

From Fiction to Science

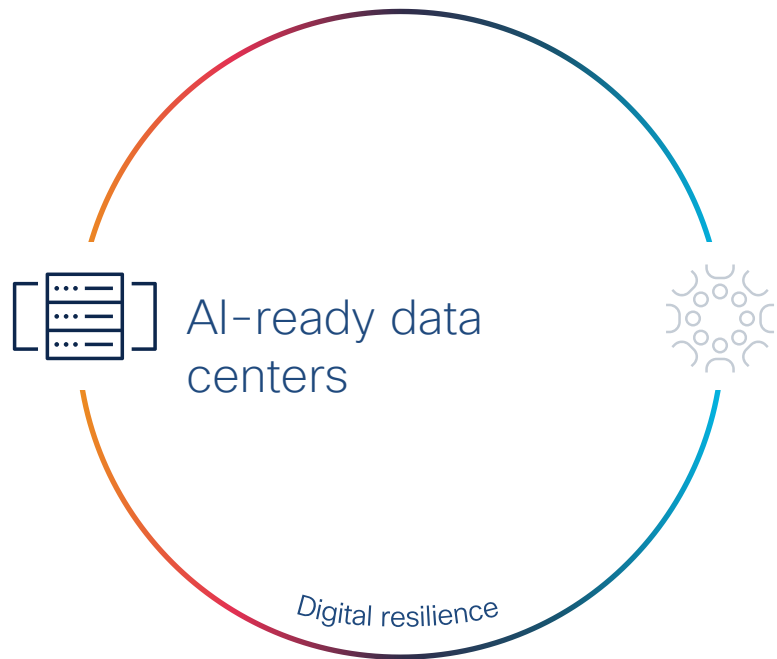
Compute for AI Deep Dive

Noah Donaldson - AI SE

Ben Flowers - SE

Transform data centers to power AI workloads anywhere

Public and private clouds, edge, on-premises



Comprehensive infrastructure

Power AI with networking, compute, and storage in fully-integrated, scalable, and modular systems for all workloads

Seamless operations and observability

Remove silos with unified management, observability, and assurance for traditional and AI workloads, across all environments

Security from ground to cloud

Protect hyper-distributed workloads by infusing security everywhere

What we thought AI would look like



Monumental Moments

Artificial Intelligence



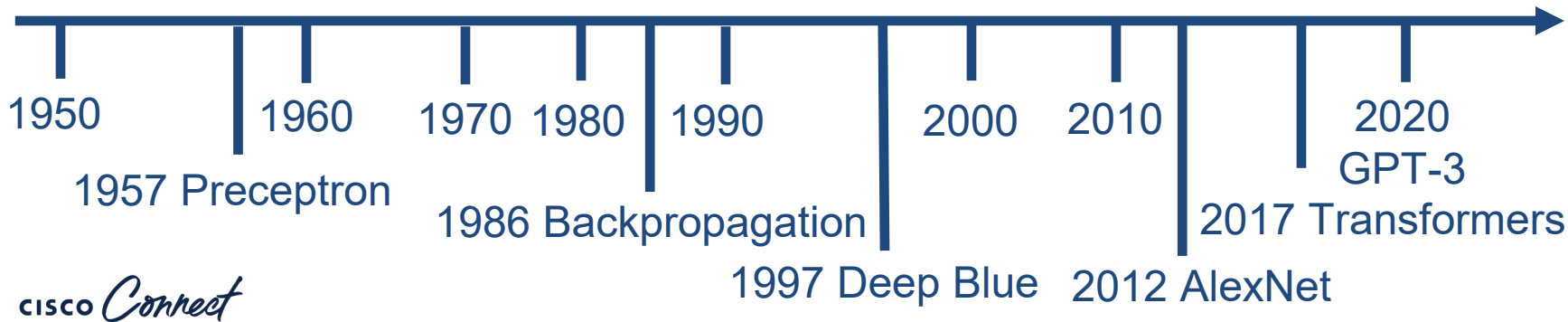
Machine Learnings



Deep Learning



Generative AI

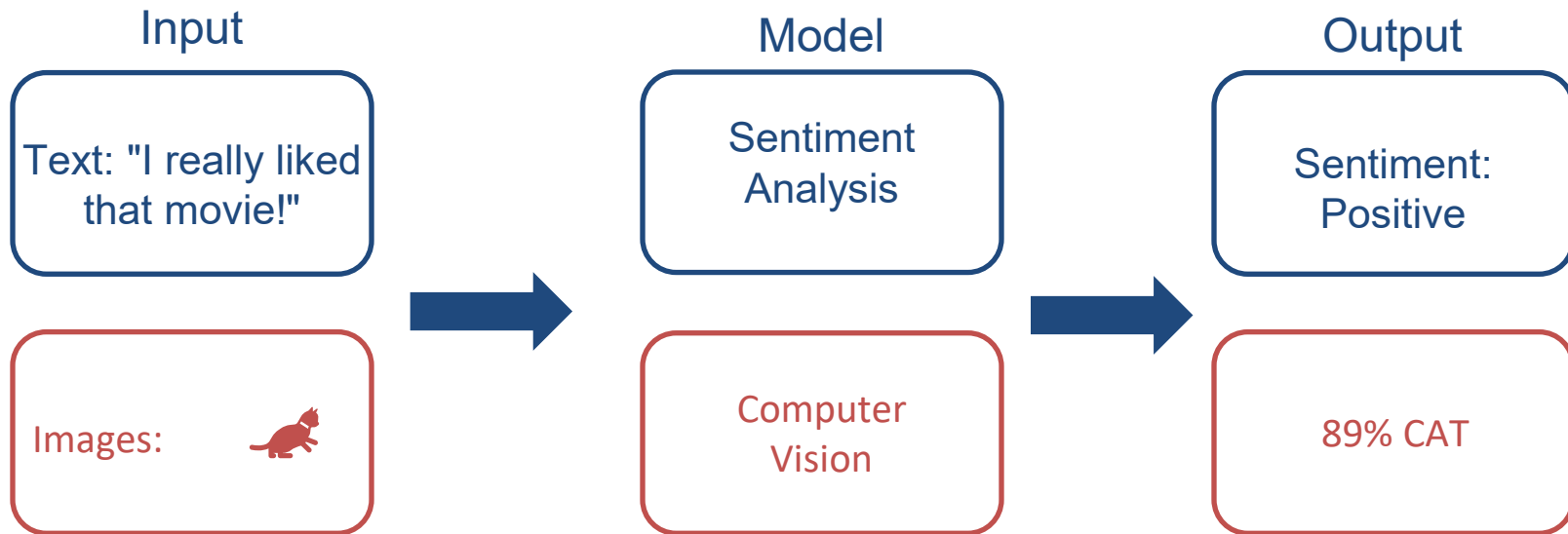


What AI looks like...today

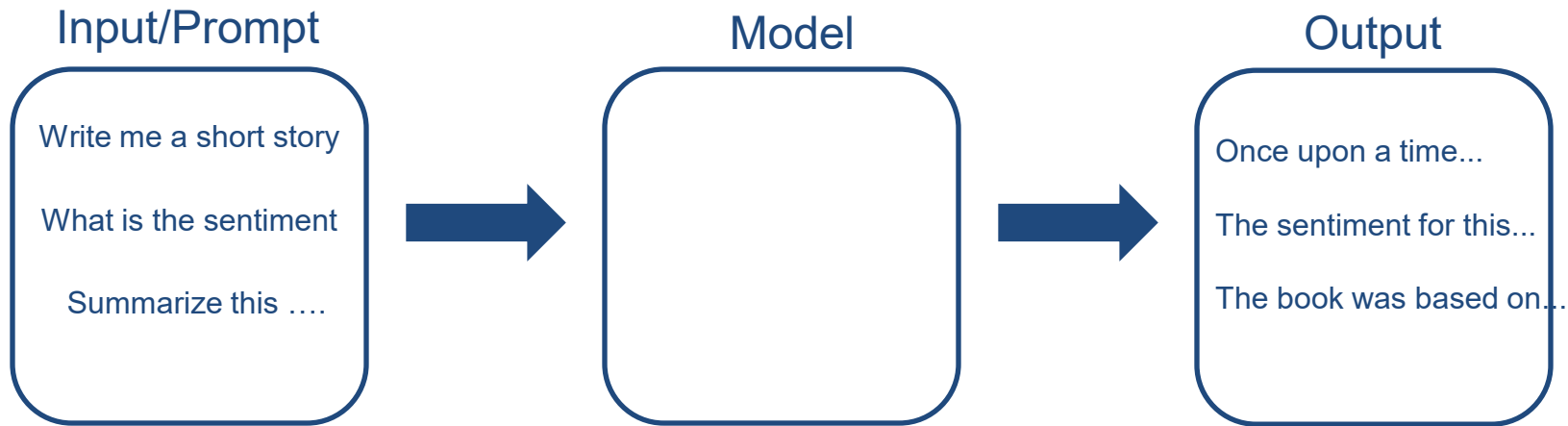
From healthcare to transportation, AI is rapidly moving from the lab to daily life. In 2023, the **FDA approved 223** AI-enabled medical devices, up from just six in 2015. On the roads, **self-driving cars** are no longer experimental: Waymo, one of the largest U.S. operators, provides over **150,000** autonomous rides each week, while Baidu's affordable Apollo Go robotaxi fleet now serves numerous cities across China

New research suggests that machine learning hardware performance, measured in 16-bit floating-point operations, has **grown 43%** annually, **doubling** every 1.9 years. Price performance has improved, with **costs dropping 30%** per year, while **energy efficiency** has **increased by 40%** annually.

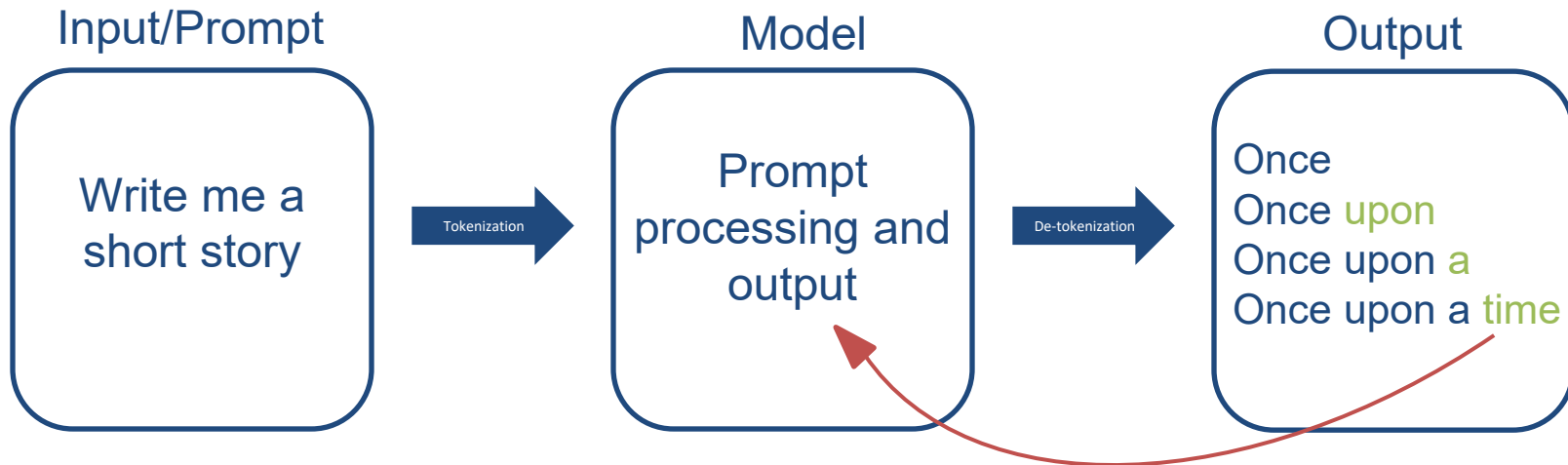
What does inference really mean?



How does LLM inferencing work



How does LLM inferencing work



How does tokenization work

- Converts input text into a vector of integers
- Each token in a vocabulary has a corresponding integer
- A token is roughly four characters

T1	T2	T3							TN
1	2	3							N

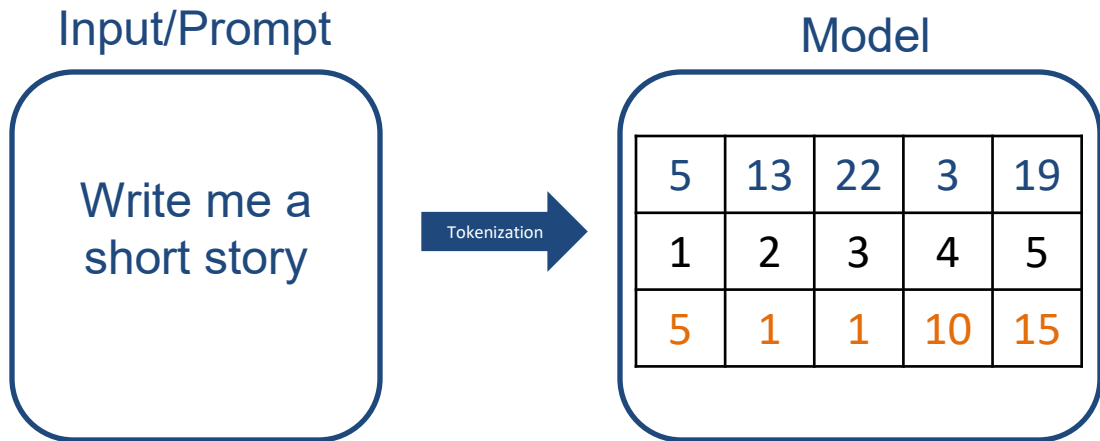
Write me a short story



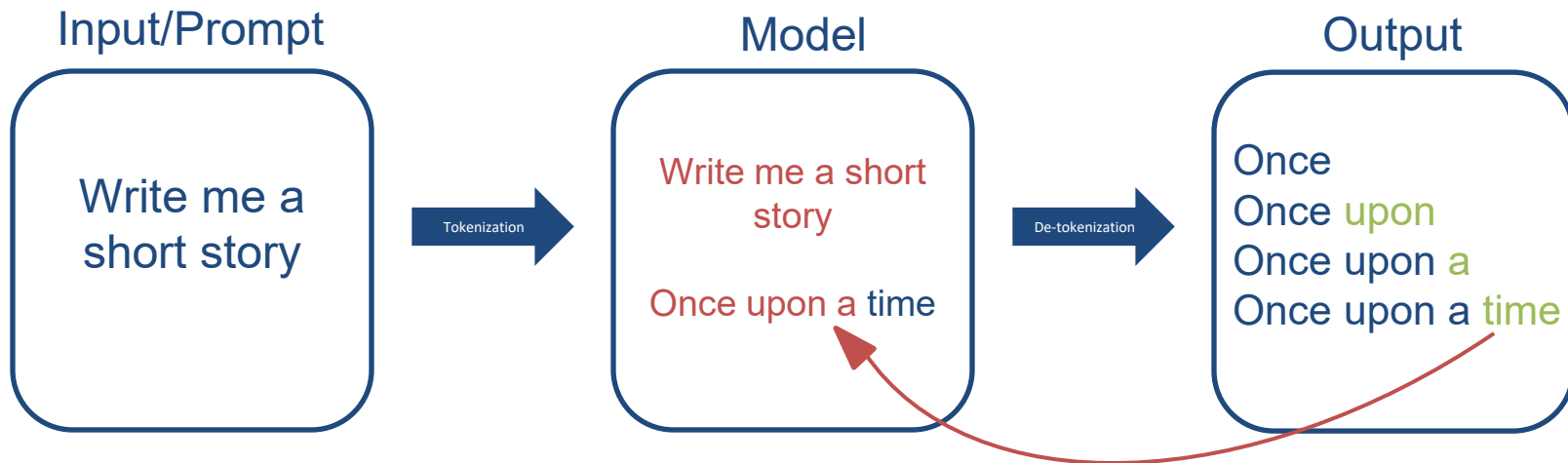
5	13	22	3	19
1	2	3	4	5

Prompt processing, Prefill, and Attention

- Numerical embeddings are supplied to the model
- The model uses these to establish context
- Attention focus on relevant parts and assigns **weights**



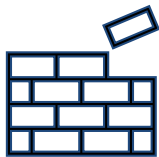
Output - Decode and Cache



- Each token gets generated using the processed prompt
- The last token generated is included in the input
- Attention vectors called KV Cache are stored in memory to speed up further token generation
- The output is the next best token based on probability

Applying this to a GPU

Model parameter size



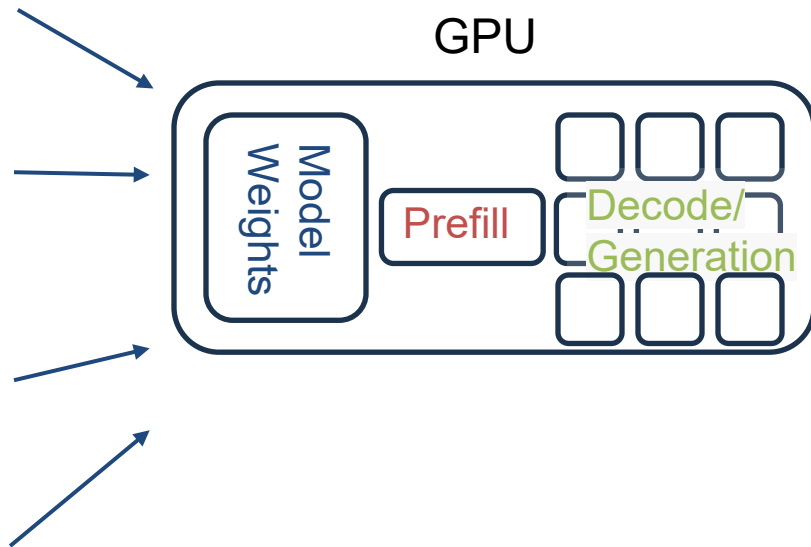
Model precision



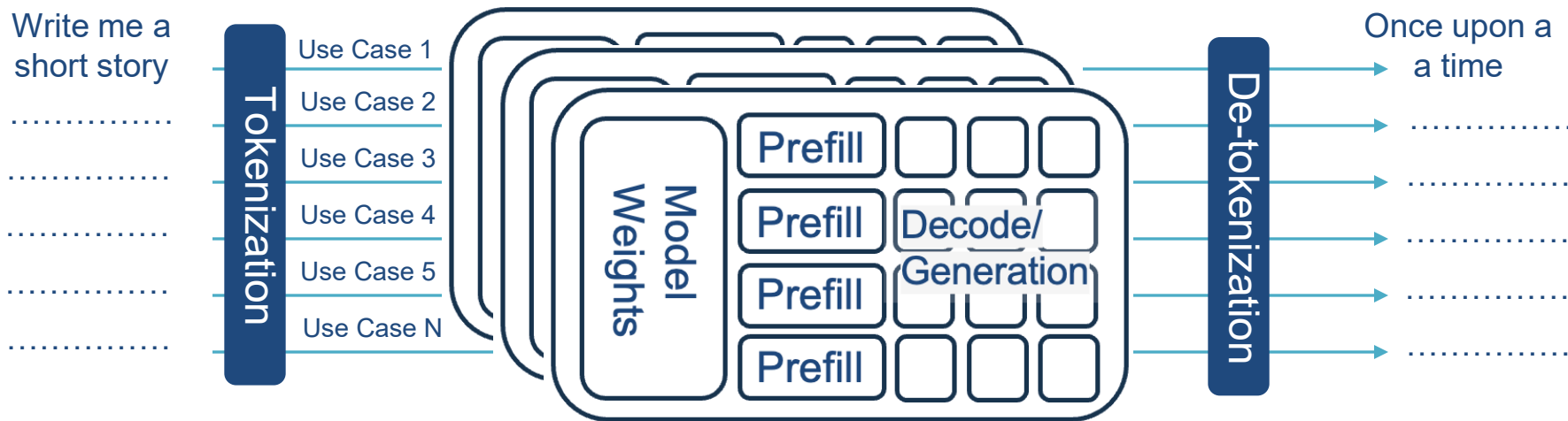
Context length



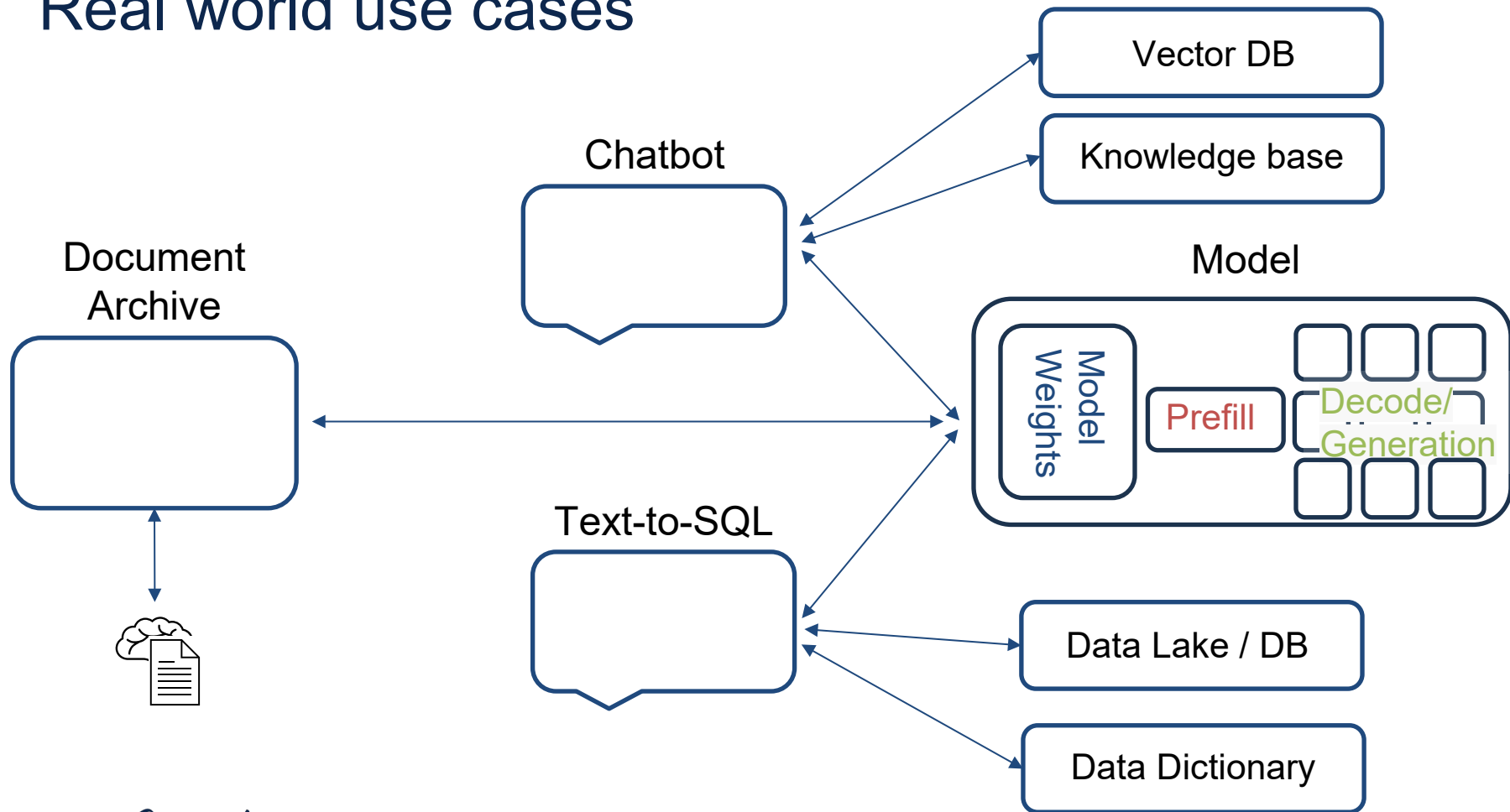
KV Cache



Serving multiple use cases



Real world use cases



Using AI Model Cards

- Details model purpose, performance and usage considerations
- Use as a guide to understand model capabilities and limitations
- Dictates server and GPU requirements

The screenshot shows the Hugging Face interface for the model **mistralai/Mistral-7B-Instruct-v0.2**. The model card is highlighted with a blue box, and the 'Files and versions' tab is also highlighted. Two blue callout boxes provide additional context:

- Model Card Callout:**
 - Contains # of parameters
 - # tokens (words) model was trained on
 - Links to whitepaper on model
- Files and versions Callout:**
 - Size in GB of model (how much GPU memory required)
 - License (IE: apache-2.0)

The interface also shows the 'Model Card for Mistral-7B-Instruct-v0.2' section, the 'Downloads last month' statistic (1,261,845), and a line graph showing the download trend.

Converting Model Card Details into GPU Requirements

Number of parameters determines GPU memory requirements (fp32)

- 4GB of GPU RAM per 1B parameters for inferencing. Parameter = 4 bytes
- 24GB of GPU RAM per 1B parameters for training. Parameter = 4 bytes + 20 bytes context

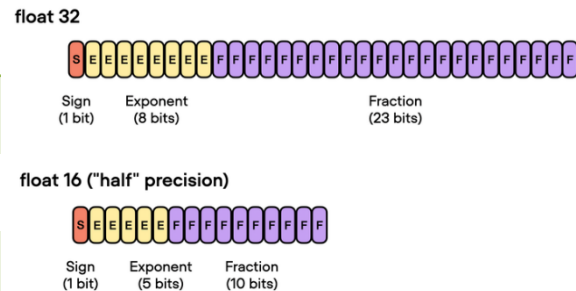
If an enterprise class GPU has 80GB of high bandwidth memory:

- Inferencing overrun at 20B parameters
- Training overrun at 3.3B parameters

Understanding Quantization

- Technique to reduce computational and memory requirements of models by lowering precision of the weights
- Changing from 32-bit to 16-bit to 8-bit etc., reduces the memory needed
- Ongoing discussions on if it is better to reduce number of parameters or change quantization

1B Params	Inferencing	Training
Float 32 (32b)	4GB	24GB
Float 16 (16b)	2GB	12GB
Int 8 (8b)	1GB	6GB



GPU Sizing

Based on Model Card

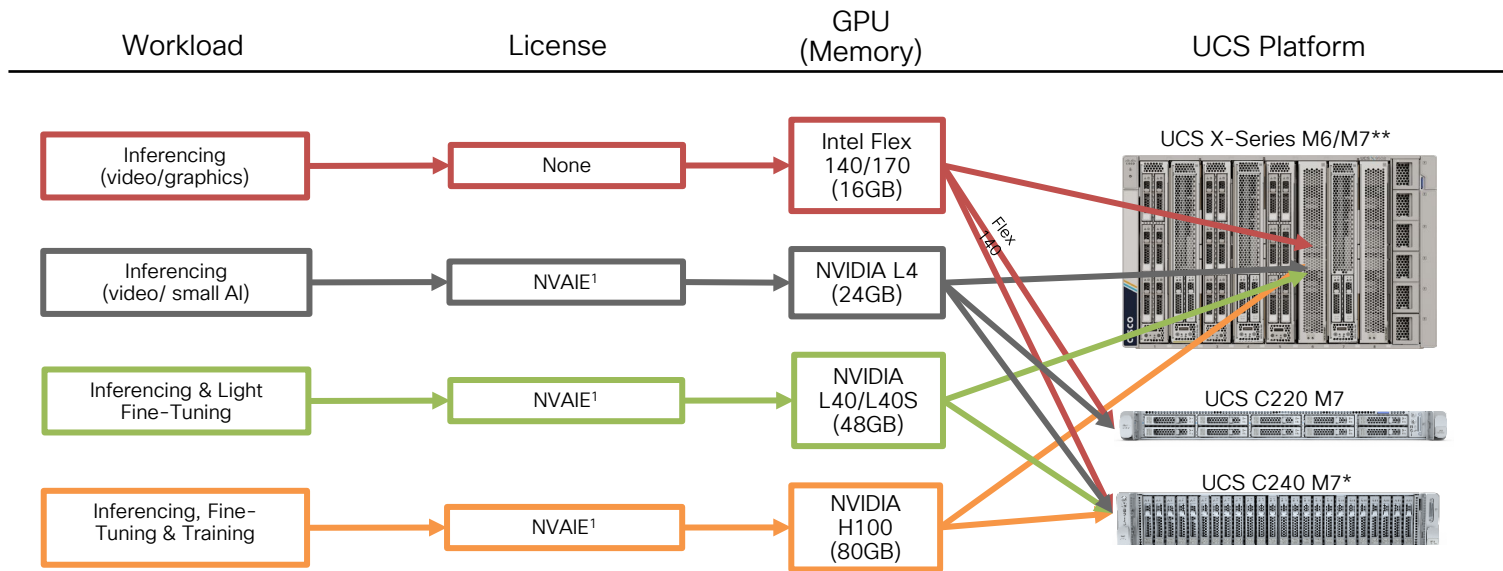
* Amount of GPU high bandwidth memory (HBM) is the most important factor

Model Size (Billions of parameters)	Float32	Float16	GPTQ 8bit	GPTQ 4bit
7B	28 GB	14 GB	7 GB	3.5 GB
13B	52 GB	26 GB	13 GB	6.5 GB
32B	128 GB	64 GB	32 GB	16 GB
65B	260 GB	130 GB	65 GB	32.5 GB

2 GB per billion parameters

AI GPUs for UCS

Platform Support



¹NVIDIA AI Enterprise (NVAIE) license required if workload virtualized

* T4 is also supported in C220 M6 and A10/A20/A40/A100/A16 are supported in C240 M6 which are not listed on this slide

** T4, A16, A40, A100 in X440p is supported with X210c M6; L4, L40, H100 in X440p is supported with X210c M7

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 A16

2x

Nvidia H100

Nvidia L40

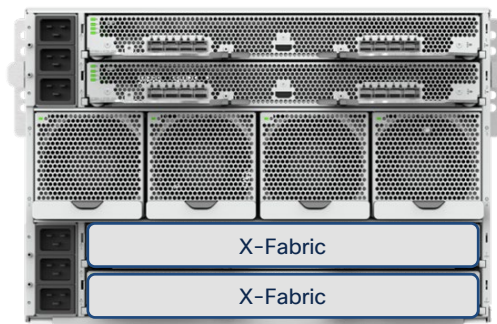
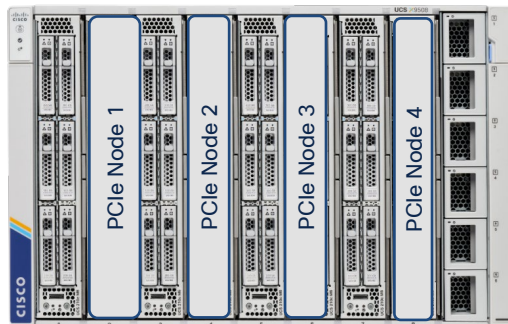
Nvidia L40S

Nvidia A40

4x

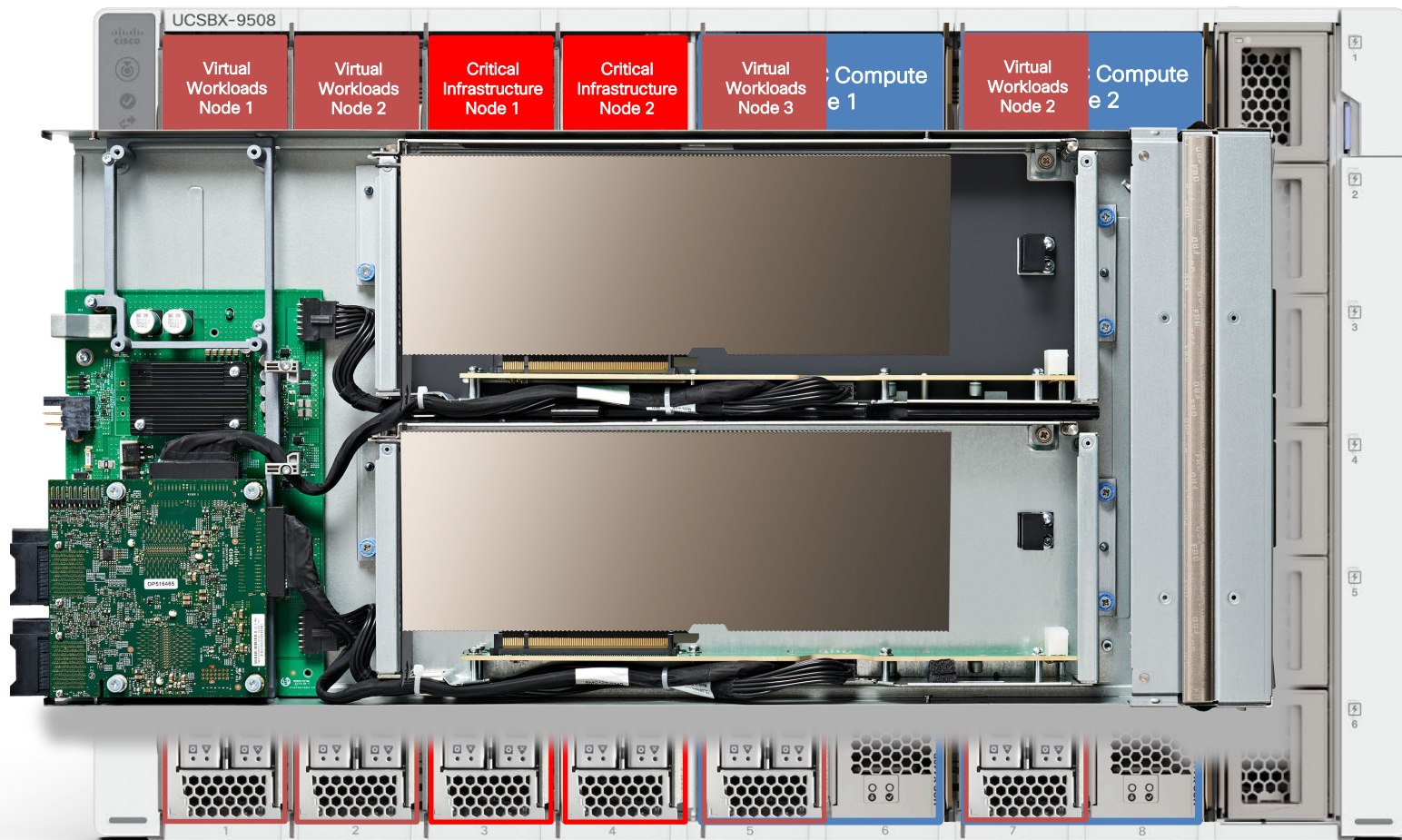
Nvidia T4

Nvidia L4



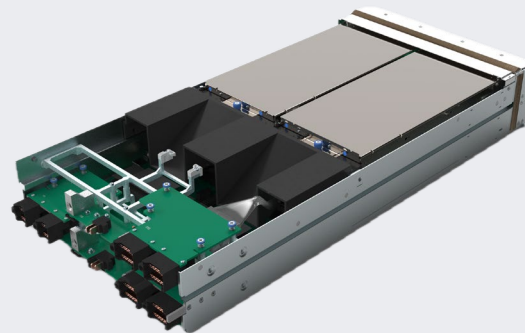
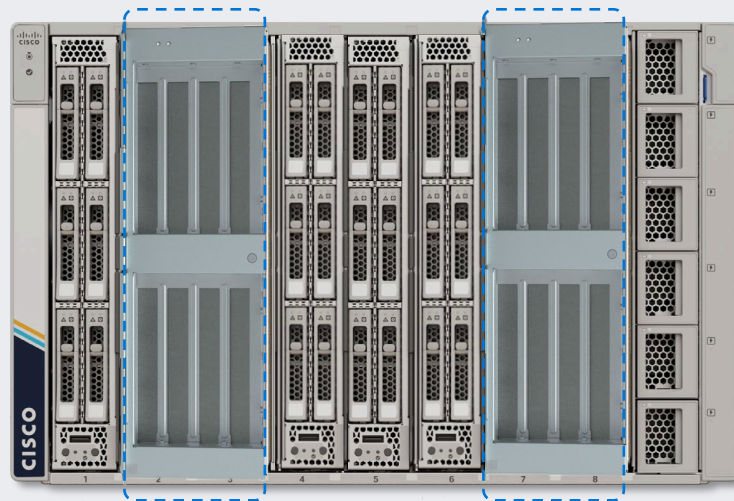
UCS X-Fabric Technology

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



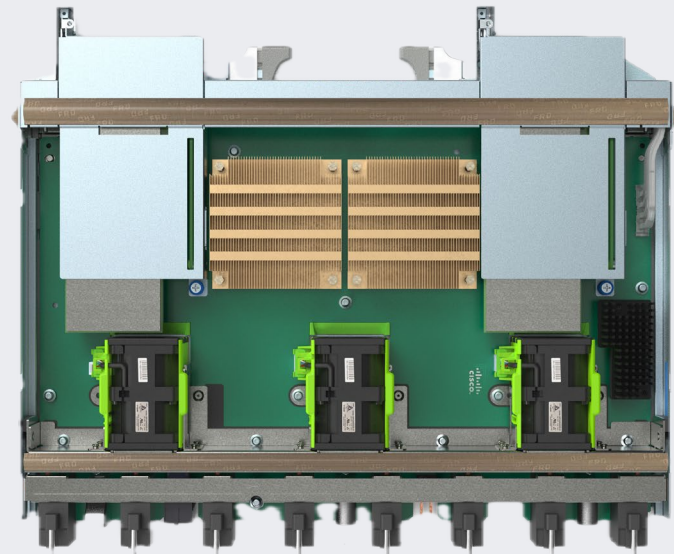
X580p PCIe Node

- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
 - Nvidia H200-NVL, B100-NVL, B40, A16
- Support multiple vendors: Nvidia, AMD*/Intel*
- 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 9516


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




AI cluster expansion

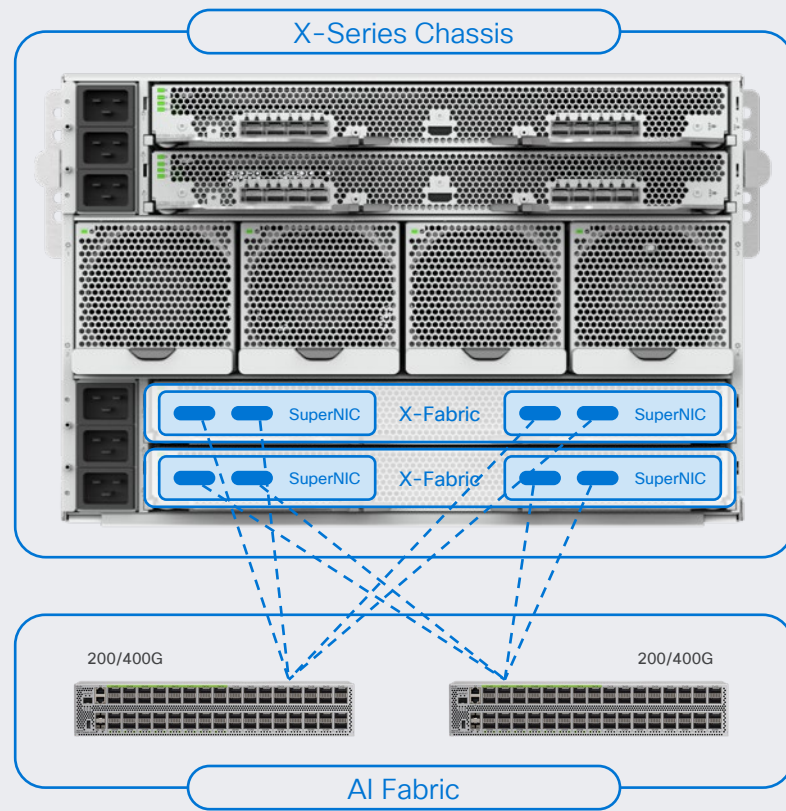
GPU-to-GPU connectivity

with XFM external ports

 X-Fabric Module with Gen5 PCIe switch

 SuperNIC Adapter for GPU East-to-West traffic

 1 or 2 external ethernet ports based on adapter



AI PODs

Deploying AI with confidence



Confidently deploy AI-ready infrastructure with pre-designed full stack architecture bundles for targeted AI use cases.



Leverage automation frameworks for rapid deployment and adoption of infrastructure.



Operate with best-in-class single-support model for your AI deployment architecture, include enterprise support for select Operations Support System (OSS) tools and libraries

AI Model

AI Tooling

Containers

Accelerated Compute

Networking

Converged Infrastructure

Management & Automation

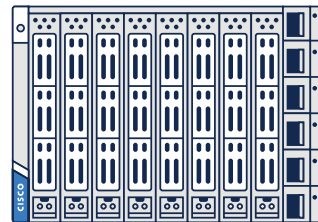
Adoption & Support Services



NVIDIA AI ENTERPRISE



OPENSIFT



PURESTORAGE

NetApp



MINT PARTNERS



CISCO Customer Experience

Cisco AI PODs for Inferencing

Typical use case	Edge Inferencing (7B-13B Parameter)	RAG Augmented Inferencing (13B-40B+ Parameter)	Large-Scale RAG Augmented Inferencing	Scale-Out Inferencing Cluster (Inferencing Multiple Models)
Hardware Specification (Required)	Small 1x X210C compute node <ul style="list-style-type: none"> 2x Intel 5th Gen 6548Y+ 512 GB System Memory 5x 1.6 TB NVMe drives 1x X440p PCIe <ul style="list-style-type: none"> 1x NVIDIA L40S X-Series FI9108 100G 	Medium 2x X210C compute nodes <ul style="list-style-type: none"> 4x Intel 5th Gen 6548Y+ 1 TB System Memory 10x 1.6 TB NVMe drives 2x X440p PCIe <ul style="list-style-type: none"> 4x NVIDIA L40S 2x Fabric Interconnect <ul style="list-style-type: none"> 6536 100G 	Large 2x X210C compute nodes <ul style="list-style-type: none"> 4x Intel 5th Gen 6548Y+ 1 TB System Memory 10x 1.6 TB NVMe drives 2x X440p PCIe <ul style="list-style-type: none"> 4x NVIDIA H100 NVL 2x Fabric Interconnect <ul style="list-style-type: none"> 6536 100G 	Scale-Out 4x X210C compute nodes <ul style="list-style-type: none"> 8x Intel 5th Gen 6548Y+ 4 TB System Memory 20x 1.6 TB NVMe drives 4x X440p PCIe <ul style="list-style-type: none"> 8x NVIDIA L40S 2x Fabric Interconnect <ul style="list-style-type: none"> 6536 100G
Software specification (Required)	Cisco Intersight <ul style="list-style-type: none"> Essentials Nvidia AI Enterprise <ul style="list-style-type: none"> Essentials 	Cisco Intersight <ul style="list-style-type: none"> Essentials Nvidia AI Enterprise <ul style="list-style-type: none"> Essentials 	Cisco Intersight <ul style="list-style-type: none"> Essentials Nvidia AI Enterprise <ul style="list-style-type: none"> Essentials 	Cisco Intersight <ul style="list-style-type: none"> Essentials Nvidia AI Enterprise <ul style="list-style-type: none"> Essentials
Default Components (Optional)	OpenShift <ul style="list-style-type: none"> OpenShift Container Platform Single-Node Controller 	OpenShift <ul style="list-style-type: none"> OpenShift Container Platform X210c Control Plane Cluster 	OpenShift <ul style="list-style-type: none"> OpenShift Container Platform X210c Control Plane Cluster 	OpenShift <ul style="list-style-type: none"> OpenShift Container Platform X210c Control Plane Cluster
Add-On	CI Storage  FlashStack ●● FlexPod	CI Storage  FlashStack ●● FlexPod	CI Storage  FlashStack ●● FlexPod	CI Storage  FlashStack ●● FlexPod

Bringing flexibility and scalability to your AI workloads

Versatile, accelerated, 4RU rack server for broad set of AI use cases



Cisco UCS® C845A M8 Rack Server

NVIDIA MGX platform with

2/4/6/8 NVIDIA H100 NVL/H200 NVL/L40S GPUs

2 AMD 5th Gen EPYC processors



H100*8+B3220+B3140H*4

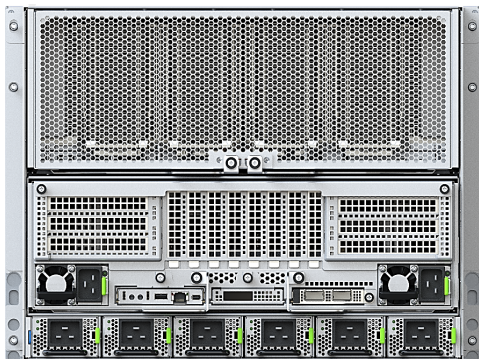


L40S*8+B3220+CX7*4

UCS C885A M8 Modular Sled Design



UCS C885A Dense GPU Server Specifications



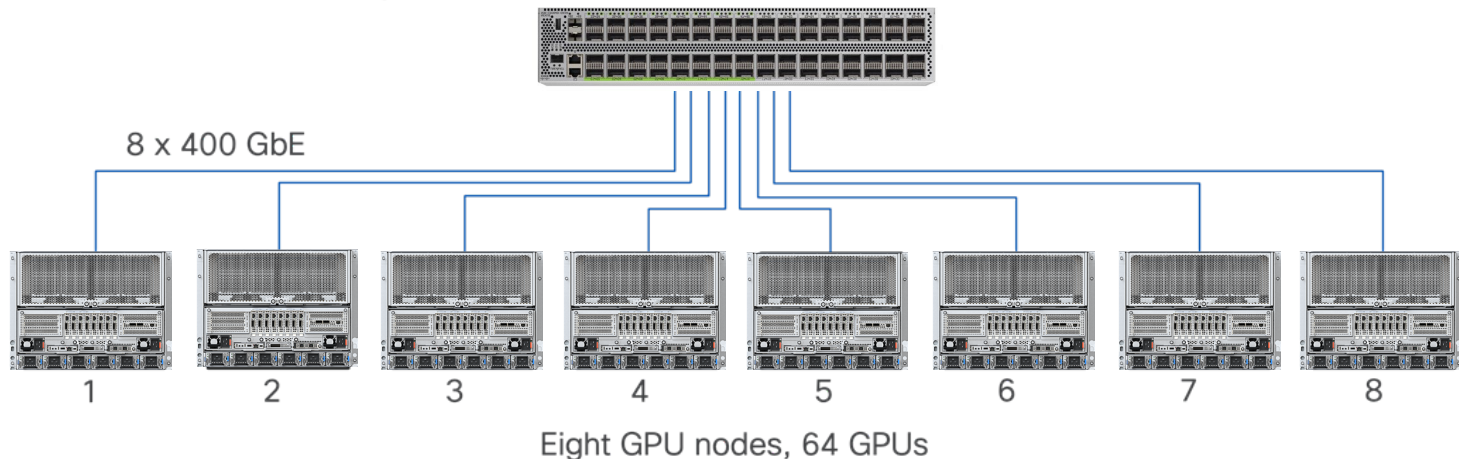
Product Specifications

Form Factor	<ul style="list-style-type: none">• HGX 8RU 19" Rack Server
Compute + Memory	<ul style="list-style-type: none">• 2x 4th Gen AMD EPYC 9554 (360W, 64 core, up to 3.75GHz) or• 2x 5th Gen AMD EPYC 9575F (400W, 64 core, up to 5GHz)• 24 DDR5 RDIMMs• Up to 6,000 MT/S
Storage	<ul style="list-style-type: none">• 1 PCIe3 x4 M.2 NVMe (Boot Device)• Up to 16 PCIe5 x4 2.5" U.2 NVMe SSD (Data Cache)
GPUs	<ul style="list-style-type: none">• Nvidia: 8 x H100 (700W) or 8 x H200 (700W) or 8 x B200A (700W)• AMD: 8 x MI300X (750W)
Network Cards	<ul style="list-style-type: none">• 8 PCIe x16 HHHL for East-West NIC ConnectX-7 or BF3 B3140H• 1 PCIe x16 FHHL for North-South NIC BF3 B3220• 1 OCP 3.0 X710-T2L for North-South or host management
Cooling	<ul style="list-style-type: none">• 12 Hot swappable (N+1) fans for system cooling• 4 fans for SSD cooling
Front IO	<ul style="list-style-type: none">• 2 USB 2.0, 1 ID Button, 1 Power Button
Rear IO	<ul style="list-style-type: none">• 1 USB 3.0 A, 1 USB 3.0 C, mDP, 1 ID Button, 1 Power Button, 1 USB 2.0 C (for debugging), 1 RJ45 (mgmt.)
Power Supply	<ul style="list-style-type: none">• Up to 6 54V 3kW and 2 12V 2.7kW MCRPS/CRPS, N+1 redundancy

Designing a Smaller Inter-GPU Backend Network

Single-switch network interconnecting 64 GPUs

Using 64-port 400 GbE Cisco Nexus 9364D-GX2A switch

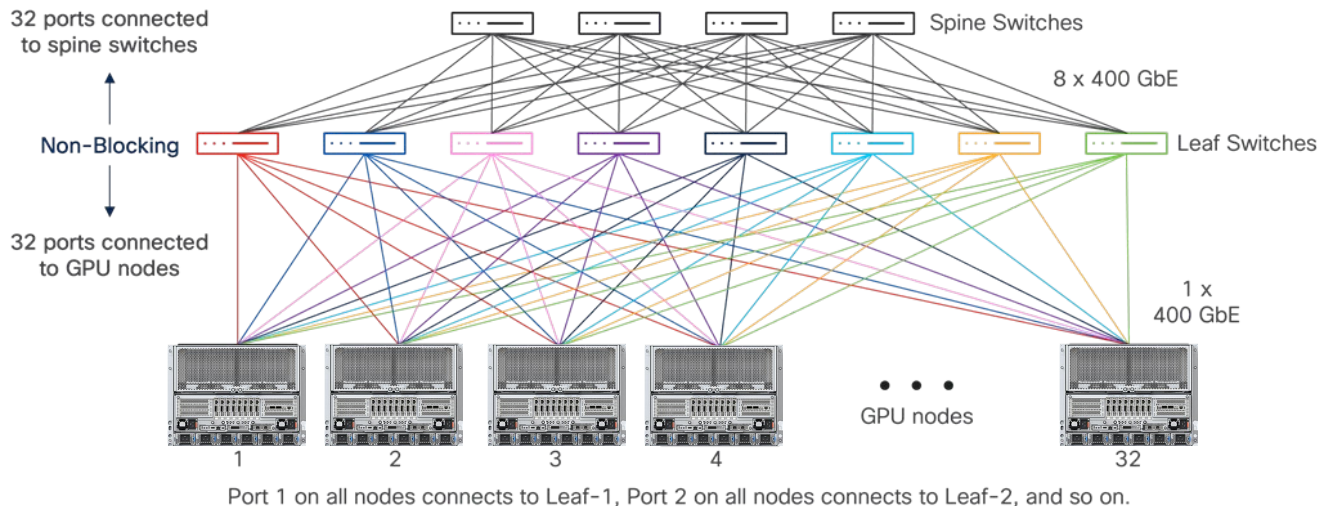


- Smaller GPU clusters can use a single-switch network. For example, up to 64 GPUs can be interconnected using the 2 RU, 64-port 400 GbE, Cisco Nexus 9364D-GX2A switch.

Designing a Larger Inter-GPU Backend Network

Rails-optimized network interconnecting 256 GPUs

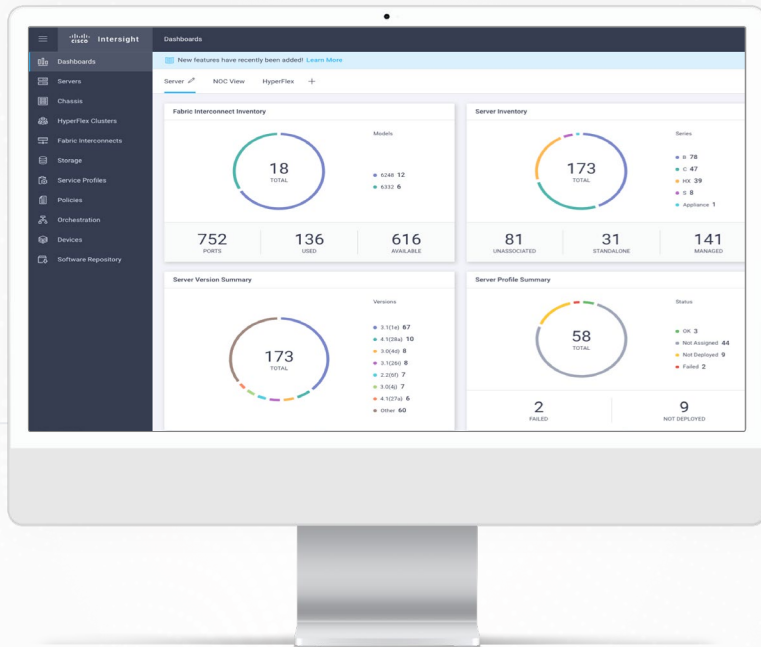
Using 64-port 400 GbE Cisco Nexus 9364D-GX2A switches



- For a larger GPU cluster, a spine-leaf network design is the best option because of its consistent and predictable performance and ability to scale.
- The edge switch ports that connect to the GPU ports should operate at the fastest supported speeds, such as 400 GbE. The core switch ports between the leaf switches and spine switches should match or exceed the speed at which the GPUs connect to the network.

Cisco Intersight

Data center
operations,
simplified



Feature Parity

- ✓ More of what you love about UCS Managed Mode, now available in Intersight Managed Mode

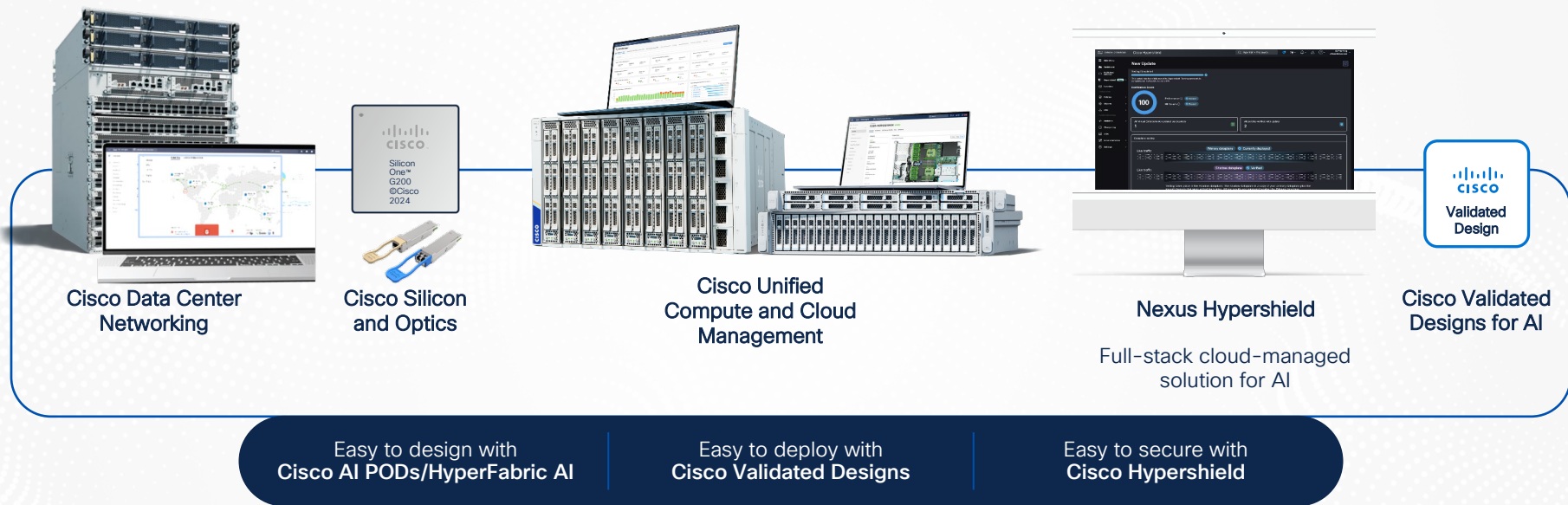
AIOps

- ✓ Enriched topology view for full visibility
- ✓ More data and telemetry than ever before
- ✓ Bringing monitoring to the virtual appliance

Sustainability

- ✓ Energy and power dashboards
- ✓ SFP metrics

Our Unified Approach to the Data Center



Only Cisco unifies networking, compute, security,
and observability to deliver AI-ready data centers.

