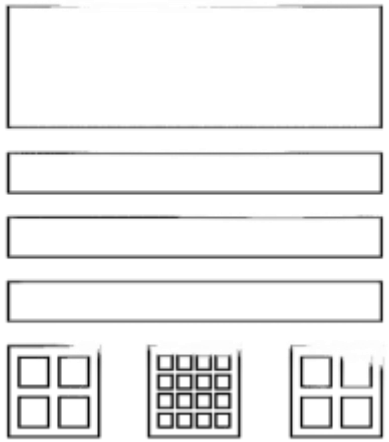
Accelerated Computing Infrastructure with NVIDIA

Dr. Pallavi Mohan

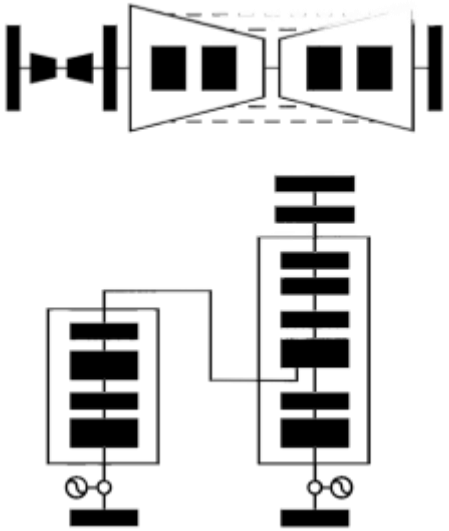
Senior Scientist & Solutions Architect, NVIDIA

Computer Industry Fundamental Transitions

ACCELERATED COMPUTING



GENERATIVE AI

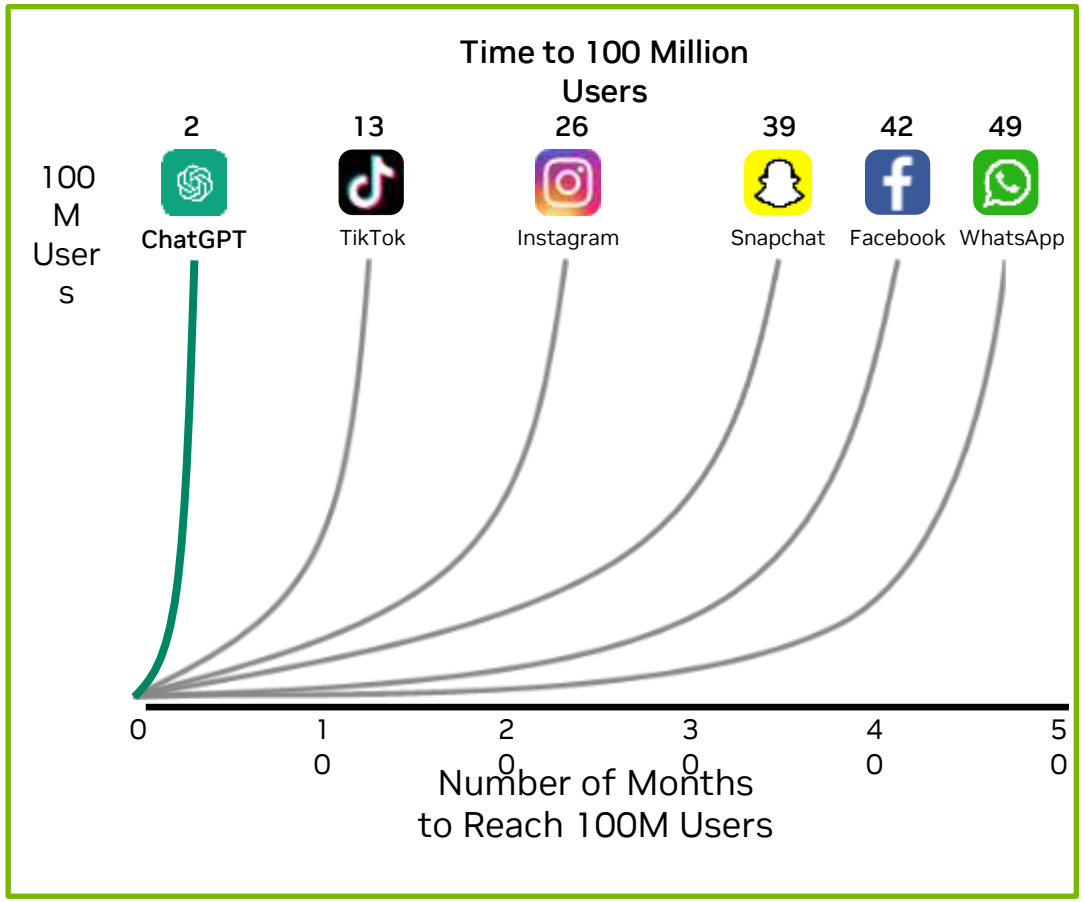


Modern Data Centers are Becoming AI Factories

Producing Intelligence from Data

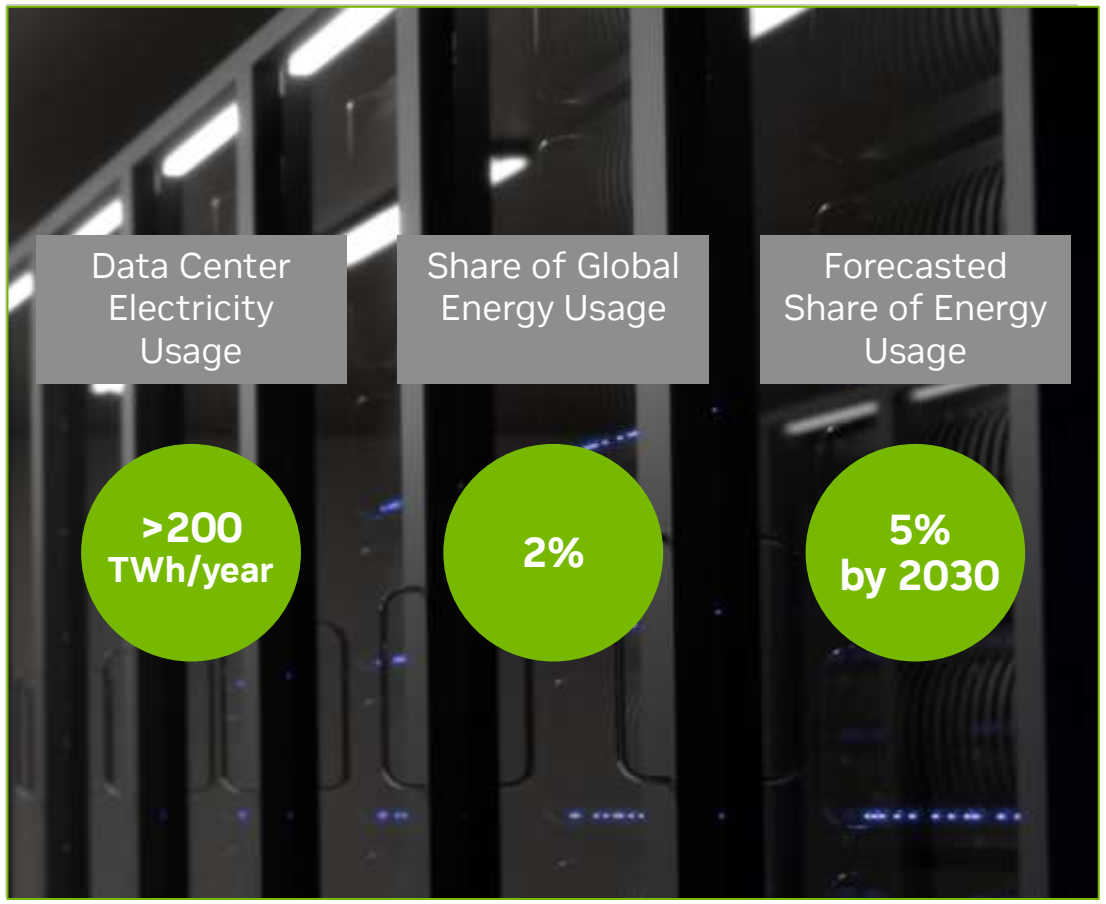
Key Data Center Trends

Demand for compute grows as data centers become power limited



Massive AI Models Drive New Use Cases

LLMs and GenAI Driving an Inflection Point

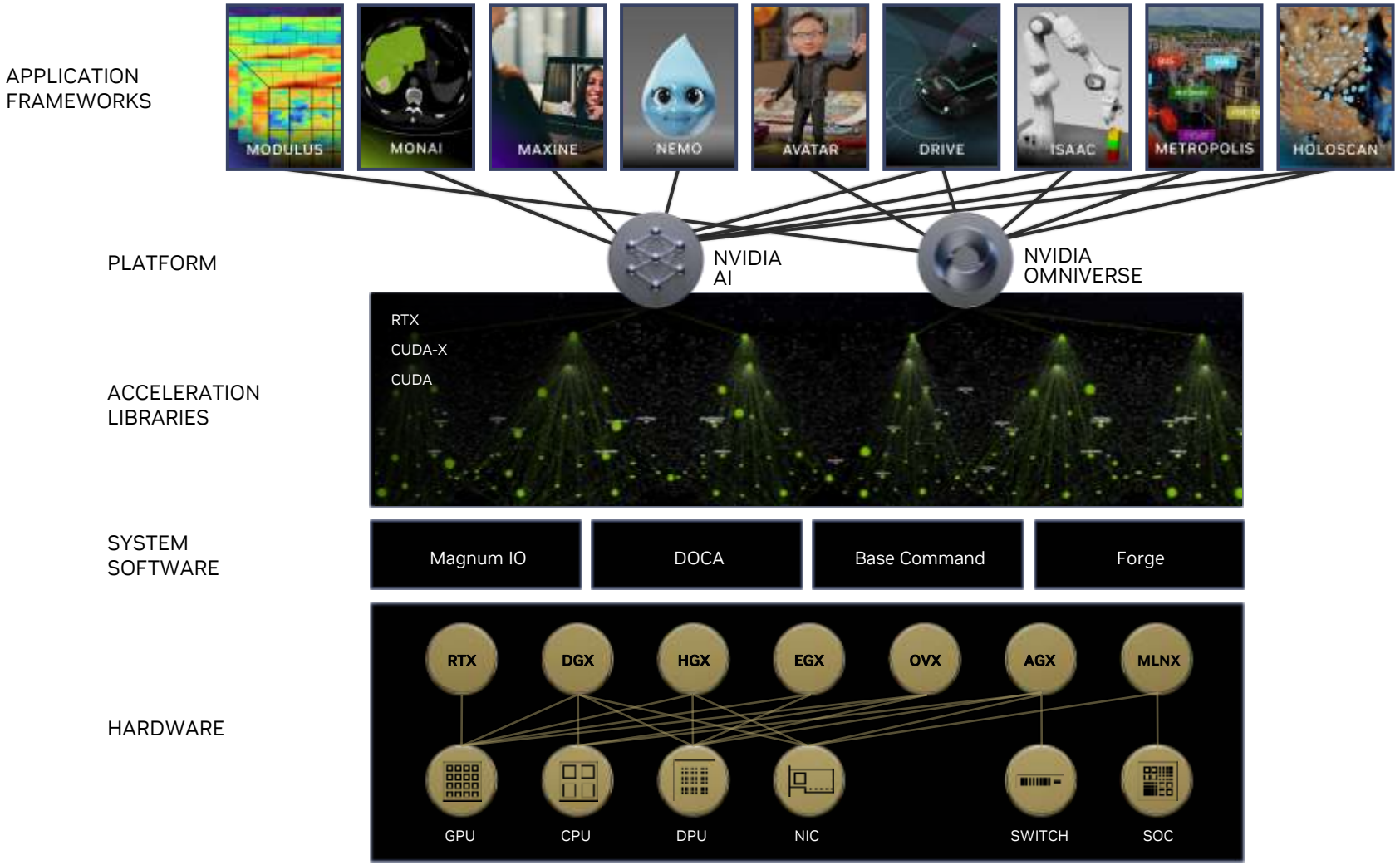


Data Centers are Power Limited

Need to Become More Efficient

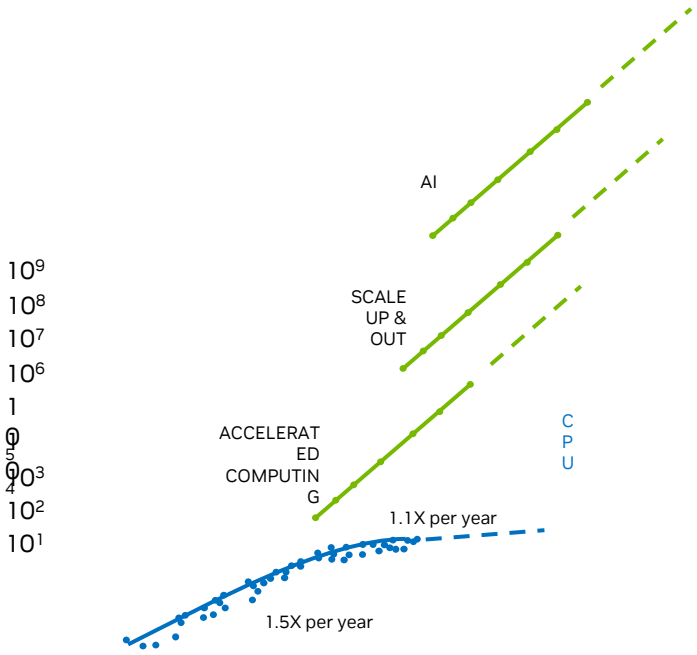
Accelerated Computing is the Path Forward

Accelerated Computing Services, Software and Systems Enabling New, Enhanced Business Models

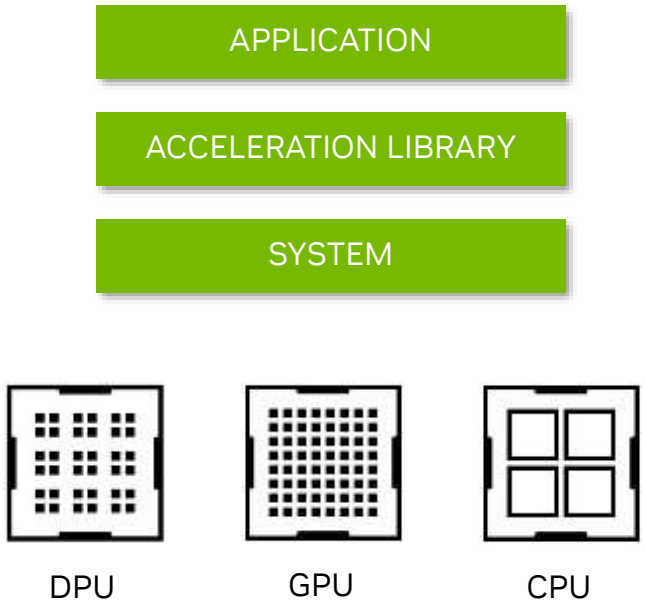


NVIDIA Accelerated Computing for Modern Data Centers

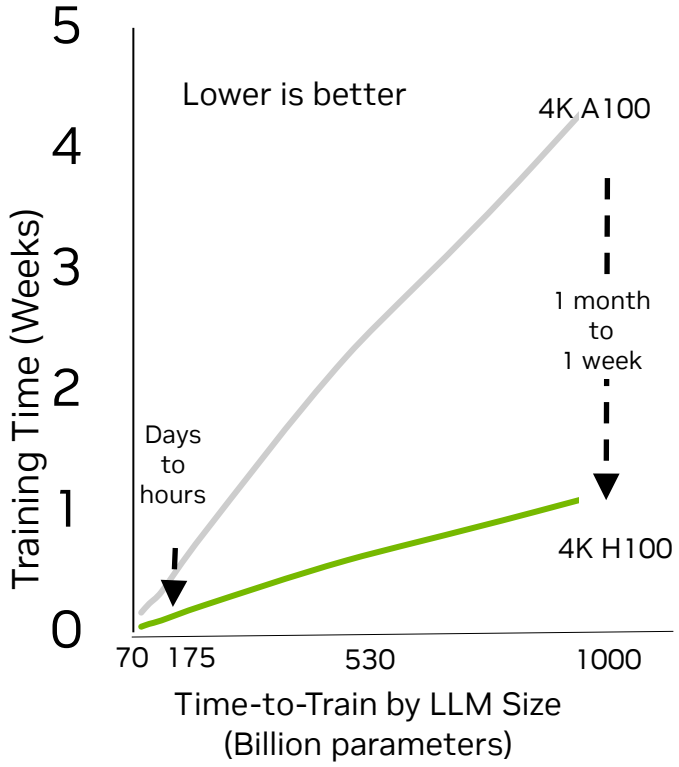
End of Moore's Law



Acceleration Takes a Full Stack



H100 Supercharges AI



NVIDIA DGX Platform

The best of NVIDIA AI—all in one place

NVIDIA DGX platform combines the best of NVIDIA software, infrastructure, and expertise in unified AI development solution that spans from the cloud to on-premises data centers.

.

Cloud

DGX Cloud
Multi-node AI training software as a service solution.

Software

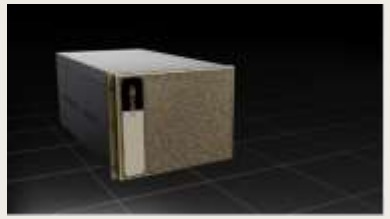
Base Command Platform
Centralized control of AI training projects across the DGX platform.

Base Command
The operating system of the DGX data center.

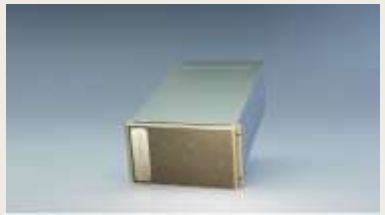
Clusters and Systems

DGX SuperPOD
Leadership-class AI infrastructure for on-premises and hybrid deployments.

DGX BasePOD
Proven reference architectures for AI infrastructure delivered with leading storage providers.



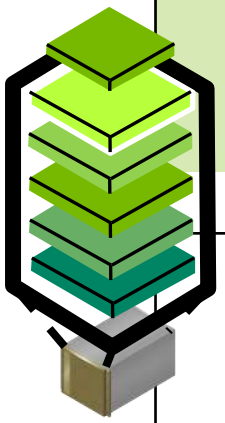
DGX H100
AI supercomputer optimized for large generative AI and other transformer-based workloads.

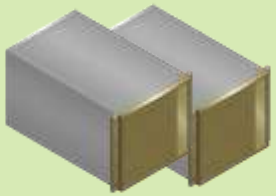


DGX A100
AI supercomputer delivering world-class performance for mainstream AI workloads.

The DGX Platform Powers Your AI Journey From End-to-End

Delivering incremental value for your AI initiatives, as your needs grow



	Day One your 1st DGX systems	Scaled Infrastructure DGX BasePOD	AI Center of Excellence DGX SuperPOD
DGX Systems powers every step in your AI journey			

Integrated software that powers AI innovation	NVIDIA Base Command
	• NVIDIA AI Enterprise:
	  • Pre-trained models, optimized frameworks
	  • Customize/fine-tune pre-trained models
	  • Optimize/accelerate inference
	• Kubernetes or Slurm scheduling
	• Add/manage DGX within your existing compute infrastructure (cloud, non-GPU)
	• Accelerate storage & network IO
	• Fully optimized OS stack
			Base Command Premium
			In addition to features on the left:
			• NVIDIA Base Command Platform:
			  • Simplify developer workflow
			  • Dataset management
			  • Batch processing
			  • Monitoring

Continuous roadmap of innovative features delivered to customers

DGX Cloud - The Cloud-First Way to Get Access to Your Own DGX AI Supercomputer

NVIDIA DGX GH200: The Trillion Parameter Instrument of AI

Massive memory supercomputing for emerging AI

World's first system built with NVIDIA NVLink Switch System

- Nearly **500X** more system memory
- **48X** GPU-to-GPU bandwidth
- **7X** CPU-to-GPU bandwidth
- **5X** interconnect power efficiency

256 Grace Hopper Superchips | **1EFLOPS** AI Performance | **144TB** unified fast memory
36 L2 NVLink switches | **900 GB/s** GPU-to-GPU bandwidth | **128 TB/s** bisection bandwidth

Available year-end 2023

NVIDIA BlueField DPU Platform

Software-Defined, Hardware-Accelerated Infrastructure Compute Platform

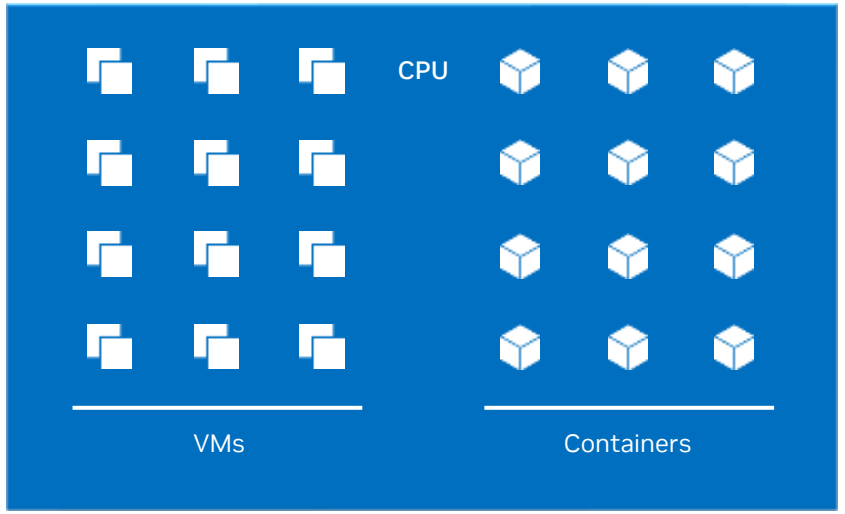
Accelerated Performance
Meet the most stringent performance requirements, run the most demanding workloads

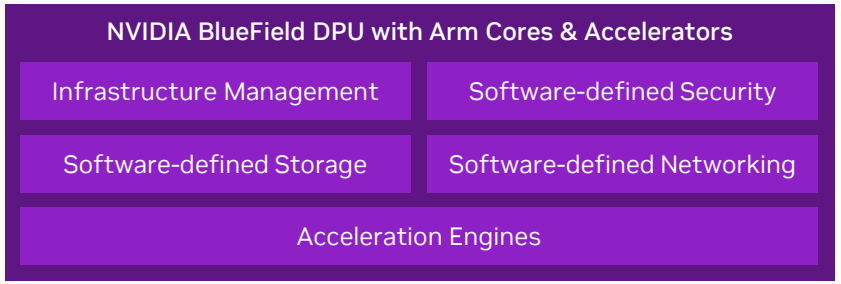
Cloud-Scale Efficiency
Free up x86 cores to business apps, achieve unprecedented scale and efficiency levels

Robust Zero-Trust Security
Ensure comprehensive data center security without compromising performance

Programmable Infrastructure
Develop and run applications consistently with maximum performance

DPU ACCELERATED SERVER





Offload | Accelerate | Isolate

NVIDIA BlueField-3 Overview

Massive Advancements, Built for Cloud Scale

400Gb Networking
2X Network Bandwidth
2X Network Pipeline
4X Host Bandwidth

Programmable Engines
4X Arm Compute
5X Memory
New Datapath Accelerator

Zero-Trust Security
4X IPsec Acceleration
2X TLS Acceleration
New MACsec Acceleration

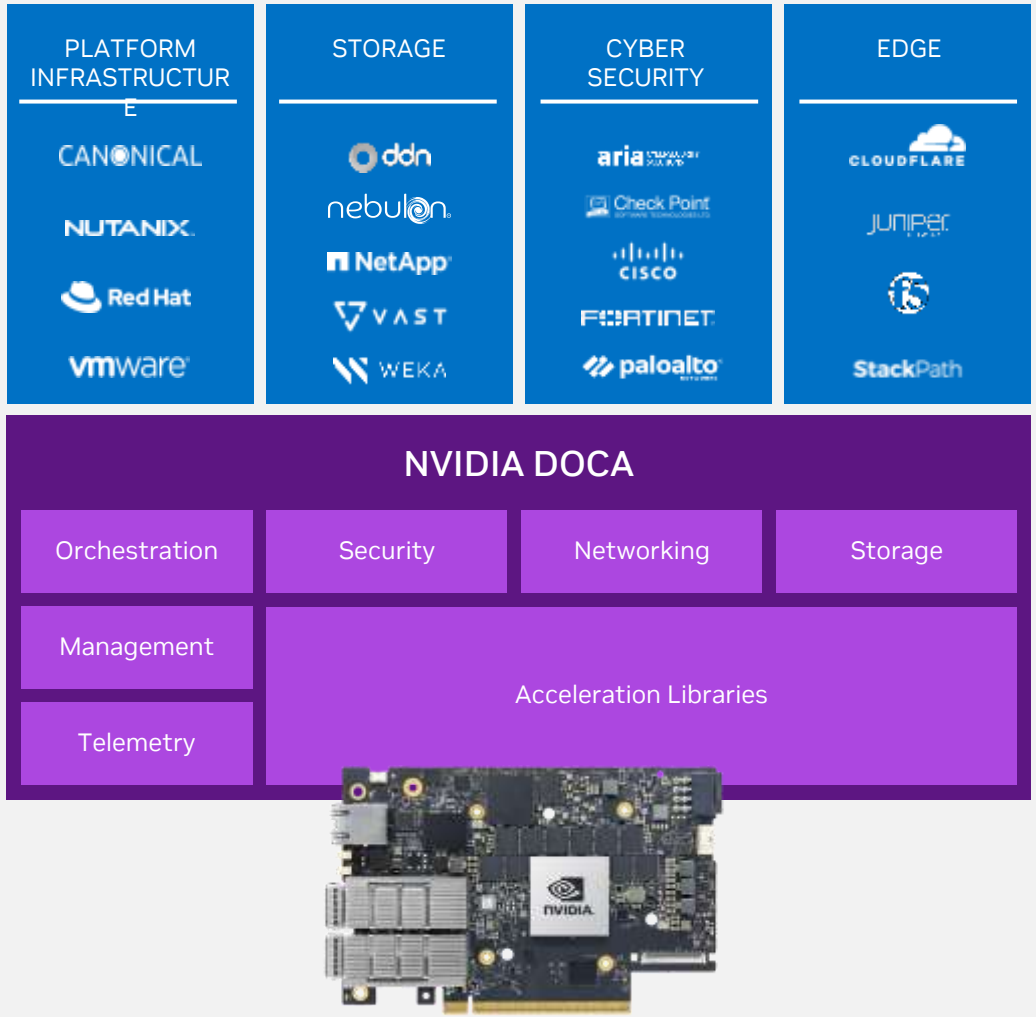
Composable Storage
2X Storage IOPs
2X Storage Encryption
New NVMe/TCP Acceleration

** Compared to previous BlueField generation*

NVIDIA DOCA

Comprehensive Acceleration SDK for BlueField DPUs

- Unified software 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
- Rich partner ecosystem



BlueField DPU

BlueField Powers NVIDIA-Accelerated Computing Systems

Full-Stack, Data Center-Scale, Multi-Domain Acceleration

Generative AI

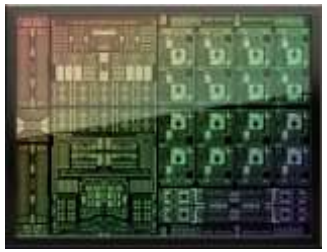
Scientific Computing

5G Networks

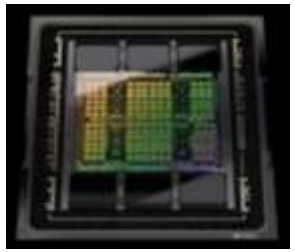
Distributed Database

Internet Services

DOCA ⟷ CUDA



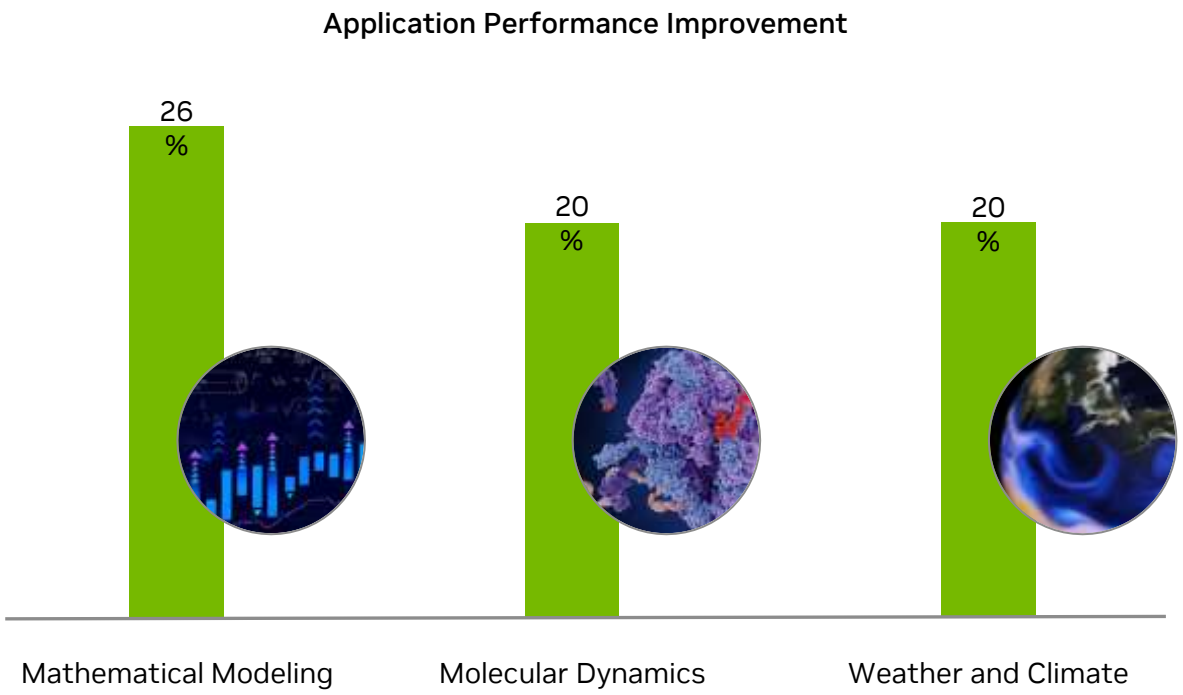
BlueField-3 DPU



H100 GPU

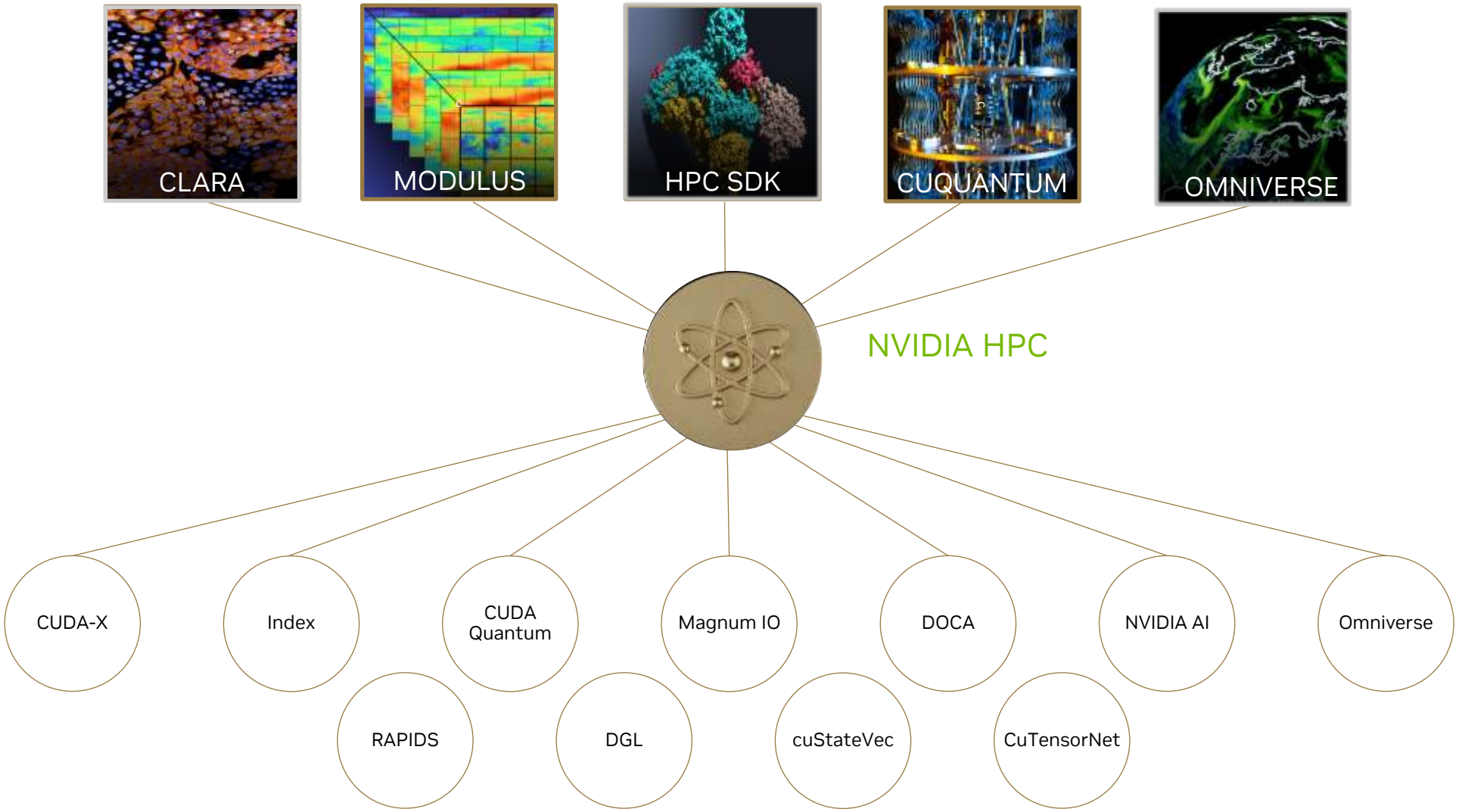
Accelerating Scientific Computing Workloads

Ignite High-Performance Computing with NVIDIA BlueField and Quantum InfiniBand



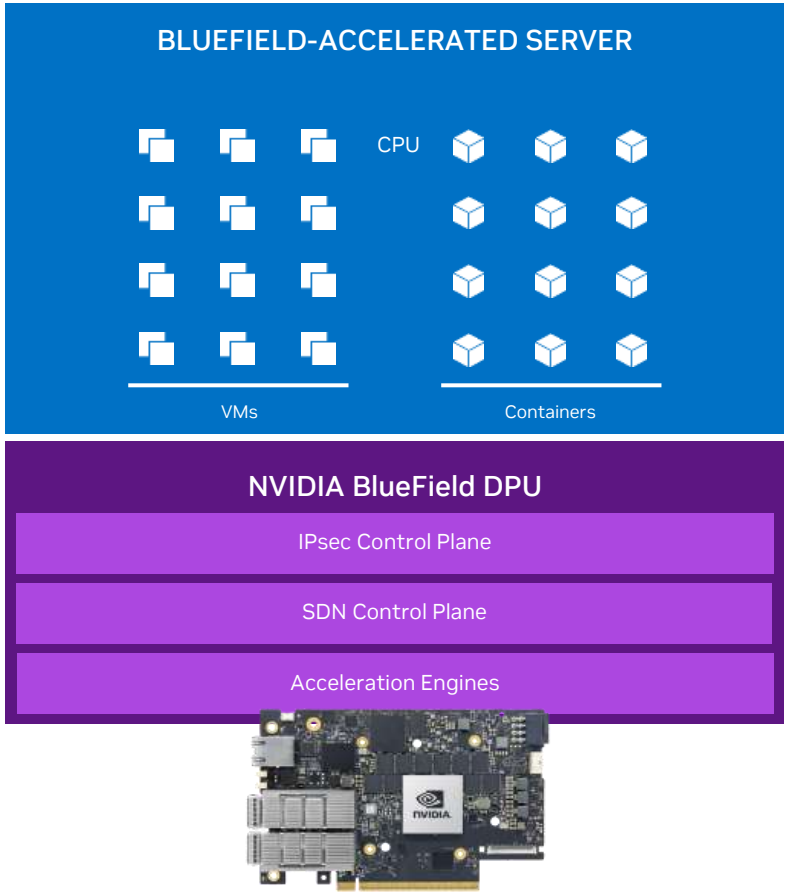
- Unleash application performance and system efficiency
- MPI performance acceleration
- Computational storage and advanced workloads
- Adaptive performance isolation

NVIDIA HPC Platform



Accelerated Computing is Sustainable Computing

BlueField-3 Enables Power-Efficient Cloud Data Centers



BlueField-3

fin