High Performance Linux Virtual Machine on Microsoft Azure: SR-IOV Networking & GPU pass-through

Kylie Liang

Microsoft Enterprise Open Source Group

Agenda

- Our Linux Journey
- Hyper-V Architecture
- PCIe Pass-through / DDA

  Accelerated GPU Experience in Azure
- SR-IOV Networking

  Accelerated Networking in Azure
- Q&A

Microsoft ❤ Linux

Business ▸ The Channel The Register

Redmond top man Satya Nadella: 'Microsoft LOVES Linux'

Open-source 'love' fairly runneth over at cloud event

CSDN首页 >

微软CEO亲口承认：微软就是热爱Linux

发表于 2014-10-21 16:44 | 7447次阅读 | 来源 CSDN CODE | 0 条评论 | 作者 CSDN CODE

微软 Satya Nadella 鲍尔默 Azure

摘要：据国外媒体报道，近日在旧金山某次活动上，微软CEO Satya Nadella介绍说，“微软喜爱Linux”，这令人惊讶不已！微软目前正在努力与Linux建立良好关系，其云计算平台Azure也在大规模使用Linux操作系统。

据国外媒体arstechnica报道，近日在旧金山微软举行的云产品发布会上，微软CEO Satya Nadella说，“微软喜爱Linux”！简直令人震惊！微软目前正在努力与Linux建立良好关系，其云计算平台Azure也在大规模使用Linux操作系统。Nadella说，Azure 20%的虚拟机使用了开源操作系统。

Software The Register 61

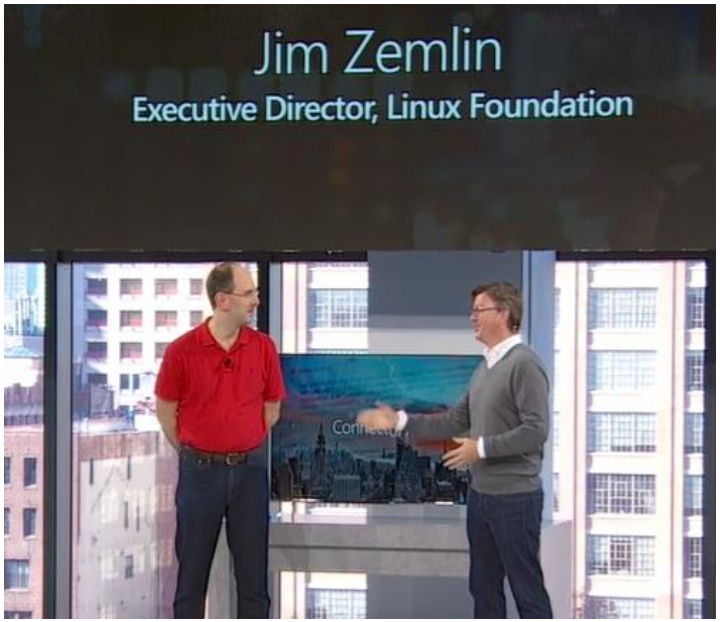
Microsoft just got its Linux Foundation platinum card, becomes top level member

More Linux love from Redmond – and a public preview of SQL Server for Linux

51CTO 系统频道 首页 资讯 Linux Windows 开源 桌面 运维&工具 技巧

支持开源！微软宣布加入Linux基金会

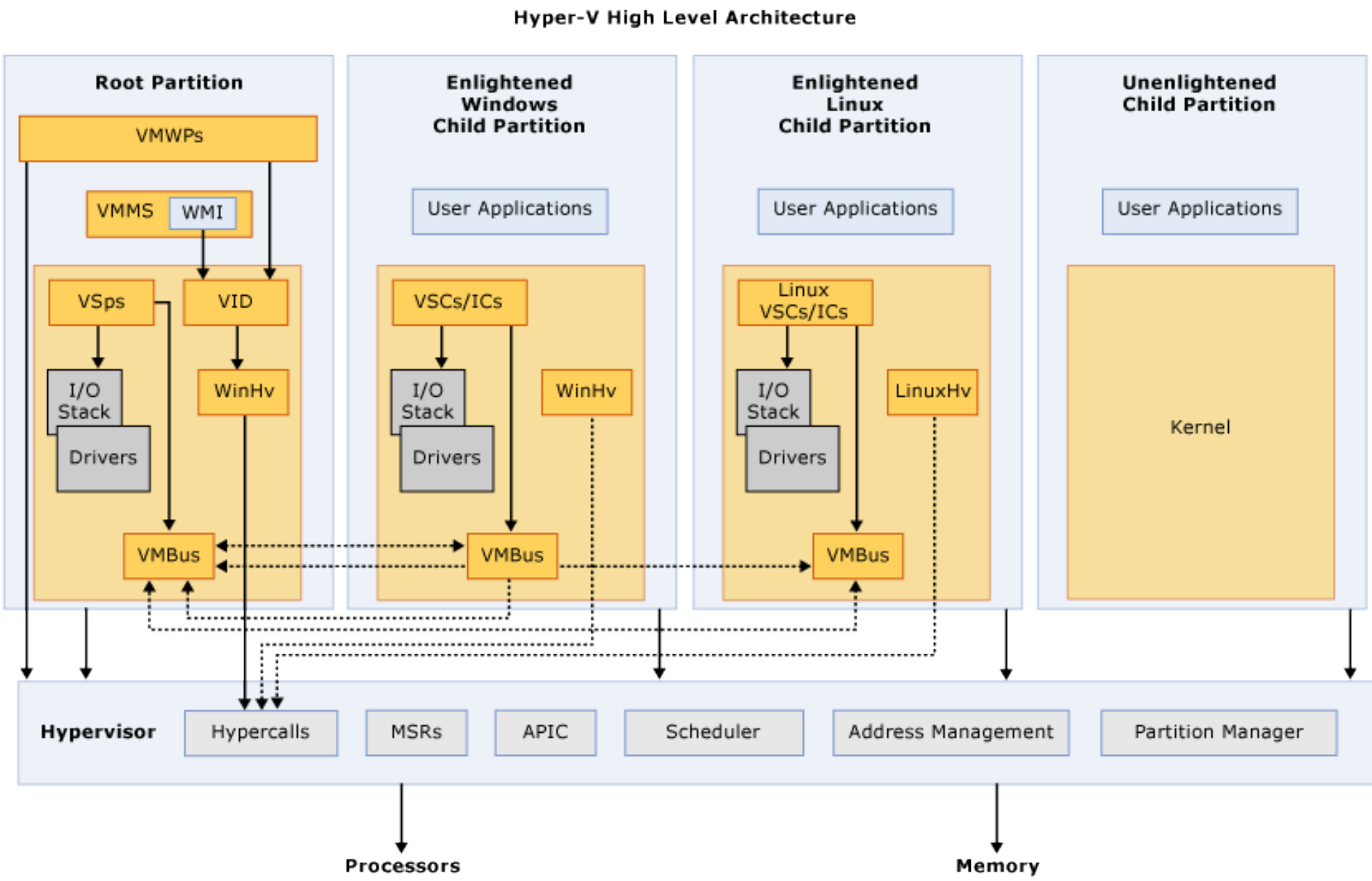
北京时间11月17日消息，微软本周在纽约举行的Connect2016开发者大会上宣布，将加入Linux基金会，并支付50万美元的年费成为该基金会最高级的白金会员。微软Azure团队的架构师约翰·格斯曼(John Gossman)将成为基金会的董事会成员。



Microsoft Azure – Open & Flexible

	Private Cloud	Hybrid Cloud	Public Cloud
Management	System Center OMS	CHEF, puppet labs, ANSIBLE, SALTSTACK, SCALR CLOUD MANAGEMENT, GitHub, juju, GRUNT, CLOUDFORMS by Red Hat Cloud	
DevOps & PaaS	Team Foundation Services	VAGRANT, Nagios, Gradle, Jelastic, Jenkins, apprenda, CLOUD FOUNDRY, libcloud, OPENSHIFT	
Applications	Dynamics SharePoint Exchange	Joomla!, Drupal, REVOLUTION ANALYTICS, Magento Open Source eCommerce	
Frameworks & Tools	.NET Visual Studio	python, Java, IntelliJIDEA, eclipse, nodeJS, R, APACHE CORDOVA, Xamarin, JBoss by Red Hat, php	
Data	SQL Server	hadoop, cloudera, MAPR, Hortonworks, redis, MySQL, cassandra, DATASTAX, splunk>, MariaDB, mongoDB	
Infrastructure	Windows Server	redhat, ubuntu, SUSE, ORACLE LINUX, debian, CentOS, bitnami, FreeBSD, CoreOS, DC/OS, docker	
	Microsoft Traditional Monetization	Microsoft + Open Source Cloud Monetization	

1/3 VMs are running Linux on Azure


Hyper-V High Level Architecture
Root Partition
VMWPs
VMMS
WMI
VSps
VID
I/O Stack
Drivers
WinHv
VMBus
Enlightened Windows Child Partition
User Applications
VSCs/ICs
I/O Stack
Drivers
WinHv
VMBus
Enlightened Linux Child Partition
User Applications
Linux VSCs/ICs
I/O Stack
Drivers
LinuxHv
VMBus
Unenlightened Child Partition
User Applications
Kernel
Hypervisor
Hypercalls
MSRs
APIC
Scheduler
Address Management
Partition Manager
Processors
Memory

PCIe Passthrough

Discrete Device Assignment (DDA)

- Discrete Device Assignment (also known as PCIe Passthrough) available as part of the Hyper-V role in Microsoft Windows Server 2016.
  - Other competitor uses different names like VMDirecPath (Vmware).

- Performance enhancement that allows a specific physical PCIe device to be directly controlled by a guest VM running on the Hyper-V instance.
  - GPU
  - Network adapter
  - NVMe device

Guest on Hyper-V vs. XEN & KVM

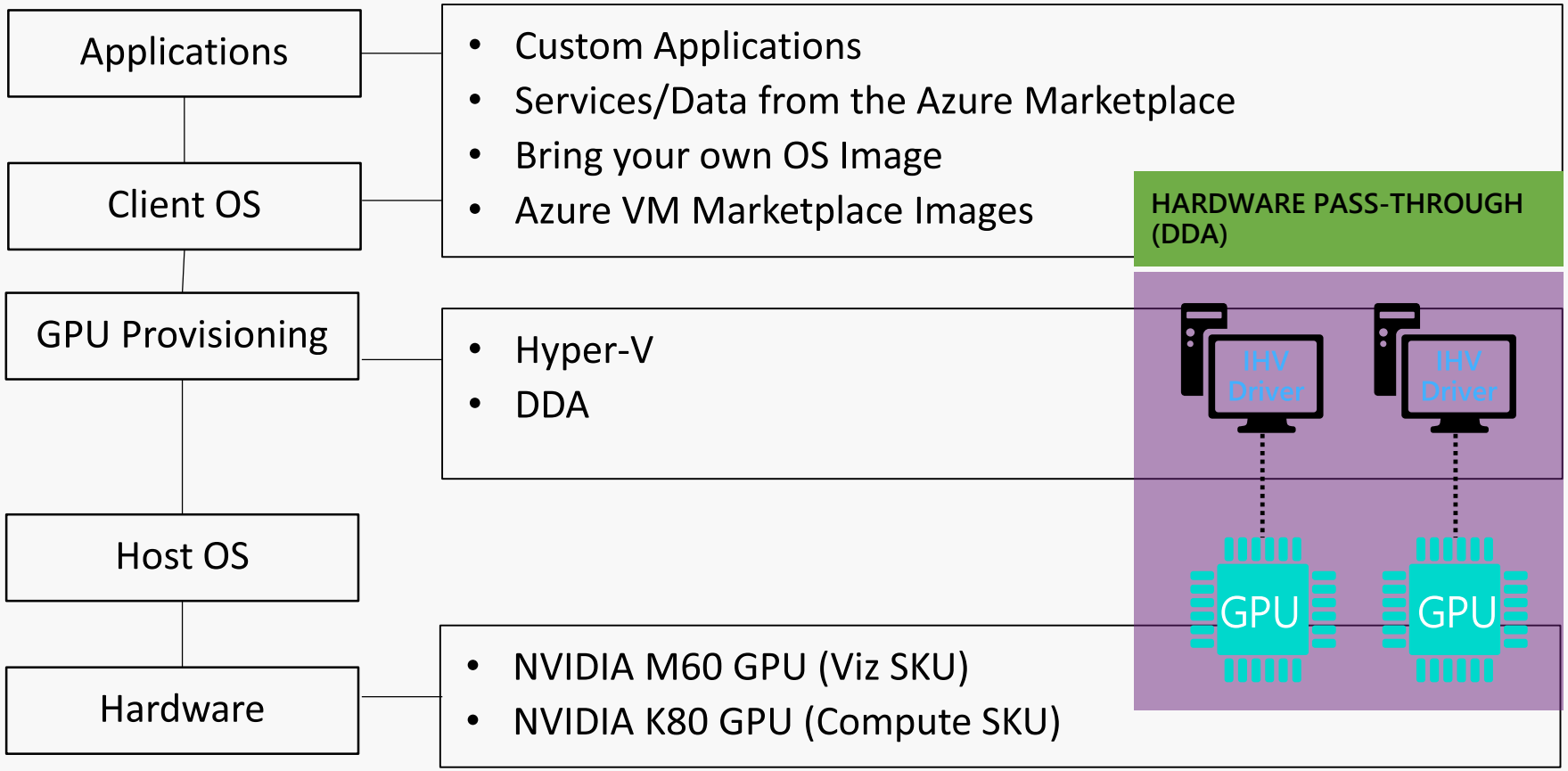
- Xen (HVM)/KVM
  - Full emulation of PCI/PCIe bus
- XEN(PV guest)/Hyper-V
  - Partial emulation of PCI/PCIe bus + PV hotplug message

The PCIe Root Bus Driver for Hyper-V PCI Express pass-through

- Not a full PCI Express bus emulation.
  - Simplify Hyper-V side implementation, thus less error prone.
- Minimum effort Linux driver for PCI Express root bus is required.
  - pci cfg space read/write: dev enumeration , dev control
  - Mapping between virtual/physical MSI data/addr

Accelerated GPU Experience in Azure

Architecture

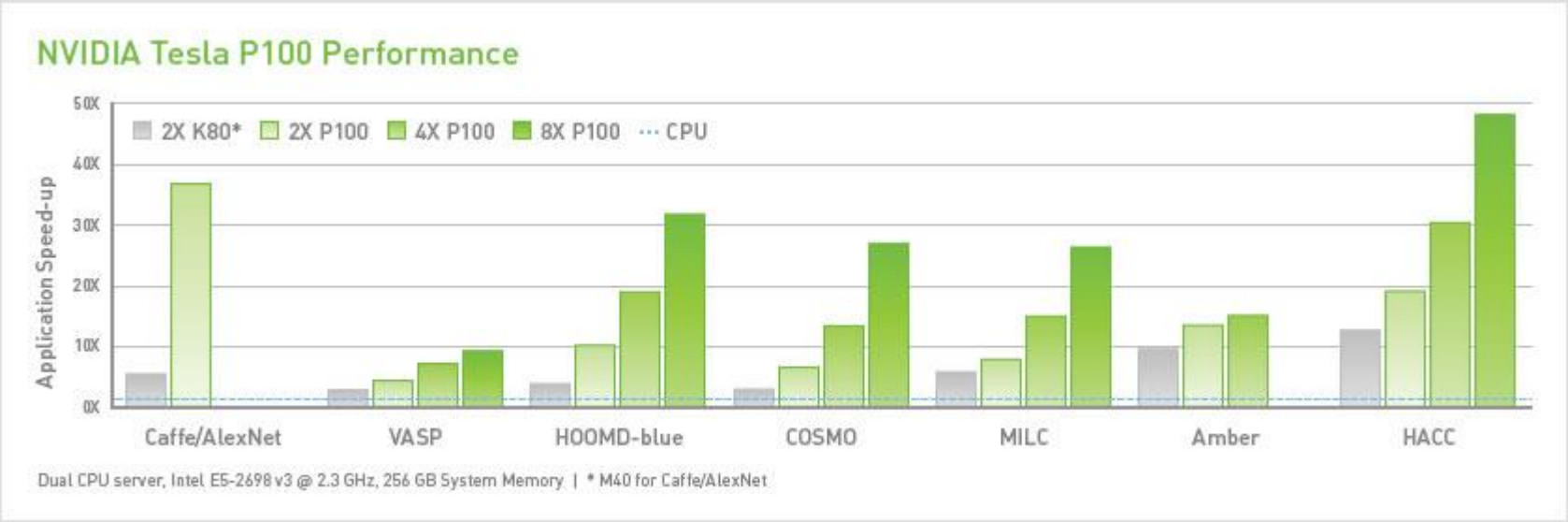


Visualization VMs - NV
Compute VMs - NC

	NV6	NV12	NV24
Cores(E5-2690v3)	6	12	24
GPU	1 x M60	2 x M60	4 x M60
Memory	56 GB	112 GB	224 GB
Disk	~380 GB SSD	~680 GB SSD	~1.5 TB SSD
Network	Azure Network	Azure Network	Azure Network

	NC6	NC12	NC24	NC24r
Cores(E5-2690v3)	6	12	24	24
GPU	1 x K80	2 x K80	4 x K80	4 x K80
Memory	56 GB	112 GB	224 GB	224 GB
Disk (SSD)	~380 GB	~680 GB	~1.5 TB	~1.5 TB
Network	Azure Network	Azure Network	Azure Network	Azure Network + RDMA (RoCE)

Announcing next generation of NC-series



	NC6s_v2	NC2s_v2	NC24s_v2	NC24rs_v2
Cores	6	12	24	24
GPU	1 x P100	2 x P100	4 x P100	4 x P100
Memory	112 GB	224 GB	448 GB	448 GB
Network	Azure Network	Azure Network	Azure Network	InfiniBand

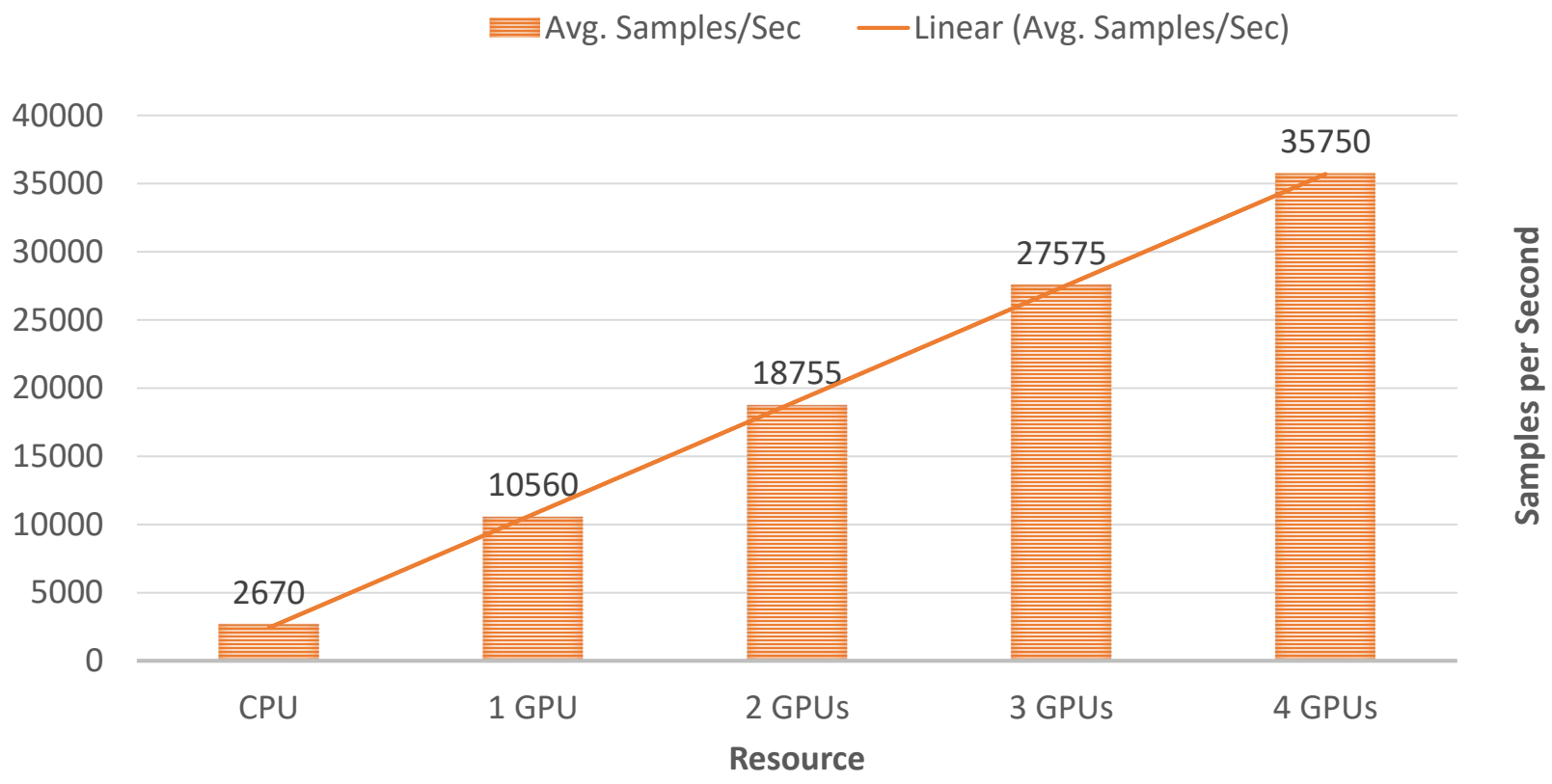
Announcing new ND-series

	ND6s	ND12s	ND24s	ND24rs
Cores	6	12	24	24
GPU	1 x P40	2 x P40	4 x P40	4 x P40
Memory	112 GB	224 GB	448 GB	448 GB
Network	Azure Network	Azure Network	Azure Network	InfiniBand

Above new sizes will be available later in the year

Open for signing up preview

CNTK Performance on DDA



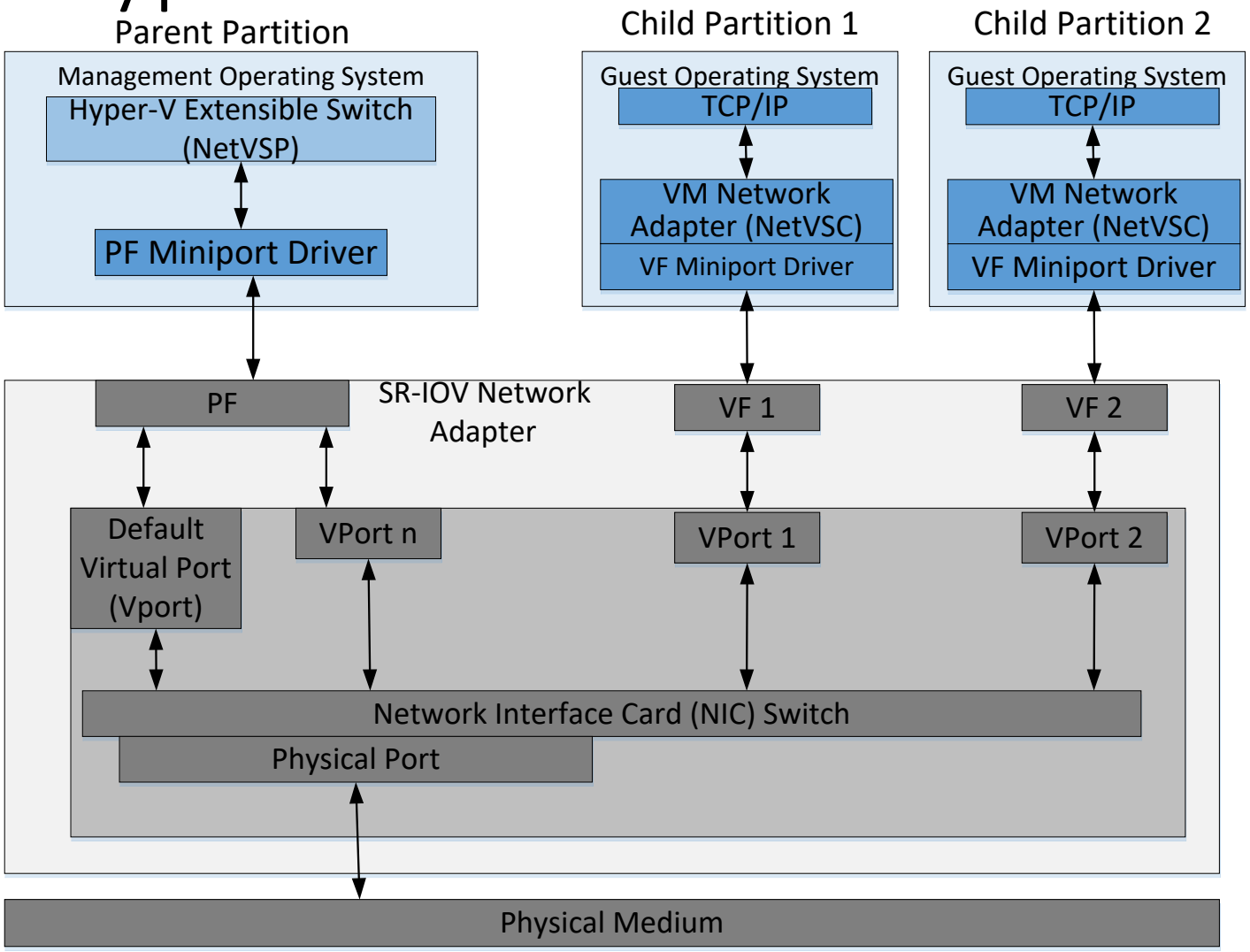
Demo

SR-IOV

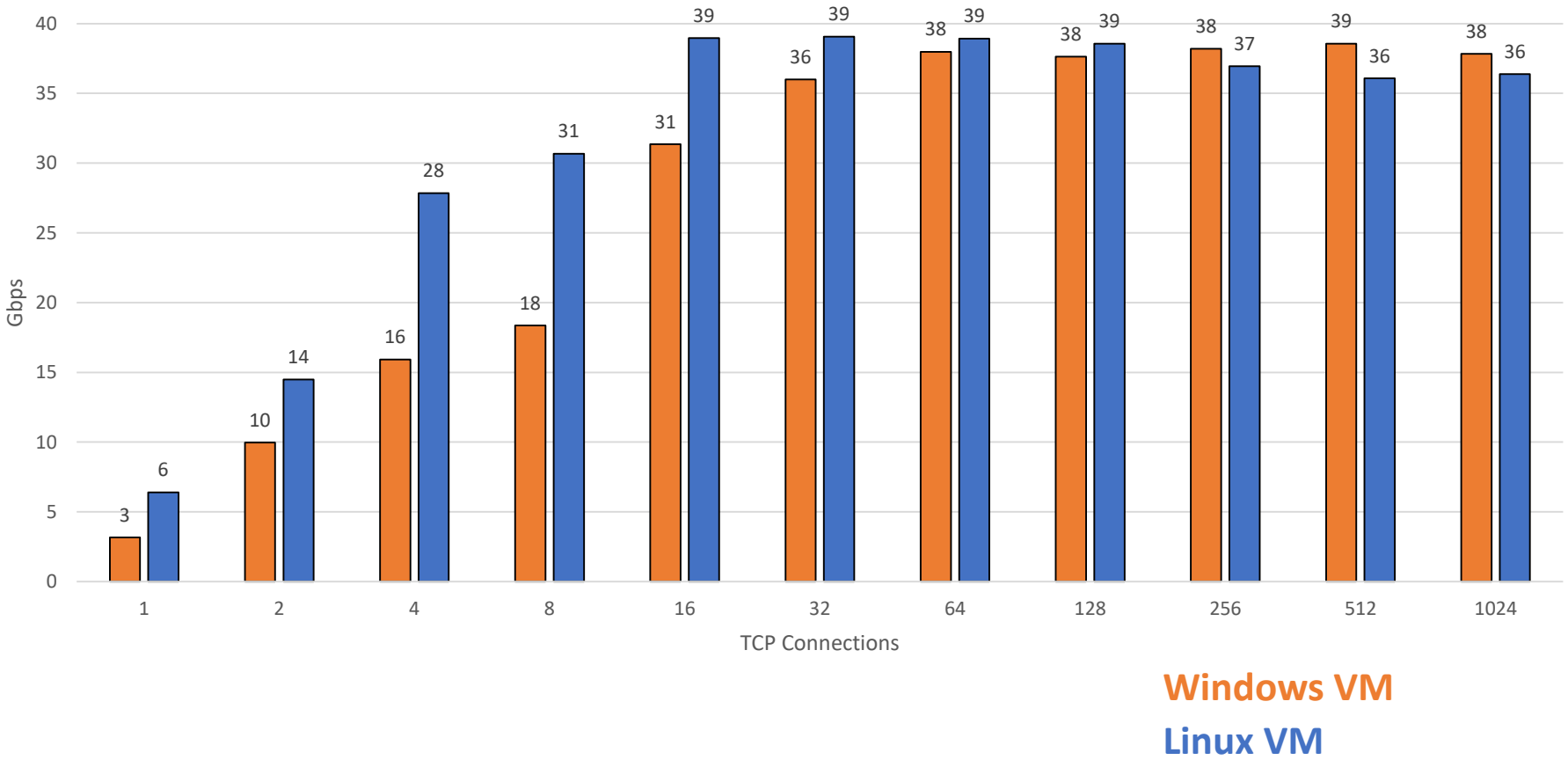
SR-IOV (Single-Root I/O Virtualization)

- HOW PCIe devices are constructed to support IO virtualization.
- Specification suggested by PCI-SIG
- Goal is to standardize on a way of bypassing the VMM's involvement in data movement by providing independent memory space, interrupts, and DMA streams for each virtual machine.
- Physical Functions(PFs) and Virtual Functions(VFs) are introduced. Multiple VFs can be mapping into one PF and one or more VF can be assigned to a VM

SR-IOV Networking Architecture on Hyper-V



SR-IOV Networking: 40Gbe on a Local Host



Accelerated Networking in Azure

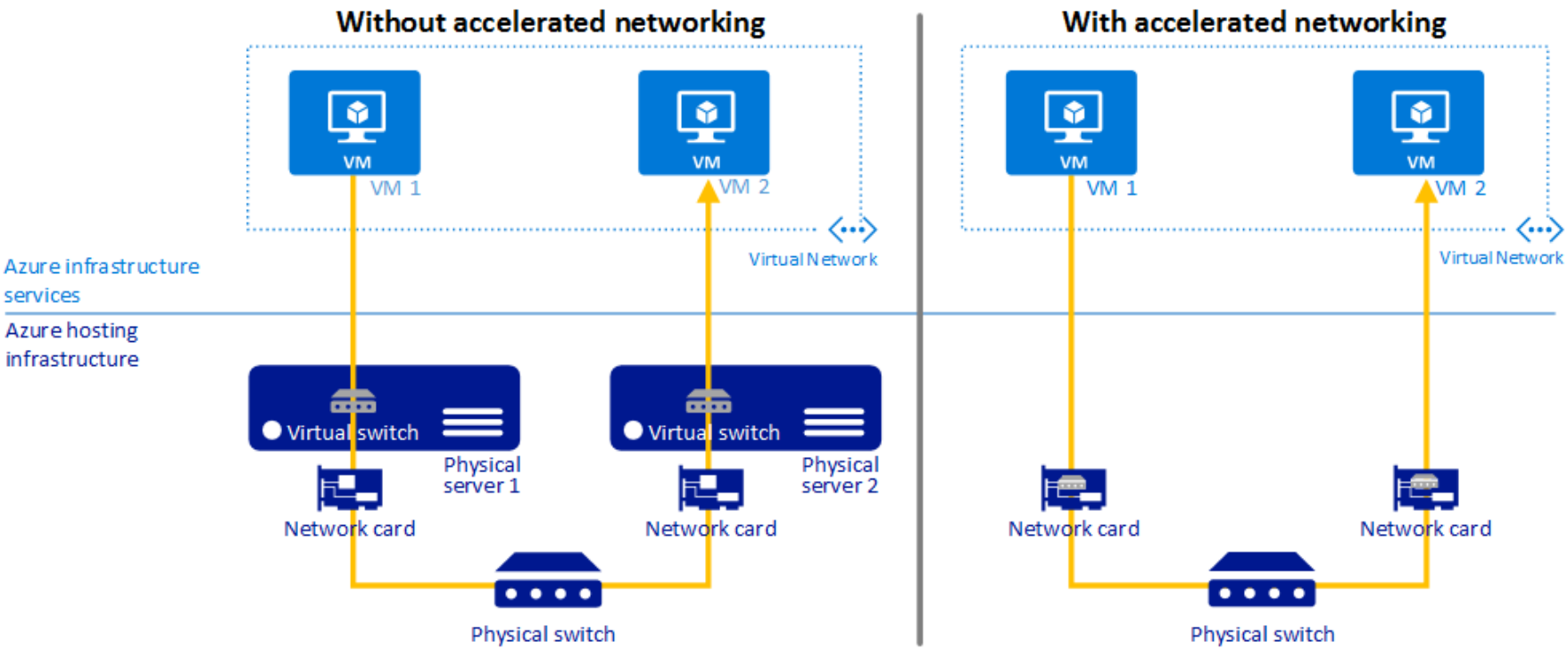
Accelerated Networking

- Highest bandwidth VMs of any cloud
  - DS15v2 & D15v2 VMs get 25Gbps
- Consistent low latency network performance
  - Provides SR-IOV to the VM
  - Up to 10x latency improvement
  - Increased packets per second (PPS)
  - Reduced jitter means more consistency in workloads
- Enables workloads requiring native performance to run in cloud VMs
  - >2x improvement for many DB and OLTP applications

Accelerated Networking – SR-IOV

SDN/Networking policy applied in software in the host

FPGA acceleration used to apply all policies



Announcing Accelerated Networking for Linux Preview

- Now offered on Dv2 Series VMs w/ 8+ cores
- Preview available in select regions
  - West US 2
  - South Central US
- Supported on multiple Linux distributions in the Azure Marketplace
  - Ubuntu
  - CentOS
  - RHEL

Get the highest bandwidth VMs of any cloud (25Gbps) with ultra-low latency on Linux!

Q & A