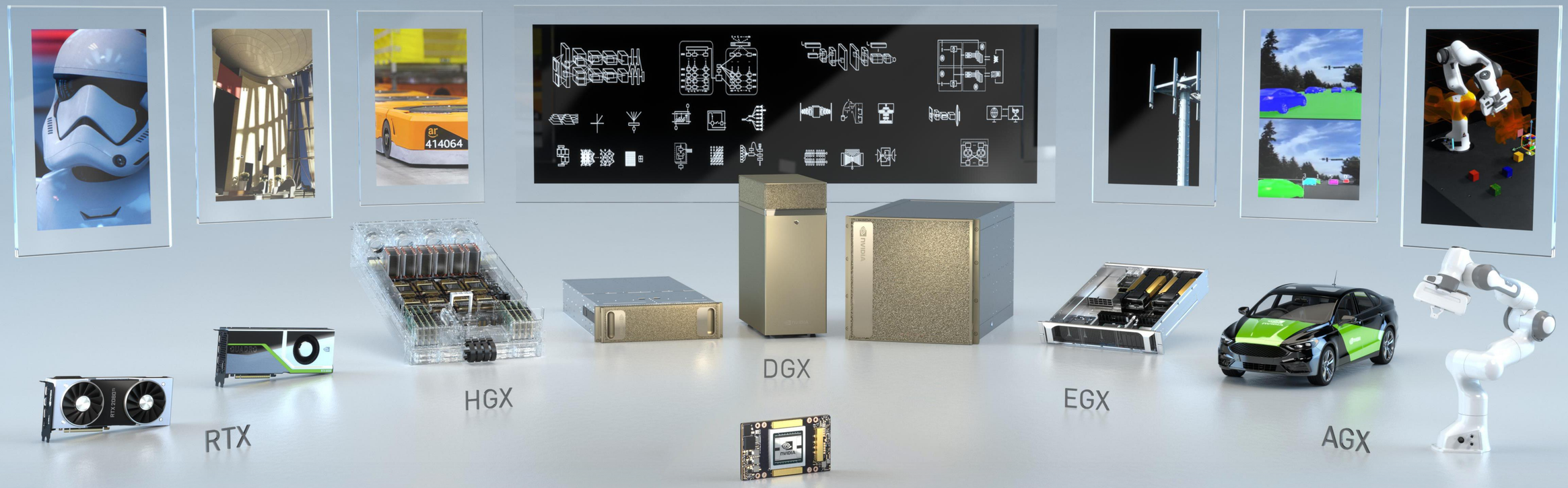
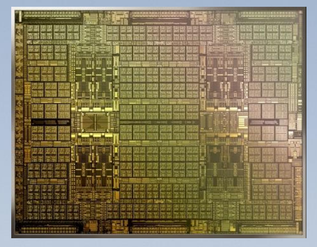
NVIDIA DATA CENTER STRATEGY
FOR ACCELERATED COMPUTING PLATFORM

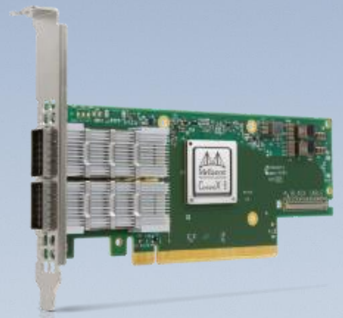
정소영 상무
Solutions Architect, Lead

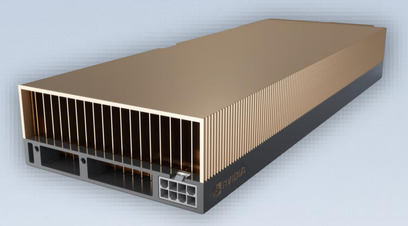
COMPUTING FOR THE DA VINCIS OF OUR TIME

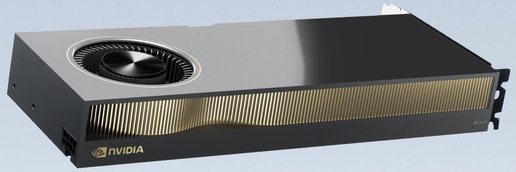


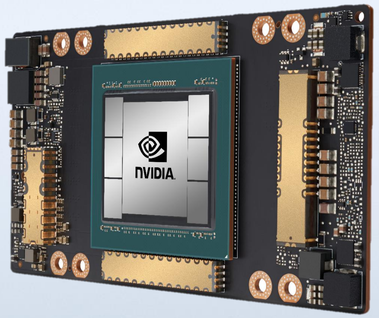
NVIDIA DATA CENTER BUSINESS











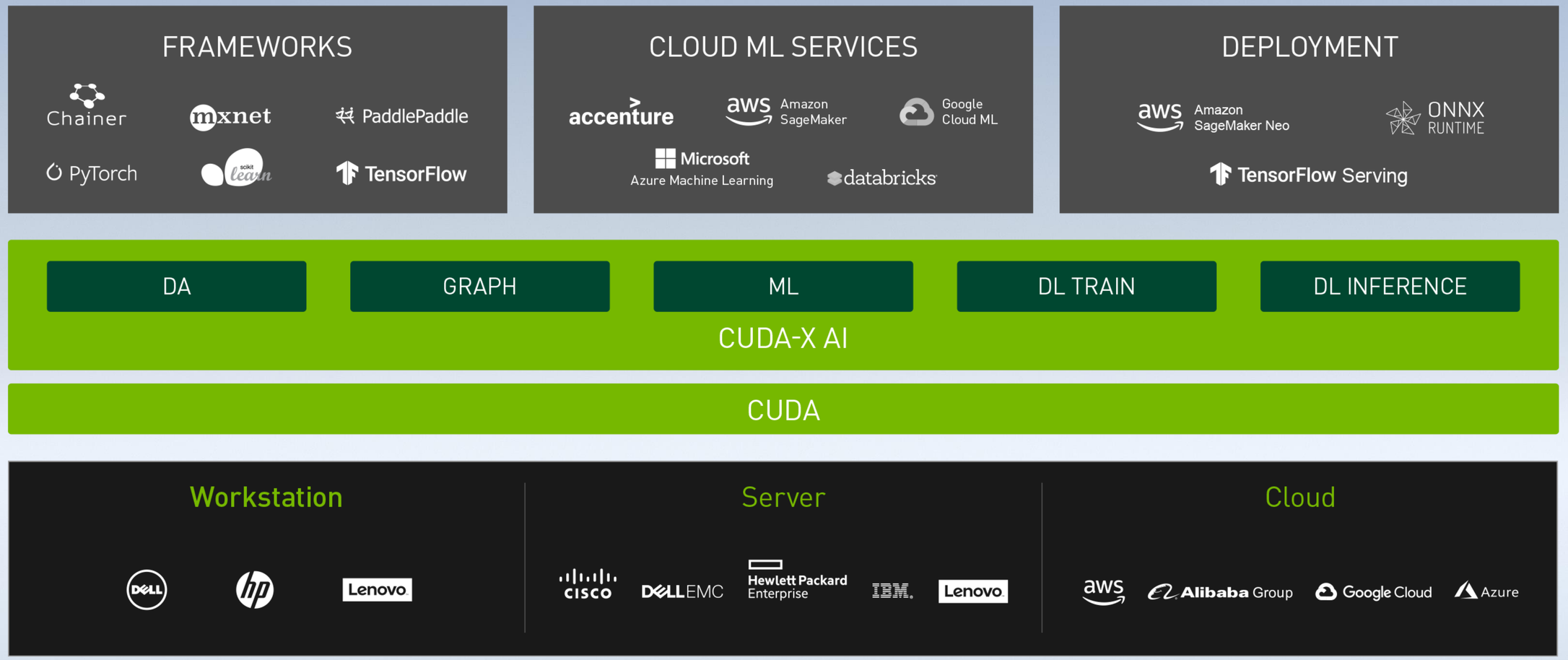
CUMULUS

SwiftStack

NVIDIA + arm

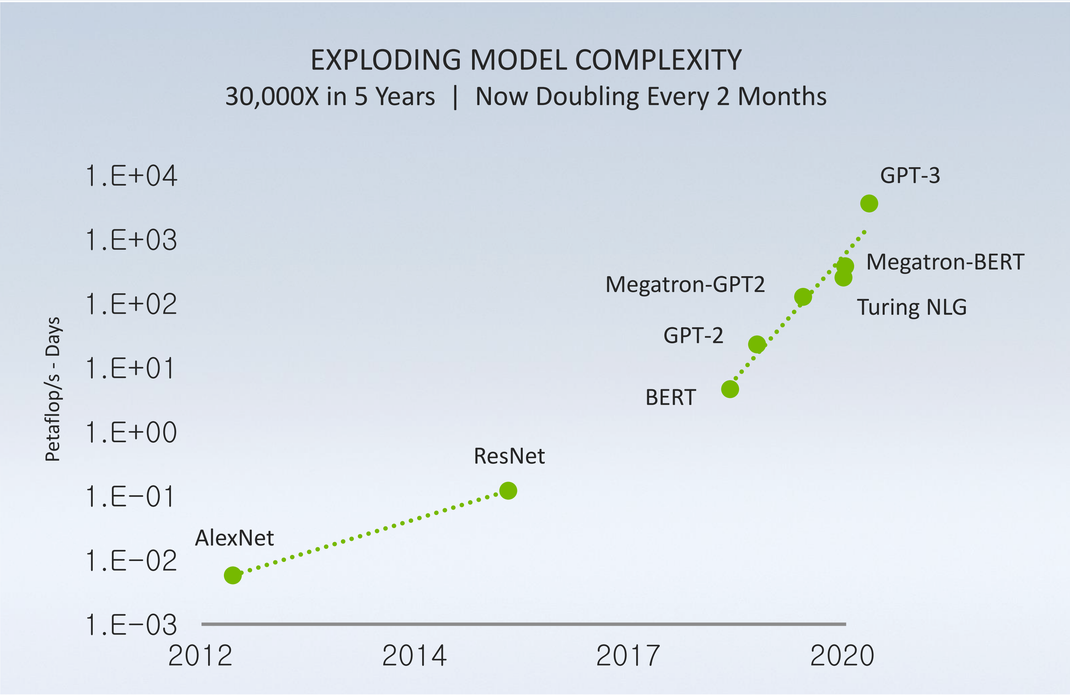
ACCELERATED COMPUTING PLATFORM COMPANY

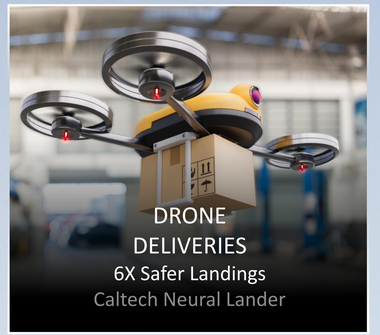
NVIDIA END-TO-END FULL STACK FOR ACCELERATED COMPUTING PLATFORM

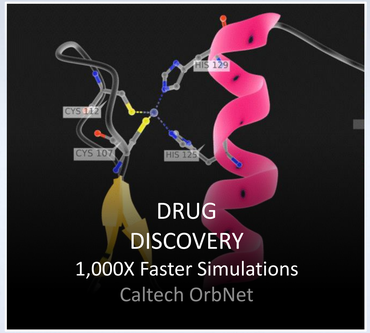


WHY DATA CENTER IS IMPORTANT?

ARTIFICIAL INTELLIGENCE —
THE MOST POWERFUL TECHNOLOGY FORCE OF OUR TIME

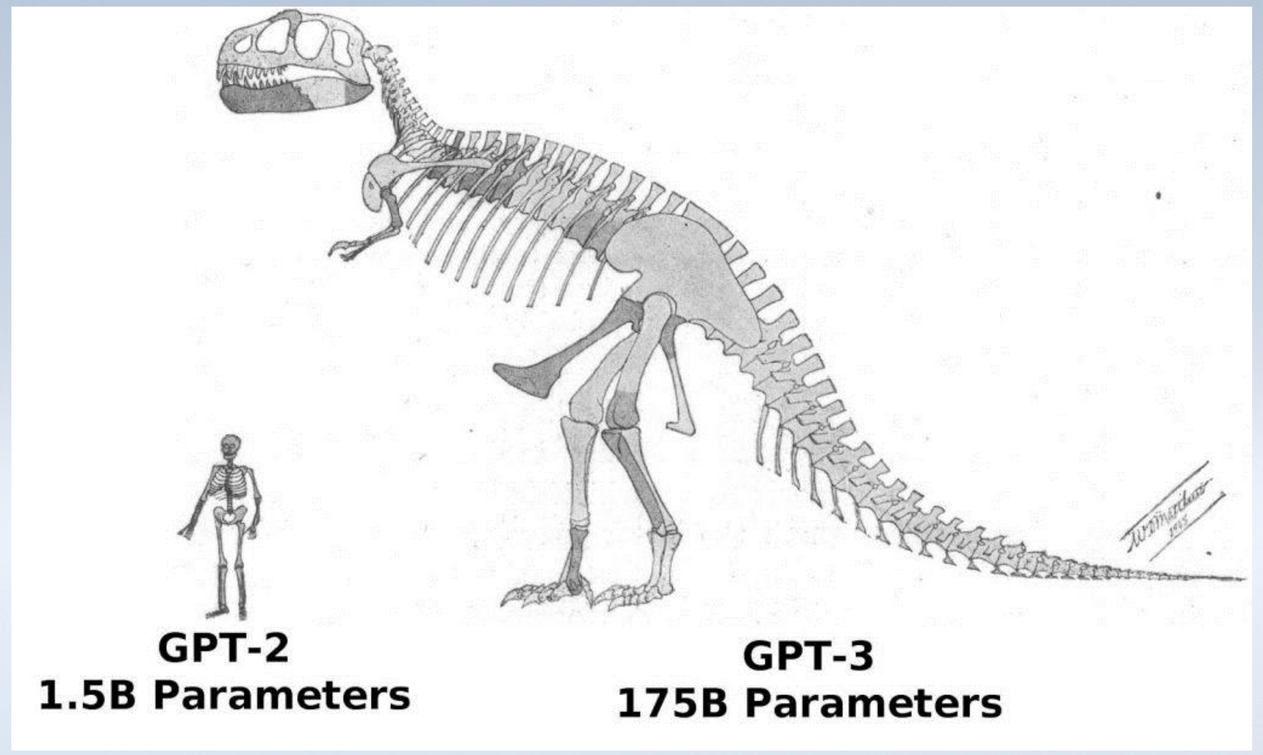






EXAMPLE – GPT-3

MORE THAN X100 BIGGER THAN GPT-2 IN 6 MONTHS



State-of-the-art Generative NLP Model by OpenAI

X1000 Bigger than BERT-base: Max. 175B parameters

300 ~ 450B tokens used for model training

Large-scale model parallel + data parallel for Training and Inference required

Image from https://blog.exxactcorp.com/what-can-you-do-with-the-openai-gpt-3-language-model/

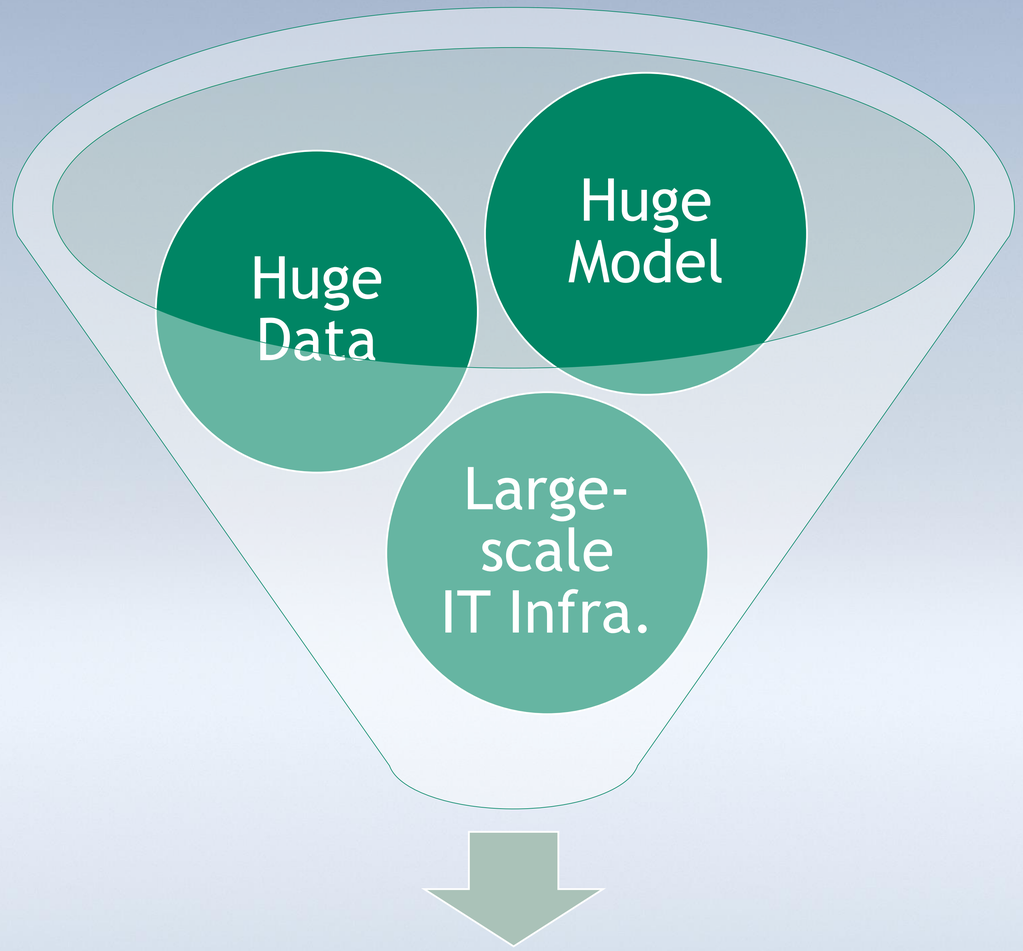
TRAINING TIME PROJECTION

V100 / A100 with Mellanox IB Interconnect Technology

# of GPUS with IB	100	200	500	1,000	2,000	5,000	**10,000**
Training days (V100)	1730.6	865.3	346.1	173.1	86.5	34.6	**17.3**
Training days (A100)	641	320.5	128.2	64.1	32	12.8	**6.4**

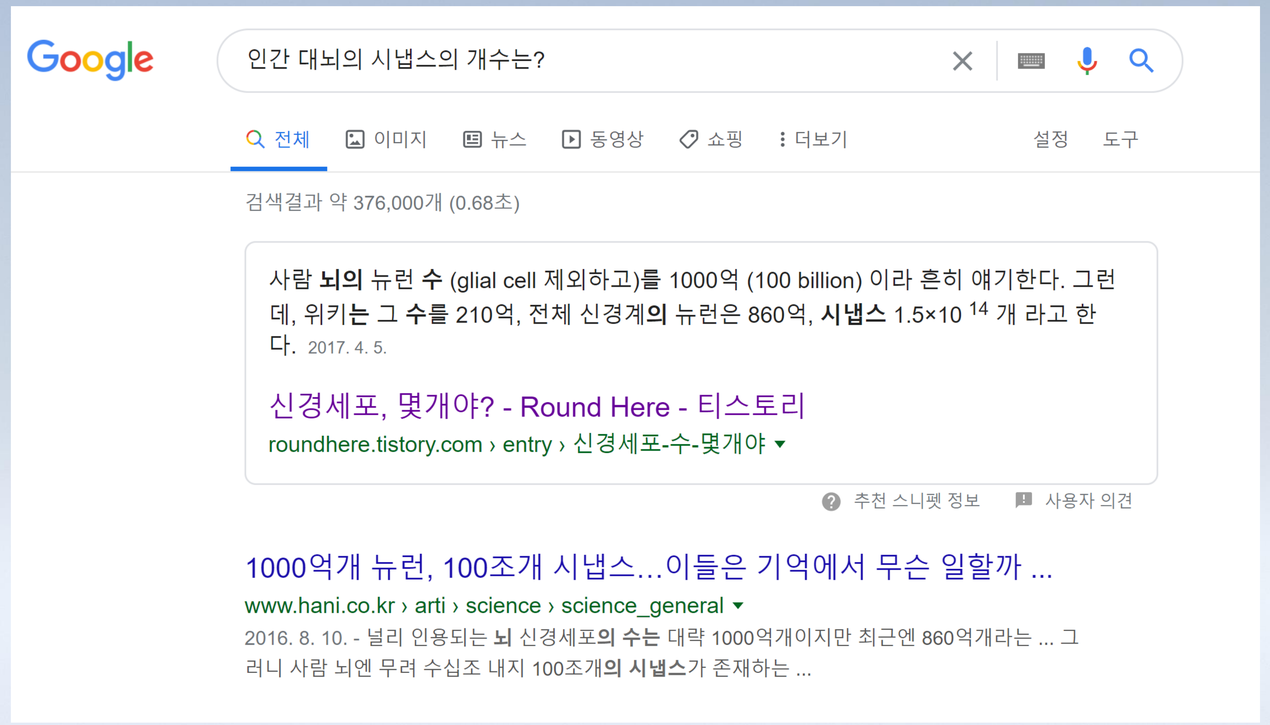
Performance projection done by NVIDIA based on the performance gap of V100 and A100 for Transformer decoder

WHAT IT MEANS…



AGI (Artificial General Intelligence)

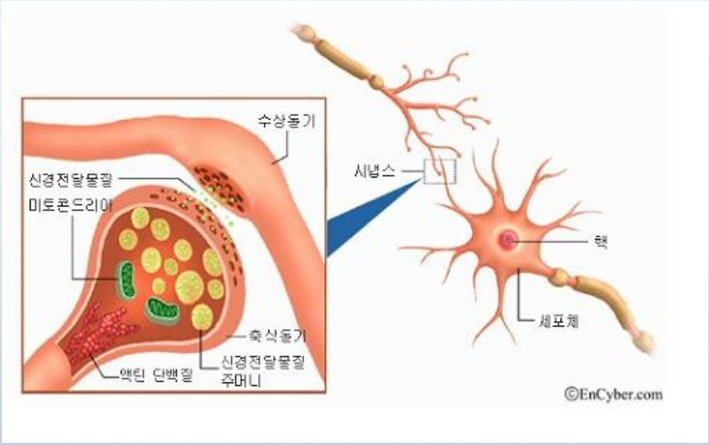
HOW BIG WILL BE THE MODEL IN THE FUTURE?



시냅스 (뉴런과 뉴런의 연결 고리) - Deep learning의 파라메터와 유사한 개념

대뇌의 시냅스 개수는 약 100 ~ 150조개로 추정

현재 GPT-3 보다 약 1000배 복잡한 모델이면?

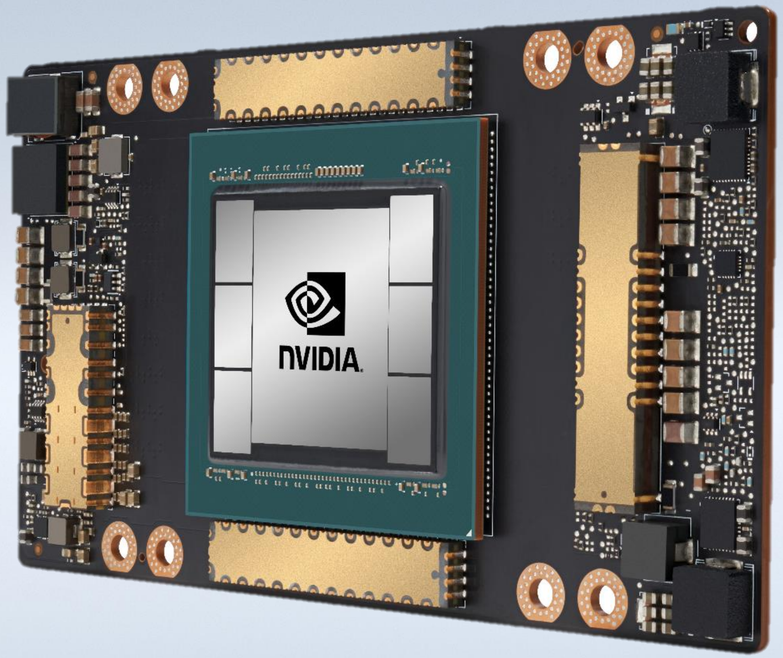


NVIDIA GPU

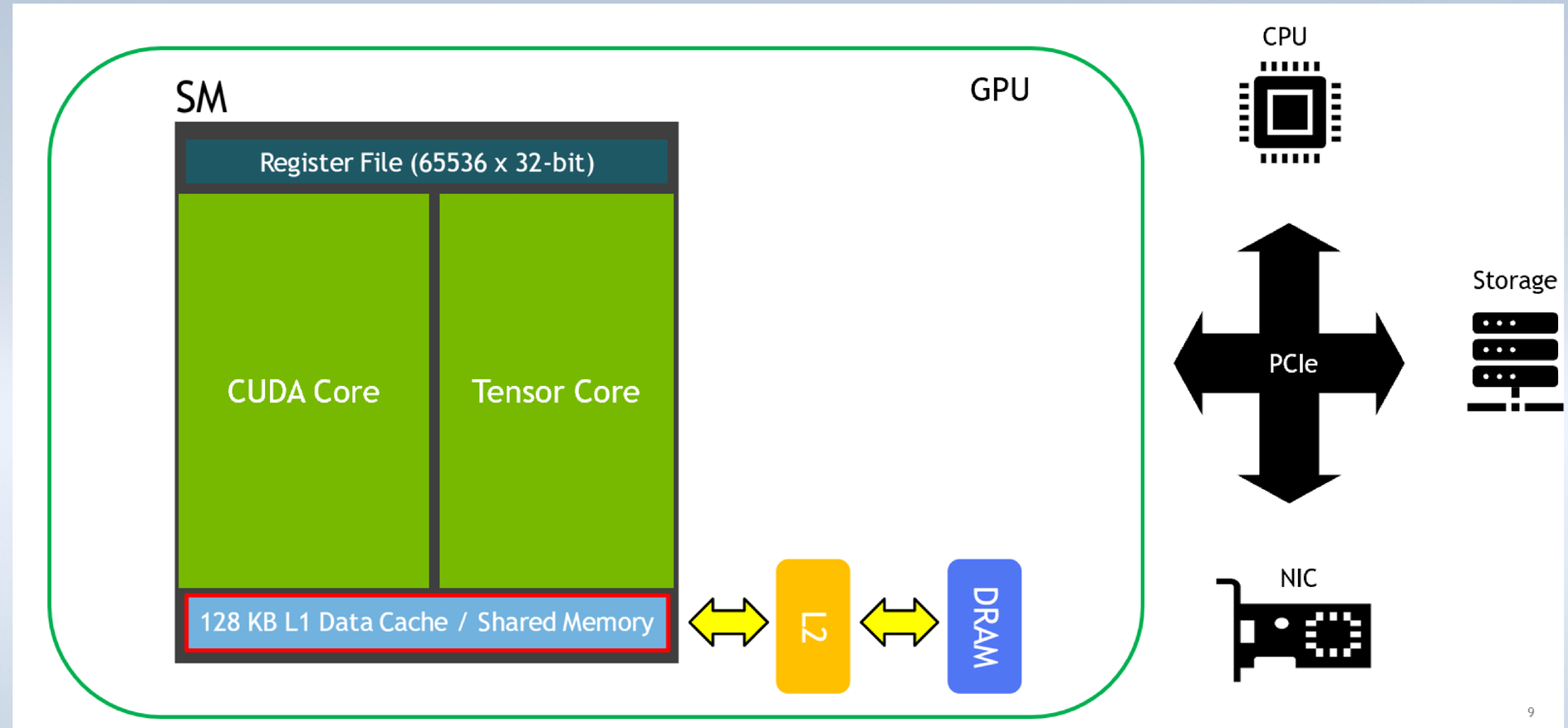
NVIDIA A100

Up to X20 Faster Computing Performance than Volta

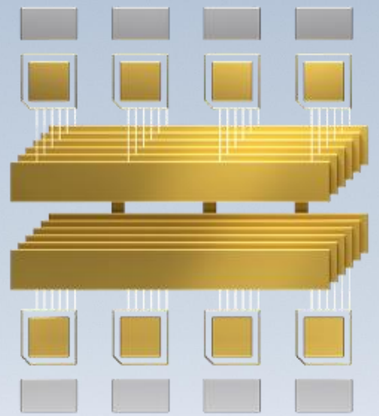
	Peak		Vs Volta
FP32 TRAINING	312	TFLOPS	20X
INT8 INFERENCE	1,248	TOPS	20X
FP64 HPC	19.5	TFLOPS	2.5X
MULTI INSTANCE GPU			7X GPUs



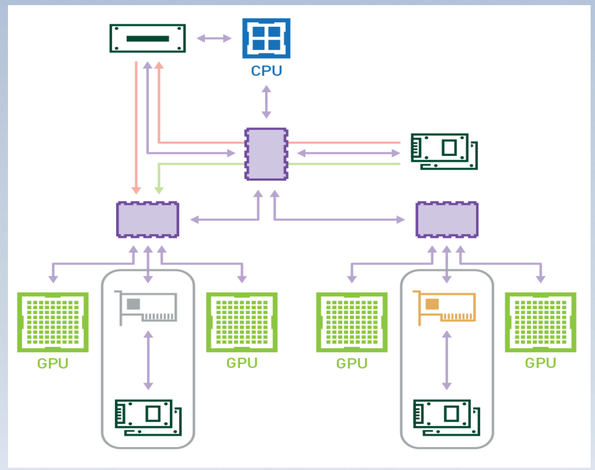
GPU SERVER


SM
GPU
Register File (65536 x 32-bit)
CUDA Core
Tensor Core
128 KB L1 Data Cache / Shared Memory
L2
DRAM
CPU
PCIe
Storage
NIC

HOW TO MAKE GPU BIGGER?

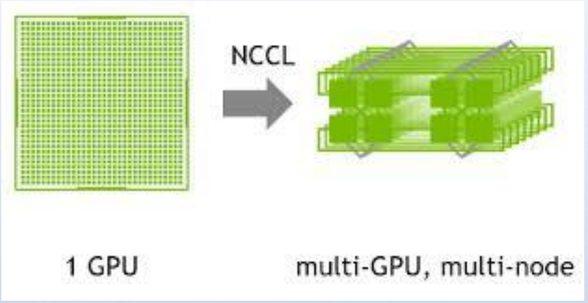


NVLink & NVSwitch



GPUDirect Technology

Mellanox Interconnect



NCCL

NVIDIA DGX A100

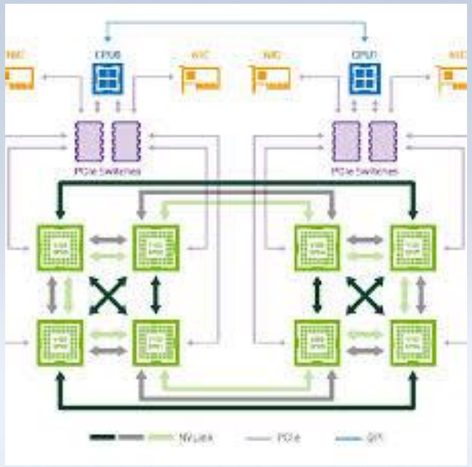
5 PetaFlops of AI Computing Performance in a Single Node

Compared to High-end CPU server

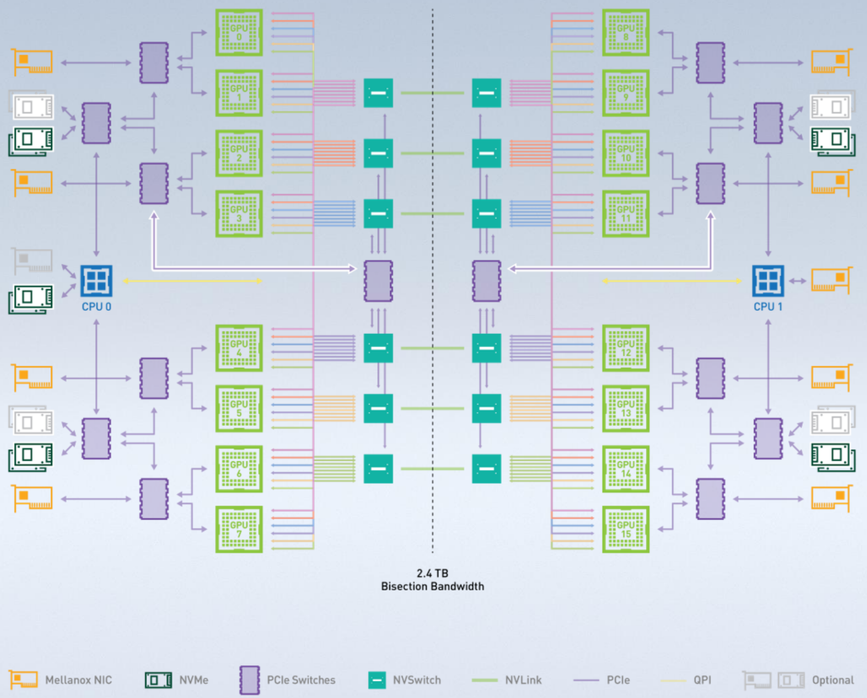
AI Compute	X 150
Memory Bandwidth	X 40
IO Bandwidth	X 40

DGX ARCHITECTURE EVOLUTION

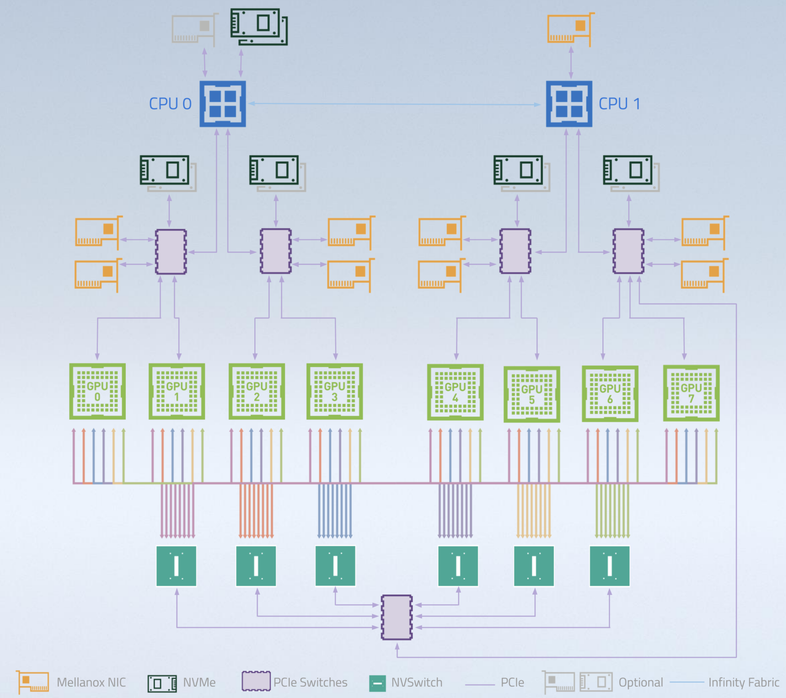
DGX-1V



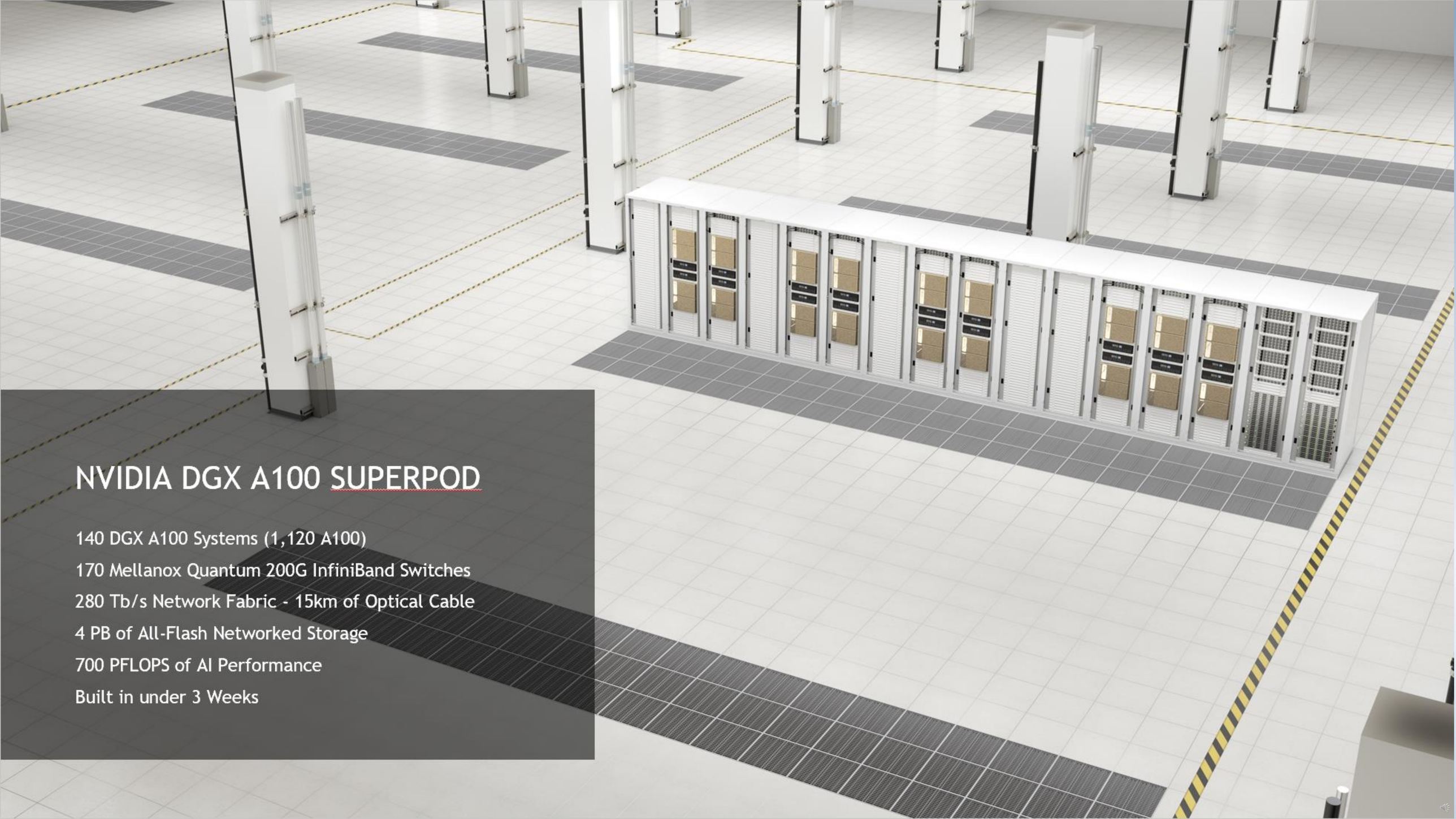
DGX-2



DGX-A100



1PF → 2PF → 5PF



NVIDIA DGX A100 SUPERPOD

140 DGX A100 Systems (1,120 A100)

170 Mellanox Quantum 200G InfiniBand Switches

280 Tb/s Network Fabric - 15km of Optical Cable

4 PB of All-Flash Networked Storage

700 PFLOPS of AI Performance

Built in under 3 Weeks

BUILD YOUR OWN SUPERCOMPUTER

https://www.nvidia.com/content/dam/en-zz/Solutions/data-center/gated-resources/nvidia-dgx-superpod-a100.pdf

NVIDIA EGX A100 FOR EDGE COMPUTING PLATFORM

NVIDIA Ampere GPU

3rd generation Tensor Core

New Security Engine for Confidential AI

Secure, Authenticated Boot

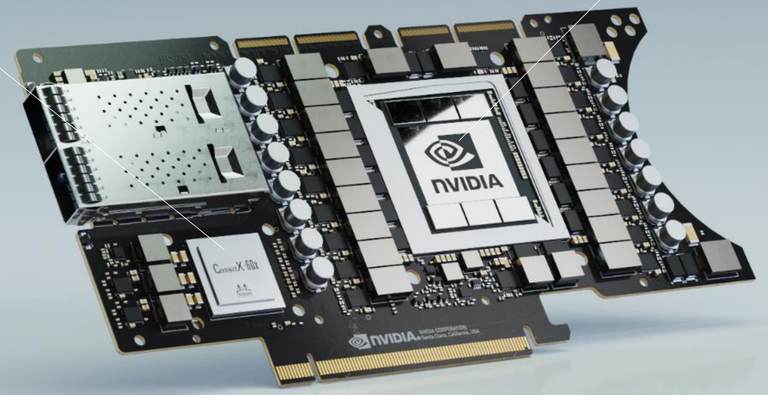
NVIDIA Mellanox ConnectX-6 DX

Dual 100 Gb/s Ethernet or InfiniBand

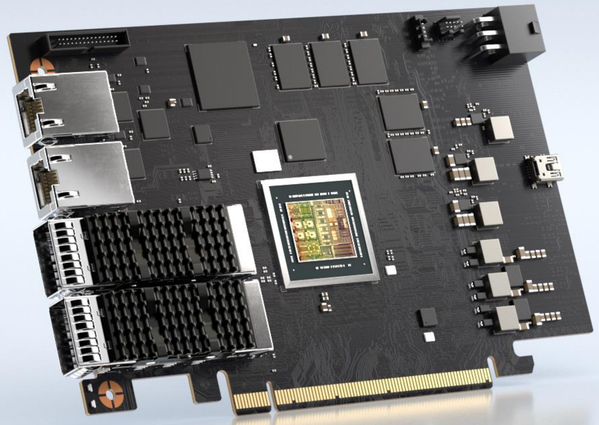
Line-speed TLS/IPSec Crypto Engine

Time Triggered Transmission Tech for Telco (5T for 5G)

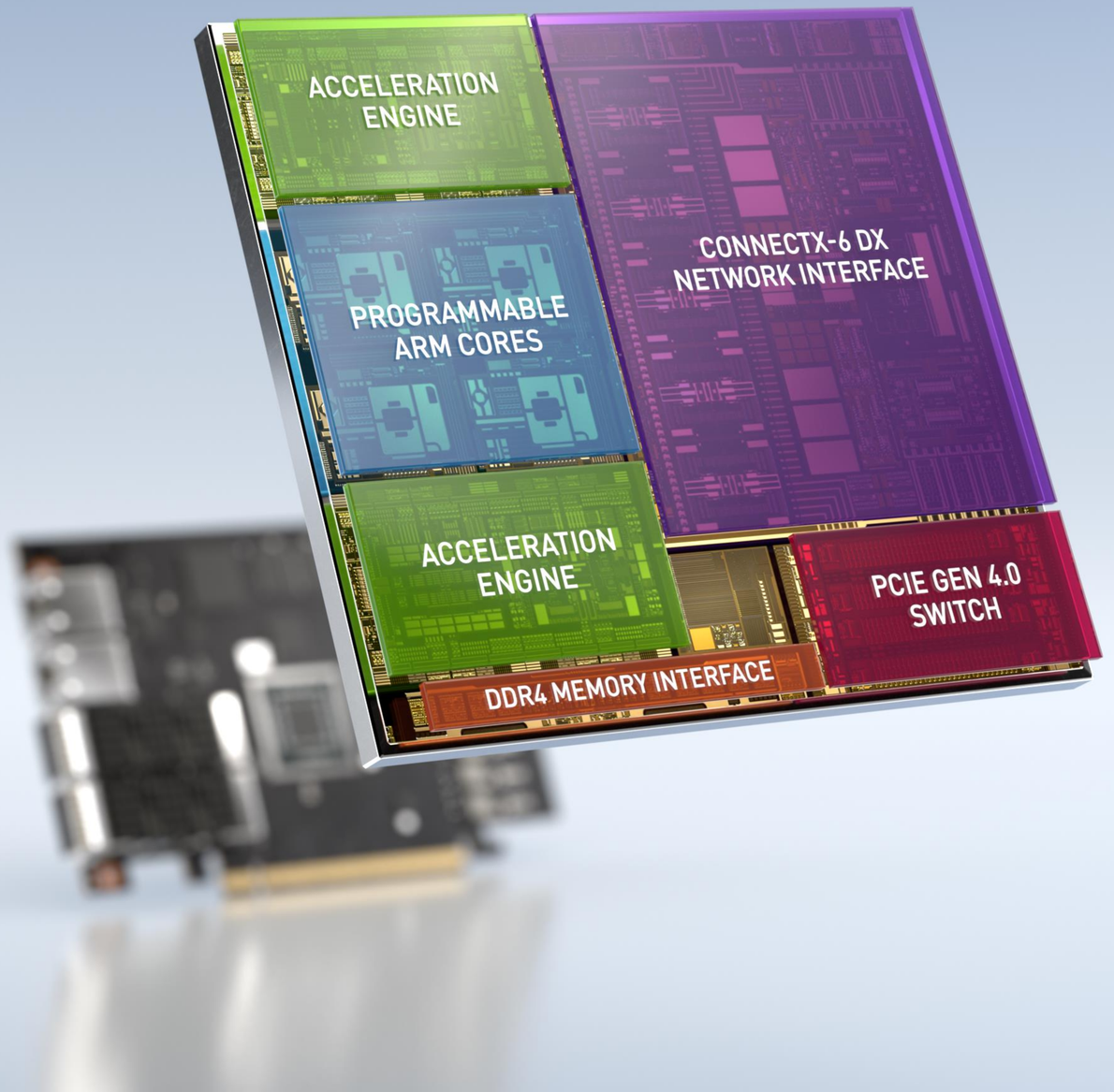
ASAP^2 SR-IOV and VirtualIO Offload



NVIDIA BLUEFIELD-2 DPU (DATA PROCESSING UNIT)



NVIDIA BLUEFIELD-2 DPU (DATA PROCESSING UNIT)



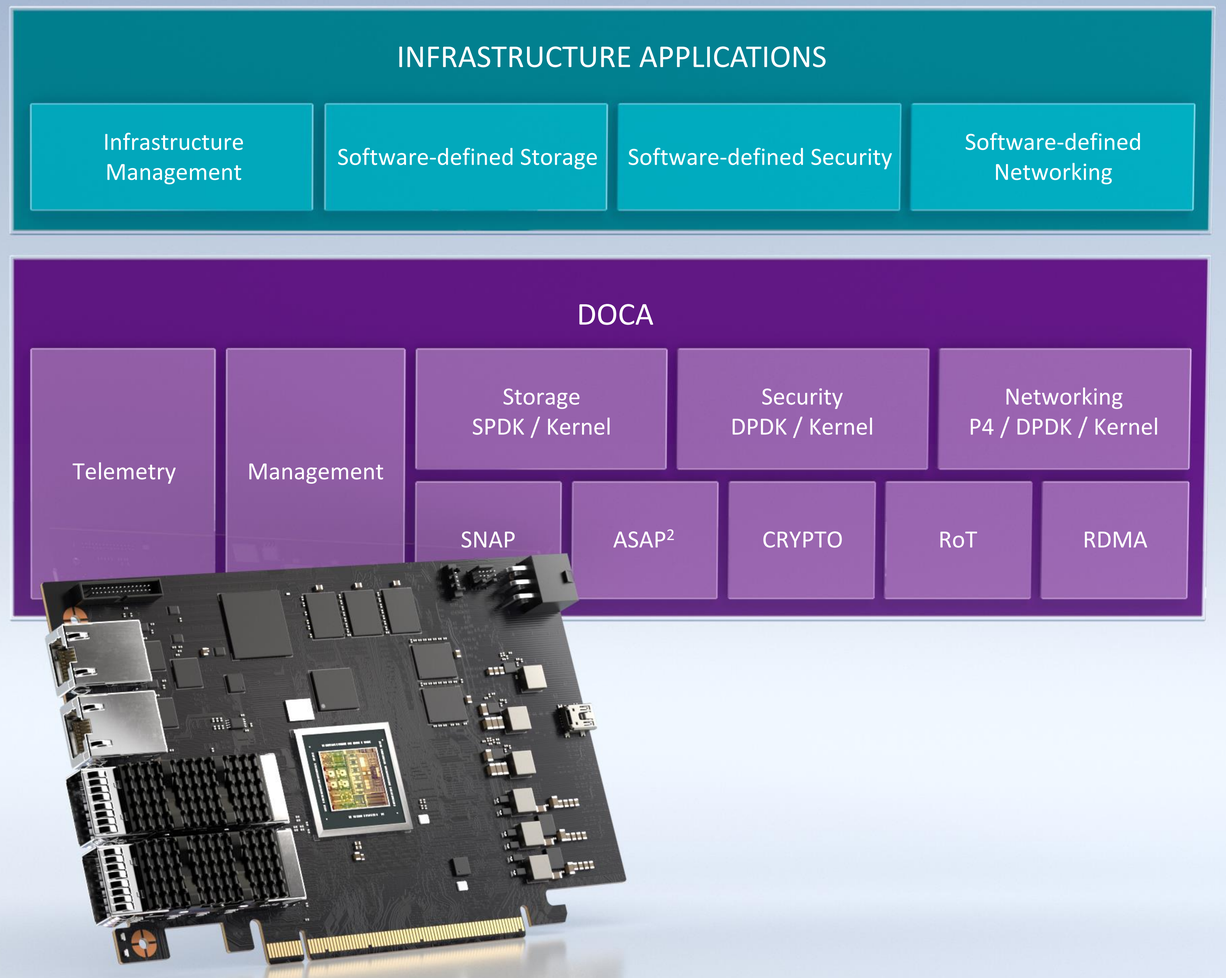
NVIDIA DOCA

DATA CENTER INFRASTRUCTURE-ON-A-CHIP ARCHITECTURE

SDK for BlueField DPUs

Open APIs – DPDK, SPDK, P4

Certified Reference Apps & Third-Party Solutions



ANNOUNCING
NVIDIA BLUEFIELD-2X

Programmable Data Center Infrastructure

Arm and CUDA Programmability

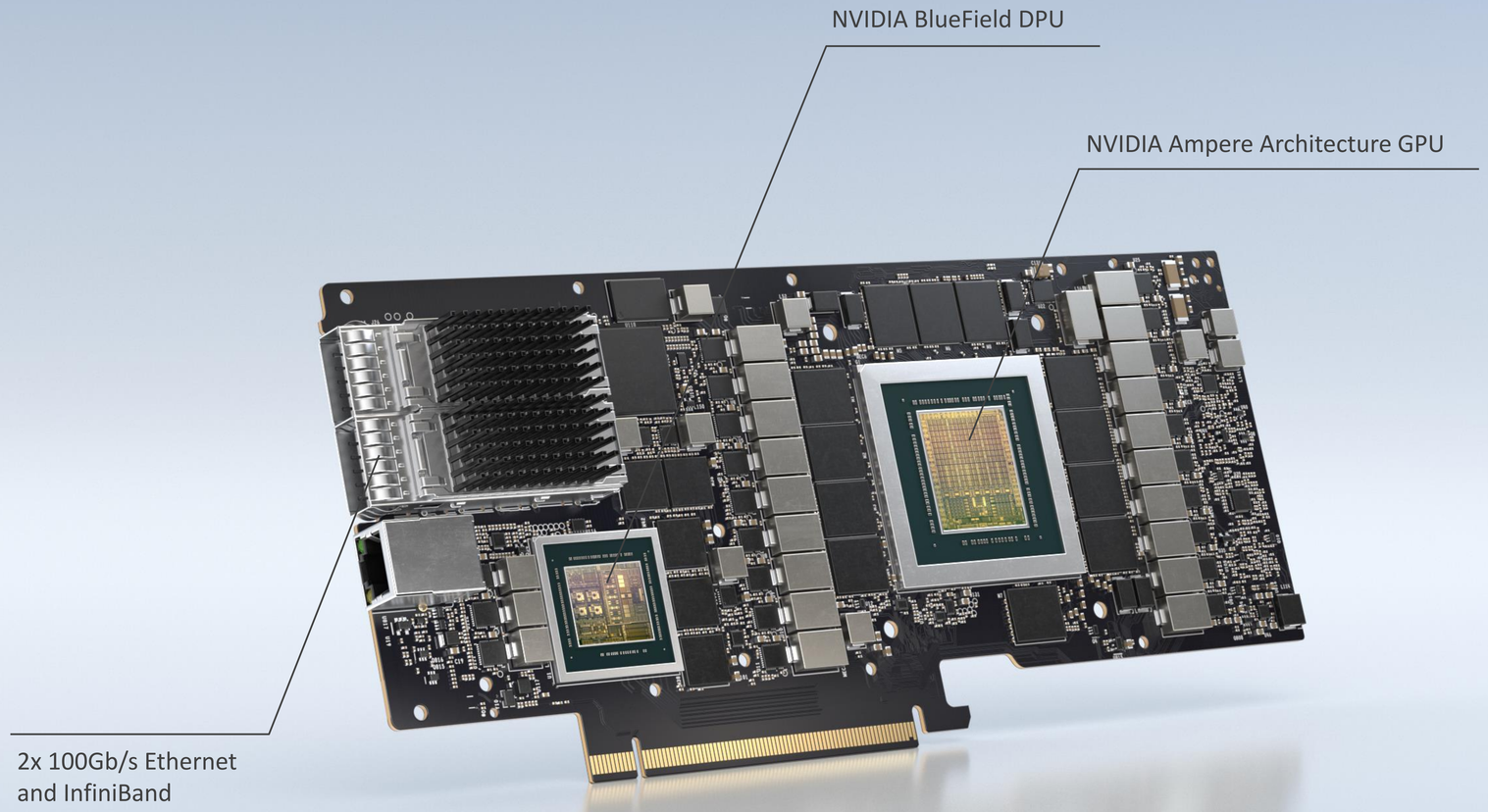
Anomaly Detection & Automated Response

Real-Time Traffic Analytics at Line Rate

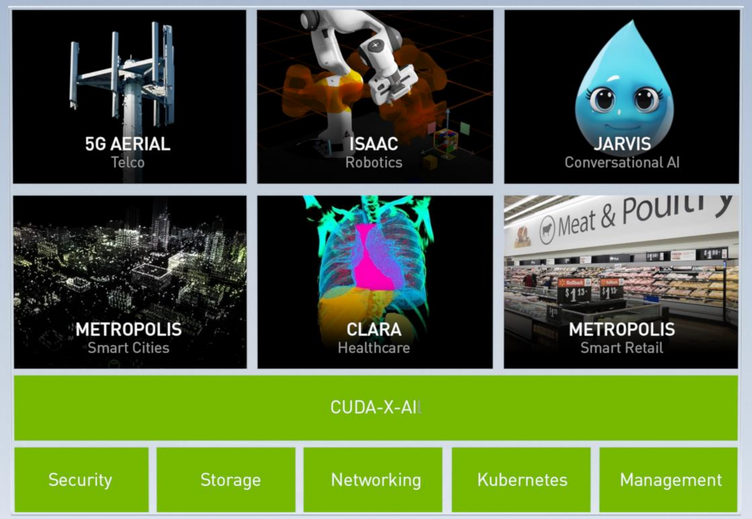
Host Introspection to Identify Malicious Activity

Dynamic Security Orchestration

Online Analytics of Uploaded Video



ANNOUNCING
EGX EDGE AI PLATFORM



NVIDIA BlueField-2

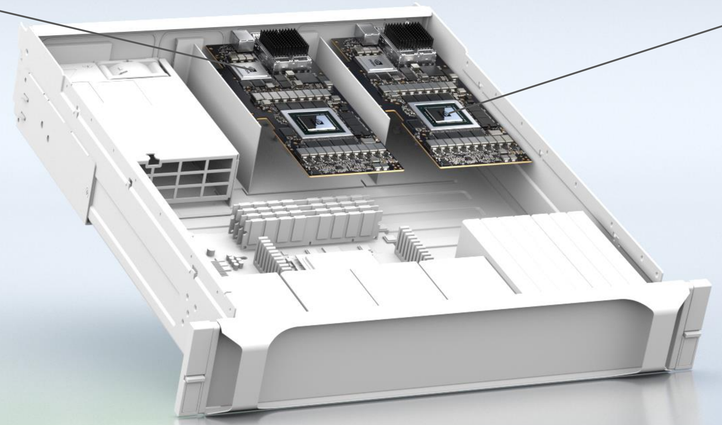
8x Arm Cores

Dual 100 Gb/s Ethernet or InfiniBand

Line-speed TLS/IPSec Crypto Engine

Time Triggered Transmission Tech for Telco (5T for 5G)

ASAP^2 SR-IOV and VirtualIO Offload



NVIDIA Ampere Architecture GPU

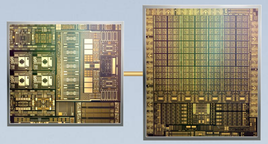
3rd generation Tensor Core

New Security Engine for Confidential AI

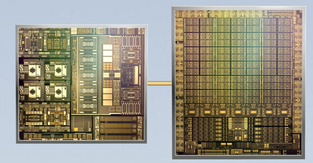
Secure, Authenticated Boot

NVIDIA DPU ROADMAP

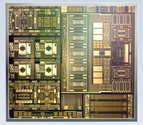
Programmable Data Center
Infrastructure-on-a-Chip



BlueField-2X
70 SPECINT
60 TOPS AI
200 Gbps

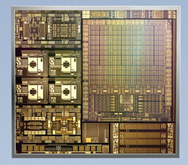


BlueField-3X
350 SPECINT
75 TOPS AI
400 Gbps



BlueField-3
350 SPECINT
1.5 TOPS
400 Gbps

BlueField-2
70 SPECINT
0.7 TOPS
200 Gbps



BlueField-4
1000 SPECINT
400 TOPS
400 Gbps

1000X
100X
10X
1X

DOCA — One Architecture

2021 2022 2023

THE PREMIER COMPUTING COMPANY
FOR THE AGE OF AI

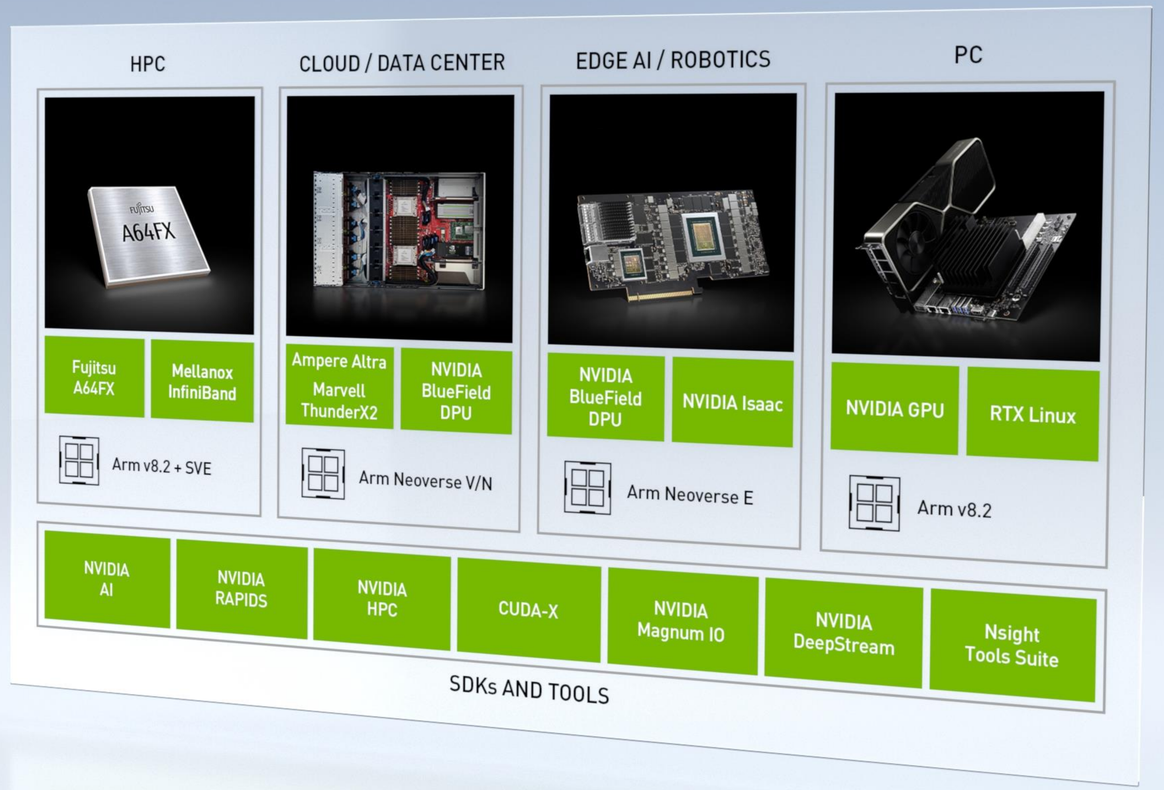
NVIDIA ACCELERATES ARM FROM CLOUD TO EDGE

Bringing GPU and DPU to Arm Ecosystem

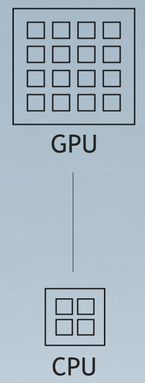
Accelerating HPC, Cloud, Edge, and PC Platforms

Offering NVIDIA's Most Advanced SDKs for Arm — AI, HPC, RTX Graphics

Partnering with Fujitsu, Ampere Computing, Marvell

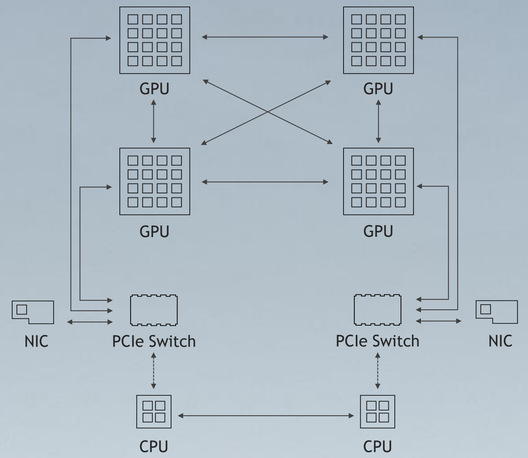


25 YEARS OF ACCELERATED COMPUTING

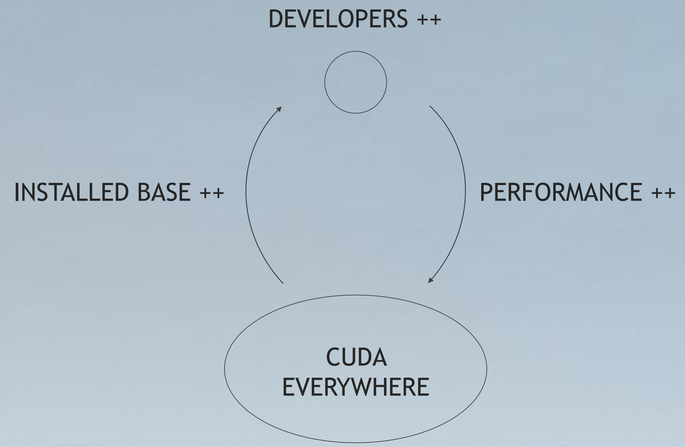


X-FACTOR SPEED-UP

FULL STACK

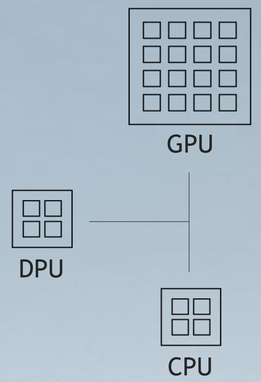


SYSTEMS



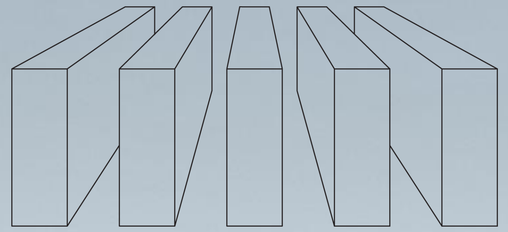
ONE ARCHITECTURE

DATA-CENTER-SCALE ACCELERATED COMPUTING

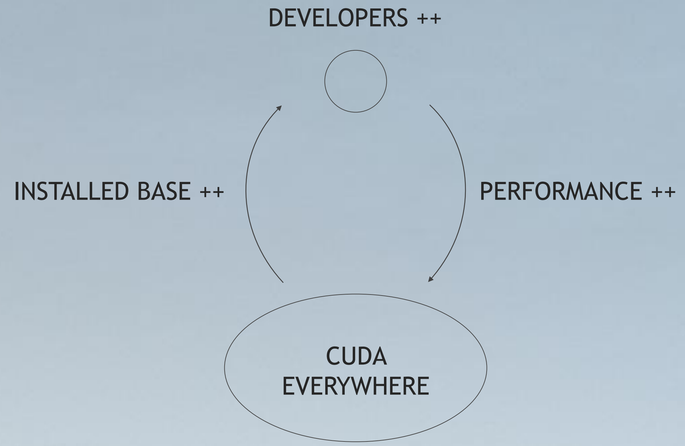


X-FACTOR SPEED-UP

FULL STACK



DATA CENTER SCALE



ONE ARCHITECTURE


NVIDIA