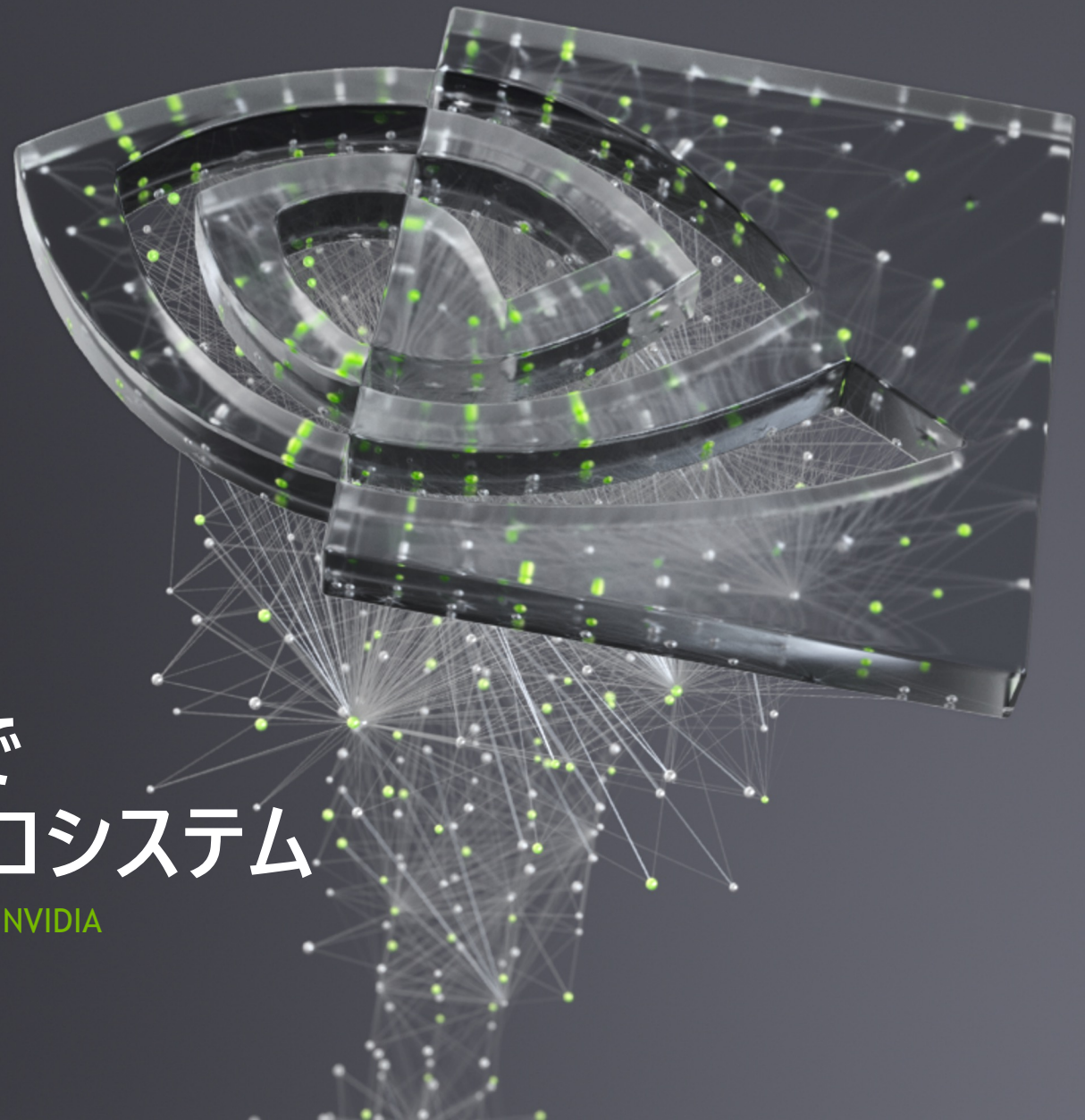


組み込みから HPC まで ARM コアで実現するエコシステム

Shinnosuke Furuya, Ph.D., HPC Developer Relations, NVIDIA

2021/08/26

AGENDA

Automotive

NVIDIA Drive

HPC

NVIDIA Grace CPU

Network

NVIDIA BlueField DPU


NVIDIA
“Best Places to Work in 2021”
GLASSDOOR
“Most Innovative Companies”
FAST COMPANY
“100 Best Companies to Work For”
FORTUNE
“50 Smartest Companies”
MIT TECH REVIEW
“World’s Best Performing CEO”
HARVARD BUSINESS REVIEW
“World’s Best CEOs”
BARRON’S
Founded in 1993
Jensen Huang, Founder & CEO
19,000 Employees
$16.7B in FY21


AUTOMOTIVE

NVIDIA ORIN

Advanced, Software-defined Platform for Autonomous Machines

24.5 billion transistors

12 A78 (Hercules) ARM64 CPUs

254 INT8 TOPS - CUDA Tensor Core GPU + DLA

205 GB/s memory bandwidth

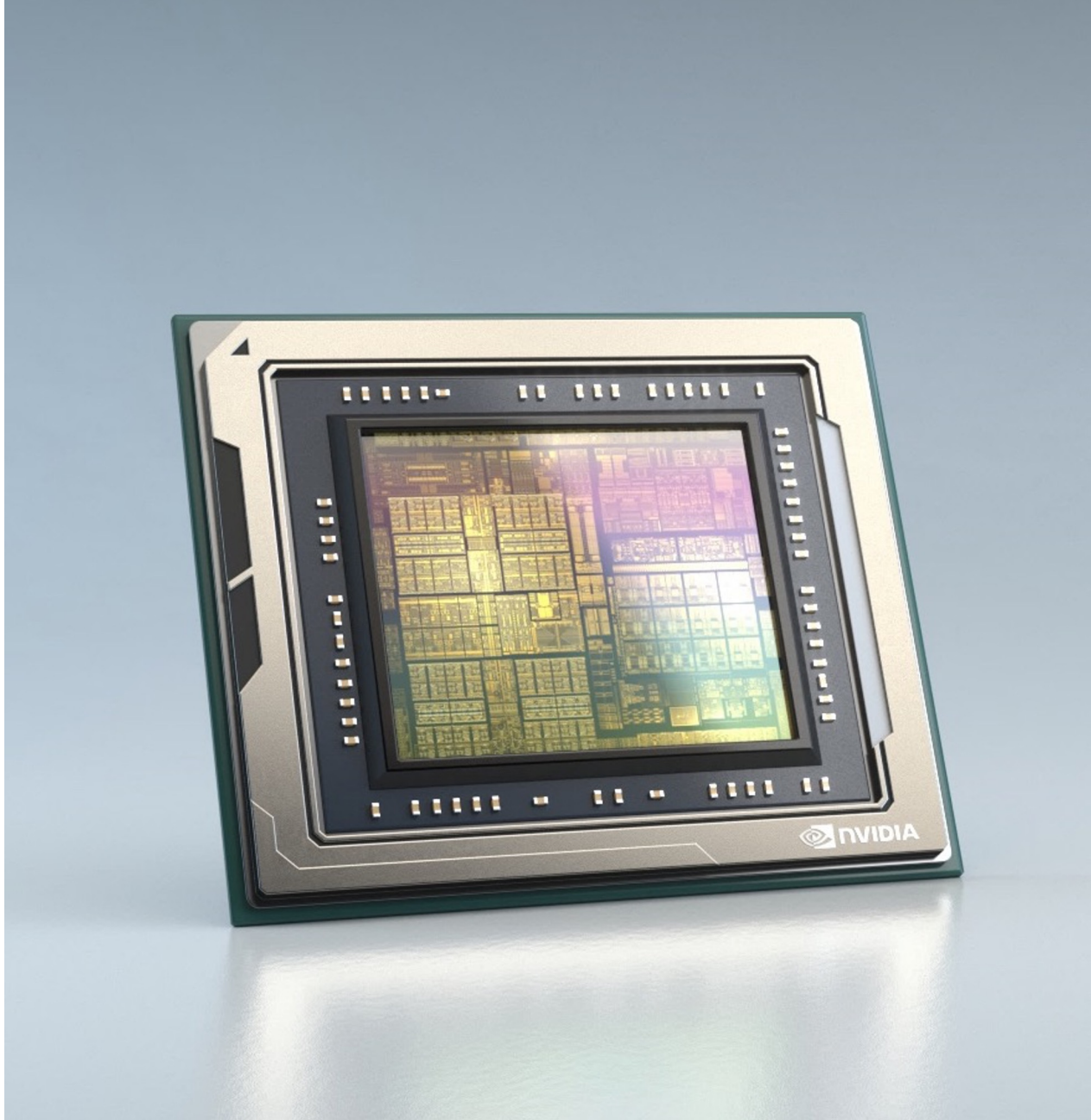
4 10Gbps ENET

8K 30 Dec | 4K 60 Enc – H264 / H265 / VP9

4 R52 Lock-step Pairs Integrated Safety Island ASIL-D

Secure key storage

FUSA ASIL-B Chip | ASIL-D Systematic



NVIDIA DRIVE ATLAN

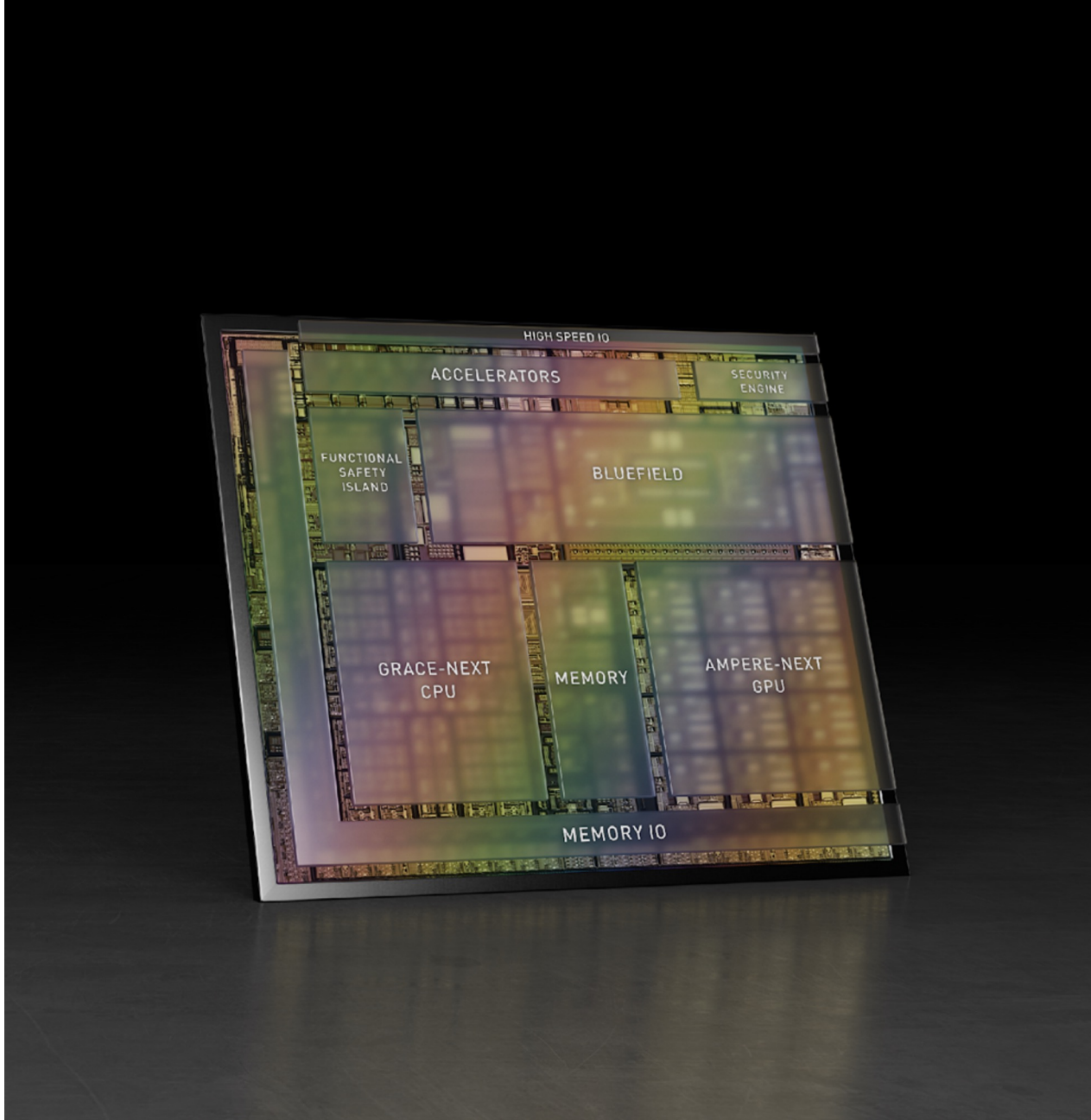
TOPS is the New Horsepower

Fusing Next Generation AI and BlueField

Industry's First 1,000 TOPS SoC

400 Gbps Networking with Secure Gateway

ASIL-D Safety Island



HYPERION 8
AV PLATFORM

State-of-the-Art Advances for Data Collection, Development and Testing

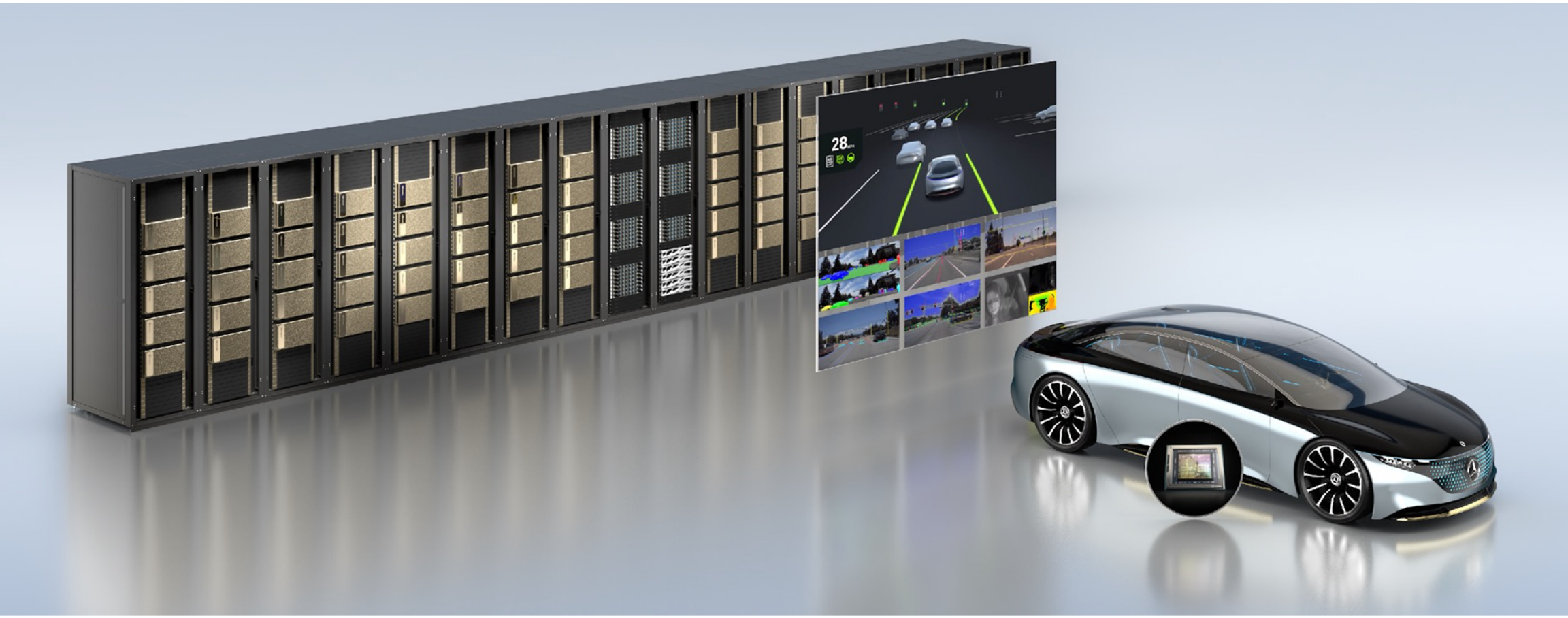
2x Orin AV Computer

1x Orin IX Computer

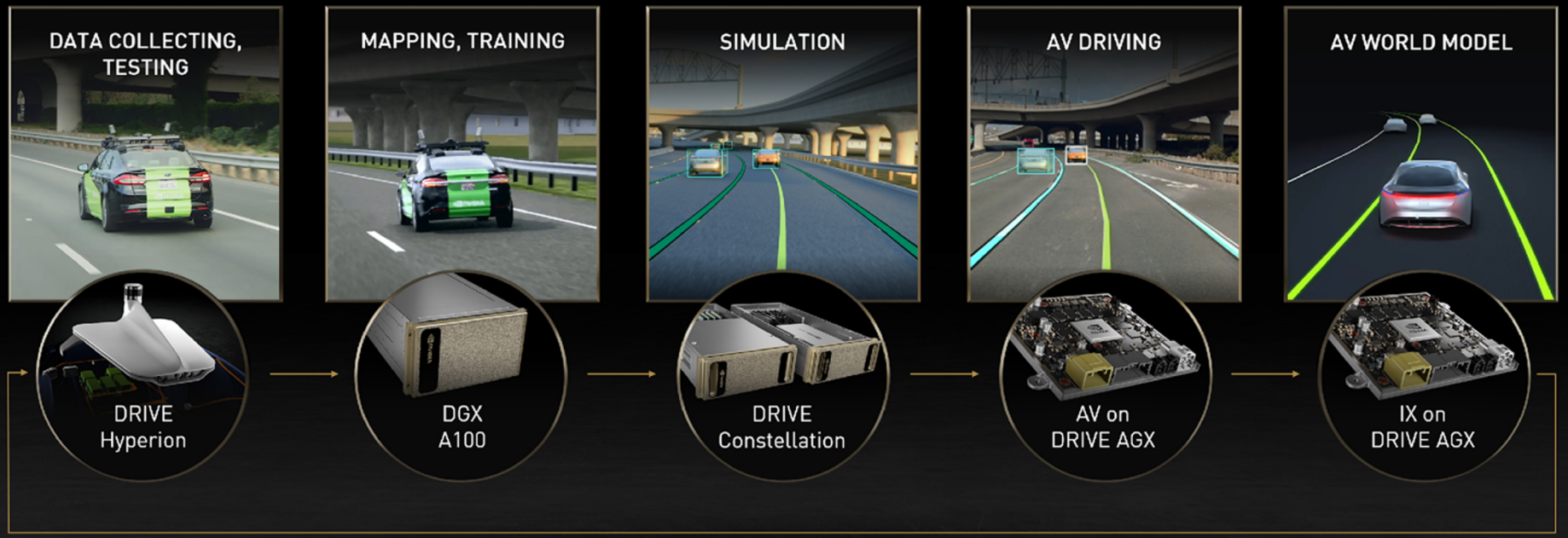
4x Orin + 4x MLNX 3D GT Data Recorder

Sensor Suite: 8 Cameras [8MP], 4 Fisheyes [3MP], 3 In-Cabin, 9 Radar, 2 Lidar

THE FUTURE CAR IS SOFTWARE DEFINED



NVIDIA DRIVE AV

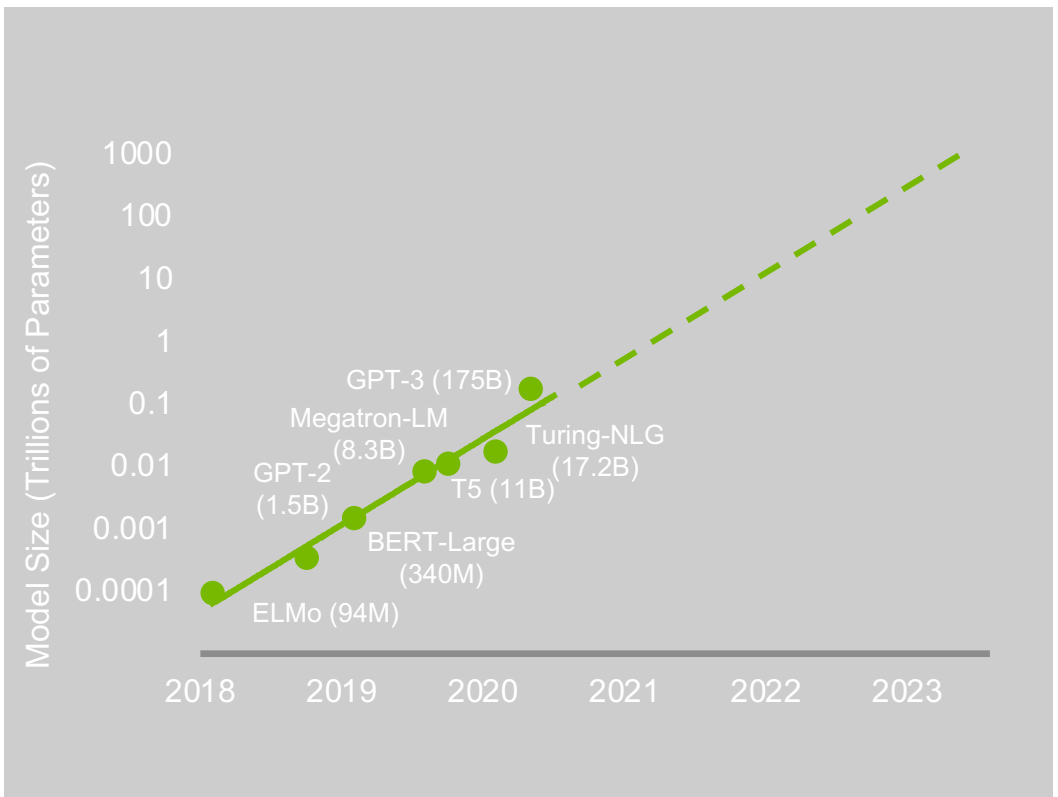



HPC

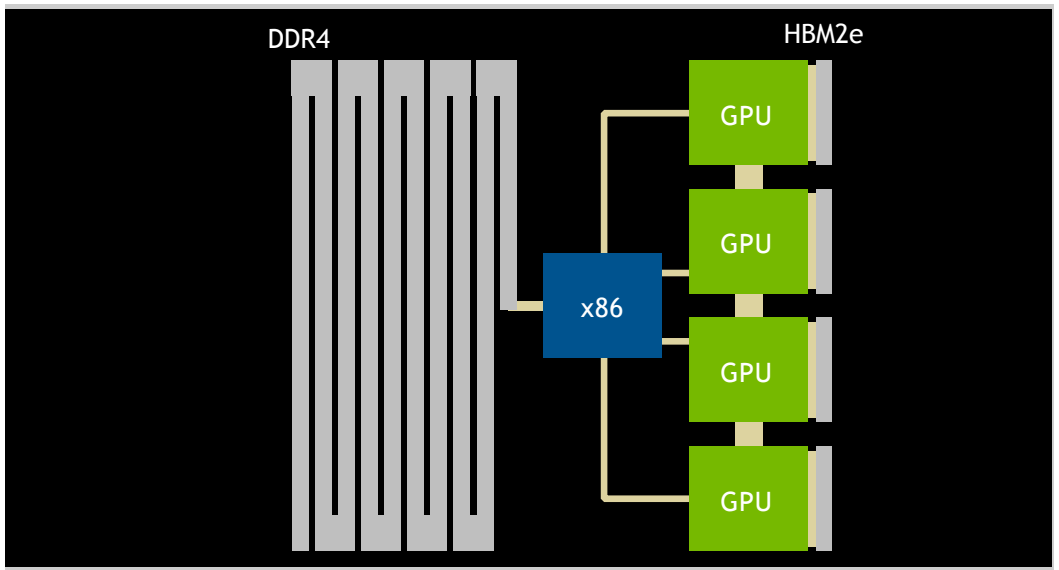
GIANT MODELS PUSHING LIMITS OF EXISTING ARCHITECTURE

Requires a New Architecture

100 TRILLION PARAMETER MODELS BY 2023



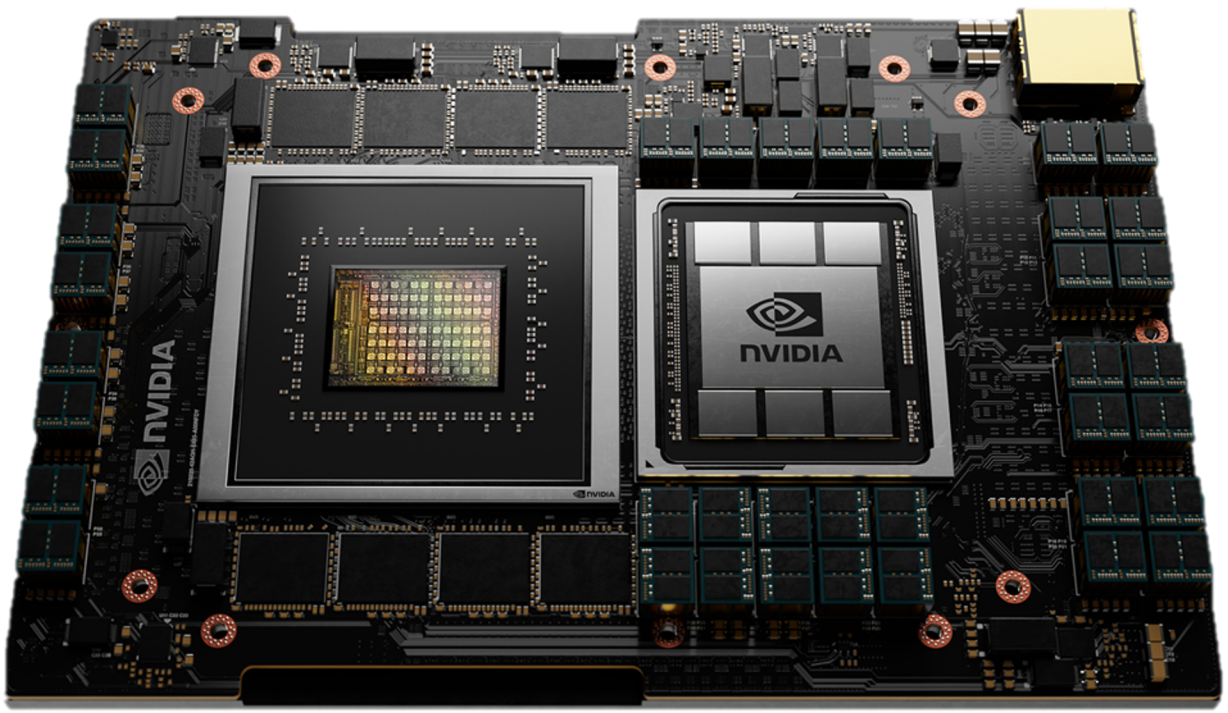
System Bandwidth Bottleneck



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
Mem-to-GPU	**64**	**GB/sec**

ANNOUNCING NVIDIA GRACE

Breakthrough CPU Designed for Giant-Scale AI and HPC Applications



FASTEST INTERCONNECTS
>900 GB/s Cache Coherent NVLink CPU To GPU (14x)
>600GB/s CPU To CPU (2x)

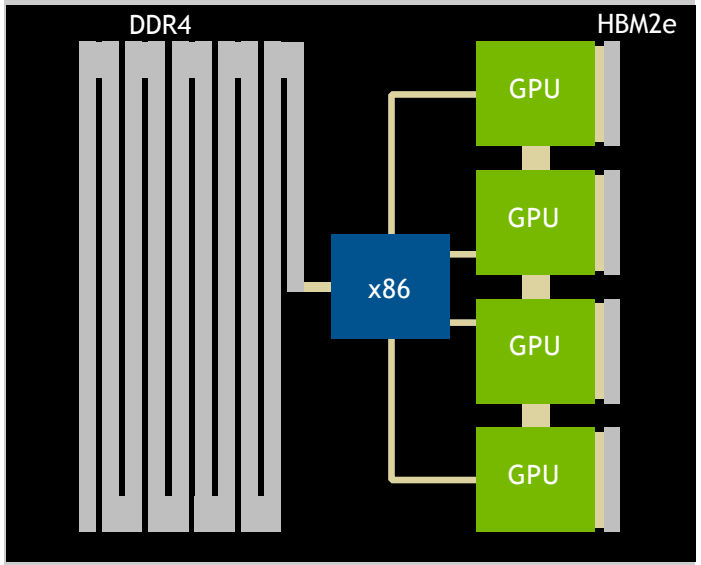
HIGHEST MEMORY BANDWIDTH
>500GB/s LPDDR5x w/ ECC
>2x Higher B/W
10x Higher Energy Efficiency

NEXT GENERATION ARM NEOVERSE CORES
>300 SPECrate2017_int_base est.
Availability 2023

TURBOCHARGED TERABYTE SCALE ACCELERATED COMPUTING

Evolving Architecture For New Workloads

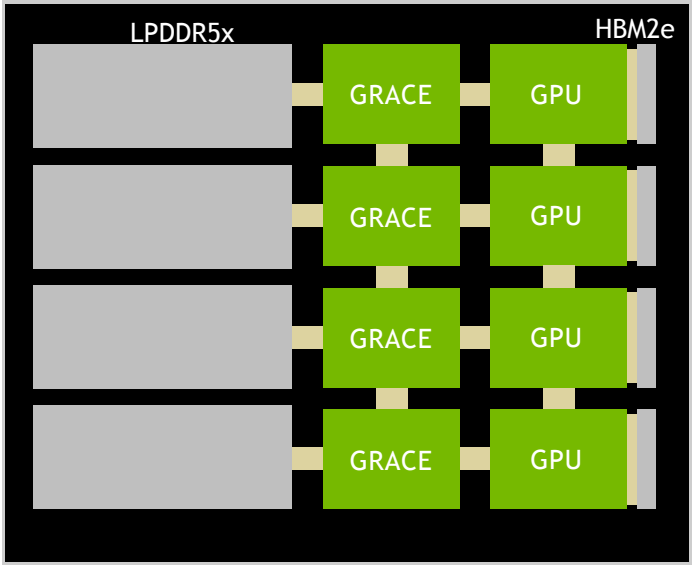
CURRENT x86 ARCHITECTURE



GPU	8,000	GB/sec
CPU	200	GB/sec
PCIE Gen4 (Effective Per GPU)	16	GB/sec
Mem-to-GPU	64	GB/sec

Transfer 2TB in 30 secs

INTEGRATED CPU-GPU ARCHITECTURE



GPU	8,000	GB/sec
CPU	500	GB/sec
NVLink	500	GB/sec
Mem-to-GPU	2000	GB/sec

Transfer 2TB in 1 secs

3 DAYS FROM 1 MONTH
Fine-Tune Training of 1T Model

REAL-TIME INFERENCE ON 0.5T MODEL
Interactive Single Node NLP Inference

Bandwidth claims rounded to nearest hundred for illustration.
Performance results based on projections on these configurations Grace : 8xGrace and 8xA100 with 4th Gen NVIDIA NVLink Connection between CPU and GPU and x86: DGX A100.
Training: 1 Month of training is Fine-Tuning a 1T parameter model on a large custom data set on 64xGrace+64xA100 compared to 8xDGXA100 (16xX86+64xA100)
Inference: 530B Parameter model on 8xGrace+8xA100 compared to DGXA100.

ANNOUNCING

THE WORLD'S FASTEST SUPERCOMPUTER FOR AI

20 Exaflops of AI

Accelerated w/ NVIDIA Grace CPU and NVIDIA GPU

HPC and AI For Scientific and Commercial Apps

Advance Weather, Climate, and Material Science

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Hewlett Packard Enterprise

NVIDIA


NETWORK

INTRODUCING NVIDIA BLUEFIELD-3 DPU

First 400Gb/s Data Processing Unit

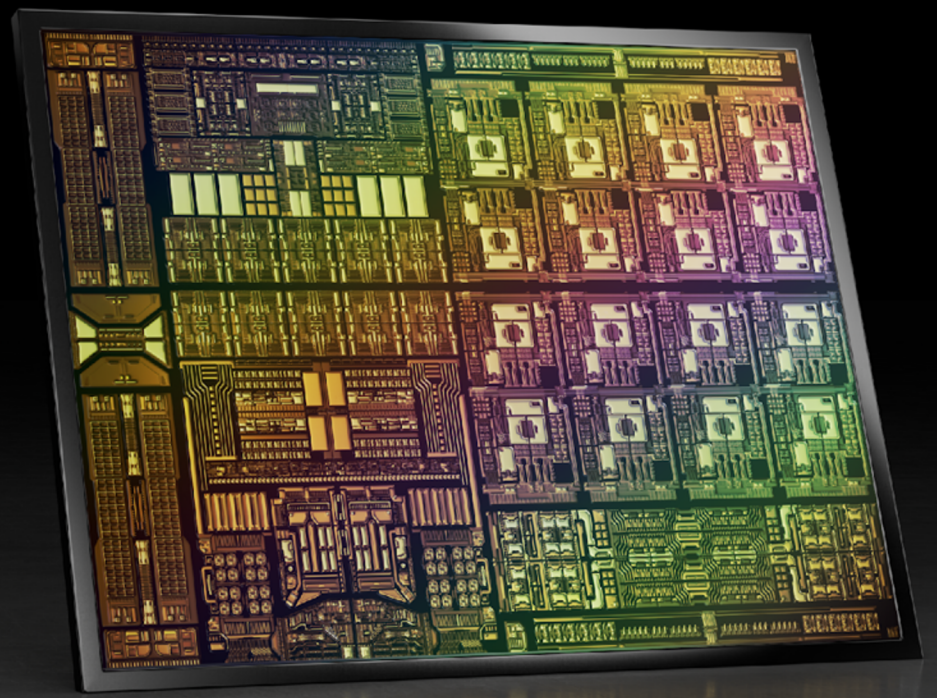
Offloads and Accelerates Data Center Infrastructure

Isolates Application from Control and Management Plane

Powerful CPU – 16x Arm A78 Cores

Datapath Accelerator – 16x Cores, 256 Threads

Process Networking, Storage, and Security at 400 Gbps



INTRODUCING NVIDIA BLUEFIELD-3 DPU

First 400Gb/s Data Processing Unit

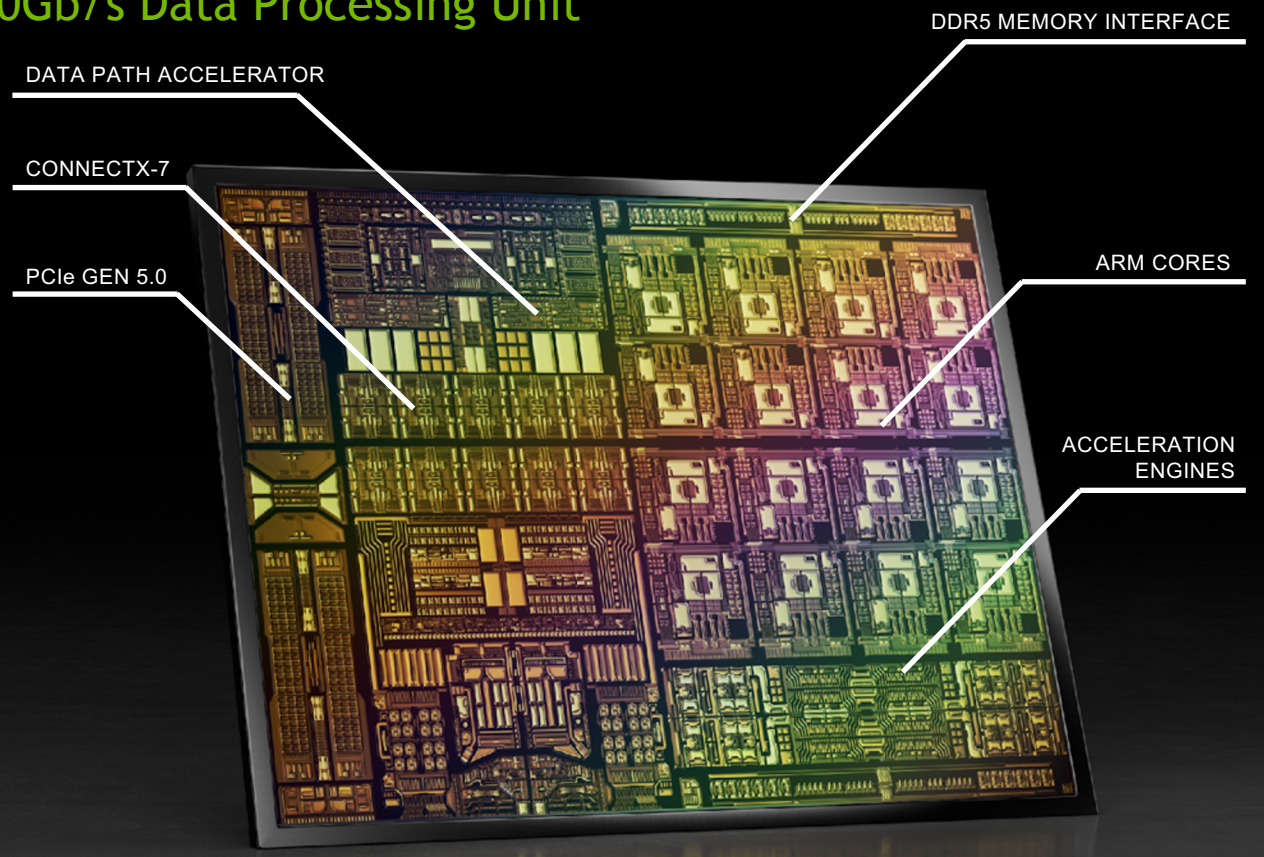
22 Billion Transistors

400Gb/s Ethernet & InfiniBand Connectivity

400Gb/s Crypto Acceleration

18M IOP/s Elastic Block Storage

300 Equivalent x86 Cores



BLUEFIELD DPU GENERATIONS

	BlueField	BlueField-2	BlueField-3
Port speed	2 x 100Gb/s InfiniBand and Ethernet	2 x 100Gb/s, 1 x 200Gb/s InfiniBand and Ethernet	1 x 400Gb/s, 2x200Gb/s InfiniBand and Ethernet
Performance	Bandwidth: 200Gb/s DPDK Max Msg Rate:150Mpps RDMA max msg rate: 200Mpps	Bandwidth: 200Gb/s DPDK Max Msg Rate: 215Mpps RDMA max msg rate: 215Mpps	Bandwidth: 400Gb/s DPDK max msg rate: 250Mpps RDMA max msg rate: 330Mpps
Modulation	NRZ	NRZ & 50G PAM4	NRZ & 100G PAM4
DDR Channels	DDR4-2400MT/s Dual channels	DDR4-3200MT/s Single channel	2 x DDR5-5600 Interfaces
Max Arm Cores	16 x A72 Arm cores	8 x A72 Arm cores	16 x A78 Arm cores (Hercules)
Embedded ASIC	ConnectX-5	ConnectX-6 Dx	ConnectX-7
PCIe	Gen3.0 x32 / Gen4.0 x16	Gen4.0 x16	Gen5.0 x32

NVIDIA DOCA

Enabling Broad BlueField Partner Ecosystem

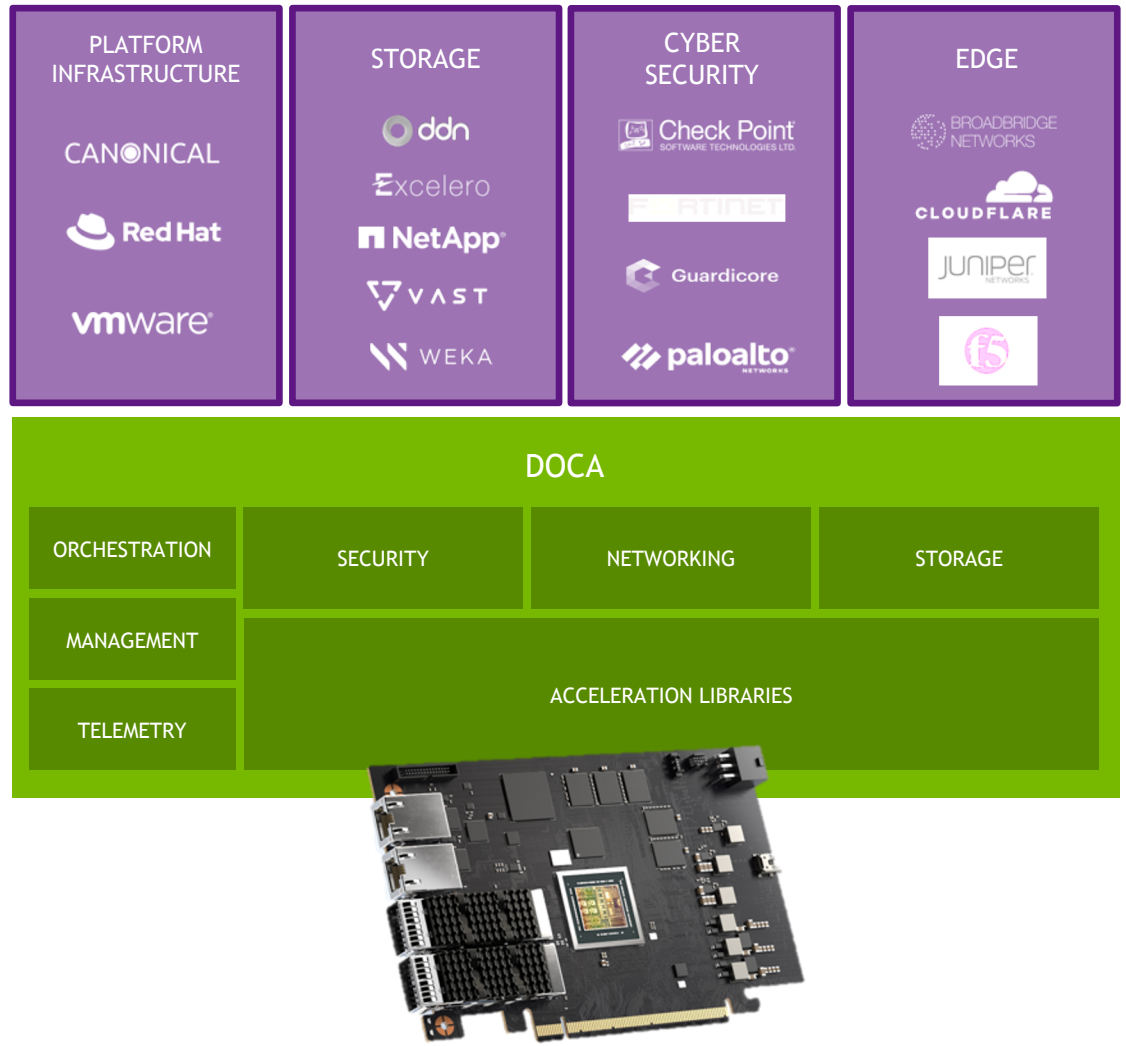
Software Development Framework for BlueField DPUs

Offload, Accelerate, and Isolate Infrastructure Processing

Support for Hyperscale, Enterprise, Supercomputing and Hyperconverged Infrastructure

Software Compatibility for Generations of BlueField DPUs

DOCA is for DPUs what CUDA is for GPUs



BLUEFIELD-3 USE CASES

Unprecedented Innovation for Modern Data Centers

Cloud Computing

Bare-Metal I Virtualized I Containerized
Private I Public I Hybrid Cloud

Cyber Security

Distributed Security | NGFW I
Micro-segmentation

HPC & AI

Cloud-Native Supercomputing |
Accelerated DLRM

Telco & Edge

Telco Cloud | CloudRAN |
Edge Compute

Data Storage

HCI I Elastic Block Storage I
Instance Storage

Media Streaming

Visual High Quality I
8K Video I CDN

BLUEFIELD ENABLES CLOUD-NATIVE SUPERCOMPUTING

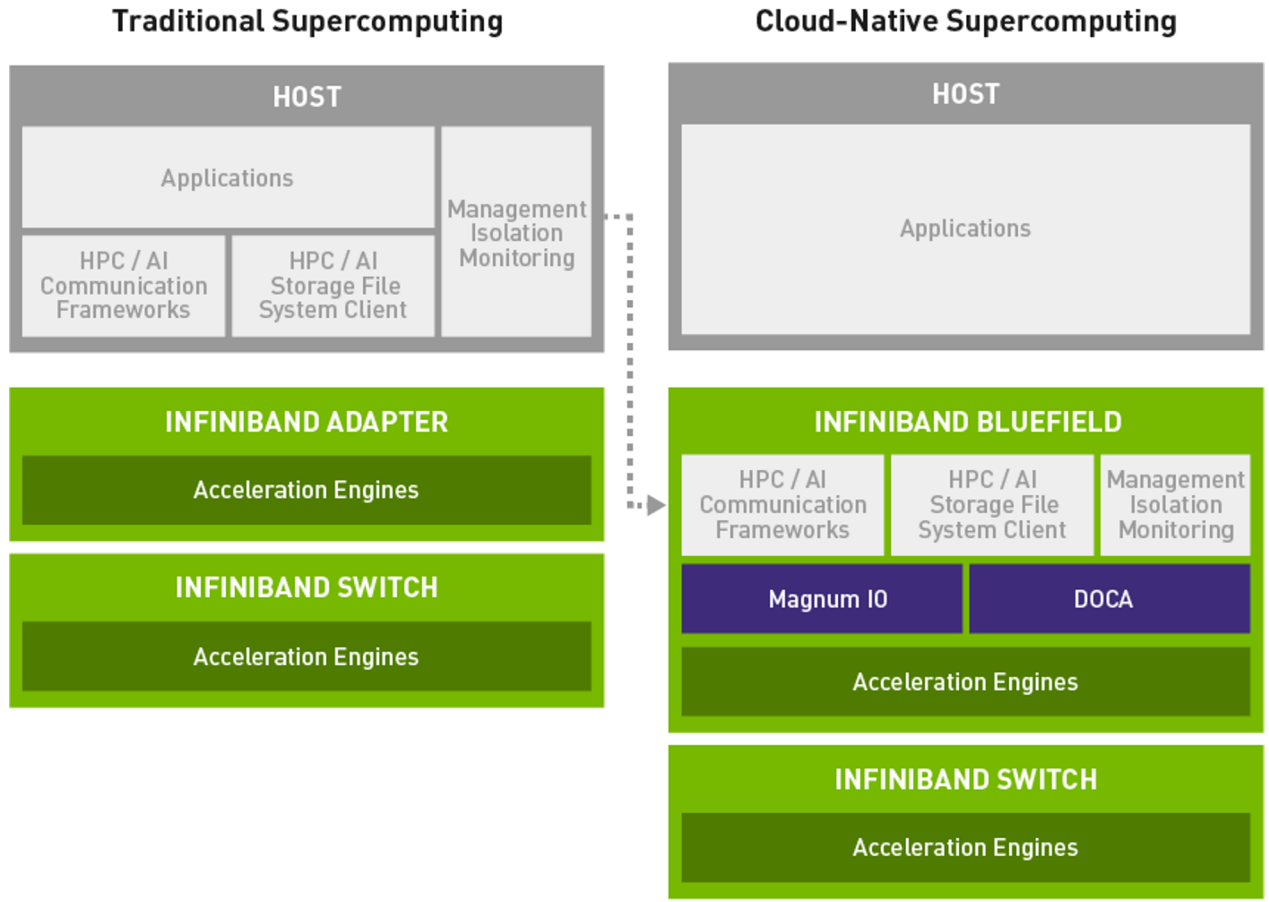
Multi-Tenancy with Zero-Trust Security

Collective offload with UCC accelerator

Smart MPI progression

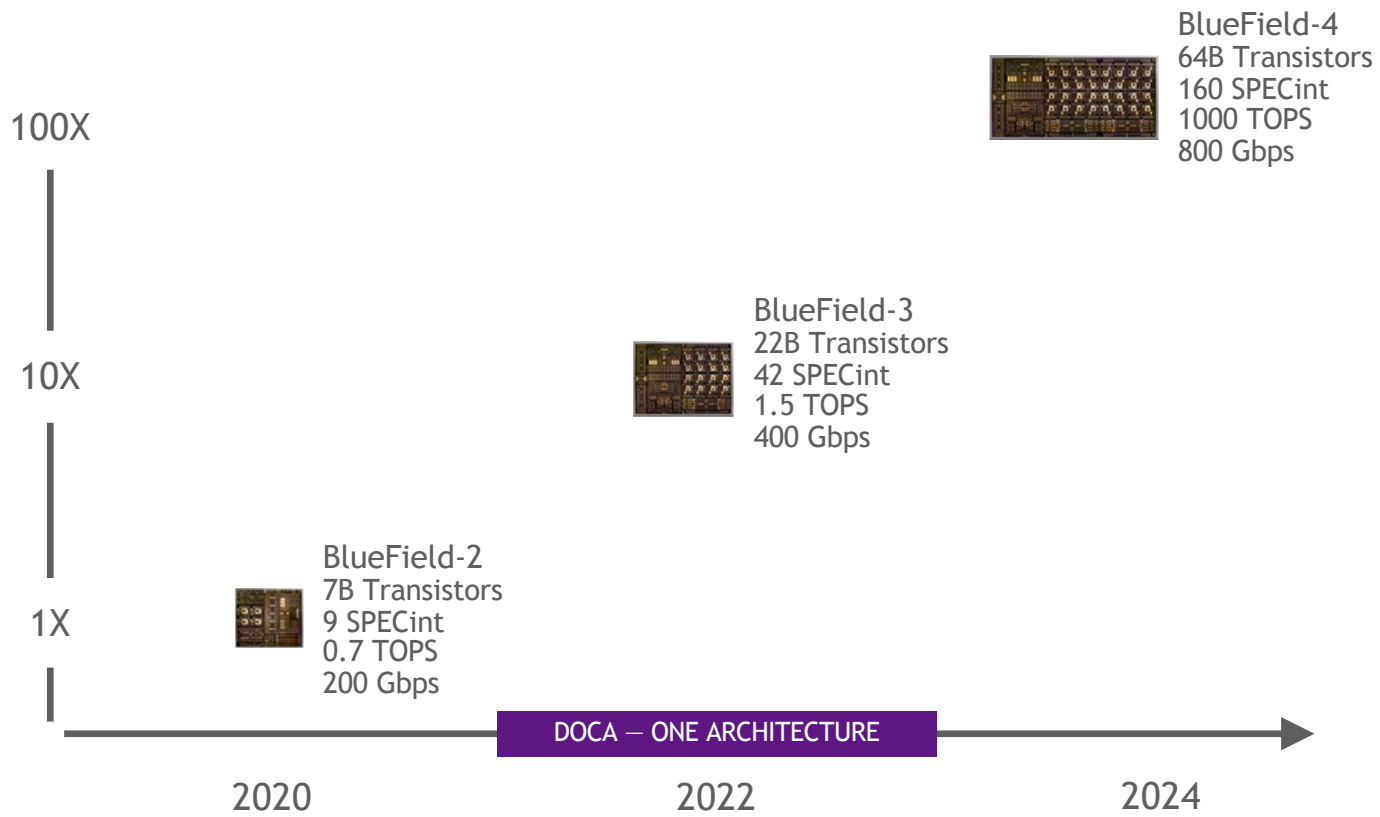
User-defined algorithms

1.4X higher application performance



NVIDIA DPU ROADMAP

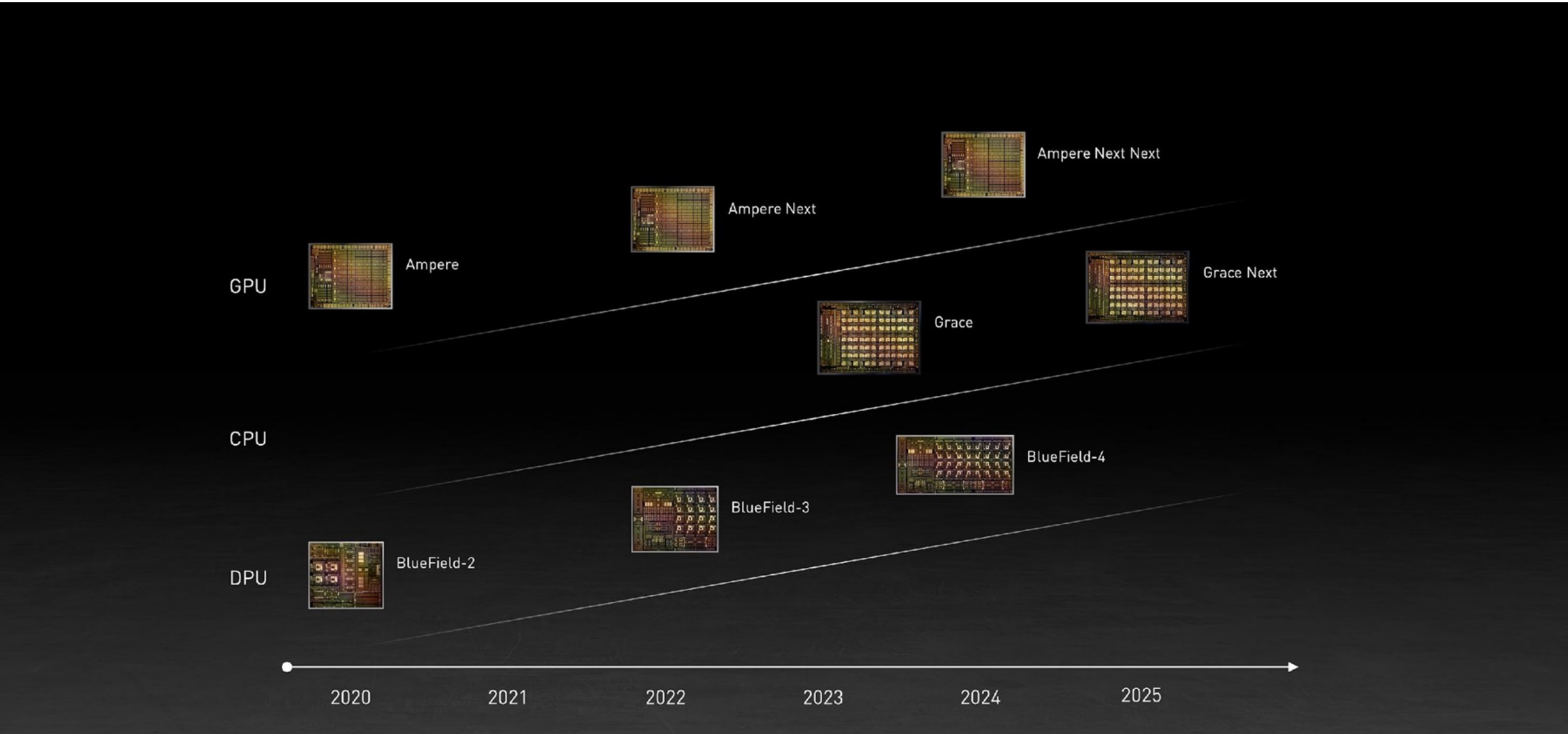
Exponential Growth in Data Center Infrastructure Processing



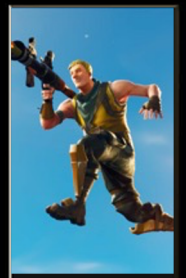
** BlueField-4 product to include opt-in GPU and non-GPU configurations*

SUMMARY

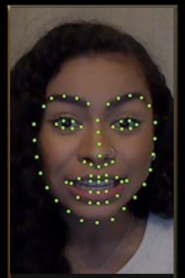
3 CHIPS. YEARLY LEAPS. ONE ARCHITECTURE.

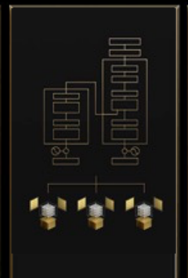


NVIDIA ECOSYSTEM PLATFORM

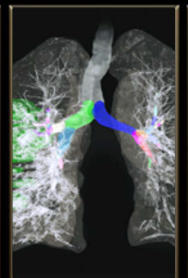

GEFORCE


RIVA


MAXINE


MEGATRON


MERLIN


CLARA

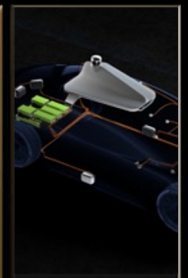

cuQUANTUM

METROPOLIS


MORPHEUS


AERIAL


ISAAC


DRIVE

APPLICATION FRAMEWORKS


VGPU


OMNIVERSE


AI ENTERPRISE


FLEET COMMAND

PLATFORM SOFTWARE


RTX


DGX

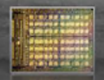

HGX


EGX


AGX

CHIPS & SYSTEMS

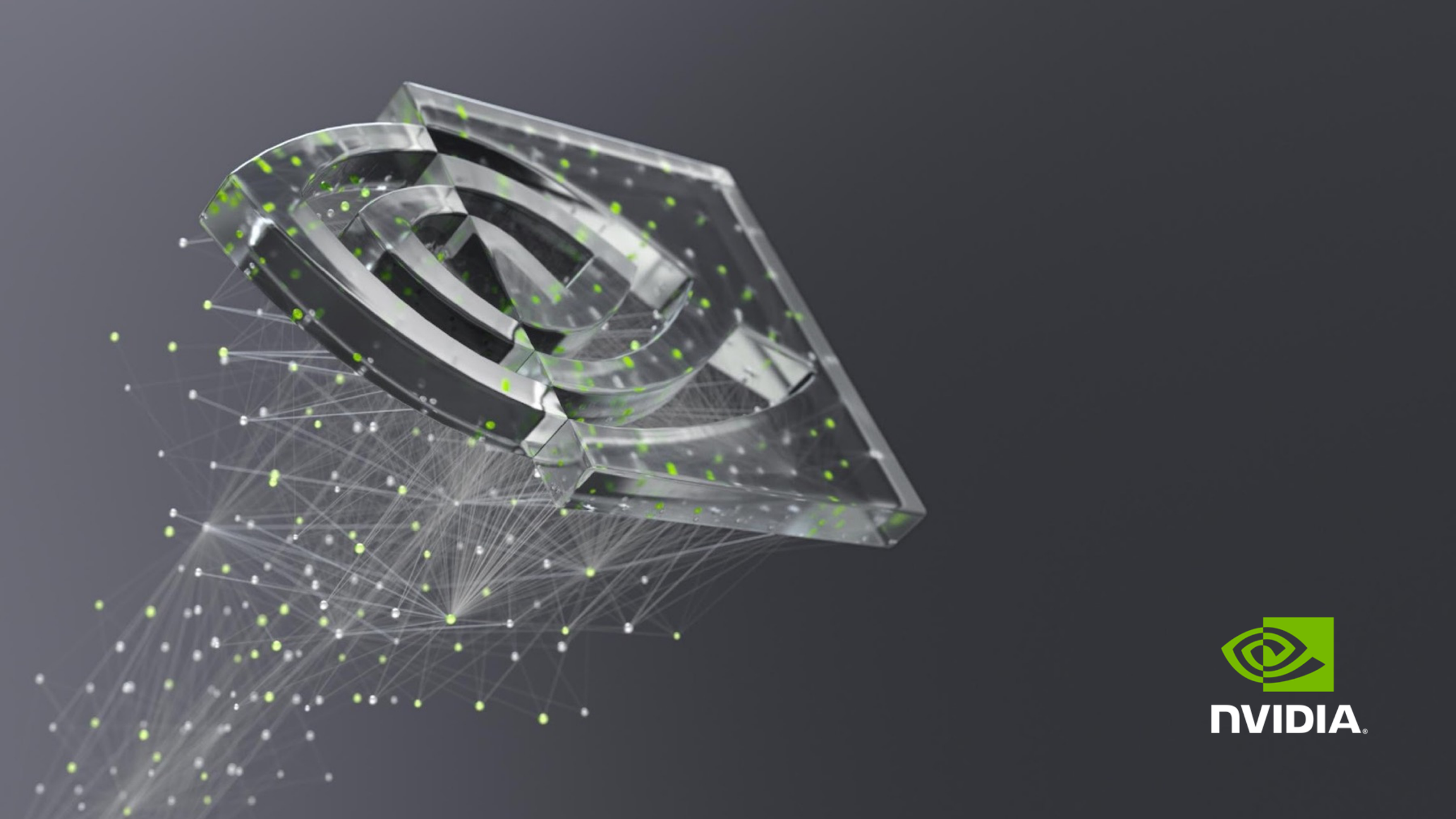

GPU


CPU


DPU

SUMMARY

- Tegra SoC has a long history, and that experience has been applied to current Xavier, the next generation Orin, and the next generation Atlan
- The future car is software defined, and NVIDIA provide whole ecosystem such as DRIVE Hyperion and DGX Systems
- Grace CPU is designed for giant-scale AI and HPC applications
- BlueField-3 DPU is the first 400 Gb/s data processing unit
- DOCA enables broad BlueField ecosystem
- GPU, CPU and DPU chips make a yearly leaps in one architecture


NVIDIA