Latest Trends and Innovation in AI

Carlo Nardone, SAE EMEA

Vertiv Engineers' Frontier Master Class, Castel Guelfo (BO) 2024/11/19

The Evolution of Deep Learning-based AI

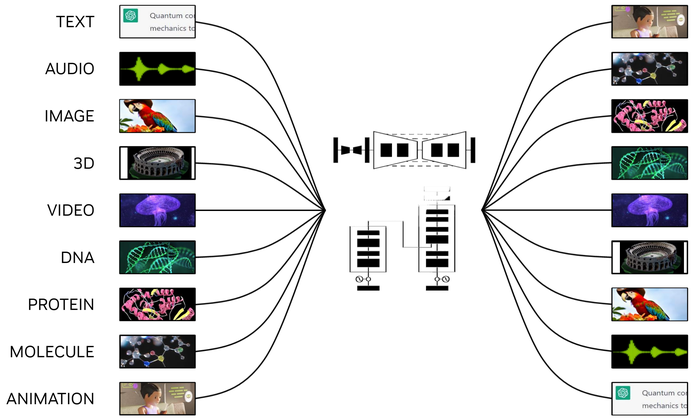
From Supervised Learning to Transformer-based Generative AI

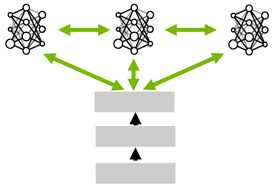
Image Classification

Transformer

Large Language Models (Transformer)

Large Language Models (Transformer)



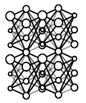


Google Gemini

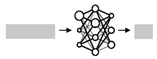
Meta NLLB

MISTRAL AI_ Mixtral

Realtime
<50ms latency



Parameters
>10T

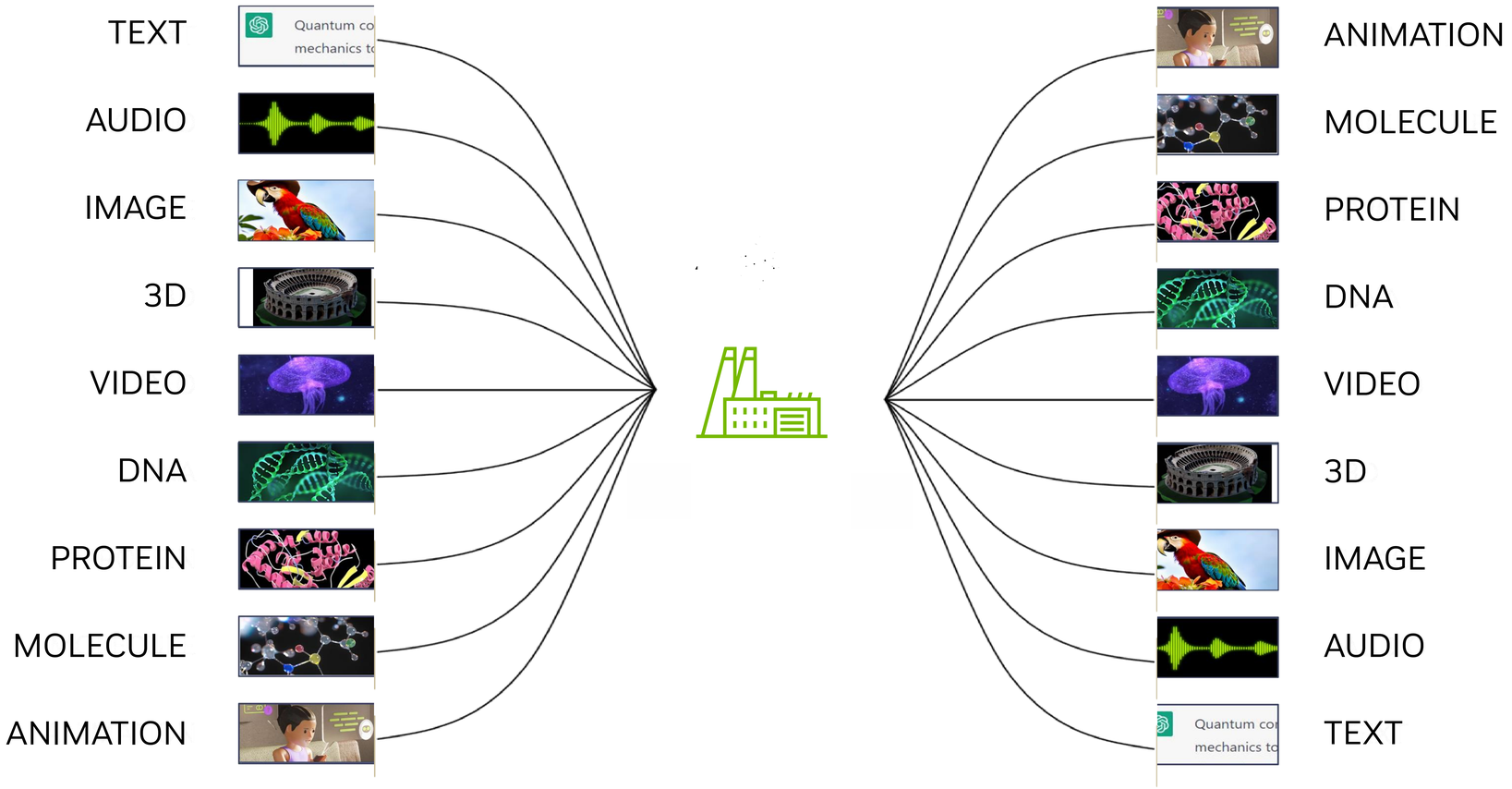


Sequence Length
>32K input

Labeled Datasets — Unlabeled Datasets — Generative AI — Multimodal Generative AI — Mixture of Experts (MoE) — Production GenAI Inference

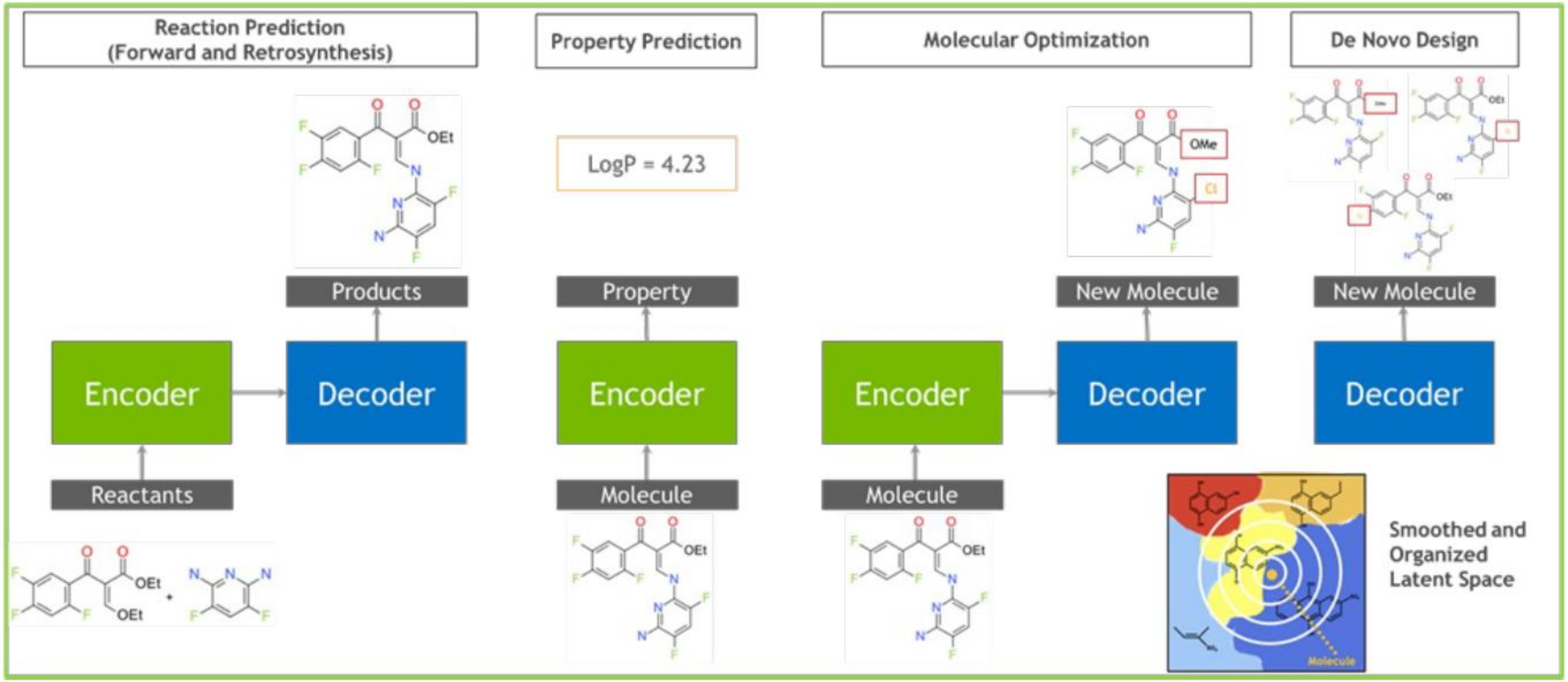
The Next Era of Generative AI

AI factories unlock $100T industries



A non-textual GenAI example: Chemistry / Drug Discovery

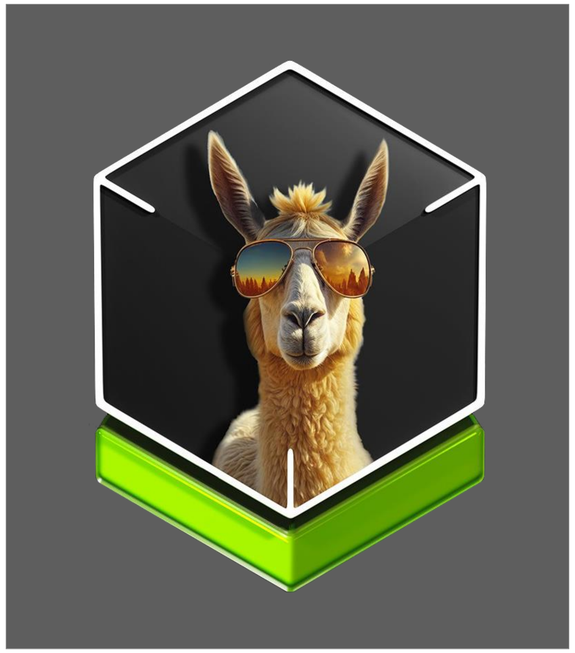
MegaMolBart



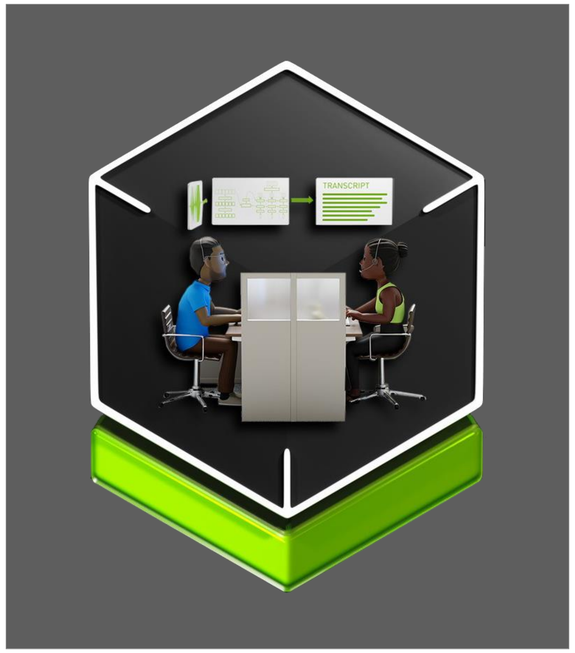
New Era of Generative AI

Unlocking unprecedented levels of productivity

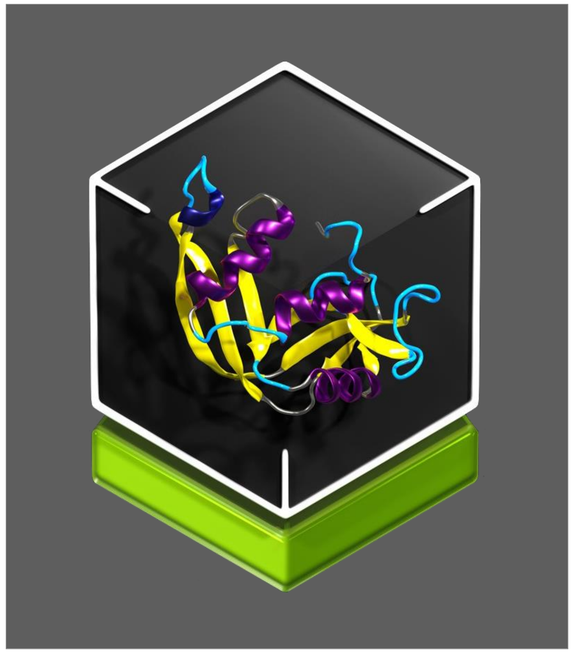
Customer Experience
Customer Self-Service
Agent Experiences



Content Creation
Personalization Domain
Specific Summarization



Software Engineering
Coding Assistant



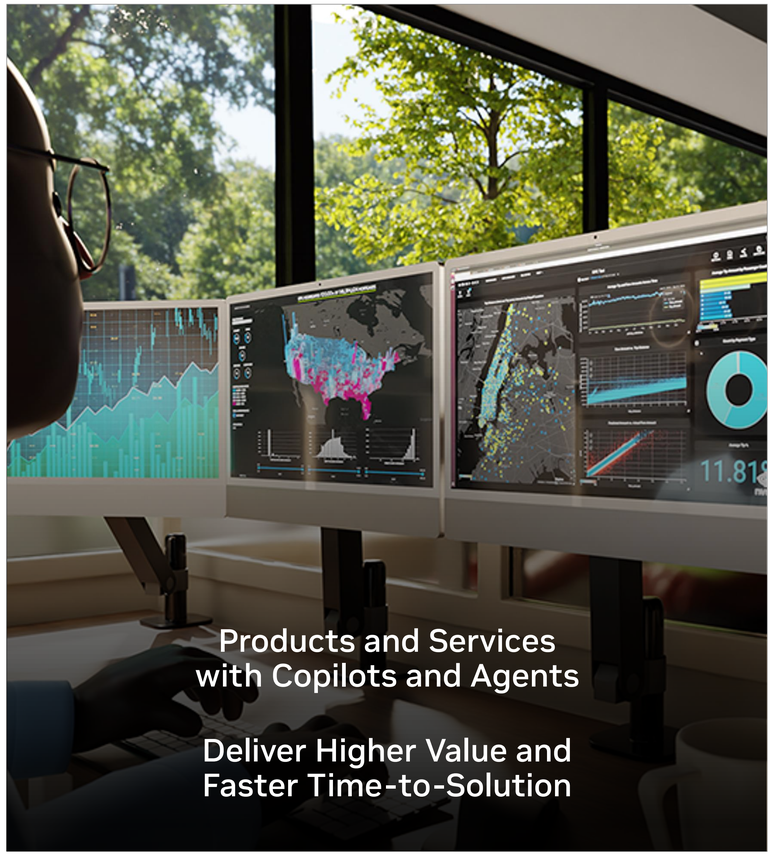
Product R&D
Enhanced Design
Simulation and Testing

"...generative AI has the potential to generate $2.6 trillion to $4.4 trillion in value across industries."

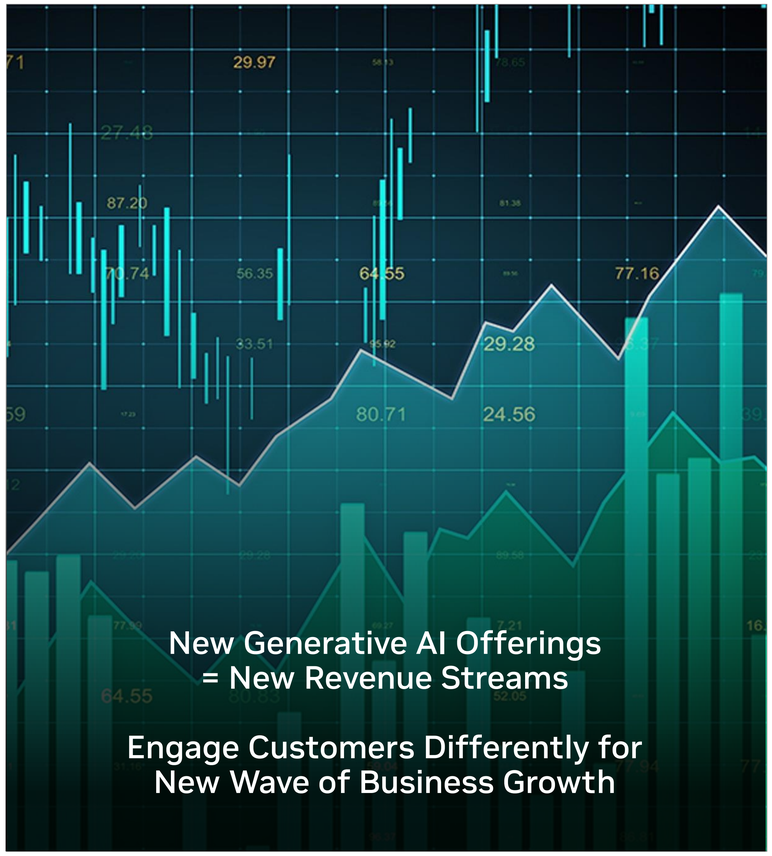
— McKinsey Digital, "The Economic Potential of Generative AI: The Next Productivity Frontier" 2023

Generative AI Helping Enterprises Grow

Operational Efficiency
NVIDIA ChipNeMo



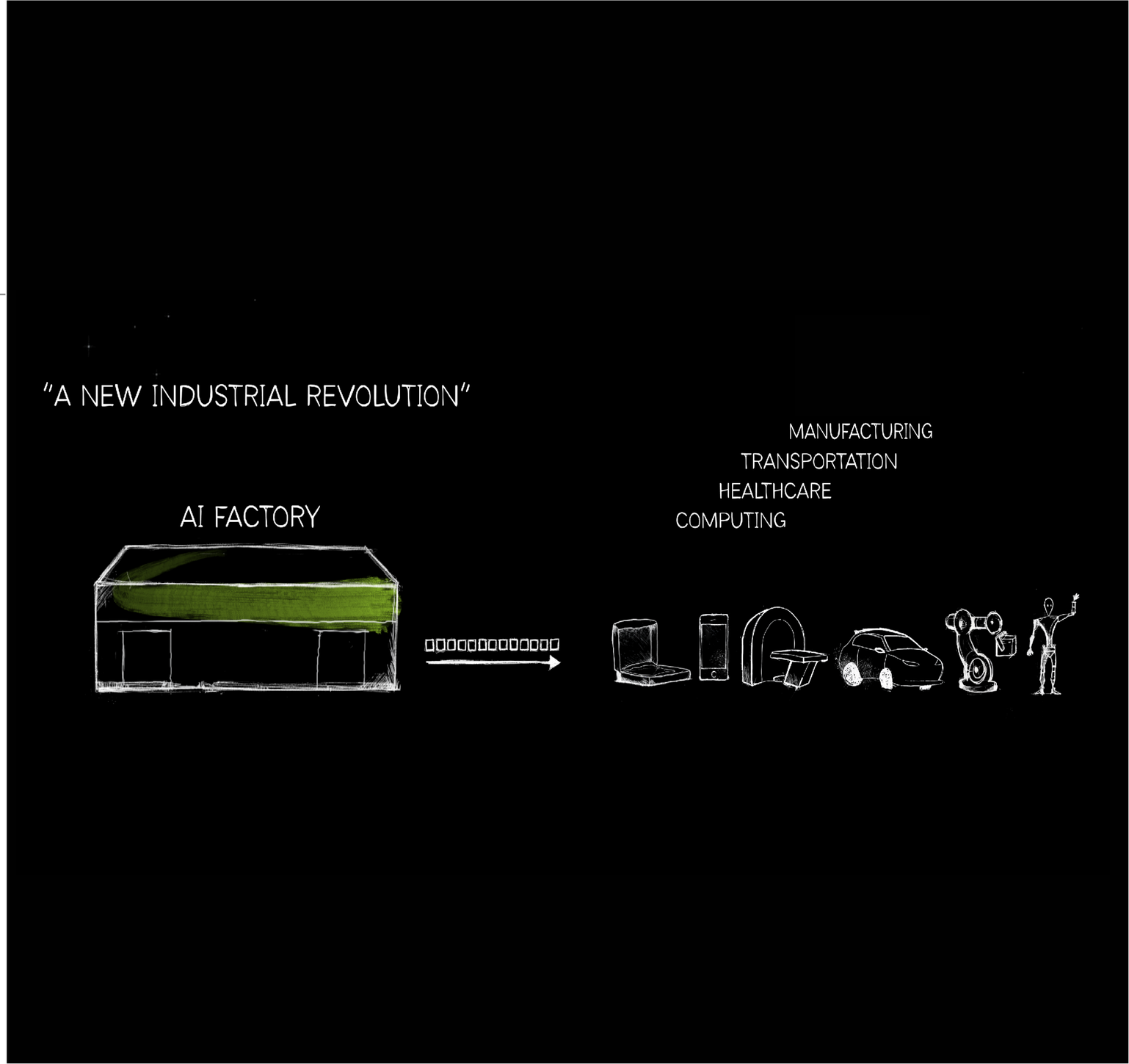
Productivity
Joule's SAP Consulting Services



Expand Business
ServiceNow Now Assist

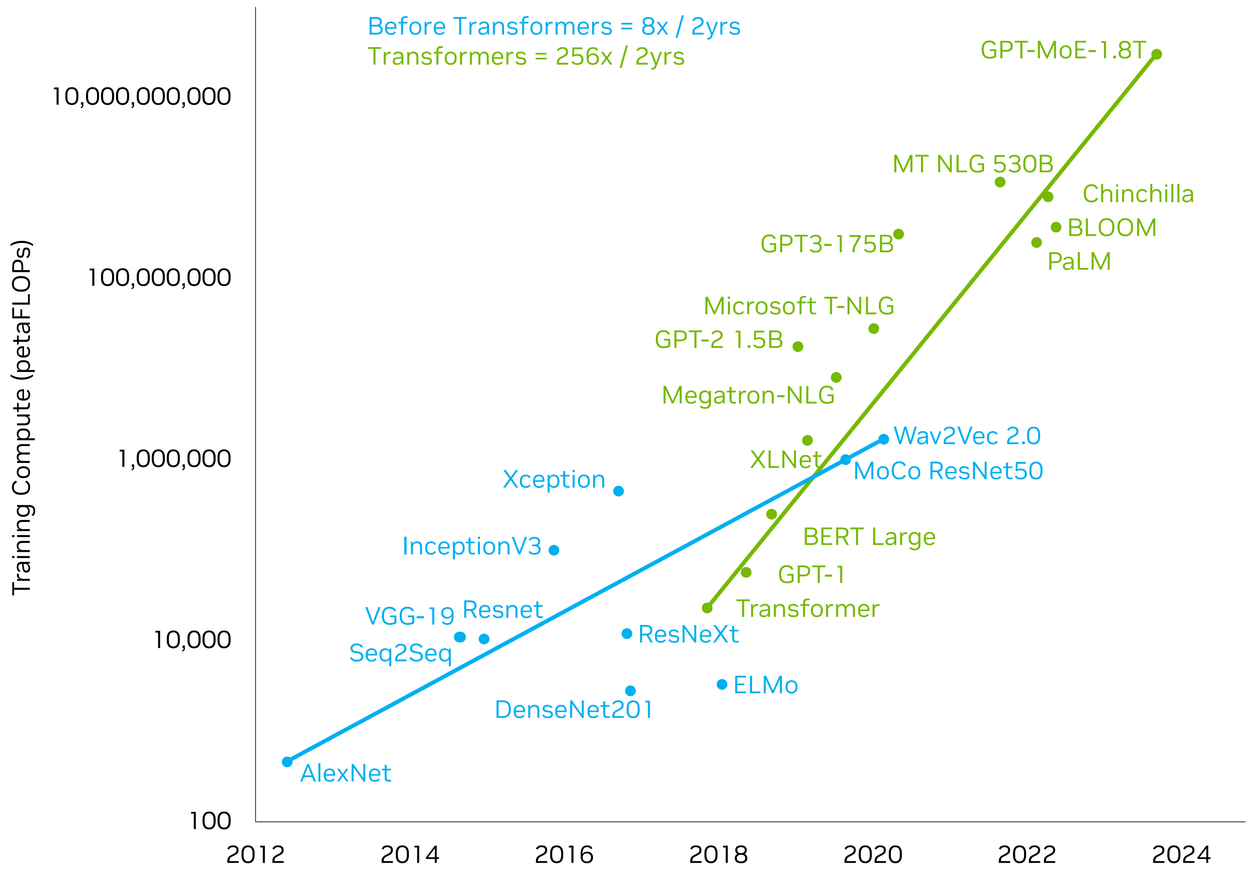
AI Factories Everywhere

- Manufacture intelligence from data
- Intelligence in the form of digital tokens
- New opportunities in all industries



Explosive Growth in AI Computational Requirements

Before and after Transformers Deep Neural Networks



CPU

CPU GPU

1t

= ~100X SPEED-UP

~3X POWER

~1.5X COST

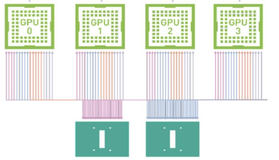
60X PERF / $ OR 98% SAVINGS

30X PERF / W OR 97% SAVINGS

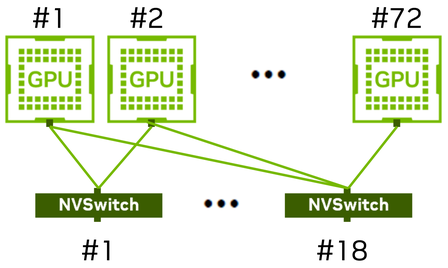
Explosive Growth in AI Computational Requirements

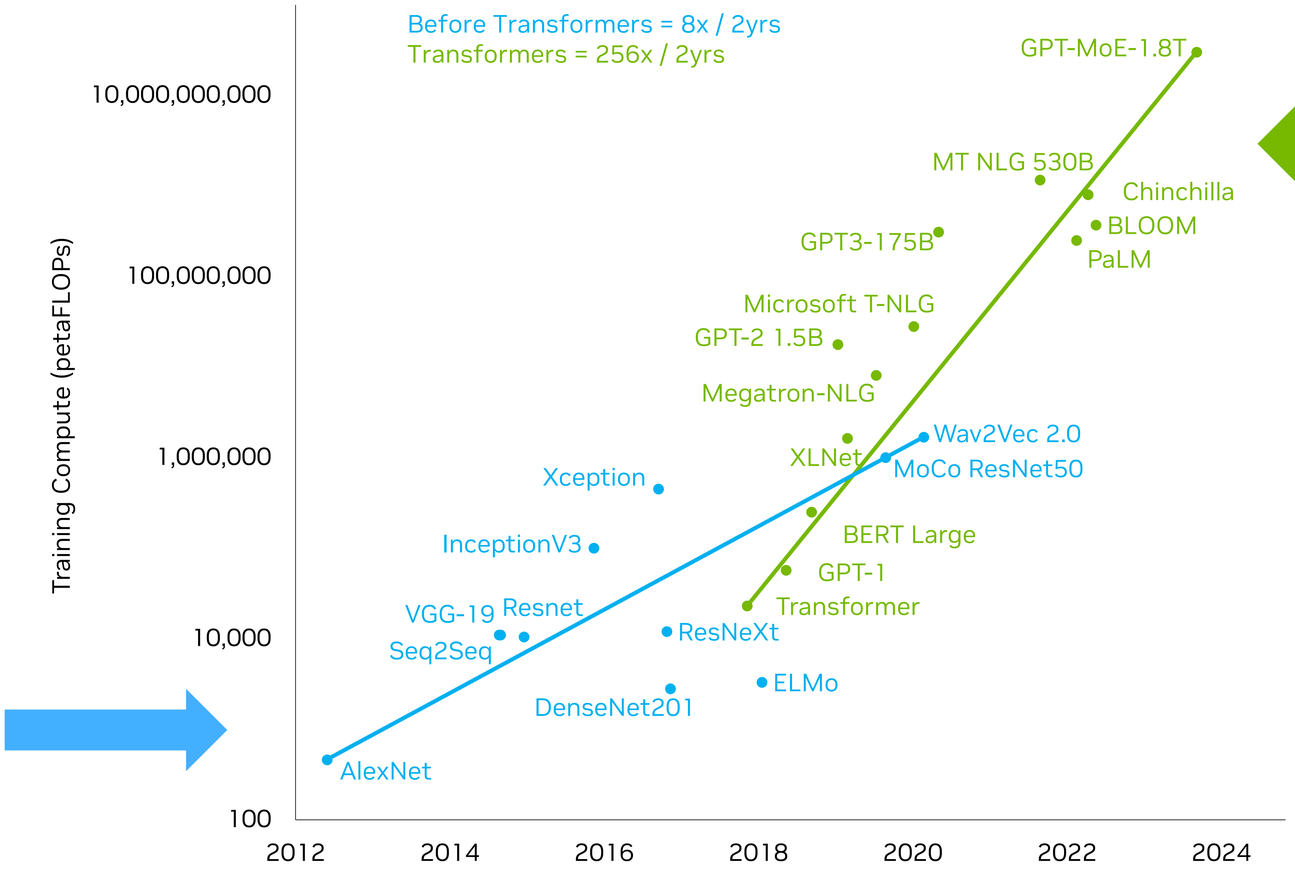
Before and after Transformers Deep Neural Networks

Fat 4/8-way GPU nodes with NVLink are needed for interactive LLM **inference**

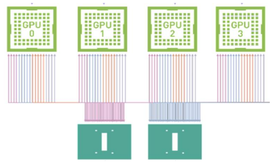


Very large GPU NVLink complexes (72-way) are needed for LLM **training**





DNN **training** drives fat 4/8-way GPU nodes with NVLink

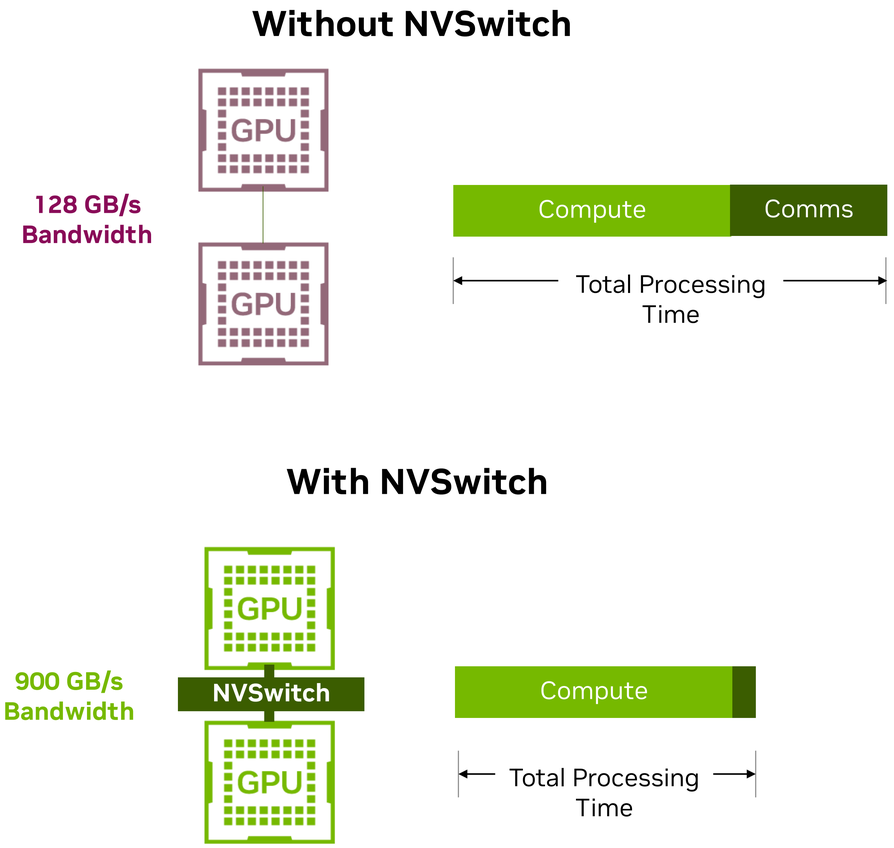


NVSwitch Supercharges LLM Training and Real-Time LLM Inference

Up to 1.5X higher inference throughput compared to point-to-point

Hopper NVSwitch

Hopper NVSwitch Provides 900 GB/s All-to-All Communication

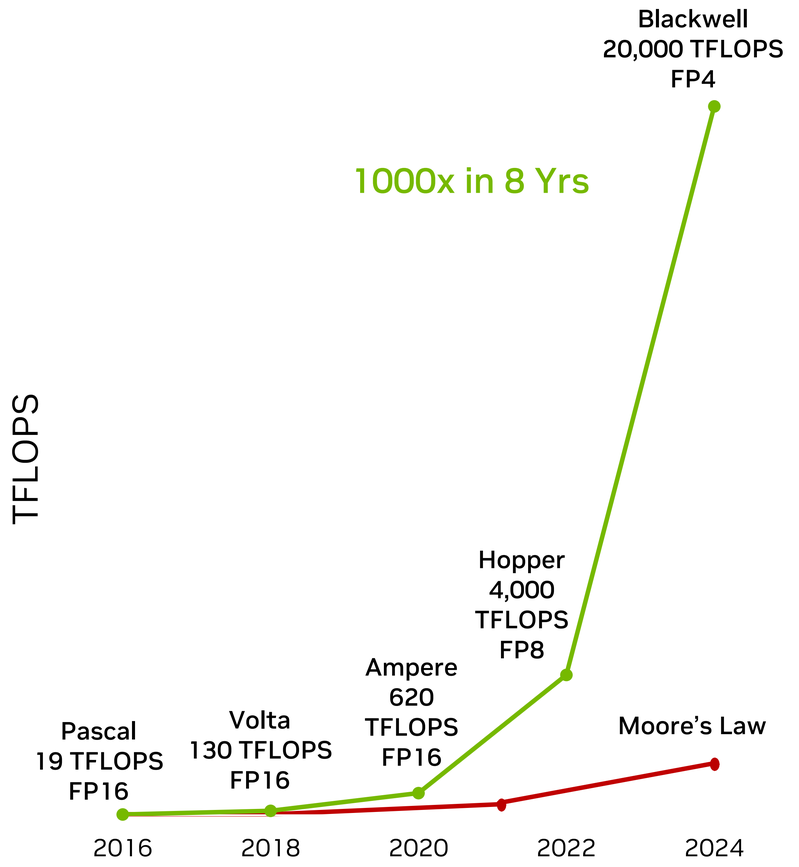


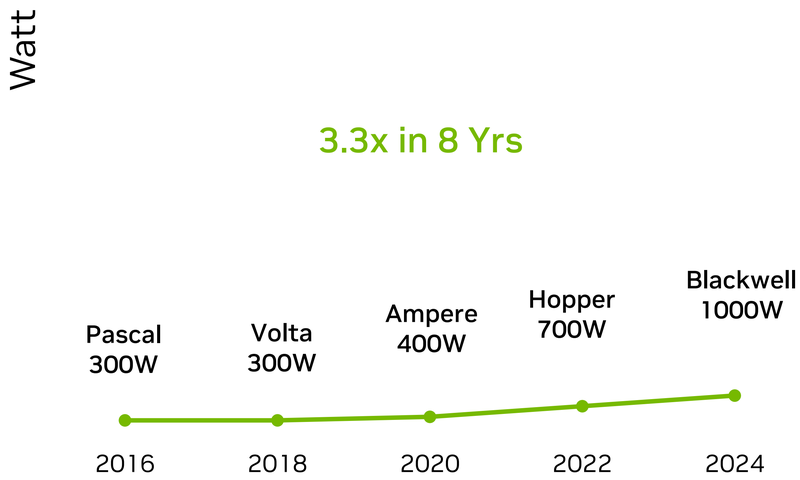
Up-to 1.5x higher Real-Time Llama3.1 70B Throughput

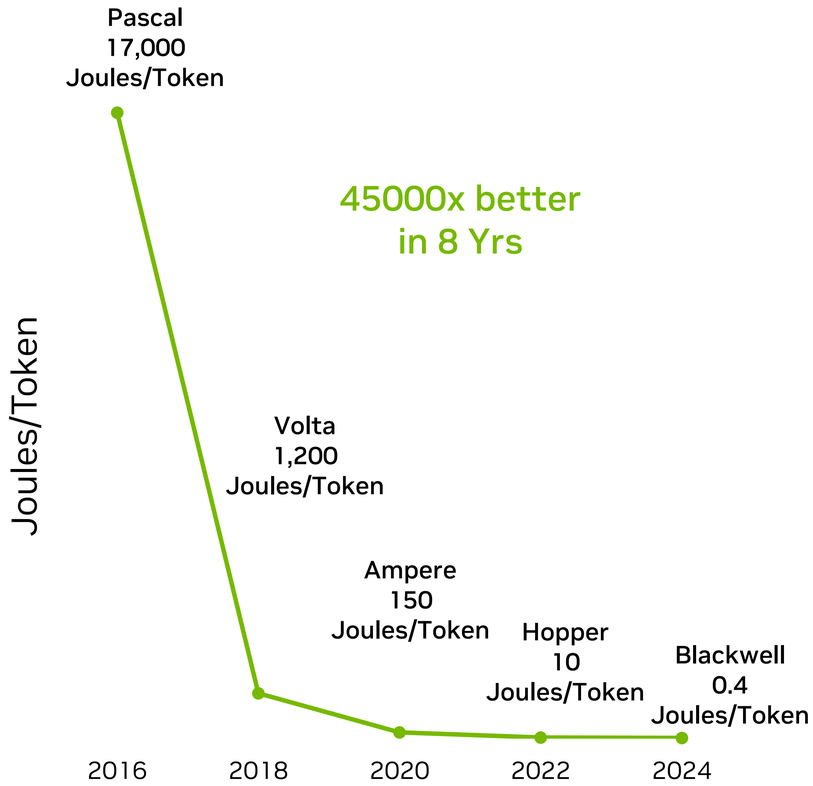
Source: https://developer.nvidia.com/blog/low-latency-inference-chapter-1-up-to-1-9x-higher-llama-3-1-performance-with-medusa-on-nvidia-hgx-h200-with-nvlink-switch/

GPU Architecture Evolution Drives Iso-Performance Energy Efficiency

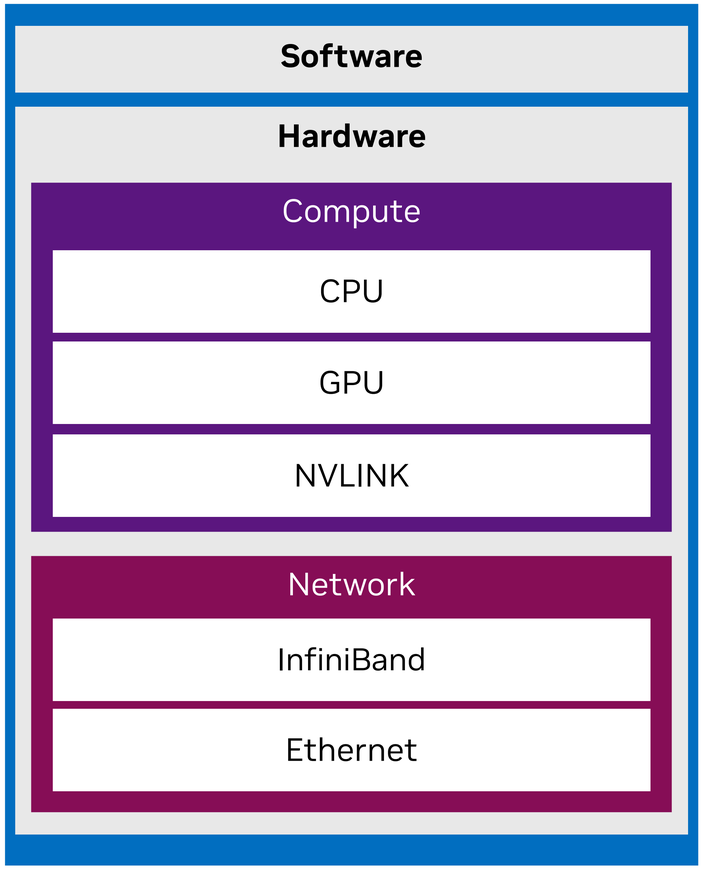
Exponential trend vs approx. linear trend in GPU TDP







AI Factory Compounds Benefits of NVIDIA Reference Architecture



5x Throughput with NIM

4X training with Blackwell

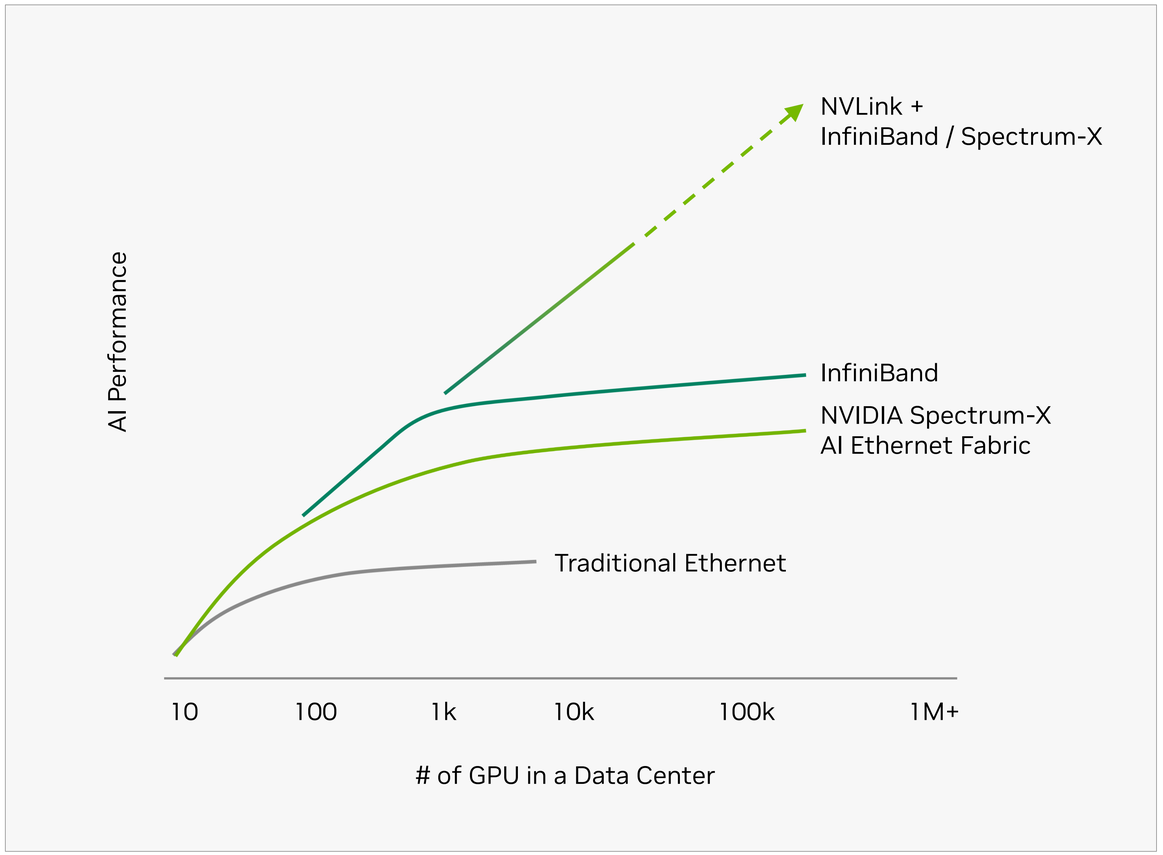
2x Performance

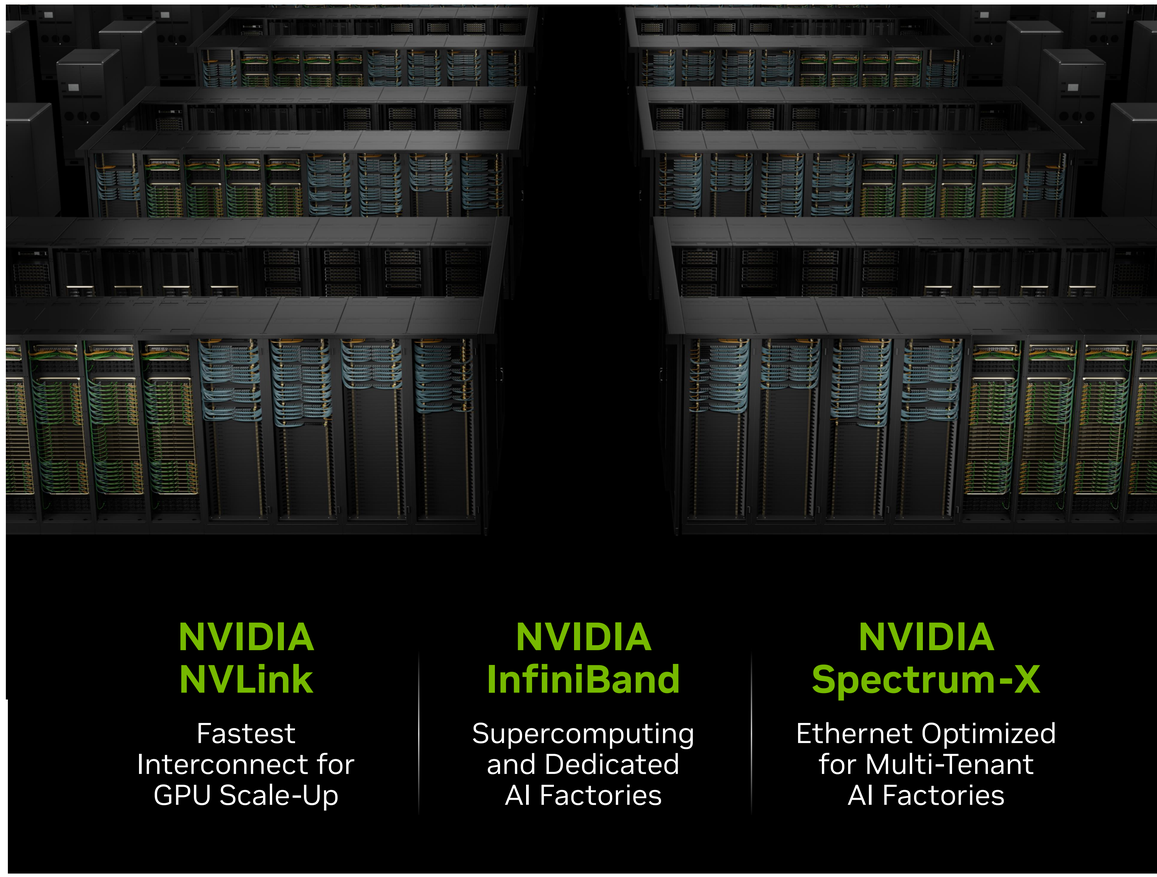
Faster Time-to-value

1.2% of energy usage for LLMs vs CPU infrastructure

Extending NVIDIA Networking to Scale-Up and Scale-Out AI in Any Datacenter

New NVLink and Spectrum-X Increase Networking Opportunity Beyond InfiniBand to Every Data Center





NVIDIA NVLink

Fastest Interconnect for GPU Scale-Up

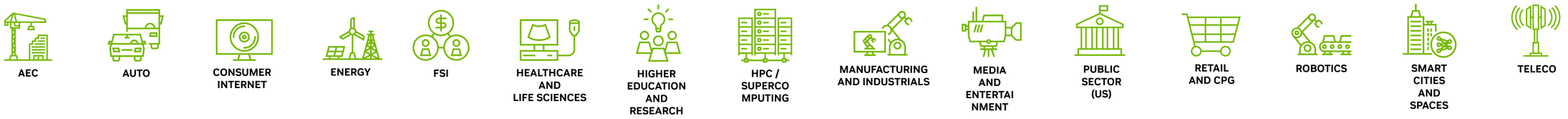
NVIDIA InfiniBand

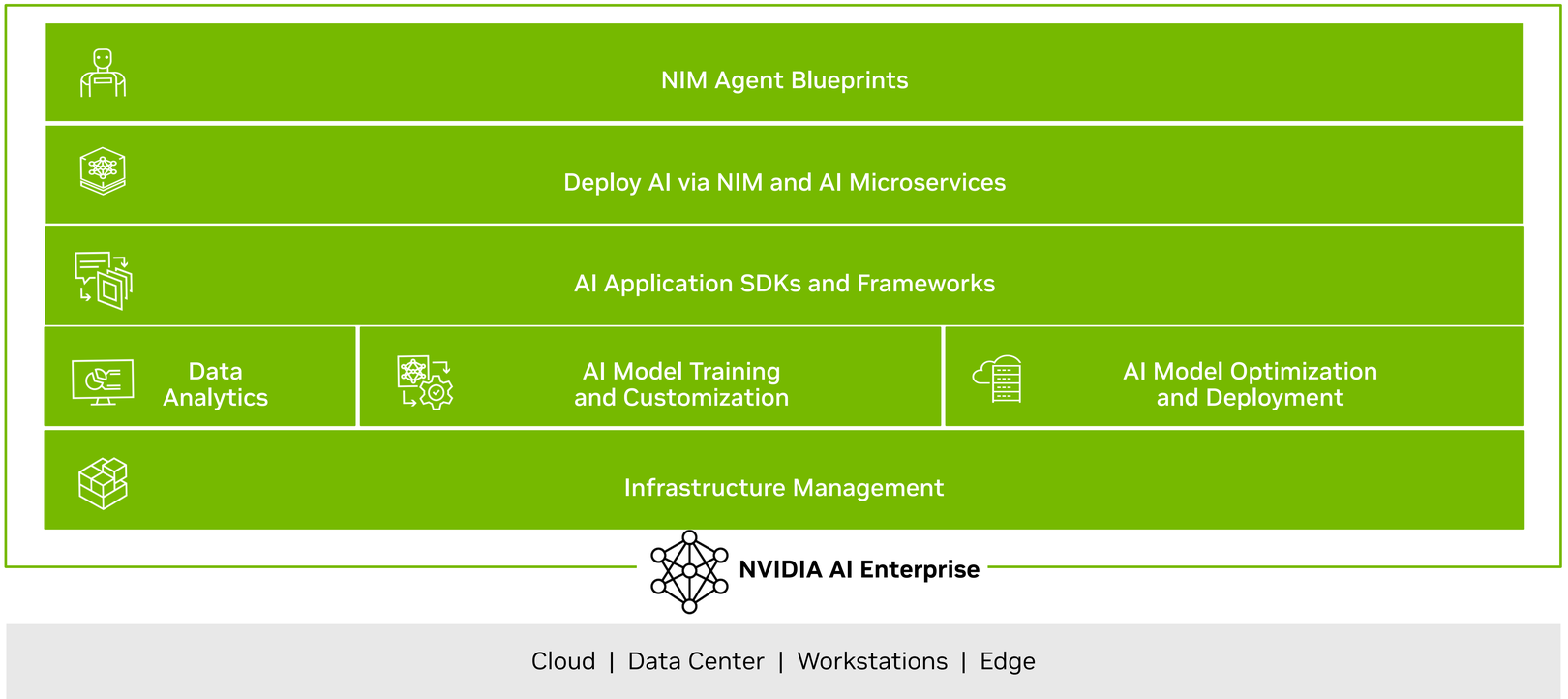
Supercomputing and Dedicated AI Factories

NVIDIA Spectrum-X

Ethernet Optimized for Multi-Tenant AI Factories

NVIDIA Software Activates AI Factories





Cloud | Data Center | Workstations | Edge

Announcing GB200 NVL72

Delivers New Unit of Compute

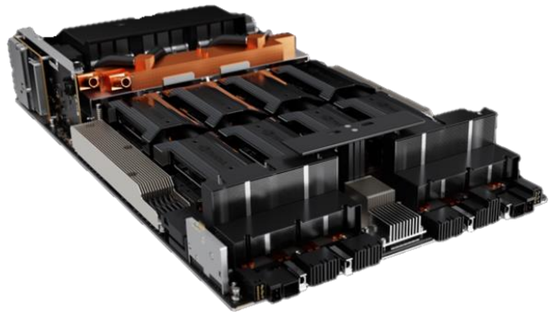
GB200 NVL72

- 36 GRACE CPUs
- 72 BLACKWELL GPUs
- Fully Connected NVLink Switch Rack

Training FP8	720 PFLOPs
Inference FP4	1,440 PFLOPs
NVL Model Size	27T params
Multi-Node All-to-All	130 TB/s
Multi-Node All-Reduce	260 TB/s

NVIDIA Liquid-Cooled, Sustainable Accelerated Compute

Less Water	Lower Power	Less Rack Space
275X	25X	4X



HGX H100
Shipping

GB200 NVL72
Coming Soon

* Energy & Space Efficiency GB200 NVL72 versus HGX H100, GPT-1.8T-MoE Inference

Introducing DGX SuperPOD With DGX GB200 Systems

Turnkey supercomputing for trillion-parameter AI

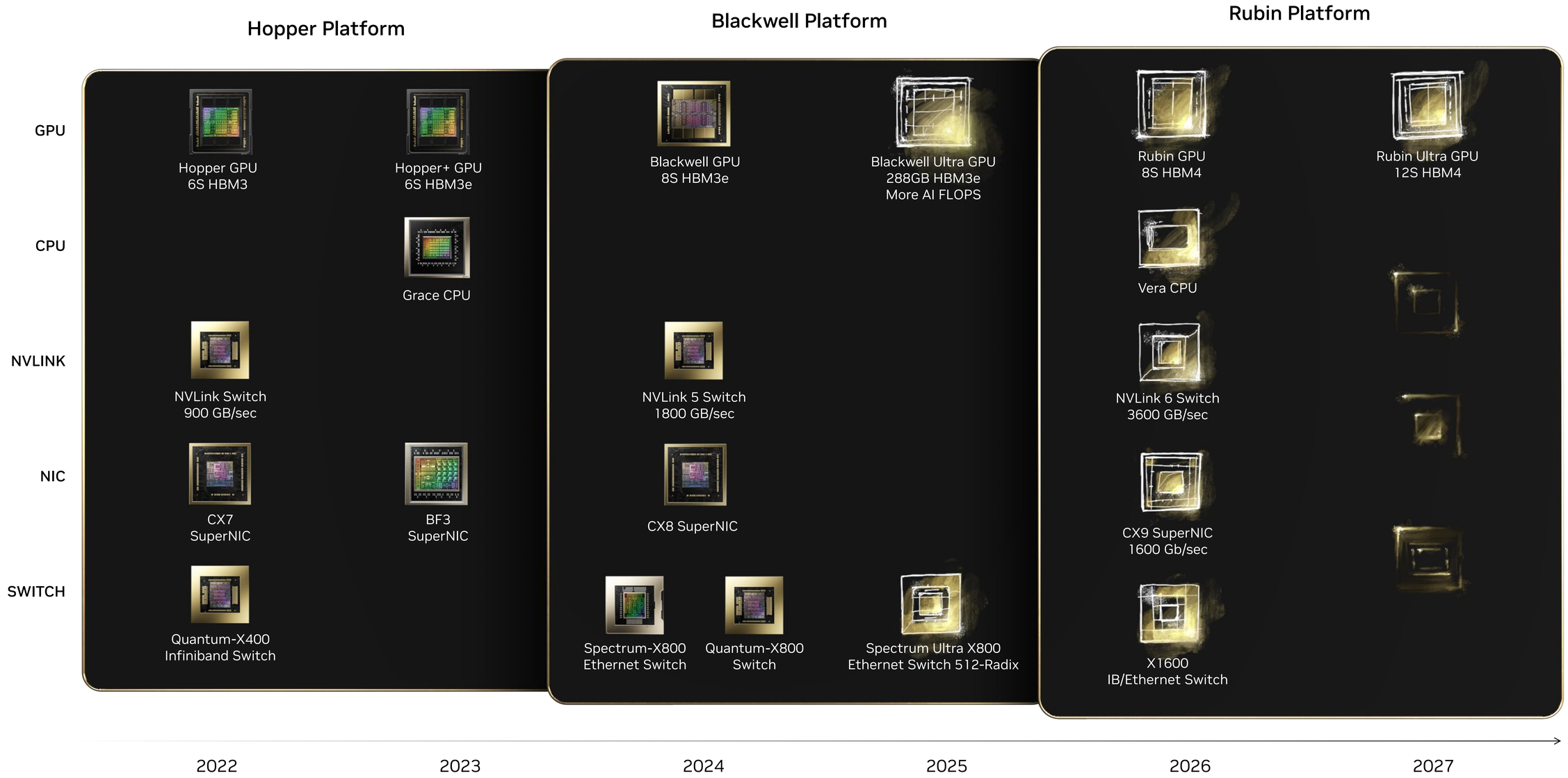
- Highly efficient, liquid-cooled, rack-scale design built with NVIDIA GB200 Grace Hopper Superchips
- **36** NVIDIA Grace CPUs and **72** NVIDIA Blackwell GPUs per rack, connected via fifth-generation NVLink
- Scale to tens of thousands of GB200 Superchips with Quantum-2 InfiniBand
- Intelligent, full-stack resilience for constant uptime
- Integrated hardware and NVIDIA AI software
- Built, cabled, and factory tested before delivery and installation
- Optional **576** NVLink configuration for memory-limited workloads

The World's Most Efficient AI Infrastructure

DGX SuperPOD with 8 DGX GB200 systems

288 Grace CPUs | 576 Blackwell GPUs
240TB Fast Memory | 11.5 ExaFLOPS FP4
30X Inference | 4X Training | 25X Energy Savings

Datacenter Scale | One-Year Rhythm | Technology Limits | One Architecture



Thank you!

cnardone@nvidia.com