Mellanox Support Efficient Virtual Networks in Cloud

Accelerate Virtual Switch in Cloud Network 을 중심으로

Mellanox TECHNOLOGIES

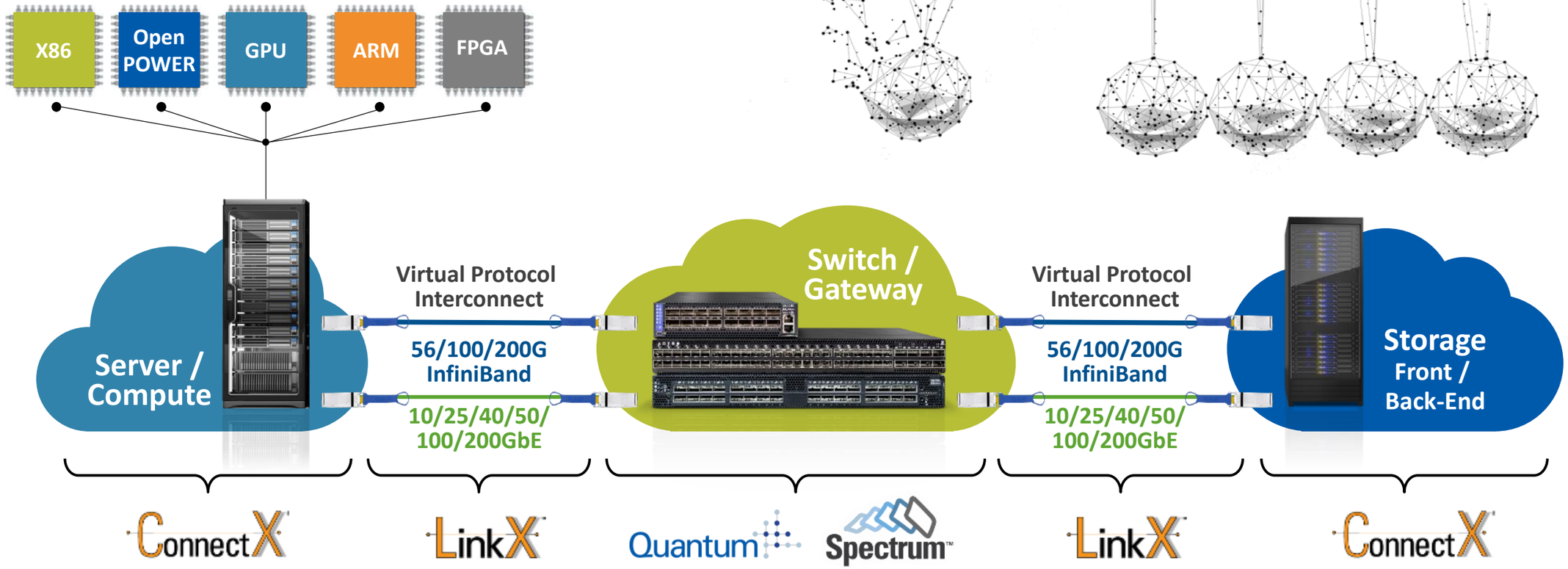
Connect. Accelerate. Outperform.™

OpenStack Day Korea 2017

Mellanox Sr.SE 정연구

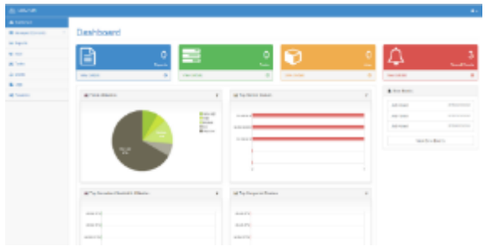
Mellanox Technologies

- Leading Supplier of
  End-to-End Interconnect Solutions



Entering The Era of 25GbE, 50GbE And 100GbE

Software



Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

10 / 25 / 40 / 50 / 56 / 100GbE

Throughput of 6.4Tb/s

Adapter

ConnectX®

100GbE Adapter

200 million messages per second

10 / 25 / 40 / 50 / 56 / 100GbE

ConnectX®-5

ConnectX®-4

Interconnect

LinkX™

Transceivers

Active Optical and Copper Cables

10 / 25 / 40 / 50 / 56 / 100GbE

VCSELs, Silicon Photonics and Copper

Mellanox supporting Industry Platform and solution.

HPC

Deep Learning /AI

Storage/ Parallel File system

Enterprise DB/ Data Analytics

BigData

Hyper-Converged system

Cloud/Web 2.0

InfiniBand H·D·R 2-0-0

- Higher performance
- Higher efficiency
- Higher scalability
- Higher speed

InfiniBand

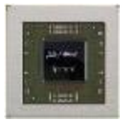
Scale-out Architecture

Network Performance

Spectrum™

Ethernet

cumulus®

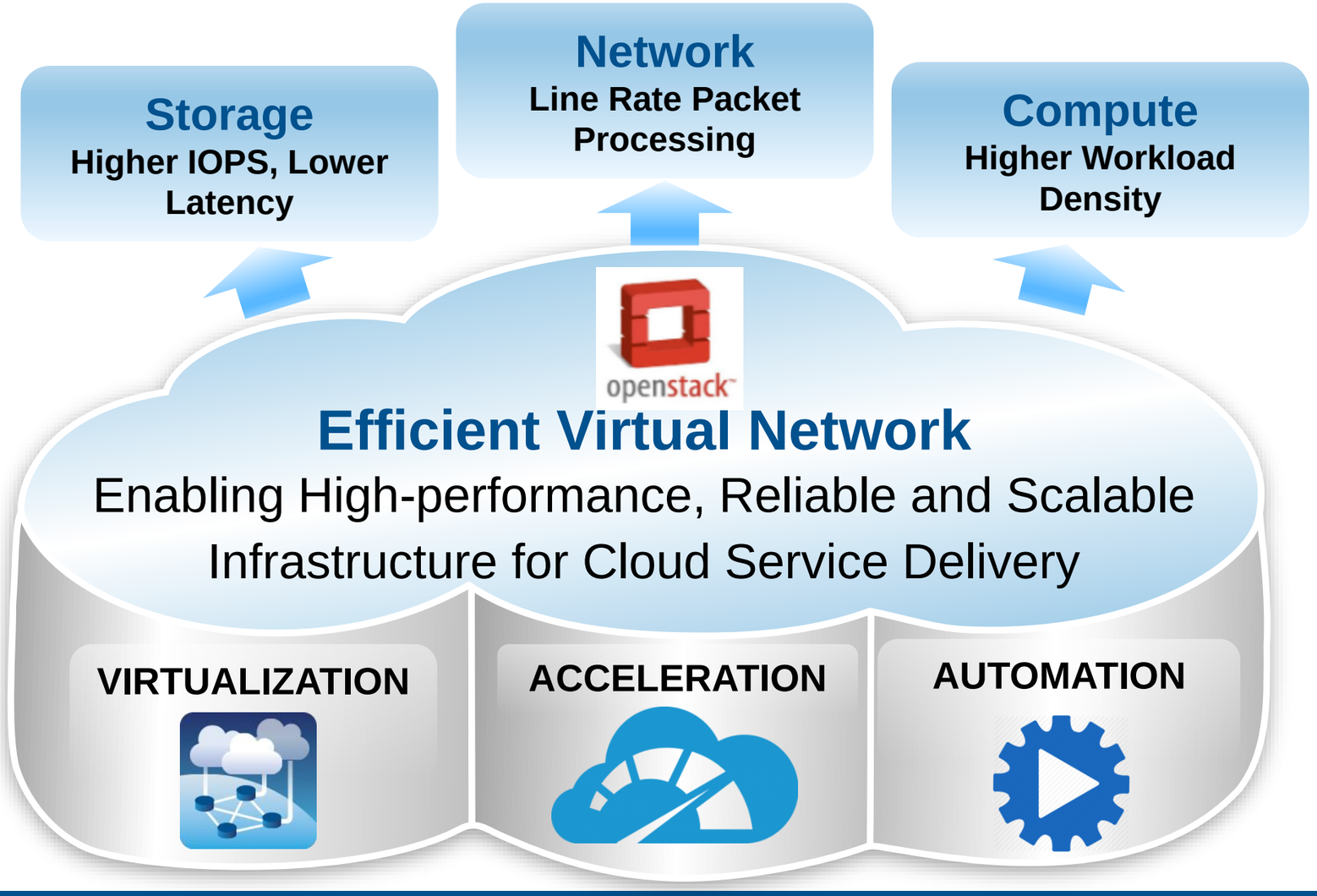


OPEN ETHERNET

SHArP

OPEN Compute Project

Cloud-Native Architecture Dictates Efficient Virtual Network



Mellanox EVN: Foundation for Efficient Cloud Infrastructure

Virtualization

Efficiency and Flexibility with Uncompromised Performance

Three Key Barriers to Achieving the Ultimate Cloud Performance

Inefficient Network Protocols

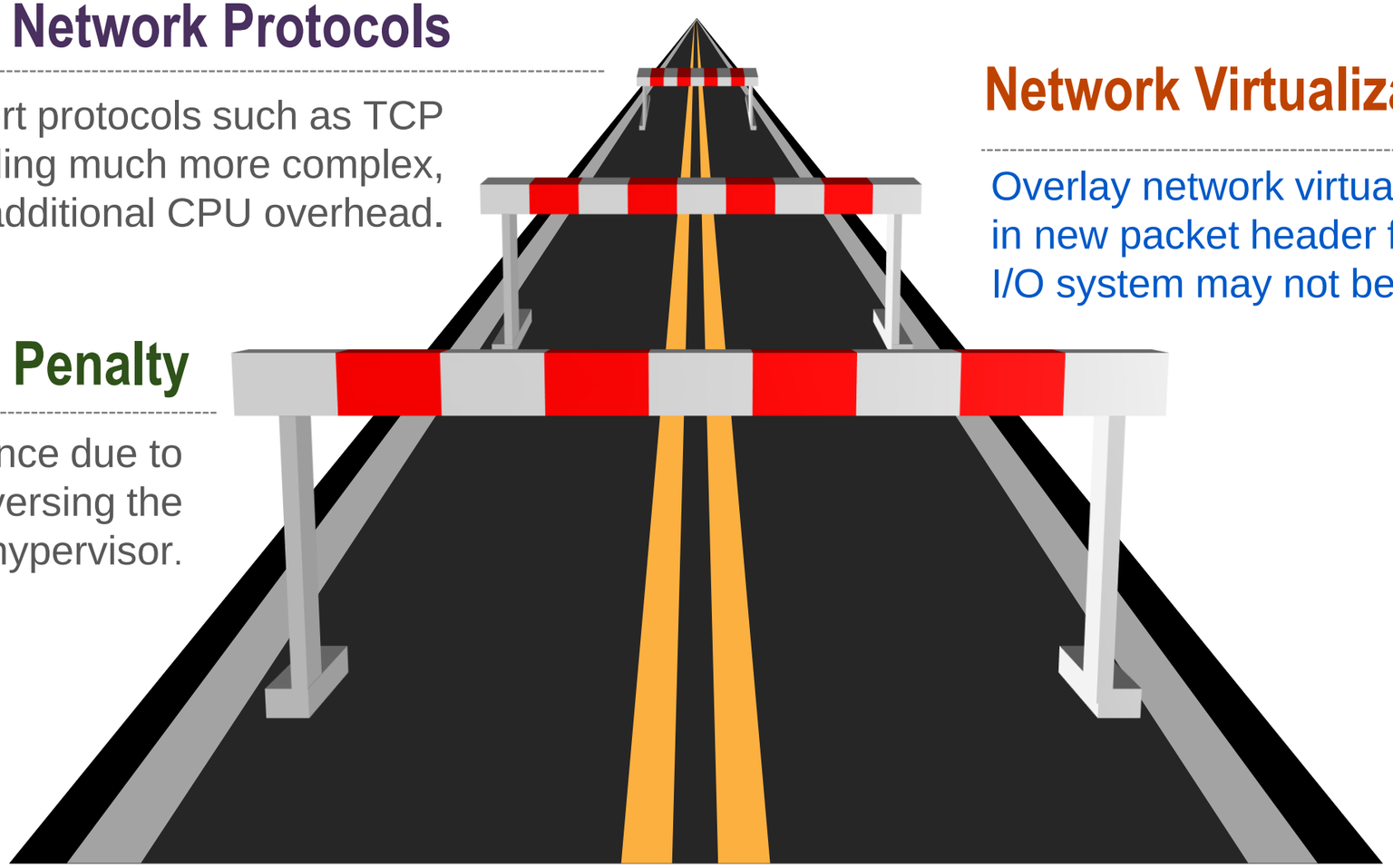
Stateful transport protocols such as TCP makes offloading much more complex, causing additional CPU overhead.

Network Virtualization Penalty

Overlay network virtualization results in new packet header format that the I/O system may not be able to handle.

Compute Virtualization Penalty

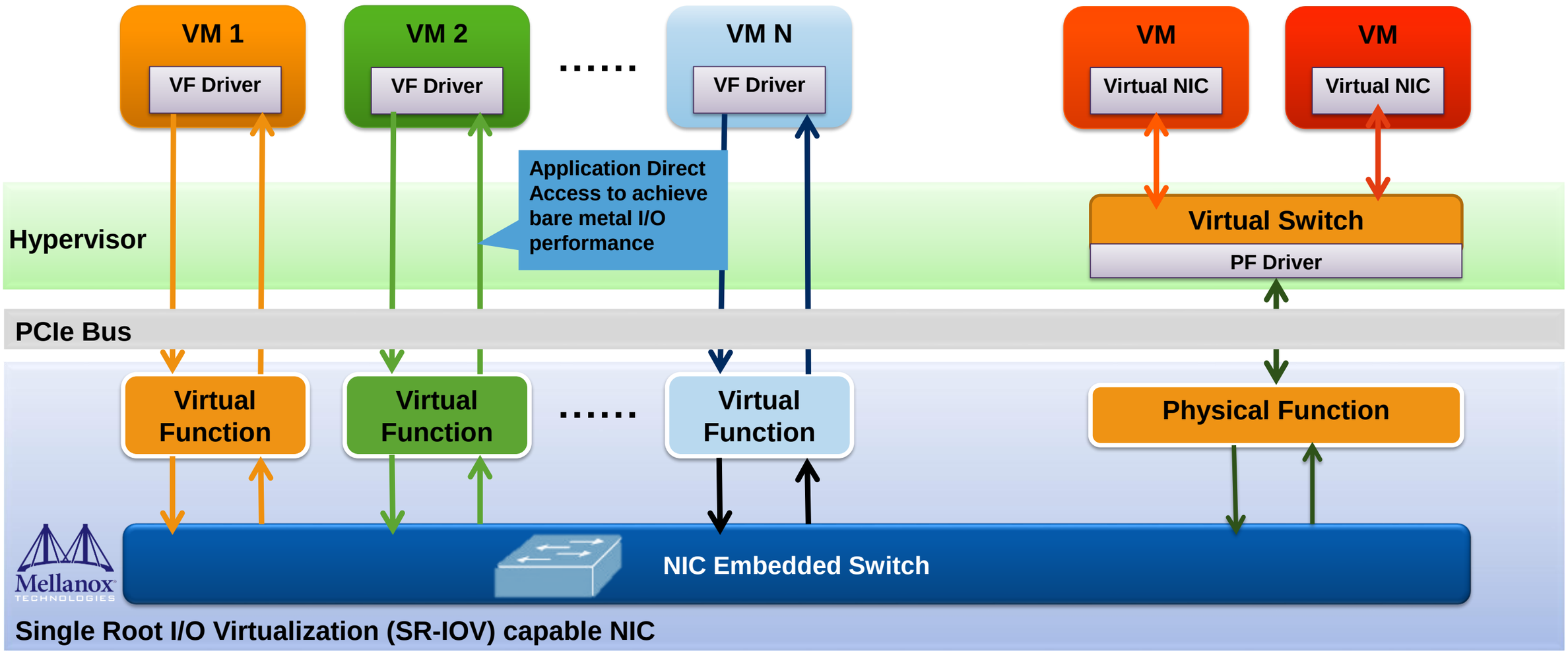
Degraded I/O performance due to overhead of traffic traversing the additional software hypervisor.



SR-IOV – Overcome Compute Virtualization Penalty

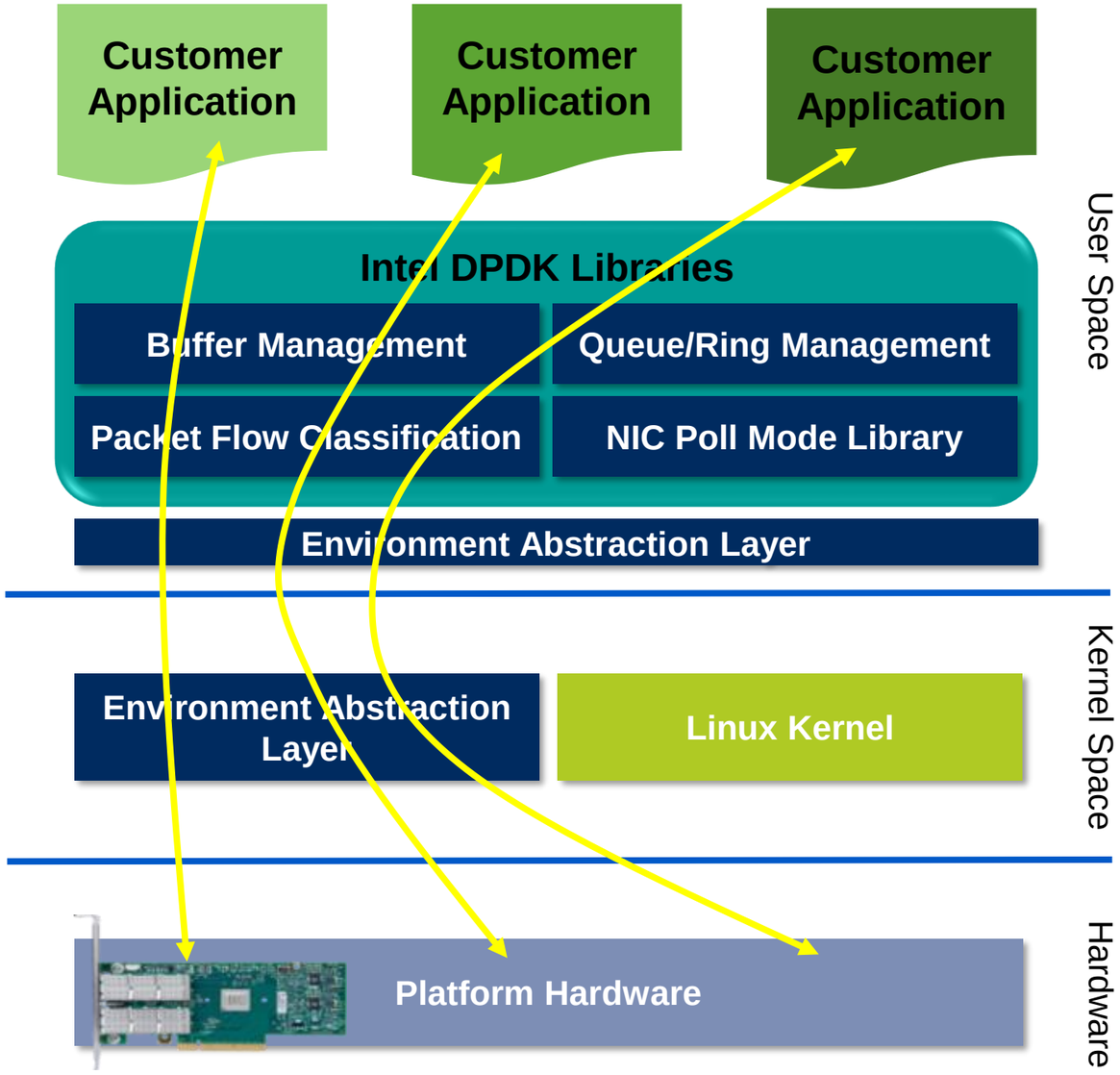
VMs leveraging SR-IOV and Mellanox eSwitch for near-line-rate performance without CPU overhead

Software-switched VMs suffering from compute virtualization penalty

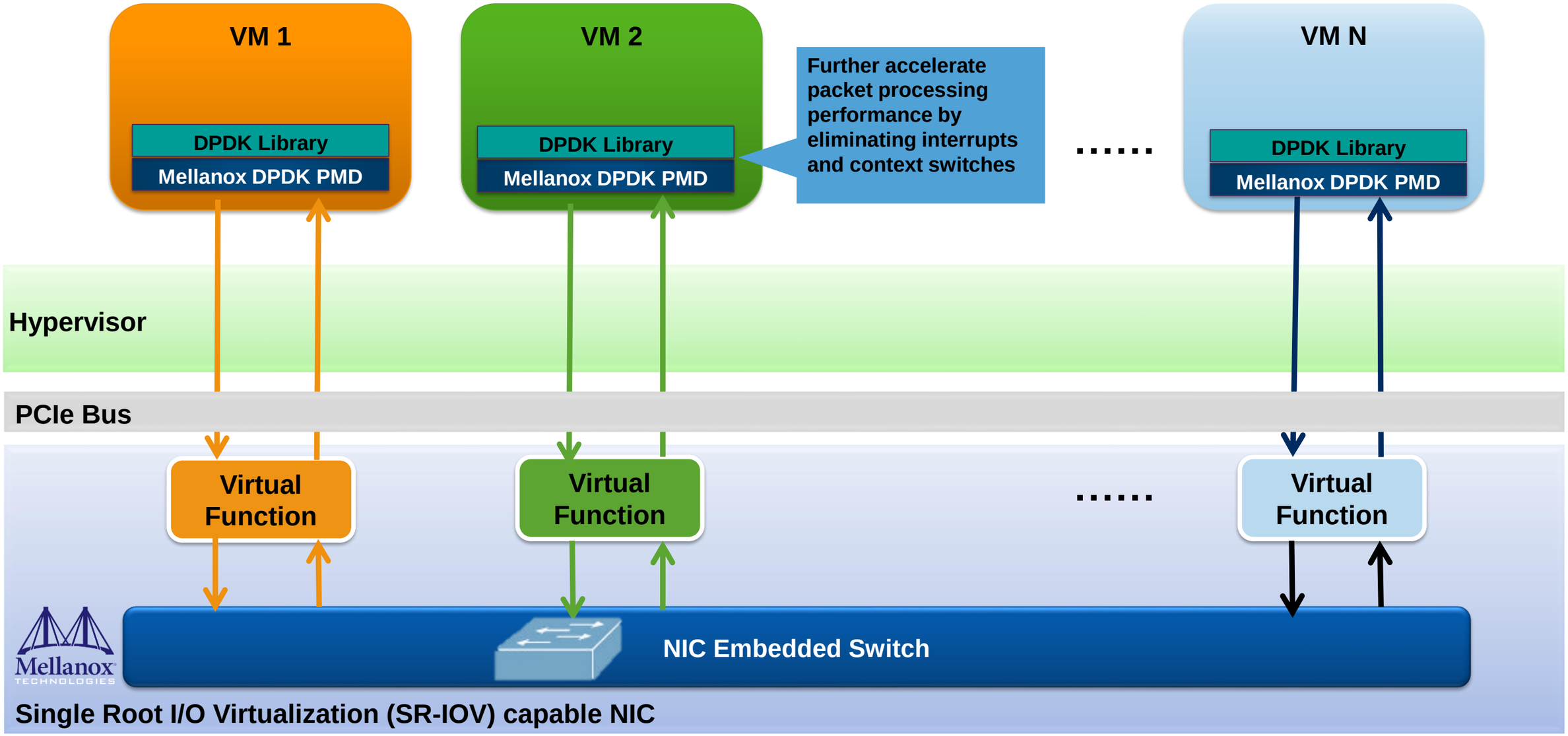


Datacom/Telecom Convergence through DPDK

- DPDK in a Nutshell
  - Data Plane Development Kit
  - Initiated by Intel to drive converged application, control and packet processing on IA
  - Expanding presence to ARM and POWER
  - Widely adopted by NFV, and gaining interests in Web2 and Enterprise sectors
- How does DPDK Enhance Packet Performance
  - Eliminate packet Rx interrupt
    - Switch from an interrupt-driven network device driver to a polled-mode driver
  - Overcome Out-of-Box Linux scheduler context switch overhead
    - Bind a single software thread to a logical core
  - Optimize Memory and PCIe Access
    - Packet batch processing
    - Batched memory read/write
  - Reduced Shared Data Structure Inefficiency
    - Lockless queue and message passing



SR-IOV + DPDK: Better Together with Mellanox PMD

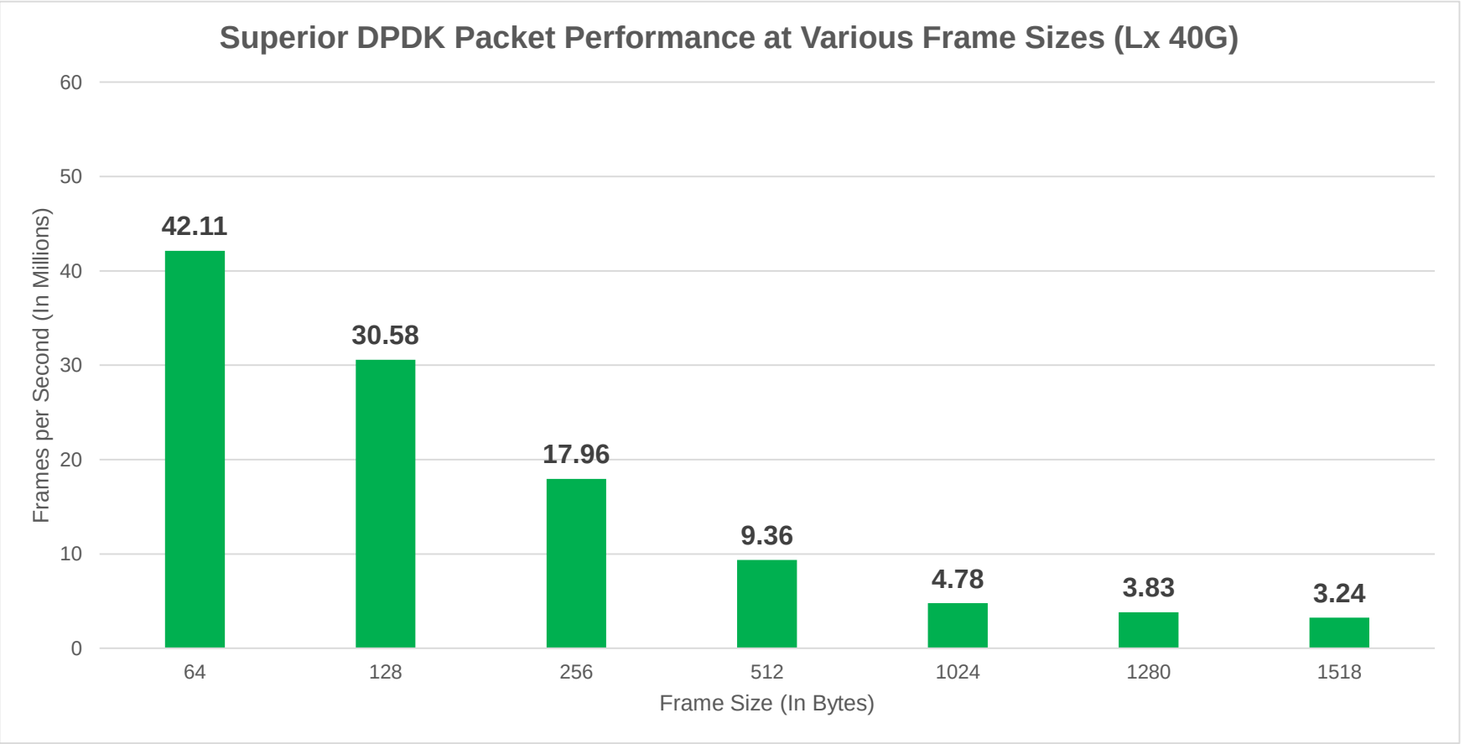


Mellanox Sets New DPDK Performance Records

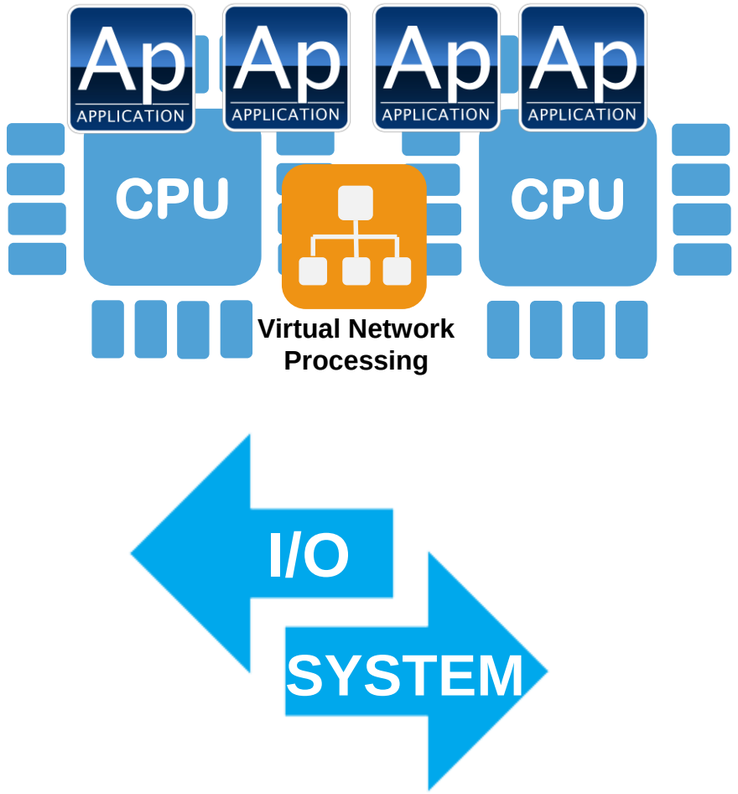
Product	ConnectX-4 100G	ConnectX-4 40G	ConnectX-4 Lx 40G	ConnectX-4 Lx 25G
Single-port TCP Throughput	93.4 Gb/s	37.6 Gb/s	37.6 Gb/s	23.5 Gb/s
DPDK 64B Packet Throughput	74.4 million p/s	56.4 million p/s	42.1 million p/s	34 million p/s

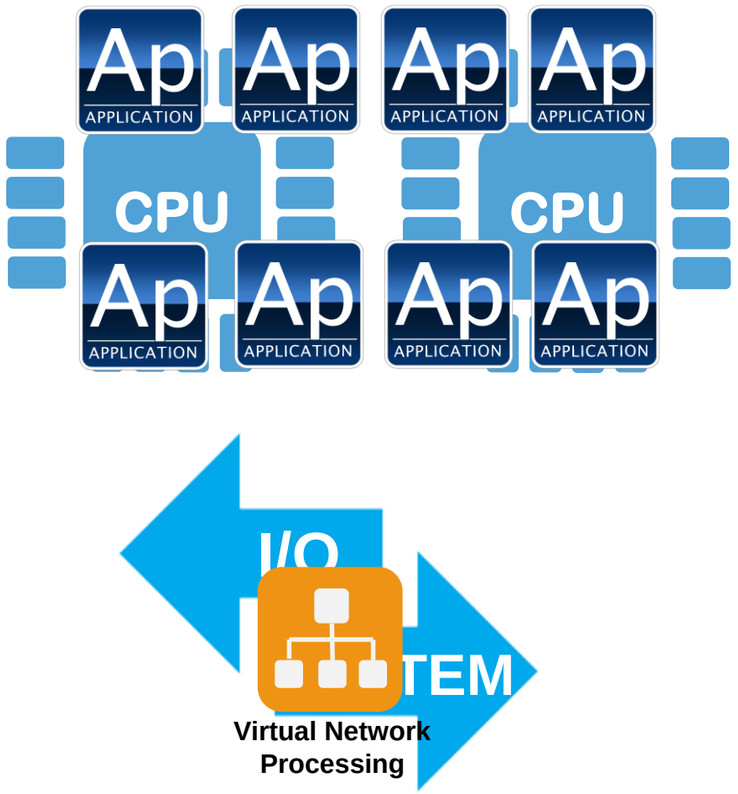
Test setup:

- ConnectX-4Lx 40GbE Single port
- 4 Cores Dedicated to DPDK



Virtual Network Processing Offload
– Overcome Network Virtualization Penalty





Slow Network Processing

Over-burdened CPU

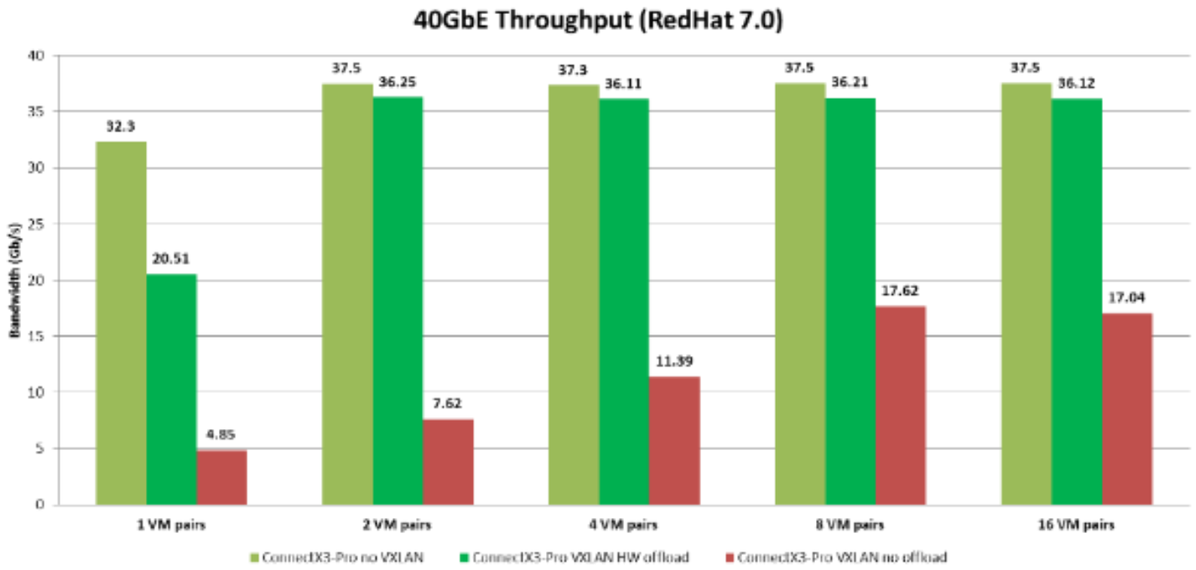
Faster Application Communication

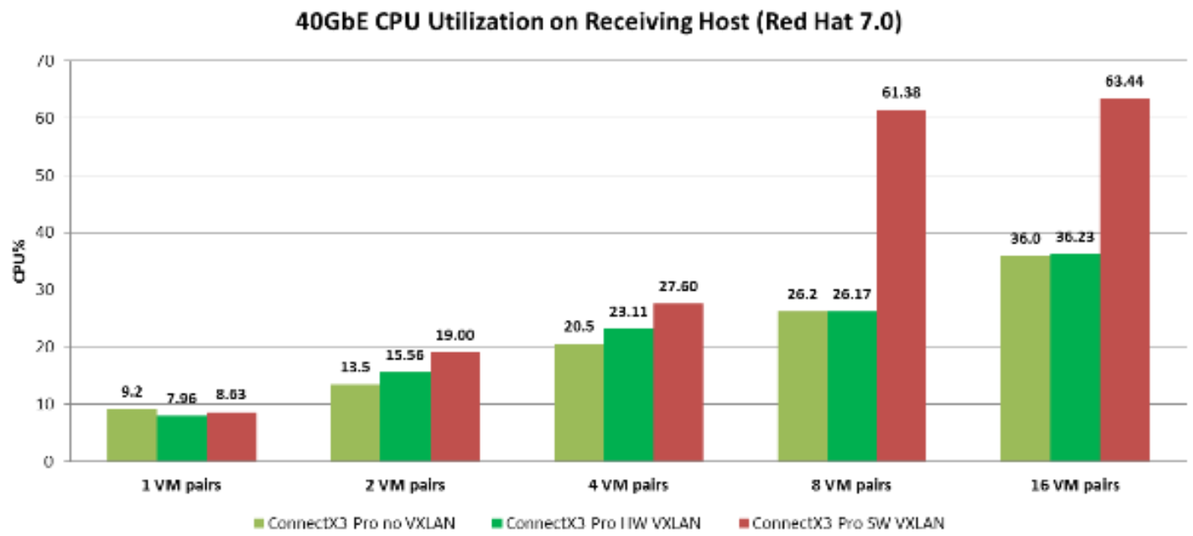
Higher Effective Workload Density

Overlay Network Virtualization: Isolation, Simplicity, Scalability

Turbocharge Overlay Networks with ConnectX-3/4 NICs

- Solution:
  - Overlay Network Accelerators in NIC
  - Penalty free overlays at bare-metal speed
  - Integrated and validated by major SDN vendors
- Benefits:
  - **37.5Gb/s** on 40G link, **>2X** compared to without VxLAN offload
  - On a 20 cores system, 7 cores are freed to run addition VMs, saving **35%** of total cores while doubling the throughput!





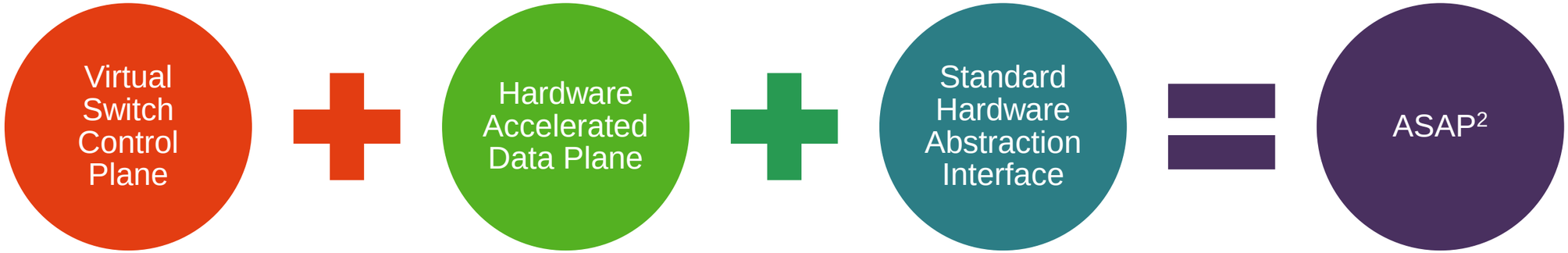
Cumulus Overlay Solution

- Switch VXLAN tunnel endpoint (VTEP) is used
  - To connect bare metal servers to VXLAN network
  - To connect VXLAN and legacy network
- Cumulus Integrated with every major Overlay Solution
- Available with Mellanox switches April 2016

VMware NSX	✓
PLUMgrid ONS	✓
Nuage VSP	✓
Midokura Midonet	✓
Juniper OpenContrail	✓
Akanda Astara	✓
Cumulus LNV	✓

Accelerated Switching And Packet Processing (ASAP^2)

Roadmap

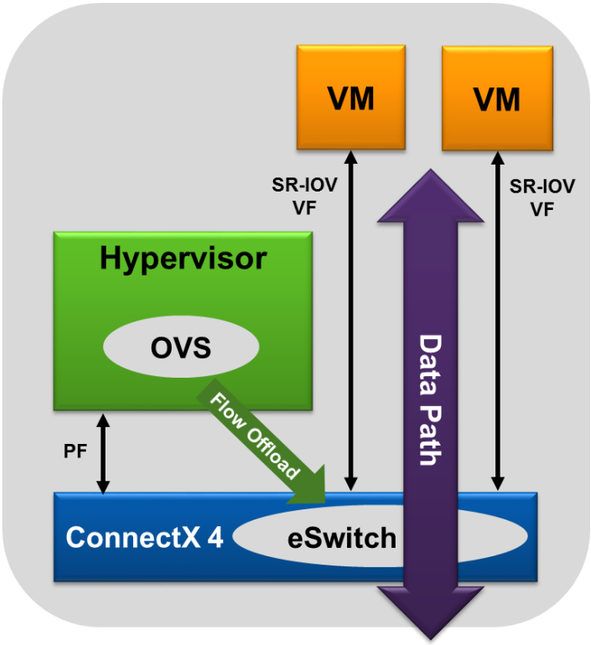


- Best of both worlds: Enable hardware accelerated data plane with SDN/virtual switch control plane
- Multiple possibilities of accelerated data plane including DPDK in CPU, embedded switch, FPGA, network processor, multi-core processor in server adaptor, TOR switch, or centralized acceleration pool
- Standard hardware API to allow control plane and data plane to operate and innovate independently

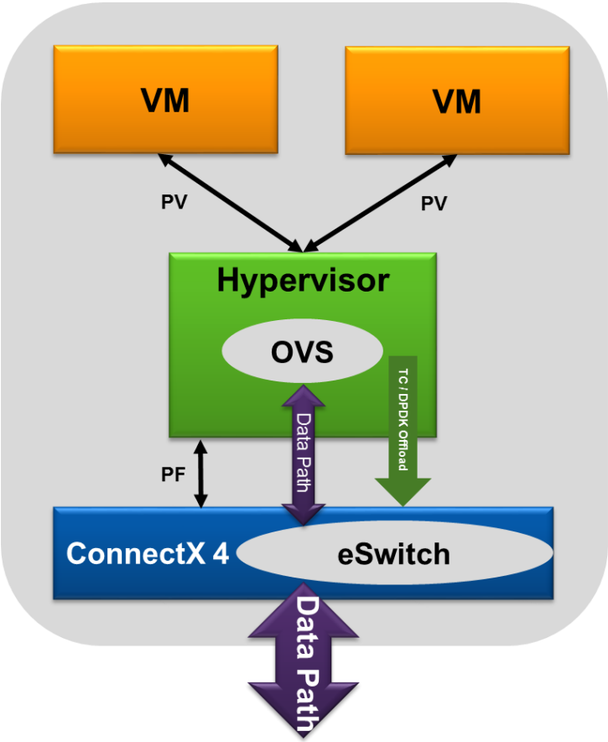
Accelerated Switch And Packet Processing (ASAP^2)

- ASAP^2 take advantage of ConnectX-4 capability to accelerate or offload “in host” network stack
- Three main use cases

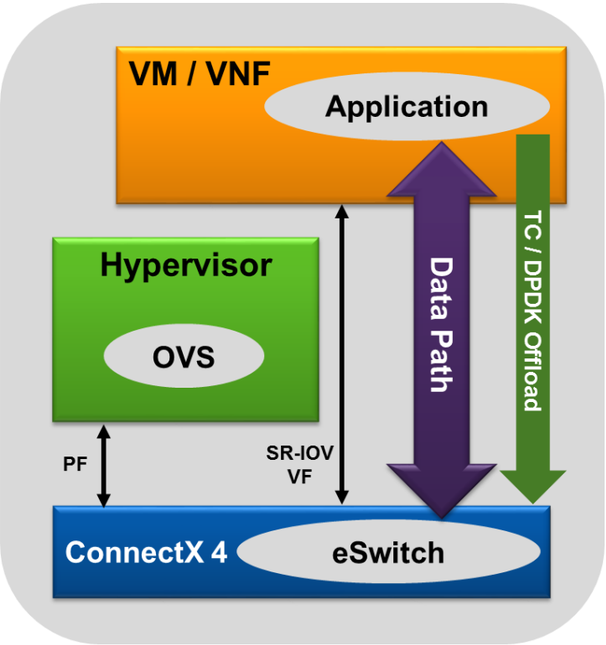
ASAP^2 Direct
Full vSwitch offload



ASAP^2 Flex
vSwitch acceleration

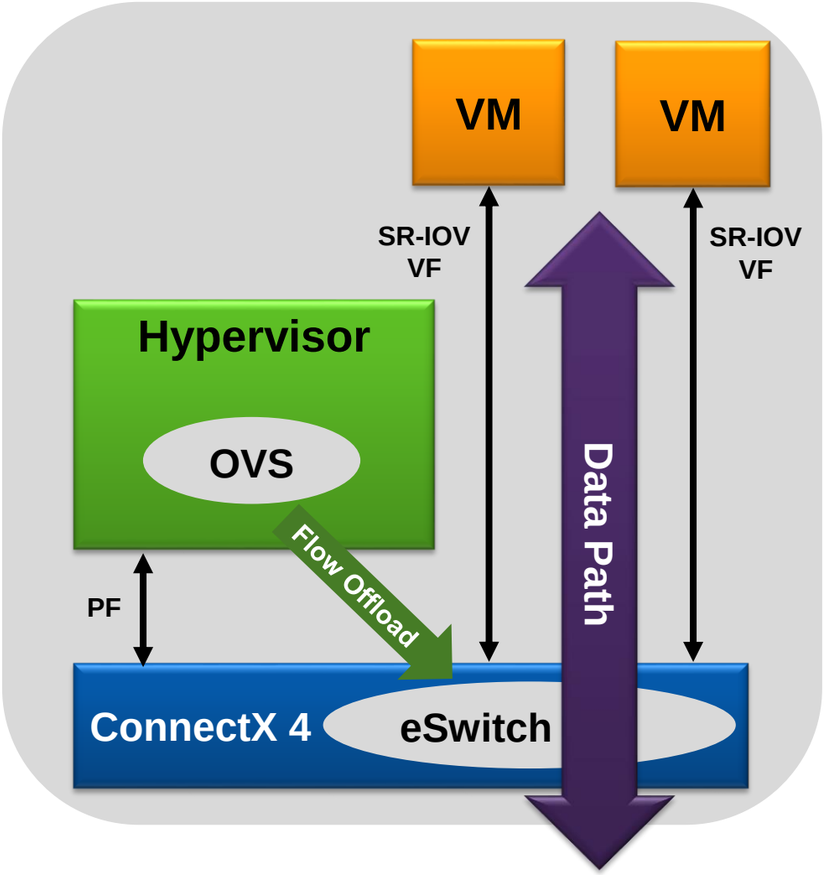


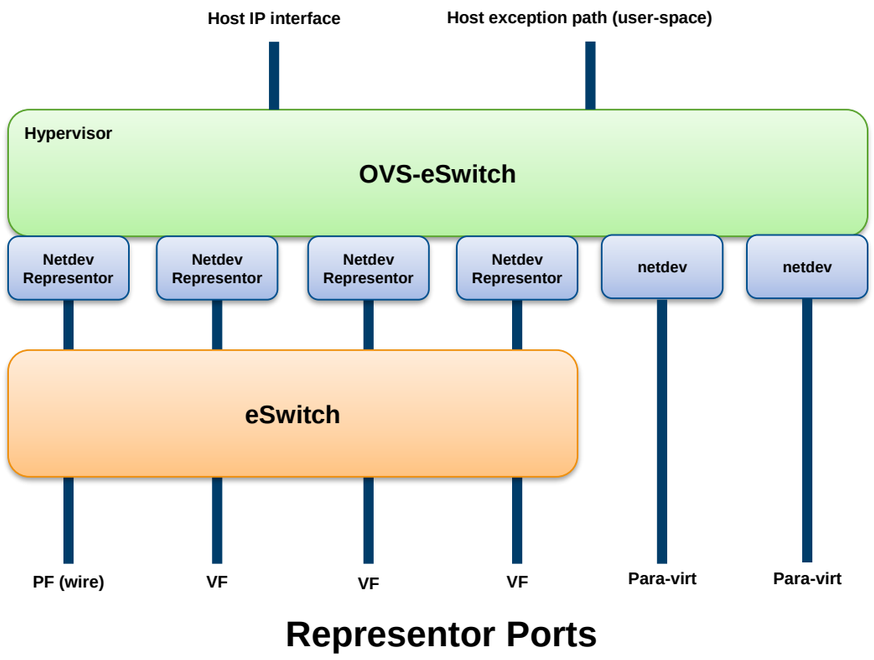
ASAP^2 Flex
VNF/VM acceleration

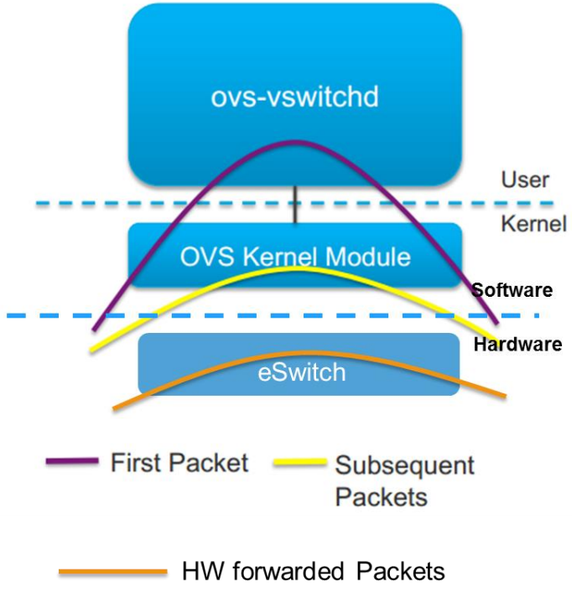


ASAP^2 Direct: Full OVS Offload

- Enable SR-IOV data path with OVS control plane
  - In other words, enable support for most SDN controllers with SR-IOV data plane
- Use Open vSwitch to be the management interface and offload OVS data-plane to Mellanox embedded Switch (eSwitch) using ASAP^2 Direct

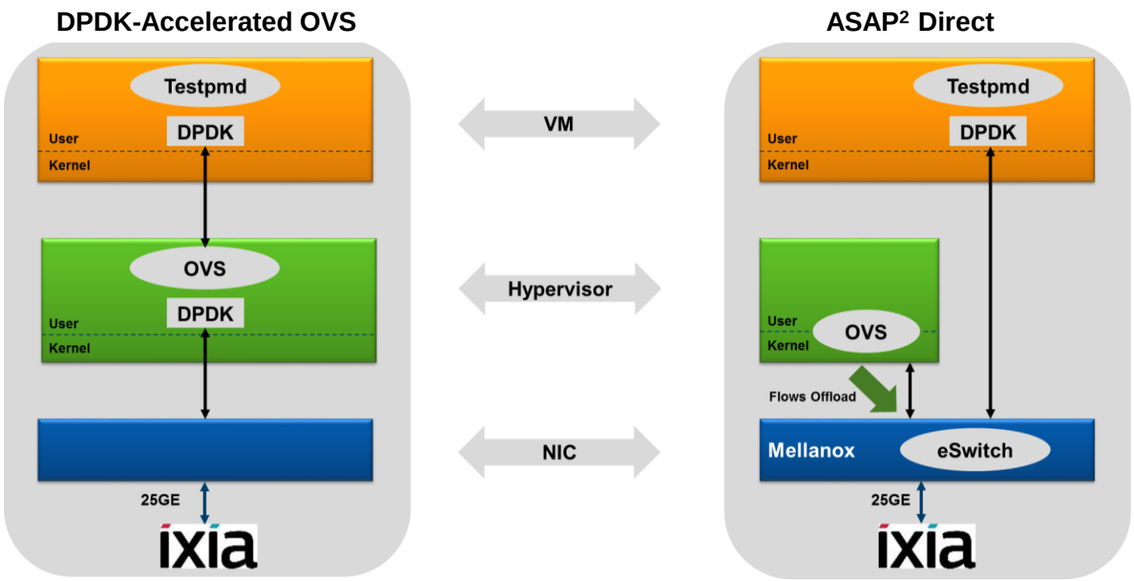


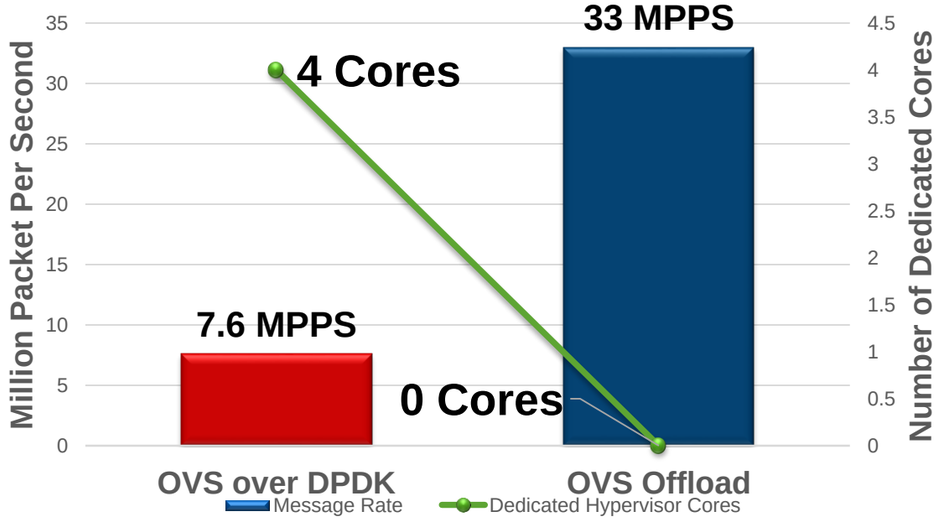




DPDK-Accelerated OVS VS. ASAP² Direct – Initial results

- 1 flow, no VXLAN
  - 330% higher message rate compared to OVS over DPDK
  - Zero! CPU utilization on hypervisor compared to 4 cores with OVS over DPDK
    - Same CPU load on VM
- 2000 flows, VXLAN HW encap/decap
  - OVS offload reach ~25MPPS
  - Still zero CPU compared to 4 cores with DPDK



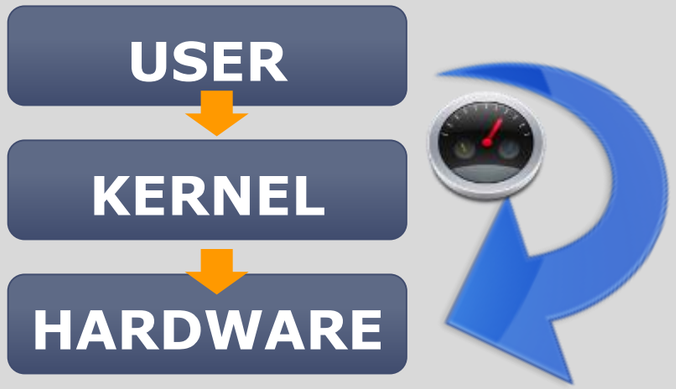


Acceleration

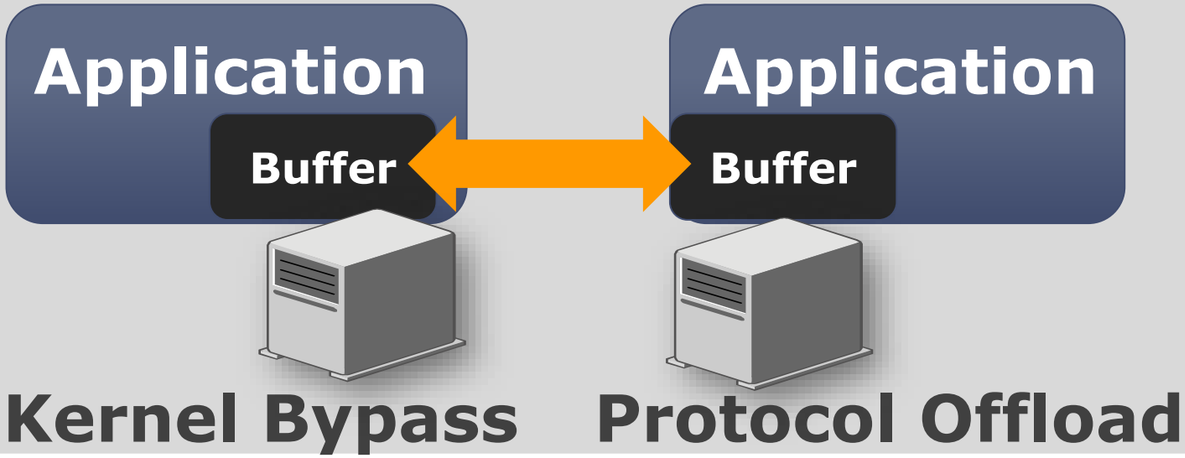
Enable Fast Networking and Storage Access for Scale-Out Applications

RDMA Acceleration – Overcome Transport Protocol Inefficiencies

ZERO Copy



Remote Data Transfer



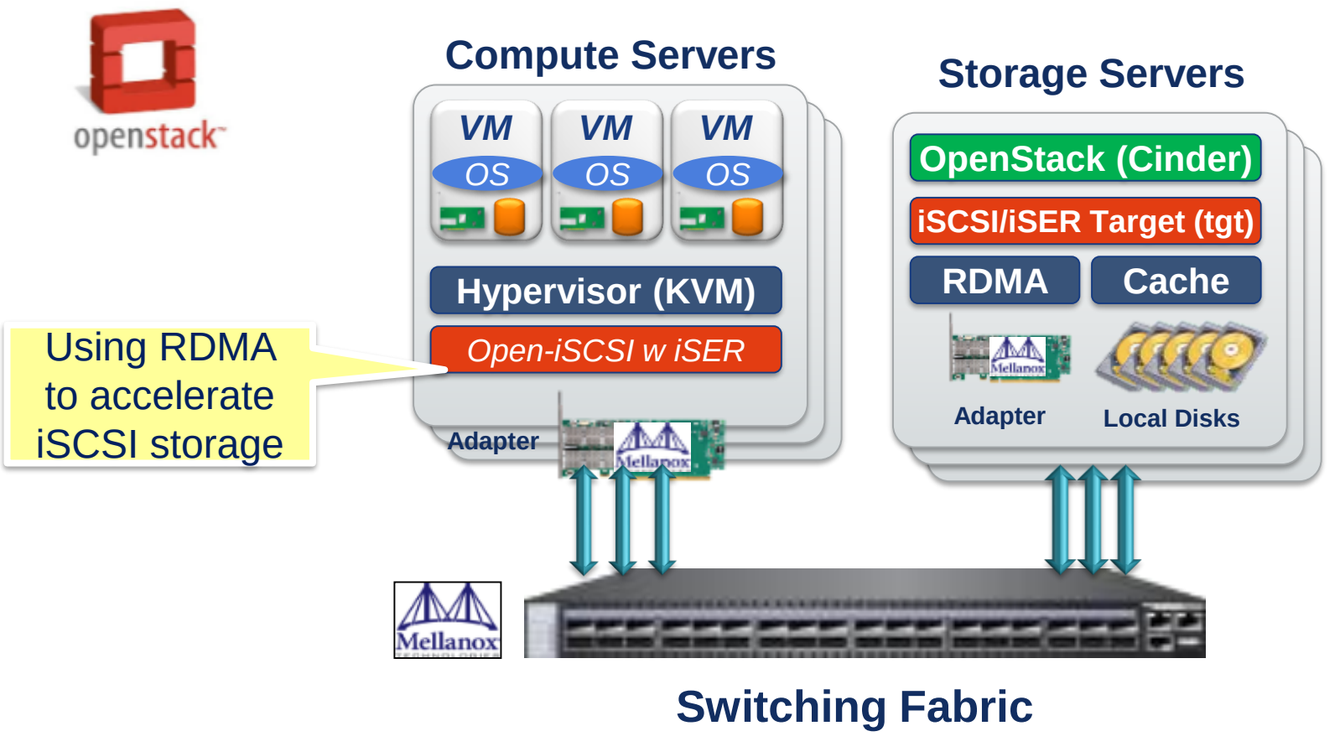
Low Latency, High Performance Data Transfers

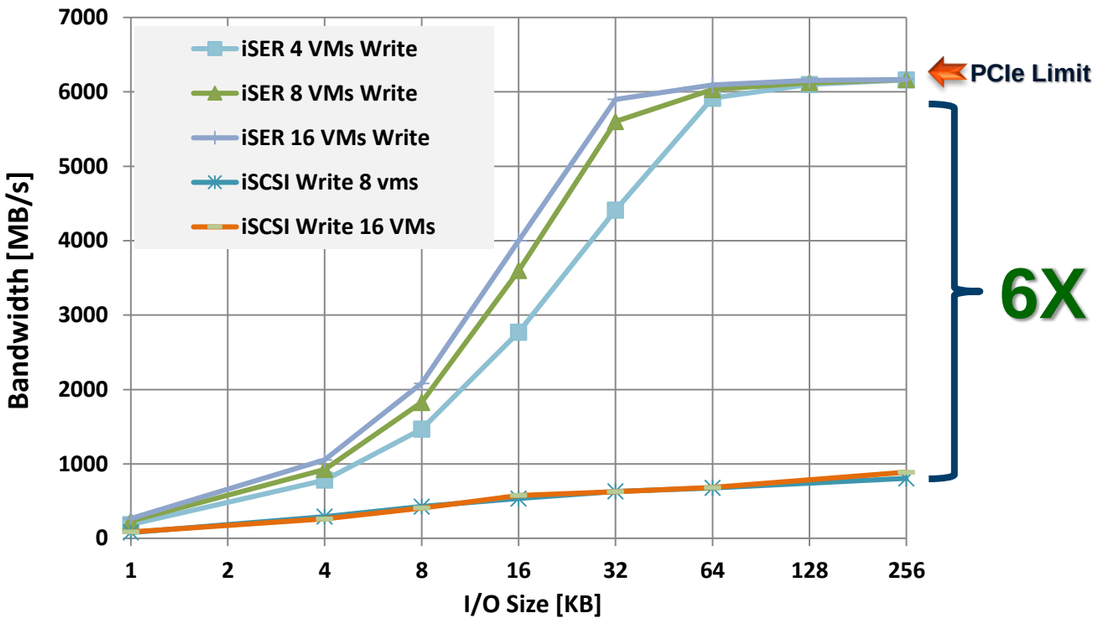
InfiniBand - 100Gb/s

RoCE* – 100Gb/s

* RDMA over Converged Ethernet

RDMA Provide Fastest OpenStack Storage Access

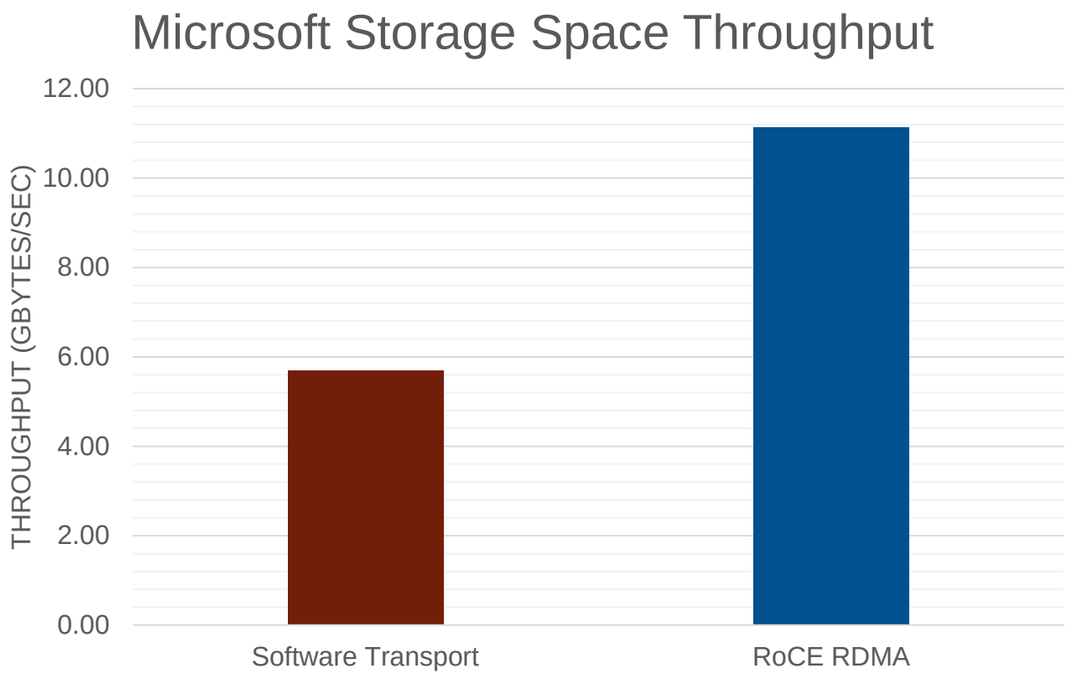


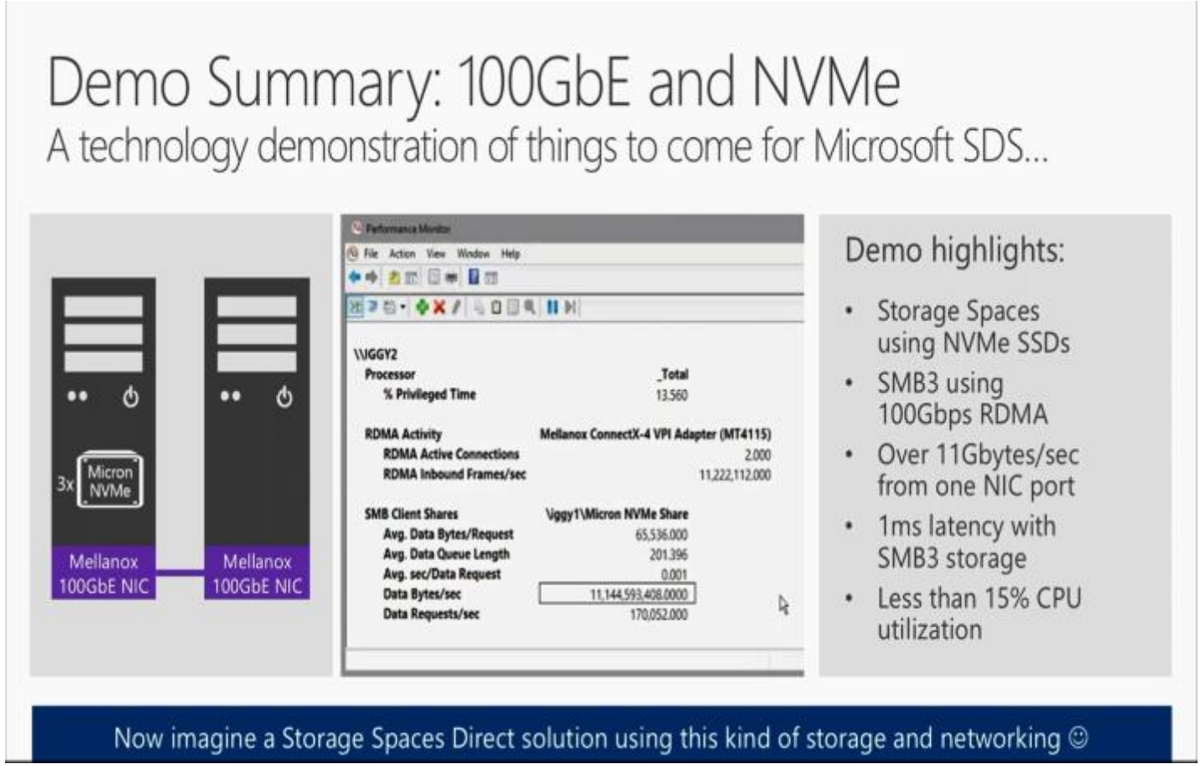


- Using OpenStack Built-in components and management (Open-iSCSI, tgt target, Cinder), no additional software is required, RDMA is already inbox and used by our OpenStack customers !

RDMA enables 6x More Bandwidth, 5x lower I/O latency, and lower CPU%

Microsoft 100Gb/s Cloud Demonstration



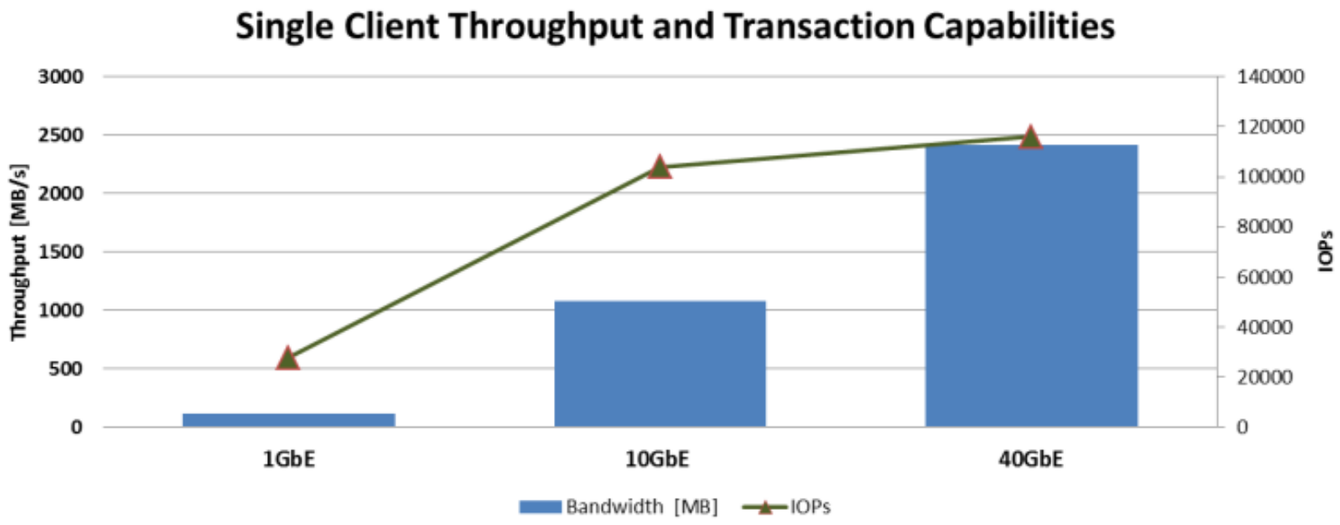


- Remote NVME Flash storage throughput
  - ConnectX-4 100Gb/s Ethernet Adapters
  - RoCE RDMA achieves full flash bandwidth
    - Remote storage without compromises
  - Twice the bandwidth & less than half the CPU utilization

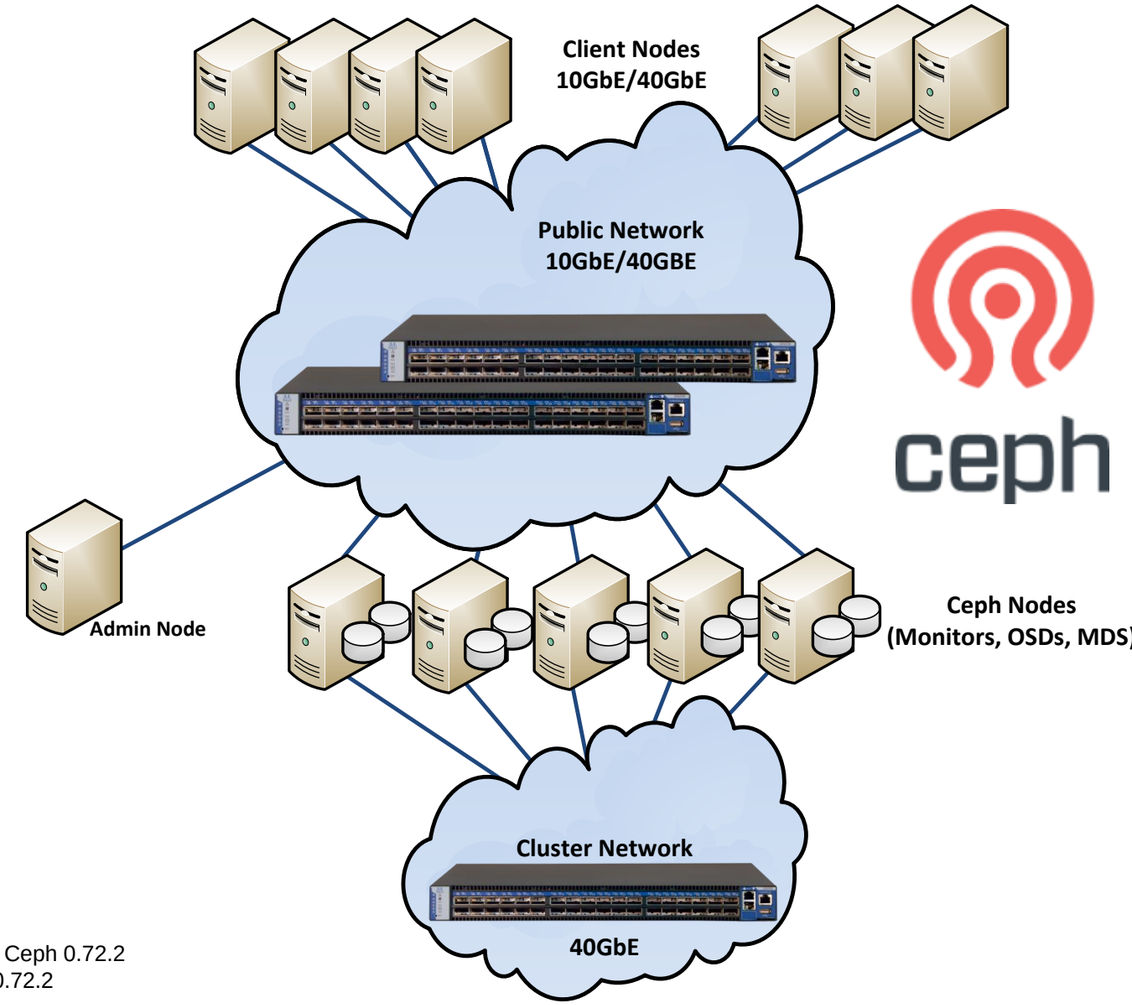
Click to Watch Microsoft 100Gb/s RoCE Presentation

Ceph Deployment Using 10GbE and 40GbE

- Cluster (Private) Network @ 40/56GbE
  - Smooth HA, unblocked heartbeats, efficient data balancing
- Throughput Clients @ 40/56GbE
  - Guaranties line rate for high ingress/egress clients
- IOPs Clients @ 10GbE or 40/56GbE
  - 100K+ IOPs/Client @4K blocks



Throughput Testing results based on fio benchmark, 8m block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2
IOPs Testing results based on fio benchmark, 4k block, 20GB file,128 parallel jobs, RBD Kernel Driver with Linux Kernel 3.13.3 RHEL 6.3, Ceph 0.72.2



20x Higher Throughput , 4x Higher IOPs with 40Gb Ethernet Clients!
(http://www.mellanox.com/related-docs/whitepapers/WP_Deploying_Ceph_over_High_Performance_Networks.pdf)

Accelerating Ceph with RDMA – Work in Progress

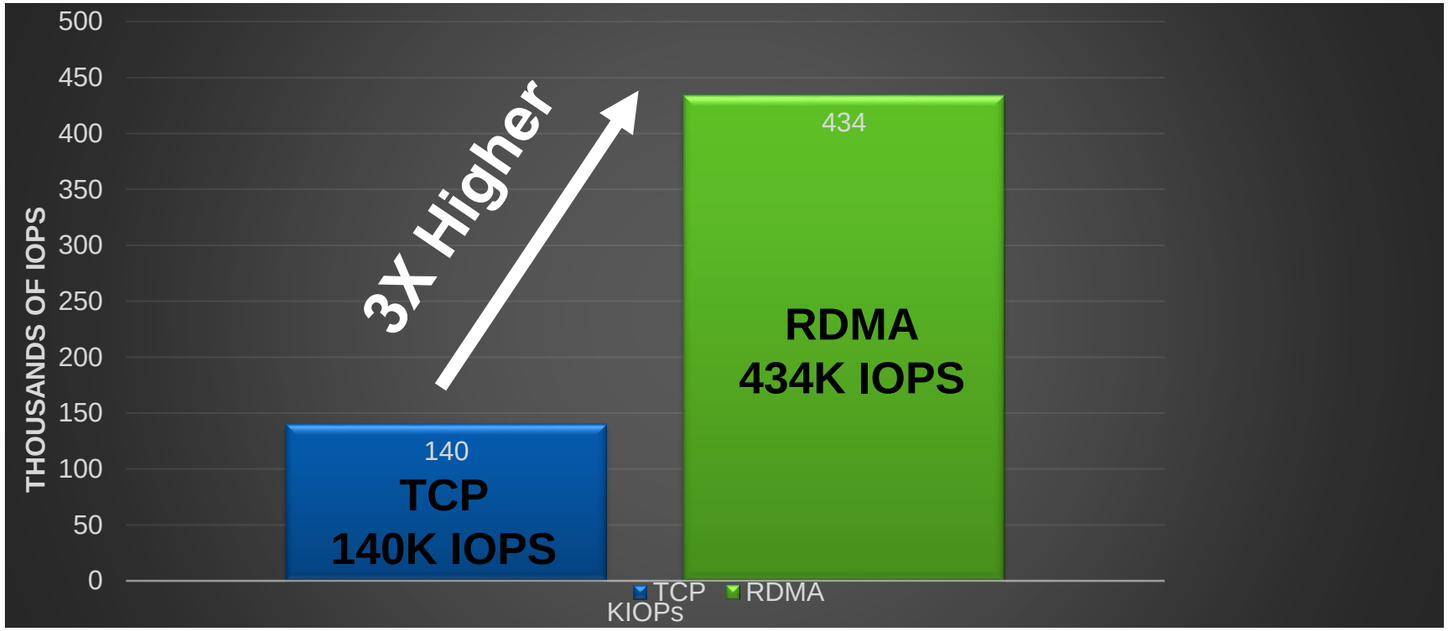
Roadmap

- Accelio, High-Performance Reliable Messaging and RPC Library

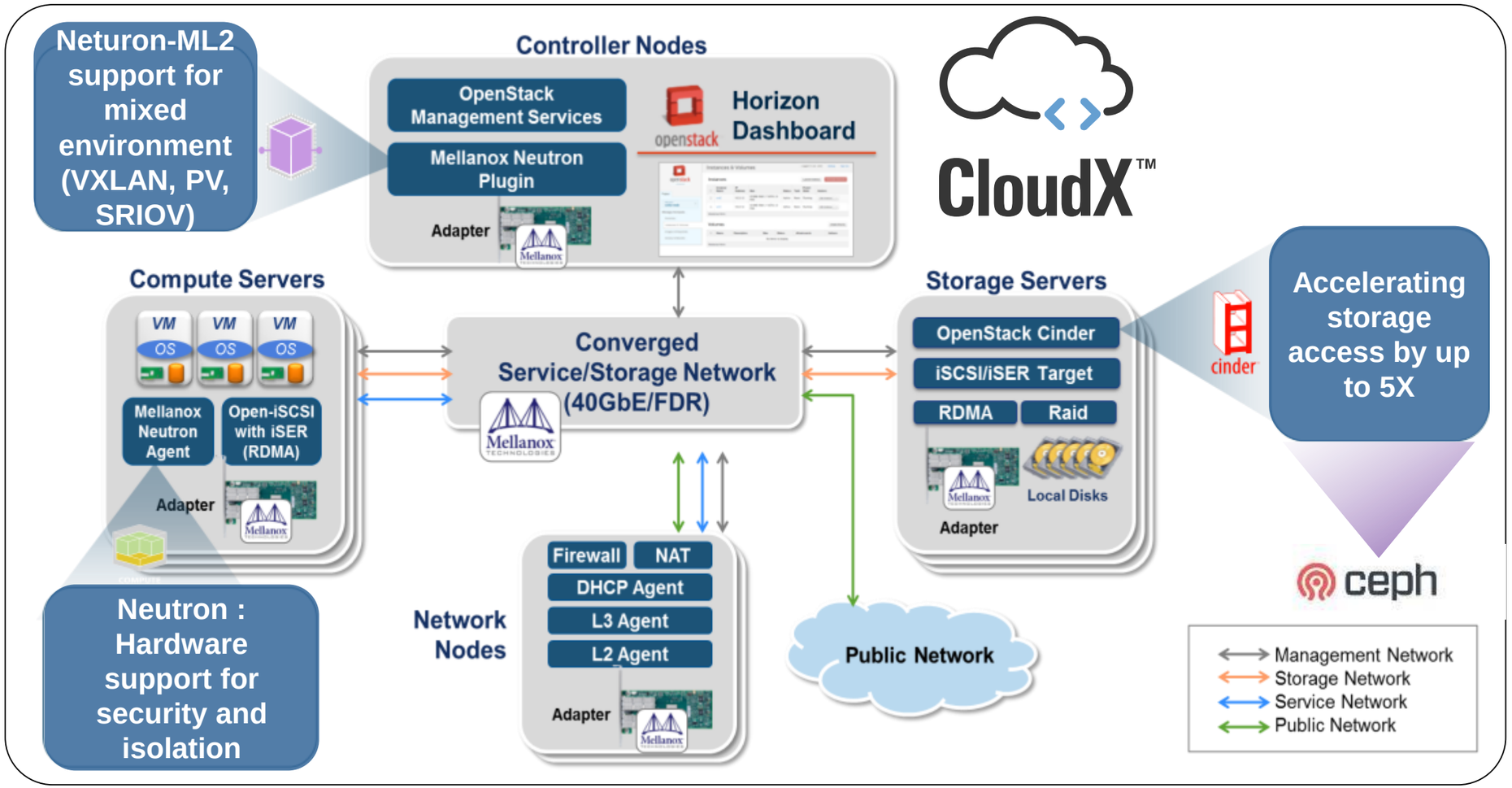
- Open source!
  - https://github.com/accelio/accelio/ && www.accelio.org

- Faster RDMA integration to application
- Asynchronous
- Maximize msg and CPU parallelism
- Enable > 10GB/s from single node
- Enable < 10usec latency under load

Ceph Read IOPS: TCP vs. RDMA



Comprehensive OpenStack Integration for Switch and Adapter



Integrated with Major OpenStack Distributions

hp
juju ubuntu Certified
MIRANTIS Pure Play OpenStack Mirantis Partner
redhat CERTIFIED Partner

In-Box

HAVANA ICEHOUSE
openstack JUNO KILO
LIBERTY mitaka

INFINIBAND TRADE ASSOCIATION
Ethernet

OpenStack Plugins Create Seamless Integration , Control, & Management

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™