Outline

• Brief history of x86 virtualization
• The case for Network Functions Virtualization
• Hypervisor performance characteristics
• Challenges and Call to Action
X86 Virtualization Timeline


VMware founded

Workstation 1.0

ESX Server 1.0

ESX 2.0 (vSMP, 2vcpus)

Workstation 5.5 (64 bit guests)

ESX 3.0 (4 vcpus)

ESX 3.5

ESX 4.0 (8 vcpus)

First KVM release in RHEL5.4

Fusion 1.0

ESXi 6.0 (128 vcpus, 4TB RAM)

Xen 1.0

x86-64 Dual core

Intel VT-x

Intel EPT

AMD-V

AMD RVI

Hardware Virtualization Support

Dual core
X86 Virtualization 1998-2006 (pre-hardware virtualization)

Diagram:
- **Direct Exec (user)**
- **VMM**
- **BT (kernel)**

Flow:
- Faults, syscalls, interrupts
- IRET, SYSRET
X86 Virtualization 2006-present (hardware virtualization)

- Apps (Ring 3)
- Guest OS (Ring 0)
- VMM

- VM exit
- VM enter
The State of Virtualization

Datacenter Workloads

- Virtualized
- Unvirtualized
## The State of Virtualization

### Datacenter Workloads

<table>
<thead>
<tr>
<th>Workload</th>
<th>Average Virtualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Exchange</td>
<td>60%</td>
</tr>
<tr>
<td>Microsoft SQL</td>
<td>59%</td>
</tr>
<tr>
<td>Microsoft SharePoint</td>
<td>58%</td>
</tr>
<tr>
<td>High Performance Computing</td>
<td>56%</td>
</tr>
<tr>
<td>Oracle Applications</td>
<td>56%</td>
</tr>
<tr>
<td>Oracle Databases</td>
<td>56%</td>
</tr>
<tr>
<td>Oracle Middleware</td>
<td>53%</td>
</tr>
<tr>
<td>Oracle Database</td>
<td>53%</td>
</tr>
<tr>
<td>Other Middleware</td>
<td>51%</td>
</tr>
<tr>
<td>Other Mission-critical</td>
<td>49%</td>
</tr>
<tr>
<td>Average</td>
<td>58%</td>
</tr>
</tbody>
</table>

### Business Critical Application Virtualization

Source: Cloud Infrastructure Annual Primary Research, 2Q13 (N=576)
Outline

• Brief history of x86 virtualization
• The case for Network Functions Virtualization
• Hypervisor performance characteristics
• Challenges and Call to Action
NFV in a nutshell

Classical Network Appliance Approach

- Message Router
- CDN
- Session Border Controller
- WAN Acceleration
- DPI
- Firewall
- Carrier Grade NAT
- Tester/QoE monitor
- SGSN/GGSN
- PE Router
- BRAS
- Radio/Fixed Access Network Nodes

- Fragmented non-commodity hardware.
- Physical install per appliance per site.
- Hardware development large barrier to entry for new vendors, constraining innovation & competition.

Independent Software Vendors

Orchestrated, automatic & remote install.

Standard High Volume Servers

Standard High Volume Storage

Standard High Volume Ethernet Switches

Network Functions Virtualisation Approach
NFV Motivation: From Margin Erosion to Sustainable Profitability

Margin Erosion

Network Investment Funded by CSP

Customer

Traffic and Infrastructure

OTT Services

Cost per Subscriber

ARPU
NFV Motivation: From Margin Erosion to Sustainable Profitability

Margin Erosion

Network Investment Funded by CSP
Customer
Traffic and Infrastructure
OTT Services

Sustainable Profitability

Cost per Subscriber
ARPU
Cloud Operating Model
Cloud Platform
NFV Motivation: From Margin Erosion to Sustainable Profitability

Margin Erosion

Network Investment Funded by CSP

Traffic and Infrastructure

Customer

Cost per Subscriber

Service Differentiation

ARPU

Cloud Platform

Service Agility

Cloud Operating Model

Sustainable Profitability

OTT Services
NFV Architecture

Operations and Business Support Systems (OSS / BSS)

Service, VNF & Infrastructure Description

Orchestrator

VNF Managers

Virtual Infrastructure Manager

NFV M&O

VNF

EMS1

VNF1

EMS2

VNF2

EMS3

VNF3

NFVI

Virtual Compute

Virtual Storage

Virtual Network

Virtualization Layer

Compute Hardware

Storage Hardware

Network Hardware
Many Use Cases
Many Use Cases
Example Use Case: 4G LTE Network

- **LTE Network Elements**

  - Evolved UTRAN (E-UTRAN)
  - Evolved Packet Core (EPC)
  - HSS
  - MME: Mobility Management Entity
  - PCRF: Policy & Charging Rule Function
  - S6a
  - S10
  - S7
  - Rx+
  - SGi
  - Serving Gateway
  - PDN Gateway
  - SAE Gateway
  - PDN

- LTE Network Elements:
  - X2
  - LTE-Uu
  - Evolved Node B (eNB)
  - LTE-UE
  - LTE-Uu
LTE: vEPC – MME, PGW, SGW

Deployed in production in multiple carriers worldwide.
LTE: vEPC – MME, PGW, SGW

LTE: vEPC – MME, PGW, SGW

Deployed in production in multiple carriers worldwide

Virtualization Layer

- ESXi Hypervisor
- VSAN
- NSX

Virtualization Layer

- Computing Hardware
- Storage Hardware
- Network Hardware

VNF

- MME/SGSN
- PGW/SGW/GGSN
- HSS
- PCRF

VNF Vendor

- EMS
- OSS/BSS

NFV Orchestrator

- Vendor’s VNF Mgr (Optional)
- Generic VNF Mgr

vCloud Director

OpenStack

vCenter
Outline

• Brief history of x86 virtualization
• The case for Network Functions Virtualization
• Hypervisor performance characteristics
• Challenges and Call to Action
Why performance is important to NFV

- NFV workloads different from enterprise / datacenter workloads
- NFV Performance & Portability Best Practices (ETSI GS NFV-PER 001) classification
  - **Data plane workloads**: tasks related to packet handling in an end-to-end communication between edge applications…expected to be very intensive in I/O operations and memory R/W operations. E.g. CDN cache, router, IPsec tunneling
    - High packet rates
  - **Control plane workloads**: communication between NFs that is not directly related to the end-to-end data communication between edge applications, like session management, routing or authentication. Compared to data plane workloads, control plane workloads are expected to be much less intensive in terms of transactions per second, while the complexity of the transactions might be higher
    - CPU and memory intensive
  - **Signal processing workloads**: NF tasks related to digital processing such as the FFT decoding and encoding in a C-RAN Base Band Unit (BBU). Expected to be very intensive in CPU processing capacity and high delay-sensitive
    - Latency and jitter sensitive
ESXi Networking Datapath Overview
ESXi Networking Datapath Overview

- Message copy from application to GOS (kernel)
ESXi Networking Datapath Overview

• Message copy from application to GOS (kernel)
• GOS (network stack) + vNIC driver queues packet for vNIC
ESXi Networking Datapath Overview

- Message copy from application to GOS (kernel)
- GOS (network stack) + vNIC driver queues packet for vNIC
- VM exit to VMM/Hypervisor
ESXi Networking Datapath Overview

- Message copy from application to GOS (kernel)
- GOS (network stack) + vNIC driver queues packet for vNIC
- VM exit to VMM/Hypervisor
- vNIC implementation emulates DMA from VM, sends to vSwitch
ESXi Networking Datapath Overview

- Message copy from application to GOS (kernel)
- GOS (network stack) + vNIC driver queues packet for vNIC
- VM exit to VMM/Hypervisor
- vNIC implementation emulates DMA from VM, sends to vSwitch
- vSwitch queues packet for pNIC
ESXi Networking Datapath Overview

- Message copy from application to GOS (kernel)
- GOS (network stack) + vNIC driver queues packet for vNIC
- VM exit to VMM/Hypervisor
- vNIC implementation emulates DMA from VM, sends to vSwitch
- vSwitch queues packet for pNIC
- pNIC DMAs packet and transmits on the wire
Guest OS bypass and hypervisor bypass

Hypervisor provides:
• Live migration (vMotion)
• Network virtualization (VXLAN, STT, Geneve)
• Virtual network monitoring, troubleshooting, stats
• High availability
• Fault tolerance
• Distributed resource scheduling
• VM snapshots
• Hot add/remove of devices, vCPUs, memory
Guest OS bypass and hypervisor bypass

Hypervisor provides:
• Live migration (vMotion)
• Network virtualization (VXLAN, STT, Geneve)
• Virtual network monitoring, troubleshooting, stats
• High availability
• Fault tolerance
• Distributed resource scheduling
• VM snapshots
• Hot add/remove of devices, vCPUs, memory
Guest OS bypass and hypervisor bypass

Hypervisor provides:
- Live migration (vMotion)
- Network virtualization (VXLAN, STT, Geneve)
- Virtual network monitoring, troubleshooting, stats
- High availability
- Fault tolerance
- Distributed resource scheduling
- VM snapshots
- Hot add/remove of devices, vCPUs, memory
Guest OS bypass and hypervisor bypass

Hypervisor provides:
- Live migration (vMotion)
- Network virtualization (VXLAN, STT, Geneve)
- Virtual network monitoring, troubleshooting, stats
- High availability
- Fault tolerance
- Distributed resource scheduling
- VM snapshots
- Hot add/remove of devices, vCPUs, memory

Guest OS bypass
Hypervisor bypass
The Latency Sensitivity Feature in vSphere 5.5
The Latency Sensitivity Feature in vSphere 5.5

- Latency-sensitivity Feature
  - Minimize virtualization overhead
  - Achieve near bare-metal performance
ESXi 5.5 Latency and Jitter Improvements

• New virtual machine property: “Latency sensitivity”
  – High => lowest latency
  – Medium => low latency

• Exclusively assign physical CPUs to virtual CPUs of “Latency Sensitivity = High” VMs
  – Physical CPUs not used for scheduling other VMs or ESXi tasks

• Idle in Virtual Machine monitor (VMM) when Guest OS is idle
  – Lowers latency to wake up the idle Guest OS, compared to idling in ESXi vmkernel

• Disable vNIC interrupt coalescing

• For DirectPath I/O, optimize interrupt delivery path for lowest latency

• Make ESXi vmkernel more preemptible
  – Reduces jitter due to long-running kernel code
ESXi 6.0 Network Latencies and Jitter

Single 4-vCPU VM to Native, Ubuntu 14.04 LTS, RTT from `ping -i 0.00001 -c 1000000`
Intel Xeon E5-2697 v2 @ 2.70 GHz, Intel 82599EB PNIC
ESXi Real-time Virtualization Performance

- cyclicstest –p 99 –a 1 –m –n –D 10m –q

<table>
<thead>
<tr>
<th>ESXi 5.0</th>
<th>ESXi 6.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min 5 µs</td>
<td>Min 2 µs</td>
</tr>
<tr>
<td>Max 1676 µs</td>
<td>Max 78 µs</td>
</tr>
<tr>
<td>Avg 13 µs</td>
<td>Avg 4 µs</td>
</tr>
<tr>
<td>99%-ile 56 µs</td>
<td>99%-ile 5 µs</td>
</tr>
<tr>
<td>99.99%-ile 118 µs</td>
<td>99.99%-ile 7 µs</td>
</tr>
</tbody>
</table>
ESXi MPI InfiniBand Latencies

MPI PingPong Latency

- Half Round-trip Latency (µs)
- Message size (bytes)

- Native
- ESX 5.5
- ESX 6.0
- ESX 6.0u1

- osu_latency benchmark
- Open MPI 1.6.5
- MLX OFED 2.2
- RedHat 6.5
- HP DL380p Gen8
- Mellanox FDR InfiniBand
Data Plane Development Kit (DPDK)

Libraries for network application development on Intel Platforms
- Speeds up networking functions
- Enables user space application development
- Facilitates both run-to-completion and pipeline models

Free, Open-sourced, BSD Licensed
- Git: http://dpdk.org/git/dpdk

Scales from Intel Atom to multi-socket Intel Xeon architecture platforms

Rich selection of pre-built example applications
Data Plane Development Kit (DPDK)

- Speeds up networking functions
- Enables user space application development
- Facilitates both run-to-completion and pipeline models

Free, Open-sourced, BSD Licensed
- Git: [http://dpdk.org/git/dpdk](http://dpdk.org/git/dpdk)

Scales from Intel Atom to multi-socket Intel Xeon architecture platforms

Rich selection of pre-built example applications
DPDK in vSphere ESXi Environment
DPDK in vSphere ESXi Environment

Virtual Machine

Unmodified App

VMware ESXi

VMXNET3

vSwitch

Linux Kernel

No bypass
DPDK in vSphere ESXi Environment

Virtual Machine

Unmodified App

Linux Kernel

VMXNET3

VMware ESXi

vSwitch

No bypass

Guest OS bypass

Virtual Machine

DPDK App

Linux Kernel

VMXNET3

VMware ESXi

vSwitch
DPDK in vSphere ESXi Environment

Virtual Machine

Unmodified App

Linux Kernel

VMware ESXi

vSwitch

VMXNET3

Guest OS bypass

DirectPath I/O

No bypass

DPDK App

IXGBE

Hypervisor bypass +

Guest OS bypass

Virtual Machine

DPDK App

Linux Kernel

VMware ESXi

vSwitch

VMXNET3

IXGBE VF

SR-IOV

Virtual Machine

DPDK App

Linux Kernel

VMware ESXi

vSwitch

Virtual Machine

DPDK App

Linux Kernel

VMware ESXi

vSwitch

Virtual Machine

DPDK App

Linux Kernel

VMware ESXi

vSwitch

IXGBE VF

SR-IOV

No bypass

Guest OS bypass

Guest OS bypass +

Hypervisor bypass
DPDK in vSphere ESXi Environment

Virtual Machine
- Unmodified App
- Linux Kernel
- VMware ESXi
- vSwitch

Virtual Machine
- DPDK App
- Linux Kernel
- VMware ESXi
- vSwitch

Virtual Machine
- DPDK App
- Linux Kernel
- VMware ESXi
- vSwitch

Virtual Machine
- DPDK App
- Linux Kernel
- VMware ESXi
- vSwitch

VMXNET3
- VMware ESXi
- vSwitch

DirectPath I/O
- Guest OS bypass + Hypervisor bypass

SR-IOV
- Guest OS bypass + Hypervisor bypass

No bypass

Guest OS bypass
Unrealistic Performance Characterization Benchmark Application

Packet Generator
64-byte UDP single flow

Physical 10G switch

pNIC1

pNIC2

Virtual network connection

vNIC1

vNIC2

VNF VM running DPDK l2fwd

Hypervisor
Unrealistic Performance Characterization Benchmark Application

1. No packet processing in the VM
Unrealistic Performance Characterization Benchmark Application

1. No packet processing in the VM
2. No memory or cache accesses in VM
Unrealistic Performance Characterization Benchmark Application

1. No packet processing in the VM
2. No memory or cache accesses in VM
3. Port-to-port packet forwarding best done on physical switch via ASICs
Unrealistic Performance Characterization Benchmark Application

1. No packet processing in the VM
2. No memory or cache accesses in VM
3. Port-to-port packet forwarding best done on physical switch via ASICs
4. No real telco network has only 64-byte UDP packets for a single flow
Unrealistic Performance Characterization Benchmark Application

1. No packet processing in the VM
2. No memory or cache accesses in VM
3. Port-to-port packet forwarding best done on physical switch via ASICs
4. No real telco network has only 64-byte UDP packets for a single flow

“Congratulations! You do nothing faster than anyone else!”
[Lori MacVittie, f5, 2010-13-10]
Realistic Performance Characterization Benchmark Application

Intel Virtualized Broadband Remote Access Server (dppd)
NUMA I/O

VMs

Hypervisor

HW

I/O Device

NUMA0

1

NUMA1

2 3

Thread
NUMA I/O

VMs

Hypervisor

HW

I/O Device

Thread

Affinity

NUMA0

NUMA1

1 2 3

1 2 3 4
Improvement due to NUMA I/O

- Test Setup:
  - 1 VM, 1vCPU
  - NUMA: 2 sockets
  - Workload: TCP bulk, 4 sessions
  - I/O Device: Mellanox 40GigE

Transmit  Receive
Improvement due to NUMA I/O

Throughput, CPU efficiency

- Throughput (Gbps)
- CPU Utilization/Gbps

16% Improvement due to NUMA I/O

Test Setup:
- 1 VM, 1vCPU
- NUMA: 2 sockets
- Workload: TCP bulk, 4 sessions
- I/O Device: Mellanox 40GigE
Improvement due to NUMA I/O

Throughput, CPU efficiency

- Test Setup:
  - 1 VM, 1vCPU
  - NUMA: 2 sockets
  - Workload: TCP bulk, 4 sessions
  - I/O Device: Mellanox 40GigE
NFV Challenges and Call to Action

- Performance focus
  - Reminiscent of the move from PSTN to VOIP
  - Shift from using unrealistic benchmarks
- Orchestration
- Matching reality to expectations
- Incumbent pushback
NFV Challenges and Call to Action

- Performance focus
  - Reminiscent of the move from PSTN to VOIP
  - Shift from using unrealistic benchmarks
- Orchestration
- Matching reality to expectations
- Incumbent pushback

Design NFV Infrastructure for both existing applications and hardware platforms as well as unknown future applications and hardware platforms

The IP Hourglass (Steve Deering, Dave Clark)
NFV Challenges and Call to Action

- Performance focus
  - Reminiscent of the move from PSTN to VOIP
  - Shift from using unrealistic benchmarks
- Orchestration
- Matching reality to expectations
- Incumbent pushback

Design NFV Infrastructure for both existing applications and hardware platforms as well as unknown future applications and hardware platforms

Improve OS and hypervisor performance instead of bypassing and losing capabilities
Thank you!

bhavesh@vmware.com