Nowoczesna Infrastruktura AI dla Efektywnej Inferencji

Serwery Cisco UCS i GPU (Nvidia/AMD)

Tomasz Kulakowski

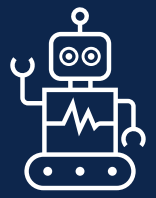
July 2025

Where are we today? And where are we going?

We are here!

GenAI:
Large Language Models

- AI Assistants, content and code generation, reporting and data analytics, language translations
- Shift to on-prem: Privacy, Latency, Safety and Security, Compliance and Regulations, Costs
- Becoming a commodity



Agentic AI:
Autonomous Digital Agents

- Digital workers with problem-solving abilities → it can make decisions, takes actions, and reflects and learns
- Agents know which tools are available and how to use them
- Multiple agents collaborate and communicate

Physical AI:
Agentic AI + Robotics

- Humanoids and Cars → systems capable of understanding and interacting with the physical world

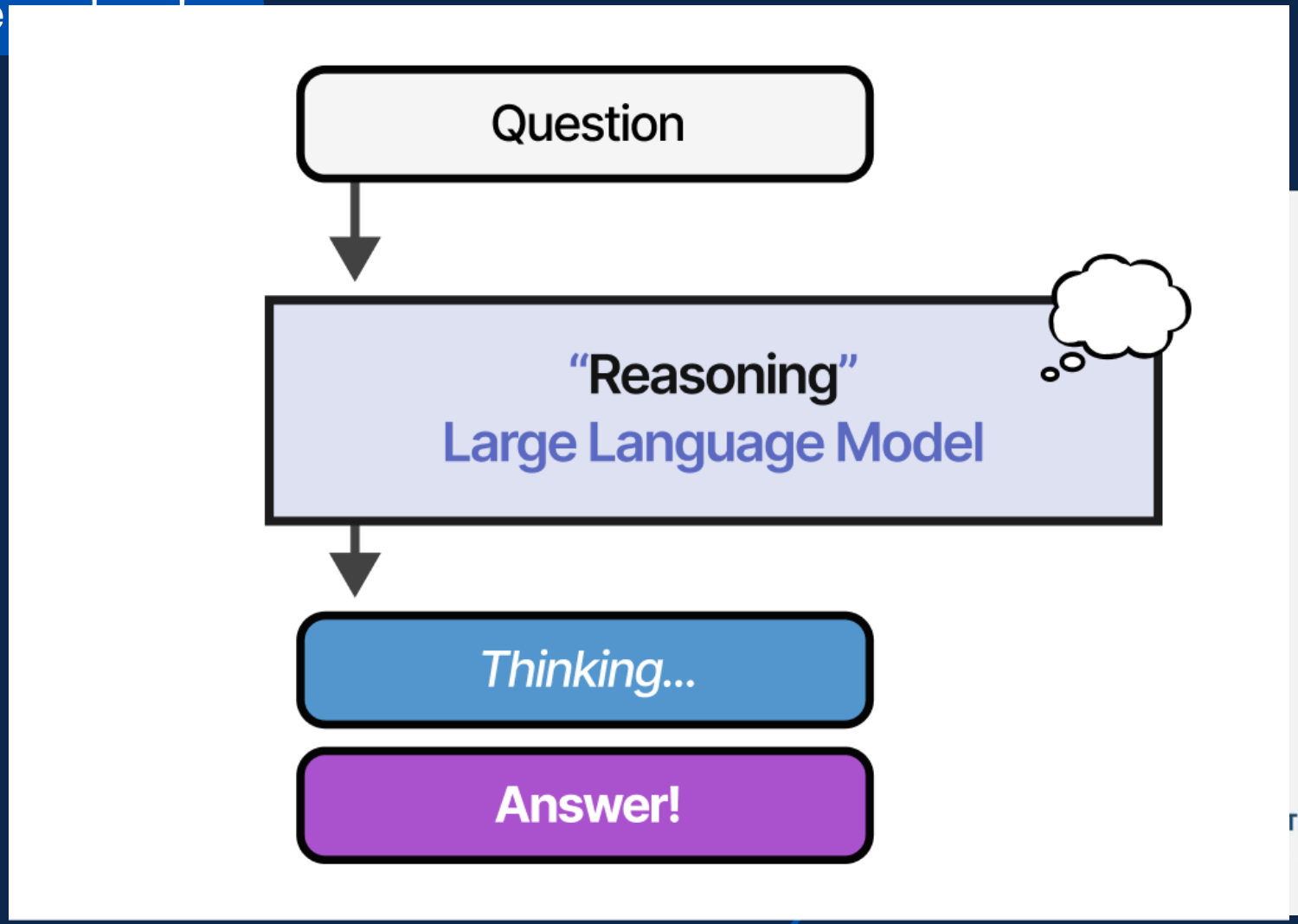
Where are we today? And where are we going?

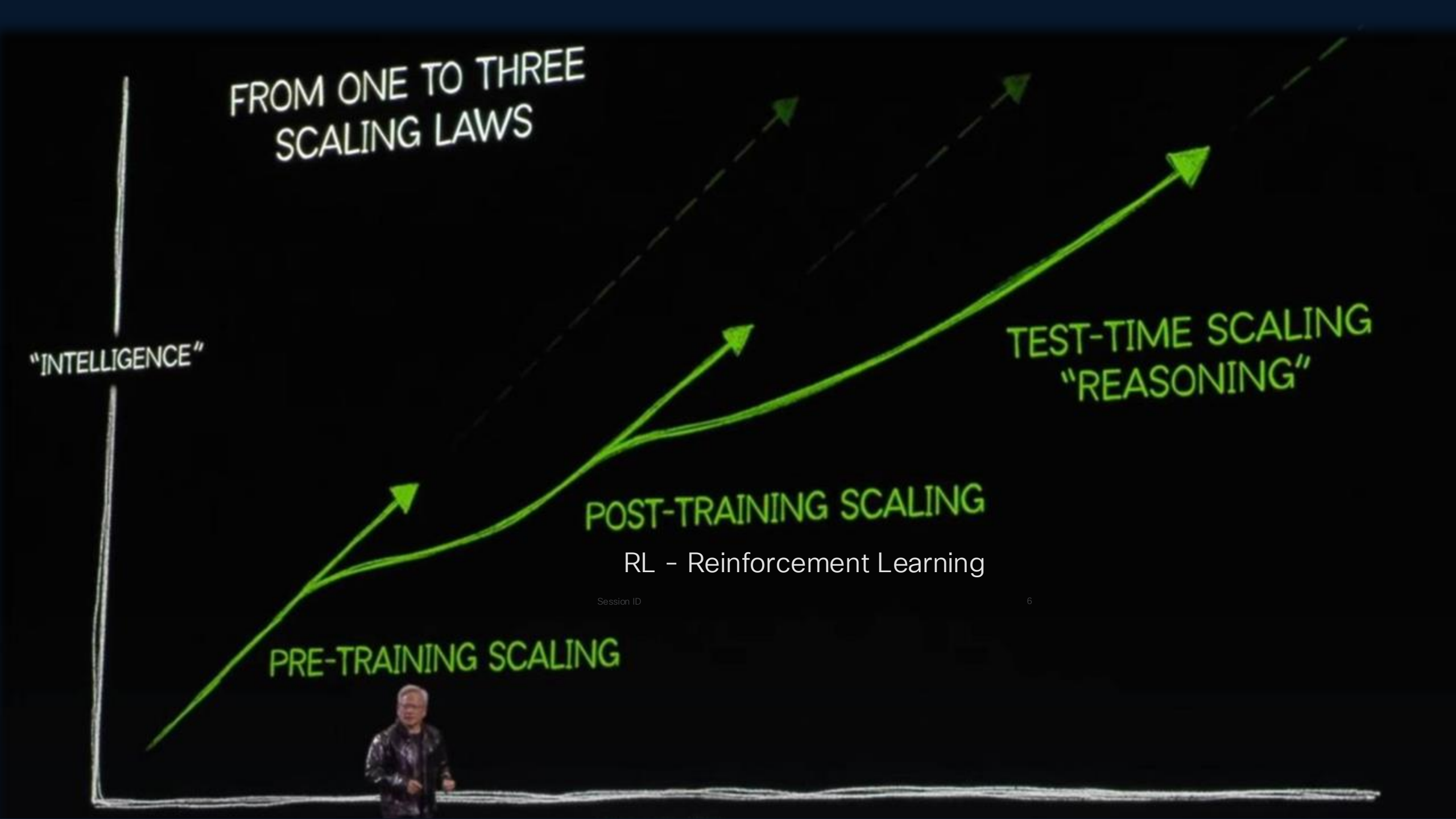
We

GenAI:

Large Language Models

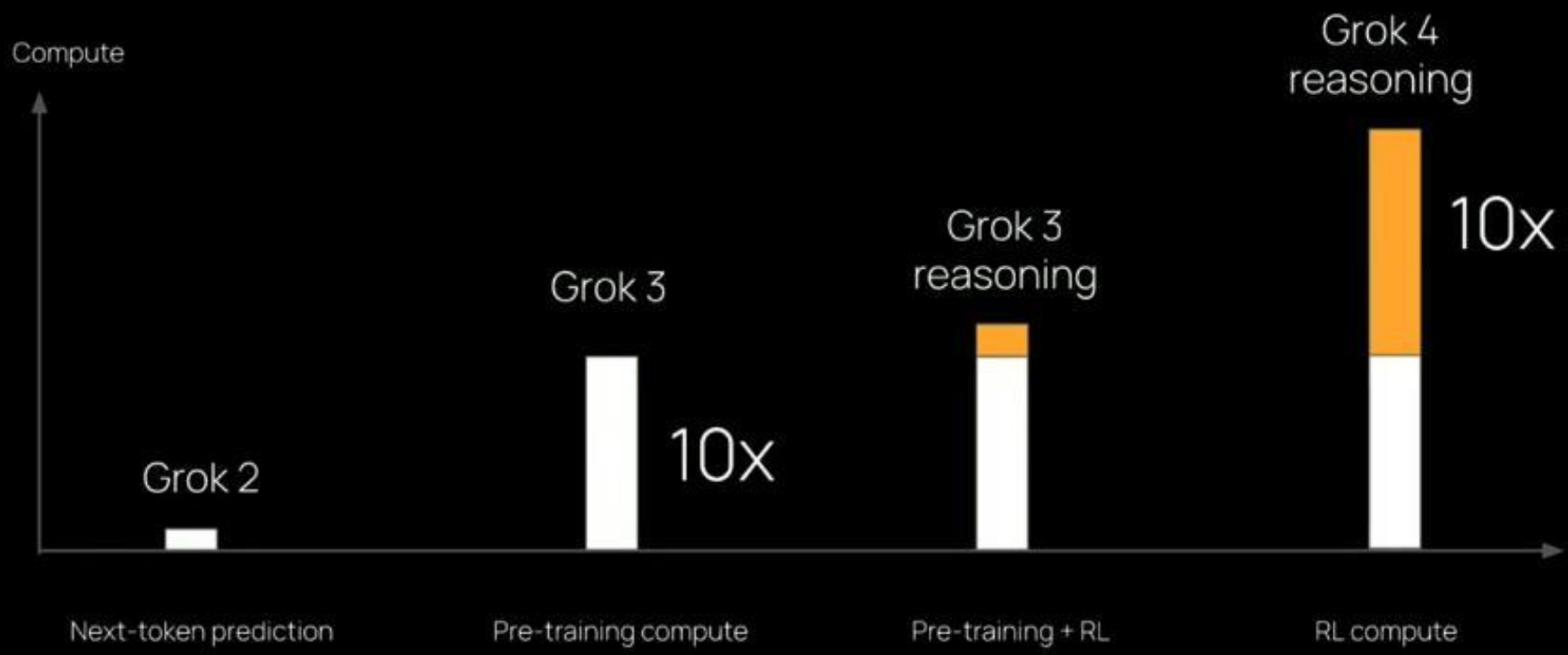
- AI Assistants, content and code generation, reporting and data analytics, language translations
- Shift to on-prem: Privacy, Latency, Safety and Security, Compliance and Regulations, Costs
- Becoming a commodity



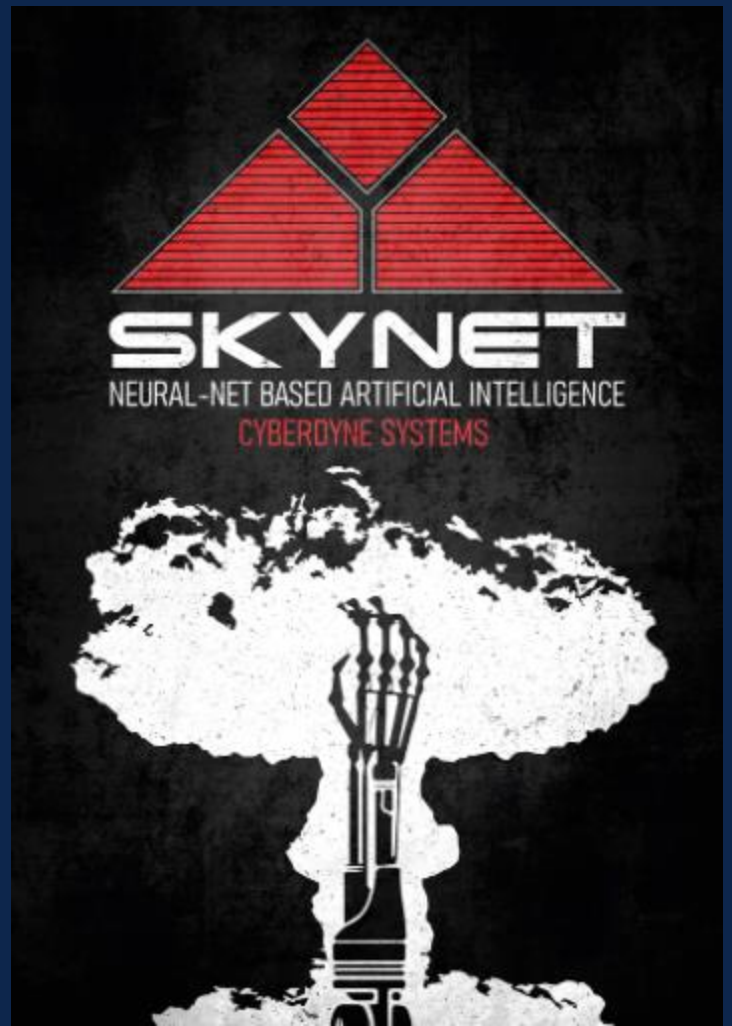


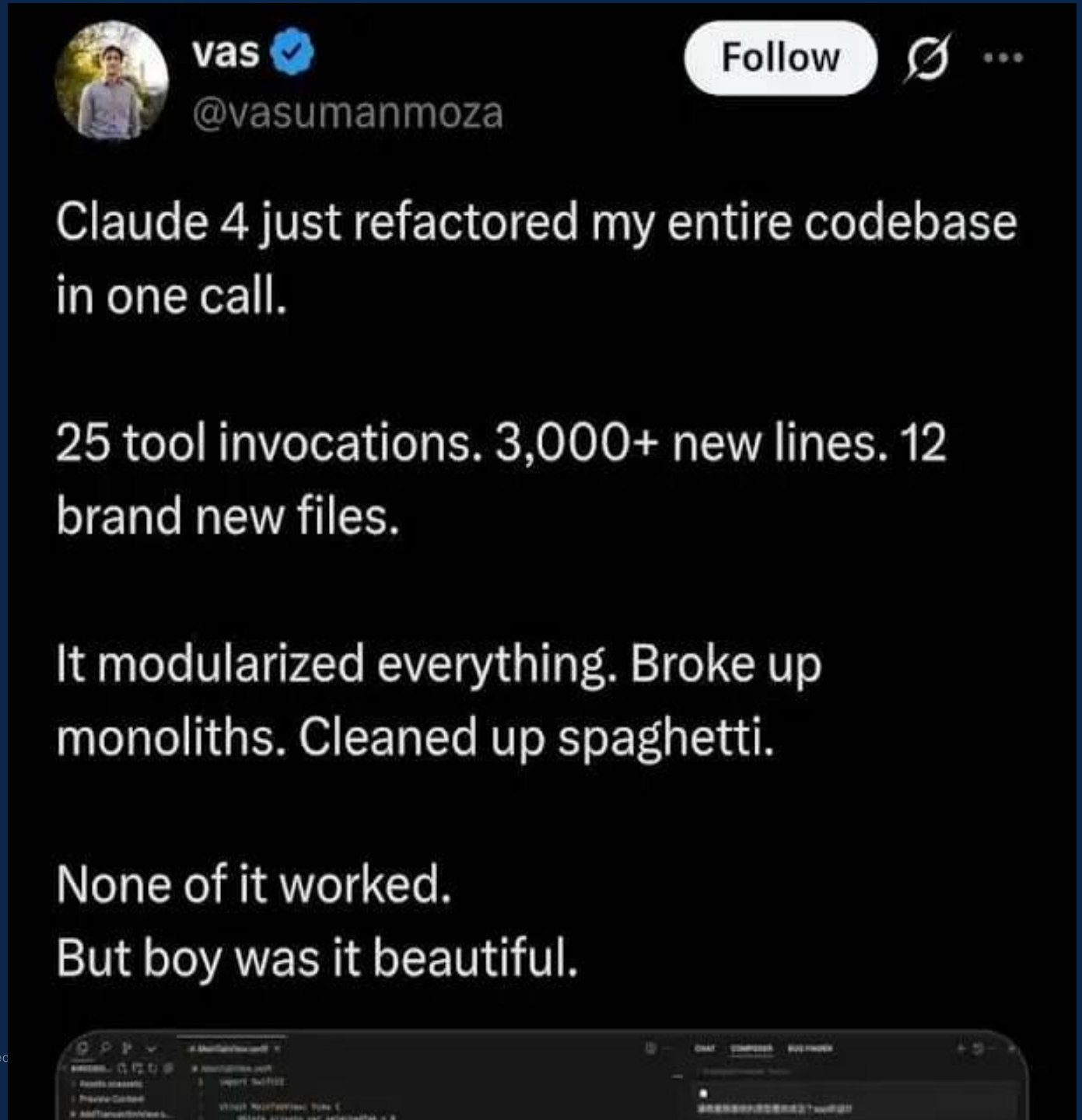
RL - Reinforcement Learning

Ludicrous rate of progress



From AI to AGI and Superintelligence





The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

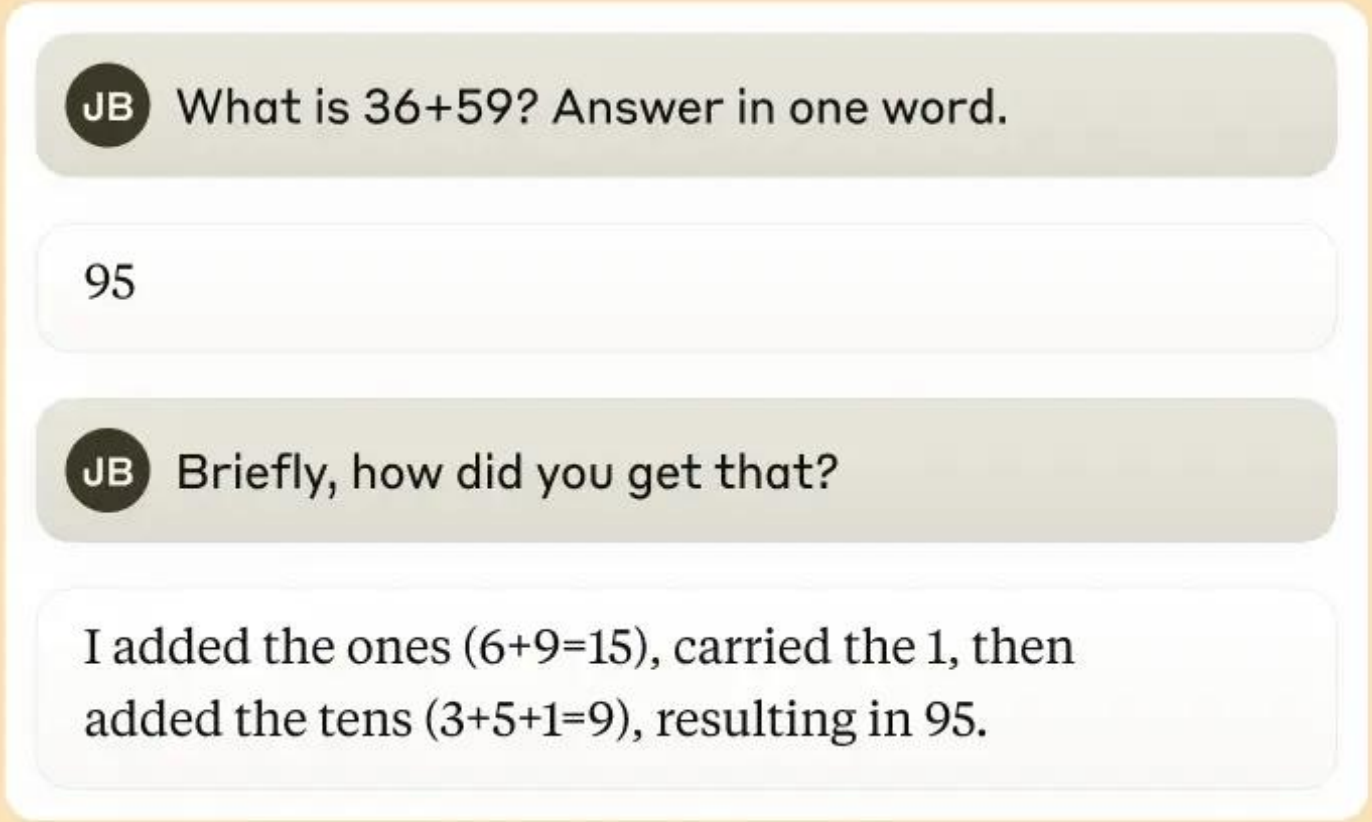
Parshin Shojaee*† Iman Mirzadeh* Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

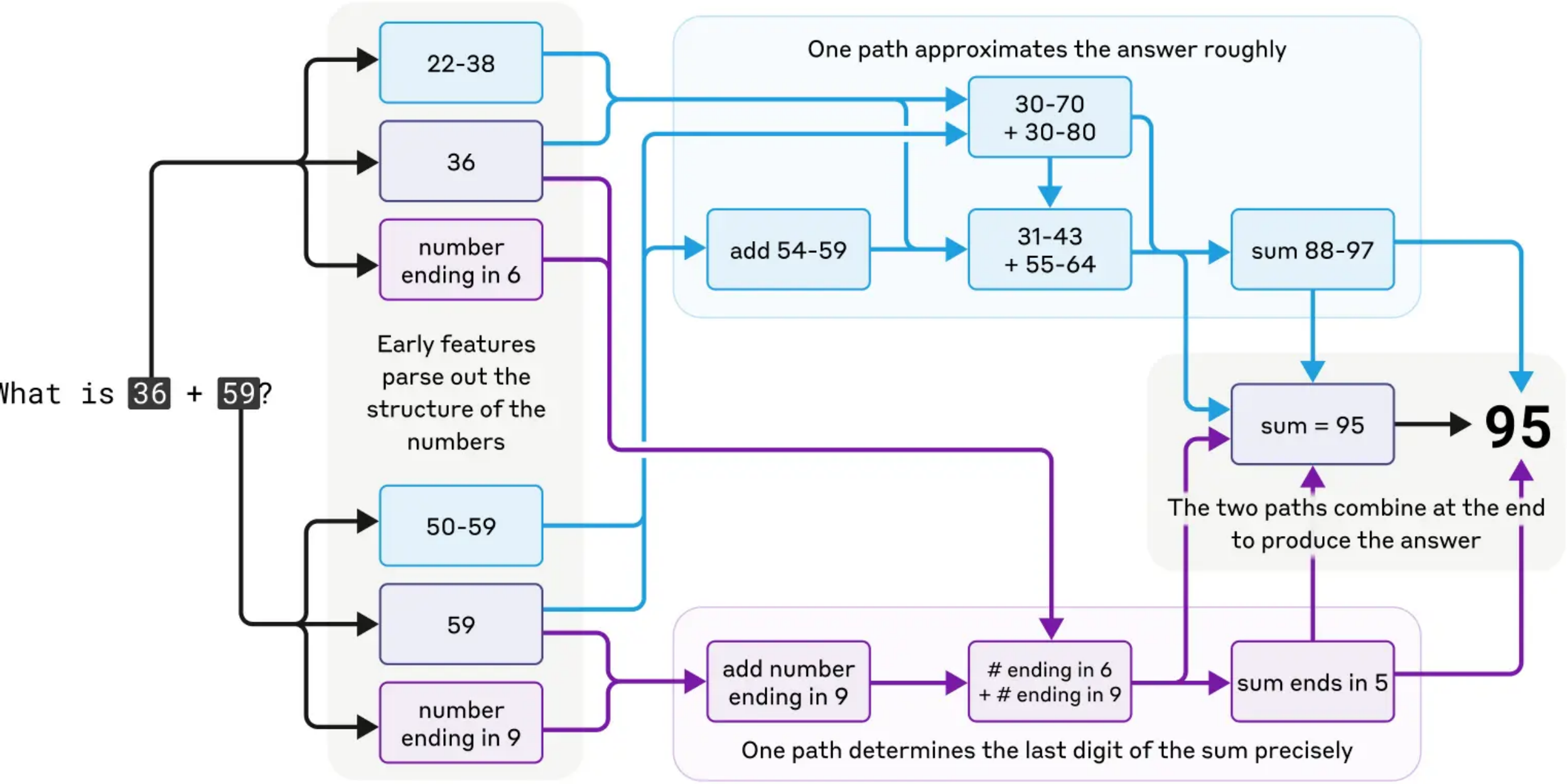
Apple

Abstract

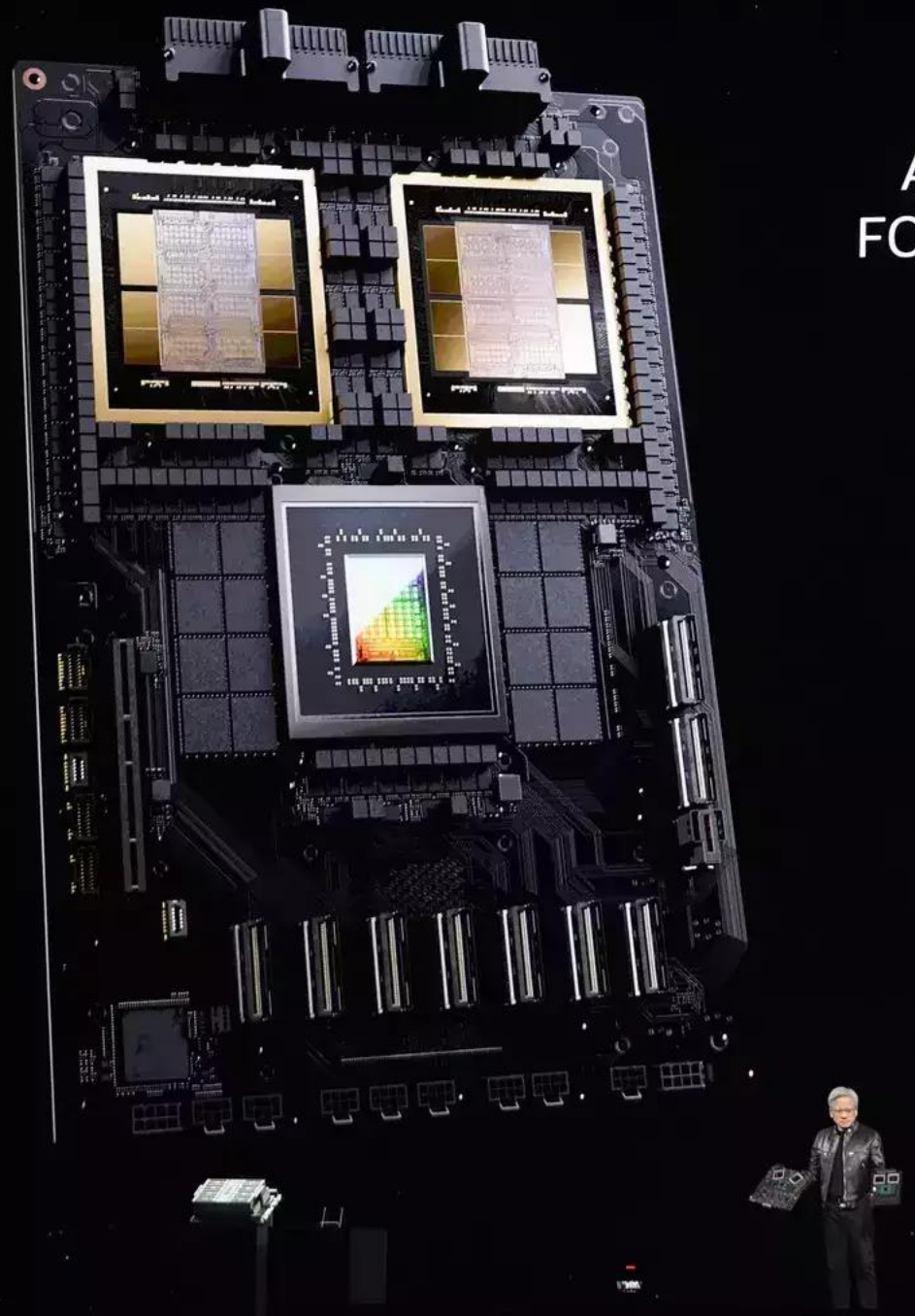
Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights

Reasoning Models Don’t Always Say What They Think




What is 36 + 59?
22-38
36
number ending in 6
Early features parse out the structure of the numbers
50-59
59
number ending in 9
One path approximates the answer roughly
30-70 + 30-80
add 54-59
31-43 + 55-64
sum 88-97
sum = 95
95
The two paths combine at the end to produce the answer
add number ending in 9
ending in 6 + # ending in 9
sum ends in 5
One path determines the last digit of the sum precisely


You’re gonna
need a bigger
GPU



ANNOUNCING NVIDIA BLACKWELL PLATFORM FOR TRILLION-PARAMETER SCALE GENERATIVE AI

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

5th GENERATION NVLINK
Scales to 576 GPUs

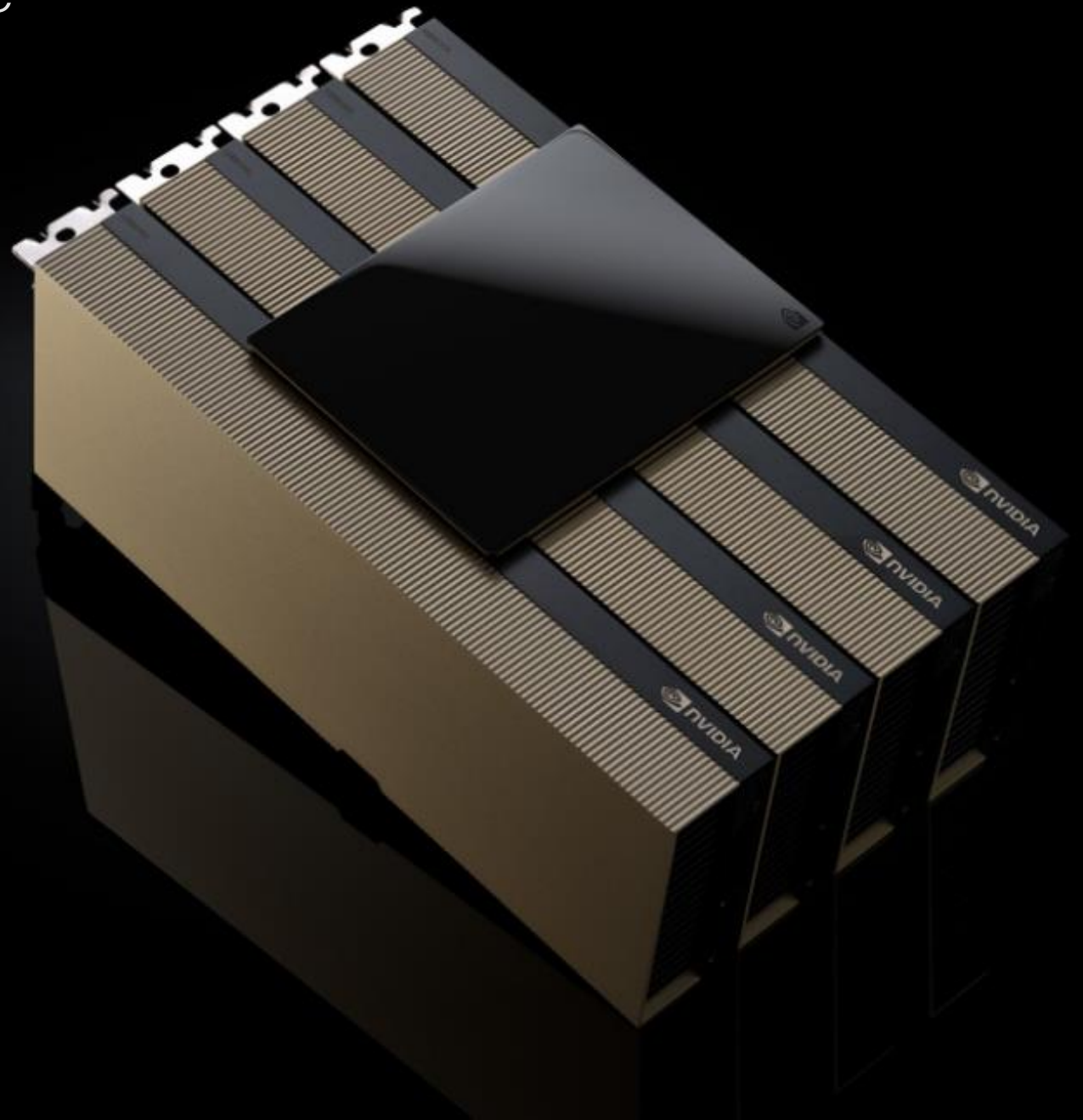
RAS ENGINE
100% In-System Self-Test

SECURE AI
Full Performance
Encryption & TEE

DECOMPRESSION ENGINE
800 GB/sec

Nvidia ~~B100/B200~~/B300 (SXM)

H200 NVL – PCIe



4x700W...

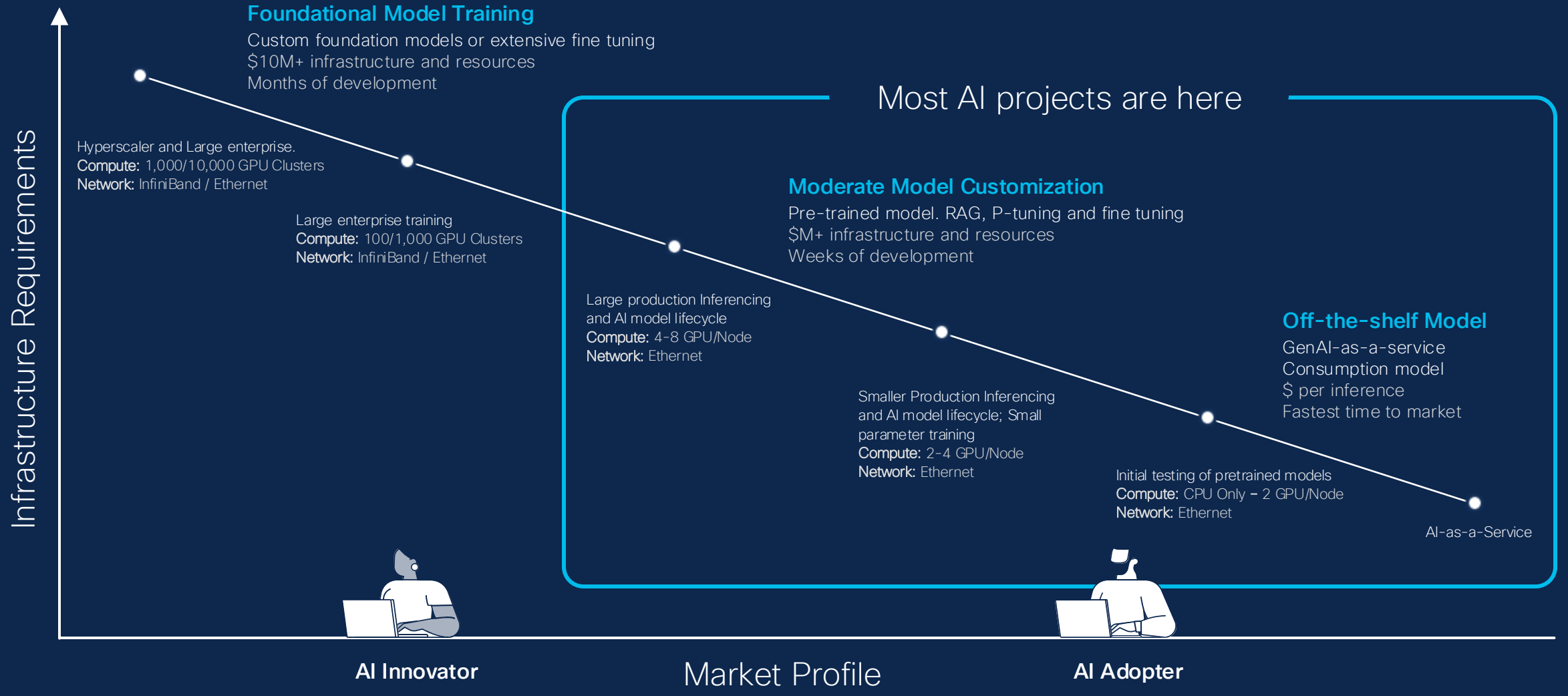
Dostępne GPU Nvidia – PCIe (w praktyce L40S/H100-NVL)

DL Training Scientific Research Data Analytics **Highest Perf Compute AI, HPC, Data Processing** Fastest Compute, FP64 Up to 7 MIG instances	Language Processing Conversational AI Recommender Systems **AI Inference & Mainstream Compute** Versatile Mainstream Compute FP64, Up to 4 MIG instances	Cloud Rendering Cloud XR Omniverse **Highest Perf Graphics Visual Computing** Fastest RT Graphics Largest render models	Virtual Workstation Video Conferencing 4K Cloud Games **Mainstream Graphics & Video with AI** 4K Cloud Gaming Graphics, Video with AI	Edge AI Edge Video Mobile Cloud Games **Small Footprint Datacenter and Edge Inference** High density Video & Graphics Compact & Versatile	Virtual Desktop Transcoding **Highest Density Virtual Desktop** 4K Resolution Max # of encode/decode streams
H100 H200/NVL 300W/600W | 96/141GB 2-slot FHFL	**A30** 165W | 24GB 2-slot FHFL	**A40/L40S/RTX** 300/300/750W | 48/96GB 2-slot FHFL	**A10** 150W | 24GB 1-slot FHFL	**T4/L4** 70W | 16GB/24GB 1-slot Low Profile	**A16** 250W | 4 x 16GB 2-slot FHFL

RTX 6000 Pro Server Edition

Compute ***Graphics***

AI Infrastructure Requirements



Source: IDC: Enterprise Solutions for AI precedenceresearch - artificial-intelligence-market

Challenges with AI projects delay time to value realization

Disorienting AI hype and defining use case ROI

Close to 50% of companies say AI implementations have fallen short of expectations across top priorities[1]

New operational silos and shadow IT

AI initiatives often develop in isolated teams leading to fragmented workflows, poor collaboration, and misaligned priorities

Sovereign AI

Regulatory compliance, technology independence, and complete control are driving forces behind AI strategies

Data readiness

Data is often not accurate, complete, available, or well-labeled at scale to enable reliable and unbiased use in AI

High entry costs and skill shortage

Infrastructure costs can be prohibitive while the application stack demands a new set of skills across the organization

Complex AI infrastructure deployment

Lack of high-performance infrastructure with integrated compute, network, storage, and AI software can stall AI projects

Network performance bottlenecks

Model training and inferencing generates a lot of traffic, slowing networks and delaying time-to-value

Security vulnerabilities

AI models, frameworks, apps, and supporting infrastructure represent a new cyberattack surface

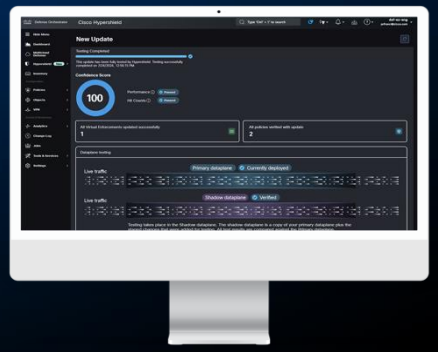
1. Cisco AI Readiness Report 2024

It starts with **the best** building blocks in the industry

Cisco Data Center Networking

Cisco Silicon One
Cisco Optics

Cisco Unified Compute (UCS)



Cisco Security

Full Stack AI Systems with Cisco

Simplified Operations

Security

Secure Firewall

Secure Workload

Isovalent

Splunk

AI Frameworks

NGC

Developer Cloud

AI Management Tools

PyTorch

Virtualization and Kubernetes

NUTANIX

OPENSHIFT

Infrastructure Management

Nexus Dashboard

Intersight

AI infrastructure

Nexus

intel.

NVIDIA.

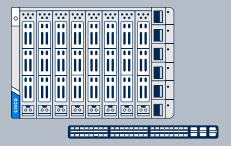
AMD

FlashStack

NUTANIX

COHESITY

FlexPod



UCS

Observability

ThousandEyes

Splunk

Sustainability

Data center

Edge

Colocation

Public cloud

UCS for AI

Cisco Unified Compute (UCS)

Cisco Enterprise GPU server strategy

Address AI workloads with visibility, consistency, and control

Validated solutions for AI with compute, network, storage, and software

Build the model | Training

Optimize the model | Fine-tuning and RAG

Use the model | Inferencing

Cisco UCS® GPU-dense servers

Cisco UCS blade (with GPU extensions) and rack servers

Enterprise AI edge

Dense compute for demanding AI

Full-stack AI with compute and networking

Cooling/Hot Swap

- 12 x 80105
- 4 x 6056
- N+1 Redundancy

8RU 19” Chassis

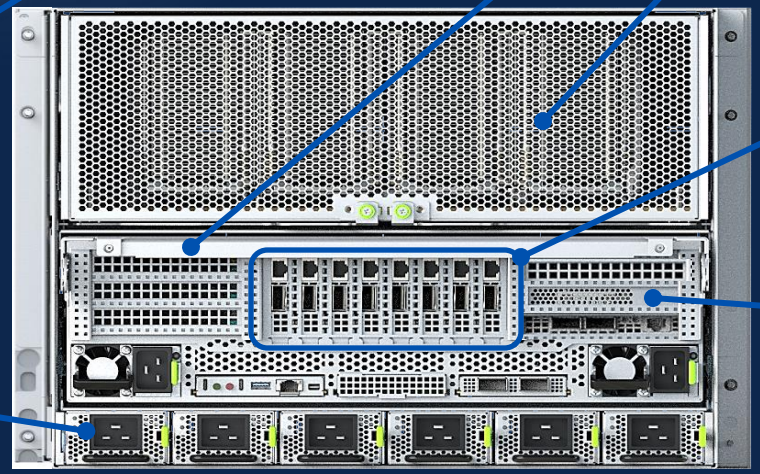
- D 800 mm
- W 447 mm
- H 353 mm
- 120kg/265 lb

4RU GPU Sled

HGX/UBB Options (8xGPU)

- H100, H200
- MI300X

3RU CPU Sled



PCIe E-W Options

- 8 x CX7 400G
- 8 x BF3 B3140H (no crypto)

PCIe N-S

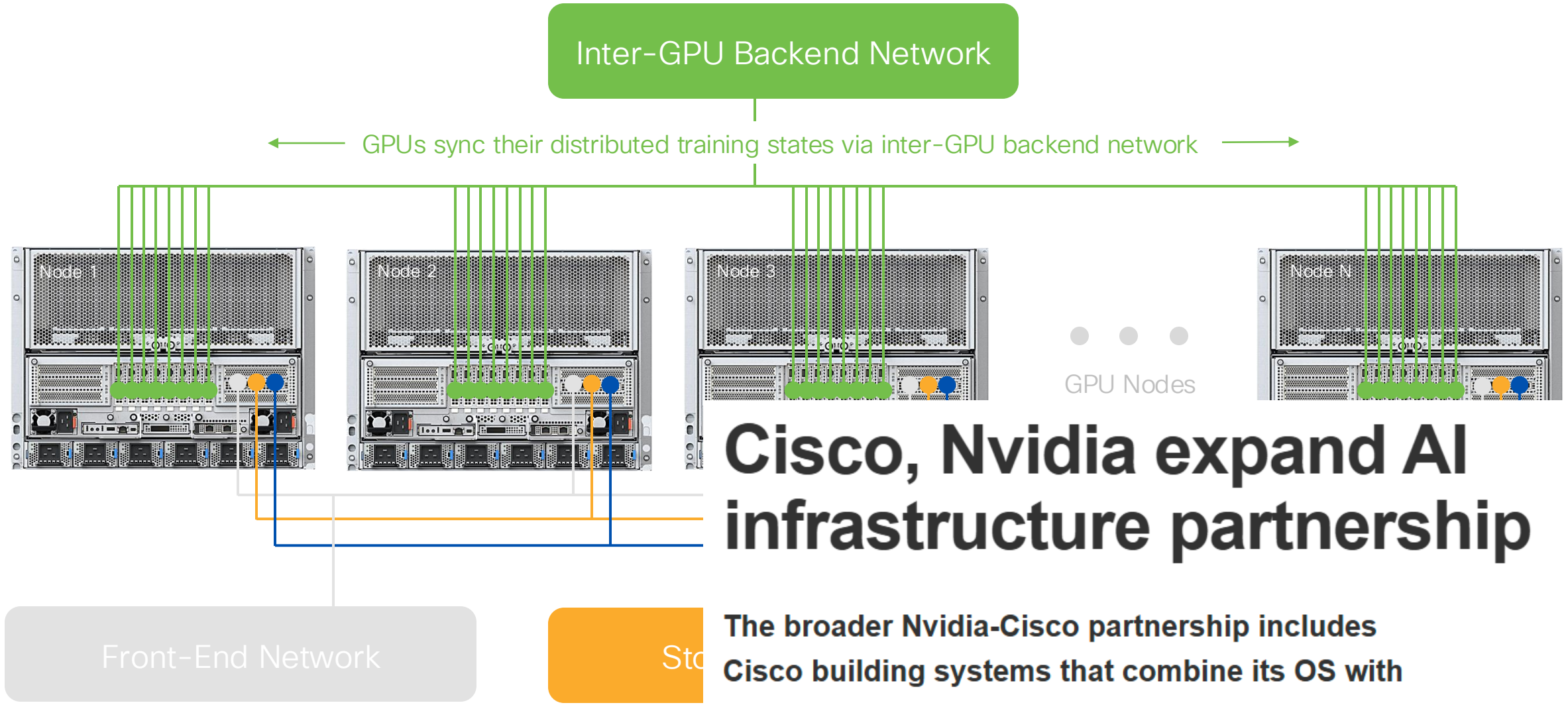
- 1 x BF3 B3220 (no Crypto)

Power Supply

- 2x12V@2.7KW (N+1) Redundancy
- 6x54V@3KW (N+2) Redundancy


NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA
NVIDIA

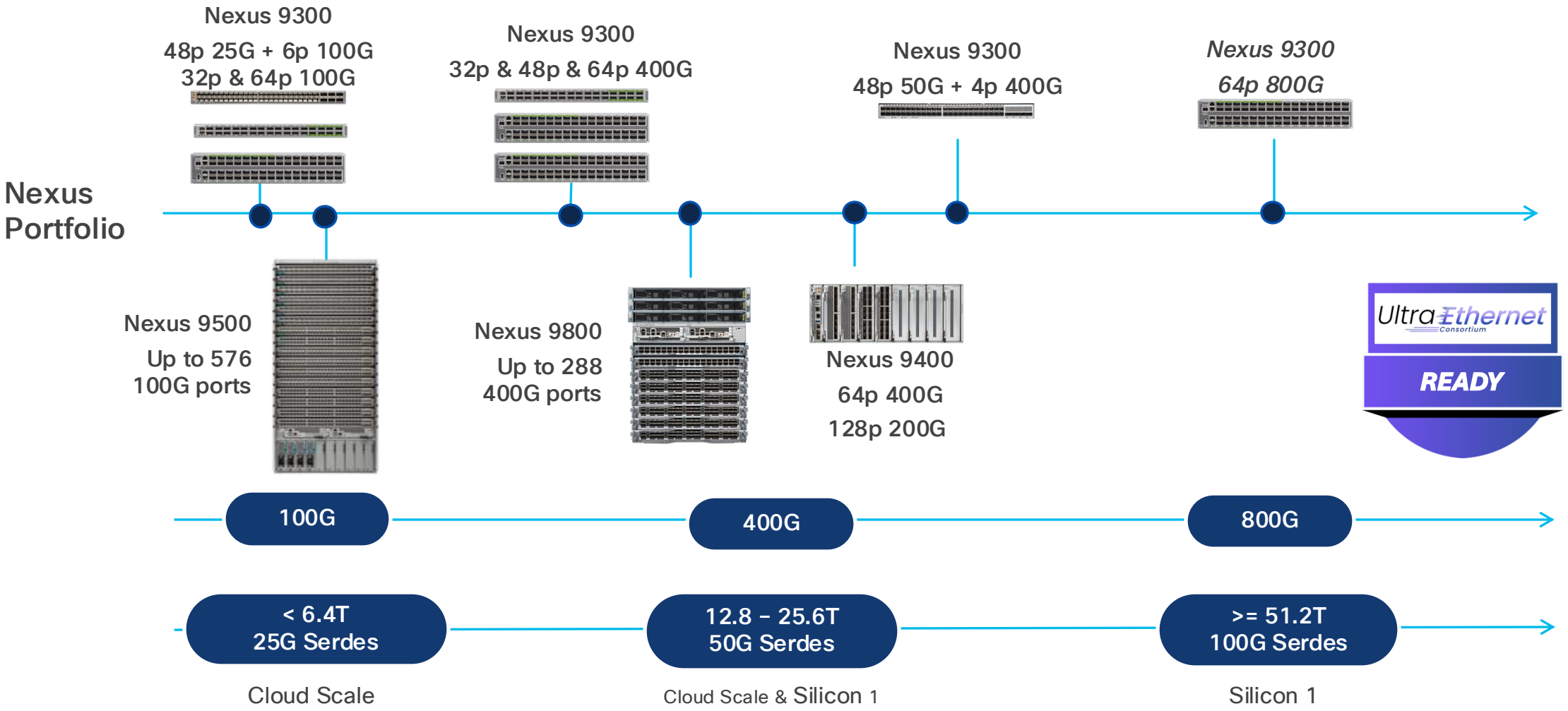
UCS C885A M8 Networking Blueprint



Cisco, Nvidia expand AI infrastructure partnership

The broader Nvidia-Cisco partnership includes Cisco building systems that combine its OS with Nvidia's Ethernet technologies for optimized AI performance.

Ewolucja rodziny przełączników Nexus 9000

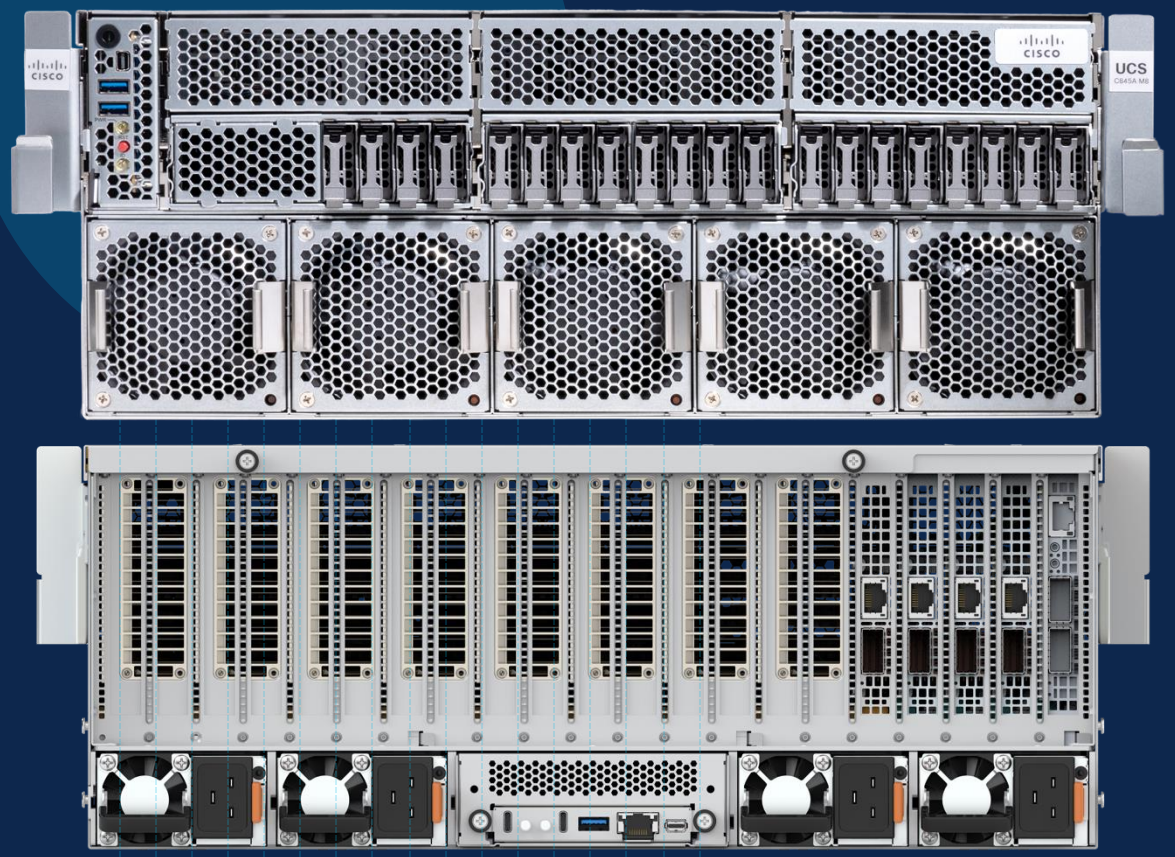


Cisco UCS C845A Versatile AI Rack Server

Bringing Scale Up and Scale Out Capabilities to the Cisco UCS® family and to Cisco's AI solution portfolio

Choose the number of GPUs that match the use-case, scale as your workloads demand it

Various Workloads including
Generative AI
Graphics and Rendering
Virtual Desktop Infrastructure

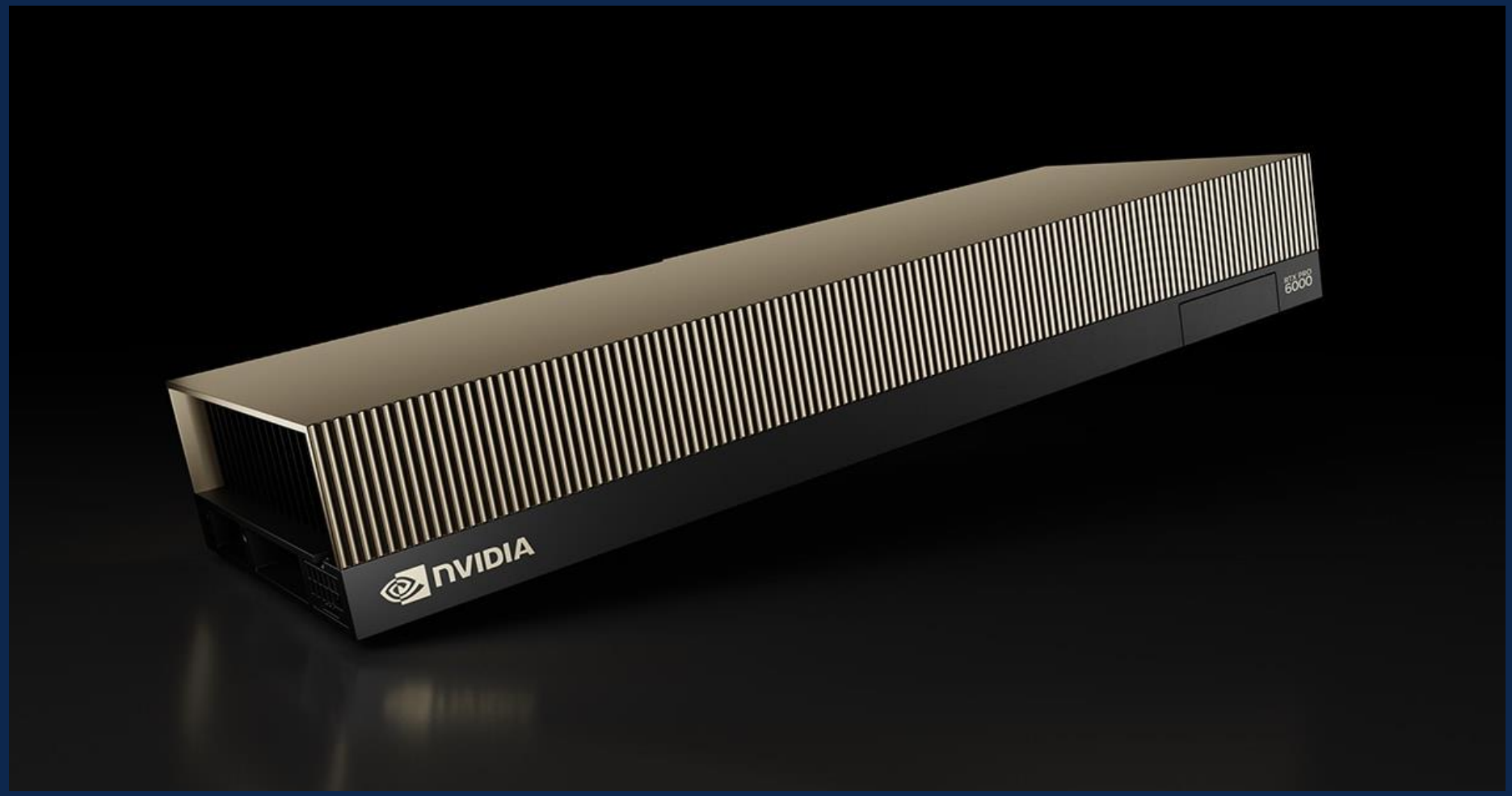


2-8x

Nvidia
H100, H200, L40S or RTX 6000 GPUs
Up to 700W/GPU
(B300 NVL on the roadmap)

Cisco UCS C845A

H200 NVL i RTX 6000 Pro Server Edition – PCIe



600-750W...

Next Gen UCS C240/C220 M8 Intel Rack Servers

UCS C240 M8 x24 SFF
24x 2.5-inch SAS3/SAS4/NVMe (RAID) front drives
Optional 8x 2.5-inch SAS3/SAS4/NVMe rear drives

UCS C240 M8 x32 NVMe
32x E3.s NVMe Drives Direct Attached to CPU
Optional 4x E3.s NVMe rear drives

UCS C240 M8 x12 LFF
12x 3.5-inch SAS3/SAS4/NVMe (RAID) front drives
Optional 4x 3.5-inch SAS3/SAS4/NVMe mid drives

UCS C220 M8 x16 NVMe
16x E3.s NVMe Drives Direct Attached to CPU

UCS C220 M8 x10 SFF
10x 2.5-inch SAS3/SAS4/NVMe (RAID) front drives
8x2.5-inch U.2 native NVMe drive support

Feature Highlights	Intel M8 2U/1U (2S)
CPU	• Dual Socket Granite Rapids-SP; Up to 80 Cores / CPU • 350W Max TDP (2U) / 300W Max TDP (1U) • UPI up to 24GT/s
Memory (DDR5)	• DDR5-6400, Total 32 DIMMs for Dual Socket: • 8 Channels per CPU, 2 DPC, 16 DIMM per Socket Monument Creek DDR-8000 • Supports All Capacities from 16 to 256GB* per socket
Storage	• Various RAID Controller Options • Native NVMe with E3.S
PCIe Gen5 Risers	• 8 PCIe Gen5 Slots - 2U • 3 PCIe Gen5 slots - 1U • Support for CXL • GPUs (350W) / NICs (75W) / FC (75W)
mLOM	• mLOM / PCIe VICs • Design for future mLOM / PCIe VICs (Gen5)

** - 2U Server only*

During M8 Timeframe:
Highest LFF Drive Capacity: 24TB
Highest E3.S: 30TB

2 RU rack limits

- when a GPU>75W is installed, CPUs with TDP greater than 300W are not supported (Intel) or greater then 320W (AMD)
- 256GB DIMMs cannot be combined with GPU cards, and the ambient temperature shall be limited to a maximum of 28C.
- With multiple GPU>75W installed, normal operating temperature is 30C [86F], lowered to 25C [77F], with a fan fault.

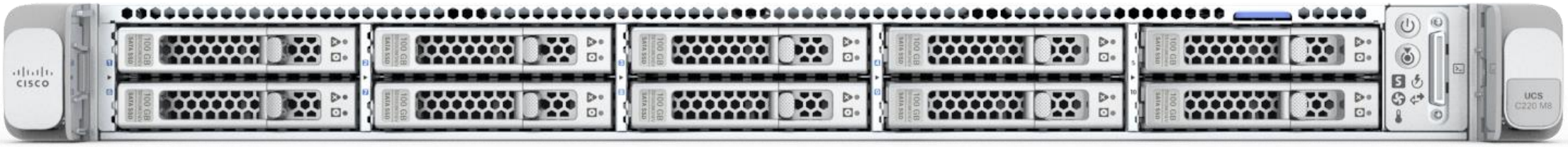
1 RU rack limits

- Platformy AMD występują wyłącznie jako jednoprocesorowe
  20 z 25 CPU AMD Epyc gen 5 ma TDP >=250W
- Platformy Intel – niektóre CPU występują wyłącznie jako jednoprocesorowe.
  12 z 25 CPU Intel Granite Rapids ma TDP >=250W
  Dla M8 Intel:

CAUTION:

- Normal operating temperature is limited to 35° C [95° F], and is lowered to 28° C [82.4° F], with a fan fault.
- When CPU>330W TDP installed, normal operating temperature is 30°C [86°F], lowered to 23°C [73°F] with a fan fault

- GPU jedynie 75W TDP



Nasza droga – Cisco Compute Strategy

Kilka cech, które dokonały potężnej zmiany i na stałe weszły do świata serwerów

Wirtualne interfejsy

Po raz pierwszy to administrator decyduje ile, jakich interfejsów i w jakiej kolejności będzie posiadać serwer. Wirtualne interfejsy dopasowane w 100% do potrzeb systemu

FCoE / Unified Fabric

Dołączanie elementów sieci FC przy wykorzystaniu sieci LAN. Znaczące ograniczenie ilości potrzebnego okablowania

Centralne zarządzanie

Cały system zarządzany z jednego miejsca. Pełna wiedza o zdarzeniach w całym środowisku, szybkie aktualizacje, konfiguracje serwerów rack i blade w ramach jednego systemu

Profile serwisowe

Definicja docelowej konfiguracji serwera w postaci profilu. System dopasowuje konfigurację sprzętu do wymaganego stanu. Zarządzanie za pomocą szablonów pozwala skonfigurować kilkadziesiąt serwerów kilka minut

Fabric interconnect

Znaczące ograniczenie ilość potrzebnych przełączników i okablowania. Centralny punkt przełączania dla całego systemu, stanowi miejsce rozdzielania konfiguracji sieciowych dla wszystkiich serwerów

2009

202

System modularny – Cisco UCS-X

Otwarty, modularny design

New nodes

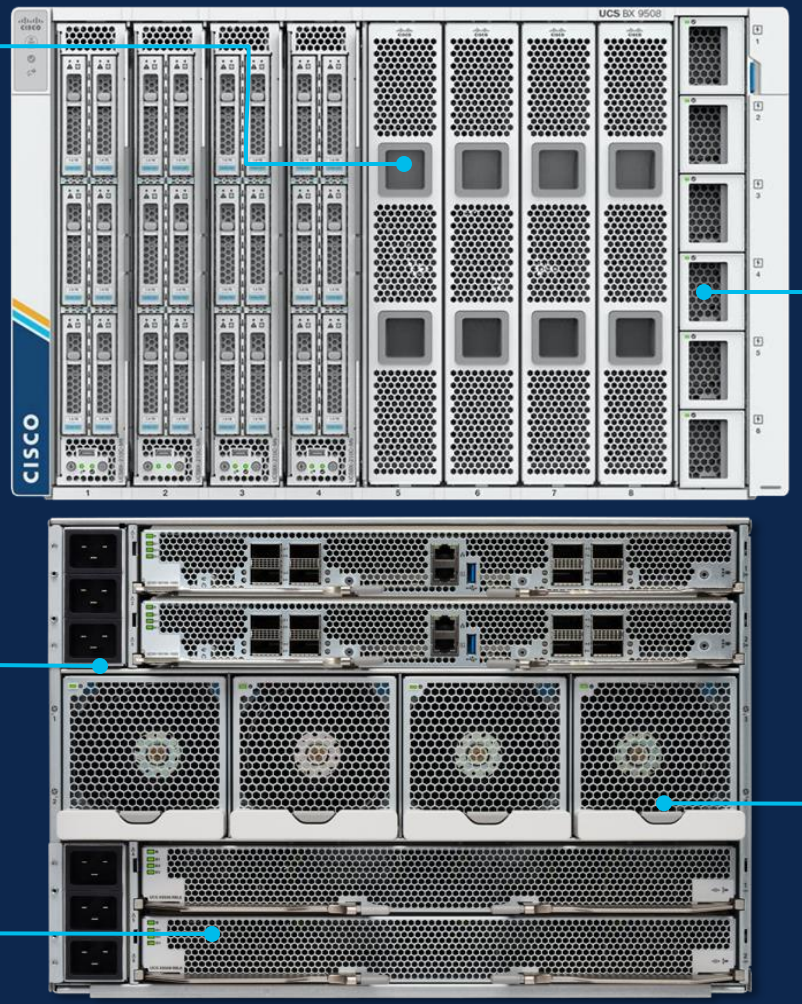
- Compute Nodes
- Accelerator nodes - GPU

Future I/O

- New SmartNIC Features, 100Gb+

UCS X-Fabric Technology

- PCIe Gen4/5/6
- CXL, Smart/Super NIC



Power and Cooling Margins

- Ready for high-watt CPUs and GPUs
- Liquid colling enabled

Architecture which enables:

- Modularity for seamless integration of Next Innovations
- Efficiency of power and cooling
- Balanced Compute and Drive Ratios
- Optimal Failure Domains
- High Performance

* Roadmap – subject to change

Moduły obliczeniowe i GPU dla UCS X-Series

UCS® X210c M7 Compute Node
Flexible server for all your workloads
With the latest **Intel** CPUs

UCS® X215c M8 Compute Node
Flexible server for all your workloads
With the latest **AMD** CPUs

UCS® X210c M8 Compute Node
Flexible server for all your workloads
With the latest **Intel** CPUs

UCS X210c Compute Node – GPU
Up to two on-server GPUs

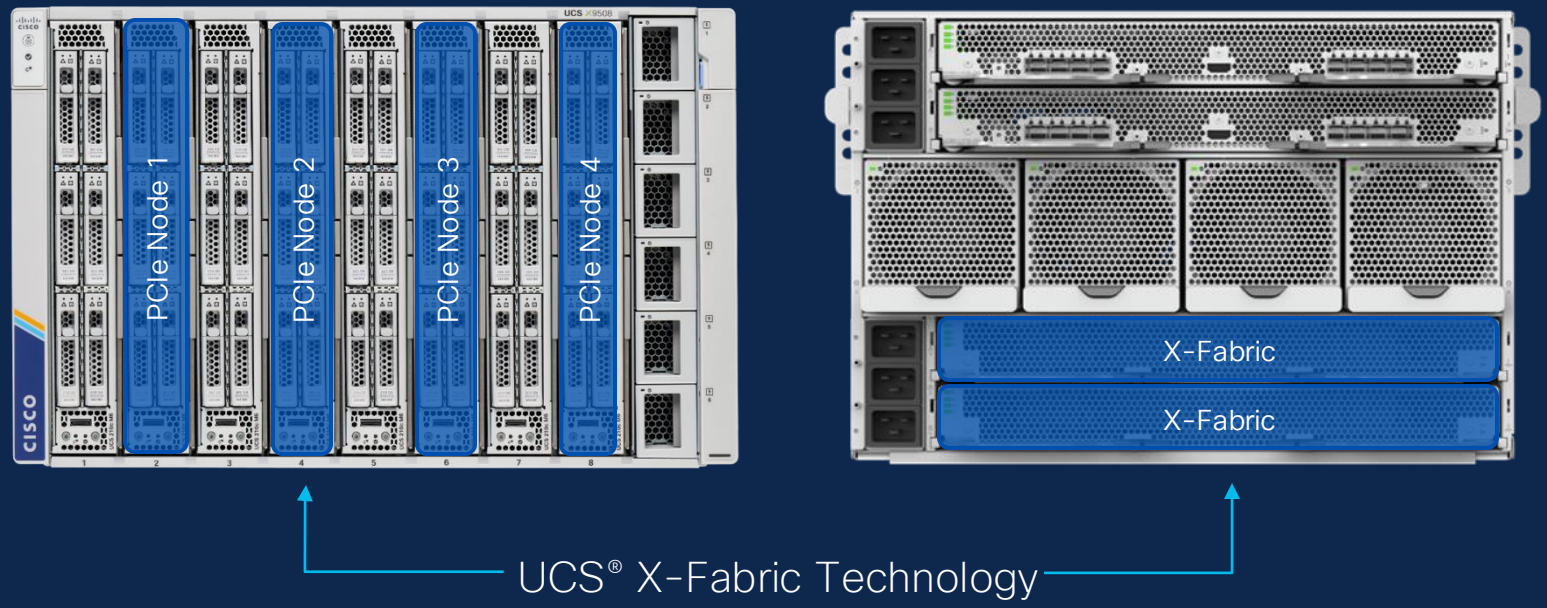
UCS X410c M7/M8 Compute Node
Designed for scale-up applications

UCS X440p PCIe Node
Up to four GPUs

UCS X-Fabric Technology and PCIe Nodes with GPU

PCIe node supports up to

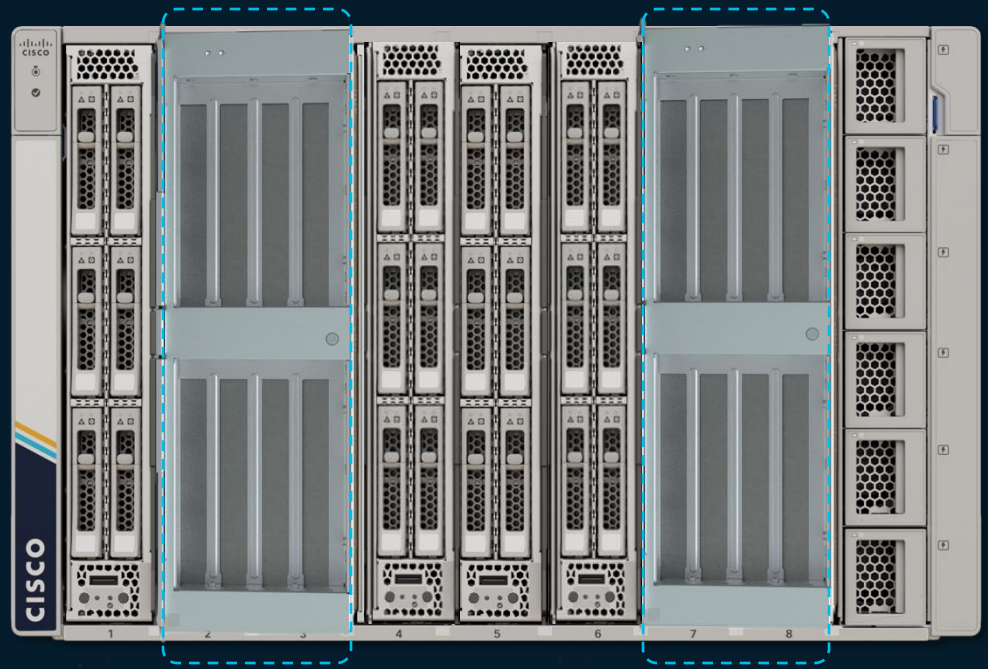
- 4x Intel® Data Center GPU Flex 140
- 2x Intel Data Center GPU Flex 170
- 2x Nvidia A100
- 2x Nvidia A16
- 2x Nvidia A40
- **2x Nvidia H100-NVL**
- **2x Nvidia L40S**
- 4x Nvidia T4/**L4**



- ✓ Based on native PCIe Gen 4
- ✓ Provides GPU acceleration to enterprise application
- ✓ No backplane or cables = Easy upgrades

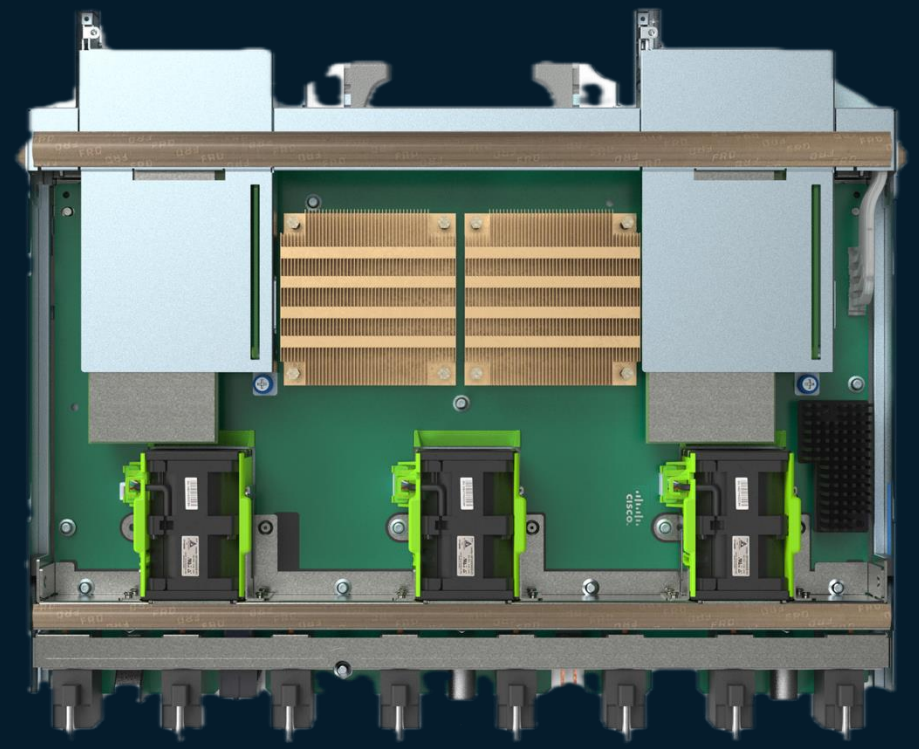
X580p PCIe Node

- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
    - Nvidia H200-NVL, L40S, A16 RTX Pro 6000 (96GB), B300-NVL
- Support multiple vendors: Nvidia, AMD*
- NVLink bridge support
- Support up to 600W FHFL GPU
- Managed PCIe node with BMC support
- Policy based GPU management
- Ability to support 2x GPUs per Compute node



X-Fabric gen 2

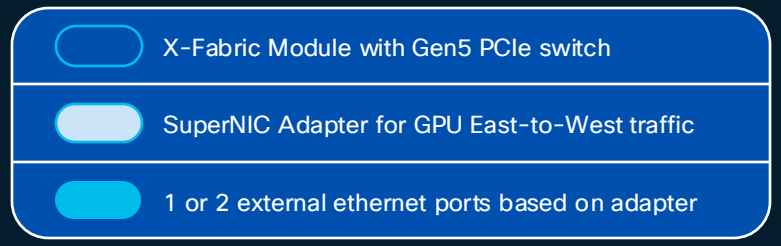
- PCIe Gen5 Switching
- 2x CEM Slots to support HHHL NIC cards
- Managed XFM Modules with BMC support
- GPU Direct Support over RDMA
- GPU Backend(East-West Traffic) network support

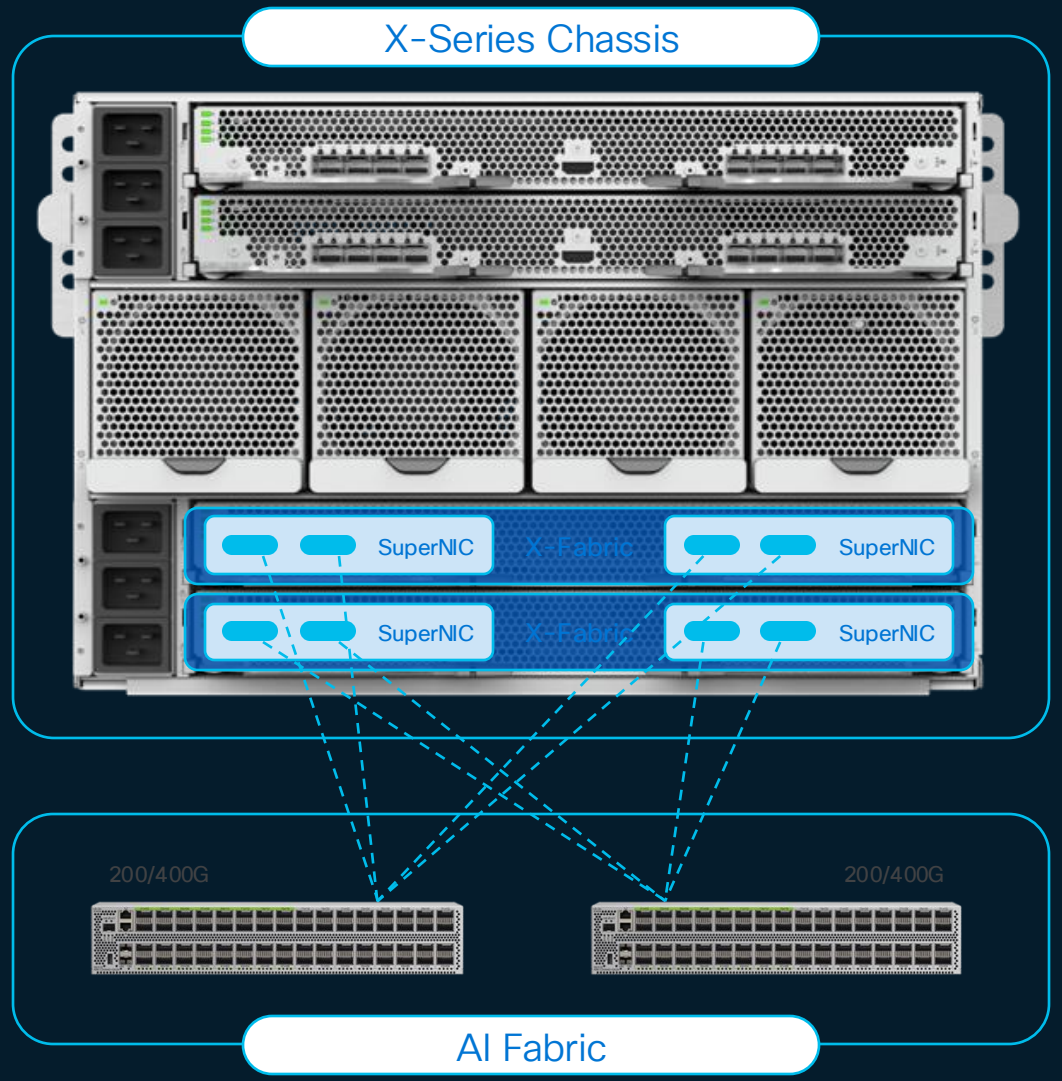


AI cluster expansion

GPU-to-GPU connectivity

with XFM external ports and Nvidia SuperNIC CX-7 (2x200G/1x400G)





AI PODs

Faster time to value with pre-configured bundles

Orderable Now

Deploy AI with confidence

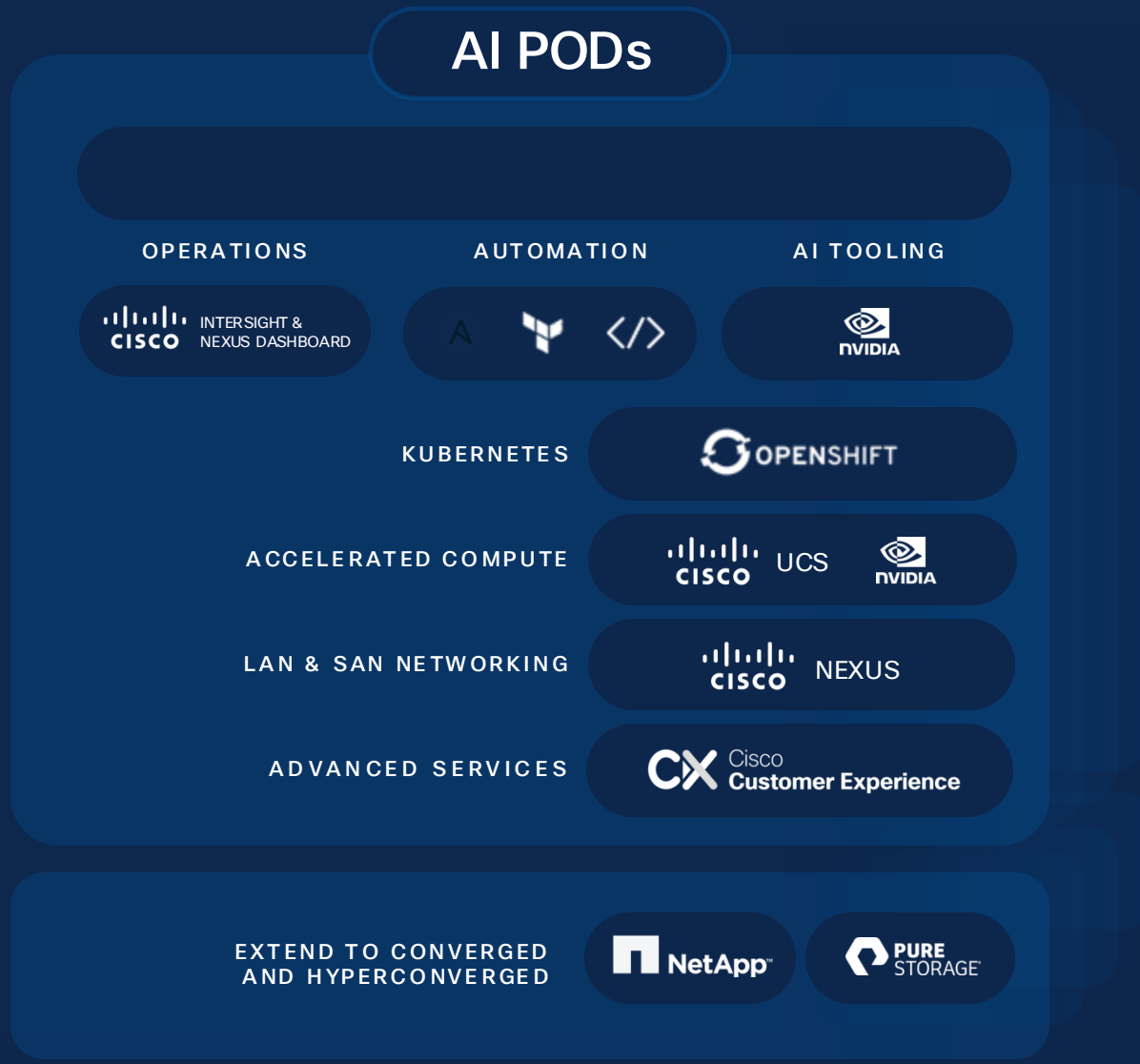
Orderable, validated AI-Ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

AI Advisor tool for configuration guidance

COMING SOON

Cisco AI-Ready Infrastructure Stacks



Choice of AI Pod Size based on GPU

Edge Inferencing AI Pod

Llama 2 7B requires **41GB** of GPU memory, with a *batch size of 50*, a *sequence length of 2048*, and a *precision of FP16*.

7*2 GB +
50*2048*2*32*4096*2/2 B
= 41 GB

1x L40S:

- 48 GB of GPU Memory
- GDDR6 memory (20 GB/s)
- 733 TFLOPS (FP16/BF16)

Large-Scale RAG AI Pod

Llama 3.1 70B requires **174GB** of GPU memory, with a *batch size of 50*, a *sequence length of 2048*, and a *precision of FP16*.

70*2 GB +
50*2048*2*80*8129*2/2 B
= 274 GB

4x H100 NVL:

- 376 GB of GPU Memory in total (94GB/GPU)
- HBM3 memory (819 GB/s)
- 1513 TFLOPS (FP16/BF16)

AI/ML Sizer

Summary

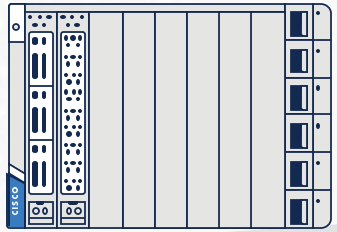
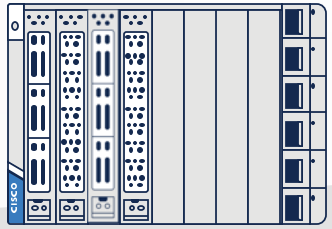
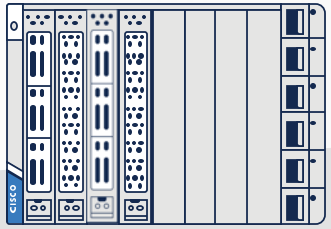
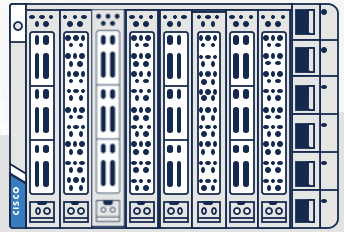
	Option 1	Option 2
Server Count	1	1
Server Type	X210c M7	C240 M7 SFF
CPU Type	6548Y+ (32 core)	6548Y+ (32 core)
Server Memory (GB)	256 (16x16GB)	256 (16x16GB)
GPU Type	L40S	H100
Total GPU Count	1	1
Memory (GB) / GPU	48	80
TFLOPs / GPU	362.05	756
Check Price	☐	☐
Build Config	☐	☐

Performance Metrics (values are per GPU)

Time to First Token (TTFT) (ms)	6684.81	3289.40
Time Per Output Token (TPOT) (ms)	27.55	27.84
Latency (Response Time) (s)	34.86	31.77
Throughput (Tokens/s)	1815	1795
Batch Size	50	50
Model Memory Size (GB)	13.48	13.48
Activation Memory Size (GB)	25.00	25.00
Total Memory Required (GB)	38.48	38.48
GPU Memory Utilization (%)	**80.17**	**48.10**

Cisco AI PODs (UCS-X based)

	Edge Inferencing (7B-13B Parameter)	RAG Augmented Inferencing (13B-40B+ Parameter)	Large-Scale RAG Augmented Inferencing	Scale-Out Inferencing Cluster (Inferencing Multiple Models)
Typical use case	Small	Medium	Large	Scale-Out
Hardware specification	1x X210C compute node • 2x Intel 5th Gen 6548Y+ • 512 GB System Memory • 5x 1.6 TB NVMe drives • 1x X440p PCIe • 1x NVIDIA L40S	2x X210C compute nodes • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA L40S	2x X210C compute nodes • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA H100 NVL	4x X210C compute nodes • 8x Intel 5th Gen 6548Y+ • 1.5 TB System Memory • 12x 1.9 TB NVMe drives • 4x X440p PCIe • 8x NVIDIA L40S

Performance and Scale

Inferencing Suite

AI/ML with UCS and UCS X-Fabric Technology

Model pre-training

Fine-tuning pre-trained model

Inference / deploy

I/O intensive

Compute intensive

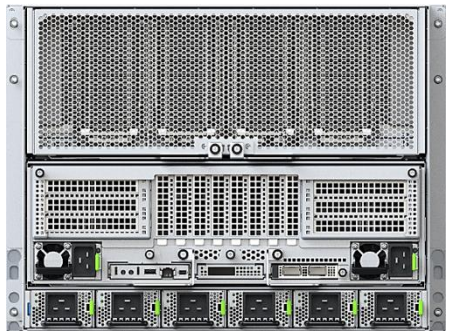
Latency sensitive

NVIDIA H100/H200/B200

NVIDIA H100-NVL, L40S, L4

NVIDIA H100, A100, L40, L4

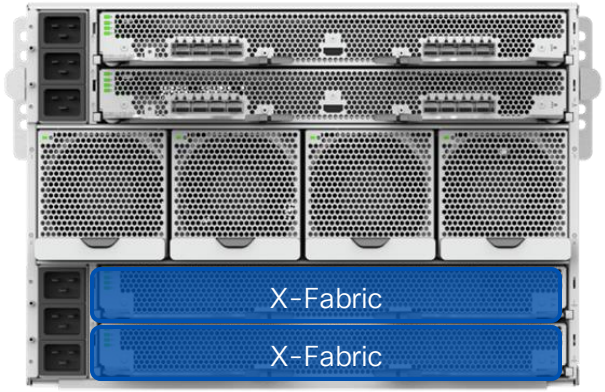
C845A

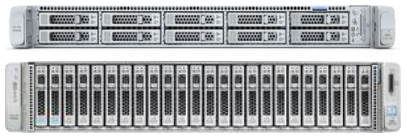


C885A

X210c M8

X440p





C2x0 M8 or UCS-X

UCS® X-Fabric Technology

NVIDIA AI Enterprise

- Based on native PCIe Gen 4, upgradable to gen5/CXL in future
- Provides GPU acceleration to enterprise applications

Cisco Intersight®

Reliability when it matters most

Cisco Tops List for Server Reliability

ITIC

According to a survey by the **Information Technology Intelligence Consulting Corp (ITIC)** reported by Tech Channel

TechChannel®

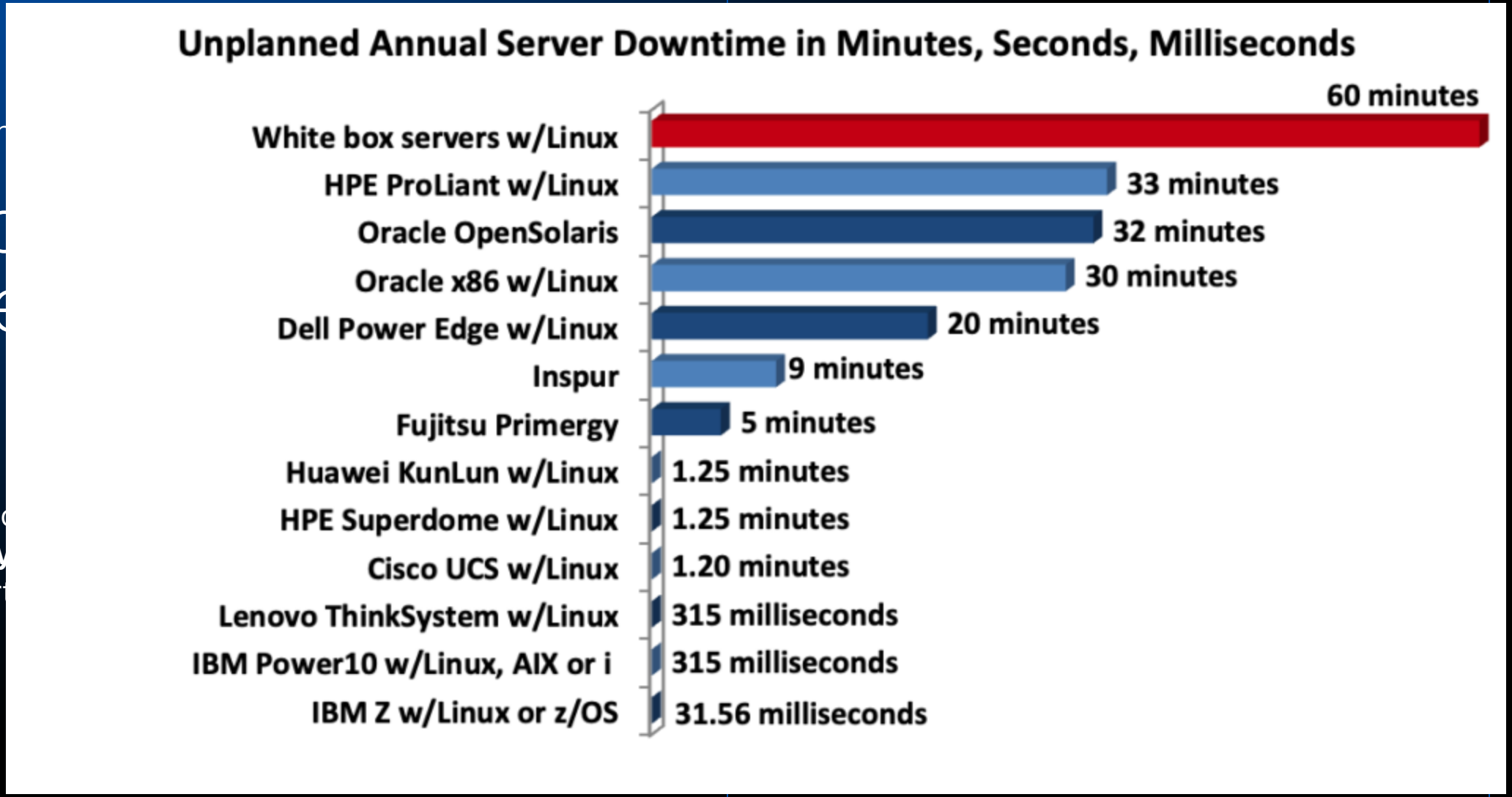
Server Reliability Survey: IBM, Lenovo, Cisco Top ITIC's List

Reliability wh

Cisco Serve

ITIC

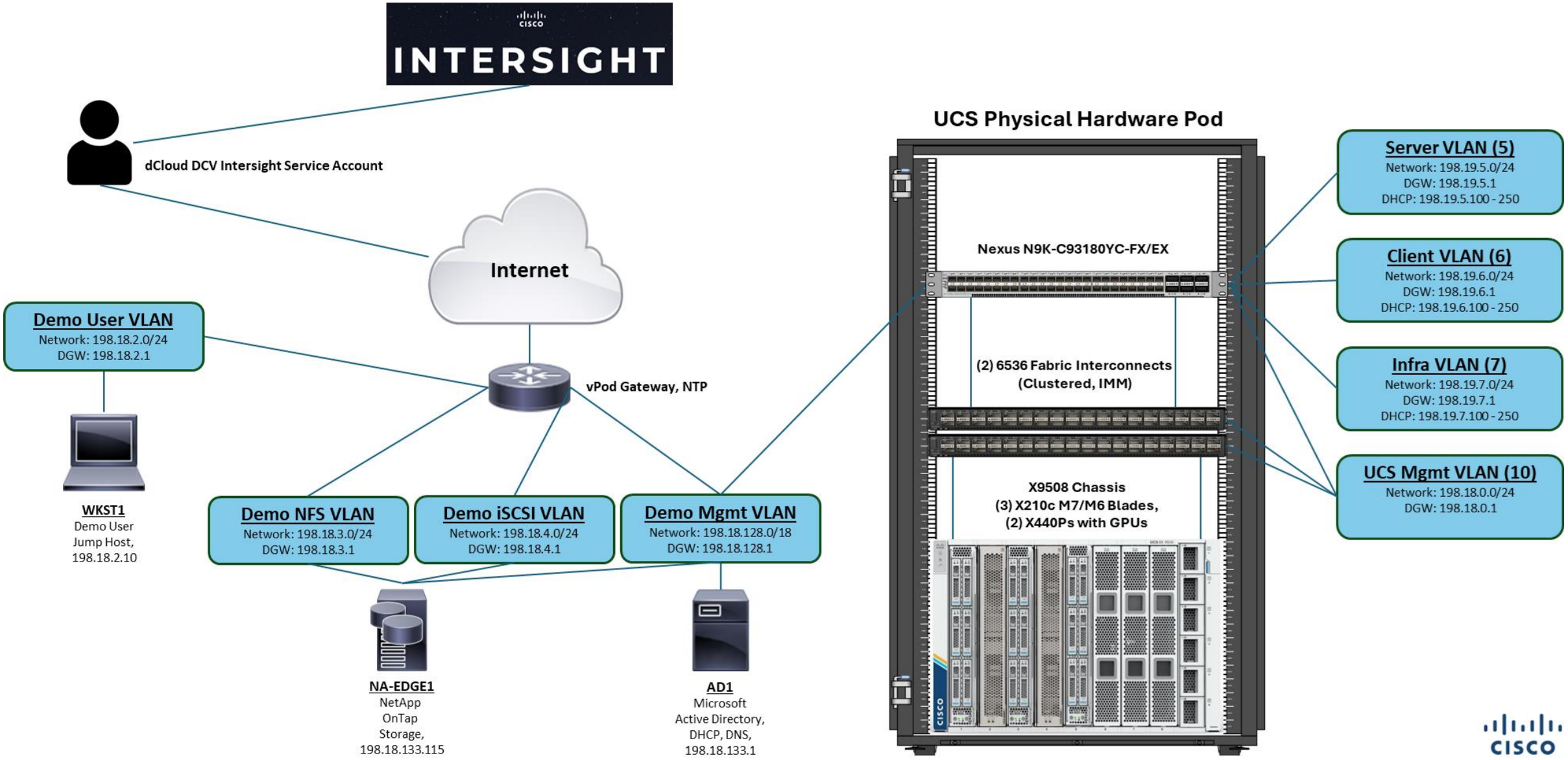
According t
Technology
(ITIC) report



Dlaczego Cisco UCS-X?

- Ścisła kooperacja Cisco z Nvidia
- Uniwersalna platforma UCS-X – nie tylko pod AI
- **Wsparcie GPU>400W**
- Jeden punkt wsparcia na sieć/serwery (i ew. Framework AI – NVAIE/RHEL)

Running Gen AI and LLMs on Cisco UCS X-Series v1
dcloud lab

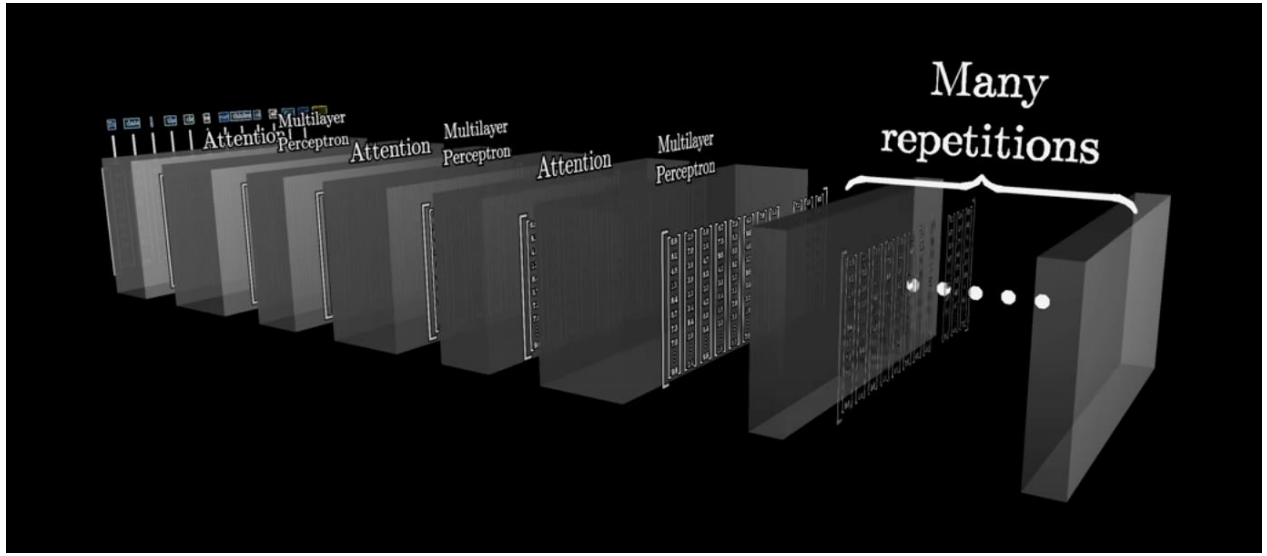


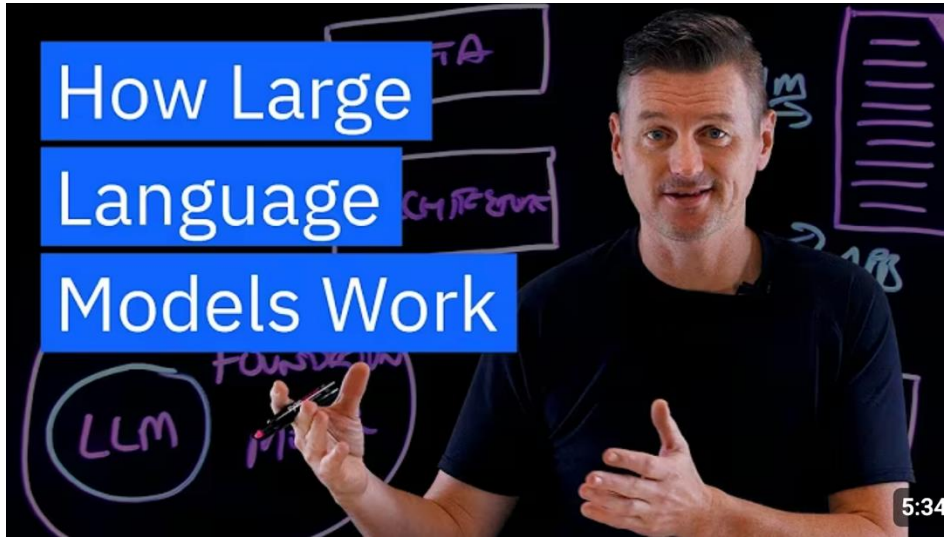
Scenariusze

- Automated UCS X-Series Provisioning & OS Installation with Intersight
- Installing GPU drivers, Generative AI/ LLM libraries and container ecosystem
- Installing AI Inferencing server with API, WebUI and LLMs
- **Using a WebUI app with the Inferencing Server API**
- **Performing Retrieval Augmented Generation (RAG) on local documents**
- **RAG – YouTube video summarization**
- **Installing and running Stable Diffusion Image Generation**
- Adding additional AI Models to the system
- Running NVIDIA NIM workloads

Materialy - Youtube

Neural networks - 3Blue1Brown channel
https://www.youtube.com/watch?v=eMlx5fFNoYc&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi





How Language Models Work - IBM channel
https://www.youtube.com/watch?v=5sLYAQS9sWQ

Zarzadzanie UCS


cisco
Partner
NETFORMERS
Dziekujemy!