Project Dell NVIDIA AMD

DELL Technologies
PARTNER PROGRAM

Explaining the A of DNA

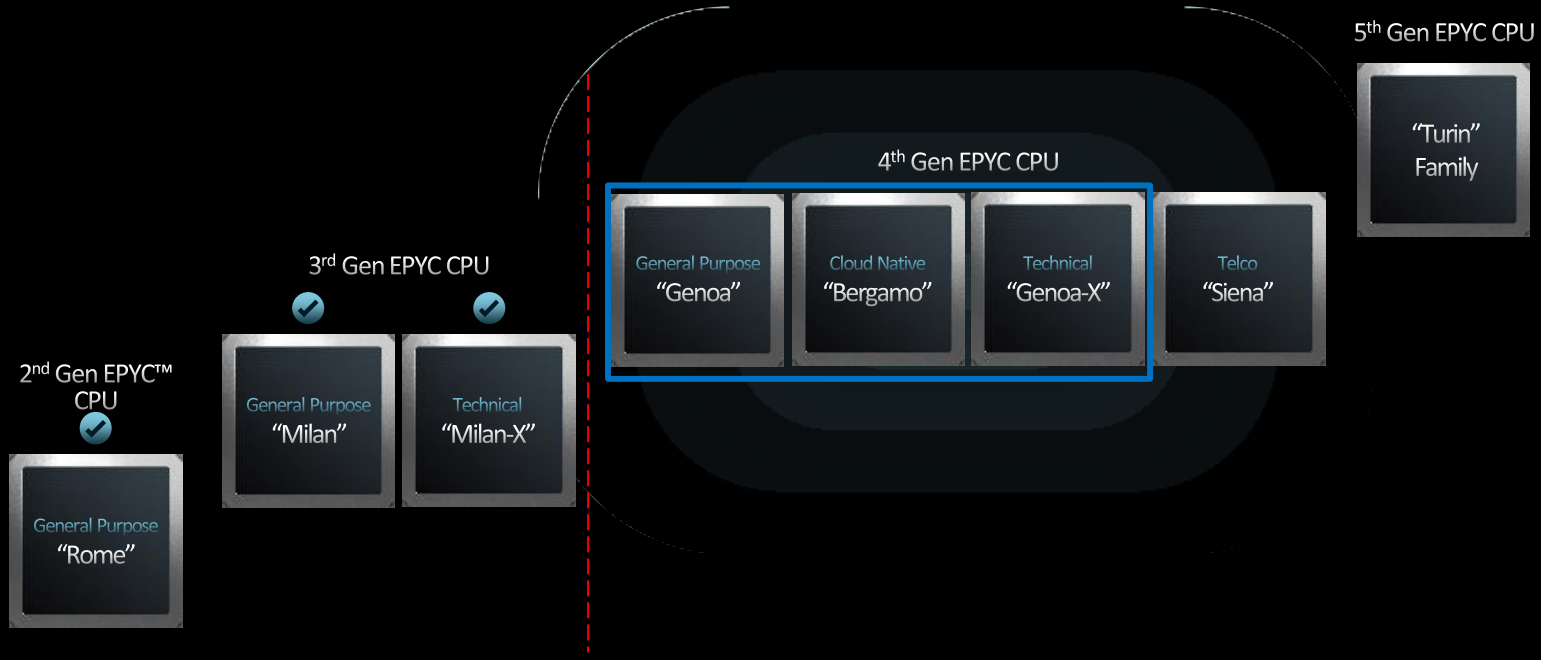
Tim Carlson, Dell AMD Lead

Dell PowerEdge AMD 4th Gen EPYC™ Performance

- 50% more cores
- Up to 121% increased performance
- Up to 55% CPU Performance per Watt improvements
- Up to 60% more storage

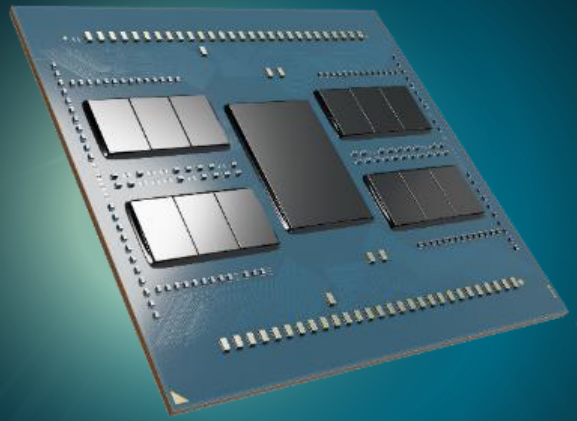
DELL Technologies

INDUSTRY LEADING OPTIMIZED SILICON



All roadmaps are subject to change.

AMD EPYC™ 9004 PROCESSORS – "GENOA"



WORLD'S HIGHEST PERFORMANCE x86 CPU*

EXTENDING COMPUTE LEADERSHIP

Leadership Socket and Per-Core Performance
Up to 96 "Zen 4" Cores in 5nm

Leadership Memory Bandwidth and Capacity
12 Channels DDR5 with up To 12TB of memory capacity**

Next Generation I/O
PCIe® Gen 5 – 128 lanes + Memory Expansion with CXL

Advances in Confidential Computing
Memory Encryption | Direct and CXL Attached

** Projected performance based on AMD internal testing and subject to change.*
*** with 256GB DIMMs and 2DPC in a 2P server*

See endnotes: SP5-001

AMD EPYC™ 9004 “GENOA” - AT A GLANCE

COMPUTE

- AMD “Zen4” x86 cores (Up to 12 CCDs / 96 cores / 192 threads)
- 1MB L2/Core, Up to 32MB L3/CCD
- ISA updates: BFLOAT16, VNNI, AVX-512 (256b data path)
- Memory addressability with 57b/52b Virtual/Physical Address

- Updated IOD and internal AMD Gen3 Infinity Fabric™ architecture with increased die-to-die bandwidth

- Target TDP range: Up to 400W (cTDP)

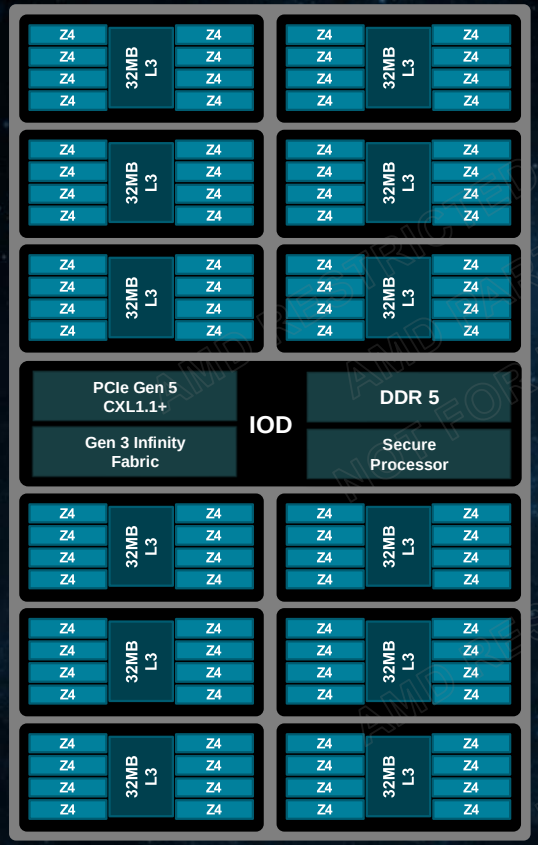
- Updated RAS

MEMORY

- 12 channel DDR5 with ECC up to 4800 MHz
- Option for 2,4,6, 8, 10, 12 channel memory interleaving[1]

- RDIMM, 3DS RDIMM

- Up to 2 DIMMs/channel capacity with up to 12TB in a 2 socket system (2DPC, 256GB 3DS RDIMMs)[1]



BLUE font indicates significant upgrades with EPYC 9004.

SP5 PLATFORM

- New socket, increased power delivery and VR
- Up to 4 links of Gen3 AMD Infinity Fabric™ with speeds of up to 32Gbps
- Flexible topology options

- Server Controller Hub (USB, UART, SPI, I2C, etc.)

INTEGRATED I/O – NO CHIPSET

Up to 160 IO lanes (2P) of PCIe® Gen5

- Speeds up to 32Gbps, bifurcations supported down to x1
- Up to 12 bonus PCIe Gen3 lanes in 2P config (8 lanes–1P)
- Up to 32 IO lanes for SATA
- SDCI (Smart Data Cache Injection) *

- 64 IO Lanes support for CXL1.1+ with bifurcations supported down to x4

SECURITY FEATURES

Dedicated Security Subsystem with enhancements

Secure Boot, Hardware Root-of-Trust

SME (Secure Memory Encryption)

SEV-ES (Secure Encrypted Virtualization & Register Encryption)

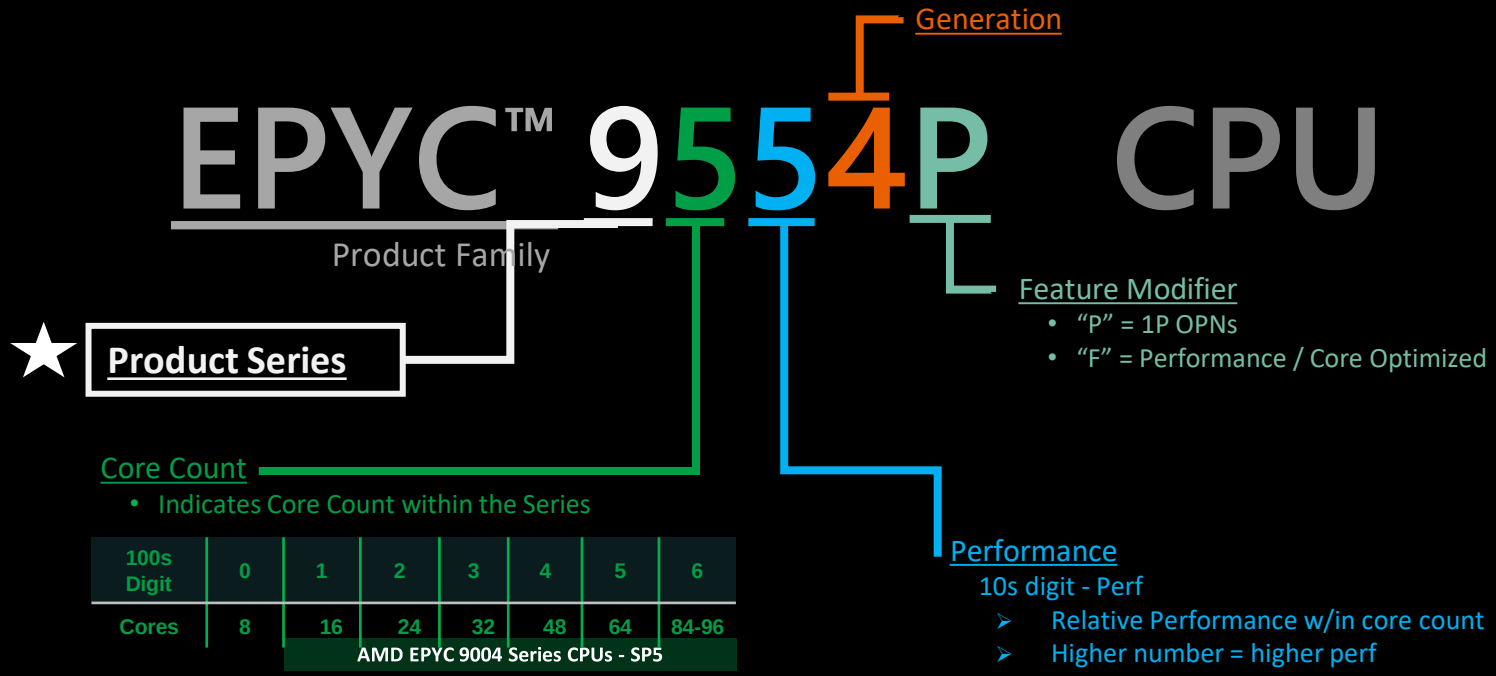
SEV-SNP (Secure Nested Paging), AES-256-XTS with more encrypted VMs

[1] With certain DIMM population rules.

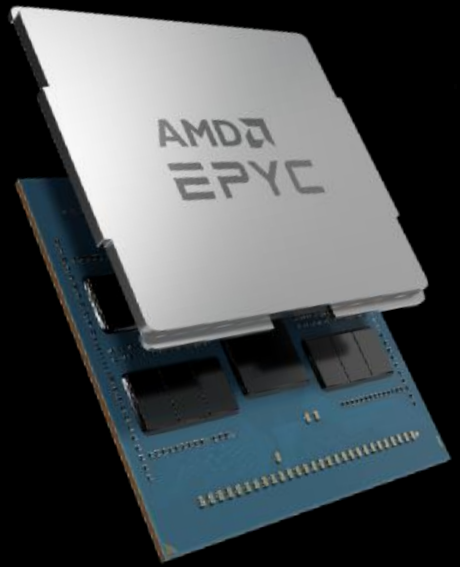
* Post launch feature.

AMD EPYC™ Processor Naming Convention

EPYC 9004 Series CPUs



AMD EPYC™ 9004 CPU MODELS



ALL-IN FEATURE SET INCLUDES

- 12 Channels of DDR5-4800
- 6TB memory capacity
- 128 lanes PCIe®5
- 64 I/O Lanes Support CXL™ 1.1
- 32gbps AMD Gen 3 Infinity Fabric™
- Flexible Topology Options
- Secure Memory Encryption*
- Secure Encrypted Virtualization*

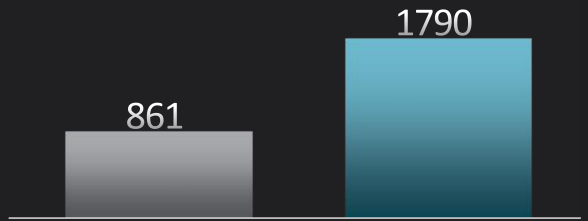
CORES	AMD EPYC	Workload Examples
96 CORES	9654/P	HPC, Cloud, VM Density
84 CORES	9634	Cost sensitive HPC-perf/watt optimized
64 CORES	9554/P	HPC/VM-perf/thread optimized
	9534	HPC/VM-perf/watt optimized
48 CORES	9474F	Technical Computing/Cloud
	9454/P	Storage/Cloud
32 CORES	9374F	Technical Computing
	9354/P	VM-cache/core ration optimized
	9334	VM-per watt optimized
24 CORES	9274F	Technical Computing
	9254	Heart of Enterprise
	9224	Value Enterprise
16 CORES	9174F	Technical Computing
	9124	Value Enterprise

Advancing AMD EPYC™ CPU Leadership

Cloud

SPECrate®2017_int_base

107% higher performance



Integer Throughput

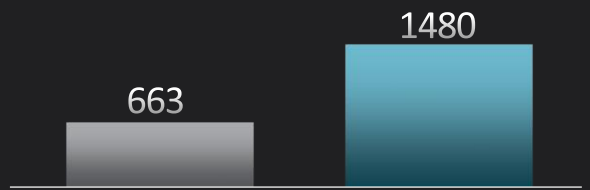
Cloud Service Providers

IaaS/PaaS | Search | Social | SaaS

HPC

SPECrate®2017_fp_base

123% higher performance



Floating-Point Throughput

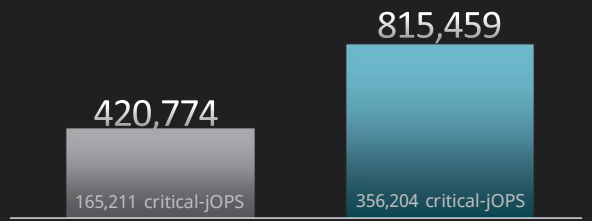
High Performance Computing

Design & Simulation | Research & Academia | Machine Learning | Super Computing

Enterprise

SPECjbb®2015 MultiJVM max-jOPS

94% higher performance



Server-Side Java® Max Throughput

Enterprise IT

Virtualization | SDS/HCI | Hadoop | NoSQL

See endnotes: SP5-001C, -002C, -005C.

HPC PERFORMANCE LEADERSHIP

FASTER TIME-TO-SOLUTIONS

3rd Gen Xeon® Platinum 8380	3rd Gen EPYC™ 7763	4th Gen EPYC™ 9654
40 Cores	64 Cores	96 Cores

4TH GEN AMD EPYC™ CPU

~2.3x

THE EST. PERFORMANCE VS. COMPETITION

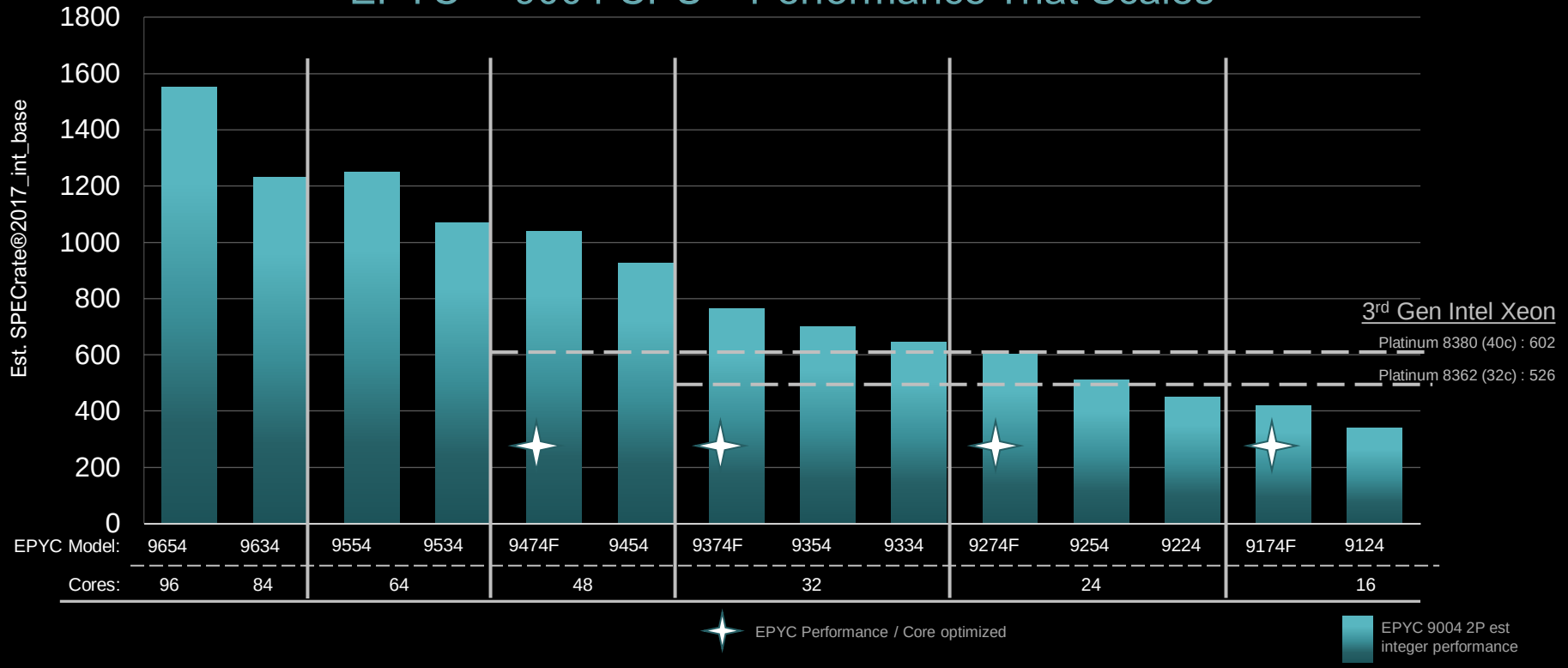
2P FLOATING POINT THROUGHPUT

SPECrate®2017_fp_base

Results may vary. See Endnotes SP5-009A.
OEM published scores will vary based on system configuration and determinism mode used (400W cTDP default performance profile).

EPIC PERFORMANCE TO FIT YOUR NEEDS

EPYC™ 9004 CPU – Performance That Scales



2P 4th Gen AMD EPYC™ CPU estimates. Best performing 2x Intel Xeon Platinum processors published at www.spec.org as of 8/22/22.
3rd Gen Intel Xeon URLs: 2x Xeon Platinum 8380, Platinum 8362.; See endnote SP5-023.

THE POWER OF ONE SOCKET

10,000 SPECrate®2017_int_base

2P INTEL® PLATINUM 8380

2P Intel Platinum 40c 8380
1,360 Cores
212k kWh per Year
602 integer score

1P AMD EPYC™ 9554P

1P AMD EPYC 64c
9554P
1088 Cores
143k kWh per Year
611 est integer score*

1P EPYC SAVINGS

~53% FEWER PROCESSORS

~25% FEWER CORES

~33% LESS POWER

~29% LOWER 3yr TCO[1]

~38 ACRES OF US FOREST ANNUALLY[2]

of equivalent carbon sequestration

Analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 6.10. AMD processor pricing based on 1KU price as of Aug 2022. Intel® Xeon® Scalable CPU data and pricing from https://ark.intel.com as of Jul 2022. All pricing is in USD.

*All AMD EPYC performance scores are estimates based on AMD internal testing, Aug 2022 on AMD reference platforms. All pricing is in USD.

[1] TCO time frame of 3-years and includes estimated costs for real estate, admin and power with power @ $0.12/kWh with 12kW/rack and a 1.7 PUE. Software cost is not included in this analysis.

[2] Values are for USA.

See endnote SP5TCO-005B

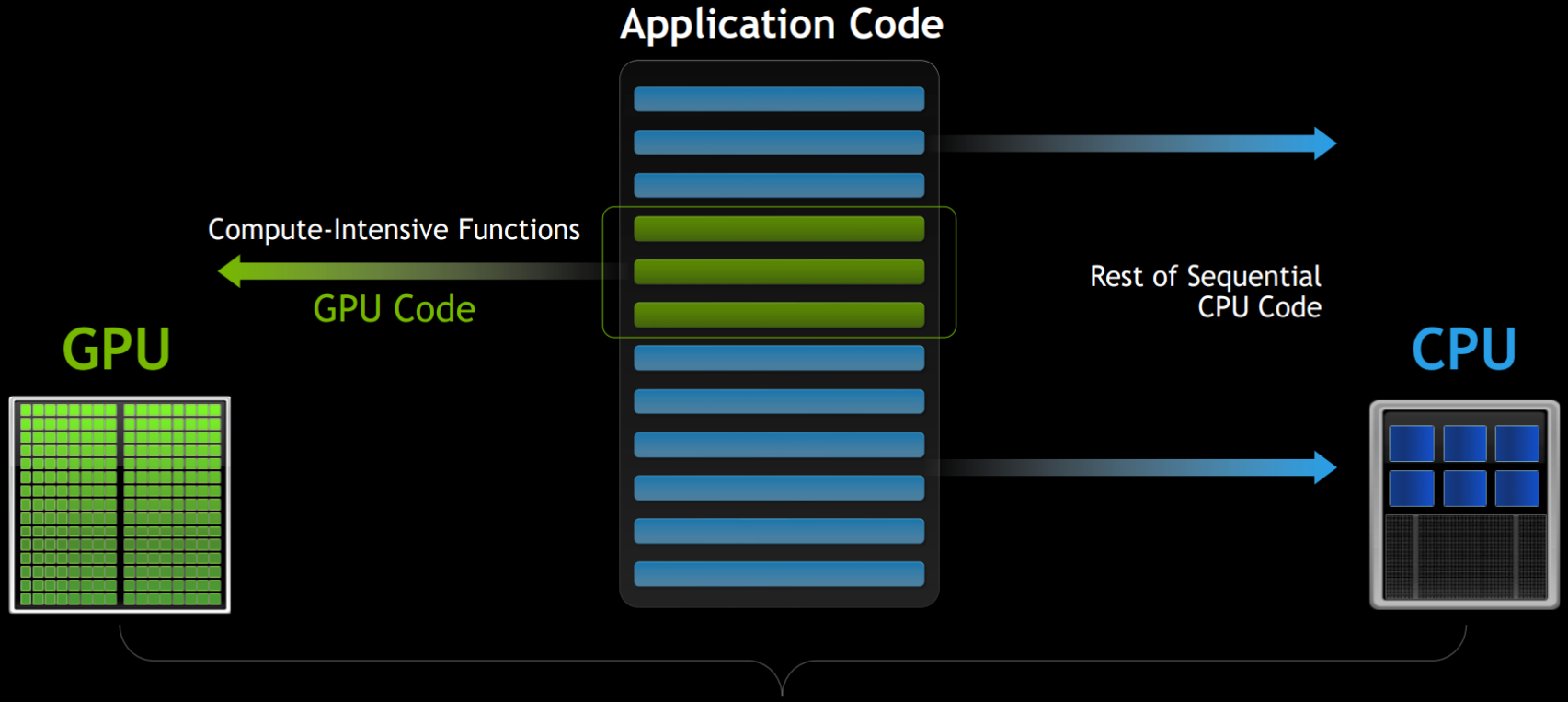
Thank You

Data Center Workloads
EMEA HPC & AI

David Westerik
BDM HPC/AI DCS NVIDIA

DELL Technologies
PARTNER PROGRAM

HOW GPU ACCELERATION WORKS

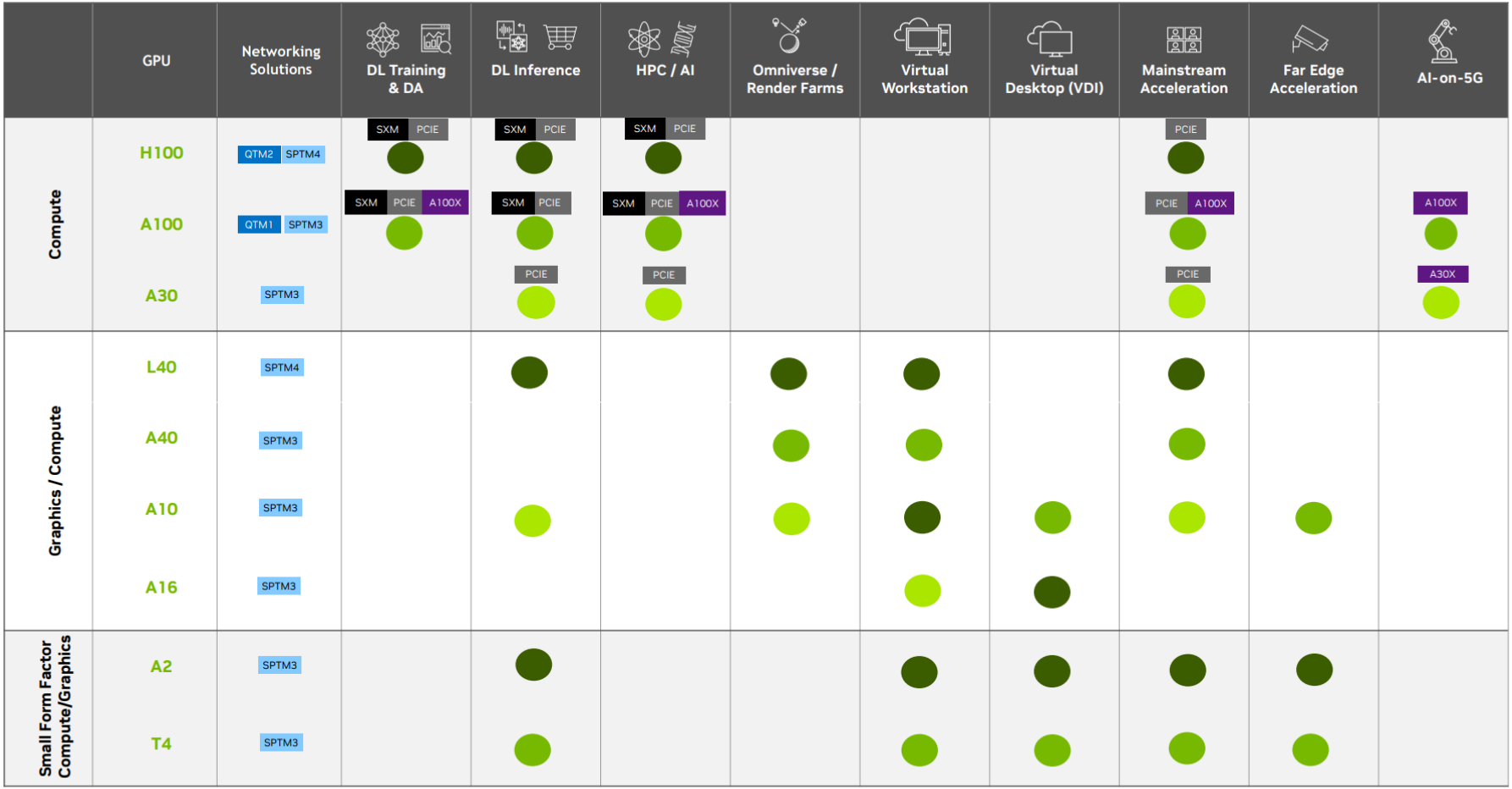


NVIDIA Datacentre grade GPU's

DELL Technologies
PARTNER PROGRAM

Innovation Built-In
intel

NVIDIA Data Center GPU Portfolio



Price-performance comparison in each product group (Compute, Graphics & Compute, SFF Compute & Graphics) and workload column

QTM1 Qauntum-1 IB switch plus BlueField2 DPUs or ConnectX-6/6 DX SmartNICs

QTM2 Qauntum-2 IB switch plus BlueField3 DPUs or ConnectX-7 SmartNICs

SPTM3 Spectrum-3 ethernet switch plus Bluefield2 DPUs or ConnectX-6 /6 Dx SmartNICs

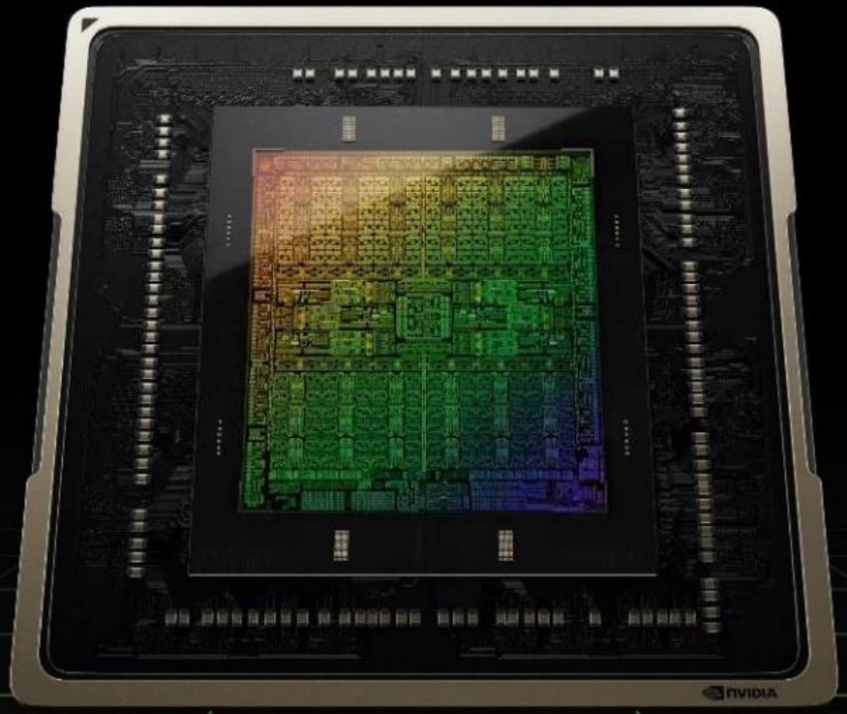
SPTM4 Spectrum-4 ethernet switch plus Bluefield3 DPUs or ConnectX-7 SmartNICs

NVIDIA

NVIDIA ADA LOVELACE

76 Billion Transistors | TSMC 4N Process | Micron G6X Memory

Shaders
New Streaming Multiprocessor
90 Shader TFLOPs
2X Power Efficiency



Ray Tracing
3rd-Gen RT Cores
200 RT TFLOPs
2X Ray-Triangle Intersection

Deep Learning
4th-Gen Tensor Cores
1,400 Tensor TFLOPs
Optical Flow Accelerator

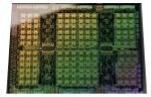
NVIDIA L40

Unprecedented visual computing performance for the data center

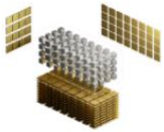
- Next-generation CUDA Cores
- 4th generation Tensor Cores
- 3rd generation RT Cores
- 48 GB GDDR6 GPU Memory with ECC
- 300W
- Secure root of trust

NVIDIA Hopper

The Engine for the World's AI Infrastructure

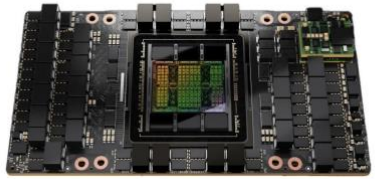


World's Most Advanced Chip



Transformer Engine

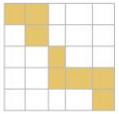
2nd Gen MIG



H100 SXM

Confidential Computing

4th Gen NVLink



DPX Instructions

H100 PCIE

Includes NVIDIA AI Enterprise

NVIDIA AI Enterprise is a 5-Year Subscription

NVIDIA

DATA CENTER & EDGE PRODUCTS TODAY

Expanding Workloads Drive The Need for Specialized Accelerators

Highest Compute Perf AI, HPC, Data Processing	AI Inference & Mainstream Compute	Small Footprint Datacenter & Edge AI	Highest Graphics Perf Visual Computing	AI & Mainstream Graphics and Cumpute	Highest Density Virtual Desktop	High Perf & Mainstream Converged Accelerators
 DL Training	 Language Processing	 Edge AI & Small Inference	 Cloud Rendering	 Cloud Rendering	 Virtual Desktop	 5G vRAN
 Scientific Research	 Conversational AI	 Edge Video	 Cloud XR and vWS	 Cloud XR and vWS	 Virtual Workstation	 AI-on-5G
 Data Analytics	 Recommender Systems	 Mobile Cloud Gaming	 Omniverse	 Virtual Desktop	 Transcoding	 AI-Based Cybersecurity
Fastest Compute, FP64 up to 7 MIG instances	Versatile Mainstream Compute FP64, Up to 4 MIG instances	Entry-level inference Video & Graphics Compact & Versatile	Fastest RT Graphics Largest render models	Rich Graphics, Video Apps and Powerful AI Compact and Flexible	4K Resolution Max # of encode/decode streams	Unified compute/network acceleration with programmable security
A100 300W | 80GB 2-Slot FHFL | Liquid | NVLINK	**A30** 165W | 24GB 2-Slot FHFL | NVLINK	**A2** 40-60W | 16GB 1-Slot Low Profile	**A40** 300W | 48GB 2-Slot FHFL | NVLINK | 3x DP	**A10** 150W | 24GB 1-Slot Full Profile	**A16** 250W | 4x 16 GB 2-Slot FHFL	**A100X | A30X** 300W | 80GB | 230W | 24GB 2-Slot FHFL | NVLINK
Compute (FP32 and FP64)	**Compute** (FP32) **& Graphics**					**GPU + DPU**

NVIDIA Data Center GPU Comparison - Sept '22

	H100		A100		A30	A2	L40	A40	A10	A16
Design	Highest Perf AI, Big NLP, HPC, DA		High Perf Compute		Mainstream Compute	Entry-Level Small Footprint	Powerful Universal Graphics + AI	High Perf Graphics	Mainstream Graphics & Video with AI	High Density Virtual Desktop
Form Factor	SXM5	x16 PCIe Gen5 2 Slot FHFL 3 NVLINK Bridge	SXM4	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x8 PCIe Gen4 1 Slot LP	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x16 PCIe Gen4 1 slot LP	x16 PCIe Gen4 2 Slot FHFL
Max Power	700W	350W	500W	300W	165W	40-60W	300W	300W	150W	250W
FP64 TC | FP32 TFLOPS[2]	67 | 67	51 | 51	19.5 | 19.5		10 |10	NA | 4.5	NA | TBD[3]	NA | 37	NA | 31	NA | 4x4.5
TF32 TC | FP16 TC TFLOPS[2]	989 | 1979	756 | 1513	312 | 624		165 | 330	18 | 36	TBD[3] | TBD[3]	150 | 300	125 | 250	4x18 | 4x36
FP8 TC | INT8 TC TFLOPS/TOPS[2]	3958 | 3958	3026 | 3026	NA | 1248		NA | 661	NA | 72	TBD[3] | TBD[3]	NA | 600	NA | 500	NA | 4x72
GPU Memory / Speed	80GB HBM3	80GB HBM2e	80GB HBM2e		24GB HBM2	16GB GDDR6	48GB GDDR6	48GB GDDR6	24GB GDDR6	4x 16GB GDDR6
Multi-Instance GPU (MIG)	Up to 7		Up to 7		Up to 4	-	-	-	-	-
NVLink Connectivity	Up to 256	2	Up to 8	2	2	-	-	2	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder		1 JPEG Decoder 5 Video Decoder		1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)		4 Video Encoder 8 Video Decoder (+AV1 decode)
Ray Tracing	-		-			Yes	Yes			
Transformer Engine	Yes		-			-		-	-	-
DPX Instructions	Yes		-			-		-	-	-
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)		For in-situ visualization (no NVIDIA vPC or RTX vWS)			Good	Top-of-Line	Best	Better	Good
vGPU	Yes						Yes*	Yes		
Hardware Root of Trust	Internal and External		Internal with Option for External				Internal	Internal with Option for External		
Confidential Computing	Yes		(1)		-	-	-	-	-	-
NVIDIA AI Enterprise	Add-on	Included	Add-on				Add-on			

1. *Supported on Azure NVIDIA A100 with reduced performance compared to A100 without Confidential Computing or H100 with Confidential Computing.*
2. *All Tensor Core numbers with sparsity. Without sparsity is ½ the value.*
3. *Precision TFLOP performance will be added in future update*

NVIDIA

Compute GPU's
H100 / A100 / A30

DELL Technologies
PARTNER PROGRAM

Innovation
Built-In
intel

NVIDIA H100 PCIE

Unprecedented Performance, Scalability, and Security for Mainstream Servers

Available today on R750, R750XA and R7525
Available soon on 16G AMD and Intel Poweredge servers

HIGHEST AI AND HPC MAINSTREAM PERFORMANCE

3.2PF FP8 (5X)| 1.6PF FP16 (2.5X)| 800TF TF32 (2.5X)| 48TF FP64 (2.5X)
6X faster Dynamic Programming with DPX Instructions
2TB/s , 80GB HBM2e memory

HIGHEST COMPUTE ENERGY EFFICIENCY

Configurable TDP - 150W to 350W
2 Slot FHFL mainstream form factor

HIGHEST UTILIZATION EFFICIENCY AND SECURITY

7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

HIGHEST PERFORMING SERVER CONNECTIVITY

128GB/s PCI Gen5
600 GB/s GPU-2-GPU connectivity (5X PCIe Gen5)
up to 2 GPUs with NVLink Bridge

ANNOUNCING HGX-H100

The World's Most Advanced Enterprise AI Infrastructure

HIGHEST PERFORMANCE FOR AI AND HPC

4-way / 8-way H100 GPUs with 32 PetaFLOPs FP8

3.6 TFLOPs FP16 in-network SHARP Compute

NVIDIA Certified High-Performance Offering from All Makers

FASTEST, SCALABLE INTERCONNECT

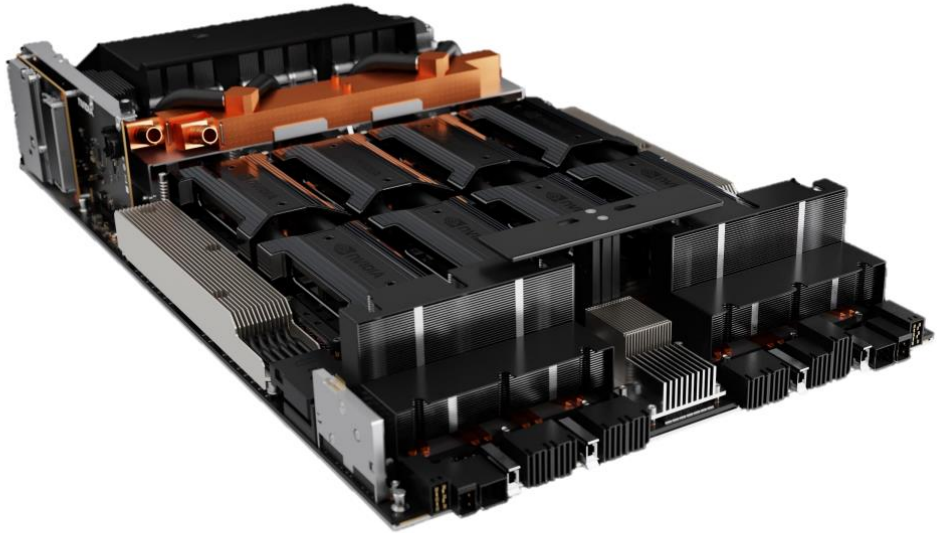
4th Gen NVLINK with 3X faster All-Reduce communications

3.6 TB/s bisection bandwidth

NVLINK Switch System Option Scales Up to 256 GPUs

SECURE COMPUTING

First HGX System with Confidential Computing

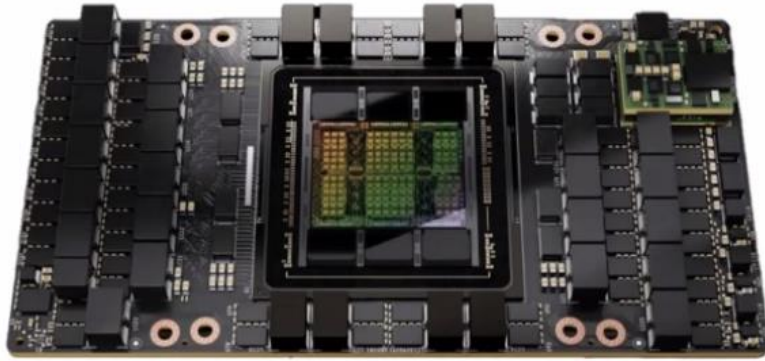


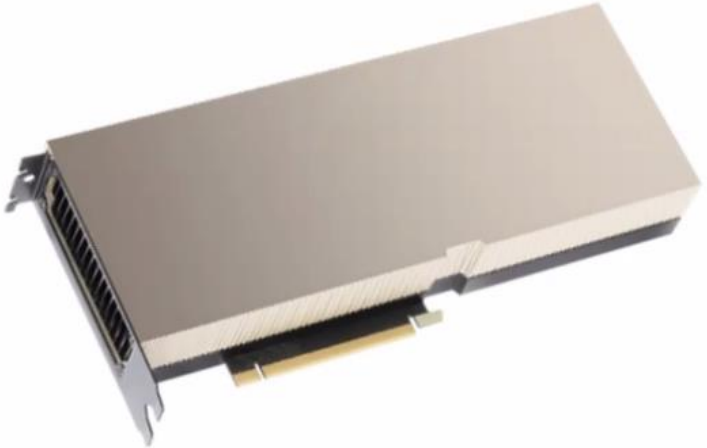
Tensor Core FLOPs shown with sparsity | Speedups compared to prior generation

NVIDIA H100 SXM5 AND PCIE

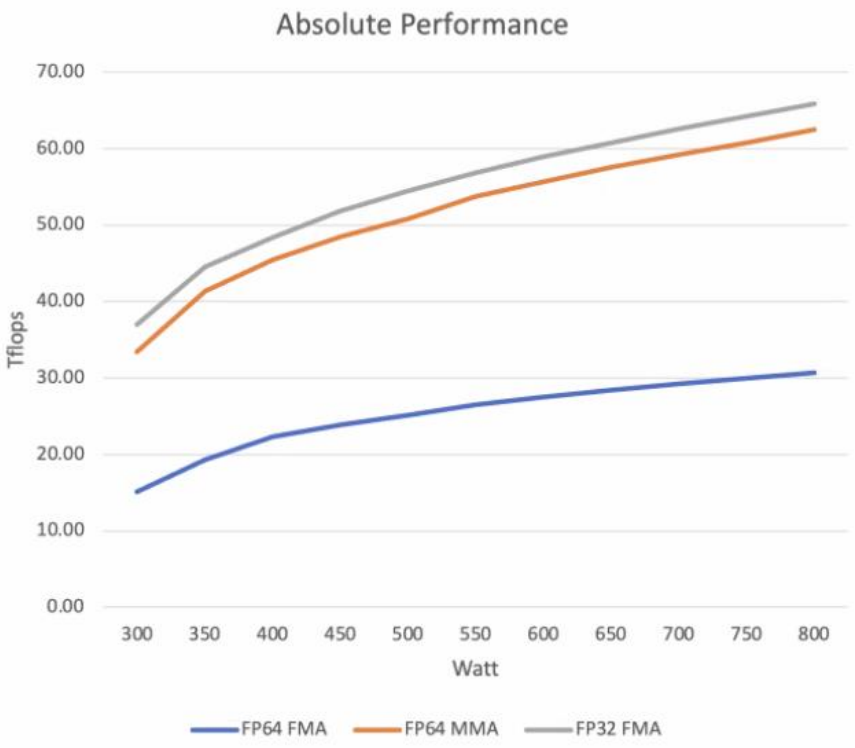
Unprecedented Performance, Scalability, and Security for Every Data Center

	H100 PCIe	H100-80 SXM5	H100-94 SXM5
New Features			
- Dynamic Programming Instructions		Supported	Supported
- Confidential Computing		Supported	Supported
- Transformer Engine with FP8		Supported	Supported
- Peak FP8 Tensor TFLOPS with FP16 Accumulate	1600/3200	2000/4000	2000/4000
- Peak FP8 Tensor TFLOPS with FP32 Accumulate	1600/3200	2000/4000	2000/4000
Compute Performance enhancement			
- Peak FP16 Tensor TFLOPS with FP16 Accumulate	800/1600	1000/2000	1000/2000
- Peak FP16 Tensor TFLOPS with FP32 Accumulate	800/1600	1000/2000	1000/2000
- Peak BF16 Tensor TFLOPS with FP32 Accumulate	800/1600	1000/2000	1000/2000
- Peak TF32 Tensor TFLOPS	400/800	500/1000	500/1000
- Peak FP64 Tensor TFLOPS	48	60	60
- Peak INT8 Tensor TOPS	1600/3200	2000/4000	2000/4000
- Peak FP16 TFLOPS (non-Tensor)	96	120	120
- Peak BF16 TFLOPS (non-Tensor)	96	120	120
- Peak FP32 TFLOPS (non-Tensor)	48	60	60
- Peak FP64 TFLOPS (non-Tensor)	24	30	30
- Peak INT32 TOPS	24	30	30
Memory			
- Memory Interface	5120-bit HBM2e	5120-bit HBM3	6016-bit HBM2e
- Memory Size	80 GB	80 GB	94 GB
- Memory Bandwidth	2000 GB/sec	3000+ GB/sec	2400 GB/sec
L2 Cache Size	50 MB	50 MB	50 MB
TDP	350 Watts	700 Watts	700 Watts





PERFORMANCE AND POWER



H100 and A100 Tensor Core GPUs

H100

PERFORMANCE
The Highest Performance

LLM
&
Transformers

HP
C

SECURITY
The Most Secure

Confidential
Computing

Secured MIG Instances

SCALABILITY
The Most Scalable

4th Gen NVLink
Quantum-2 IB

A100 & Ampere

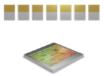
FOR ALL OTHER NEEDS

- Best for CNN models (RESNET-50, R-CNN…)

Non-Transformer
AI

Mainstream
Acceleration

3rd Gen NVLink
Quantum-1 IB



Multi-instance
GPU

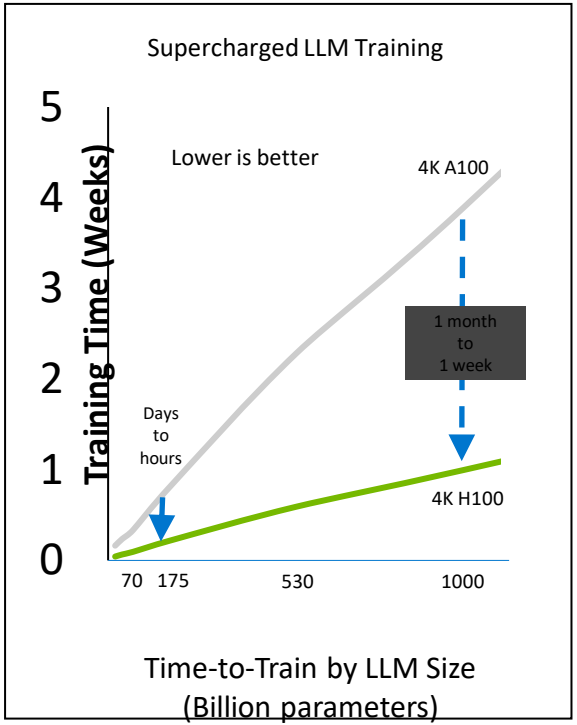
Includes NVIDIA AI Enterprise Software Suite and Support

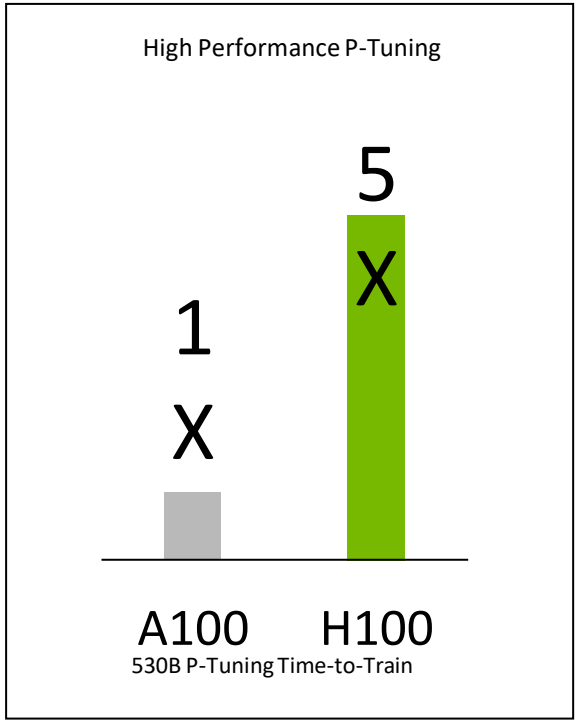
5-Year Subscription and Support included for H100 PCIe

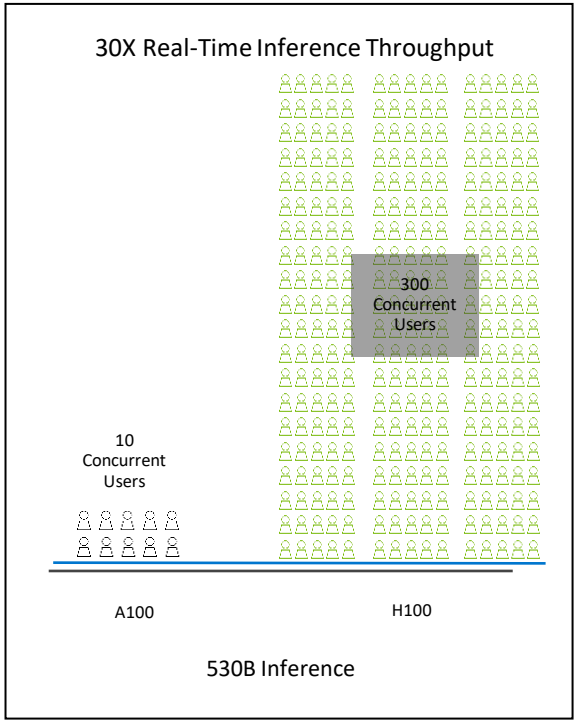
NVIDIA AI Enterprise Sold Separately

DELL Technologies

NVIDIA H100 Supercharges LLMs



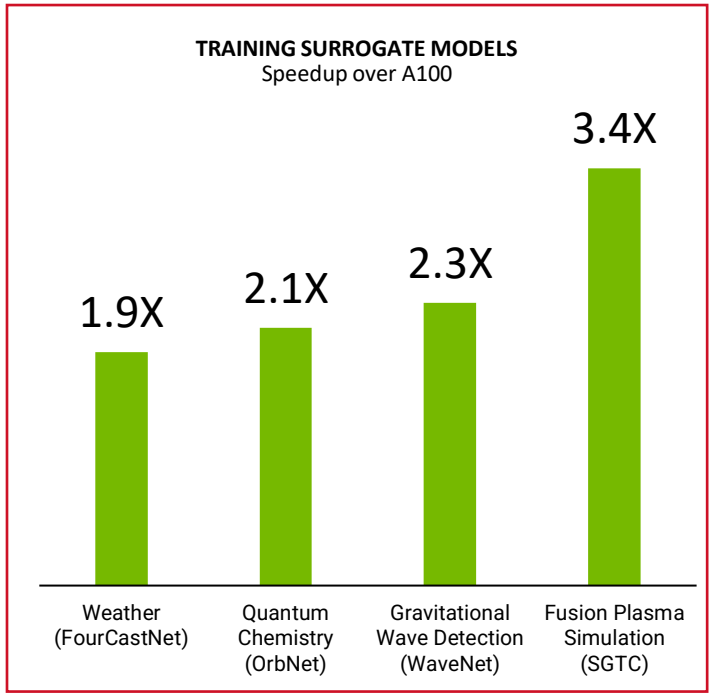


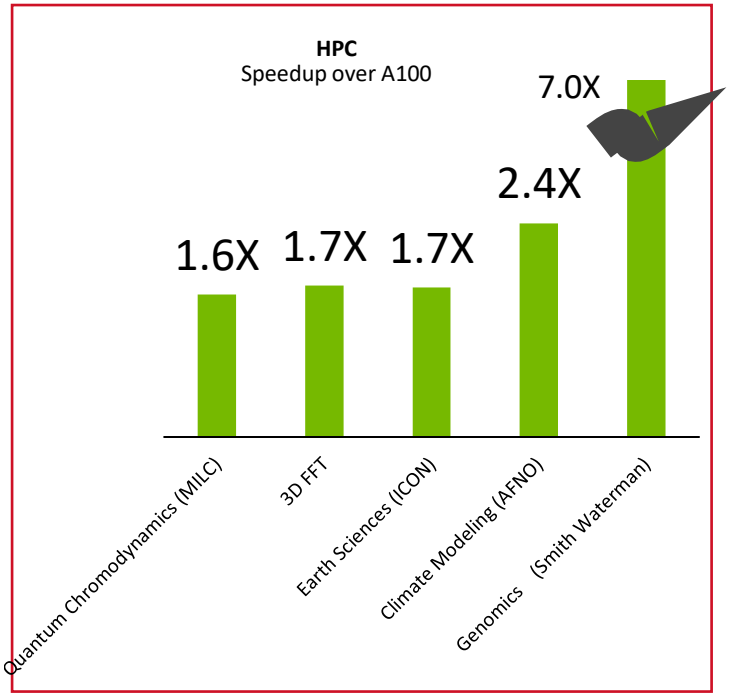


LLM Training | 4096 GPUs | H100 NDR IB | A100 HDR IB | 300 Billion tokens. P-Tuning | DGX H100 | DGX A100 | 530B Q&A tuning using SQuAD dataset
Inference | chatbot| 10 DGX H100 NDR IB | 10 DGX A100 HDR IB | <1 sec latency | 1 inference/second/user.
H100 data center projected workload performance, subject to change

H100 Performance Training

Performance and scalability for the next generation of AI and HPC breakthroughs





Projected performance subject to change. ESTIMATES ONLY | HPC Applications: Climate Modelling 1K, LQCD 1K, Genomics 8, 3D-FFT 256, MT-NLG 32 (batch sizes: 4 for A100, 60 for H100 at 1sec, 8 for A100 and 64 for H100 at 1.5 and 2sec), MRCNN 8 (batch 32), GPT-3 16B 512 (batch 256), Training Surrogate AI Models: FourCastNet running 30 sq km spatial resolution model, Wavenet training using Keras SavedModel , Orbnet training using nsys timeline for H<->D transfer, SGTC model was extracted from the PyTorch source, trained an ensemble of 10 models Comparison to DGX-A100 (4x A100 SXM4 80 GB) | HPC Application ICON v2.6.5 running dataset QUBICC r2b6 40km global resolution

NVIDIA A30

Versatile Compute Acceleration for Mainstream Enterprise Servers

Purpose built for Inference and Flexible Enterprise Compute
- 4X T4 Delivered Application perf

Multi-Instance GPU
- Up to 4 concurrent instances per GPU (QoS)

Compute
- 3rd Gen Tensor cores, Fast FP64

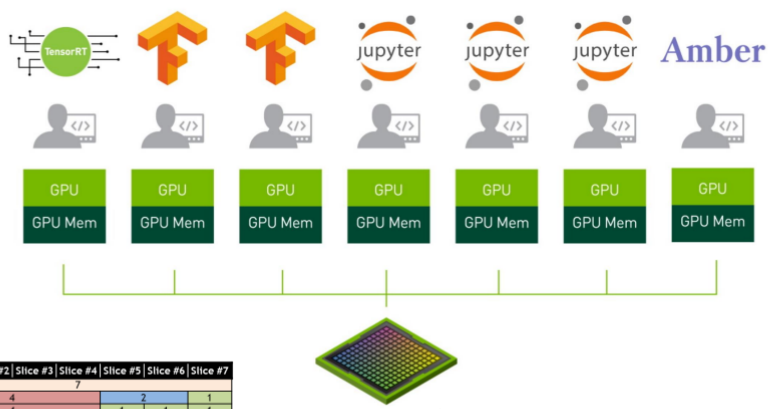
High Bandwidth Memory
- Ultra-low latency

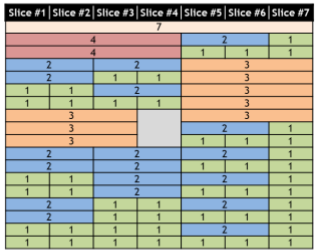
Power Efficient
- Excellent Perf/W

Sparsity Acceleration
- Further 2X speed up

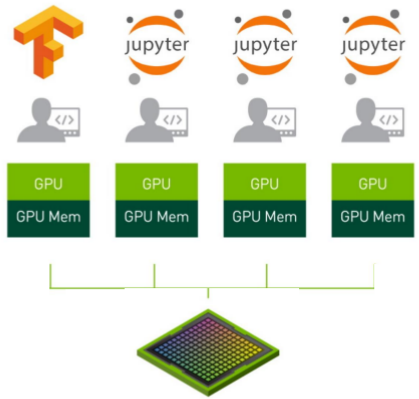
Multi Instance GPU - MIG

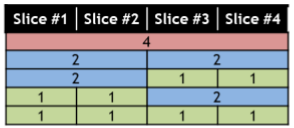
NVIDIA A100 & A30





NVIDIA A100
(40 GB & 80GB)





NVIDIA A30
(24 GB)

Graphics/Compute GPU’s

NVIDIA L40

Unprecedented visual computing performance for the data center

- Next-generation CUDA Cores
- 4th generation Tensor Cores
- 3rd generation RT Cores
- 48 GB GDDR6 GPU Memory with ECC
- 300W
- Secure root of trust

NVIDIA A40

World's most powerful data center GPU for visual computing

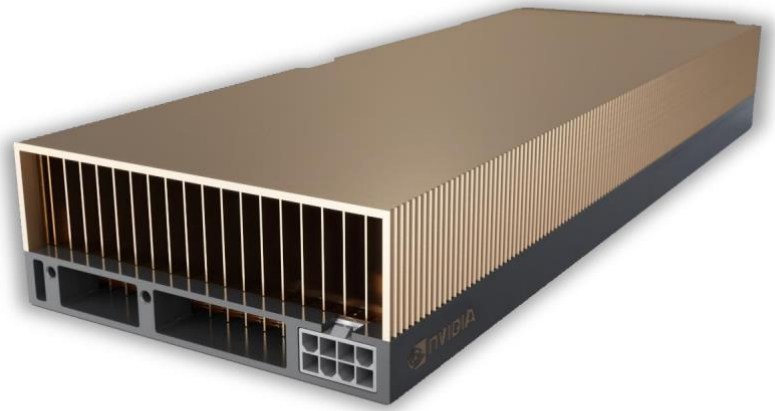
NVIDIA Ampere Architecture CUDA Cores
Up to 2X FP32 throughput of previous generation*

2nd Generation RT Cores
Up to 2X throughput of previous generation*

3rd Generation Tensor Cores
Up to 5X throughput with TF32*

48 GB GDDR6 Memory
Largest frame buffer for professional graphics

PCIe Gen 4
2X bandwidth of PCIe Gen 3



- 3x Display Port 1.4 outputs**
- 2-way NVLink
- Quadro Sync support
- vGPU software support
- Hardware secure boot

NVIDIA

*Performance measures gen to gen comparison of RTX 6000 to NVIDIA A40

** A40 is configured for virtualization by default with physical display connectors disabled. The display outputs can be enabled via management software tools.

NVIDIA A10

High-performance graphics & video with AI

NVIDIA Ampere architecture
2nd gen RT Cores, 3rd gen Tensor Cores

24GB GDDR6 Memory
1.5X memory versus previous generation*

Improved Performance
Up to 2.5X faster graphics and inferencing*

High-density, Power Efficient
Single-slot form factor, 150W

Media Acceleration
AV1 Decode, multiple 4K streams, 8K HDR

Flexibly accelerate multiple data center workloads
Deploy virtual workstations & desktops or AI inference

NVIDIA

*Gen to gen comparison of NVIDIA T4 to NVIDIA A10

NVIDIA A16

Unprecedented user experience and density for graphics-rich VDI

Purpose-built for high user density
2X density versus previous generation[1]

Lowest Cost per virtual workstation user
Affordable entry virtual workstations[2]

4x 16GB GDDR6 Memory
Up to 64 multimedia-rich virtual desktops per board
Larger framebuffer per user for entry CAD virtual workstations[3]

Flexibility of heterogenous users
Simultaneously host different user profiles on one board

Highest Quality Video
Supports H.265 encode/decode, VP9, and AV1 decode

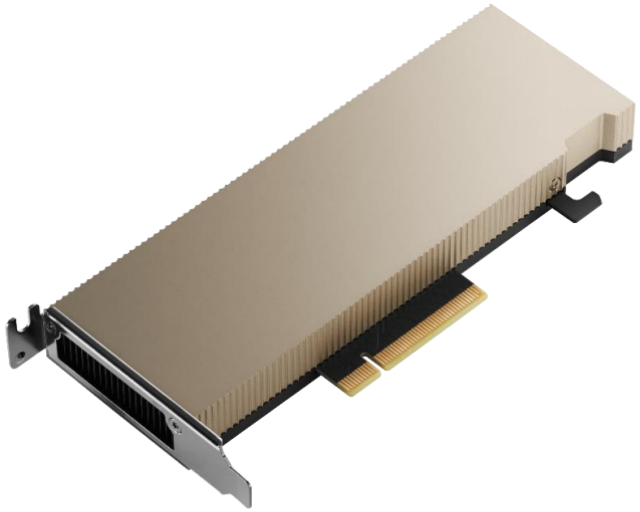
Multiuser performance for streaming video & multimedia
More than 2X encoder throughput[1]

Latest NVIDIA Ampere architecture
2nd gen RT cores, 3rd gen Tensor Cores

1. *Gen to gen comparison of NVIDIA M10 to NVIDIA A16*
2. *Comparison of NVIDIA A16 vs. T4, RTX 6000, RTX 8000, and A40*
3. *Gen to gen comparison of NVIDIA T4 to NVIDIA A16*

NVIDIA A2

Entry-level GPU bringing NVIDIA AI to any server



Compact, Entry-Level Inference

- Single slot LP, lower power – fits any server
- Optimal for thermally constrained systems

Latest Ampere Architecture Features

- 3rd gen Tensor cores, 2nd gen RT cores, Secure RoT

Higher Intelligent Video Analytics (IVA) Performance

- 1.3X better performance vs T4

Up to 20X Higher Performance versus CPU

- Speedups for AI inference and cloud gaming

3 Reasons to Transition from Previous Gen to A2

Superior value and lower power on NVIDIA Ampere architecture

Best Perf/$ For Compact Edge

Superior ROI

Up To 40% Lower Power

TDP: A2 (40-60W) vs T4 (70W)

30% Higher Iva Performance

A2 Superior Video Decode

NVIDIA Data Center GPU Comparison - Sept '22

	H100		A100		A30	A2	L40	A40	A10	A16
Design	Highest Perf AI, Big NLP, HPC, DA		High Perf Compute		Mainstream Compute	Entry-Level Small Footprint	Powerful Universal Graphics + AI	High Perf Graphics	Mainstream Graphics & Video with AI	High Density Virtual Desktop
Form Factor	SXM5	x16 PCIe Gen5 2 Slot FHFL 3 NVLINK Bridge	SXM4	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x8 PCIe Gen4 1 Slot LP	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x16 PCIe Gen4 1 slot LP	x16 PCIe Gen4 2 Slot FHFL
Max Power	700W	350W	500W	300W	165W	40-60W	300W	300W	150W	250W
FP64 TC | FP32 TFLOPS[2]	67 | 67	51 | 51	19.5 | 19.5		10 |10	NA | 4.5	NA | TBD[3]	NA | 37	NA | 31	NA | 4x4.5
TF32 TC | FP16 TC TFLOPS[2]	989 | 1979	756 | 1513	312 | 624		165 | 330	18 | 36	TBD[3] | TBD[3]	150 | 300	125 | 250	4x18 | 4x36
FP8 TC | INT8 TC TFLOPS/TOPS[2]	3958 | 3958	3026 | 3026	NA | 1248		NA | 661	NA | 72	TBD[3] | TBD[3]	NA | 600	NA | 500	NA | 4x72
GPU Memory / Speed	80GB HBM3	80GB HBM2e	80GB HBM2e		24GB HBM2	16GB GDDR6	48GB GDDR6	48GB GDDR6	24GB GDDR6	4x 16GB GDDR6
Multi-Instance GPU (MIG)	Up to 7		Up to 7		Up to 4	-	-	-	-	-
NVLink Connectivity	Up to 256	2	Up to 8	2	2	-	-	2	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder		1 JPEG Decoder 5 Video Decoder		1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)		4 Video Encoder 8 Video Decoder (+AV1 decode)
Ray Tracing	-		-			Yes	Yes			
Transformer Engine	Yes		-			-		-	-	-
DPX Instructions	Yes		-			-		-	-	-
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)		For in-situ visualization (no NVIDIA vPC or RTX vWS)			Good	Top-of-Line	Best	Better	Good
vGPU	Yes						Yes*	Yes		
Hardware Root of Trust	Internal and External		Internal with Option for External				Internal	Internal with Option for External		
Confidential Computing	Yes		(1)		-	-	-	-	-	-
NVIDIA AI Enterprise	Add-on	Included	Add-on				Add-on			

1. *Supported on Azure NVIDIA A100 with reduced performance compared to A100 without Confidential Computing or H100 with Confidential Computing.*
2. *All Tensor Core numbers with sparsity. Without sparsity is ½ the value.*
3. *Precision TFLOP performance will be added in future update*

NVIDIA

Backup & Use Cases

ACCELERATING TIME TO TREATMENT

Contouring is a time-consuming but important task in radiation therapy. It helps radiologists determine how much radiation to use to treat tumors without damaging surrounding healthy organs.

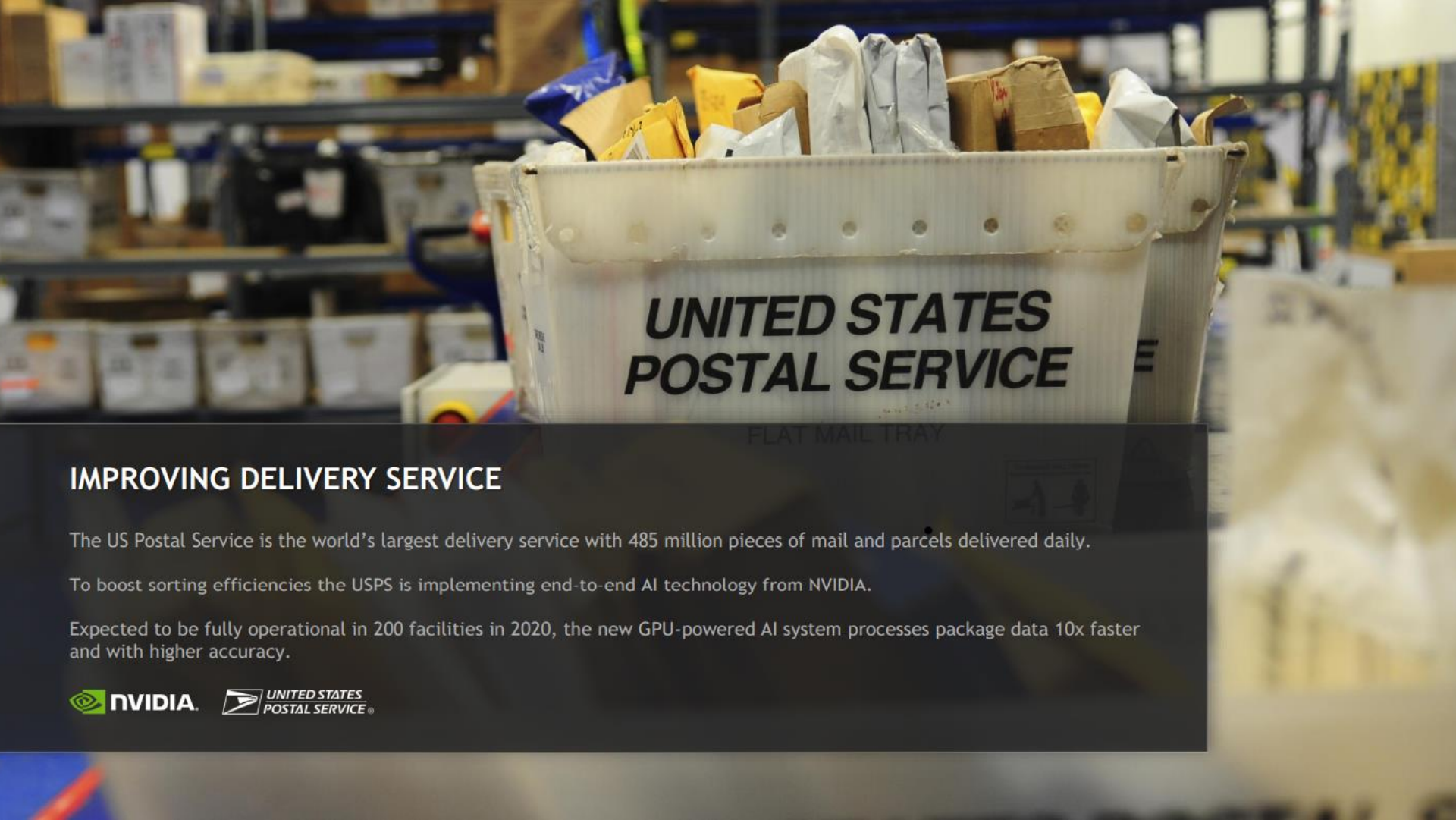
Siemens Healthineers uses *syngo.via* RT Image Suite to automatically outline organs using AI-assisted AutoContouring. Trained on >4.5M images on the GPU-powered Sherlock supercomputer, the AI model saves radiation oncologist time and accelerates radiation treatment for patients.

PAVING THE WAY TO FUSION ENERGY

Fusion is the future of energy, but today's reactors cannot run for more than a few seconds without a disruption — which is not long enough to produce a net energy gain. Physicists need better tools to monitor and understand the plasma physics.

Scientists at Técnico Lisboa are using NVIDIA GPU-powered AI to visualize several phenomena —such as plasma heating and disruptions— inside the reactor and can predict time-to-disruption and its probability with 85% accuracy.

NVIDIA. TÉCNICO LISBOA EUROfusion ipfn



IMPROVING DELIVERY SERVICE

The US Postal Service is the world's largest delivery service with 485 million pieces of mail and parcels delivered daily.

To boost sorting efficiencies the USPS is implementing end-to-end AI technology from NVIDIA.

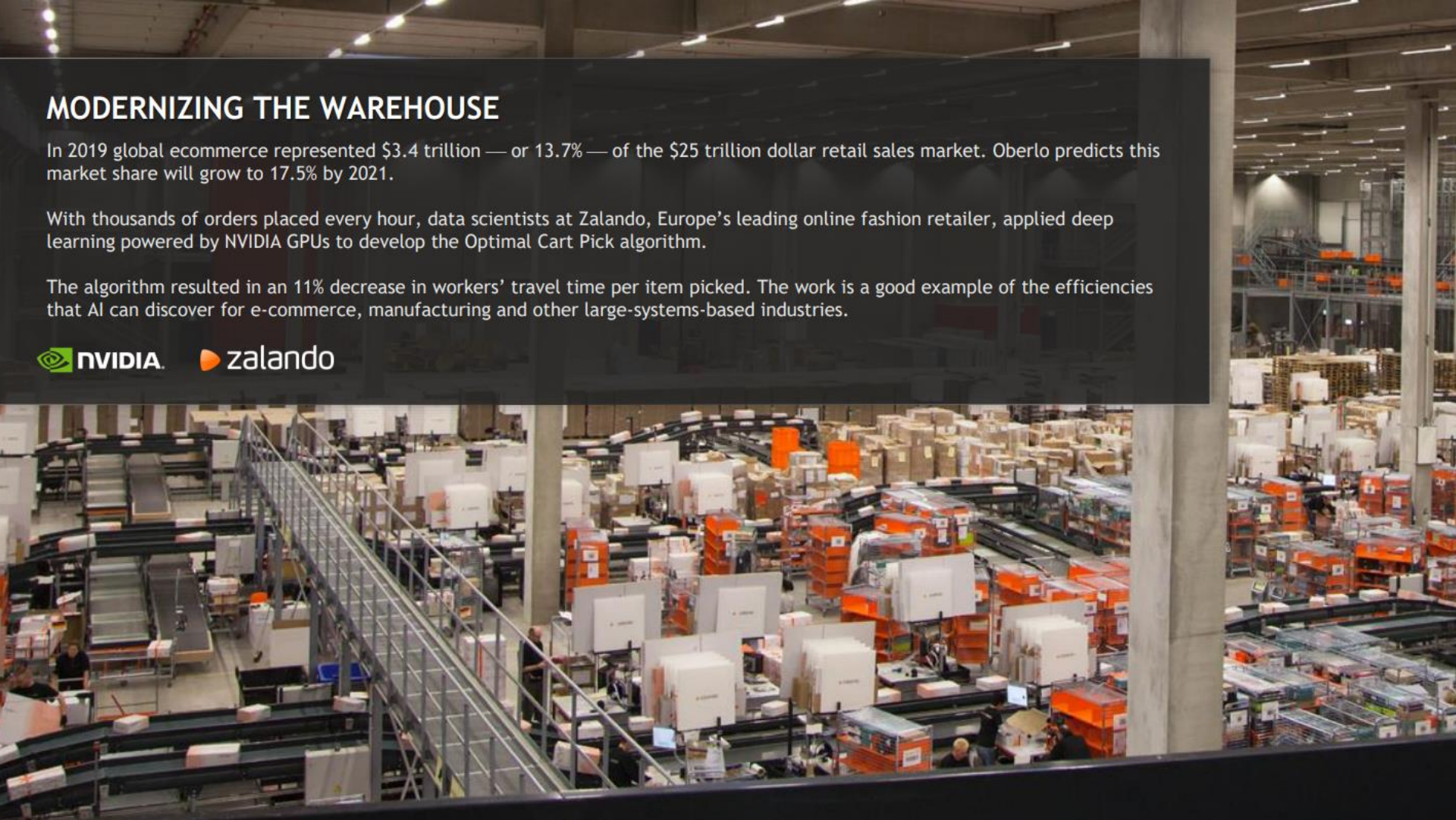
Expected to be fully operational in 200 facilities in 2020, the new GPU-powered AI system processes package data 10x faster and with higher accuracy.

NVIDIA. UNITED STATES POSTAL SERVICE®

THE BRAINS BEHIND SMART CITIES

Verizon's Smart Communities Group is on a mission to make cities safer, smarter and greener. Using NVIDIA Metropolis, an edge-to-cloud video platform for building smarter, faster AI-powered applications, Verizon is working to collect and analyze multiple streams of video data to improve traffic flow, enhance pedestrian safety, optimize parking and more.

NVIDIA. verizon



MODERNIZING THE WAREHOUSE

In 2019 global ecommerce represented $3.4 trillion — or 13.7% — of the $25 trillion dollar retail sales market. Oberlo predicts this market share will grow to 17.5% by 2021.

With thousands of orders placed every hour, data scientists at Zalando, Europe's leading online fashion retailer, applied deep learning powered by NVIDIA GPUs to develop the Optimal Cart Pick algorithm.

The algorithm resulted in an 11% decrease in workers' travel time per item picked. The work is a good example of the efficiencies that AI can discover for e-commerce, manufacturing and other large-systems-based industries.

NVIDIA zalando

Backup Slides AMD