Moving HPC Workloads to the Cloud

Asaf Wachtel, Sr. Director of Business Development

HPC for Wall Street | April 2016

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Leading Supplier of End-to-End Interconnect Solutions

Comprehensive End-to-End InfiniBand and Ethernet Portfolio (VPI)

ICs

Adapter Cards

NPU & Multicore

Switches/Gateways

Software

Metro / WAN

Cables/Modules

Mellanox InfiniBand Proven and Most Scalable HPC Interconnect

OAK RIDGE National Laboratory

“Summit” System

U.S. DEPARTMENT OF ENERGY

Lawrence Livermore National Laboratory

“Sierra” System

Paving the Road to Exascale

Mellanox Ethernet Enables the Most Efficient Azure / Azure Stack

Microsoft

Ignite'15 Demo Highlights:

- ❑ Mellanox 100Gb/s
- ❑ Mellanox RDMA
- ❑ 2X Cloud Efficiency

"Compute intensive VMs – more memory, more virtual machines, InfiniBand access with RDMA within region and across regions at Azure, enable you to build high performance high scale applications"

Brad Anderson, Corporate Vice President, Microsoft

"To make storage cheaper we use lots more network! How do we make Azure Storage scale? RoCE (RDMA over Converged Ethernet) enabled at 40GbE for Windows Azure Storage, achieving massive COGS savings"

Albert Greenberg, Microsoft, SDN Azure Infrastructure

Is the Cloud Ready for HPC Workloads?

- Cloud computing would seem to be an HPC user's dream offering almost unlimited storage and instantly available and scalable computing resources, all at a reasonable metered cost
- Typical clouds offer:
  - Instant availability
  - Large capacity
  - Software choice
  - Virtualized
  - Service-level performance

- HPC users generally have a different set of requirements, mainly as it relates to system performance
- Currently, enterprise use represents 2% to 3% of the HPC in the cloud market, mostly used for "bursts", but that is expected to grow fast in the coming years
- This presentation will focus on the performance aspects, as they relate to different use cases:
  - Traditional HPC
  - Telco NFV (Network Function Virtualization)
  - Financial services

Traditional HPC

Traditional HPC

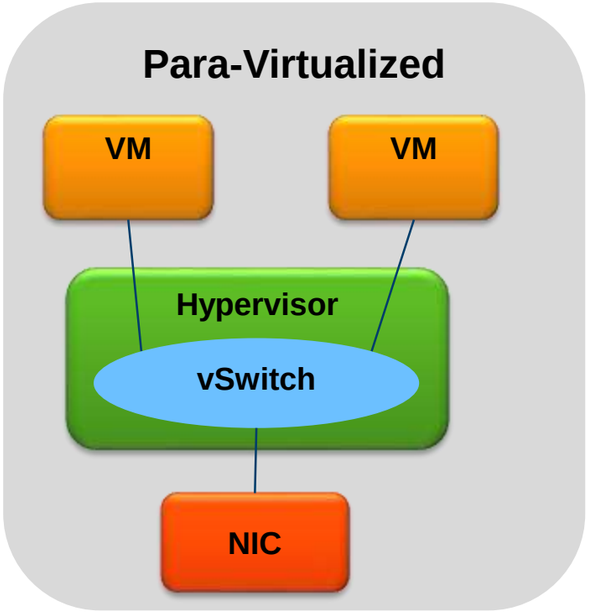
- Government, Defense, Research, Academia, Manufacturing, Oil & Gas, Bio-sciences

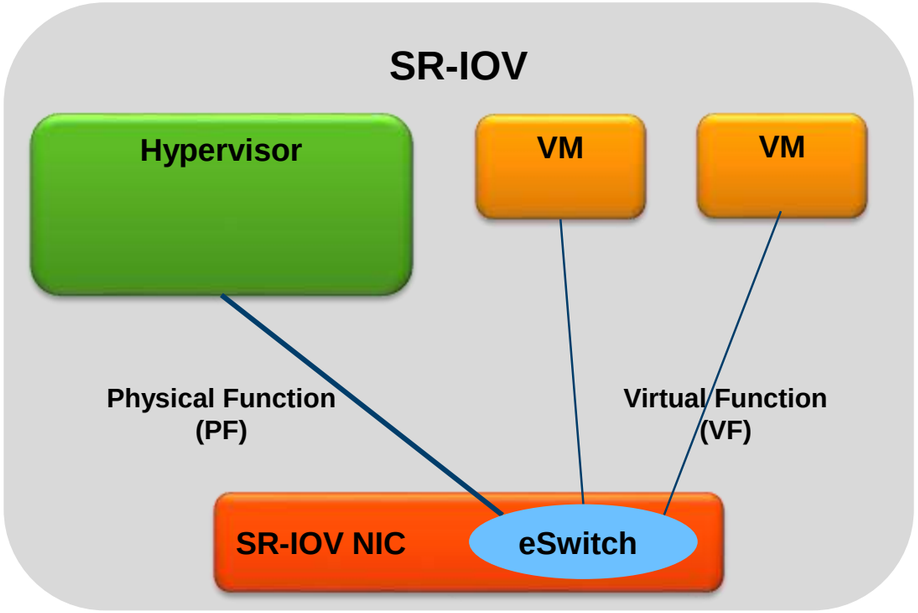
- Large, distributed and synchronized parallel compute jobs
  - Very intense on all fronts – compute, network and storage

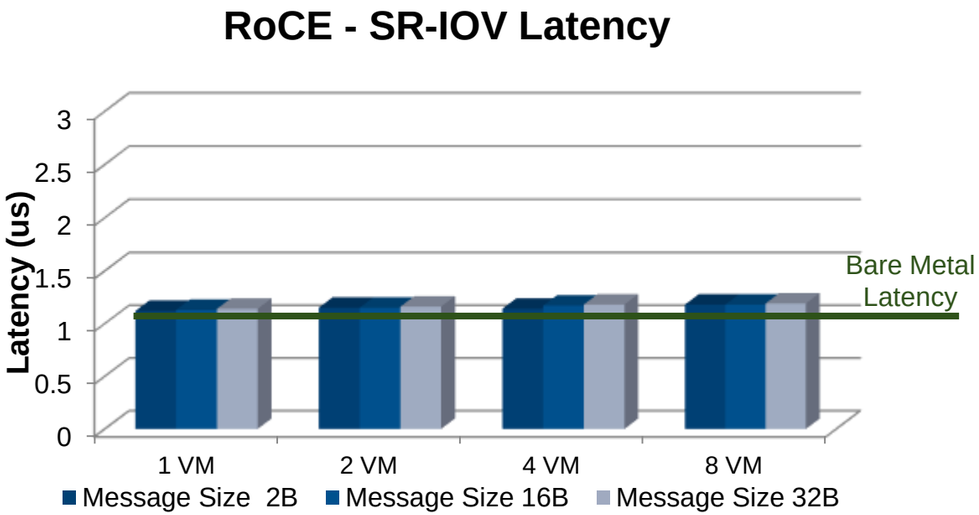
- Cloud Solutions need to address unique technology requirements
  - High End Compute
    - Fastest processors & Memory
    - GPUs
  - Seamless Interconnect
    - High Bandwidth, Low latency
    - OS bypass
  - High performance parallel file systems
    - Lustre, GPFS

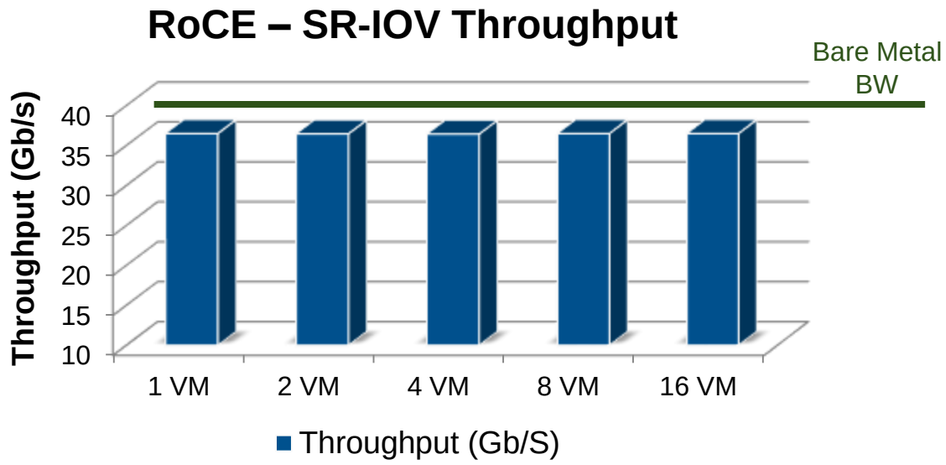
Single Root I/O Virtualization (SR-IOV)

- PCIe device presents multiple instances to the OS/Hypervisor
- Enables Application Direct Access
  - Bare metal performance for VM
  - Reduces CPU overhead
- Enable RDMA to the VM
  - Low latency applications benefit from the Virtual infrastructure
- Now supports also HA & QoS





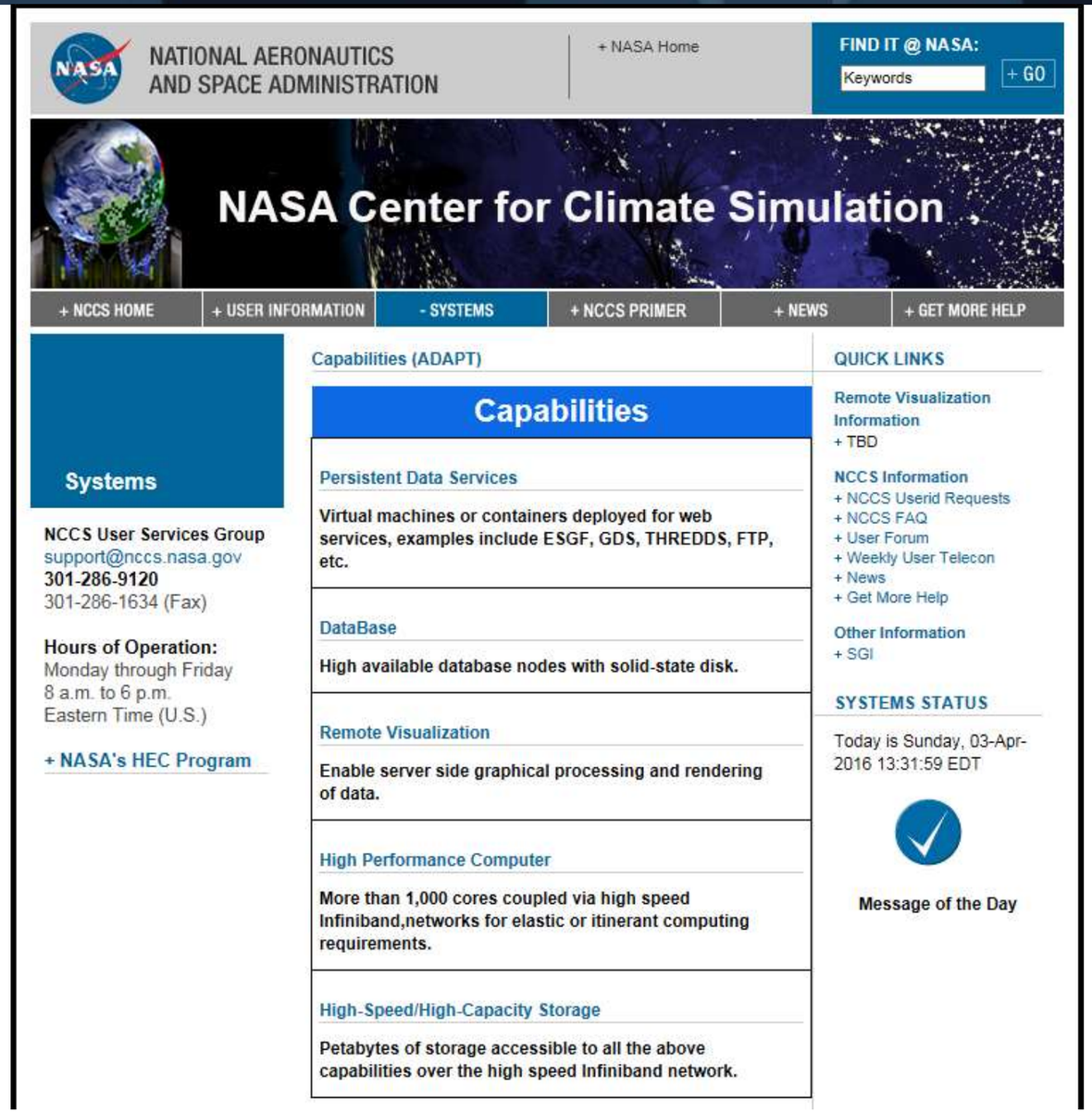




HPC Private Cloud Case Study:

Advanced Data Analytics Platform at the NASA Center for Climate Simulation

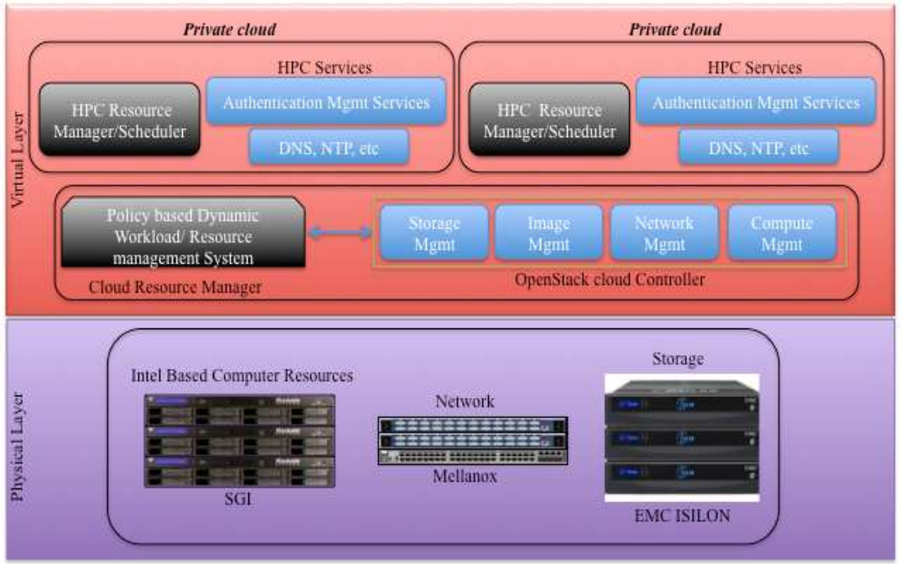
- Usage: Climate Research
- System Capabilities
  - PaaS, VMs, OpenStack
  - 1,000 compute cores
  - 7PB of storage (Gluster)
  - QDR/FDR InfiniBand
  - SR-IOV
- Strategic Objective: Explore the capabilities of HPC in the cloud and prepare the infrastructure for bursting to the public cloud



HPC Private Cloud Case Study:

HPC4Health Consortium, Canada

- Collaborative effort between Toronto's Downtown Hospitals and related health research institutions to address high performance computing (HPC) needs in research environments encompassing patient and other sensitive data.
- System Capabilities
  - 340 SGI compute nodes, 13,024 compute threads
  - 52.7 terabytes of RAM, 306 terabytes of total local disk space and 4 PB of storage
  - InfiniBand, SR-IOV
  - OpenStack
  - Adaptive Computing, Moab HPC Suite
- Each organization has their own dedicated resources that they control plus access to a common shared pool.



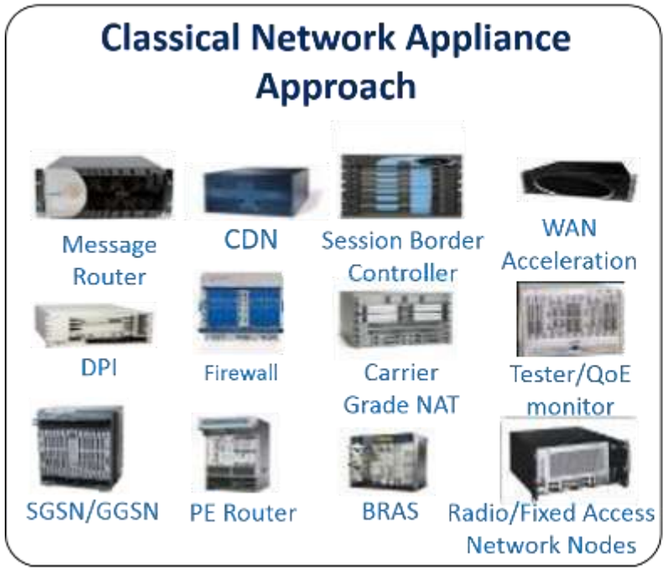
HPC options in the Public Cloud

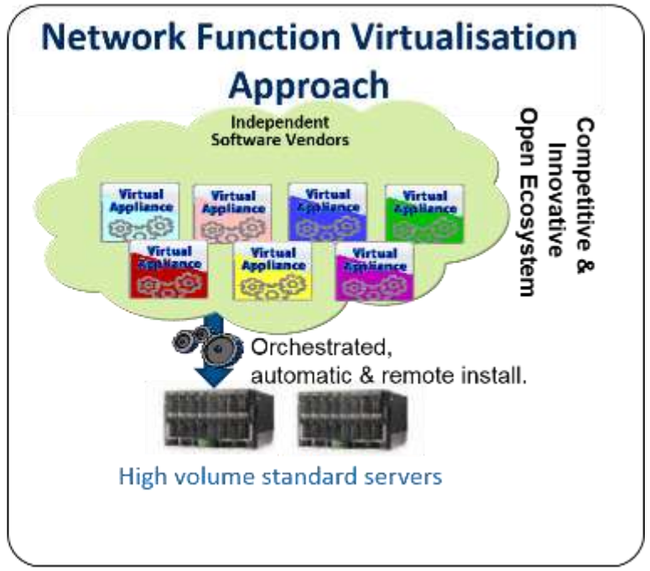
	amazon web services™	Microsoft Azure
Reference	https://aws.amazon.com/hpc/	https://azure.microsoft.com/en-us/documentation/scenarios/high-performance-computing/
High End Compute Nodes	Yes (EC2 C4)	Yes (A8 & A9)
GPU Nodes	Yes	Yes
High Speed Interconnect	10GbE	10GbE and InfiniBand
Non-Blocking Fabric	Yes	Yes
SR-IOV	Yes	Yes
Native RDMA	No	Yes
Parallel File System	Yes	Yes
OS Support	Linux + windows guests	Linux + windows guests
Usage	High End Compute	High End Compute + MPI

Telco / NFV

Network Function Virtualization (NFV) in the Telco Space

- The NFV (Network Function Virtualization) revolution
  - Telcos are moving from proprietary hardware appliances to virtualized servers
  - Benefits:
    - Better time to market: VM bring-up is faster than Appliance procurement and installation
    - Agility and flexibility: Scale up/down, add/enhance services faster at lower cost
    - Reduce Capex and Opex, eliminate vendor lock-in
  - DPDK and line-rate packet processing allow NFV to meet Appliances performance

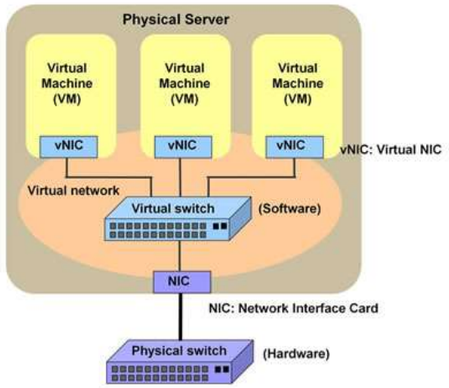


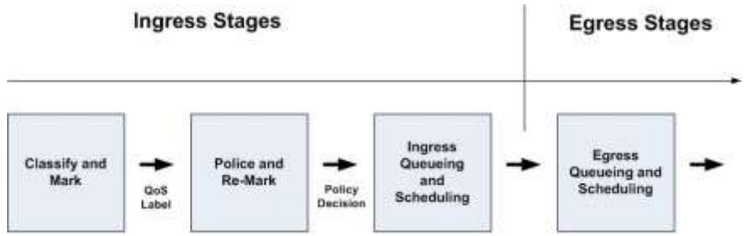


at&t Sprint verizon Deutsche Telekom Telefonica SDN NFV FORUM

NFV vs. Traditional HPC – Key Differences

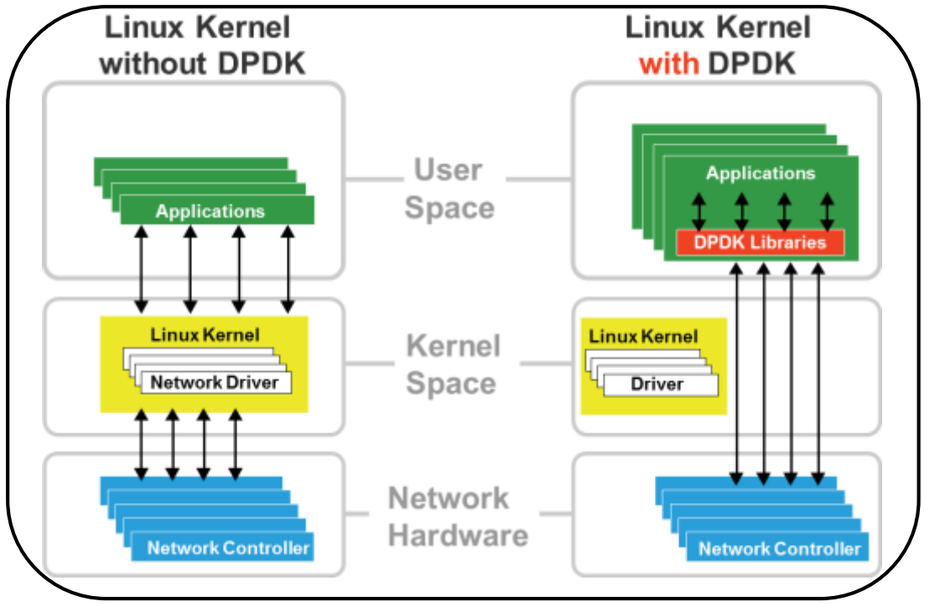
- Small packets → High PPS
- OVS becomes main bottleneck
  - Each packet requires Lookup, classification, encap/decap, QoS, etc in software
  - Linux Kernel today can handle max of 1.5 – 2M PPS in software
- No storage
- Individual I/O – no sync between servers
- Ecosystem: New, from Data Center/ETH vs. IB/MPI Legacy of traditional HPC
- Only Private Cloud at this point

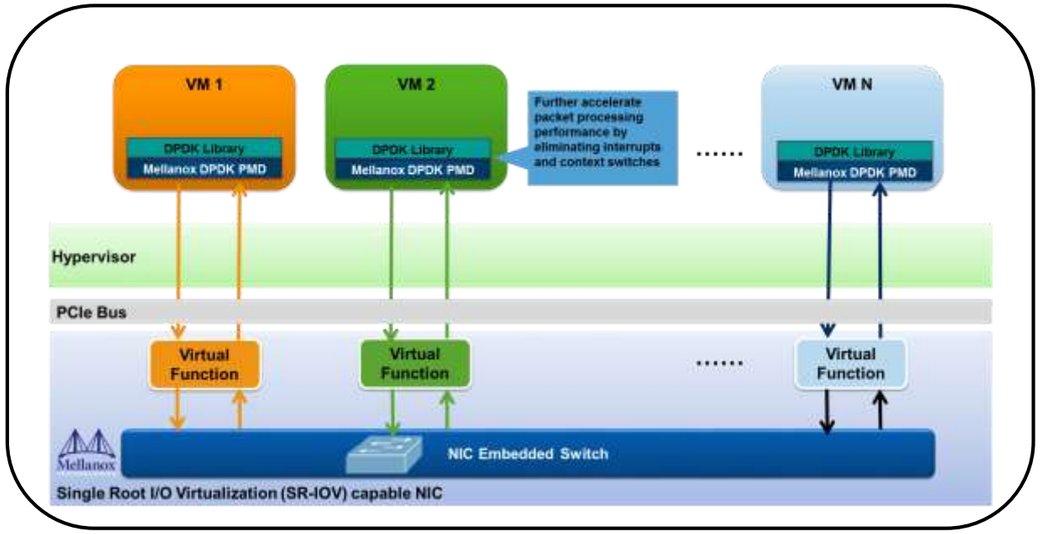




Data Plane Development Kit (DPDK)

- DPDK in a Nutshell
  - DPDK is a set of open source libraries and drivers for fast packet processing (www.dpdk.org)
  - Receive and send packets within the minimum number of CPU cycles
  - 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
- Common Use Cases
  - Router, Security, DPI, Packet Capture
- DPDK in the cloud
  - Accelerate virtual switches (i.e., OVS over DPDK – eg 6Wind)
  - Enable Virtual Network Functions (VNFs)





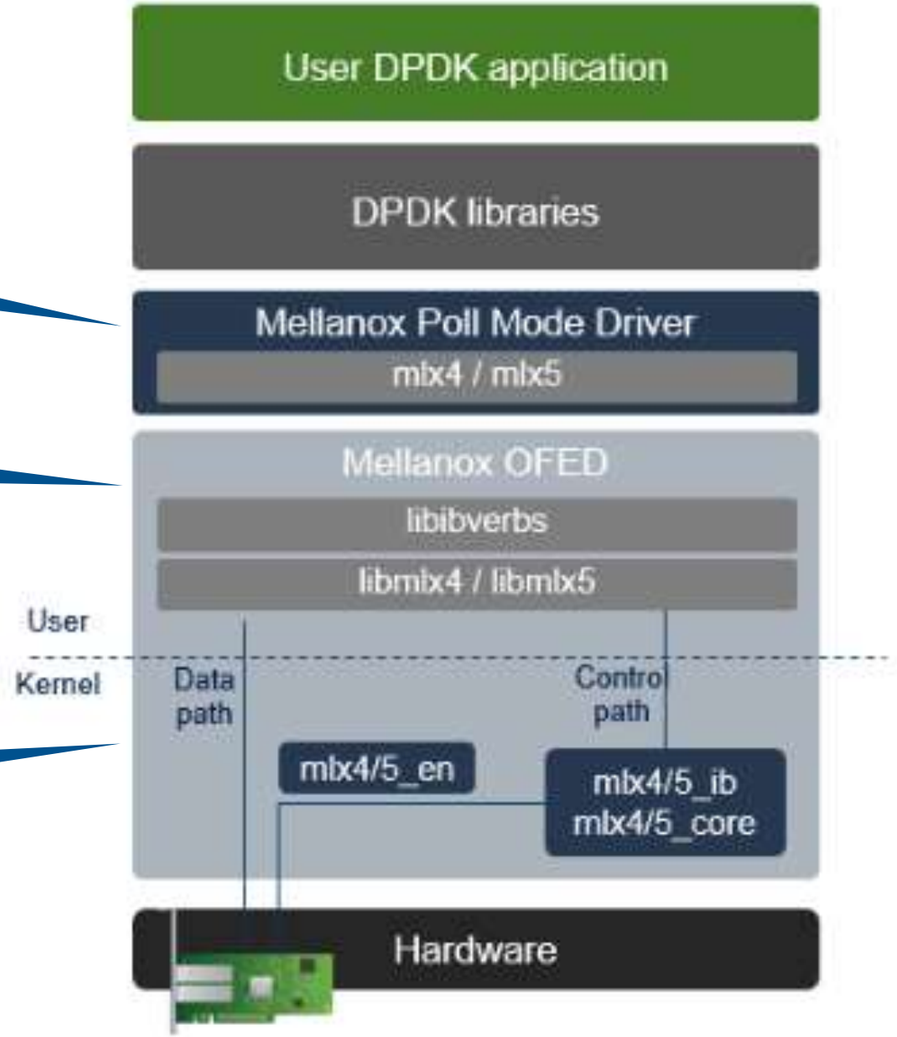
Mellanox DPDK Arch

Mellanox Poll Mode Driver (PMD)

- Running in user space
- Accesses the RX and TX descriptors directly without any interrupts
- Receives, process and deliver packets
- Built on top of libibverbs using the Raw Ethernet verbs API

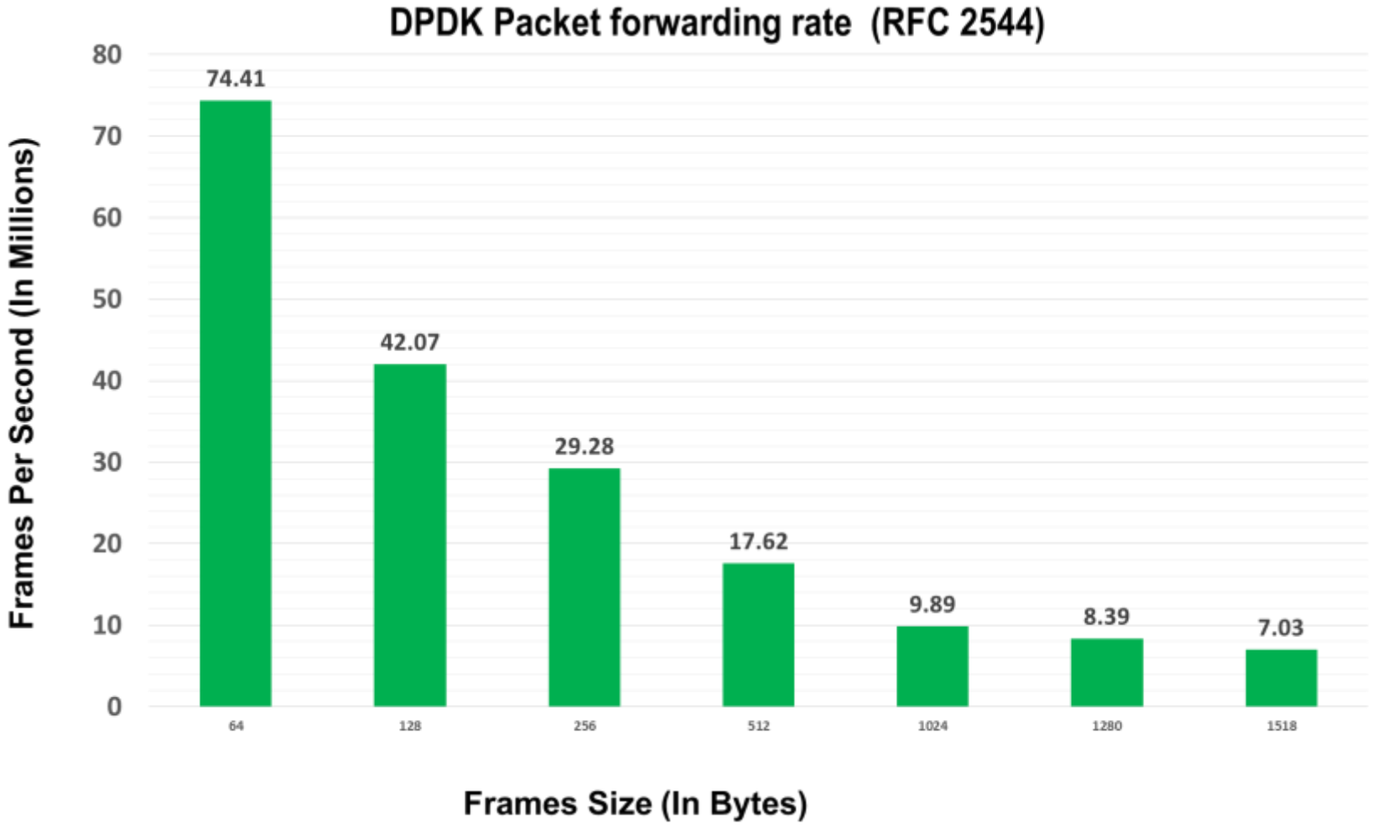
- libmlx4 / libmlx5 are the Mellanox user space drivers for Mellanox NICs
- mlx4_ib / mlx5_ib and mlx4_core / mlx5_core kernel modules used for control path
- mlx4_en / mlx5_en are used for Interface Bring up

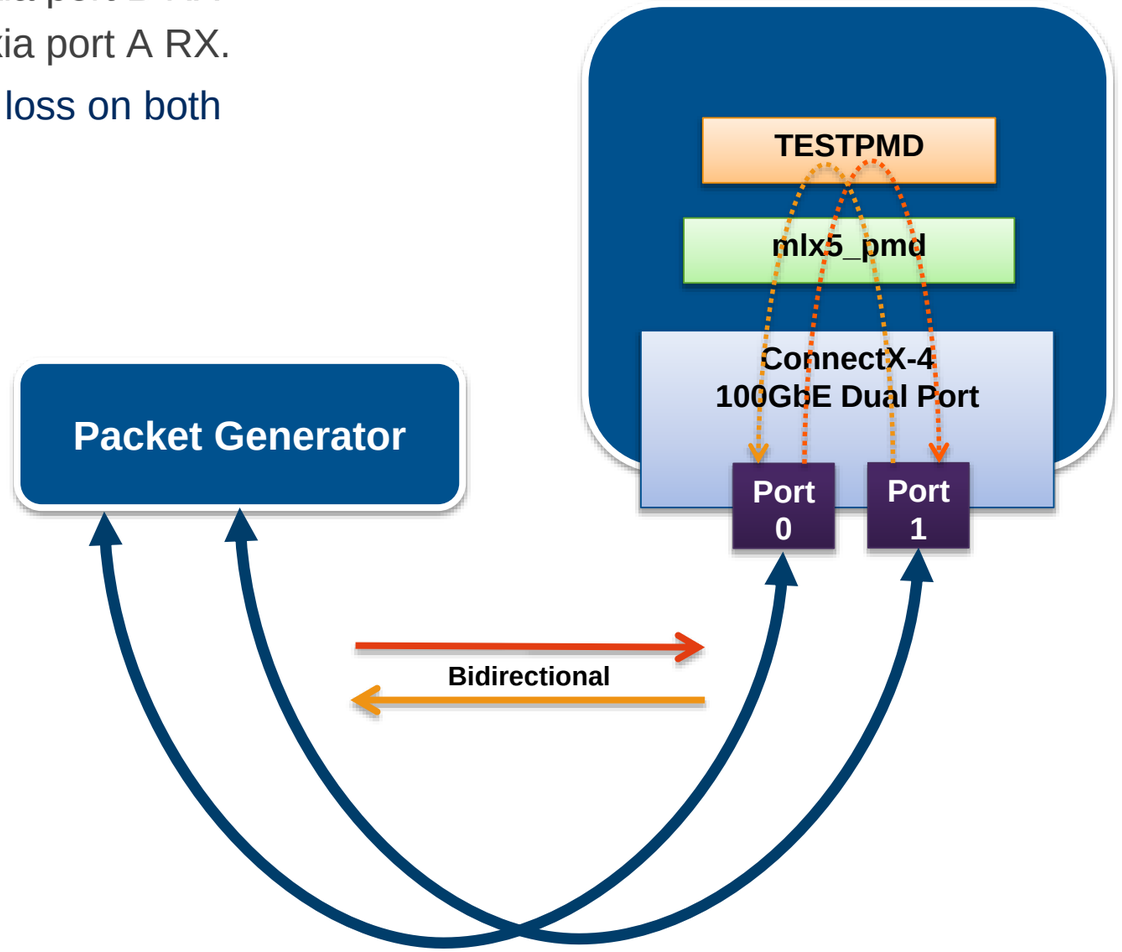
- Mellanox PMD coexists with kernel network interfaces which remain functional
- Ports that are not being used by DPDK can send and receive traffic through the kernel networking stack



Packet Forwarding Rate
ConnectX-4 100GbE dual port, 4 Cores per port

- DPDK IO forwarding, 0 packet loss
- **ConnectX-4 Dual-port Bidirectional:**
  - Ixia port A TX-> ConnectX-4 port 1 RX -> ConnectX-4 port 2 TX -> Ixia port B RX
  - Ixia port B TX-> ConnectX-4 port 2 RX -> ConnectX-4 port 1 TX -> Ixia port A RX.
- Results: Max Ixia port A TX rate + Max Ixia port B TX rate with 0 packet loss on both





Full Virtual Switch Offload

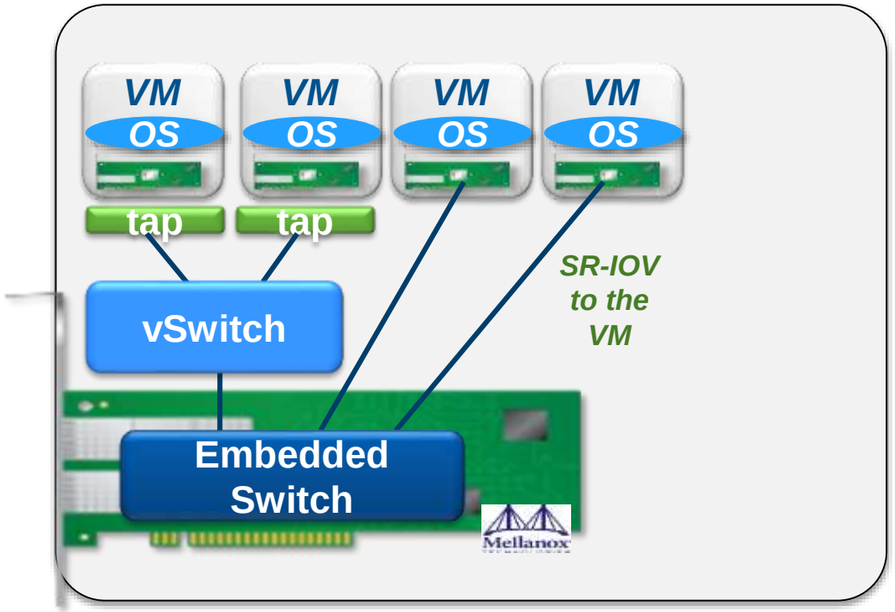
ASAP^2-Direct

Accelerated Switching And Packet Processing (ASAP^2)

- Virtual switches are used as the forwarding plane in the hypervisor
- Virtual switches implement extensive support for SDN (e.g. enforce policies) and are widely used by the industry
- SR-IOV technology allows direct connectivity to the NIC, as such, it bypasses the virtual switch and the policies it can enforce

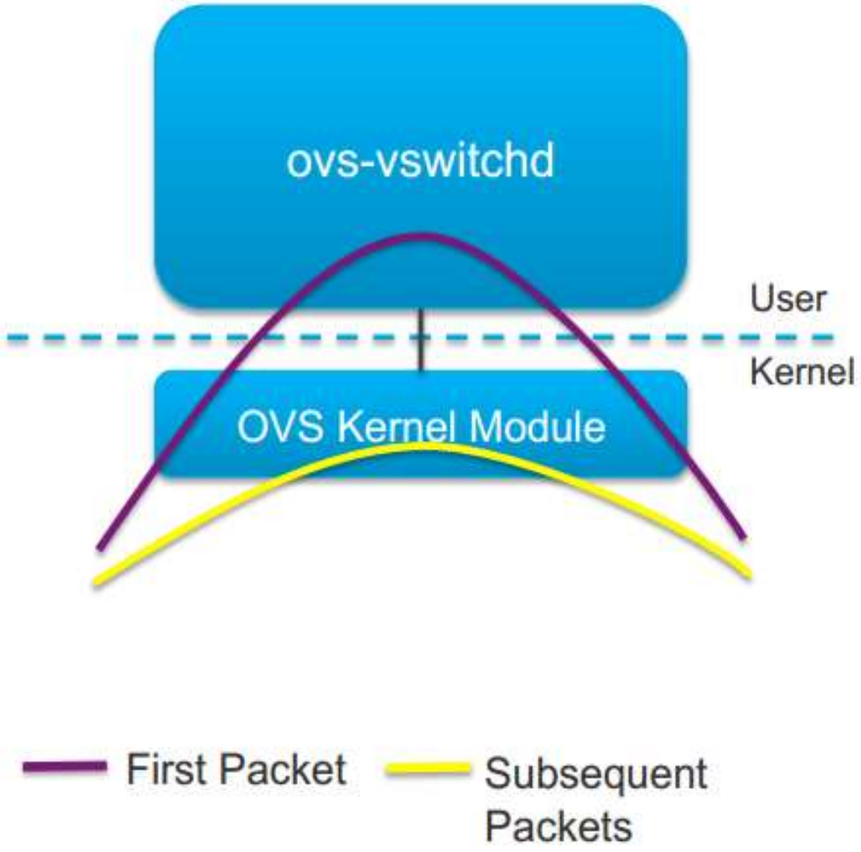
Goal

- Enable SR-IOV data plane with OVS control plane
  - In other words, enable support for most SDN controllers with SR-IOV data plane
- Offload OVS flow handling (classification, forwarding etc.) to Mellanox eSwitch



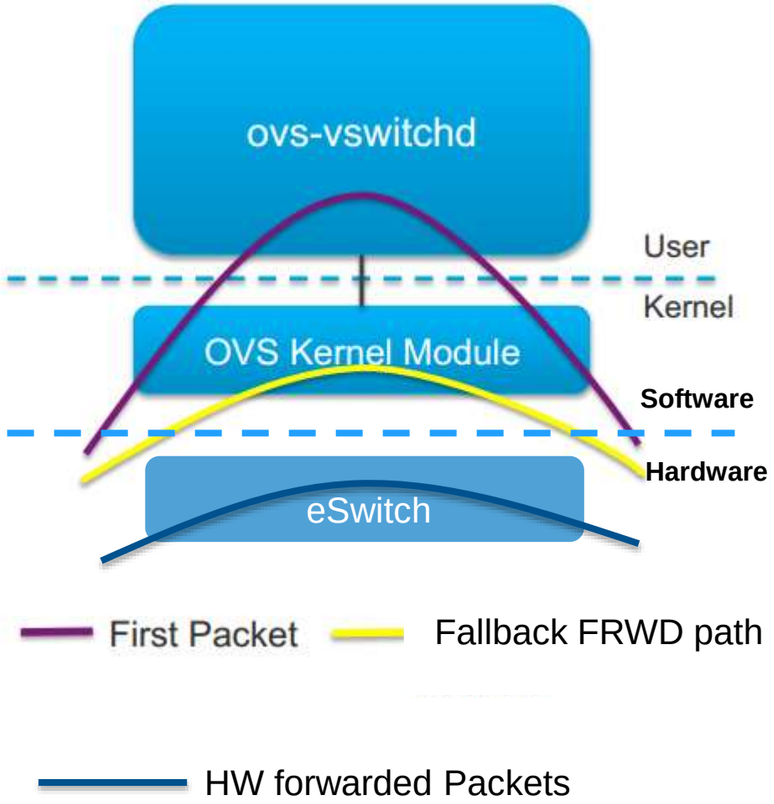
Open vSwitch

- Forwarding
  - Flow-based forwarding
  - Decision about how to process a packet is made in user space
  - First packet of a new flow is directed to ovs-vswitchd, following packets hit cached entry in kernel
- OVS Overview
  - http://openvswitch.org/slides/OpenStack-131107.pdf



Adding the Hardware Layer to the Forwarding Plane

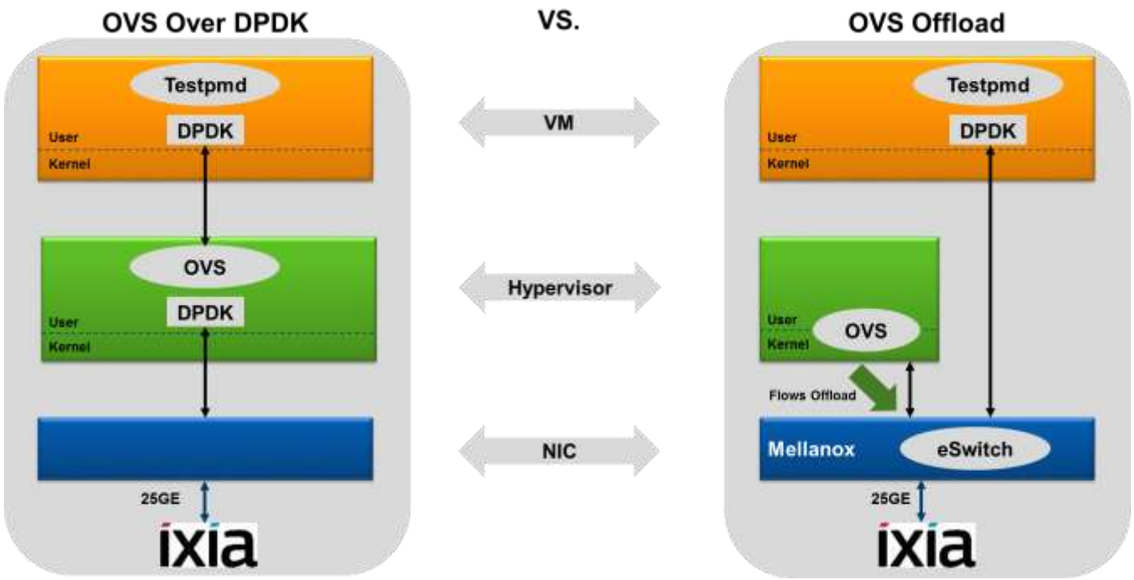
- The NIC Embedded Switch is layered below the kernel datapath
- The Embedded Switch is the first to ‘see’ all packets
- New flow (‘miss’ action) is directed to OVS kernel module
  - Miss in kernel will forward the packet to user space as before
- Decision if to offload the new flow to HW is done by “Offload Policer” based on device capabilities
- Following packets of flow are forwarded by eSwitch -- if offloaded

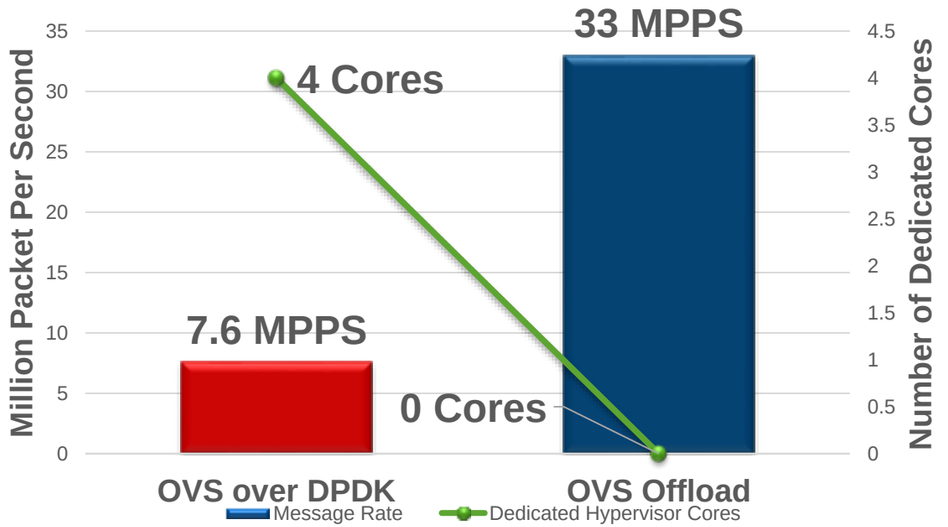


Retain the “first packet” concept (slow path) while enabling the “fast-est” path – via the HW switch by installing the proper flows

OVS over DPDK VS. OVS Offload

- 330% higher message rate compared to OVS over DPDK
  - 33M PPS VS. 7.6M PPS
  - OVS Offload reach near line rate at 25G (37.2M PPS)
- Zero! CPU utilization on hypervisor compared to 4 cores with OVS over DPDK
  - This delta will grow further with packet rate and link speed
- Same CPU load on VM





Summary & Applications for Wall Street

Summary & Conclusions for Wall Street

- Identify your workloads

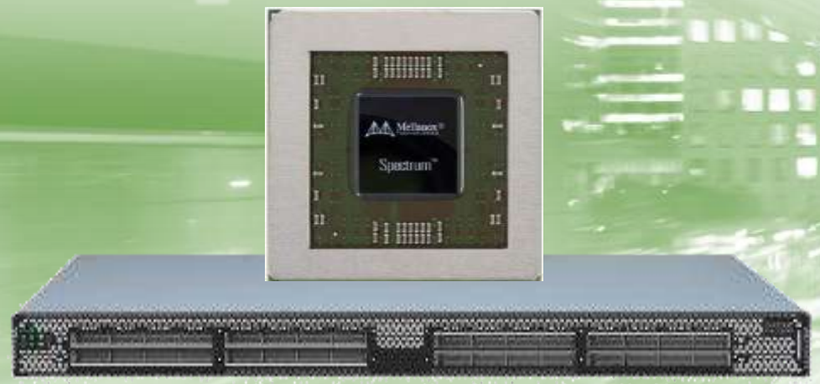
Workload Type	Single / Multi Job	Compute	Network	Storage	Location (Co-lo)
MPI-based Research	Single	Yes	Yes	Yes	No
NFV (Security, Capture)	Multi	Yes	Yes	No	No
Monte-Carlo (Risk/Pricing)	Multi	Yes	Depends	Yes	No
Big Data	Single/Multi	Yes	Depends	Yes	No
High Frequency Trading	Multi	Yes	Yes	No	Yes

- Public, Private or “Burst”
  - TCO
  - Security
  - Performance

- Look at accumulated experience in other industries

Come Visit our Booth @ HPC on Wall Street:
25Gb/s is the new 10, 50 is the new 40, and 100 is the Present

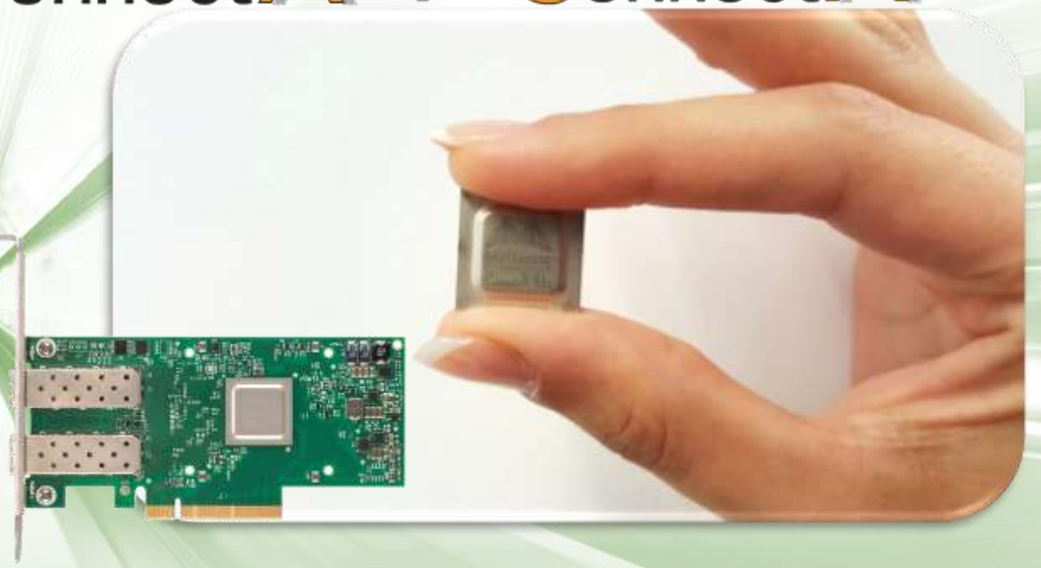
Spectrum™



Flexibility, Opportunities, Speed

Open Ethernet, Zero Packet Loss

ConnectX®-4 ConnectX®-4 Lx



Most Cost-Effective Ethernet Adapter

Same Infrastructure, Same Connectors

One Switch. A World of Options.

25, 50, 100Gb/s at Your Fingertips

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™