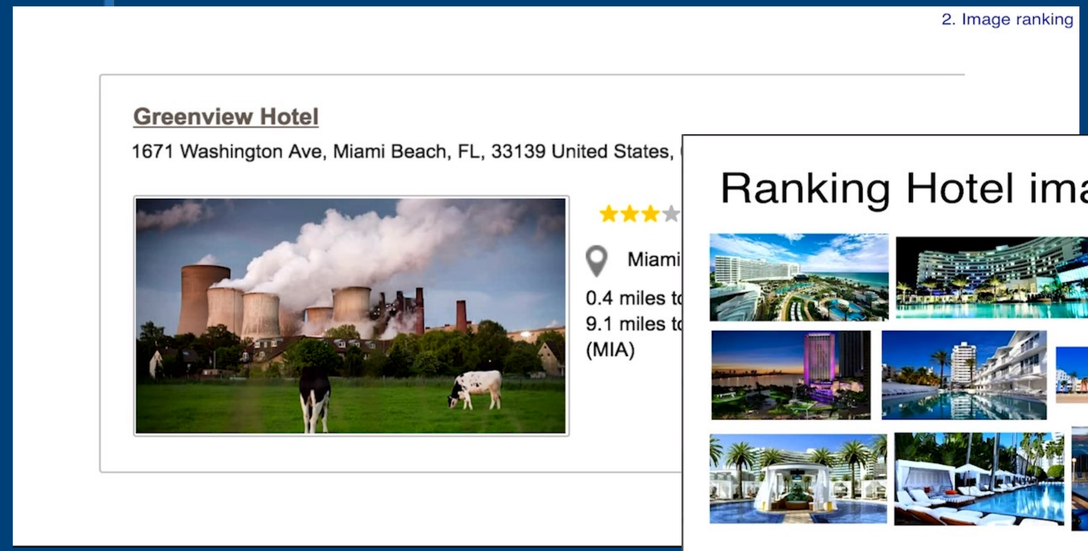
Deep Dive on EC2 GPU instances

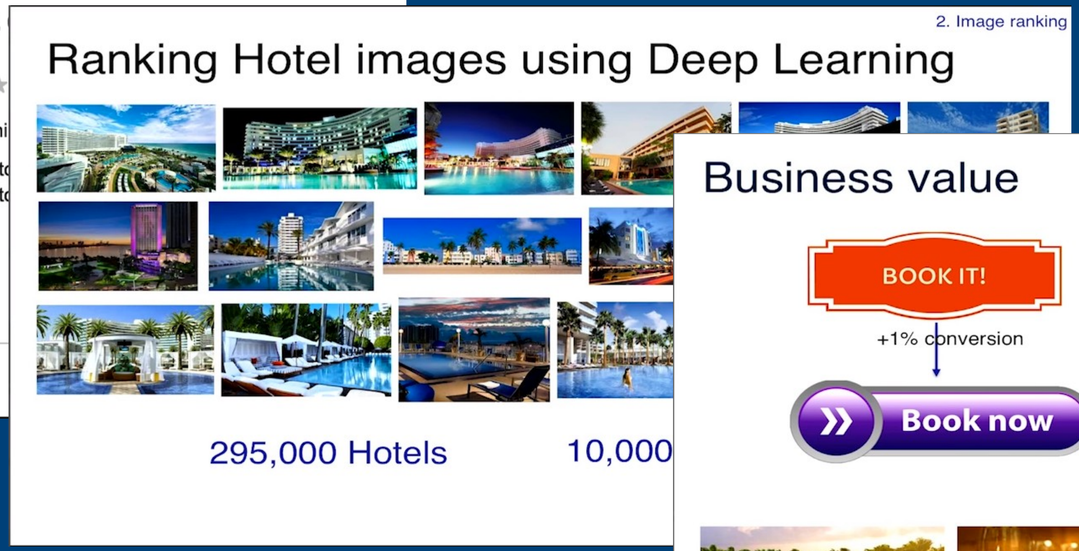
Chetan Kapoor | Senior Product Manager – EC2

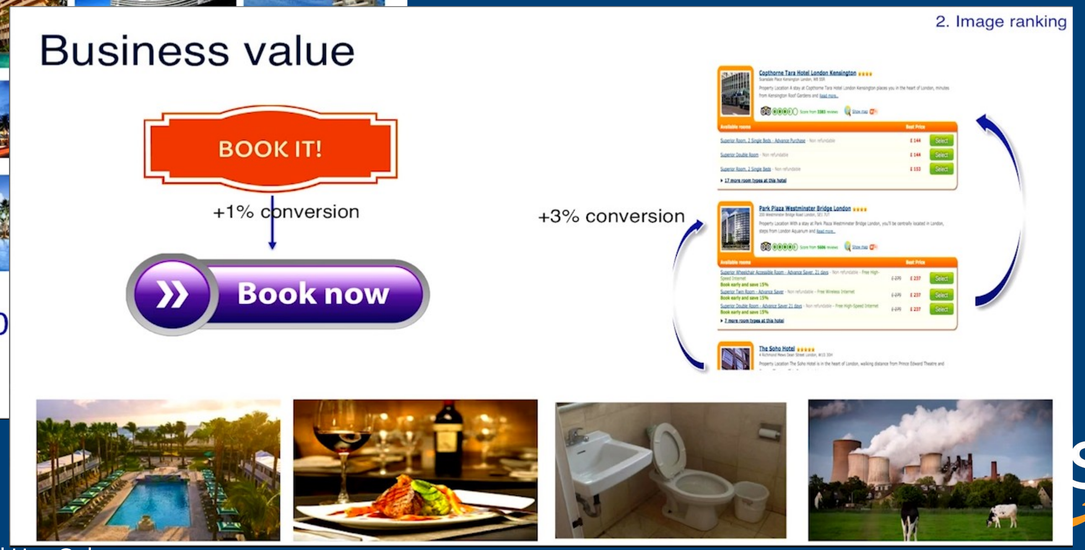
JANUARY 2018 | LAS VEGAS, NEVADA

Expedia – Ranking Hotel Images with Deep Learning

- “Expedia has over 10 million images from nearly 300,000 hotels, and ranking/sorting these manually would be difficult and time consuming
- Insert Artificial Intelligence to do this automatically.”







EC2 Compute Instance Types

General Purpose	Compute Optimized	Storage and IO Optimized	Memory Optimized	Accelerated Computing
M5	C5	I3	X1e	G3
T2	C4	H1	R4	F1
				P3

Making EC2 Instances Real...versus imaginary objects in the Cloud

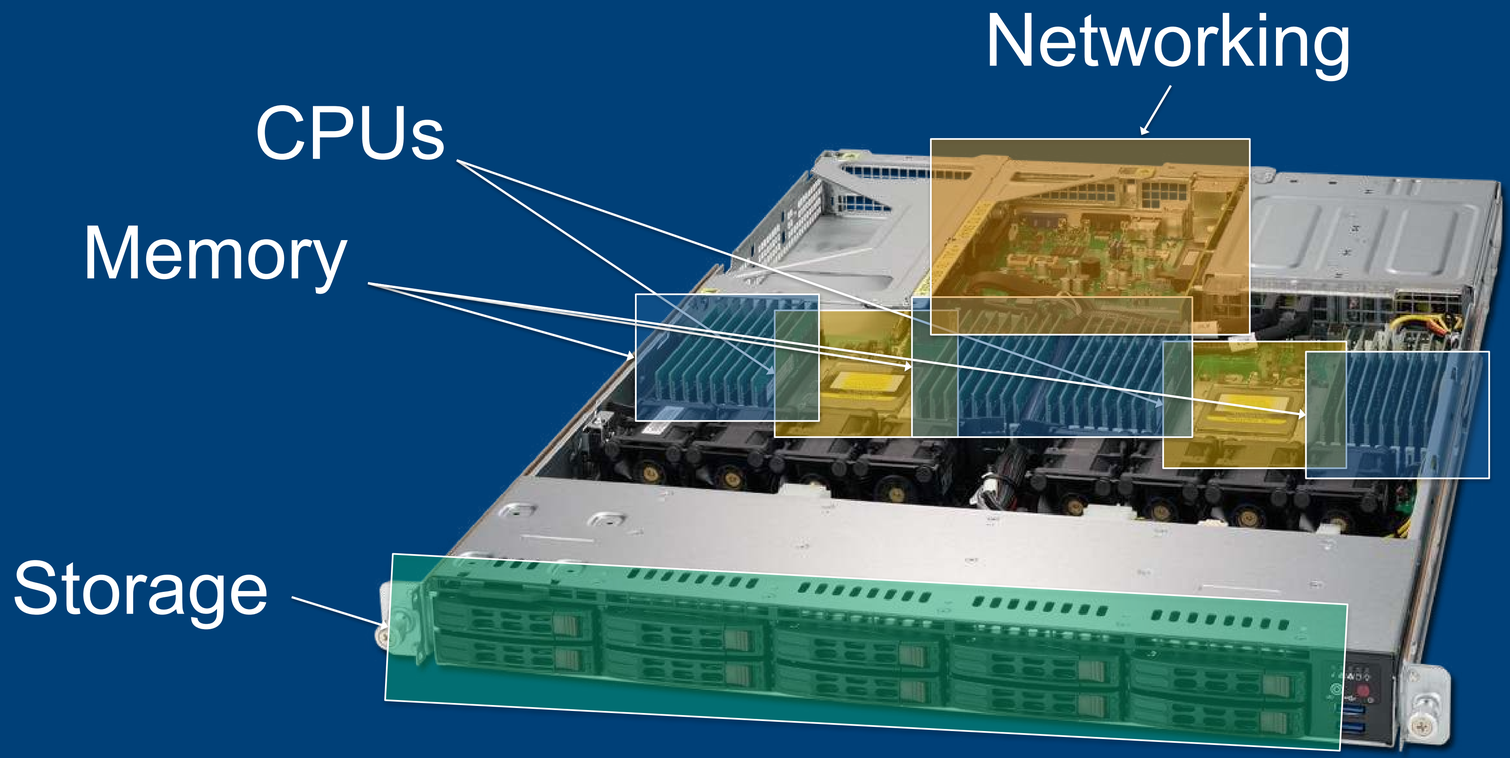


Image for illustration only. Does not represent an actual EC2 server

EC2 Accelerator Platforms

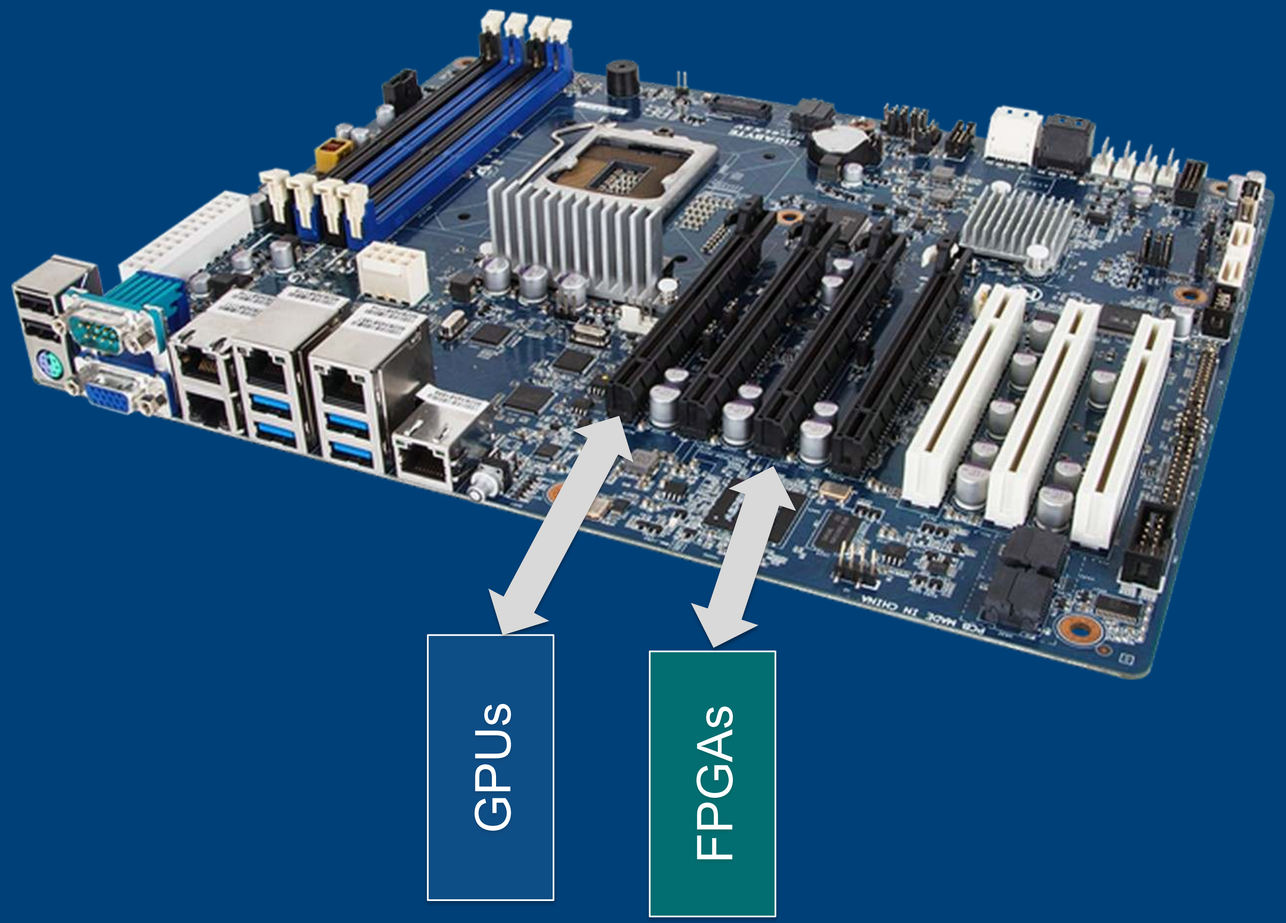
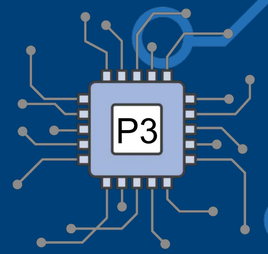


Image for illustration only. Does not represent an actual EC2 server

EC2 Accelerated Computing Instances

P3: GPU Compute Instance

- Up to 8 NVIDIA V100 GPUs in a single instance, with peer-to-peer NVLInk GPU interconnect.
- Industry's most powerful platform for Machine Learning and HPC application

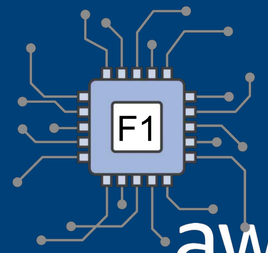


G3: GPU Graphics Instance

- Up to 4 NVIDIA M60 GPUs, with GRID Virtual Workstation features and licenses
- Designed for workloads such as 3D rendering, 3D visualizations, graphics-intensive remote workstations, video encoding, and virtual reality applications

F1: FPGA instance

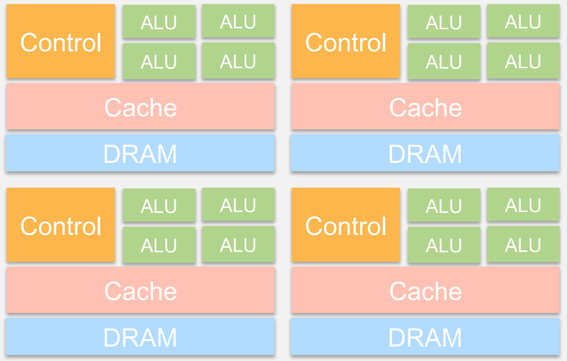
- Up to 8 Xilinx Virtex UltraScale+ VU9P FPGAs in a single instance. Programmable via VHDL, Verilog, or OpenCL
- Designed for hardware-accelerated applications including financial computing, genomics, accelerated search, and image processing



GPU - Compute

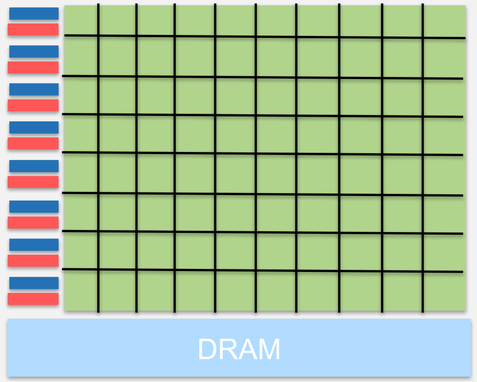
CPUs vs GPUs vs FPGA for Compute

CPU



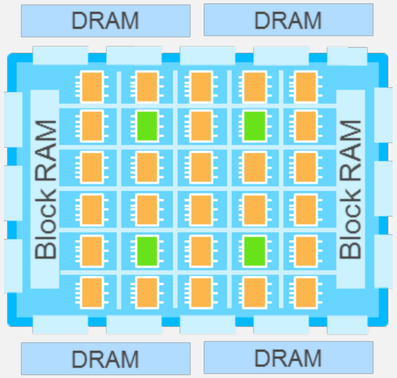
- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and datapath widths
- Highly effective at parallel execution

FPGA



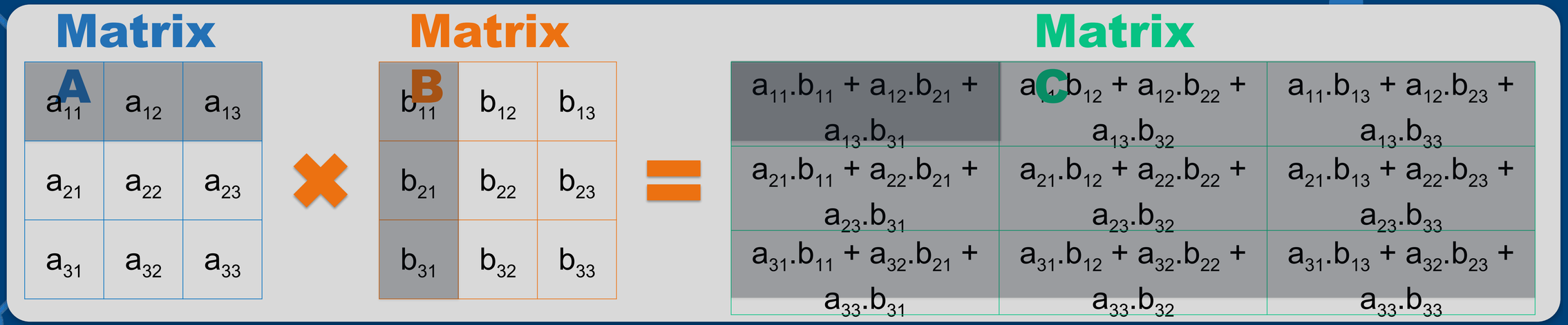
- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed execution

Next Generation of GPU Compute Instances– P3 Instances!

- Based on NVIDIA's latest GPU Tesla V100
- Industry's most powerful GPU-based platform
- Up to 6X more powerful than competitive cloud-based offering
- 1 PetaFLOP of computational performance in a single instance
- Provides up to 14X performance improvement over P2 for Machine Learning use-cases
- Up to 2.6X performance improvement over P2 for High-Performance Computing use-cases

GPUs for Compute

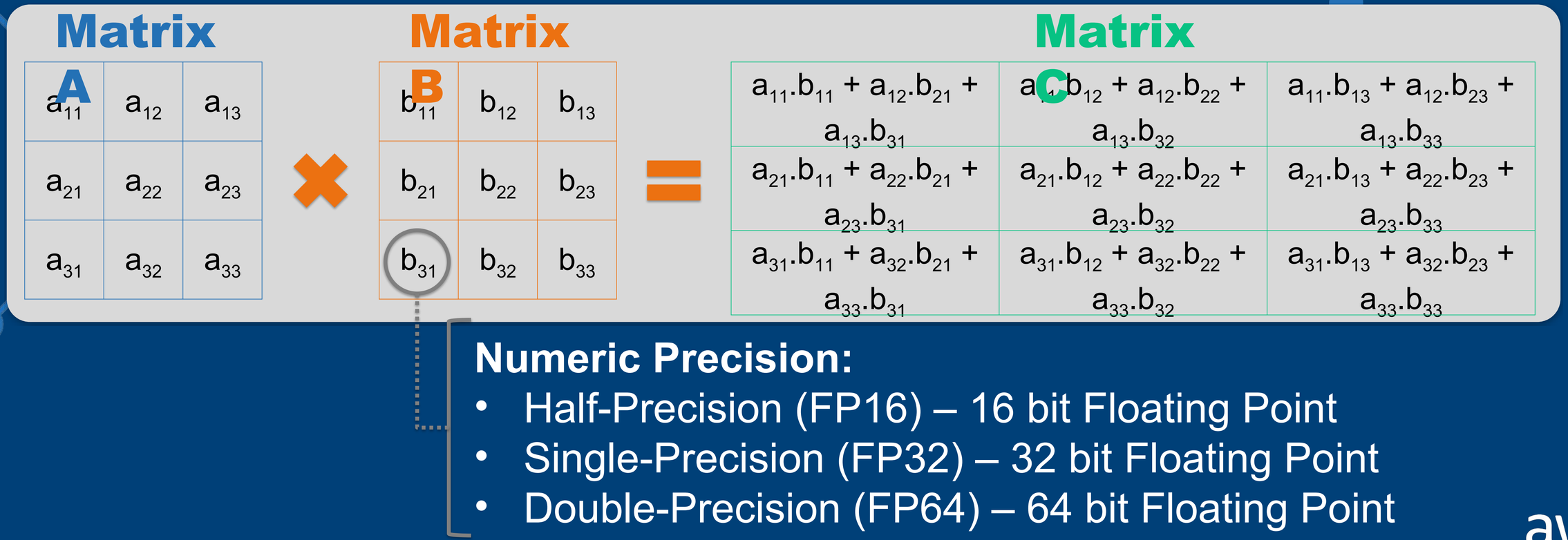
Matrix algebra, image processing, physics simulations benefit greatly with GPUs



These multiply and accumulate operations can be parallelized across 1,000s of core available in a typical GPU

GPUs for Compute

Matrix algebra, image processing, physics simulations benefit greatly with GPUs



Numeric Precision:

- Half-Precision (FP16) – 16 bit Floating Point
- Single-Precision (FP32) – 32 bit Floating Point
- Double-Precision (FP64) – 64 bit Floating Point

New Tensor Cores

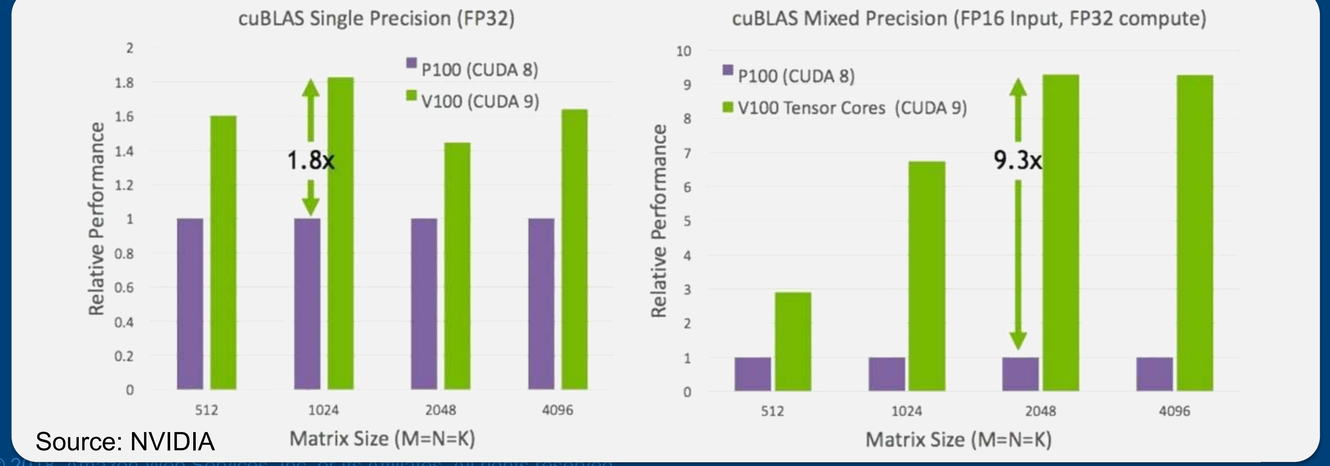
- In addition to 5,120 CUDA Cores, each Tesla V100 has 640 Tensor Cores with 125 TFLOPs of mixed-precision performance per GPU
- Each Tensor Core provides a 4x4x4 matrix processing array, which performs the operation D = A X B + C.

D (FP16 or FP32) = **A** (FP16) × **B** (FP16) + **C** (FP16 or FP32)

a_{11}	a_{12}	a_{13}	a_{14}
a_{21}	a_{22}	a_{23}	a_{24}
a_{31}	a_{32}	a_{33}	a_{34}
a_{41}	a_{42}	a_{43}	a_{44}

b_{11}	b_{12}	b_{13}	b_{14}
b_{21}	b_{22}	b_{23}	b_{24}
b_{31}	b_{32}	b_{33}	b_{34}
b_{41}	b_{42}	b_{43}	b_{44}

c_{11}	c_{12}	c_{13}	c_{14}
c_{21}	c_{22}	c_{23}	c_{24}
c_{31}	c_{32}	c_{33}	c_{34}
c_{41}	c_{42}	c_{43}	c_{44}

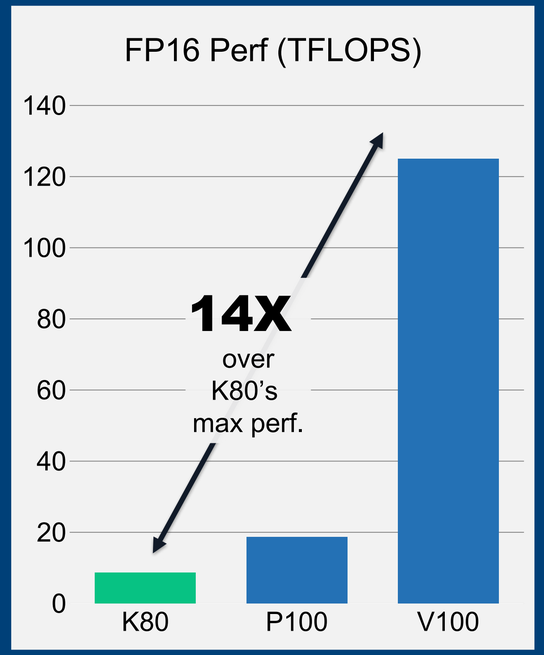


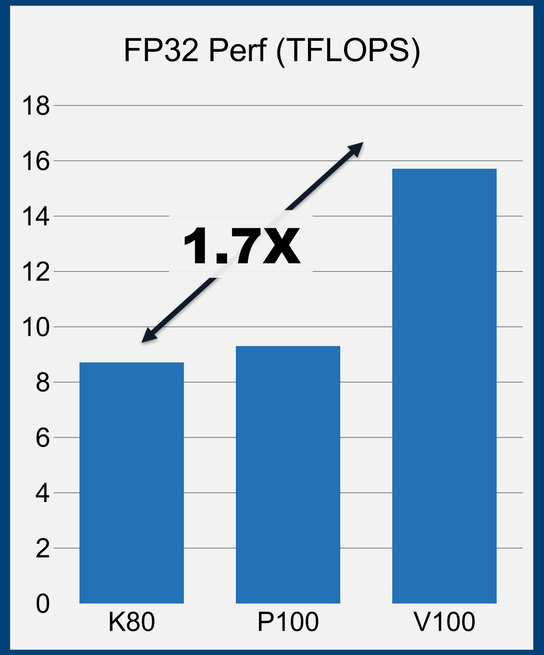
Source: NVIDIA

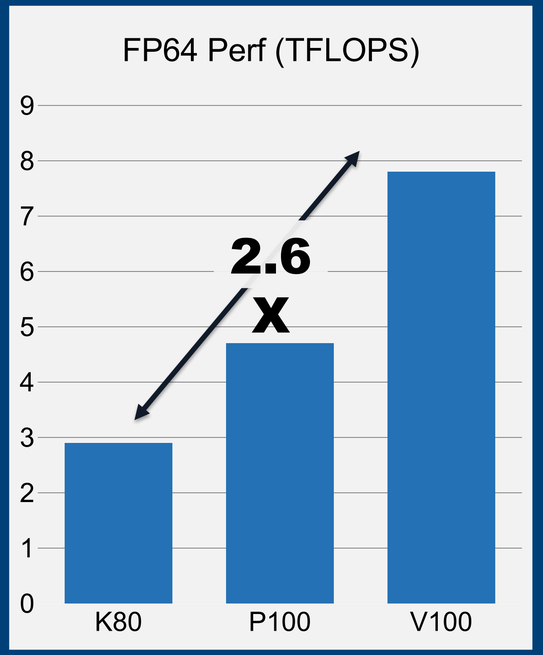
Tesla V100 GPU

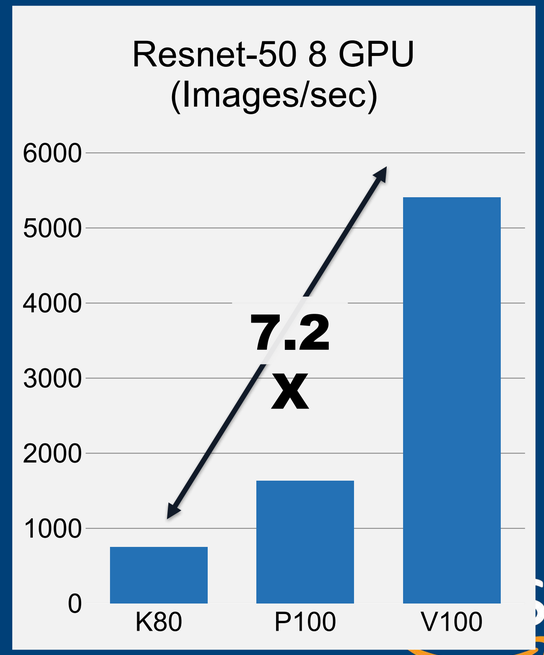
GPU Performance Comparison

- NVIDIA GPU Architectures:
    - Kepler > Maxwell > Pascal > Volta
- P2 Instances use K80 Accelerator (Kepler Architecture)
- P3 Instances use V100 Accelerator (Volta Architecture)









P3 Instance Sizes and Specifications

Instance Size	GPUs	Accelerator (V100)	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	Amazon EBS Bandwidth
P3.2xlarge	1	1	No	16	8	61	Up to 10 Gbps	1.7 Gbps
P3.8xlarge	4	4	NVLink	64	32	244	10 Gbps	7 Gbps
P3.16xlarge	8	8	NVLink	128	64	488	25 Gbps	14 Gbps

- P2 Instances used K80 accelerator, each with 2 GPUs
- P2.16xlarge provided access to 8 x K80s (16 GPU total)

P3 Instance Sizes and Specifications

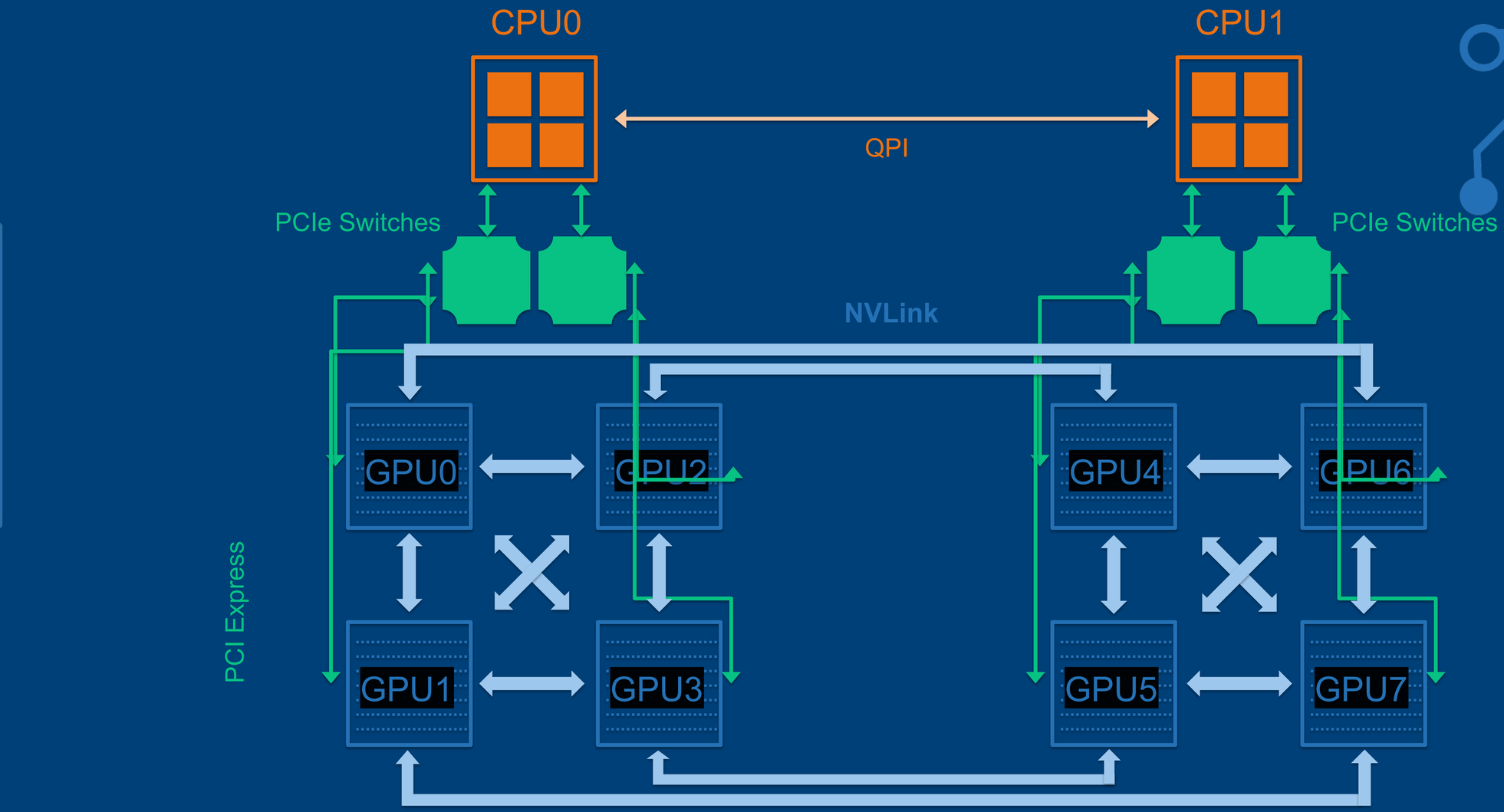
Instance Size	GPUs	Accelerator (V100)	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	Amazon EBS Bandwidth
P3.2xlarge	1	1	No	16	8	61	Up to 10 Gbps	1.7 Gbps
P3.8xlarge	4	4	NVLink	64	32	244	10 Gbps	7 Gbps
P3.16xlarge	8	8	NVLink	128	64	488	25 Gbps	14 Gbps

- P3 instances provide GPU-to-GPU data transfer over **NVLink**
- P2 instanced provided GPU-to-GPU data transfer over **PCI Express**

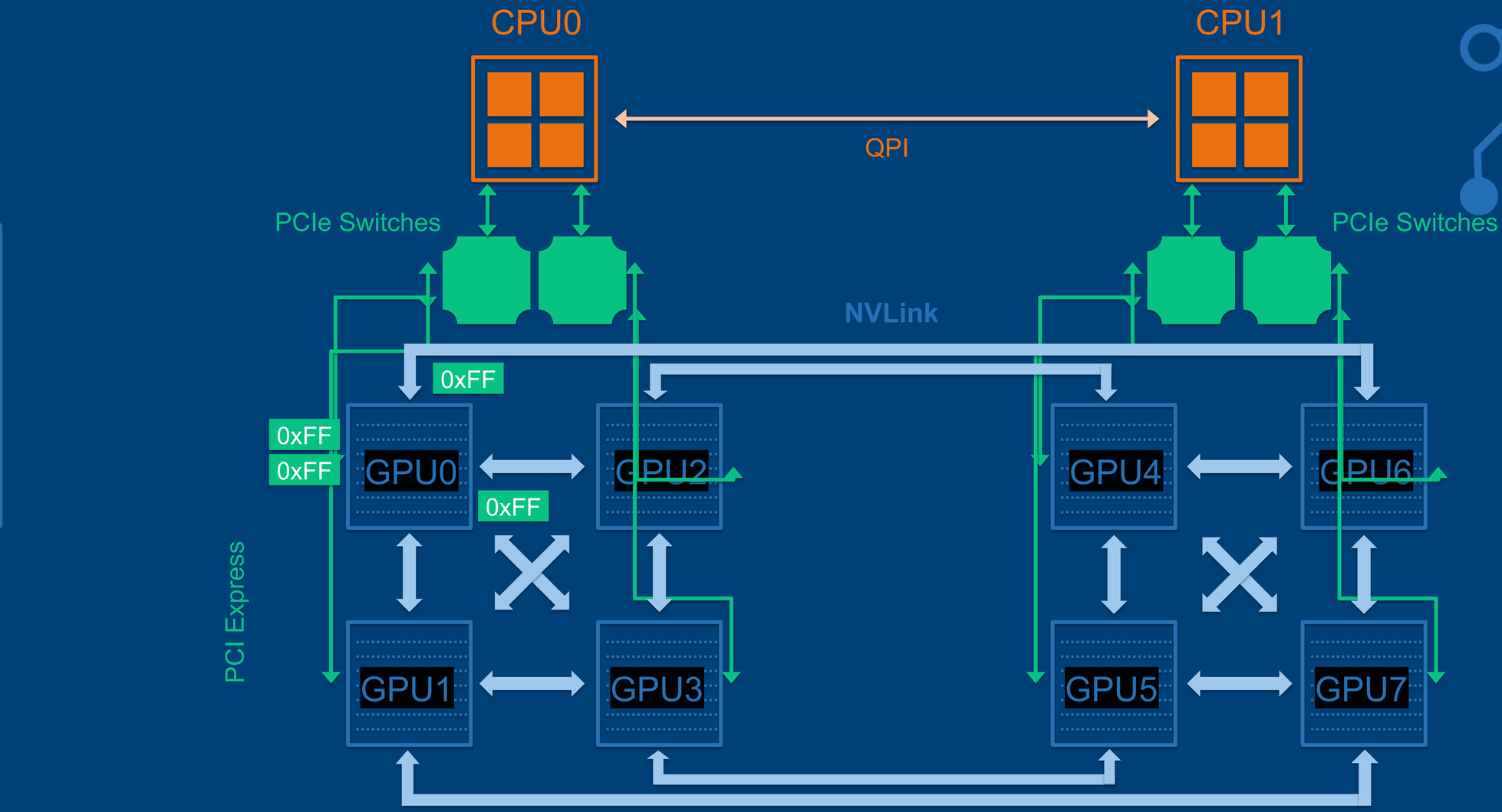
P3 vs P2 Peer-to-Peer Configurations

Description	P3.16xlarge	P2.16xlarge	P3 GPU Performance Improvement
Number of GPUs	8	16	-
Number of Accelerators	8 (V100)	8 (K80)	
GPU – Peer to Peer	NVLink – 300 GB/s	PCI-Express - 32 GB/s	9.4X
CPU to GPU Throughput PCIe throughput per GPUs	8 GB/s	1 GB/s	8X
CPU to GPU Throughput Total instance PCIe throughput	64 GB/s (Four x16 Gen3)	16 GB/s (One x16 Gen3)	4X

P3 PCIe and NVLink Configurations



P3 PCIe and NVLink Configurations



Data Ingestion Options

- Within a P3 instance, we have maxed out the data throughput in to GPUs (PCI Express to/from host CPUs) and between GPUs (NVLink)
- For customers to maintain high utilization of GPUs, they need high throughput data stream coming in to P3 instances
- **Option 1:** Use Multiple EBS Volumes
  - Each Provisioned IOPS SSD (io1) EBS volume and provide about **330 MB/s** of read or write throughput (need to be provisioned with 20,000 IOPS)
  - Customer can use independent EBS volume or combine multiple volumes via RAID to create a single logical volume (5 io1 volumes can support **1.65 GB/s)**
  - http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/raid-config.html
- **Option 2**: Amazon S3 -> EC2
  - We have increased data transfer from Amazon S3 directly in to EC2 from 5 Gbps to 25Gbps
  - Need to parallelize connections to Amazon S3 by using the TransferManager available in Amazon S3's Java SDK
  - https://docs.aws.amazon.com/sdk-for-java/v1/developer-guide/examples-s3-transfermanager.html

Software Support for P3

Required Drivers & Libraries

- Hardware Driver version 384.81 or newer
- CUDA 9 or newer
- CuDNN 7 or newer & NCCL 2.0 or newer
    - Generally packaged with CUDA

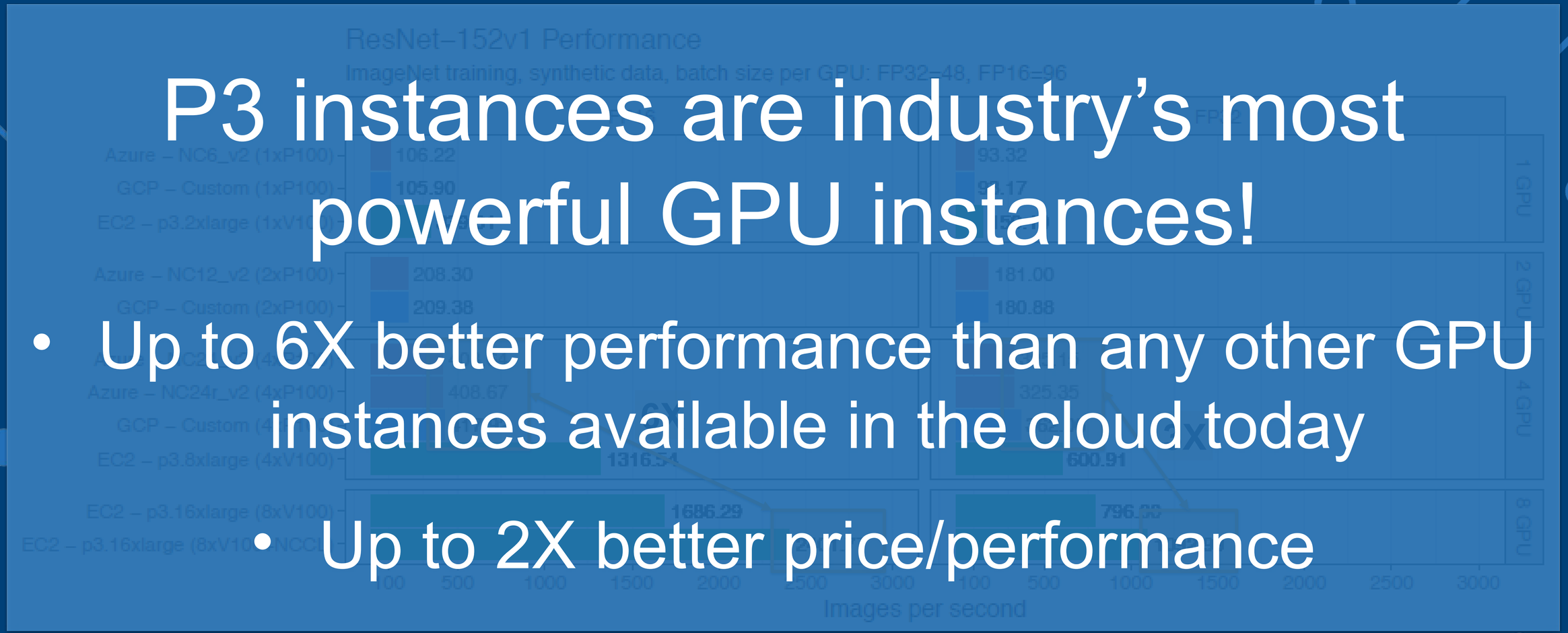
Machine Learning Frameworks

- For customers to take advantage of the new Tensor Cores in V100 GPUs, they will need to use latest distros of ML framework
- **MXNet** and **Caffe 2** have formally released support for V100 GPUs
- **PyTorch** has the relevant PRs merged in the master but does not have a released build yet. Customers will need to build from source
- **Tensorflow** has most of the PRs merged in 1.4 master. Some issues are still outstanding. TF 1.5 is targeting to have stable support for V100.
- http://docs.nvidia.com/deeplearning/sdk/pdf/Training-Mixed-Precision-User-Guide.pdf

Competition

	K80	P100	V100
GCE	Available	Available	Announced
Azure	Available	Available	Announced
Oracle	N/A	Announced	Announced
IBM	Available	Announced	N/A
NVIDIA	N/A	Available	Available

P3 ML Benchmarking

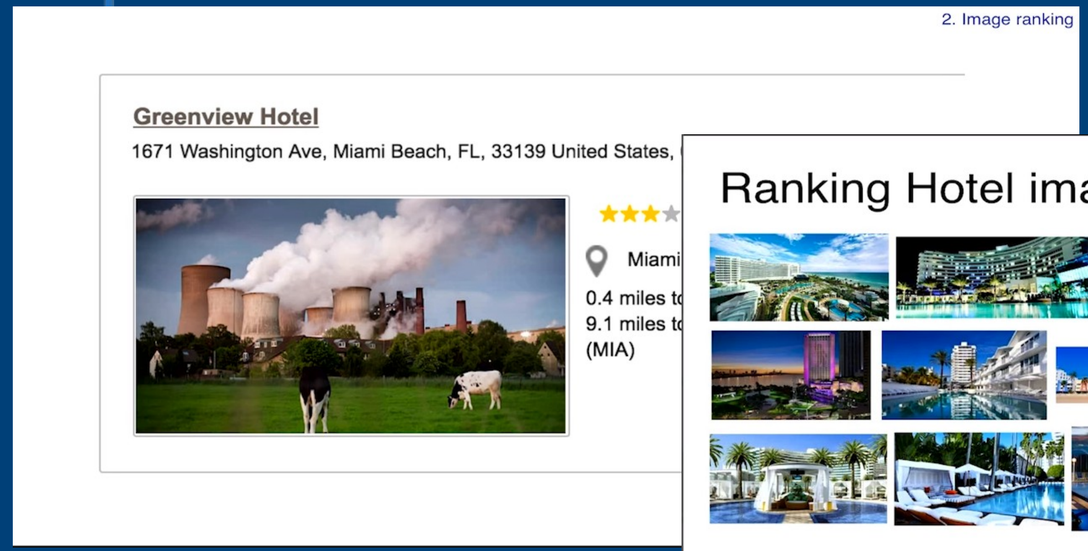


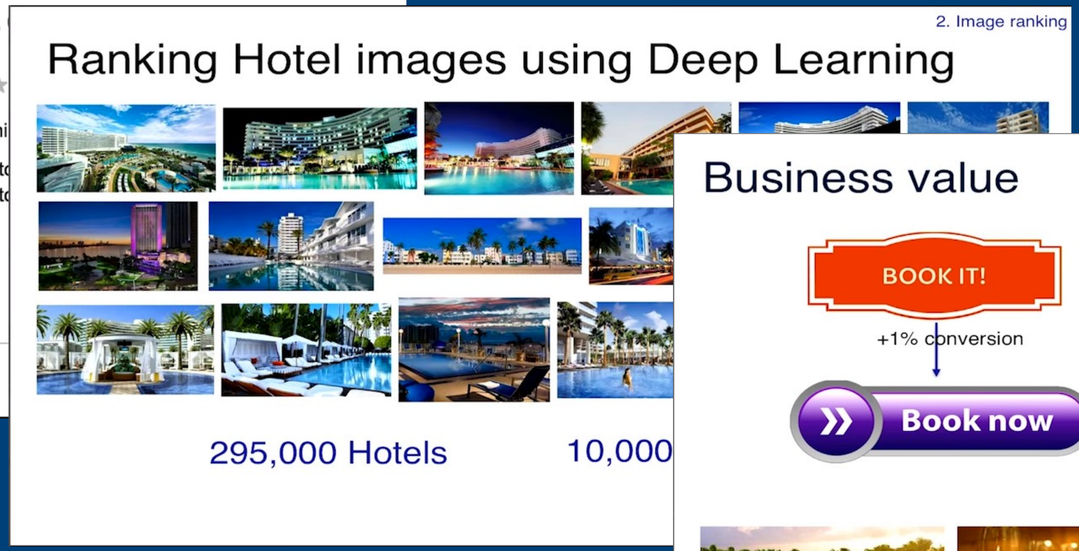
P3 instances are industry's most powerful GPU instances!

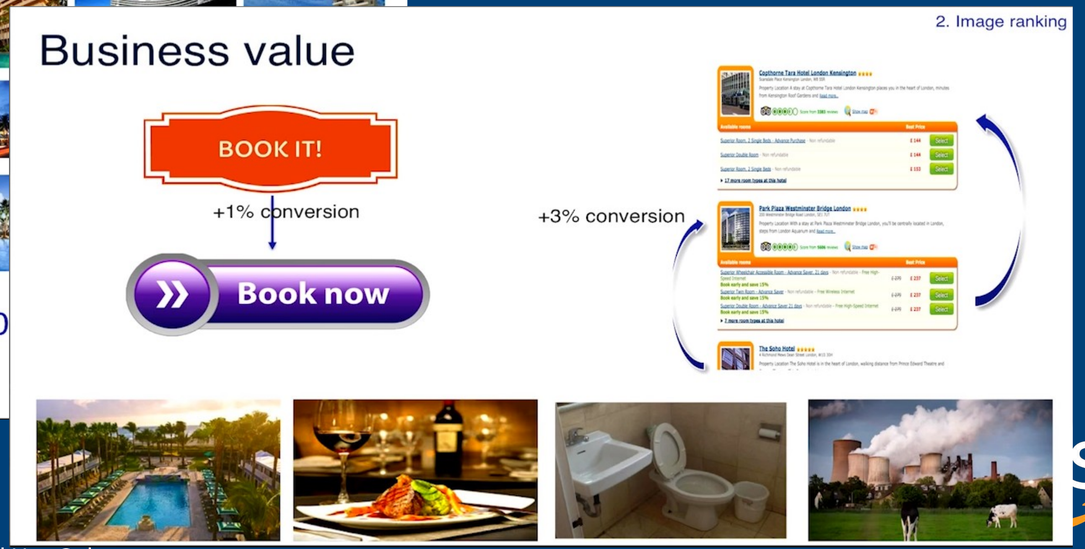
- Up to 6X better performance than any other GPU instances available in the cloud today
- Up to 2X better price/performance

Expedia – Ranking Hotel Images with Deep Learning

- Expedia has over 10 million images from nearly 300,000 hotels, and ranking/sorting these manually would be difficult and time consuming
- Insert Artificial Intelligence to do this automatically.



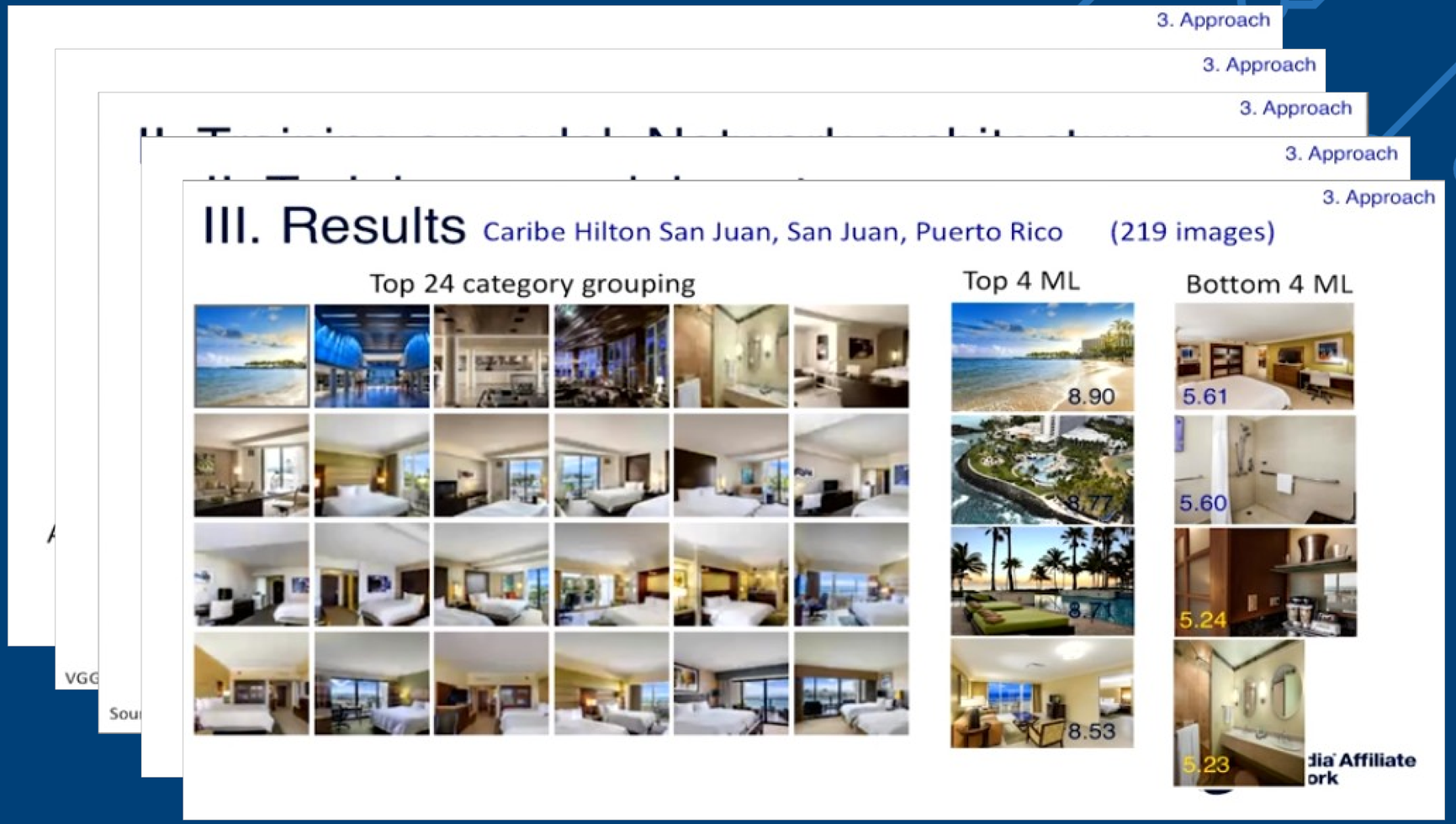




Expedia – Ranking Hotel Images with Deep Learning

Three core steps in building this solution

1. Building a dataset
2. Training the model
3. Visualizing the results



WD – HPC Use-case

Western Digital

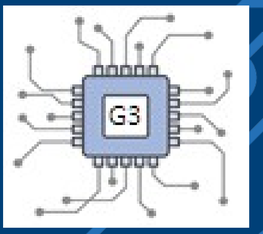
Western Digital is an industry-leading provider of storage technologies and solutions that enable people to create, leverage, experience, and preserve data. David Hinz, Senior Director Cloud, and Data Center Operations, said:

"Our engineering and product development teams use high performance computing to run 10s of thousands of simulations for all areas needed to deliver new hard disk drive (HDD) and solid-state storage solutions. The simulations include materials sciences, heat flows, magnetics, and data transfer simulations to improve disk drive and storage solution performance and quality. Based upon early testing, the new P3 instances can allow engineering teams to run GPU-accelerated modeling and simulations at least three times faster than currently deployed GPU solutions. We are looking forward to using the P3 instances in production as a cost-effective and performant way to provide HPC solutions to our engineering teams."

aws

GPU - Graphics

AWS G3 GPU instances



- Up to four NVIDIA M60 GPUs
- Includes GRID Virtual Workstation features and licenses, supports up to four monitors with 4096x2160 (4K) resolution
- Includes NVIDIA GRID Virtual Application capabilities for application virtualization software like Citrix XenApp Essentials and VMWare Horizon, supporting up to 25 concurrent users per GPU
- Hardware encoding to support up to 10 H.265 (HEVC) 1080p30 streams, and up to 18 H.264 1080p30 streams per GPU
- Designed for workloads such as 3D rendering, 3D visualizations, graphics-intensive remote workstations, video encoding, and virtual reality applications

Instance Size	GPUs	vCPUs	Memory (GiB)	Linux price per hour (IAD)	Windows price per hour (IAD)
g3.4xlarge	1	16	122	$1.14	$1.88
g3.8xlarge	2	32	244	$2.28	$3.75
g3.16xlarge	4	64	488	$4.56	$7.50

G3 GRID Workstation vs. Virtual Application Modes

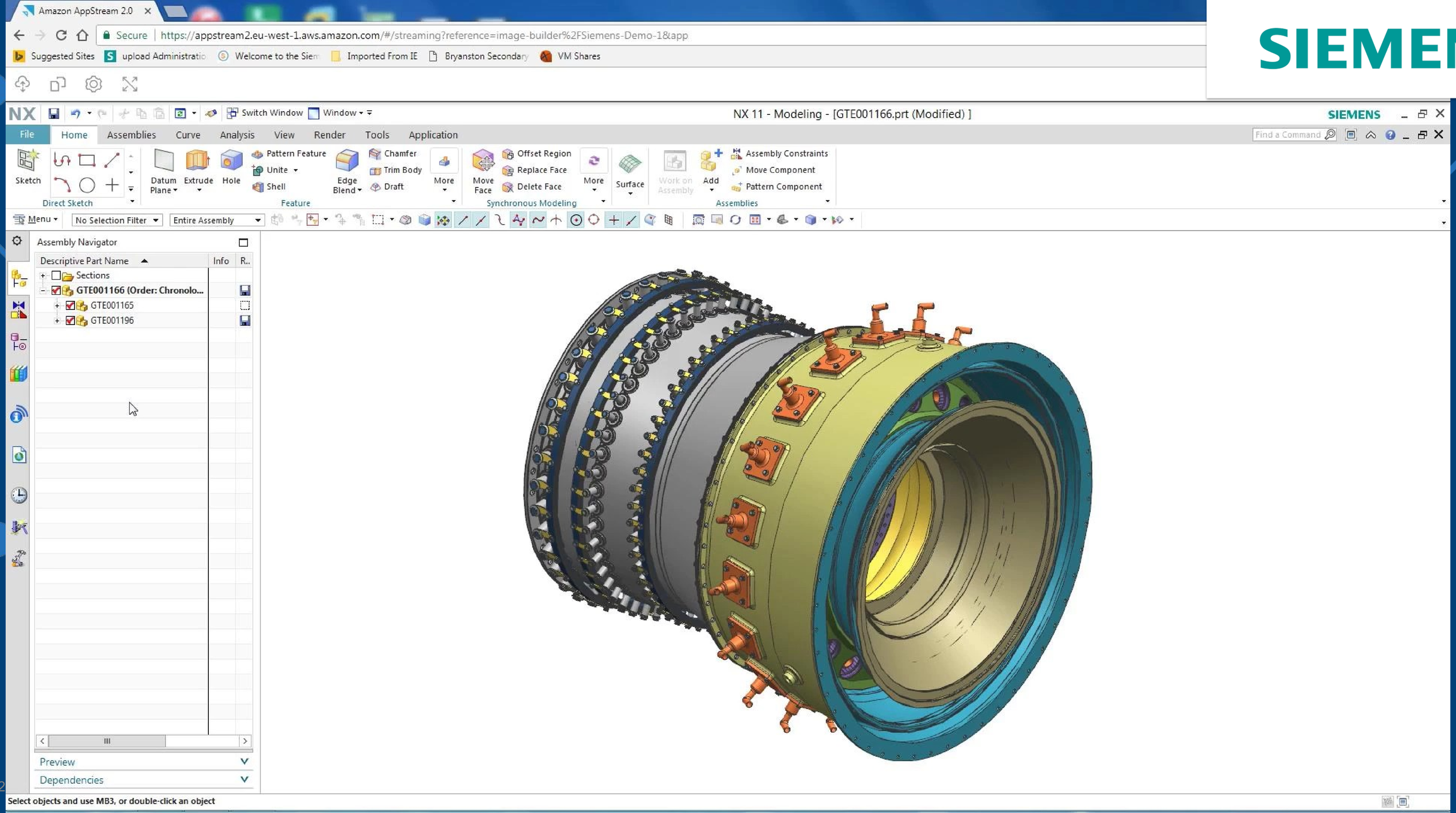
Feature	Workstation	Virtual Applications
	For professional 3D graphics applications at full performance	For PC-level applications, server-hosted RDSH desktops, XenApp
Concurrent users per GPU	1	25
NVIDIA Quadro feature	Yes	No
Desktop virtualization	Yes	No
Display & Resolution	4 monitors with 4096 x 2160 resolution	N/A
CUDA, OpenGL, DirectX and OpenCL	Yes	Yes

How to switch between the modes: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/activate_grid.html

G3 Use Cases

- Desktop and Application Virtualization
  - Productivity and consumer apps
  - Design and engineering
  - Media and entertainment post-production
- Media and entertainment: video playout/broadcast, encoding/transcoding
- Cloud GPU rendering & visualization, such as high end car configurators, AR/VR
- Cloud Gaming

Desktop and Application Virtualization

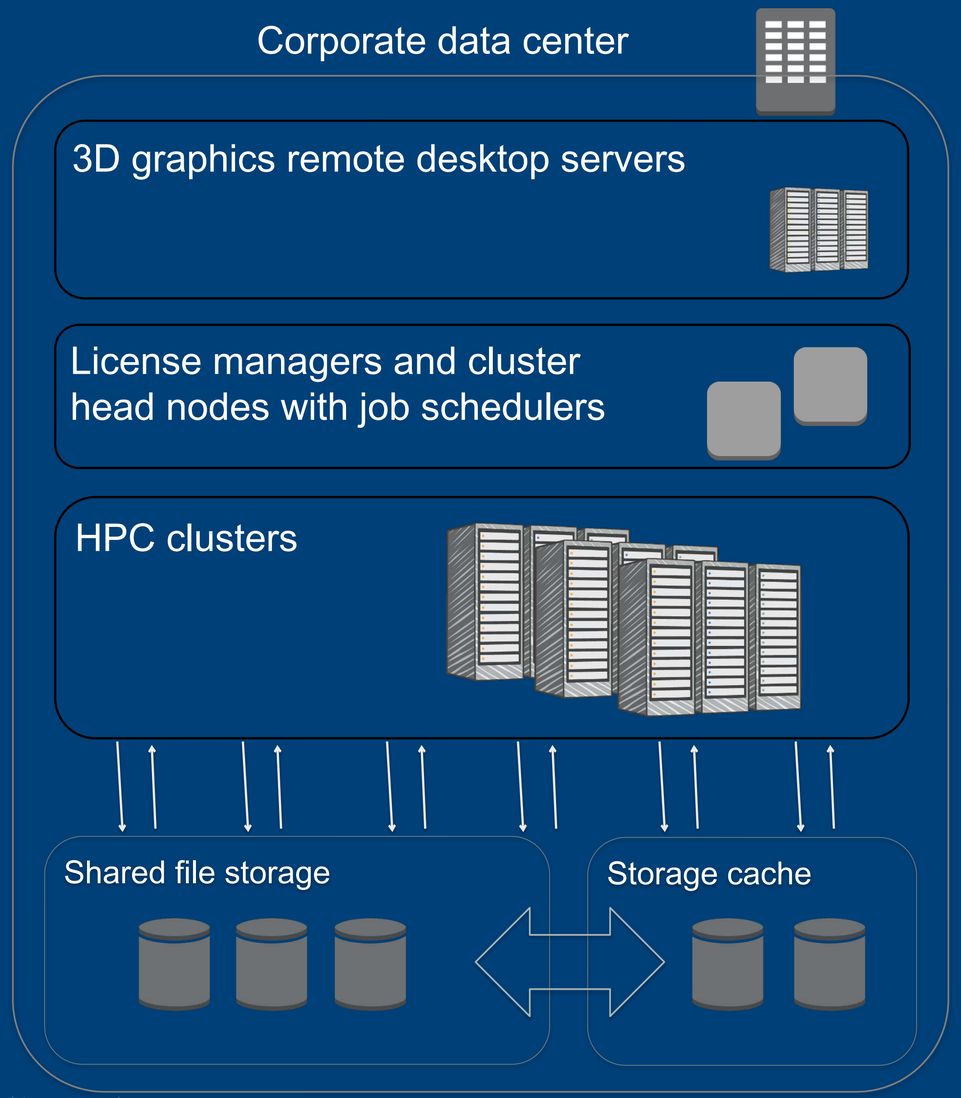

Amazon AppStream 2.0
Secure
https://appstream2.eu-west-1.aws.amazon.com/#/streaming?reference=image-builder%2FSiemens-Demo-1&app
Suggested Sites
upload Administratio
Welcome to the Sien
Imported From IE
Bryanston Secondary
VM Shares
NX
Switch Window
Window
NX 11 - Modeling - [GTE001166.prt (Modified)]
SIEMENS
File
Home
Assemblies
Curve
Analysis
View
Render
Tools
Application
Find a Command
Sketch
Direct Sketch
Datum Plane
Extrude
Hole
Pattern Feature
Unite
Shell
Edge Blend
Chamfer
Trim Body
Draft
More
Feature
Move Face
Offset Region
Replace Face
Delete Face
More
Synchronous Modeling
Surface
Work on Assembly
Add
Assembly Constraints
Move Component
Pattern Component
Assemblies
Menu
No Selection Filter
Entire Assembly
Assembly Navigator
Descriptive Part Name
Info
R..
Sections
GTE001166 (Order: Chronolo...
GTE001165
GTE001196
Preview
Dependencies
Select objects and use MB3, or double-click an object

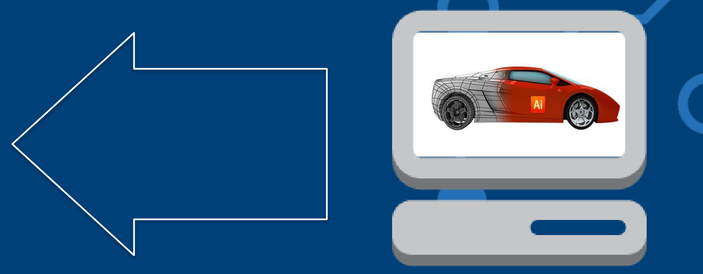
SIEMENS

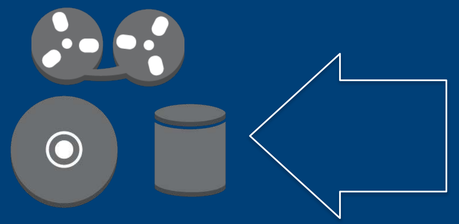

FRAMESTORE
MARVEL STUDIOS
GUARDIANS OF THE GALAXY VOL. 2
IN 3D, REAL D 3D MAY 5 AND IMAX 3D

Traditional On-Prem Stack

Traditional CAD/CAE/EDA infrastructure is inflexible, often poorly utilized, and must be managed through a years-long life cycle.







Stack on AWS

On AWS, secure and well-optimized clusters and design chambers can be created, operated, and torn down in just minutes.

Virtual private cloud on AWS

VPC

3D graphics virtual workstations (backed by GPUs in EC2)

DCV

amazon AppStream 2.0

License managers and cluster head nodes with job schedulers

Cloud-based, Auto-Scaling clusters

AWS Batch

Shared file storage

Storage **cache**

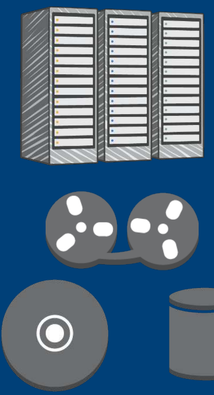
Amazon S3 and Amazon Glacier

Or Citrix, Frame, Teradici, etc.

Thin or zero client
- No local data -

Corporate data center

On-premises IT resources



AWS Snowball

AWS Direct Connect

Opportunity: Desktop Apps

ISVs:

- Autodesk
- Adobe
- Siemens
- Dassault
- PTC
- ESRI
- Bentley
- Ansys
- Altair
- The Foundry
- SideFX

Streaming Solutions:

- Amazon Appstream
- Amazon Workspaces
- Amazon NICE DCV
- Citrix XenApp/XenDesktop
- VMWare Horizon
- Frame
- Teradici
- OTOY X.IO
- UberCloud
- Leostream
- Bebop

- We’re actively working with several ISVs and streaming solution providers to build/certify AWS-based solutions
- **CTA**: Engage customers to understand their challenges in moving to the cloud

Media and Entertainment Workflows

Discovery COMMUNICATIONS™

5-DAY FORECAST

MON	TUE	WED	THU	FRI
56	52	65	60	58
38	35	45	40	25

MARK JOHNSTON LIVE IN CONCERT
AUGUST 21 - AT THE BROWNS AMPHITHEATRE. TICKETS: TIX.COM

WEST 456 HWY EAST & WEST

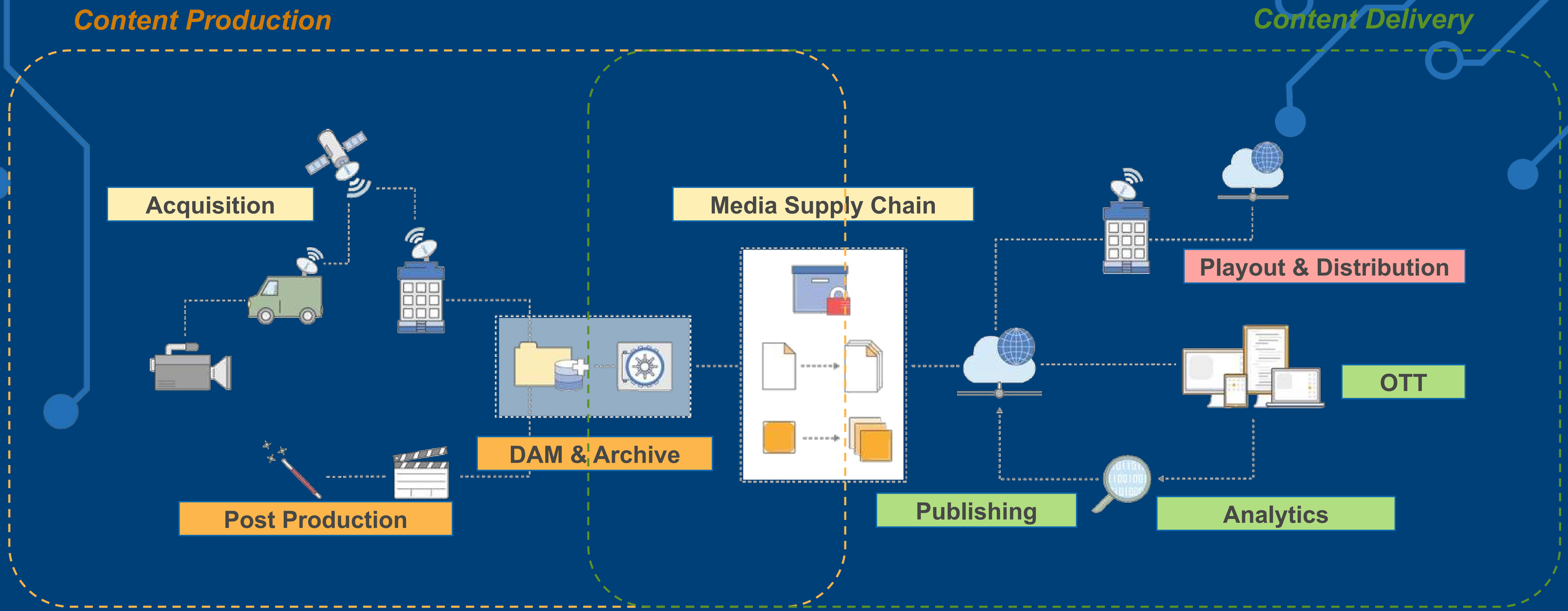
- Downtown industrial fire rages on into second day. Locals urged to evacuate. “This is not the time for heroics or hesitation.” - Police Comissioner Darko Lucic.

NBA
RAPTORS
LA LAKERS
1:10 PM

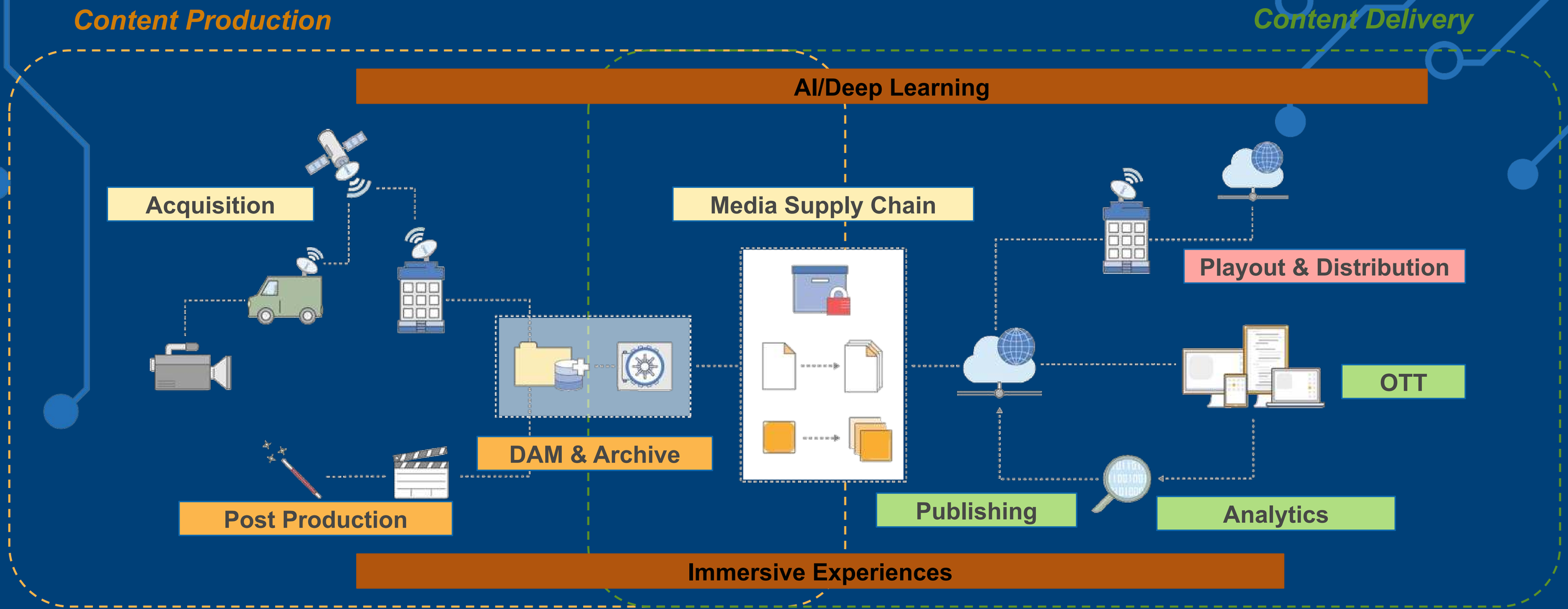
MLB
NFL
CFL

Toronto closer R. Osuna out indef. for ‘Tommy-John’ surgery. “Huge loss,” - John Gibbons.

M&E workload segments



M&E workload segments



M&E workload segments & GPU relevance

Content Production

Content Delivery

AI/Deep Learning

Image/Video Deep Learning
- Custom AI models

Text and Speech Models

Analytics/ML Model

Mostly with P2 instances

Media Supply Chain

Real-time Rendering
- Rendering overlays
- Playout scheduling

Playout & Distribution

Remote Application Streaming
- Video editing
- QC
- Graphics artist workstations
- VFX/Rendering GPU based

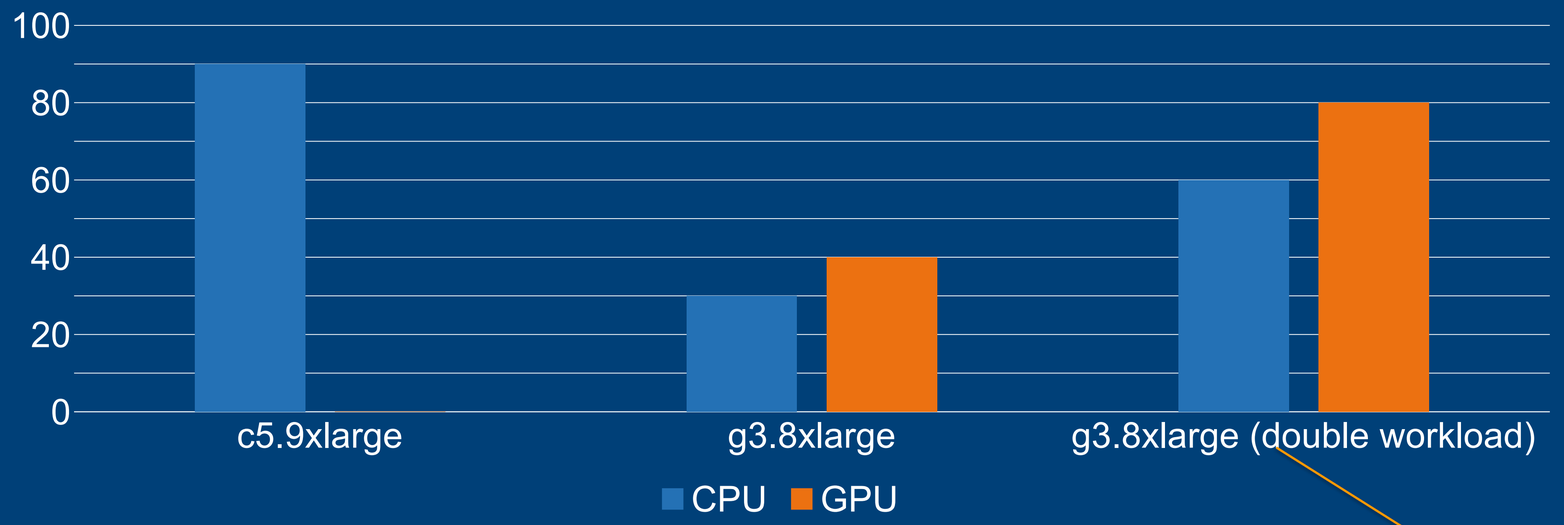
Workstation GPU (G3 Required)

Batch workloads
- Encoding/Transcoding
- Just in-time
- Live scenarios
- Stitching in the case or VR/AR

Post Production

Immersive Experiences

Pipeline cost



Model	vCPU	Memory (GiB)	On-demand price/hour
g3.8xlarge	32	244	$2.28
r4.8xlarge	32	244	$2.13
m5.12xlarge	48	192	$2.30
c5.9xlarge	36	72	$1.53

$1.14/hour

ZeroLight

Web Configurator

Experience the Audi Q2 in Cloud-streamed 3D with a 2D On Demand Option for high latency users


parsec
0:25.686
POTENTIAL
DELL
SIMPLE, LOW-LATENCY GAME STREAMING

Common Challenges

Graphics Certification

EULA of ISV Software

Pricing

Virtualization Expertise/Skills

What We're Doing About Them

Certification: Graphics Certification Program

EULA: ISV Engagements and framework for AMIs

Pricing: Evaluating

Virtualization Expertise/Skills: Evaluating Expansion of End User Computing ProServ Specialty Practice

Summary

- AWS provides our customers with the widest choice of **accelerated computing** instances in the market
  - P3 – Industry's most powerful GPU Compute instances
  - F1 – Industry's first (and currently only) FPGA accelerated instances
  - G3 – Ideal instances for Graphics workloads
- These instances are ideal for **high growth markets** such as:
  - Machine Learning
  - High Performance Computing
  - Graphics
- In combination with our portfolio of AWS Services (Networking, Storage, etc.), broad regional availability, we have industry's best platform to **help our customers build exciting applications.**

Thank you