OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP

RDMA in Virtualized and Cloud Environments

#OFADevWorkshop

Aaron Blasius, ESXi Product Manager
Bhavesh Davda, Office of CTO
VMware

Takeaways

- It is possible to bring the benefits of virtualization to low latency environments
- VMware is working on virtualization support for host and guest services over RDMA
- Early performance numbers are promising

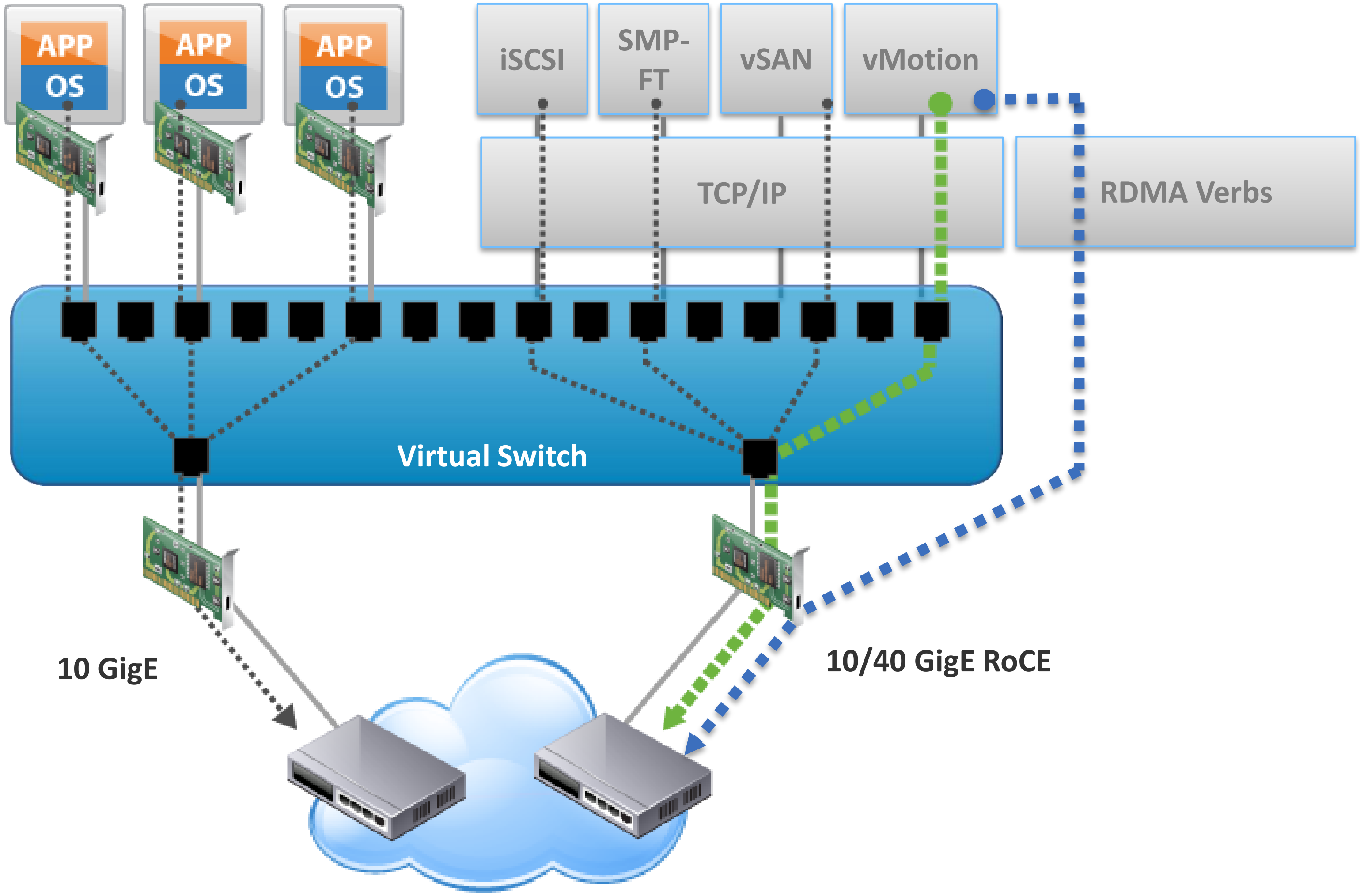
Virtualization of Latency-Sensitive Applications on ESXi

- Historically, virtualization was not suitable for latency-sensitive workloads
- vSphere ESXi 5.5 (2013) introduced an “easy button” for running extremely latency-sensitive workloads
  - Disables Interrupt Coalescing
  - Pins vCPUs to pCPUs
  - Pins down VM memory on local NUMA node
  - Reduces idle guest (HALT) wake-up latencies in VMM

Host-Level RDMA

- Physical RDMA interconnect on ESXi hosts:
  - Support for physical RDMA connections on ESXi hosts (RoCE, iWARP, IB)
  - OFED RDMA stack in ESXi vmkernel
- Use cases:
  - vMotion (Live migration of virtual machines between ESXi hosts)
  - vSAN (Scale-out clustered storage from direct-attached HDDs and SSDs on ESXi hosts)
  - SMP-FT (Lock-step fault tolerance of SMP VMs)
  - NFS
  - iSCSI

RDMA for hypervisor services

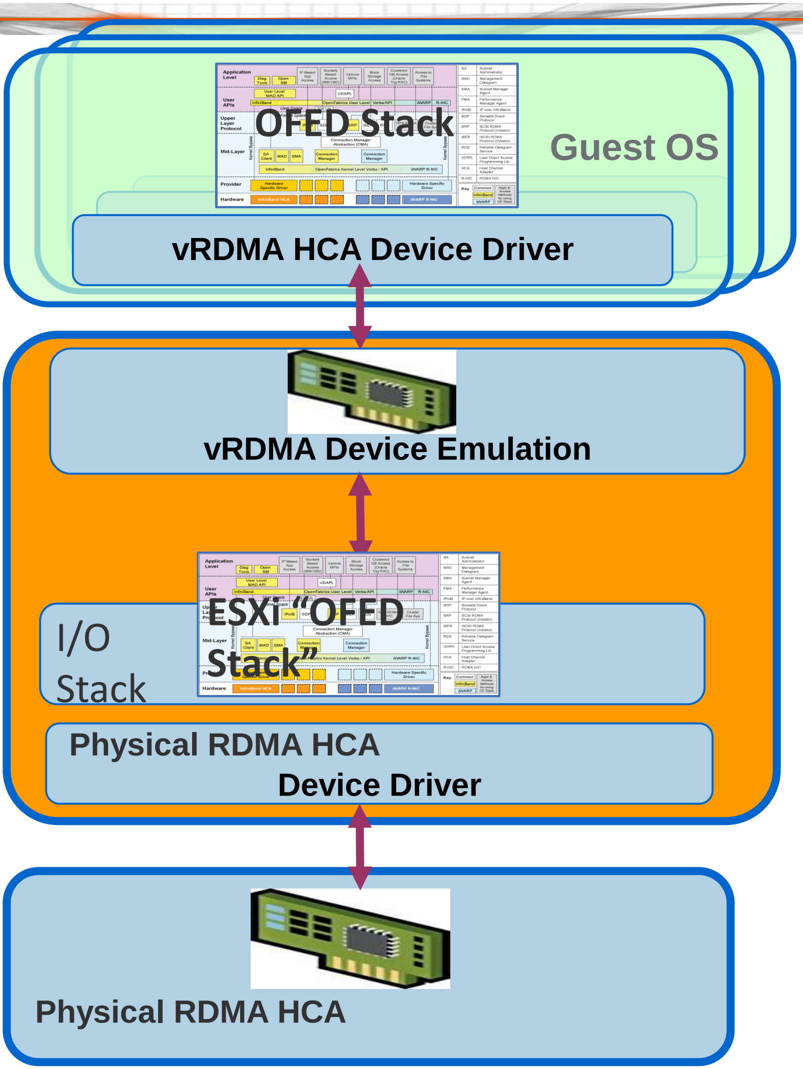


Guest-Level RDMA

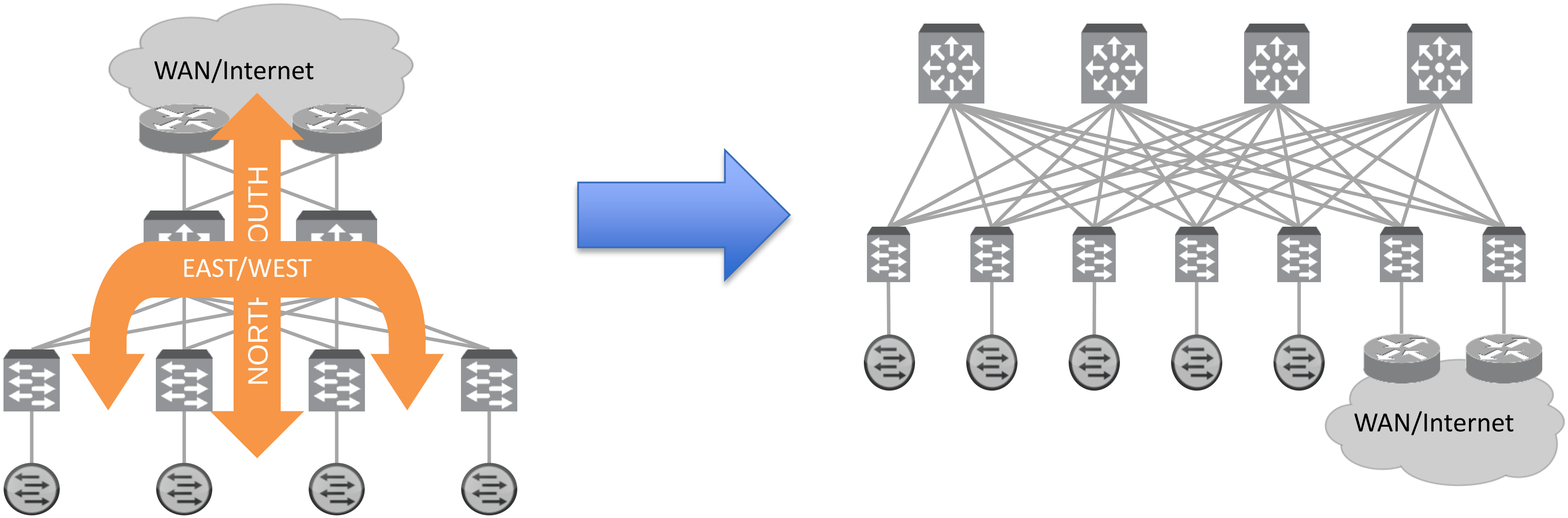
- Proposed paravirtual vRDMA device supports Verbs
  - Compatible with all virtualization features like vMotion, snapshots and checkpoints
  - Lowest latencies for a pure virtual environment, without relying on pass through direct assignment
- Use cases:
  - Scale-out databases
  - Enterprise distributed applications
  - MPI-based HPC applications
  - Faster network attached storage
  - Big data applications

Proposed Paravirtual RDMA HCA (vRDMA) offered to VM

- Paravirtualized device exposed to Virtual Machine
  - Implements Verbs interface
- Device emulated in ESXi hypervisor
  - Translates Verbs from Guest to Verbs to ESXi OFED Stack
  - Guest physical memory regions mapped to ESXi and passed down to physical RDMA HCA
  - Zero-copy DMA directly from/to guest physical memory
  - Completions/interrupts proxied by emulation

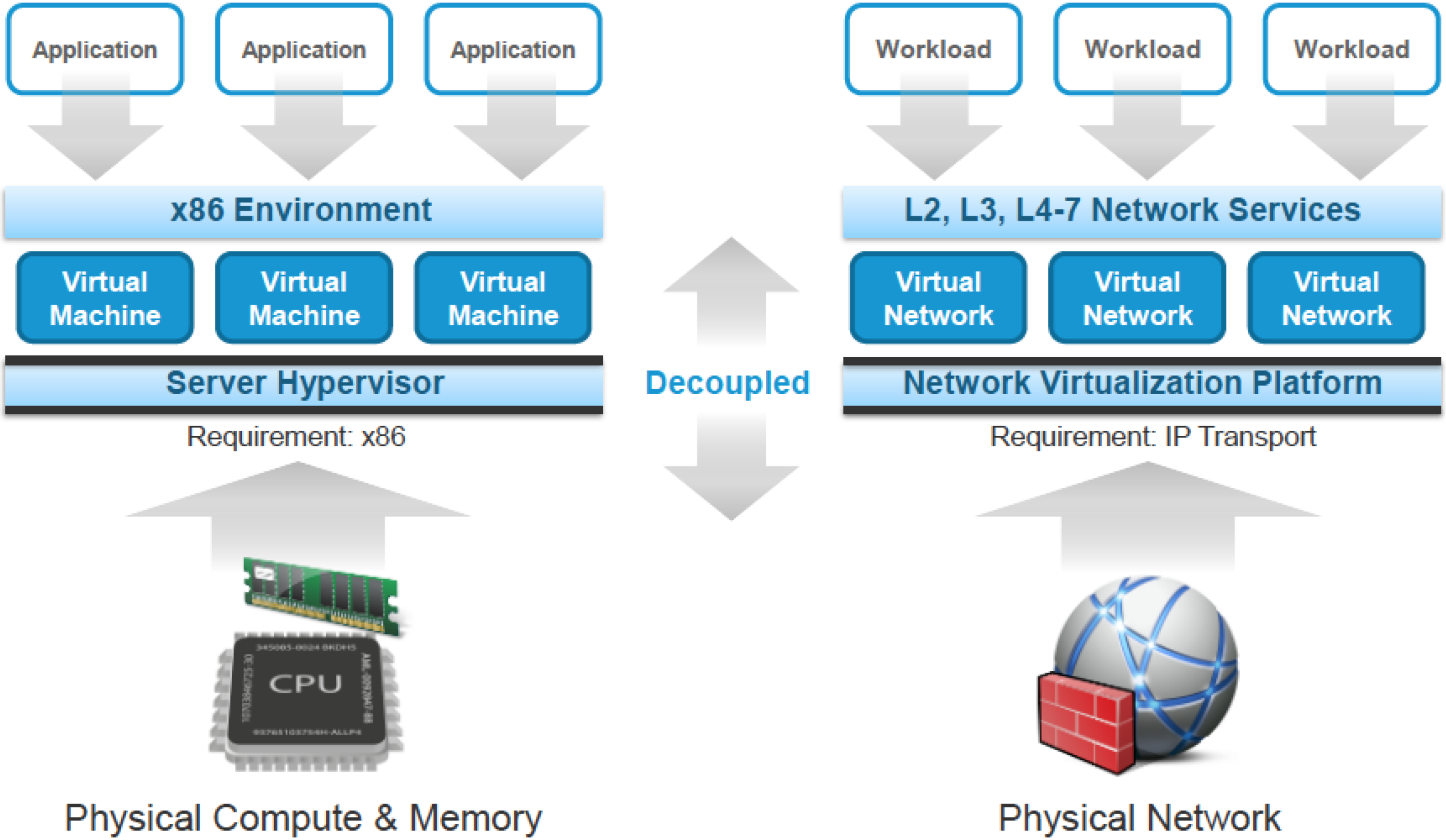


Data Center Networks – the Trend to Fabrics

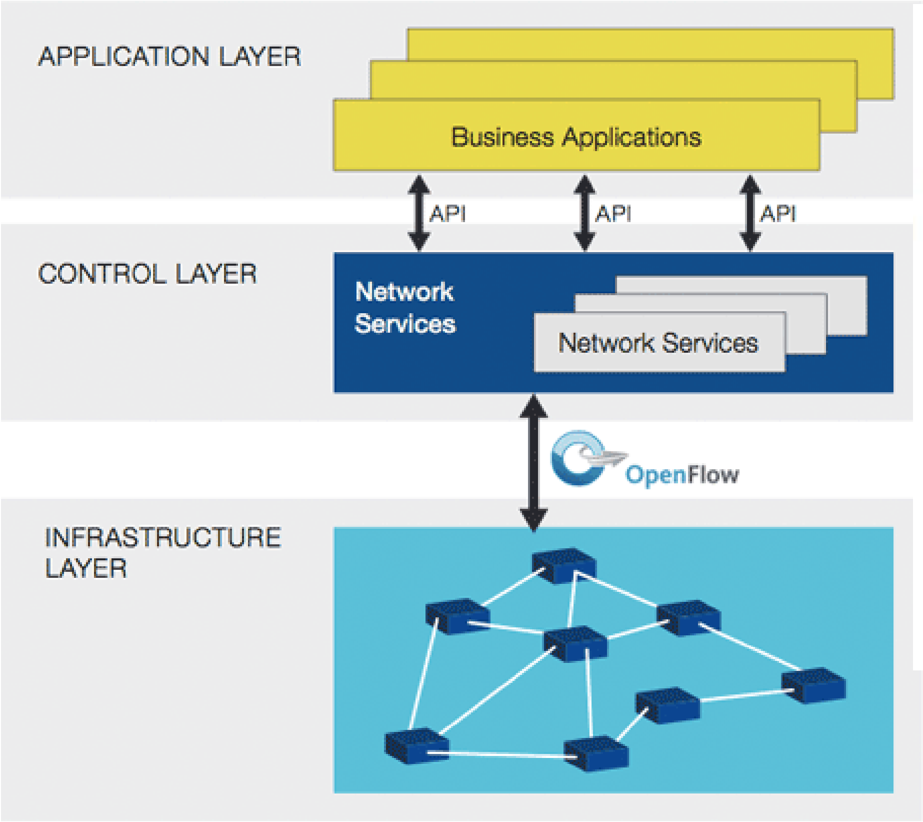


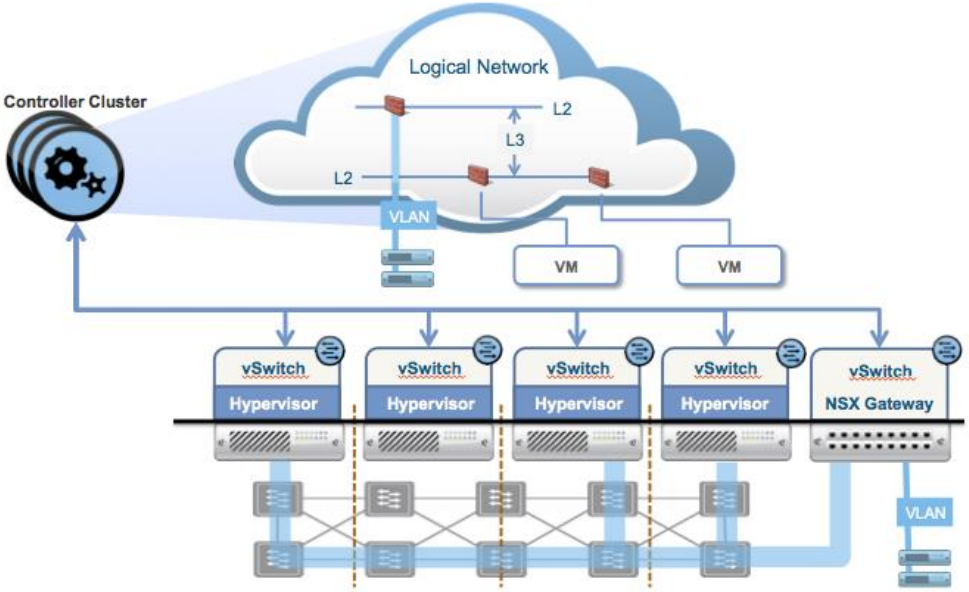
- Increase in East-West traffic due to:
    - Virtualization leading to flexible placement of applications within datacenter
    - Scale-out applications
    - Scale-out hypervisor services
- More uniform bandwidth and latencies
    - Very Similar to HPC network topologies

Network Virtualization



Software Defined Network

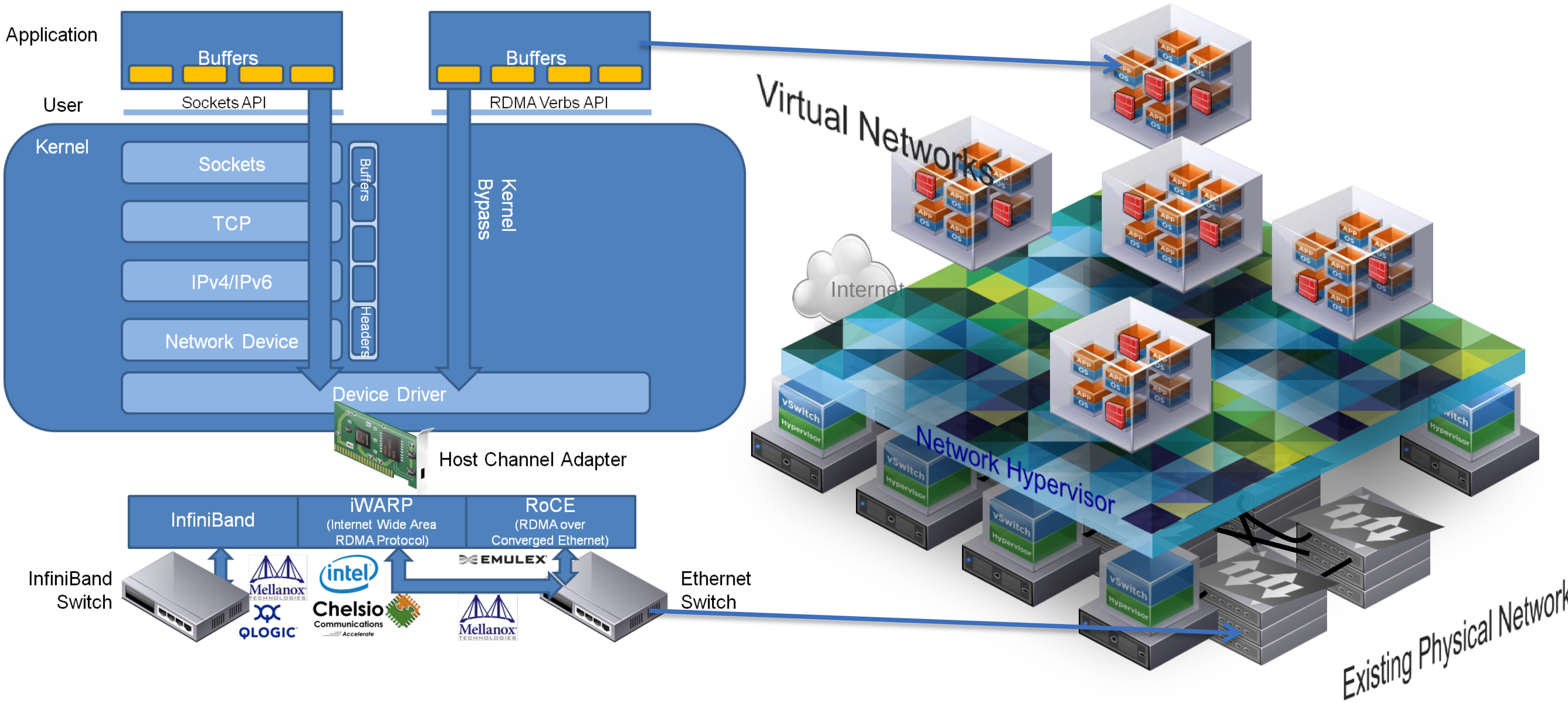




Open Networking Foundation's SDN Architecture

VMware NSX Network Hypervisor Architecture

Impedance Mismatch?



RDMA Requirements for Enterprise and Cloud

- Enterprise applications usually written to socket(2) based frameworks
  - Need to exploit the benefits of RDMA while keeping the socket(2) based API compatibility
  - R-sockets? SDP? IBM JSOR? IBM SMC-R?
- How to exploit the benefits of RDMA (high bandwidth, low latency, CPU offload) in virtualized applications, without losing the benefits of compute (e.g. ESXi) and network (e.g. NSX) virtualization?

Thank You

OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP