vmworld

VAP2340BU

POSSIBLE BEGINS WITH YOU

Driving Organizational Value by Virtualizing AI/ML/DL and HPC Workloads

Anthony Foster, Dell EMC
Gina Rosenthal, VMware, Inc.

#vmworld #VAP2340BU

vmware

Disclaimer

This presentation may contain product features or functionality that are currently under development.

This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

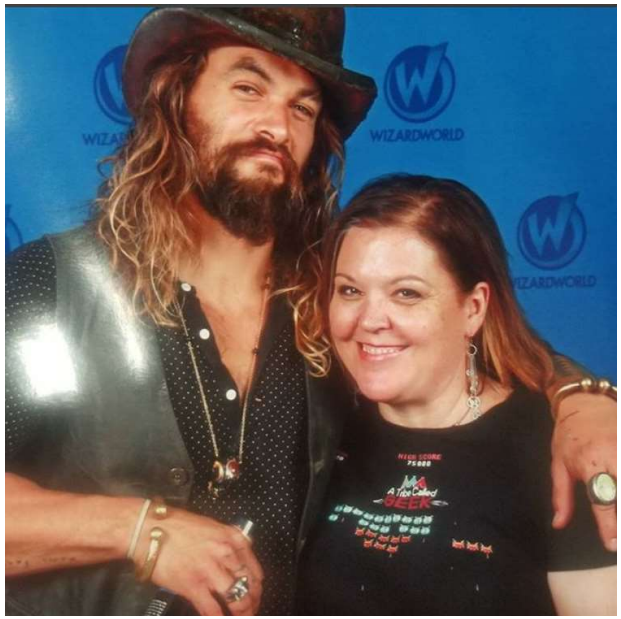
Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new features/functionality/ technology discussed or presented, have not been determined.

About Gina

- Virtualized since ~2008 (ESX 3.?)
- 7th VMworld (2011)
- Employer: VMware
- Role: Product Marketing
- Passions: Tying tech history to current trends, thinking of the impact of new tech on disadvantaged communities
- Three random things:
  - I can pickles and jellies
  - I have a HUGE rubber band ball
  - I still collect vinyl
- Find out more: I blog at https://24x7itconnection.com and podcast at http://wideworldoftech.com
- Twitter: @gminks

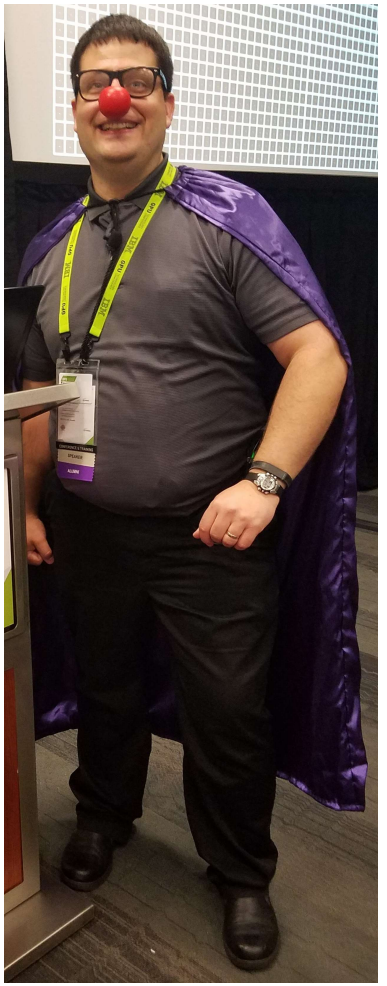


About Tony

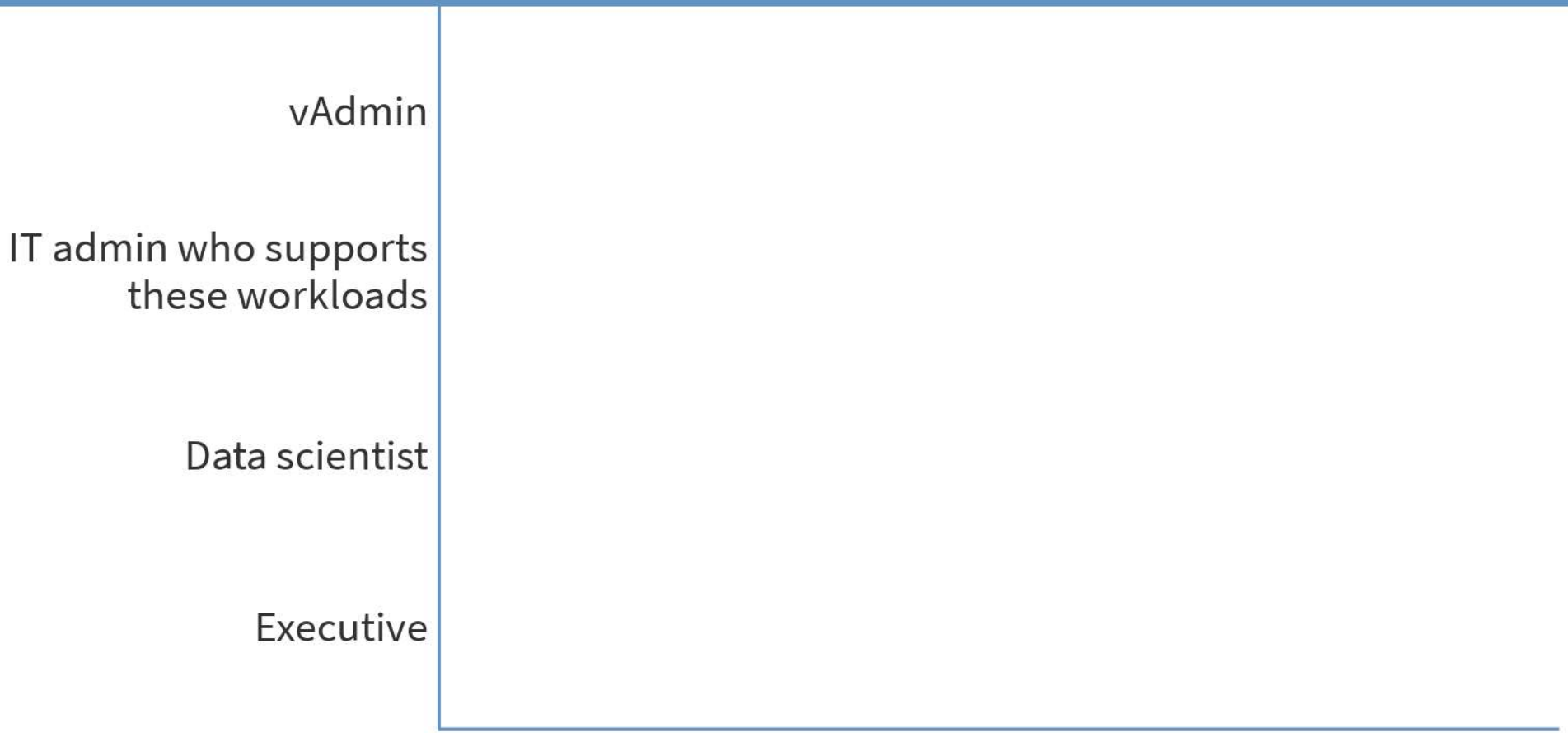
When you live in Kansas…

Where Hollywood thinks I live

- Aliases: ‘WonderNerd’ and “Hey You’
- Virtualized since 2005 (ESX 2.0)
- 10th VMworld (2008)
- Employer: Dell EMC
- Role: Technical Marketing
- Passions:
  Crazy next generation ideas not found on road maps; VDI/EUC; GPUs
- Three random things:
  - Trained storm spotter
  - Part owner of a rodeo
  - Builder of high power rockets
- Find out more: www.wondernerd.net
- Twitter: @wonder_nerd



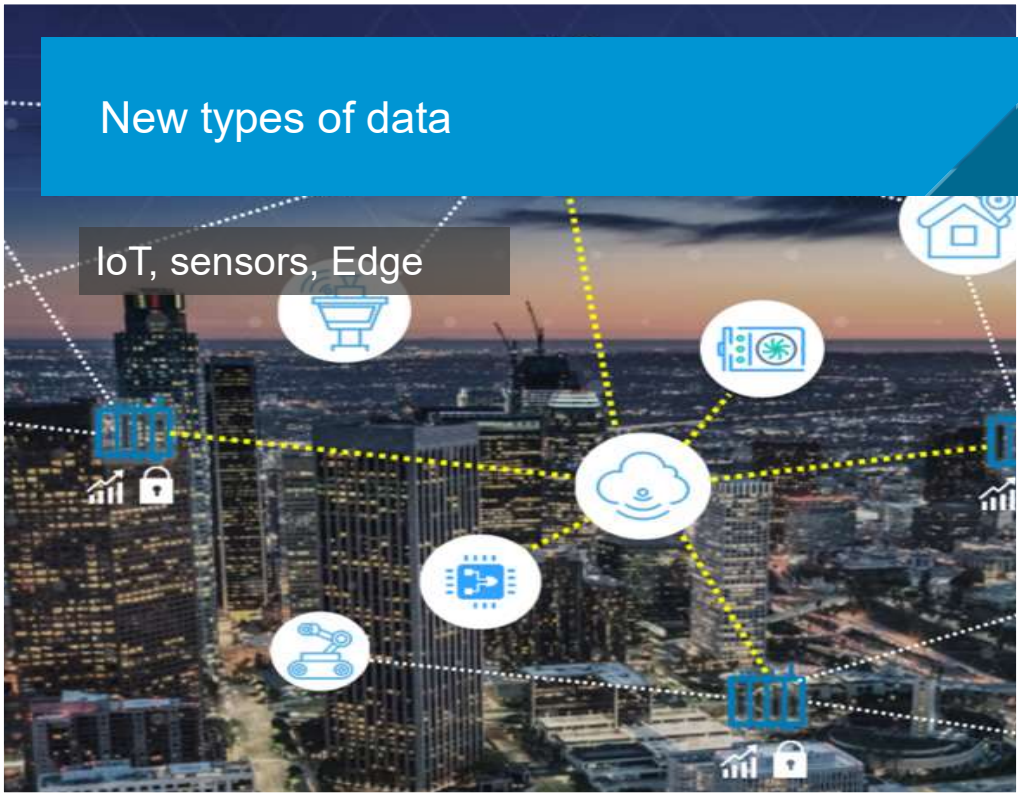
Who are you?



Artificial Intelligence is the New Buzzword That's Existed since the 1950s

Data Continues to Grow at an Incredible Pace

Patterns in the data can help us do amazing things



AI is Powered by HPC and Big Data Architectures

It is fueled with big data

AI
The capability of a machine to imitate intelligent human behavior

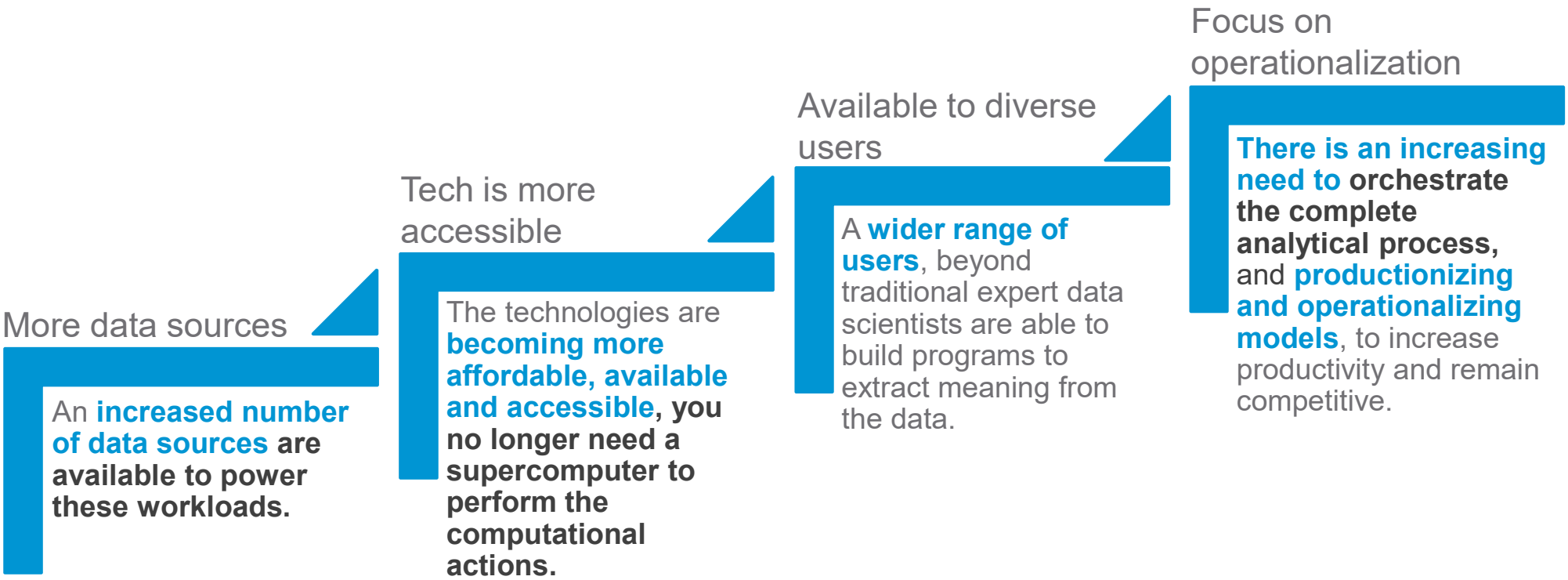
HPC
Aggregates computing power to deliver much higher performance to solve large computational problems

Machine learning
Process of "training" a machine, feeding large amounts of data into algorithms that give it the ability to learn how to perform a task.

Deep learning
A machine learning technique that uses neural networks as the underlying architecture for training models.

Big Data
Large sets of data, used to gain insights and/or generate value.

Why the Sudden Focus on these Workloads?



AI & IoT Compute Use Cases in Multiple Verticals







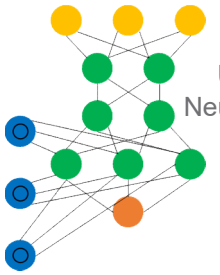
VERTICAL	Facilities	Energy	Health & Life Sciences	Manufacturing & Industrial	Transportation	Retail	Smart Cities
LOCATIONS	Office Airports Education	Generation Distribution Rigs	Hospitals Ambulances Clinics or labs	Oil rigs Mines	Taxis Air Rail Marine	Distribution Hotels Gas stations	Roads Towns Highways
DEVICES	HVAC Lighting Fire alarms Security Access	Turbines Generators Fuel cells Windmills	Implants Pumps Monitors Telemedicine	Motors Pipelines Assembly Tanks	Tools Sensors	Terminals Tags Vending	Traffics lights Road sensors Alarms

How Machines Learn?



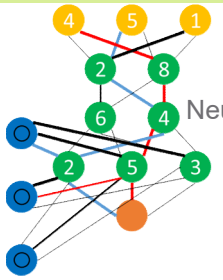
Training Set | Validation Set | New Data

Tagged Data



Untrained Neural Network

Maui	Momoa
X	✓



Trained Neural Network

Maui	Momoa
93%	95%

Trained Model

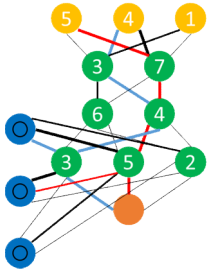
Acceptable

No

Yes

Training

(Learning)



Maui	Momoa
95%	98%

Trained Model

Inferencing

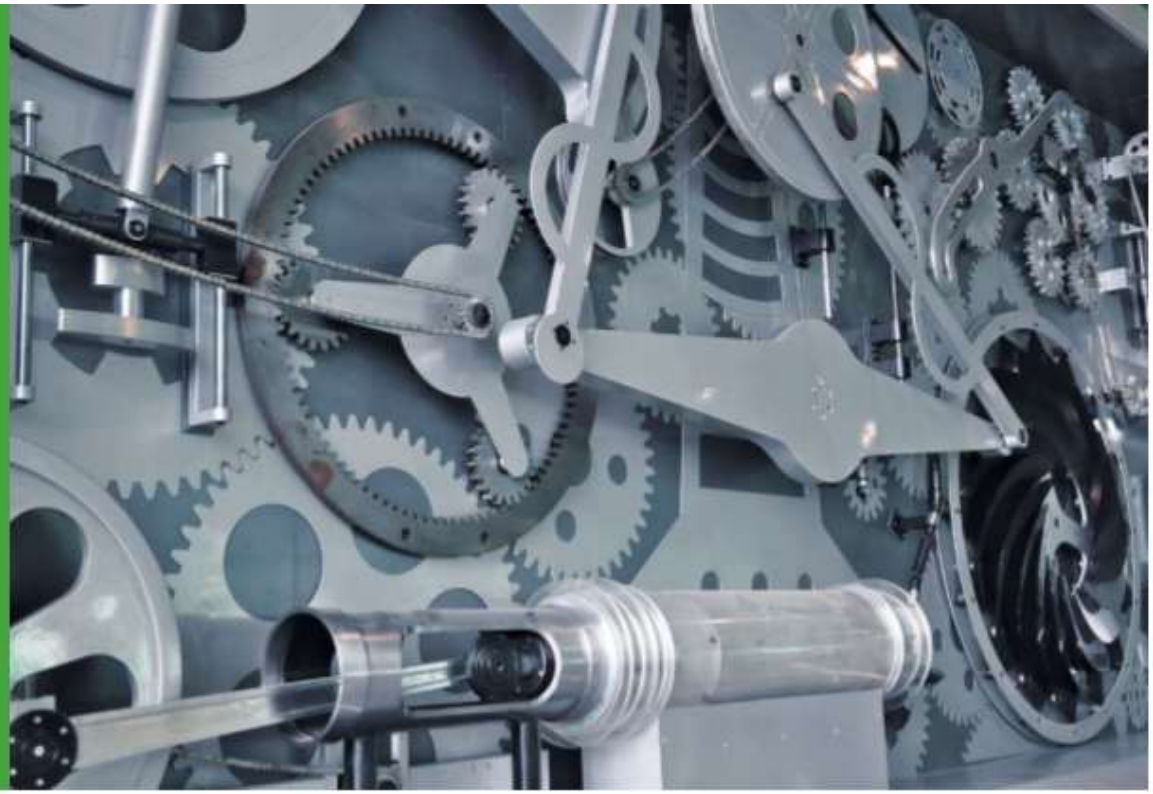
(Applying what was learned)

Traditionally, Teaching Machines Took Lots of Metal

Compute Acceleration Hardware Made Deep Learning Possible

HPC Drives Innovation, but Requires Massive Compute Resources

By 2022, HPC-driven simulations and deep learning will be the core innovation engines driving 10,000x increase in compute requirements



Gartner

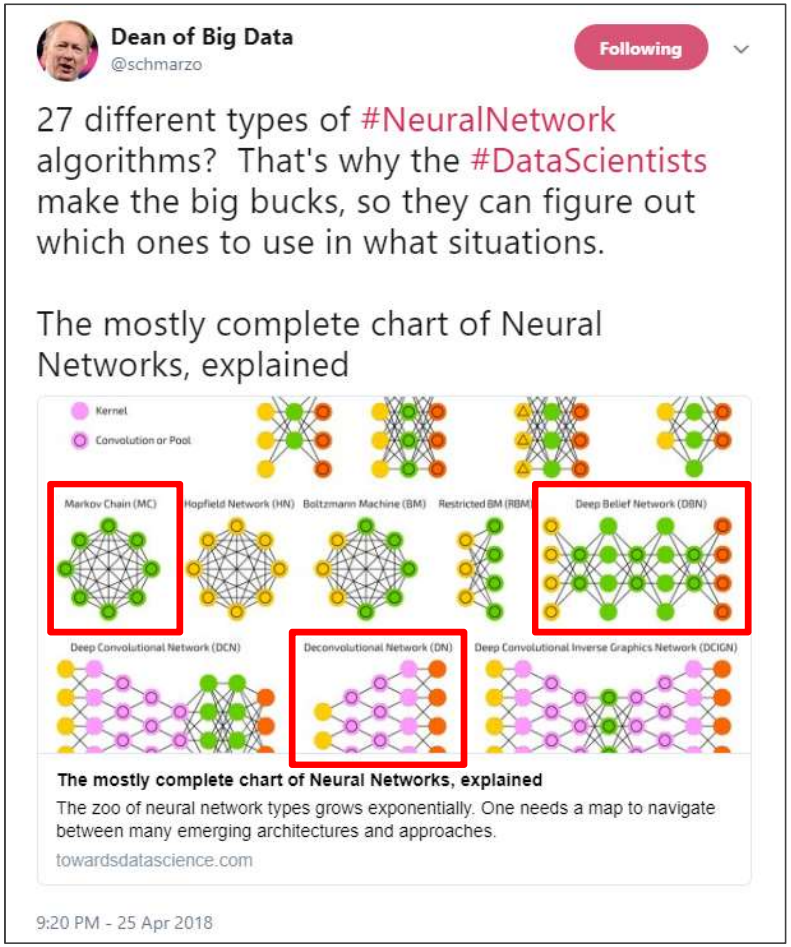
1 © 2018 Gartner, Inc. and/or its affiliates. All rights reserved.

Data Scientists Determine Which Type of Neural Network to Use

The neural networks rely on agile compute, storage, network, and software environments

Neural networks are resource intensive *workloads*.

Workloads require *architectures*.



Where do your AI, ML, DL, or HPC workloads run?

On premises in my data center on bare metal systems

On premises with powerful workstations

On premises with virtualization (private cloud)

In the public cloud

All of the above

We don't have any of these workloads yet

Next Evolution in AI: Virtualize the Infrastructure

vSphere can help data scientists get to answers faster

Operational Flexibility

- Simple cluster expansion and contraction
- Rapidly reproduce research environments
- Higher resiliency and less downtime with vMotion
- Fault-isolation (hardware and software)

Reduced Complexity

- Cluster resource-sharing
- Minimize setup and configuration time with centralized management capabilities
- Simultaneously support mixed software environments
- Industry-leading virtualization platform that your IT already knows

Secure Sensitive Workloads

- Easy, secure data access and sharing
- Security Isolation
- Multi-tenant data security

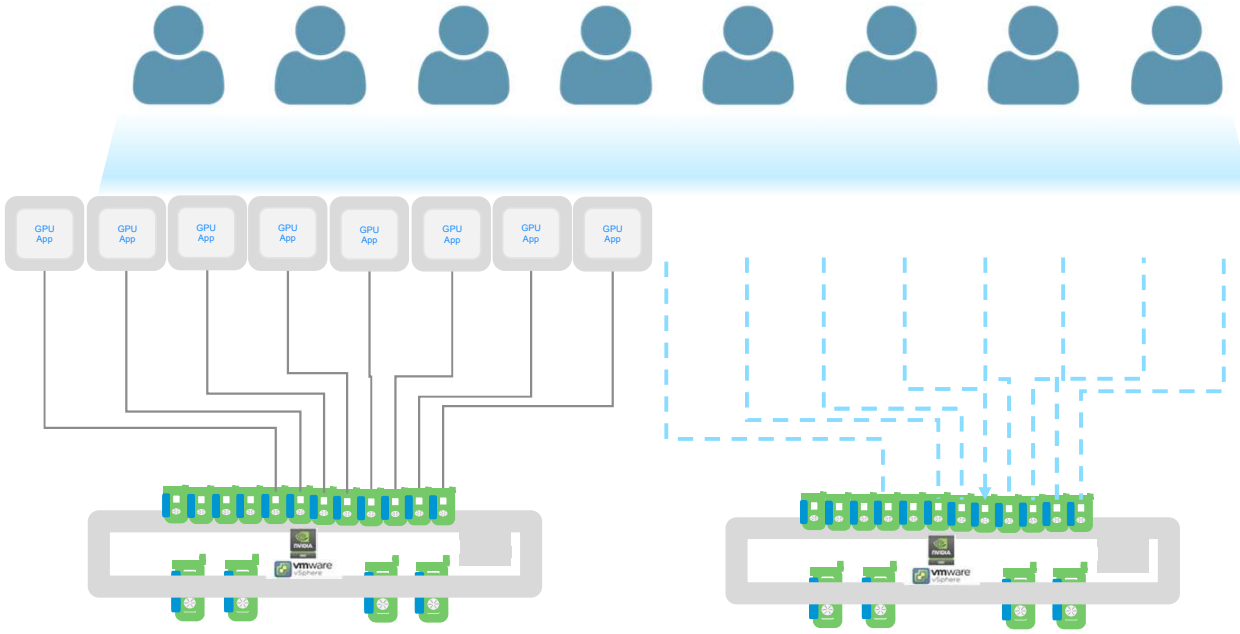
VMware vSphere 6.7 Update 1 with NVIDIA Quadro vDWS

Share single GPU among multiple VMs

- Provision partial or up to 1 full GPU

GRID vGPU VM vMotion support

- Rapidly repurpose GPU infrastructure
  - VDI/Data Science by Day
  - Compute (ML) by Night



Seamless to end-users and applications

- Host maintenance, patching or upgrade
- Rebalance Desktop pools without user flow disruption
- Provision different vGPU profile
- Reuse GPUs in DirectPath IO

Application-level RDMA (Remote Direct Memory Access) with vSphere

Passthrough mode (InfiniBand or RoCE)

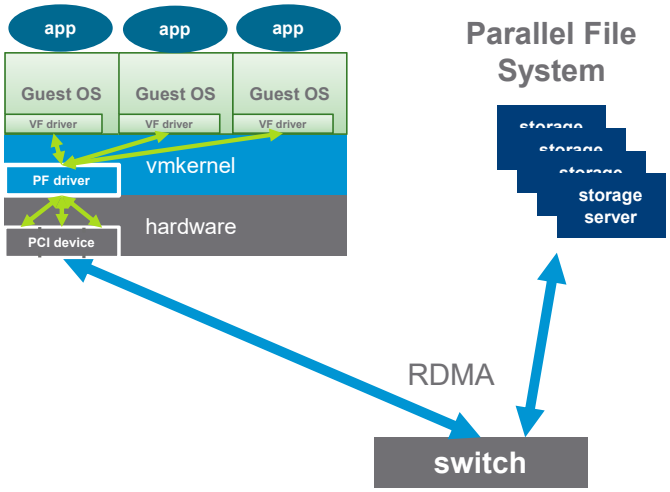
- The most common case when running MPI applications, since it matches bare-metal practices
- Only the standard guest driver is required – no ESXi driver

SR-IOV mode (InfiniBand or RoCE)

- Commonly used in throughput environments when shared access to parallel file systems (like Lustre or IBM Spectrum Scale) is required
- SR-IOV requires both an ESXi driver and standard guest driver
- InfiniBand SR-IOV requires Mellanox Connect-X4 or later

pvRDMA (Para-virtualized RoCE)

- When RDMA is required without compromising VM mobility (vMotion) and snapshots
- ESXi level support (included in ESX 6.5) plus standard guest driver
- Currently, all endpoints must be virtual



Para-Virtualized Remote Direct Memory Access (PVRDMA)

VM

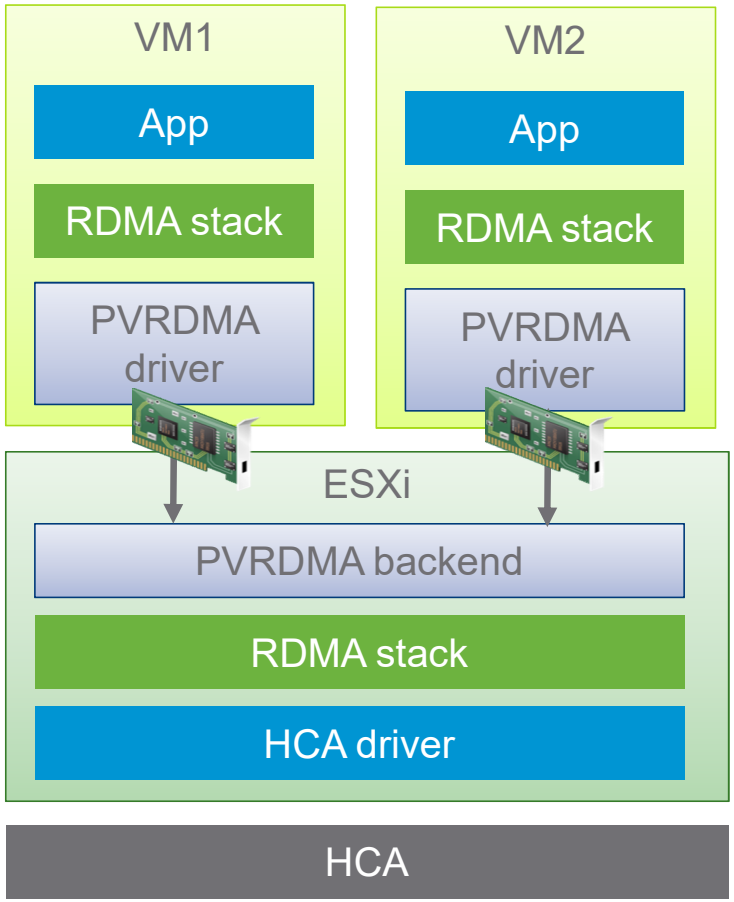
- Expose a virtual PCIe device
- Device Driver
  - Full support for the RDMA Verbs API
- User Library
  - Provides direct HW access for data path
- RDMA API calls proxied to PVRDMA backend

PVRDMA backend

- Creates virtual RDMA resources for VM
- Guests operate on these virtual resources

ESXi

- Leverage native RDMA stack
- Create corresponding resources in HCA
- Physical HCA services all VMs



1drnrd.me/PVRDMA

Using vSphere to Virtualize AI infrastructures

Which edition is right for me?

	STD	ENT+	vSOM ENT+	Scale-Out
Hypervisor	Yes	Yes	Yes	Yes
vMotion	Yes	Yes	Yes	Yes
High Availability	Yes	Yes	Yes	
Data Protection & Replication	Yes	Yes	Yes	
vShield Endpoint	Yes	Yes	Yes	
Fault Tolerance	Yes	Yes	Yes	
Storage vMotion	Yes	Yes	Yes	Yes
DRS and DPM		Yes	Yes	
Storage APIs		Yes	Yes	Yes
Reliable Memory		Yes	Yes	
Distributed Switch		Yes	Yes	Yes
Storage DRS Profile-Driven Storage		Yes	Yes	
I/O Controls & SR-IOV		Yes	Yes	Yes
Host Profiles & Auto Deploy		Yes	Yes	Yes
Encryption		Yes	Yes	
Ops Management			Yes	
Includes	**1 vS CPU lic.**	**1 vS CPU lic.**	**1 vS CPU lic.**	**8 vS CPU lic.**

Use vSphere Scale-Out if:

- You do not need DRS or HA from the hypervisor layer
- You are looking to reduce cost

Use vSphere Enterprise Plus if:

- Your HPC/Big Data/ML/DL applications don't provide DRS and HA

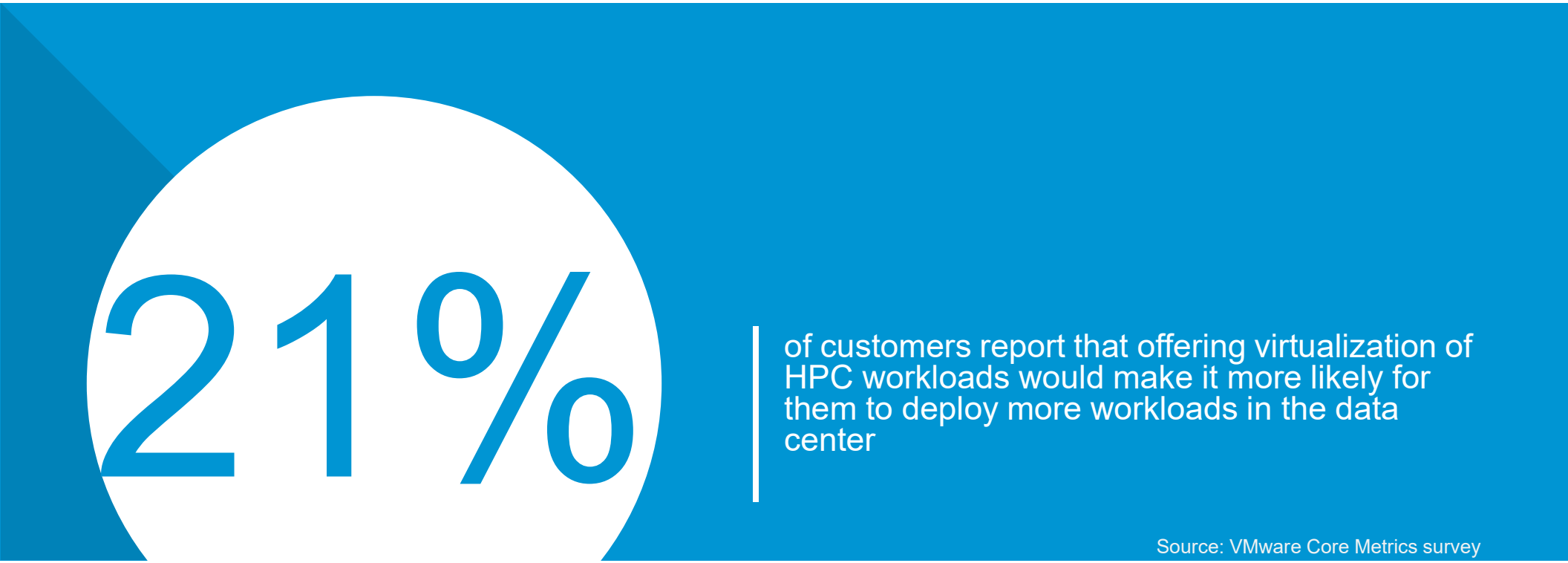
BUZZWORD

B	I	N	G	O
‘Hey Alexa’	Tensor Flow	Unstructured Data	Market Transformation	Predictive Analytics
Scary	Data Warehouse	Neural Network	Logistic Regression	AlphaGo
Data Science	IoT / Edge Sensors	AI Does Everything	Voice Assistant	Chat Bot
Smart City	Shared Resources	‘Ok Google’	Mass Surveillance	Intellectual Property
Turing Machine	GPU	Replaced by AI	Modeling	Data Mining

Council on Competitiveness

Solve report

Two-thirds of all U.S.-based companies that use HPC say that “increasing performance of computational models is a matter of competitive survival.”



Our Customers are already considering virtualizing HPC

What are your biggest challenges to adopting these workloads?

I need to provide IaaS/PaaS for all of my data scientists. **A**

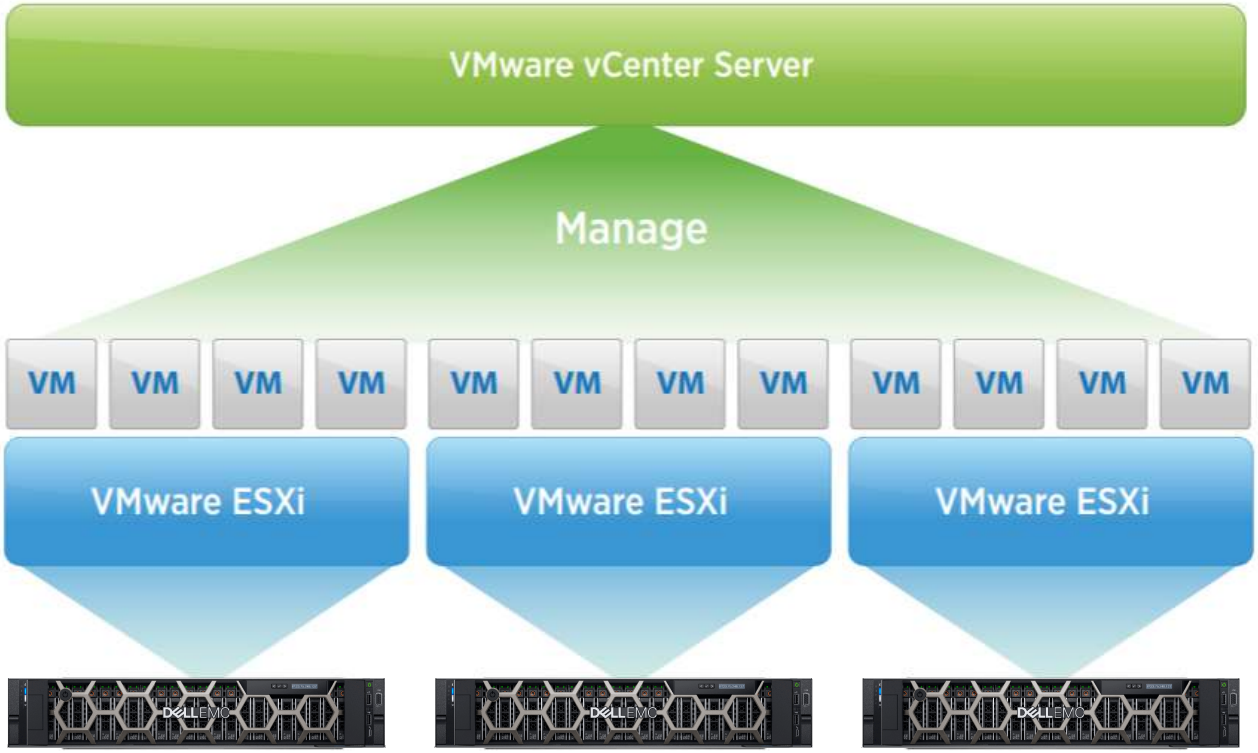
I would like to integrate disparate hardware (include acceleration cards). **B**

It needs to be easier to provision. **C**

I need this to be operationalized and integrated with my existing IT infrastructure. **D**

Architectures

vSphere Benefits Can Be Applied to AI Architectures



vSphere offers:

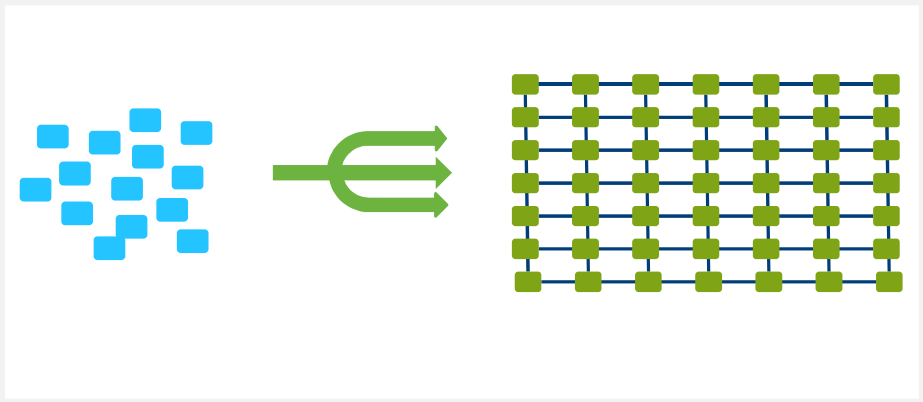
- Heterogeneity
- Multi-tenant data security
- Fault isolation
- Reproducibility
- Fault resiliency
- Performance

Virtualizing the Workloads

Two main architectures to consider

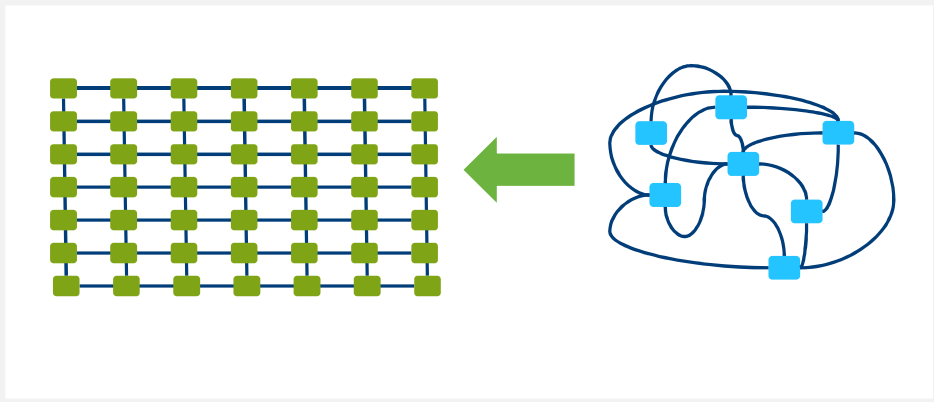
MPI

- Many processes that run in parallel, but also require intense intercommunication
- Performance depends on the message sizes being passed within the application.

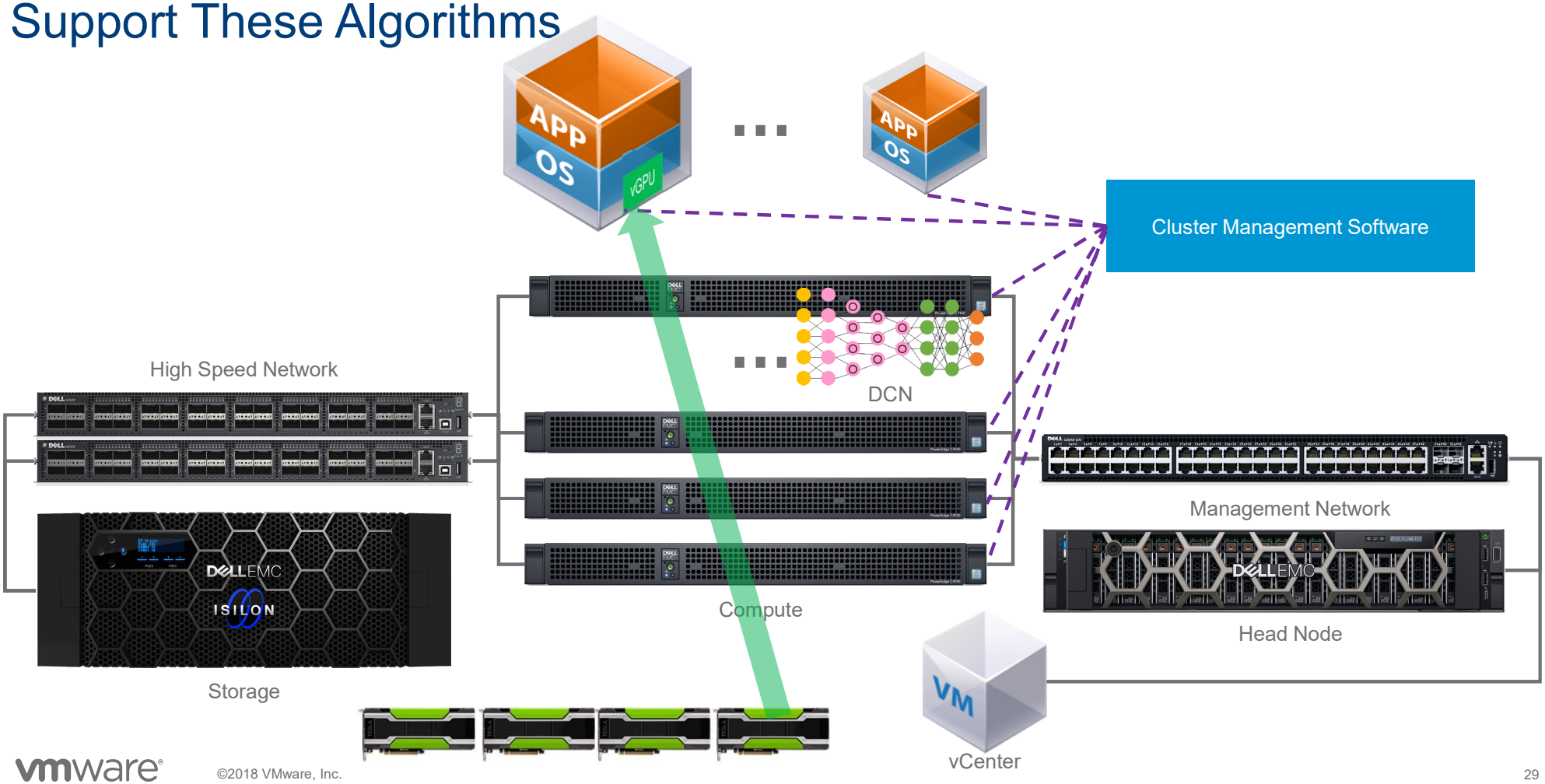


Throughput

- Many independent jobs that run in parallel.
- Minimal performance impact and near bare-metal performance



IT Architects Design the Best Architecture to Support These Algorithms



Deep Learning Virtualization Use Case: Cycle Harvesting

Challenge:

Data Scientists submit jobs in traditional batches, because of compute availability

- Submit jobs one day
- Wait until the next day for the job results

What if…

The VDI environment has unused cycles. Could HPC jobs be run in the environment when it is not needed to run VDI?

Will it blend?

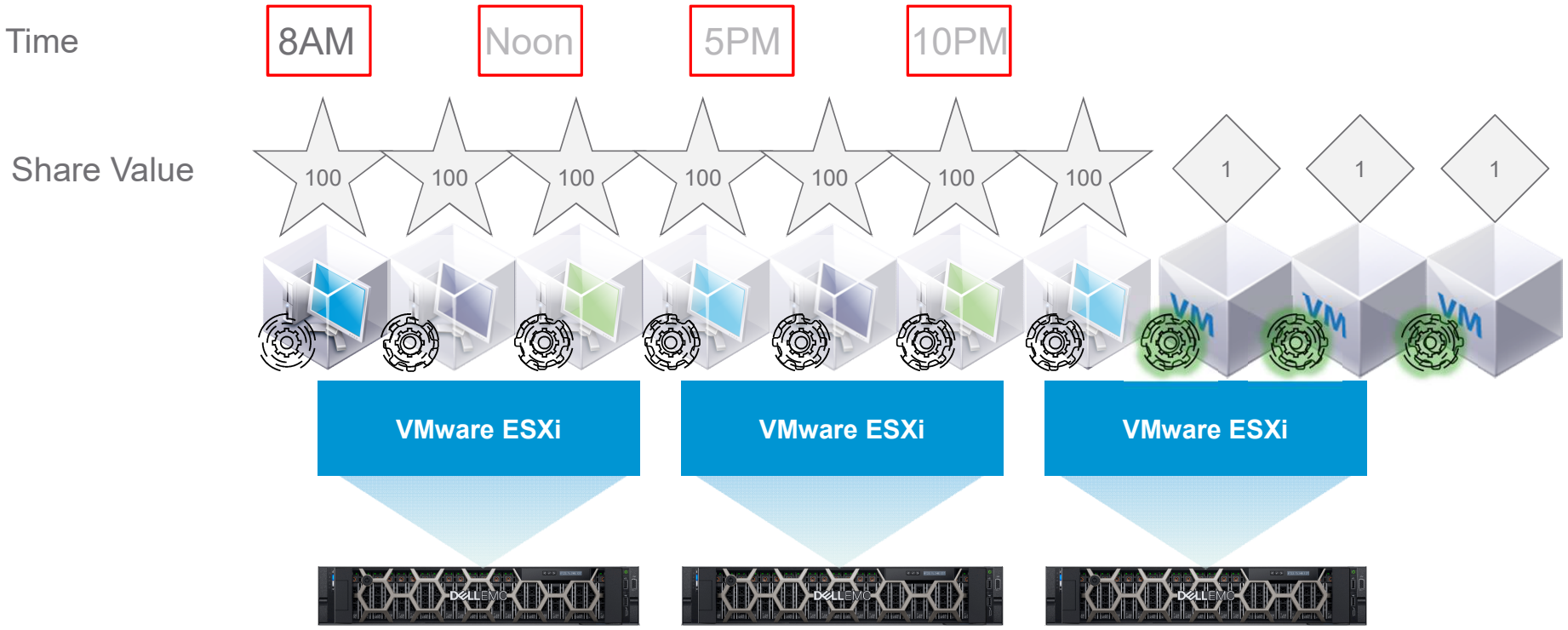
Outcome

Enable HPC compute jobs to harvest cycles from a VDI compute environment.

Benefit

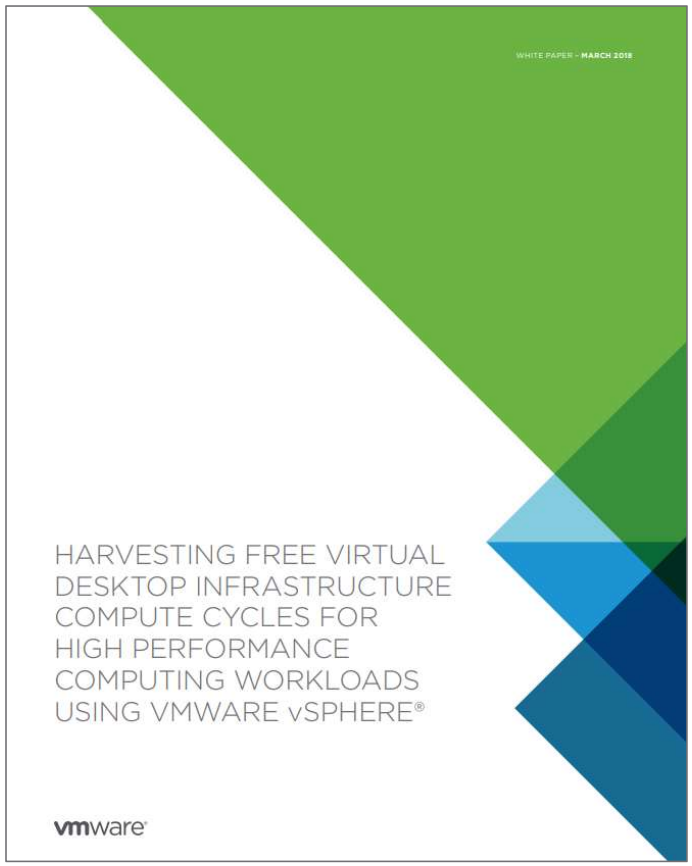
Go beyond a traditional batch-processing to viewing HPC resources as an engine for returning results in real time.

Cycle Harvesting



Cycle Harvesting Case Study

https://bit.ly/2MrBngH



Common Concerns about Virtualizing These Workloads

Performance Concerns

Adding a virtualization layer will impact performance

Reality: In most cases, performance is on-par or better than bare metal

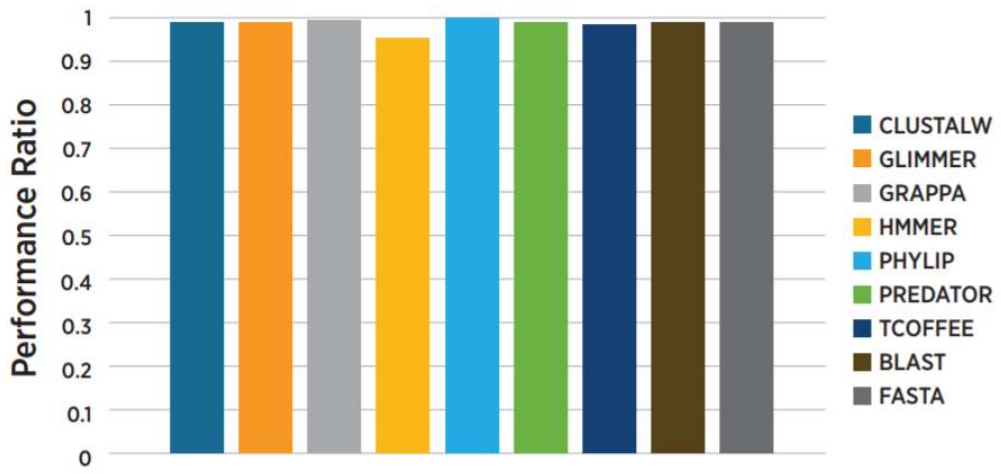


Figure 1. Performance Comparison of Typical HPC Throughput Applications – Virtualized Against Unvirtualized

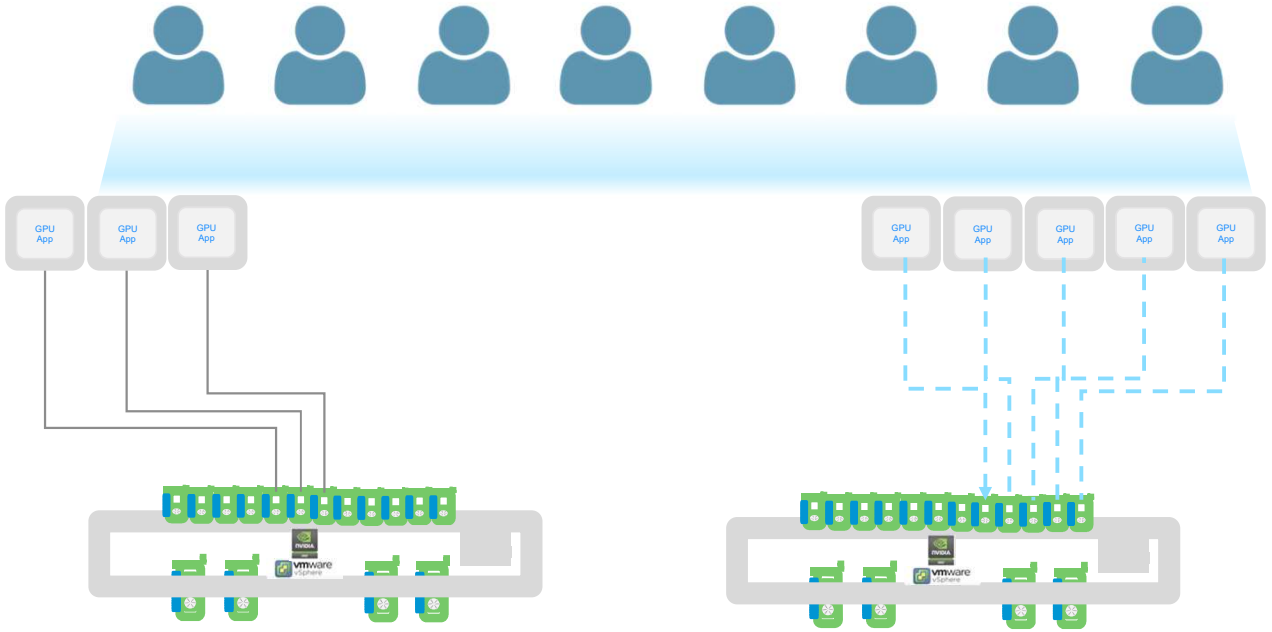
NOTE: Higher is better.

Common Concerns about Virtualizing These Workloads

Feature limitations

Reality: Almost any physical part of a cluster can be virtualized, including GPUs and RDMA.

I won't be able to use features, such as acceleration hardware

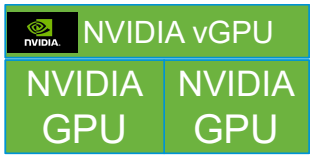


Common Concerns about Virtualizing These Workloads

Costs

Virtualization will increase my costs

Reality: Virtualization helps you fully utilize your hardware investments.



vSphere Scale-Out


vmworld
POSSIBLE BEGINS WITH YOU
PLEASE FILL OUT YOUR SURVEY.
Take a survey and enter a drawing for a VMware company store gift card.
#vmworld
#VAP2340BU
vmware


Questions?

Next Steps

Learn More This Week! Session Schedule

All Week

SPL-1947-01-EMT_U *
- **Hands on Lab**

Machine Learning Workloads in vSphere using GPUs – Getting Started

Wednesday

CTO2390BU
- 10:00 – 11:00; **Virtualize and Accelerate HPC/Big Data with SR-IOV, vGPU and RDMA**

SAI3243BU *
- 11:30 – 12:30; **Use Artificial Intelligence and Machine Learning to Simplify Security**

Thursday

VIN2067BU *
- 12:00 – 1:00; Accelerating &Optimizing Machine Learning on vSphere leveraging NVIDIA GPU

CTO1917BU *
- 1:30 – 2:30; Emerging Technologies in the Real World

* ML Virtual Track

Learn More!

vSphere for HPC

- https://www.vmware.com/solutions/high-performance-computing.html

vSphere for Big Data

- https://www.vmware.com/solutions/big-data.html

Dell EMC AI

- http://dellemc.com/ai

Deep Learning Institute

NEW!

Learn how to build Artificial Intelligence (AI) and accelerated computing applications with hands-on training

Fundamentals

Industry/domain-specific courses

Workshops and labs

Industry collaborations

Videos and podcasts

Now from Dell EMC

Access the World-class HPC and AI Innovation Lab

Dedicated to research and development of HPC solutions so you can make discoveries faster

Leverage HPC engineering expertise

Test new technologies

Tune your applications for performance and efficiency


vmworld
POSSIBLE
BEGINS
WITH YOU
THANK YOU!
#vmworld
#VAP2340BU
vmware