5 uses 38 CPU-cycles per packet
- Intel i40e uses 51 CPU cycles per packet

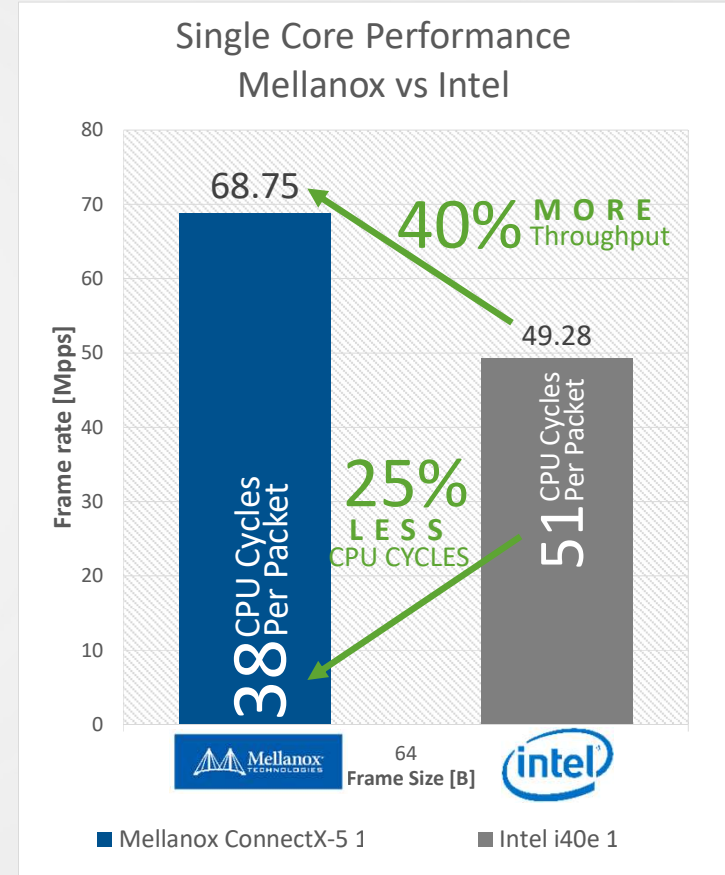
40% MORE THROUGHPUT

25% LESS CPU CYCLES

Save CPU Cores
Save CAPEX and OPEX

Save Power
Reduce Heat

Higher throughput with Less Cores

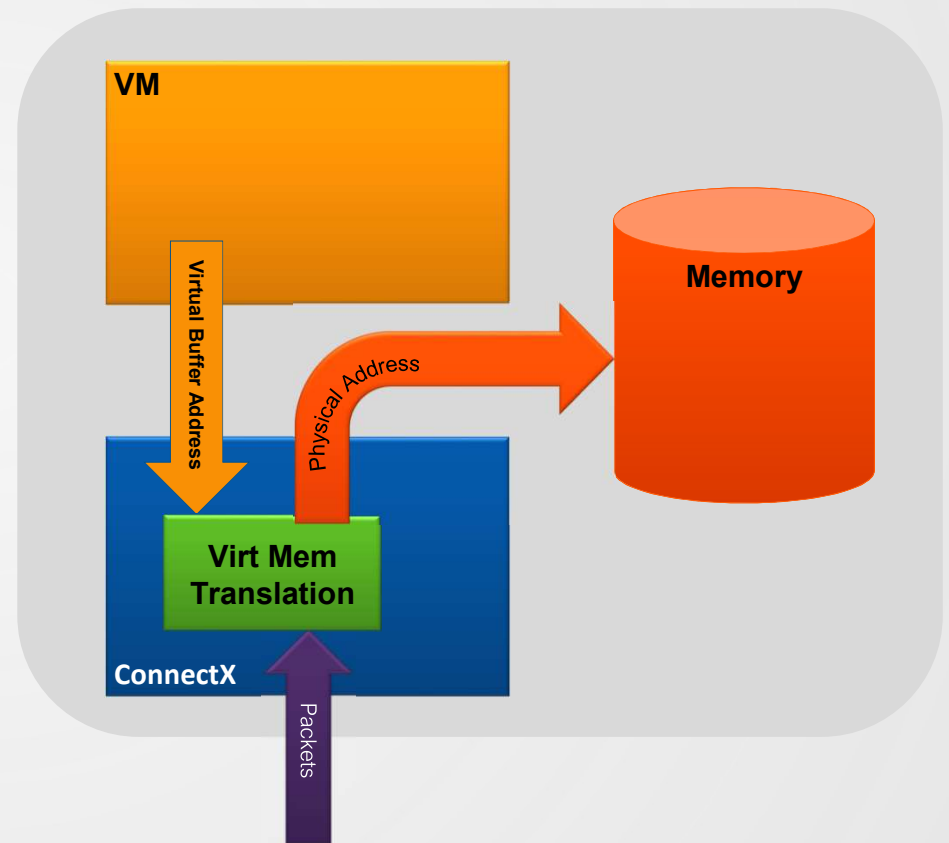


Source - official dpdk.org performance reports: [Intel Perf Report](#) [Mellanox Perf Report](#)

Memory Translation for DPDK / VMA

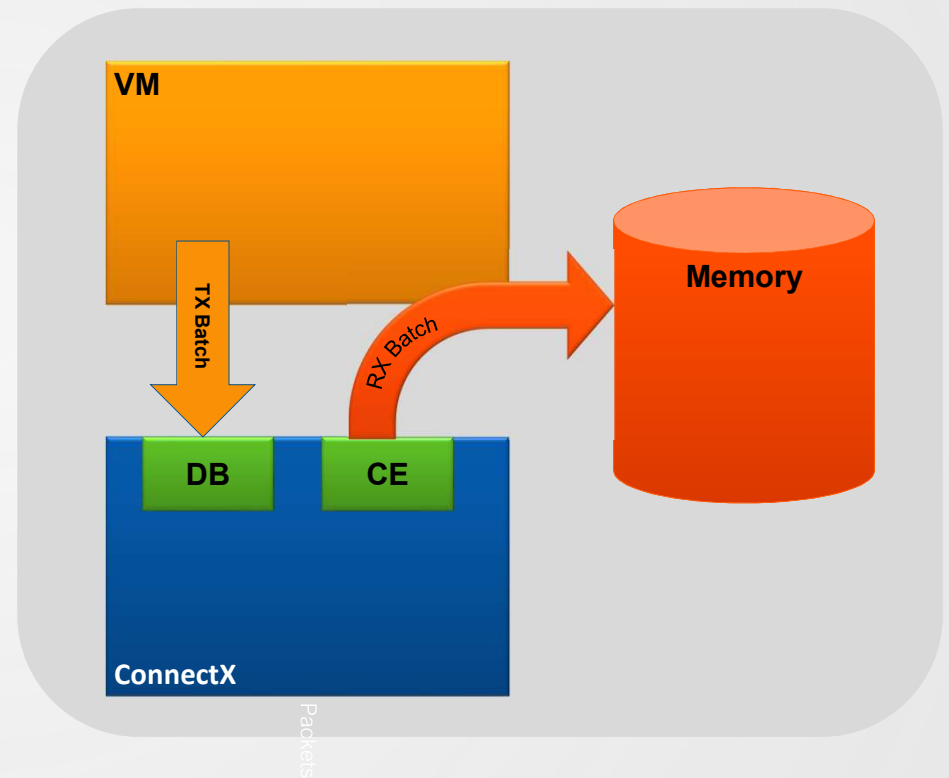
- Receive directly to VM memory requires address translation
- But at 150 MPPS @ 100Gbps
- What about 200Gbps, 400Gbps...

- Only the hardware can do it!

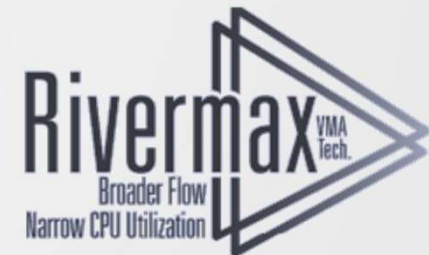


Efficient Software Interface DPDK / VMA

- The impact of batch processing was clearly demonstrated with Vector Packet Processing (VPP)
- To get great performance it also relies on hardware support for
 - Posting batches of sent packets to the device
 - Reporting arrival of batches of packet to the host/VM
- The hardware modules enabling batches are now even more critical as their latency and capacity is required to increase by batch size

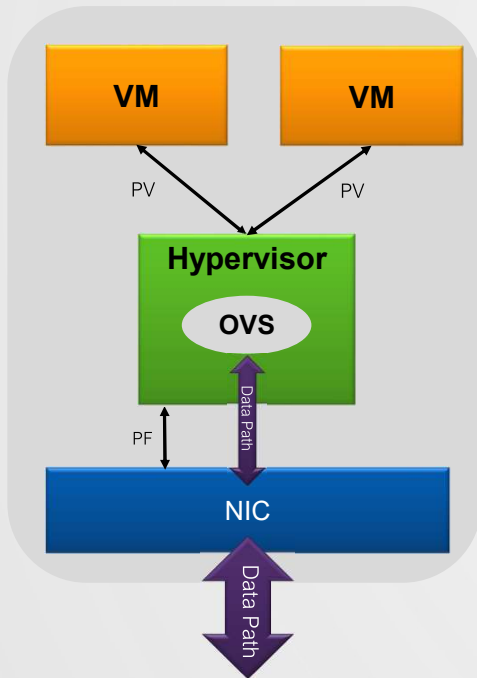


Network Acceleration Technologies Beyond RDMA



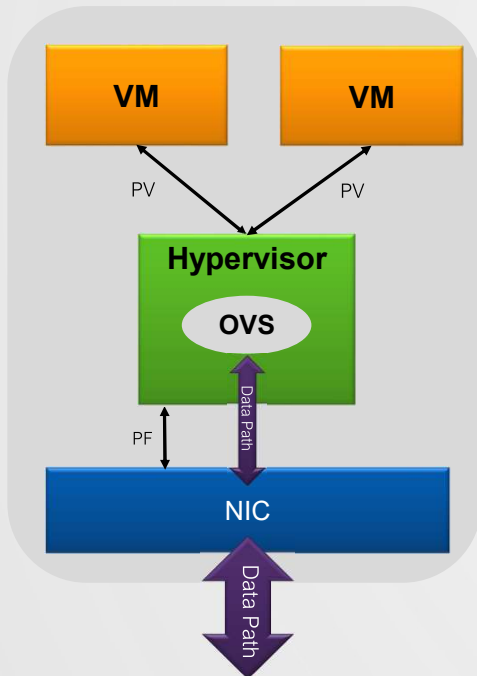
Network Virtualization Offloading

Hypervisor Software
vSwitch

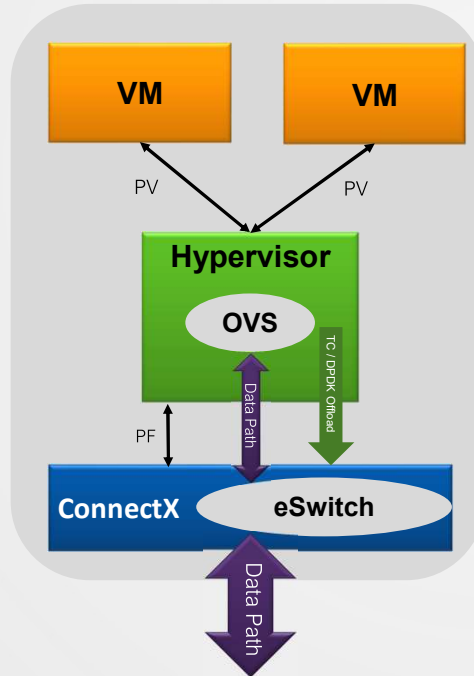


Network Virtualization Offloading

Hypervisor Software
vSwitch

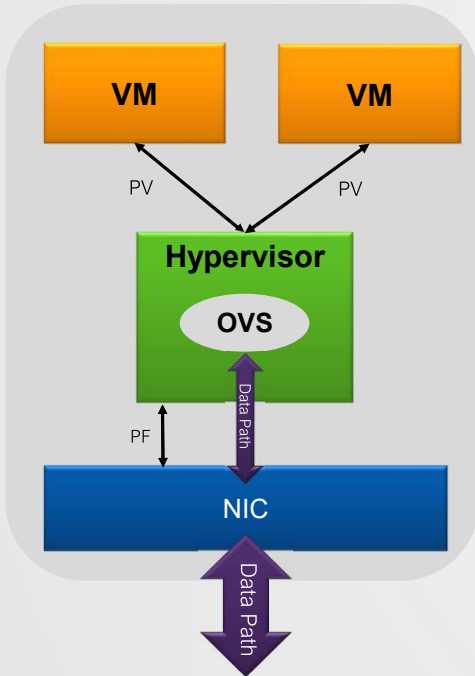


ASAP² Flex
vSwitch acceleration

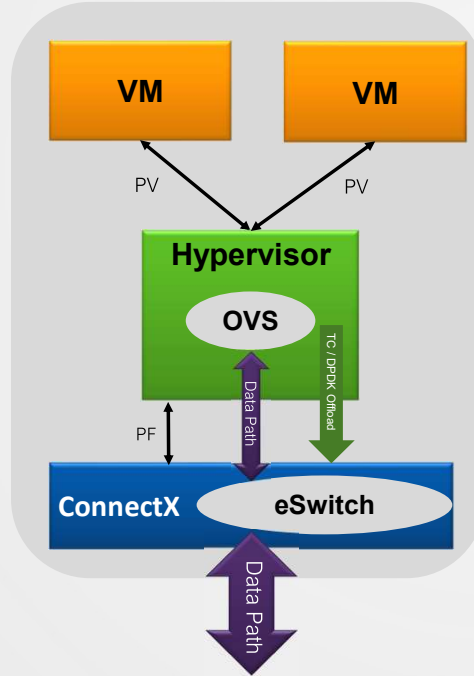


Network Virtualization Offloading

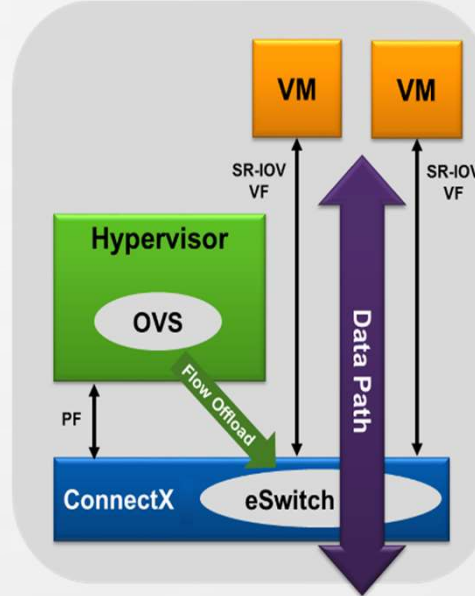
Hypervisor **Software**
vSwitch



ASAP² **Flex**
vSwitch acceleration

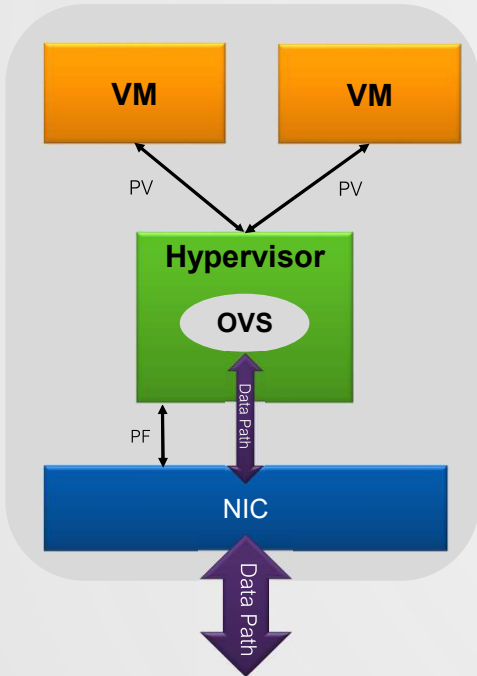


ASAP² **Direct**
Full vSwitch offload

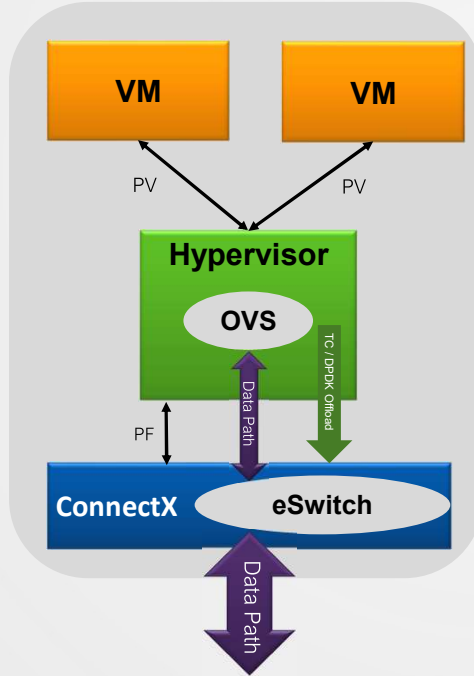


Network Virtualization Offloading

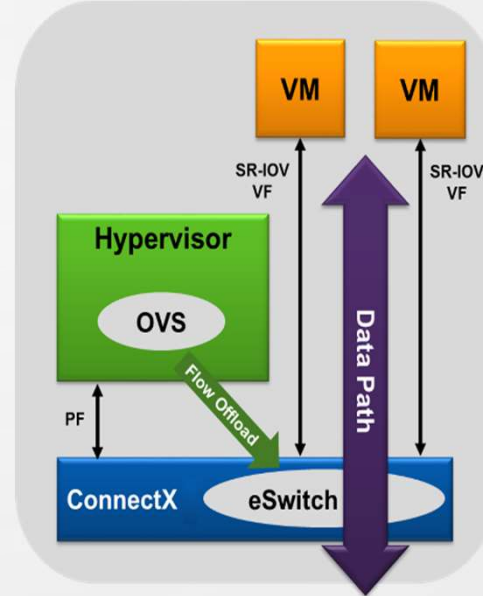
Hypervisor **Software**
vSwitch



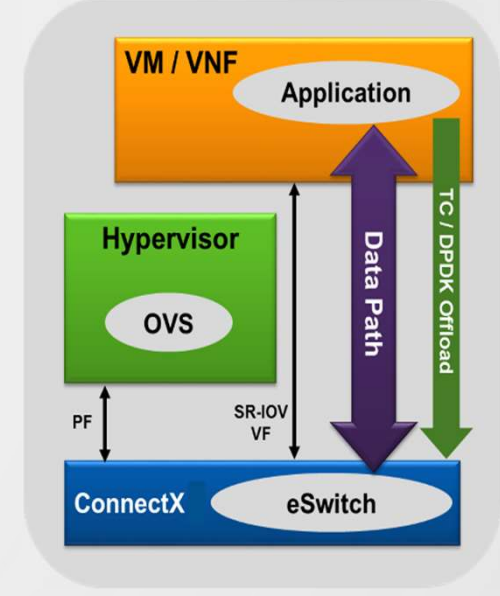
ASAP² **Flex**
vSwitch acceleration



ASAP² **Direct**
Full vSwitch offload



ASAP² **NFV**
VNF/VM acceleration

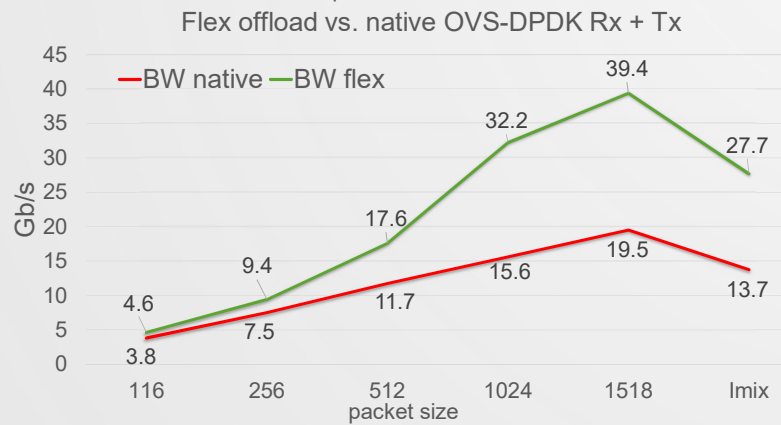
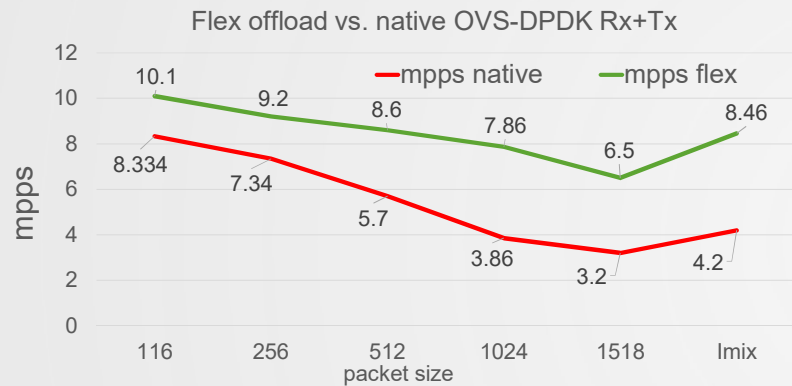


ASAP² Performance



ASAP² Flex

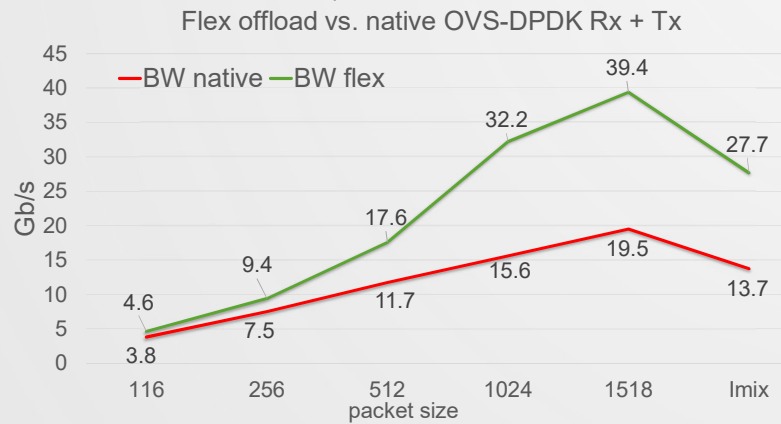
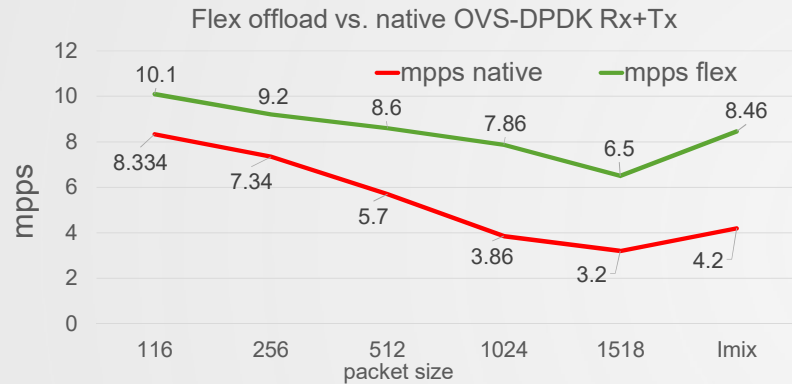
Classification test



ASAP² Performance

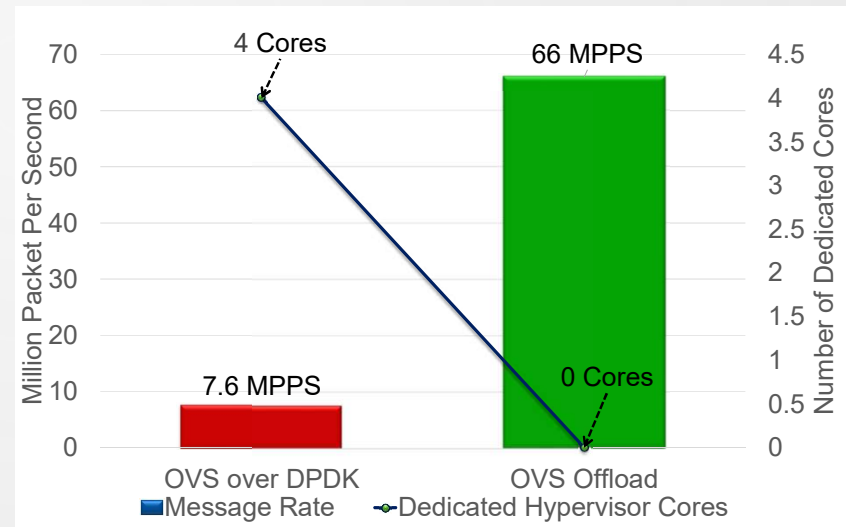
ASAP² Flex

Classification test



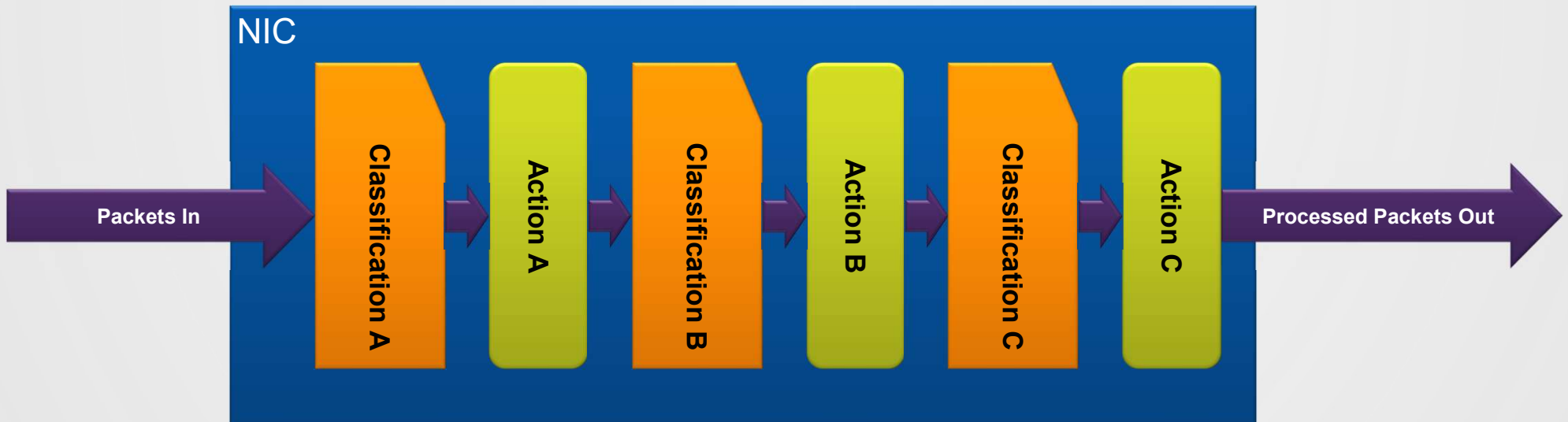
ASAP² Direct

Test	ASAP2 Direct	OVS DPDK	Benefit
1 Flow VXLAN	66M PPS	7.6M PPS (VLAN)	8.6X
60K flows VXLAN	19.8M PPS	1.9M PPS	10.4X



Offloading eSwitch to the NIC

- Most network functions share some data-path operations
 - Packet classification (into flows)
 - Action based on the classification result
- Mellanox NIC has the capability to offload both the classification and the actions in hardware



Example ASAP² Classifications

- L2: Ethernet Layer 2
 - Destination MAC
 - 2 outer VLANs / priority
 - Ethertype

- L3: IP (v4 /v6)
 - Source address
 - Destination address
 - Protocol / Next header

- L4: TCP /UDP
 - Source port
 - Destination port
 - TCP flags
 - Connection State

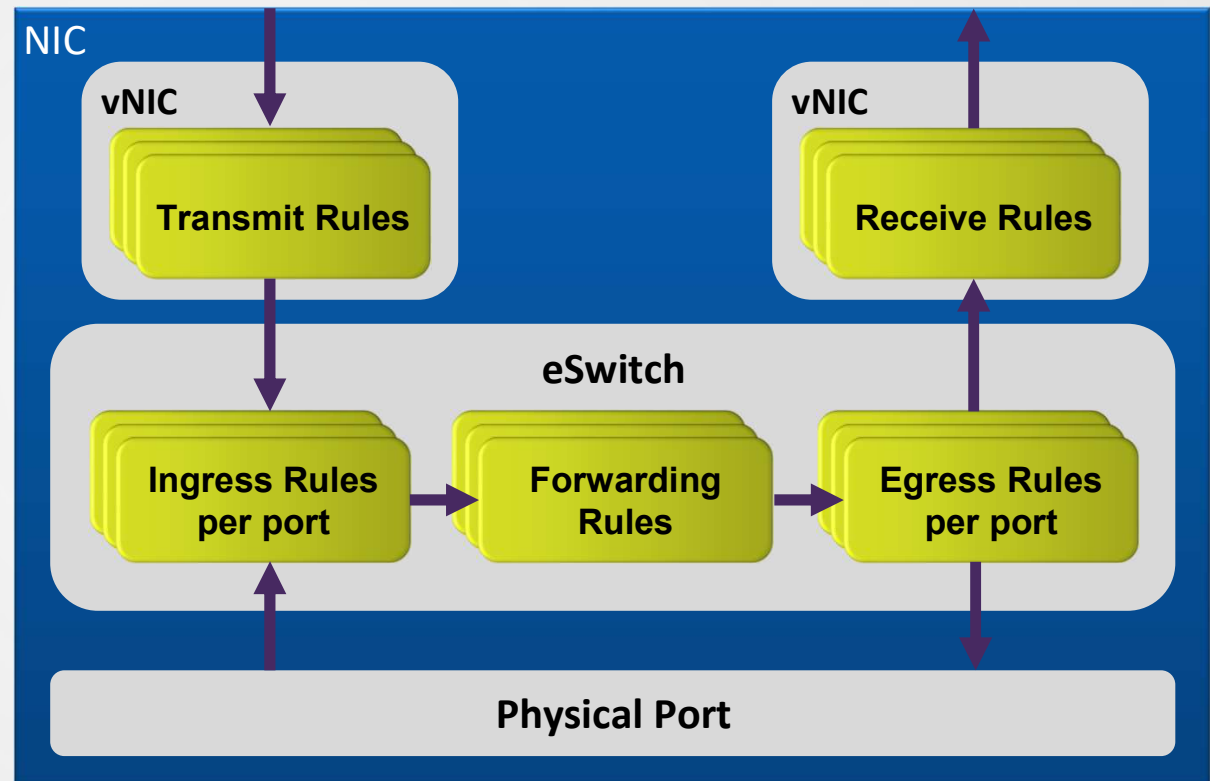
- L5: Tunnel VxLAN/GRE

Classify by Every Packet Header Field



Flow Tables Overview

- Multiple tables
- Programmable table size
- Programmable table cascading
- Dedicate, isolated tables for hypervisor and/or VMs
- Practically unlimited table size
 - Can support million of rules/flows



Network Acceleration Technologies Beyond RDMA

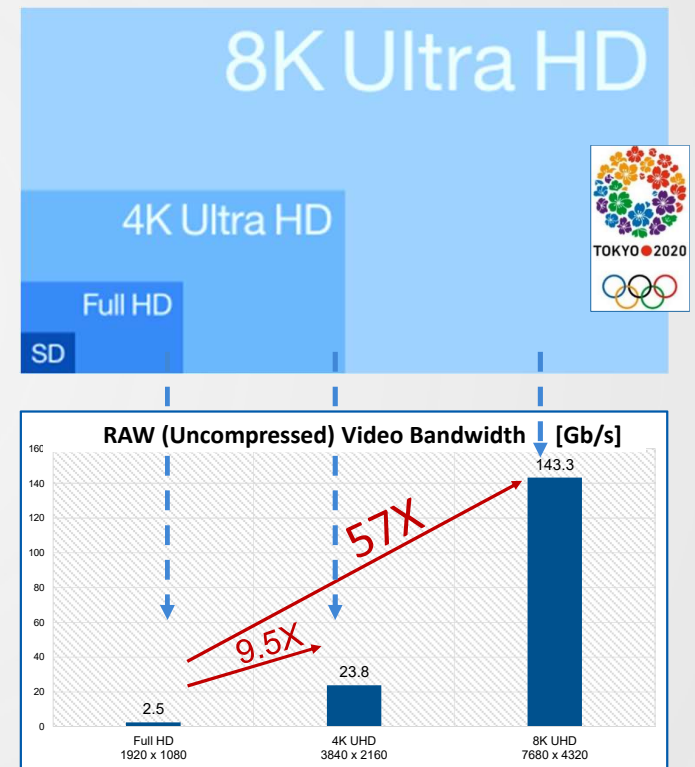


Ultra HD video resolution drives high-bandwidth requirements

- Raw uncompressed video bandwidth is rising
- Bandwidth exceeding 100Gb/s with 8K UHD high frame rate

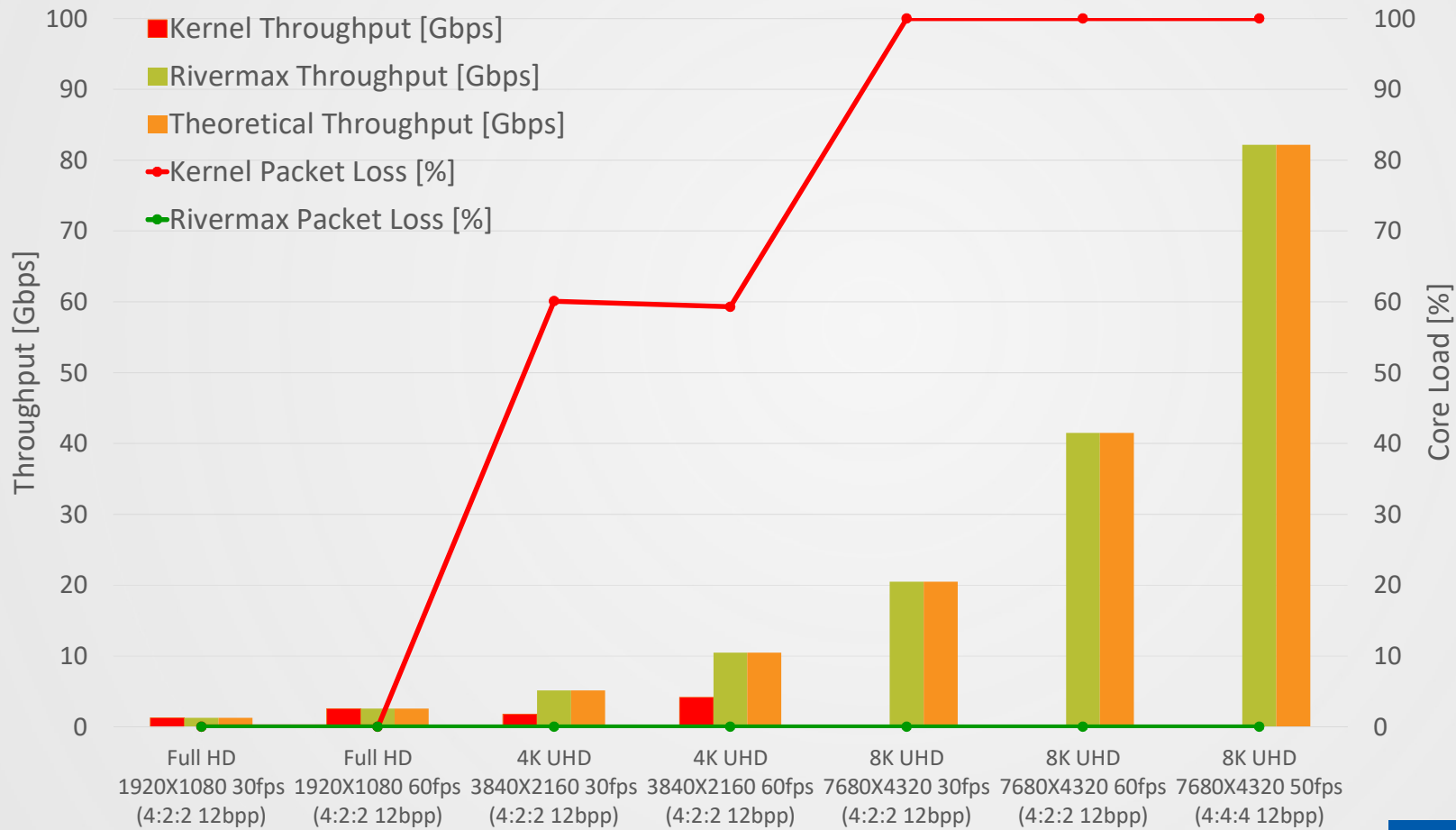
	Bit per pixel (bpp)	Frame rate	Data Rate (Gb/s)	Packets/sec
Full HD 1920X1080	20	25	1.04	90K
	20	30	1.25	108K
	20	50	2.09	180K
	20	60	2.5	216K
4K UHD 3840X2160	20	50	8.3	720K
	20	60	10	865K
	20	100	16.6	1.44M
	20	120	20	1.73M
	24	50	9.96	720K
	24	60	11.95	865K
	24	100	19.92	1.44M
	24	120	23.8	1.73M
8K UHD 7680X4320	20	50	33.2	2.88M
	20	60	40	3.46M
	20	100	66.4	5.76M
	20	120	81	6.92M
	36	50	59.7	2.88M
	36	60	71.7	3.46M
	36	100	119.4	5.76M
	36	120	143.3	6.92M

Over 100Gb/s for a single video stream



Rivermax Single Video Stream Single Core

Linux UDP stack vs Rivermax based Application



Rivermax Key Hardware Features

Packet Pacing

- Leverages ConnectX-5 hardware based Packet Pacing
- SMPTE ST 2110-21 compliance at any bit rate
 - No dependency on CPU Strength, OS interrupt level or Application



Kernel Bypass

- Reduced Kernel overhead with direct network adapter access
- Selective bypass – enables to select traffic bypasses and which flows to kernel
 - Reduced latency
 - Reduced CPU utilization
 - Increased throughput



Application Pkt Placement

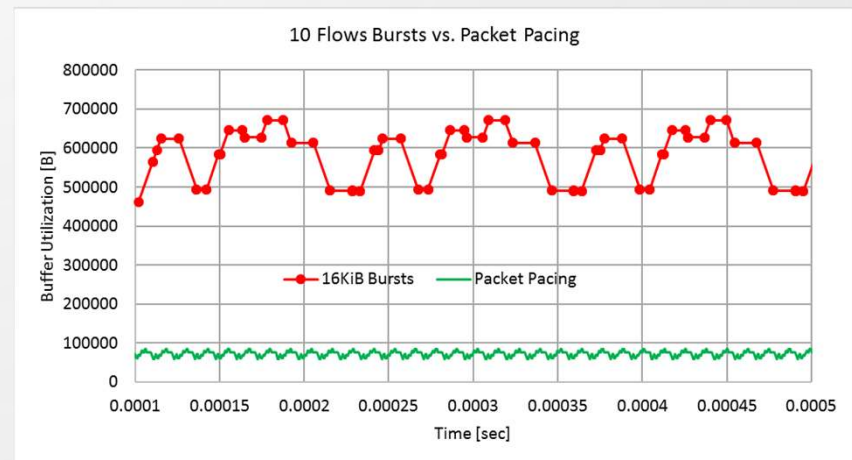
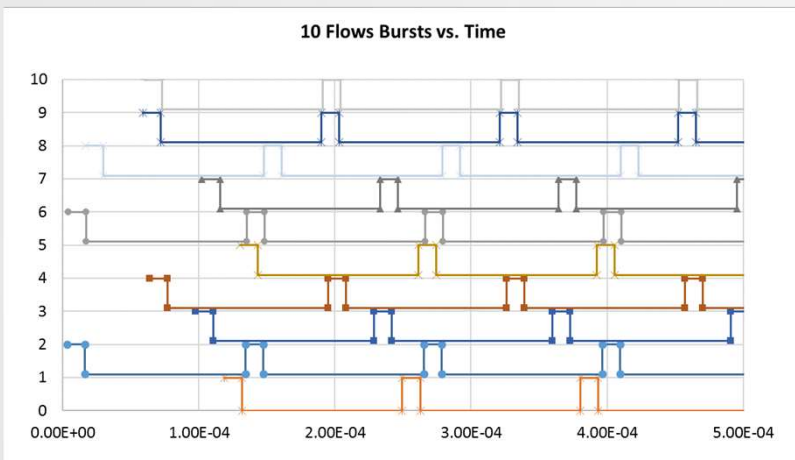
- At Frame/Line(s) level
- Receive: fully assembled frame/lines(s) in memory
- Transmit: synchronously transmit packet paced full frames/lines (/chunks)

Packets vs Frames

Based on ConnectX-5 Technology

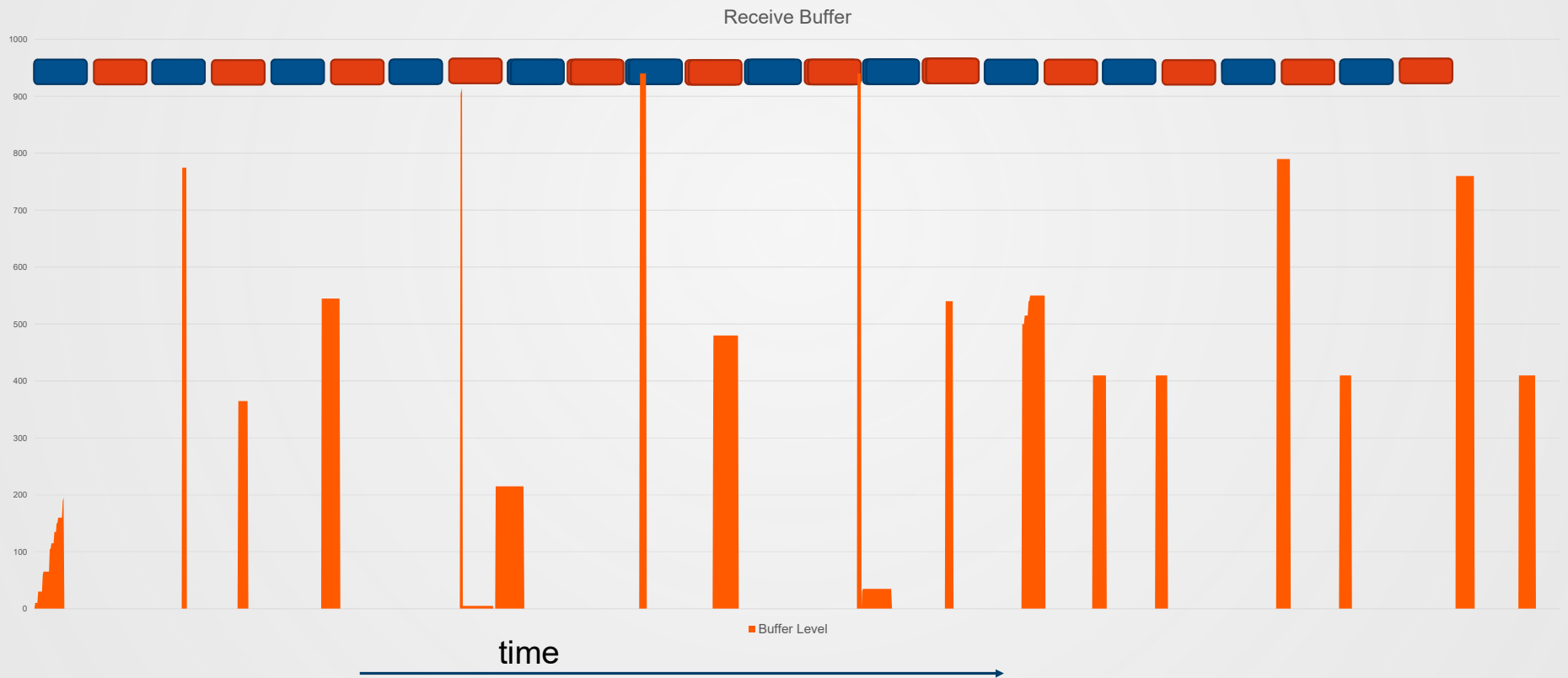
Packet Pacing, avoiding switch buffer congestion

- Consider N Streaming Video flows entering a network switch filling up the output bandwidth
- Their arrival times are not correlated – but each flow sends for T_b then waits $T_b(N-1)$
- The number of concurrently injecting flows to a switch is somewhat related to the Max Load problem
- Network switches buffer demands are thus directly related to the flows burstiness
 - Even if the average bandwidth is the same, the higher the burst size the higher the required buffer
- Packet Pacing is the ability to completely avoid any burst of traffic
 - i.e. separate each flow packets evenly



Bursty Traffic Receive Buffer Behavior

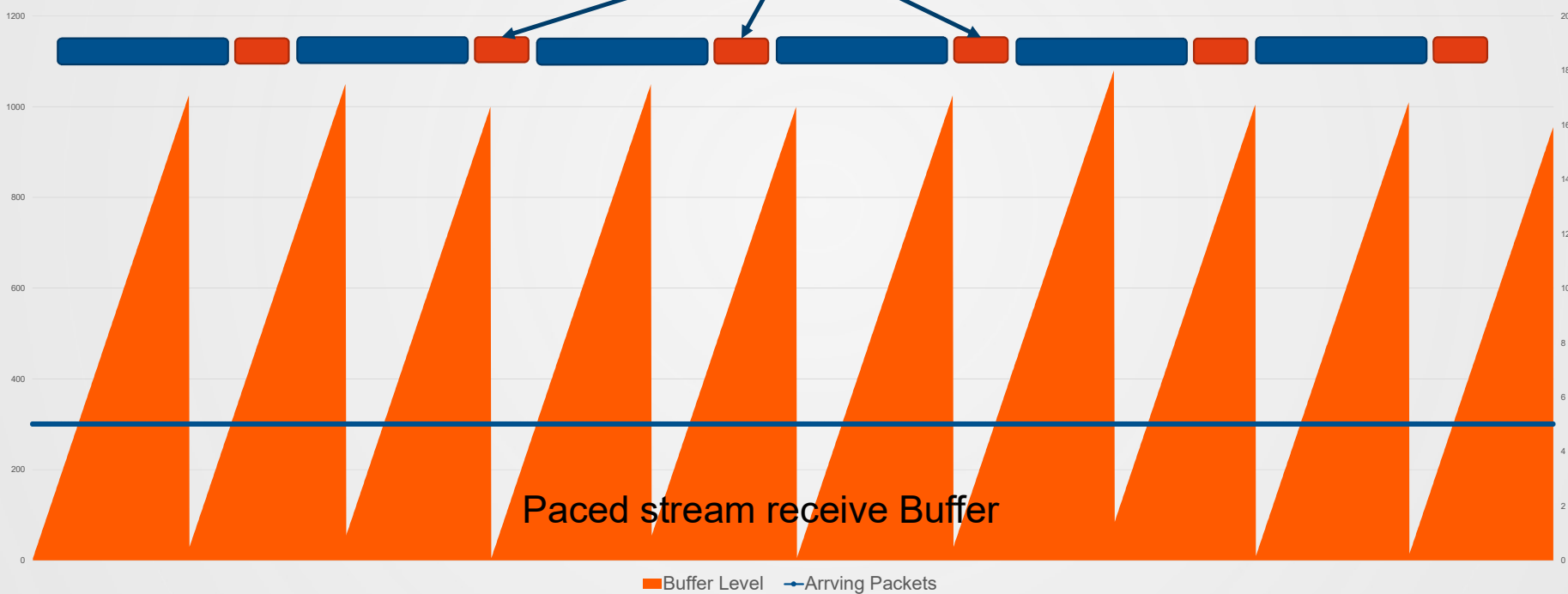
- Application
- Peek Buffer



Paced Behavior – Just In Time Polling

- Application
- Peek Buffer

Every buffer peek is efficient

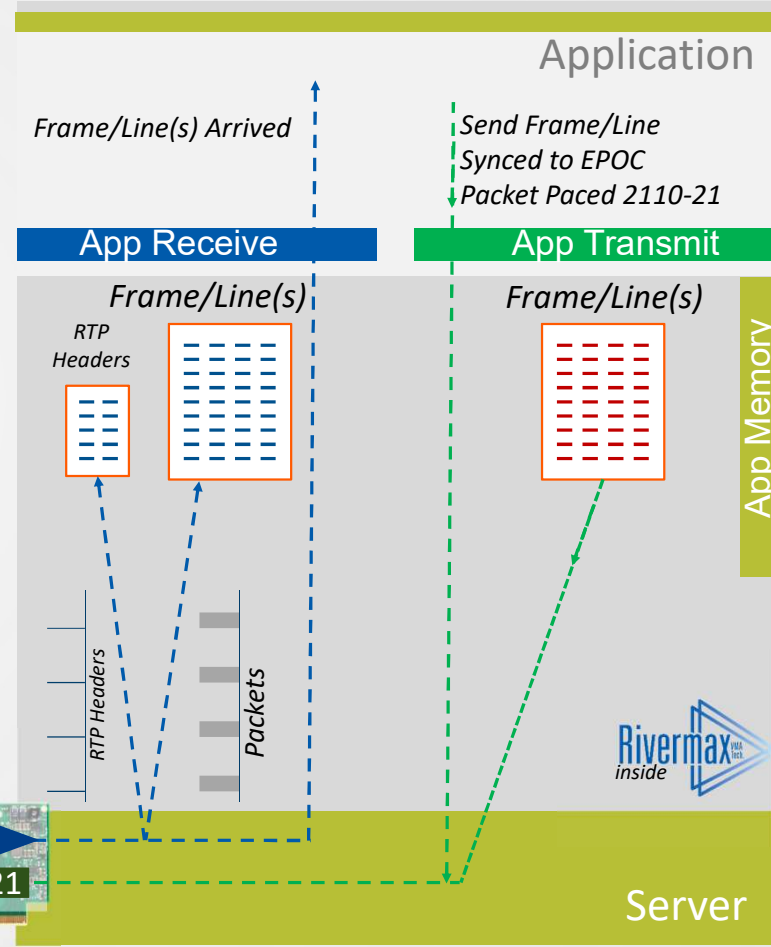
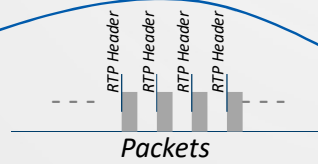


time →

Keeping Application at line(s)/frames level

- Rivermax enables offloading packet handling to ConnectX-5
- On receive (RX network to application)
 - Application receives fully assembled frame/lines(s) in memory
 - RTP Headers stripped to separate buffer
 - Notification to application at full frame/line(s)
 - Built on User Memory Regions (UMR) API

UHD IP Camera



Summary and Conclusion



CPU's do not meet the exponential growth of Network Bandwidth

Network Acceleration Features does

- For example
 - Kernel Bypassing to user space via DPDK / VMA
 - eSwitch offloading
 - Video Processing
- Dedicated Hardware Accelerators are **Key** to Network Performance
 - Not just for RDMA but for almost every network functionality



Thank You

