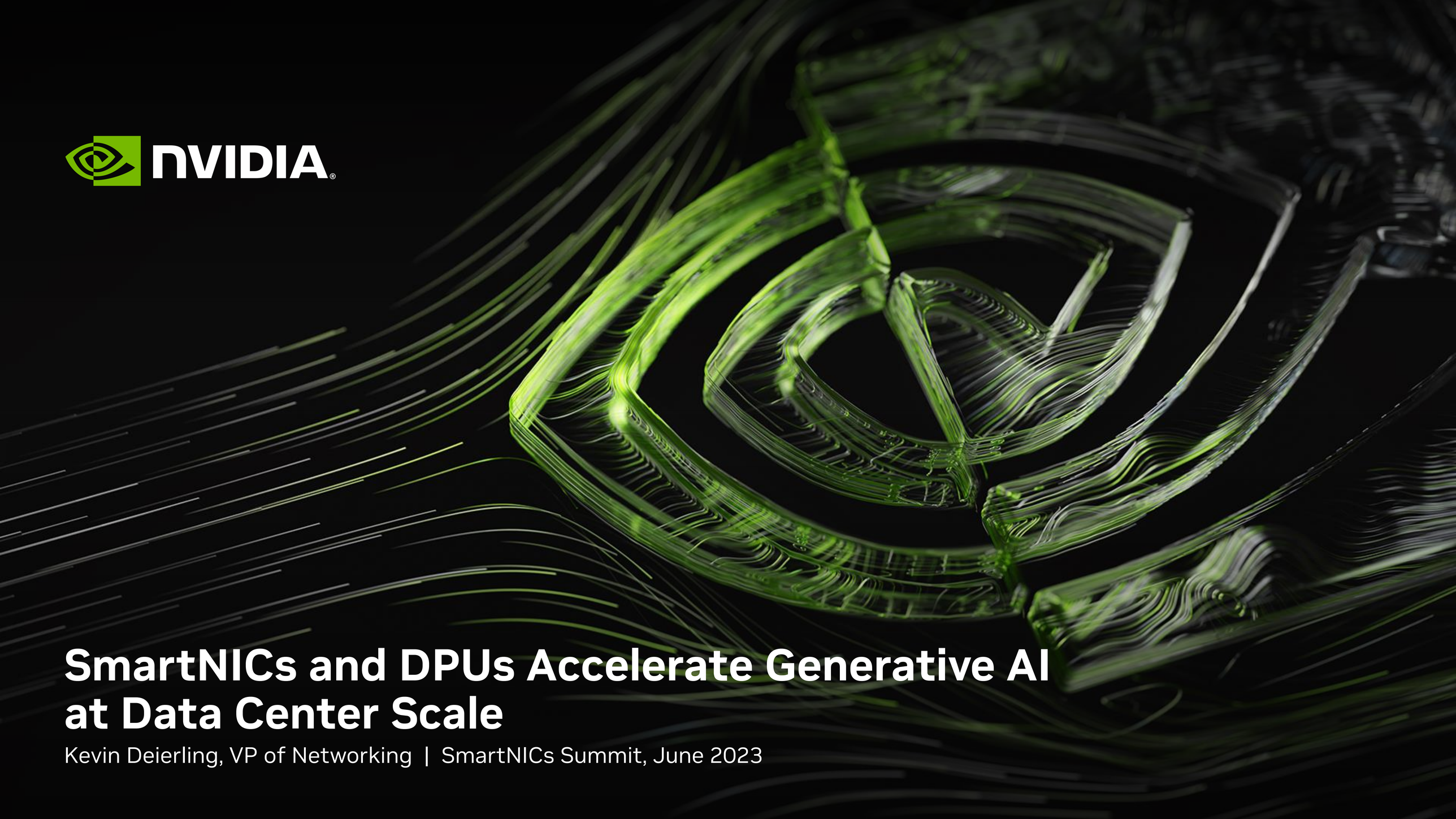
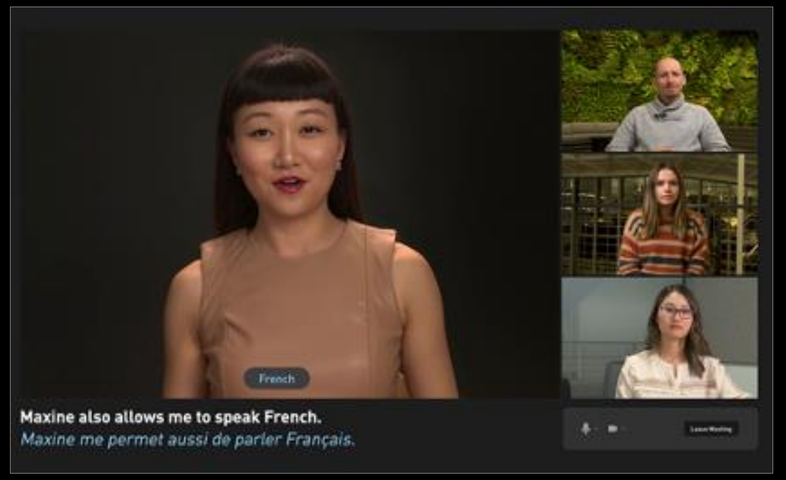
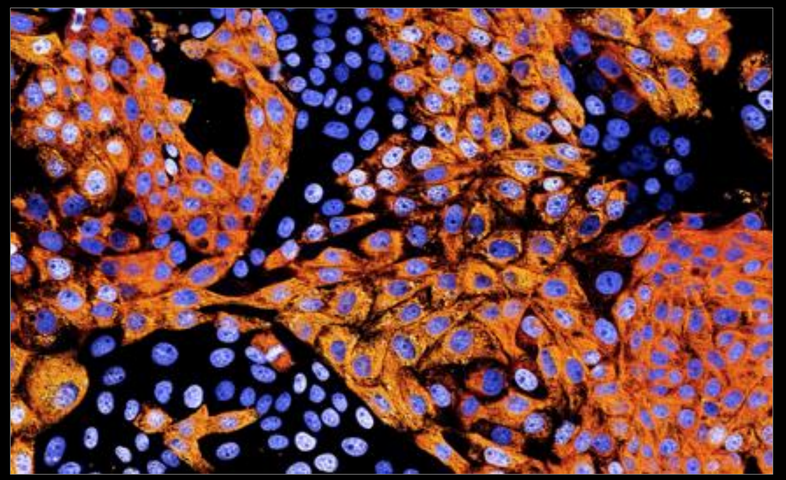


SmartNICs and DPUs Accelerate Generative AI at Data Center Scale

Kevin Deierling, VP of Networking | SmartNICs Summit, June 2023

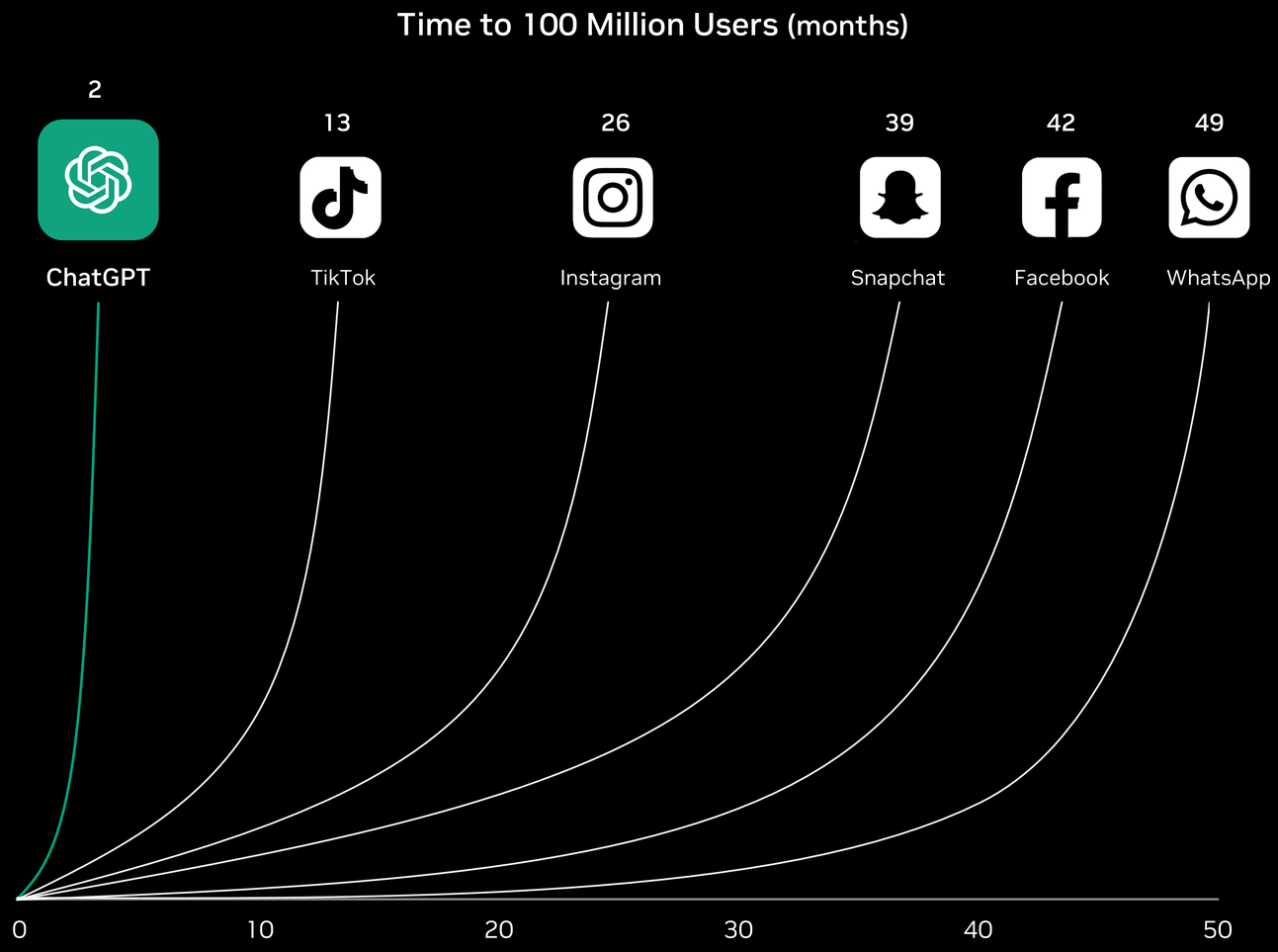
Democratizing AI Across Diverse Fields



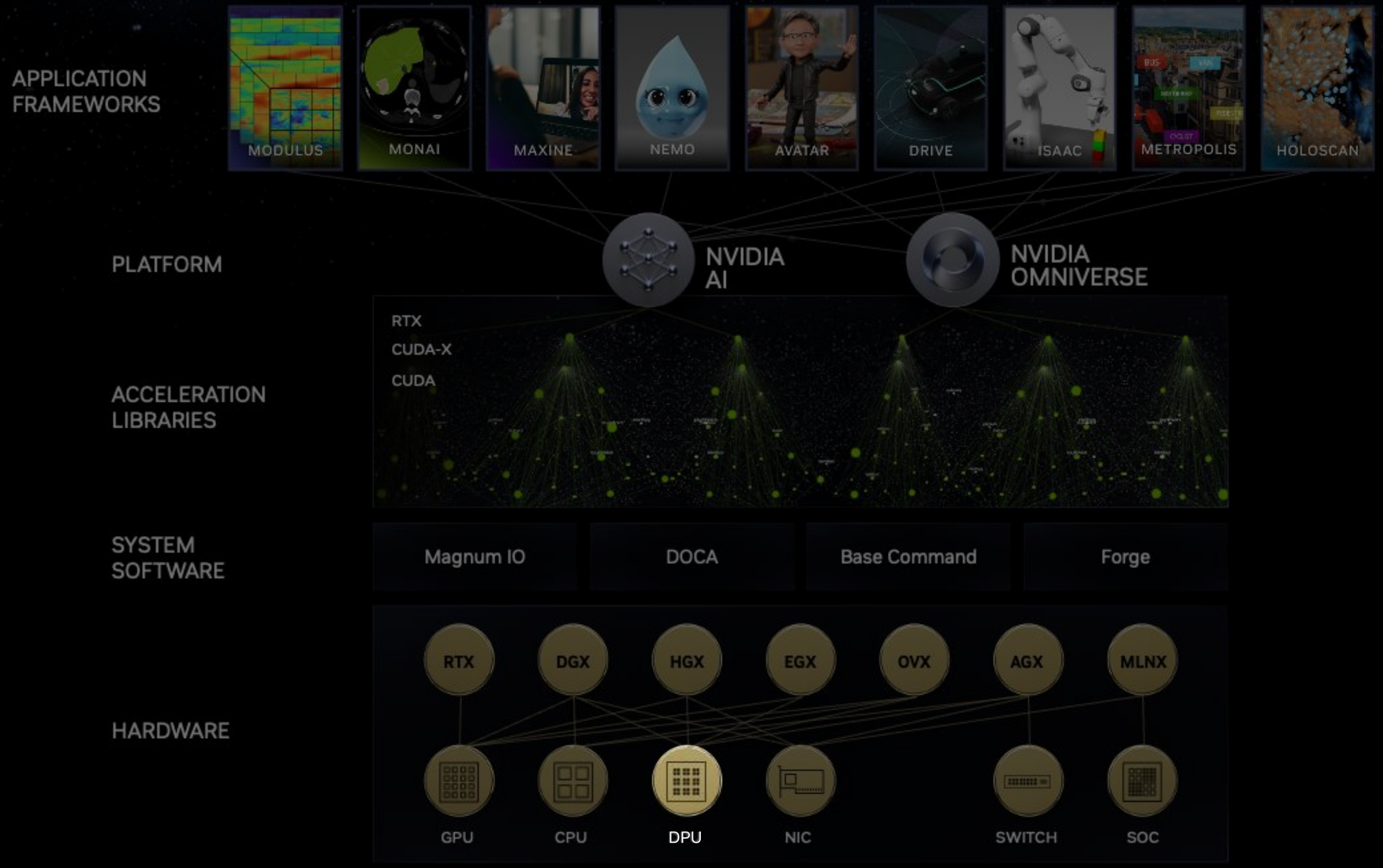


AI Workloads Accelerating Data Center Transformation

ChatGPT is the fastest-growing application in history




APPLICATION FRAMEWORKS
MODULUS
MONAI
MAXINE
NEMO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
PLATFORM
NVIDIA AI
NVIDIA OMNIVERSE
ACCELERATION LIBRARIES
RTX
CUDA-X
CUDA
SYSTEM SOFTWARE
Magnum IO
DOCA
Base Command
Forge
HARDWARE
RTX
DGX
HGX
EGX
OVX
AGX
MLNX
GPU
CPU
DPU
NIC
SWITCH
SOC

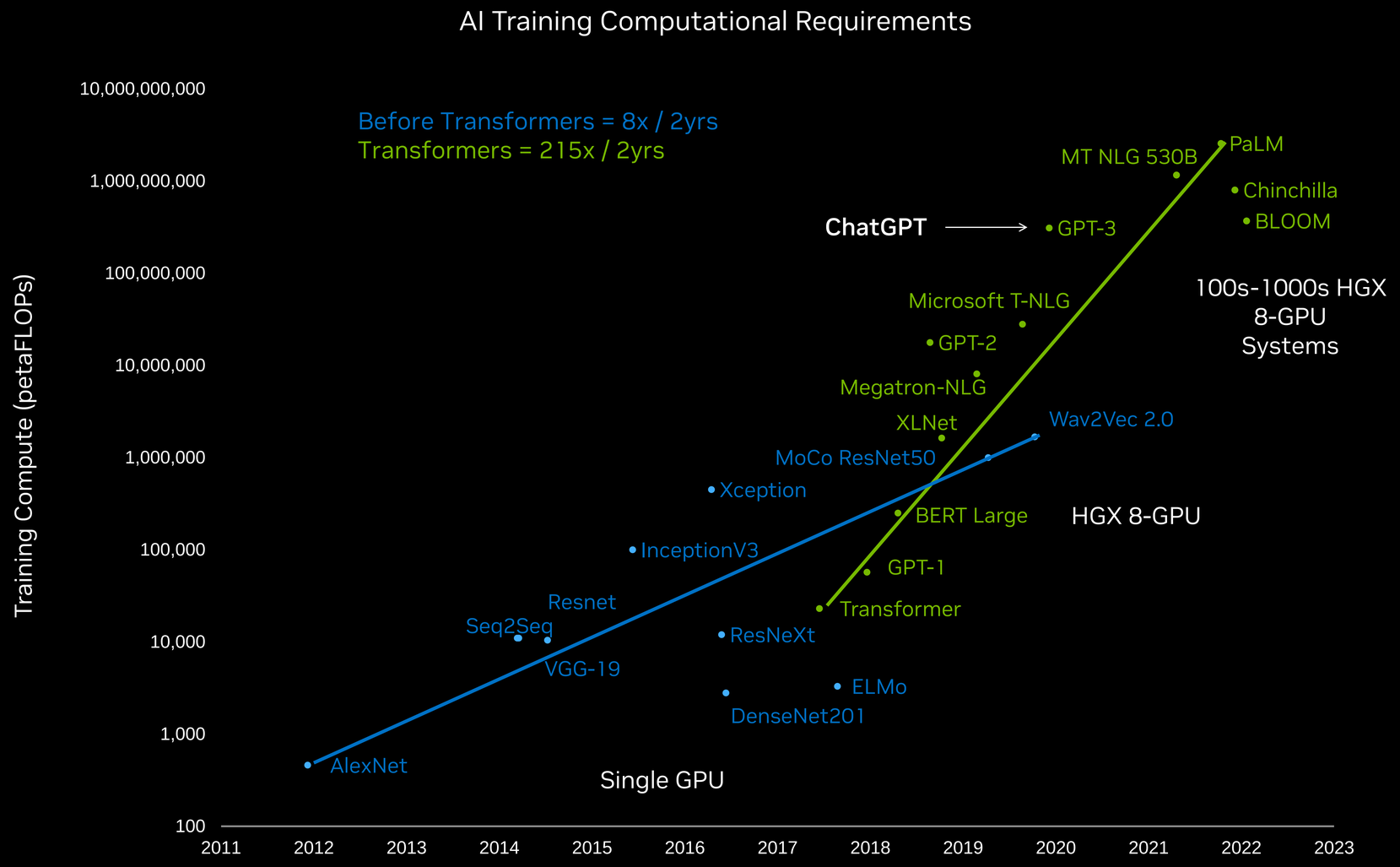

APPLICATION FRAMEWORKS
MODULUS
MONAI
MAXINE
NEMO
AVATAR
DRIVE
ISAAC
METROPOLIS
HOLOSCAN
PLATFORM
NVIDIA AI
NVIDIA OMNIVERSE
ACCELERATION LIBRARIES
RTX
CUDA-X
CUDA
SYSTEM SOFTWARE
Magnum IO
DOCA
Base Command
Forge
HARDWARE
RTX
DGX
HGX
EGX
OVX
AGX
MLNX
GPU
CPU
DPU
NIC
SWITCH
SOC

NVIDIA Full Stack Compute and Networking

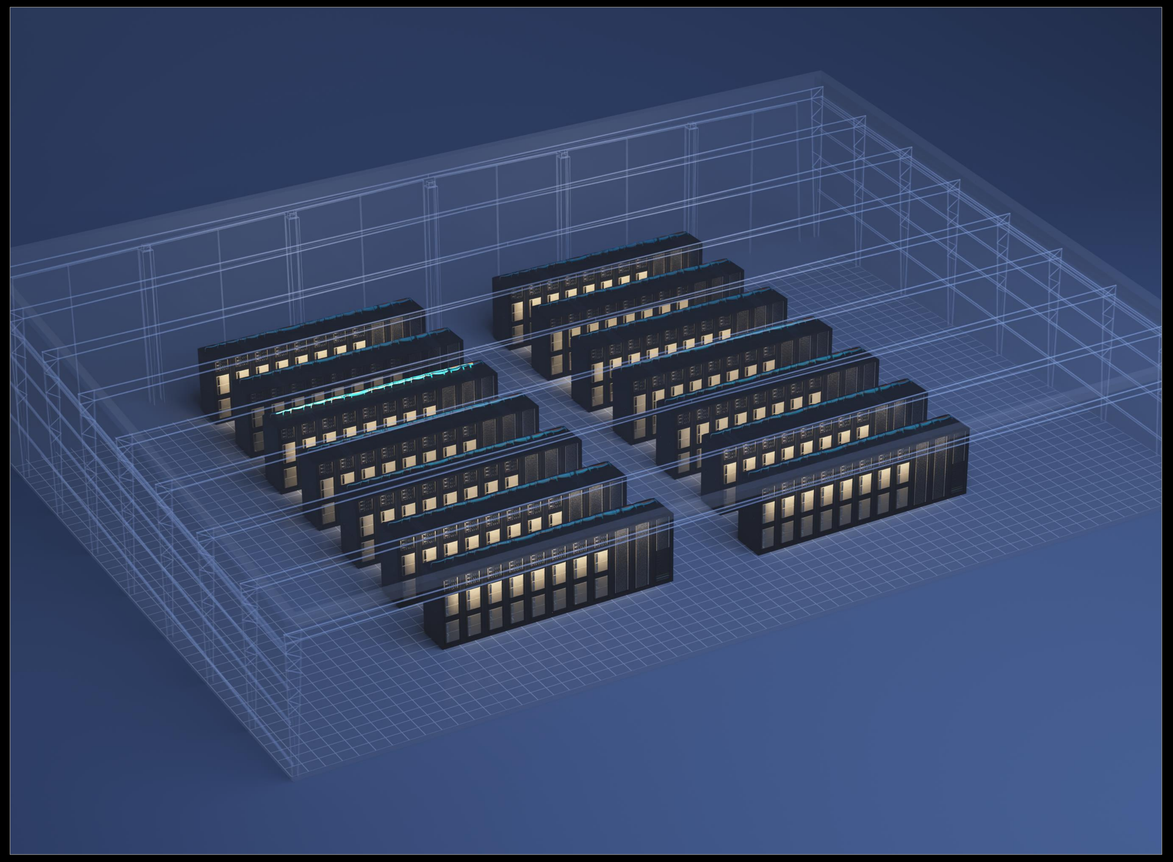
Fueling giant-scale AI infrastructure

Modern AI is a Data Center Scale Computing Workload

Data centers are becoming AI factories: data as input, intelligence as output



Networking for AI Data Centers



AI Factories

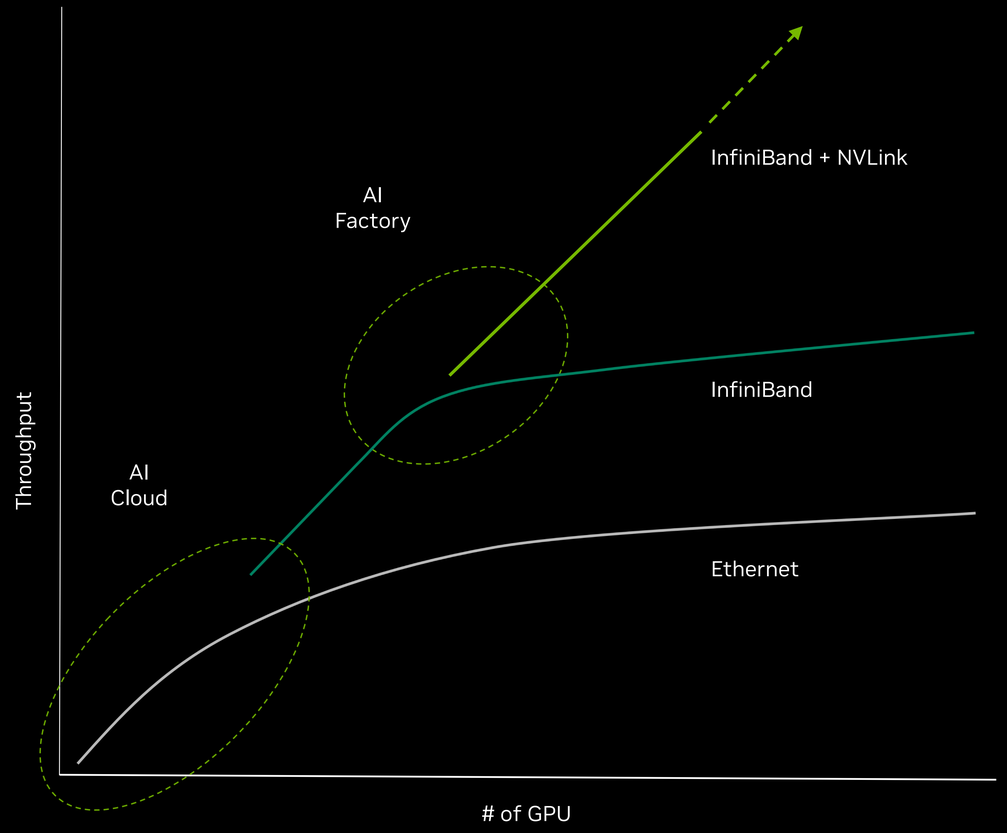
Single or few users | Extremely large AI models | NVLink and InfiniBand AI fabric

AI Cloud

Multi-tenant | Variety of workloads | Ethernet network

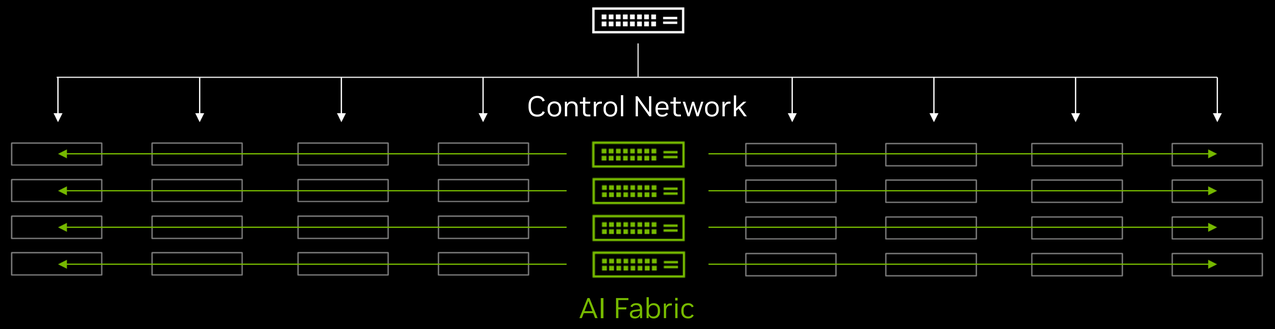
The Core of AI Factories – NVIDIA AI Compute Networking

AI Factories and Clouds Require Different Infrastructure Networking



AI Clouds Going Through A Major Change

Generative AI workloads require new class of Ethernet



Control / User Access Network (North-South)	AI Fabric (East-West)
Loosely Coupled Applications	Distributed Computing
TCP (Low Bandwidth Flows and Utilization)	RoCE (High Bandwidth Flows and Utilization)
High Jitter Tolerance	Low Jitter Tolerance
Oversubscribed Topologies	Performance Optimized Topologies
Heterogeneous Traffic Average Multi-Pathing	Bursty Network Capacity Predictive Performance

NVIDIA Spectrum-X Platform

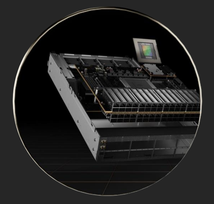
World's first high-performance Ethernet for AI

Full stack optimized for Generative AI clouds

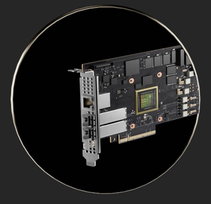
Spectrum-4 Ethernet Switch, BlueField-3 DPU

RoCE adaptive routing and performance isolation

End-to-end cloud provisioning and security



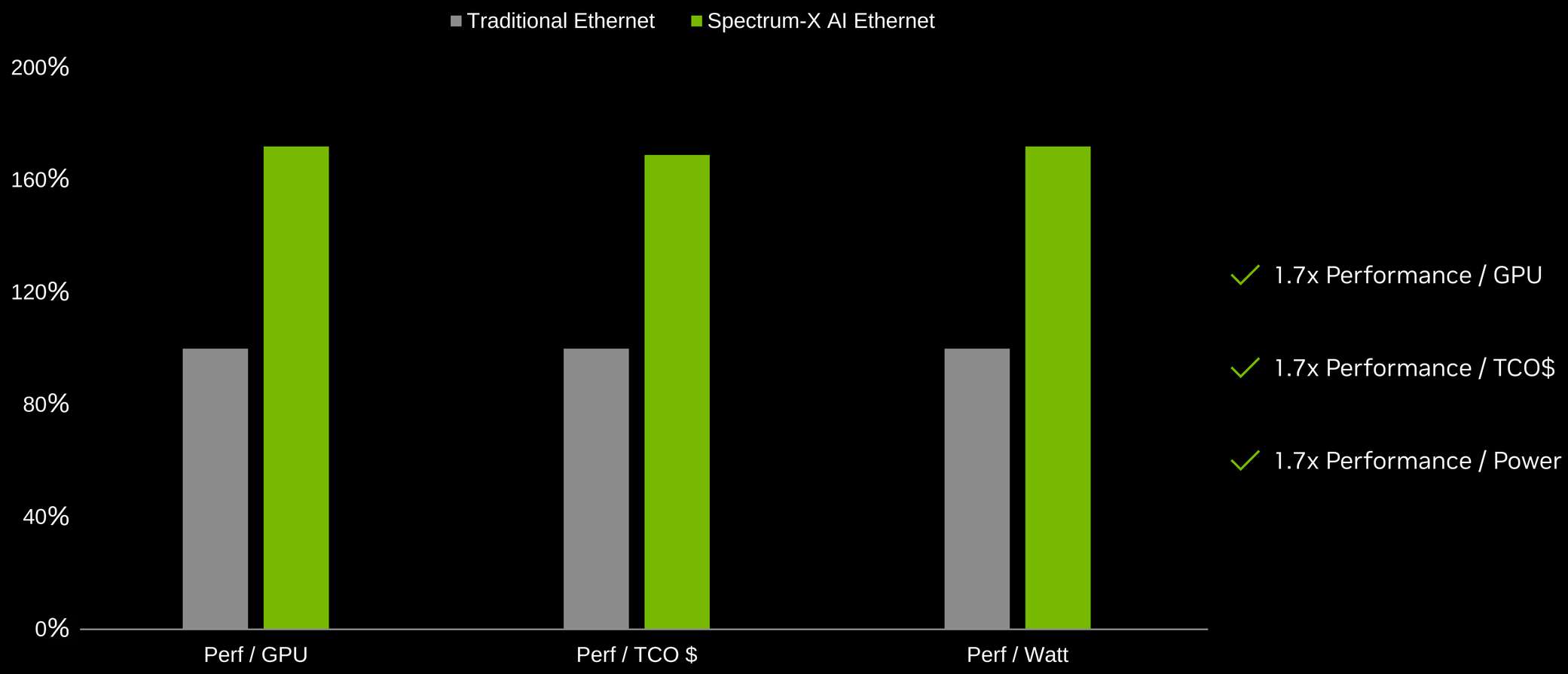
Spectrum-4 Ethernet Switch



BlueField-3 DPU

NVIDIA Spectrum-X Delivers the Highest GPT-3 Performance

400GbE/GPU

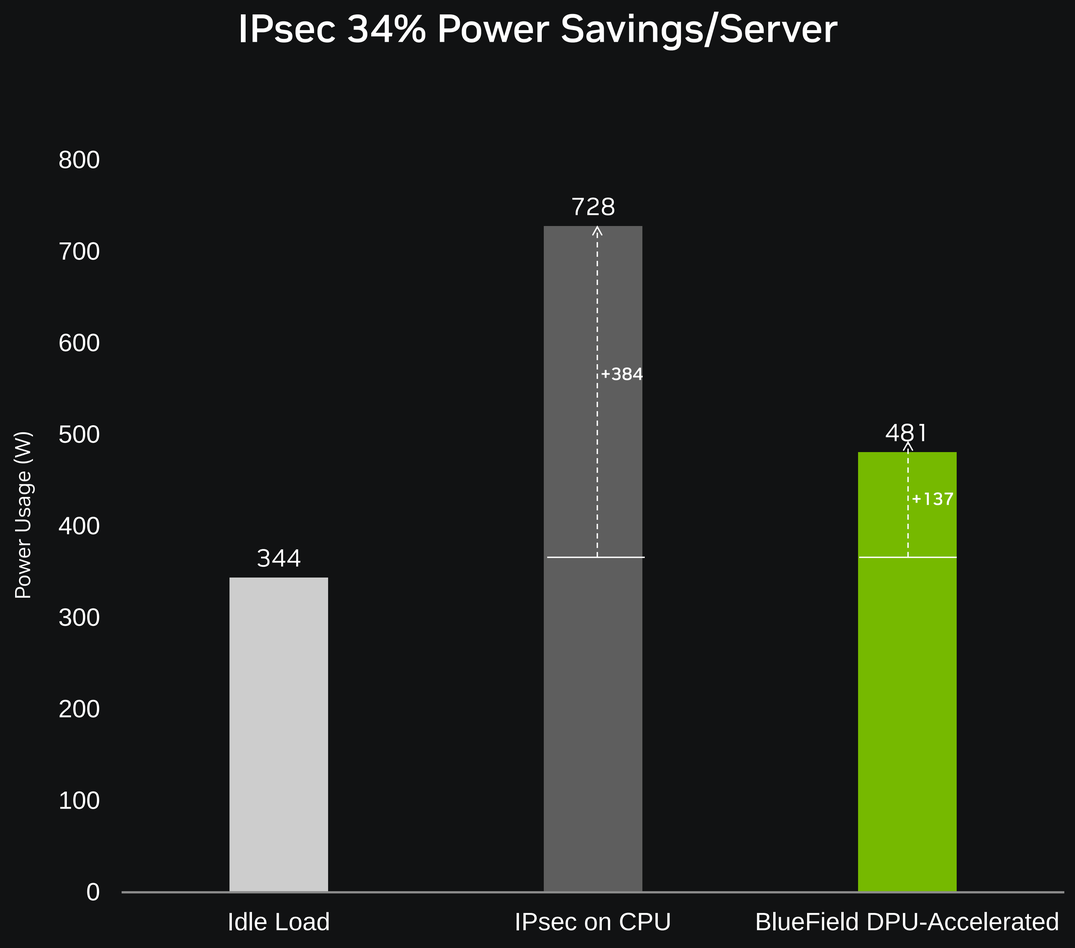


GPT3 175B – Performance per GPU, TCO$, and Watts with 16K H100 GPUs

Sustainable Cloud Computing

BlueField-3 enables power-efficient cloud data centers

- Nearly all data centers are power limited
- There are hard limits on power inputs in existing data centers
- Increasing electricity costs are becoming a long-term trend
- Need to get more out of your data centers
- Energy efficiency becomes a high priority
- BlueField DPU enables more while consuming less power



** Compared to idle load power consumption*

- OVS Networking: 29% Power Savings/Server
- Redis: 34% Less Power per Transaction

Israel-1

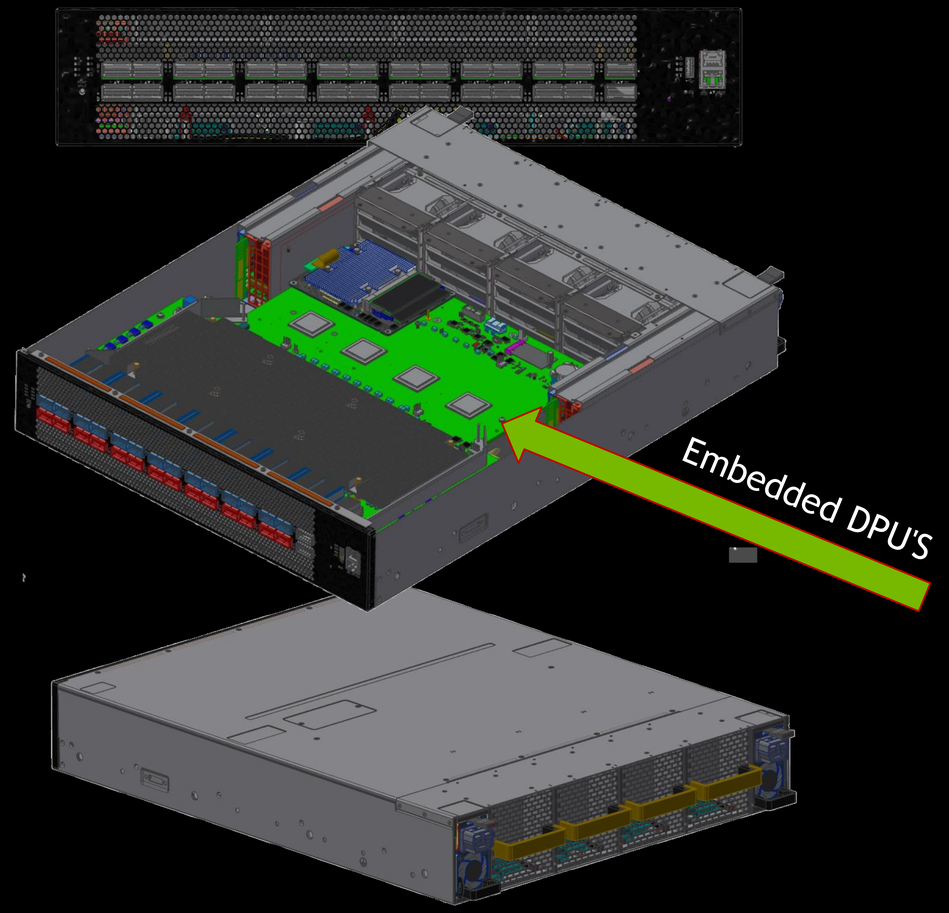
Hyperscale generative AI full stack performance optimization platform

256 Dell PowerEdge XE9680 Servers | 2048 H100 GPUs
80 Spectrum-4 Switches | 2560 BlueField-3 DPUs

Smart Switch

High Throughput Switch + Fully Programmable DPU Accelerated Networking

- FLEXIBLE SWITCH SYSTEM
- QSFP-DD
- SWITCH ASIC
- DPU'S == CPS/PPS
- Replaces Traditional T1 Switch



Rapidly Evolving BlueField Ecosystem

Unlocking the potential of the DPU

Cloud

Alibaba.com
Baidu 百度
Cirrascale
JD.COM 京东
Microsoft Azure
Oracle Cloud Infrastructure
OVHcloud
STACKPATH
UCLOUD 优刻得

Cybersecurity

Akamai
aria
Check Point Software Technologies Ltd.
Cloudflare
Custodio Technologies
Fortinet
f5
IMVISION
Juniper Networks
Nucleon
paloalto Networks
SentinelOne
Trend Micro
Versa Networks

Platform

Canonical
Nutanix
Red Hat
vmware

Storage

ddn
DELL EMC
IBM
nebulon
NetApp
VAST
WEKA


NVIDIA