CISCO Engage | Tech Day

Accelerate Your AI-Ready Data Center

With Cisco CX Services

Vamsidhar A
Business Development Manager - AI Factory
vamsa@cisco.com

Dec 17th 2025

What We'll Cover

Agenda

The AI Horizon

- *Are we Ready for AI*
- *Cisco AI Ready Data Center*
- *Cisco's AI Services portfolio*
- *AI Ready Enterprises*
- *Future of AI*

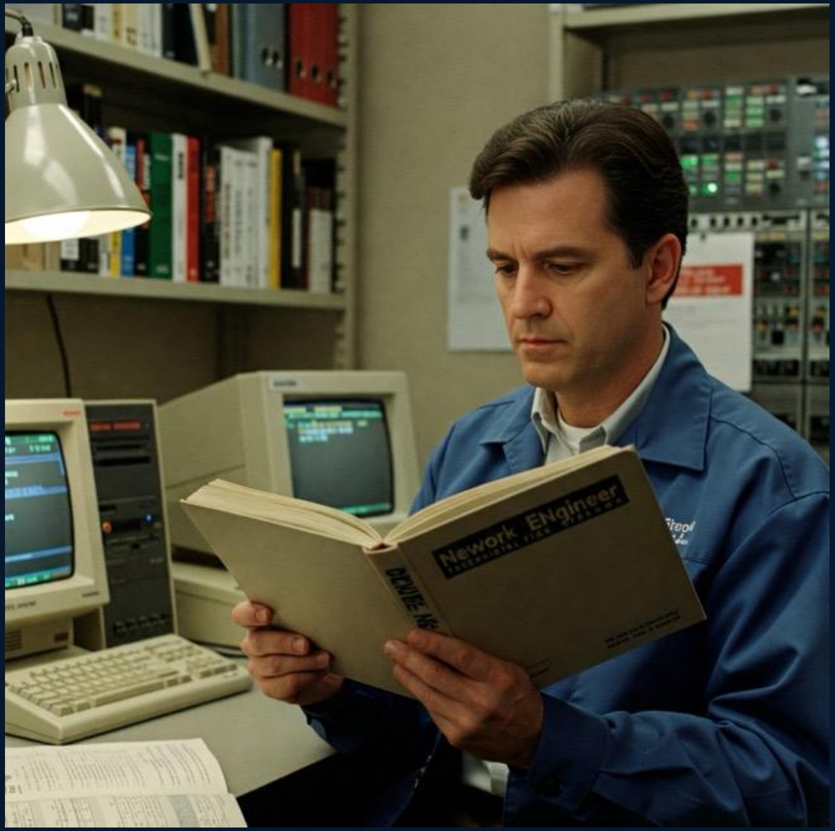
Are we Ready for Future !

World has shifted 3 Years ago!

How Powerful is your Brain !

* Ref 30th Nov 2022

In the 90s, before the internet, we had to read books



Fundamental Shift on How we work

With internet and search engines

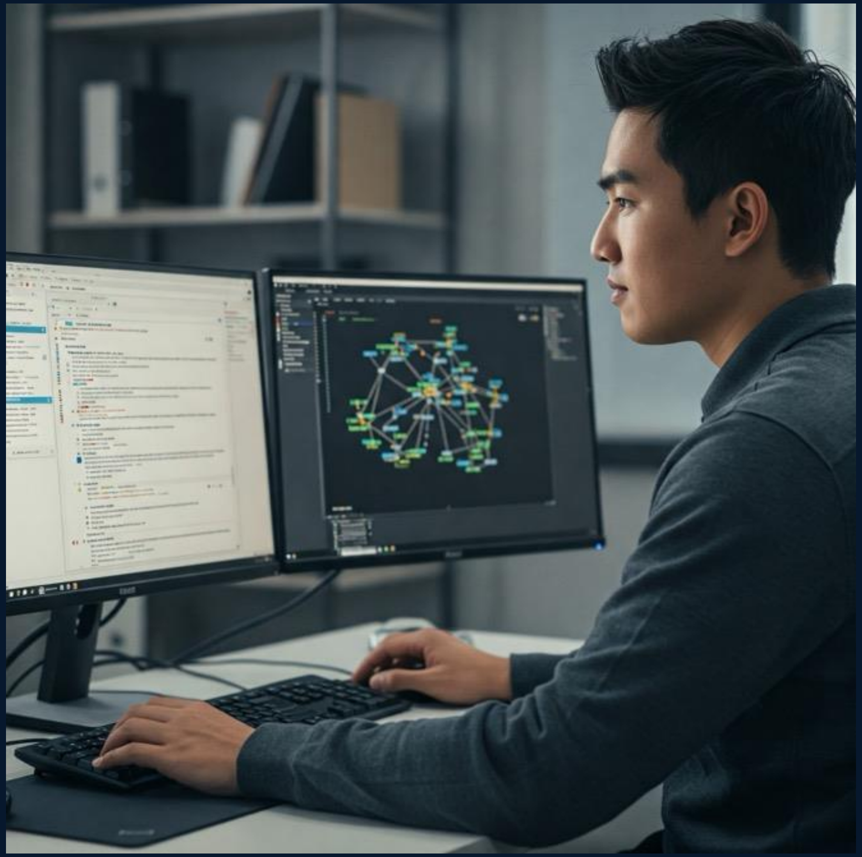


image generated by AI

What we thought AI would look like

AI: The Next Fundamental Transformation in How We Live and Work

Will Robots Pamper us !

Are we Ready for AI

AI is Everywhere

$15.7T

Potential contribution to global economy by 2030

$300B

Global spending on AI by 2026

75%

Of large enterprises will rely on AI-infused processes by 2026

Healthcare and Life Sciences

Diagnosis
Drug discovery
Personalized medicine

Financial Services

Fraud detection
Risk assessment
Trading

Retail

Personalization
Inventory optimization
Virtual agents

Manufacturing

Predictive maintenance
Quality control
Demand forecasting

Agriculture

Yield optimization
Automated irrigation
Pest prediction & prevention

Transportation

Route optimization
Autonomous vehicles
Predictive maintenance

Energy

Distribution optimization
Fault prediction
Demand forecasting

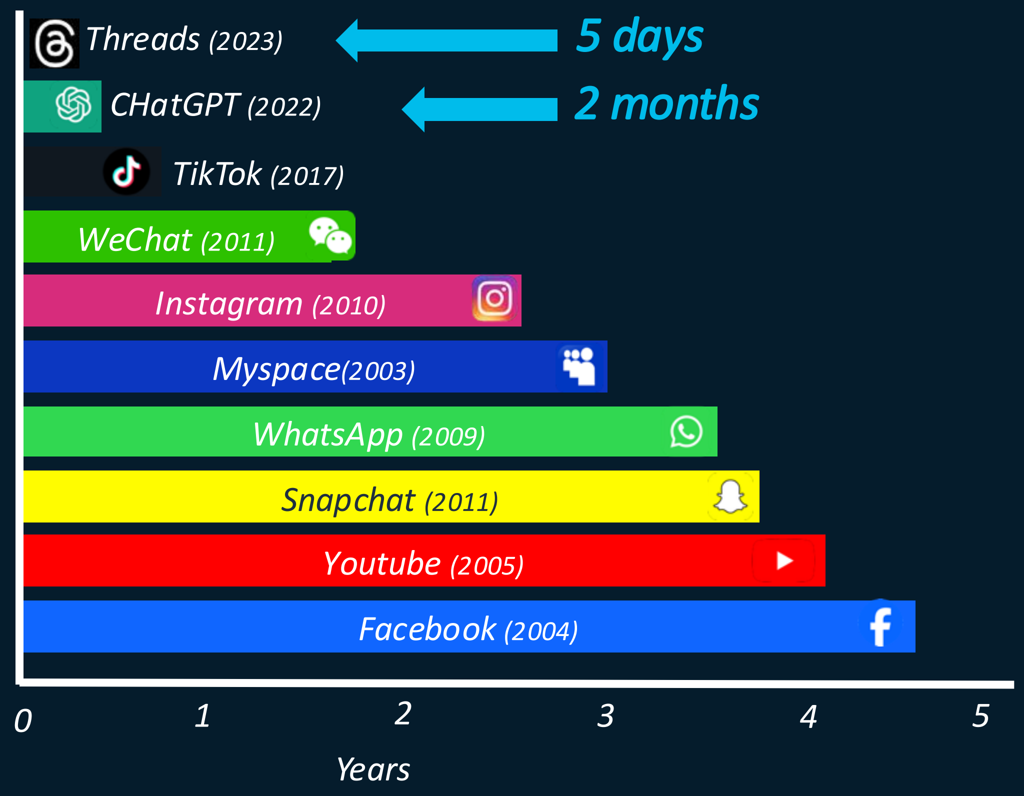
Public Sector

Smart cities
Security
Services improvement

Sources: PWC, IDC

Pace of Adoption

Time to Reach 100M users

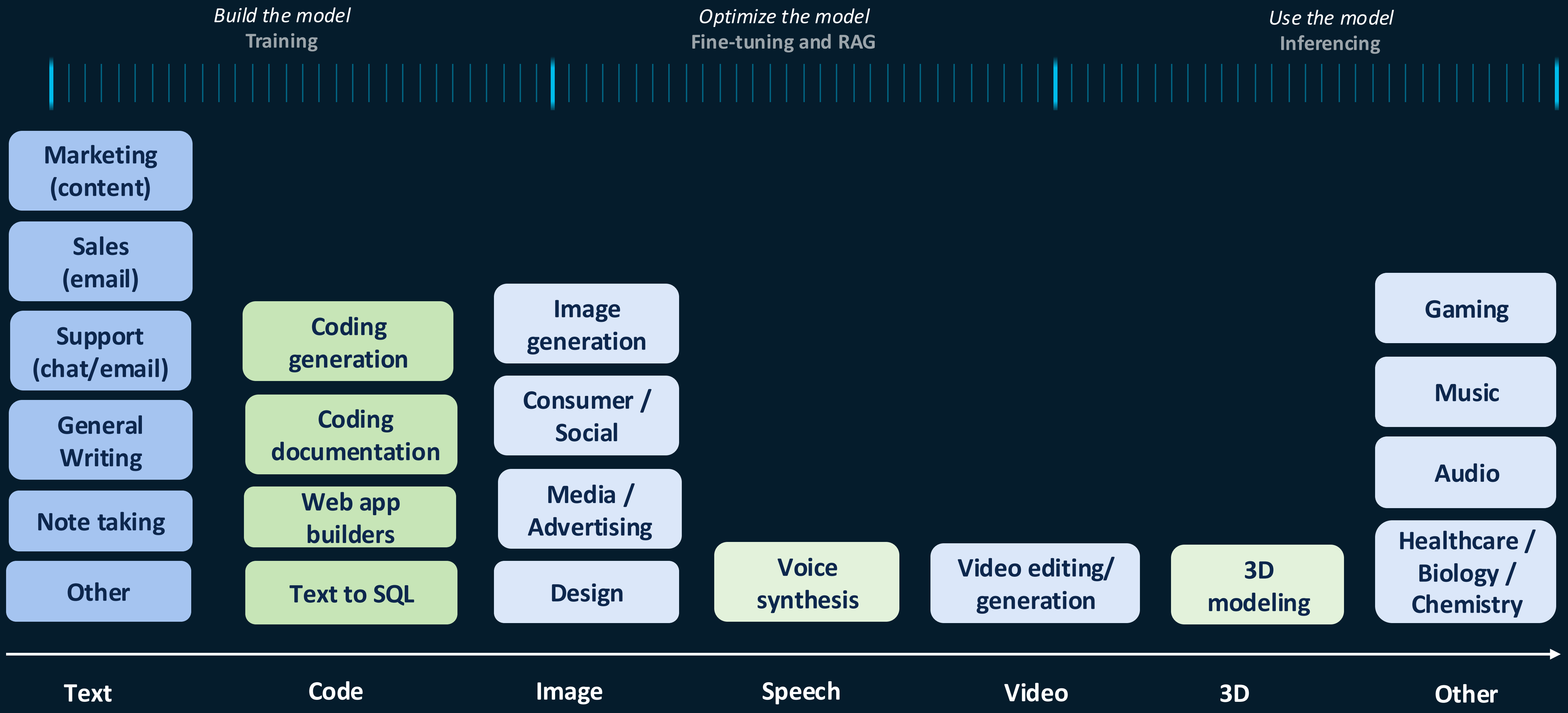


visualcapitalist.com

86% of companies not fully ready to integrate AI into their businesses.

Cisco AI Readiness Index

Every organization’s needs are different



Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Text
- Marketing (content)
- Sales (email)
- Support (chat/email)
- General Writing
- Note taking
- Other

Code
- Coding generation
- Coding documentation
- Web app builders
- Text to SQL

Image
- Image generation
- Consumer / Social
- Media / Advertising
- Design

Speech
- Voice synthesis

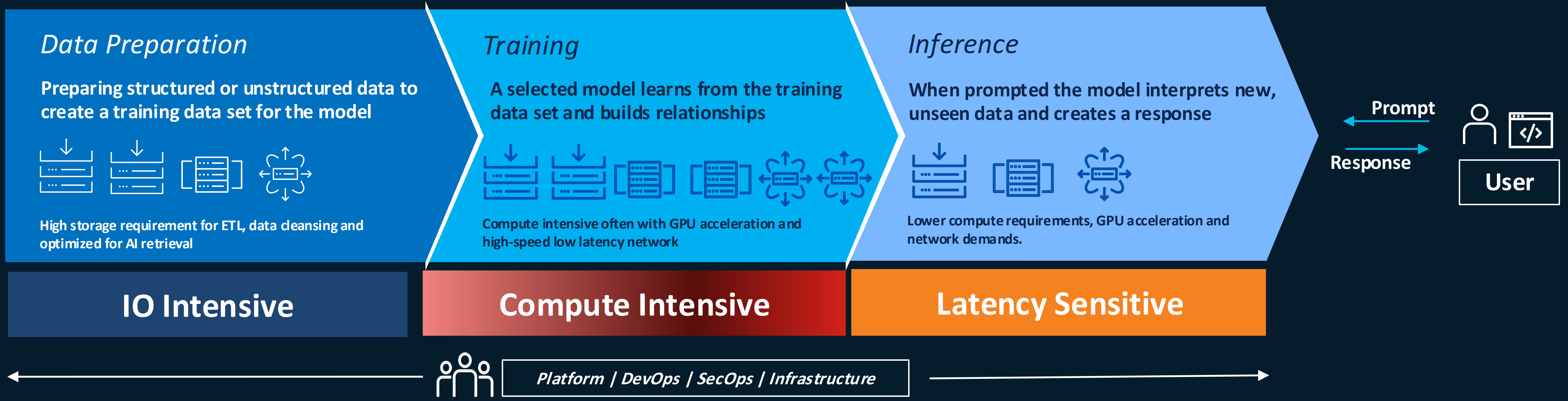
Video
- Video editing/ generation

3D
- 3D modeling

Other
- Gaming
- Music
- Audio
- Healthcare / Biology / Chemistry

* Multimodal LLMs

AI and Infrastructure



Data Preparation

Preparing structured or unstructured data to create a training data set for the model

High storage requirement for ETL, data cleansing and optimized for AI retrieval

Training

A selected model learns from the training data set and builds relationships

Compute intensive often with GPU acceleration and high-speed low latency network

Inference

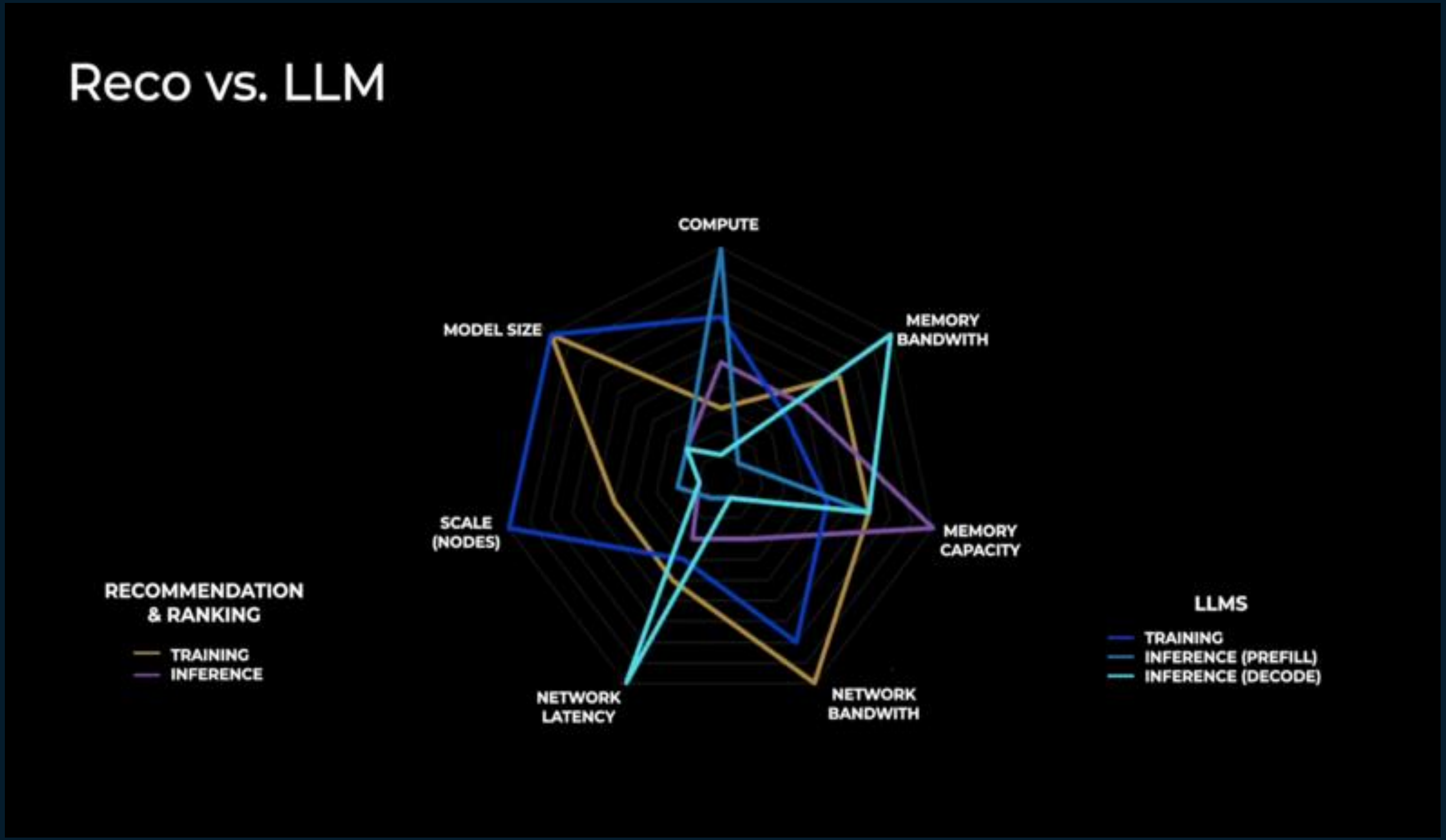
When prompted the model interprets new, unseen data and creates a response

Lower compute requirements, GPU acceleration and network demands.

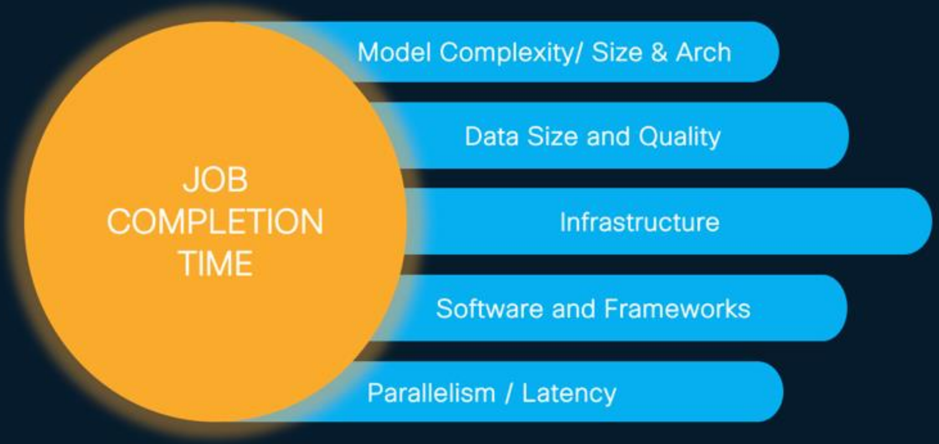
Model performance tightly coupled to infrastructure

AI/ML Requirements

Training vs Inference



Reco vs. LLM



JOB COMPLETION TIME

- Model Complexity/ Size & Arch
- Data Size and Quality
- Infrastructure
- Software and Frameworks
- Parallelism / Latency

AI Ready Data Center

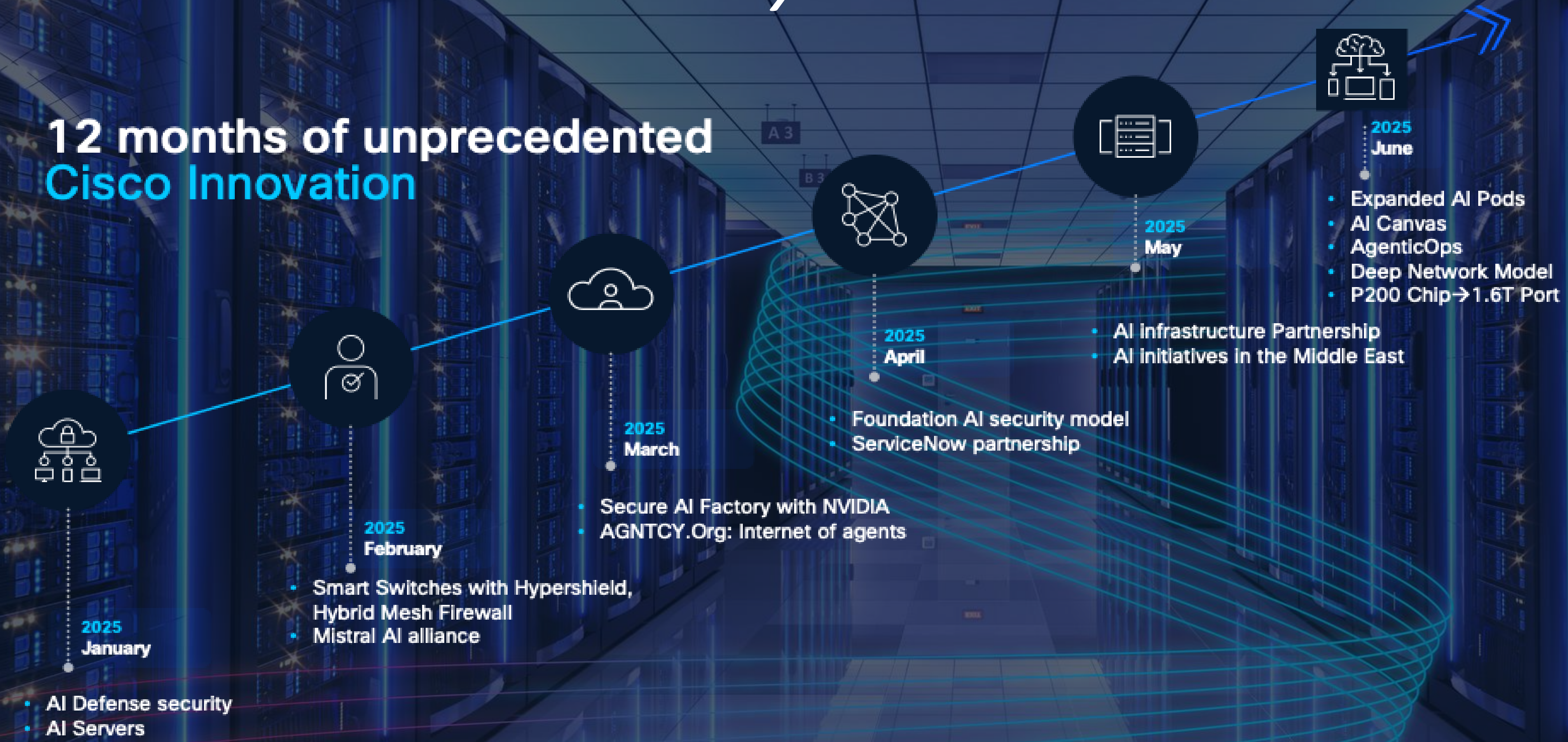
The world is experiencing the *largest expansion of data centers* in history

Cisco is foundational to the world’s data center buildouts

Enterprises | Neoclouds | Service Providers | Hyperscalers

The AI-ready data center

12 months of unprecedented Cisco Innovation



2025 January

- AI Defense security
- AI Servers

2025 February

- Smart Switches with Hypershield, Hybrid Mesh Firewall
- Mistral AI alliance

2025 March

- Secure AI Factory with NVIDIA
- AGNTCY.Org: Internet of agents

2025 April

- Foundation AI security model
- ServiceNow partnership

2025 May

- AI infrastructure Partnership
- AI initiatives in the Middle East

2025 June

- Expanded AI Pods
- AI Canvas
- AgenticOps
- Deep Network Model
- P200 Chip→1.6T Port

AI is network bound

Private data center “re-acceleration”

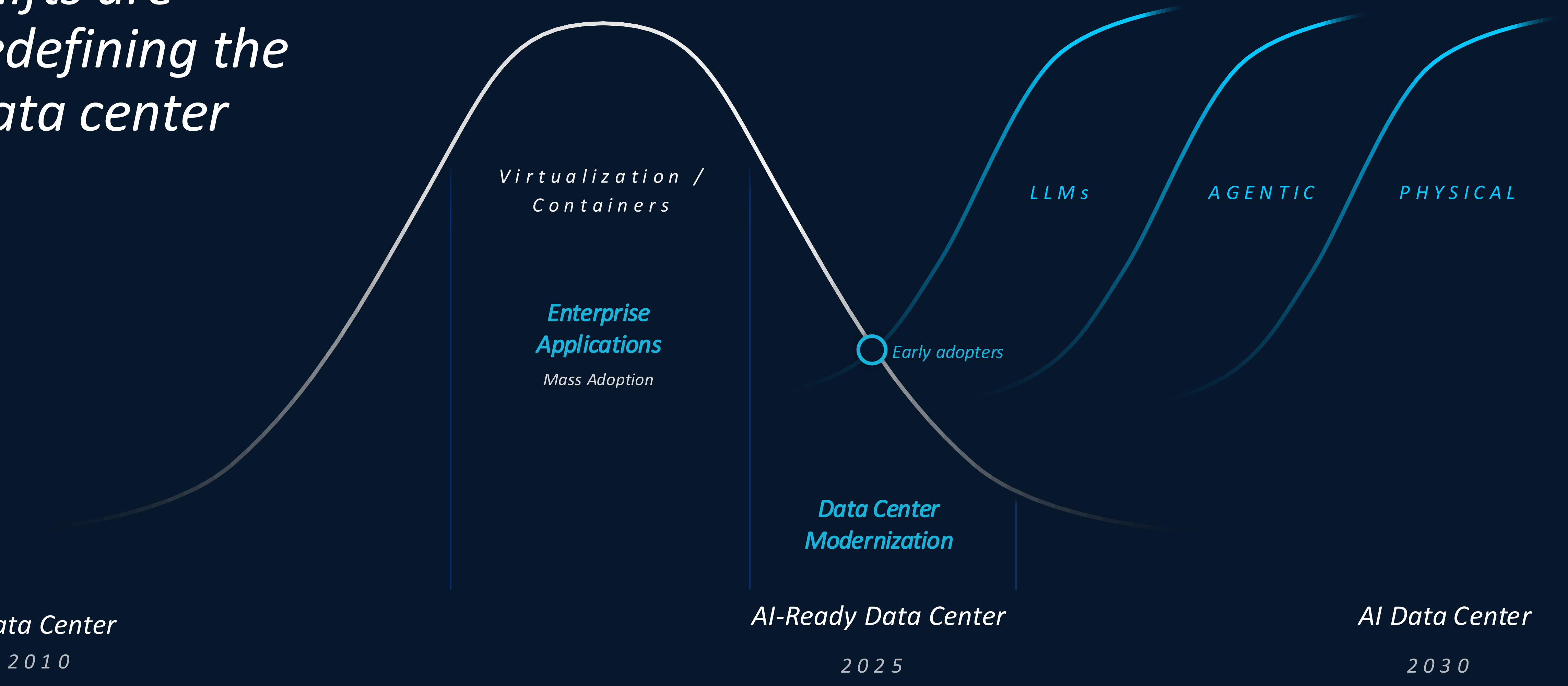
Common foundation for AI safety

Hyper-distributed security

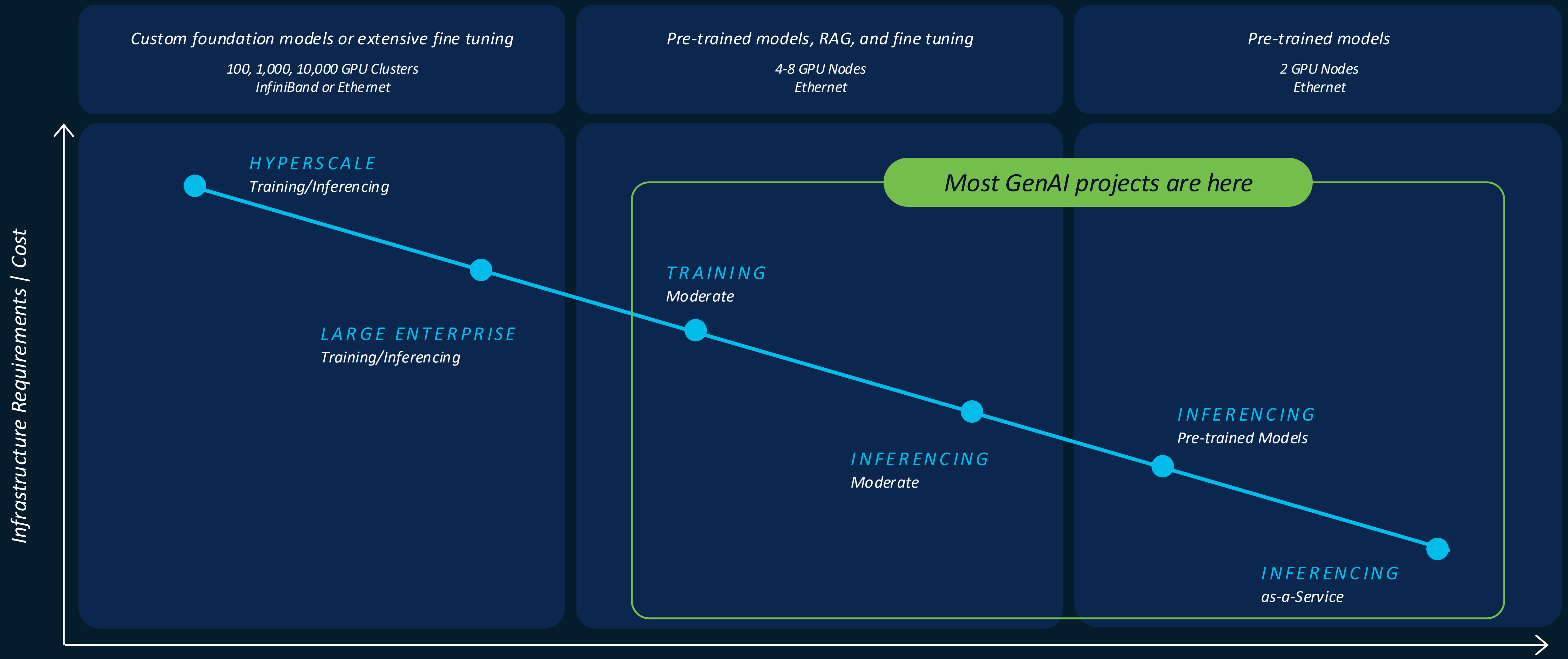
What we know

Architectural shifts are redefining the data center

8B to 80B



AI Use Case Spectrum



Custom foundation models or extensive fine tuning

100, 1,000, 10,000 GPU Clusters
InfiniBand or Ethernet

Pre-trained models, RAG, and fine tuning

4-8 GPU Nodes
Ethernet

Pre-trained models

2 GPU Nodes
Ethernet

The pace of AI innovation is staggering

Complexity has exceeded **Human scale**

2022
ChatGPT LLM 3.5

2023
Multimodal
(GPT 4)

2024
Long Context AI
(Claude 3)

2024
Memory-based AI

2023
Retrieval Augmented Generation

2025
Agentic AI

2025
DeepSeek R1

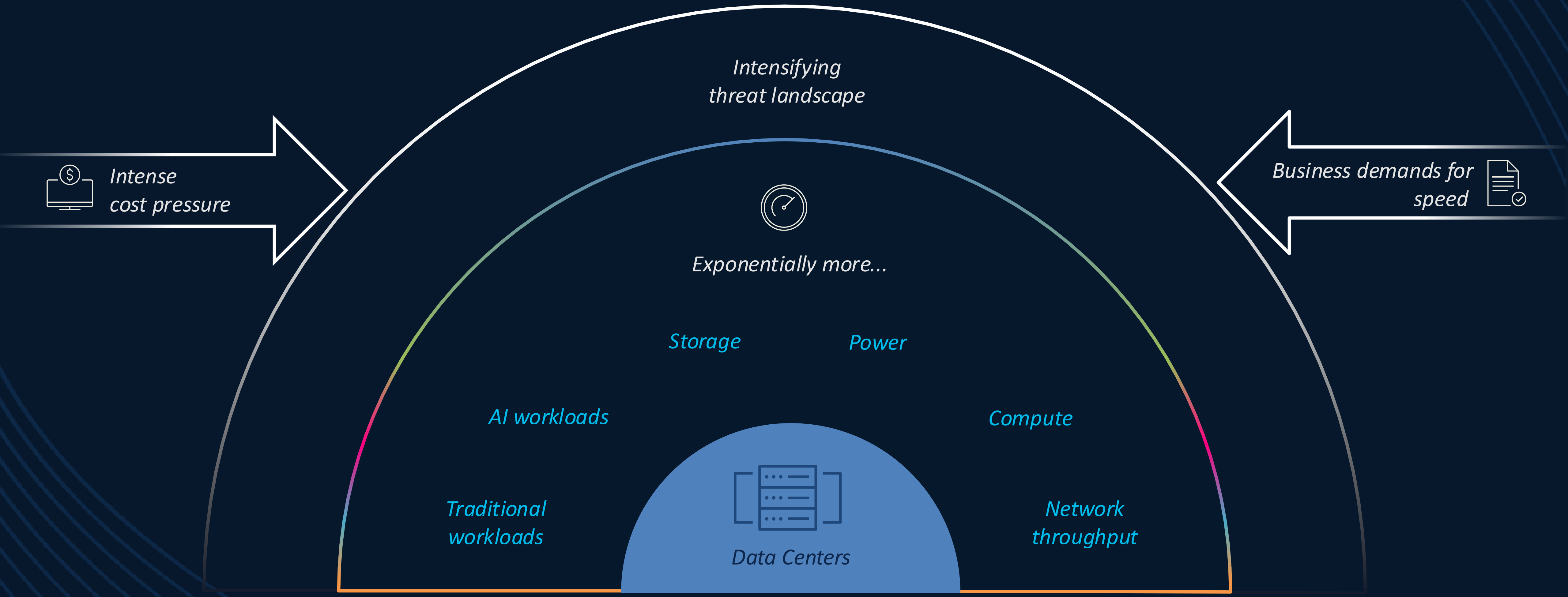
2025
OpenAI Deep Research

What's Next?
Physical AI /Synthetic AI

2026-2030 AI Runs out of Public Digitized data to train

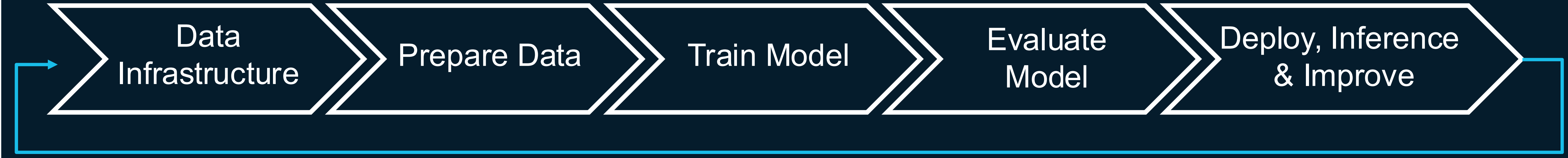
Reimagine the data center for the AI era

While managing cost pressures, intensifying threats, and business demands



AI Infrastructure Stack

Compute, Network, Security, Observability



Data Infrastructure

Prepare Data

Train Model

Evaluate Model

Deploy, Inference & Improve

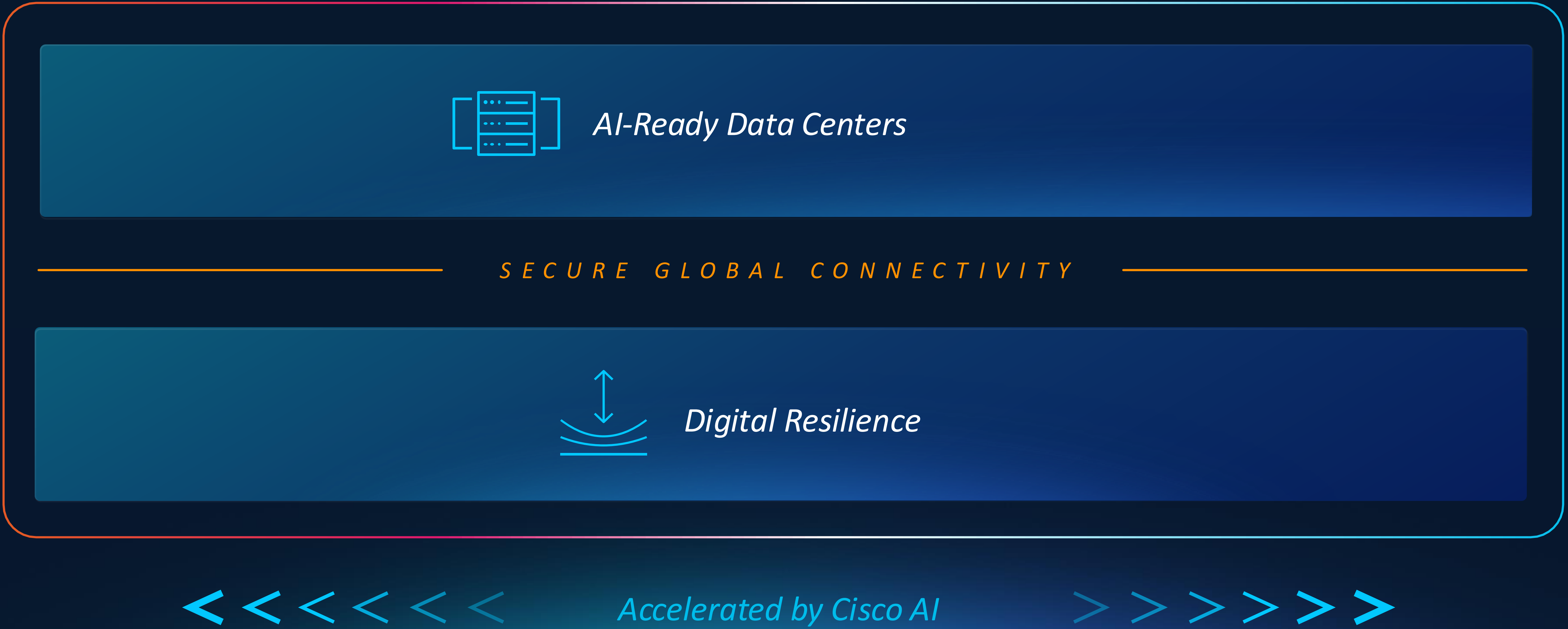
Network, Compute: Infrastructure Optimized for GenAI Compute

Security: Data Integrity, Attack Prevention, Policy Management

Observability: Model and Infra Monitoring

A Multi-architecture play

Cisco's AI Ready DC



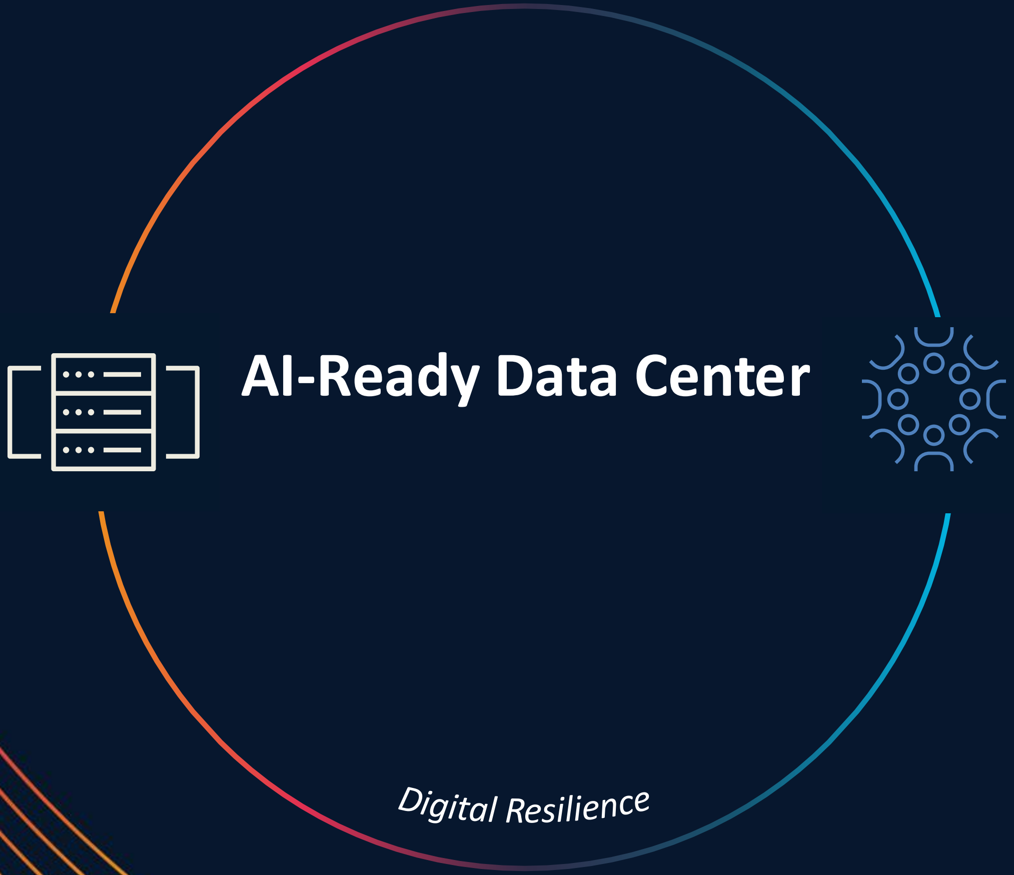
AI-Ready Data Centers

SECURE GLOBAL CONNECTIVITY

Digital Resilience

Accelerated by Cisco AI

Cisco's AI Ready DC Portfolio



AI-Ready Data Center

Digital Resilience

Distributed AI infrastructure

- UCS Accelerated Compute
- Nexus 9K Series
- Cisco Optics
- Cisco Silicon One
- Cisco 8K series (hyperscalers)
- Storage (partners)

Full stack systems*

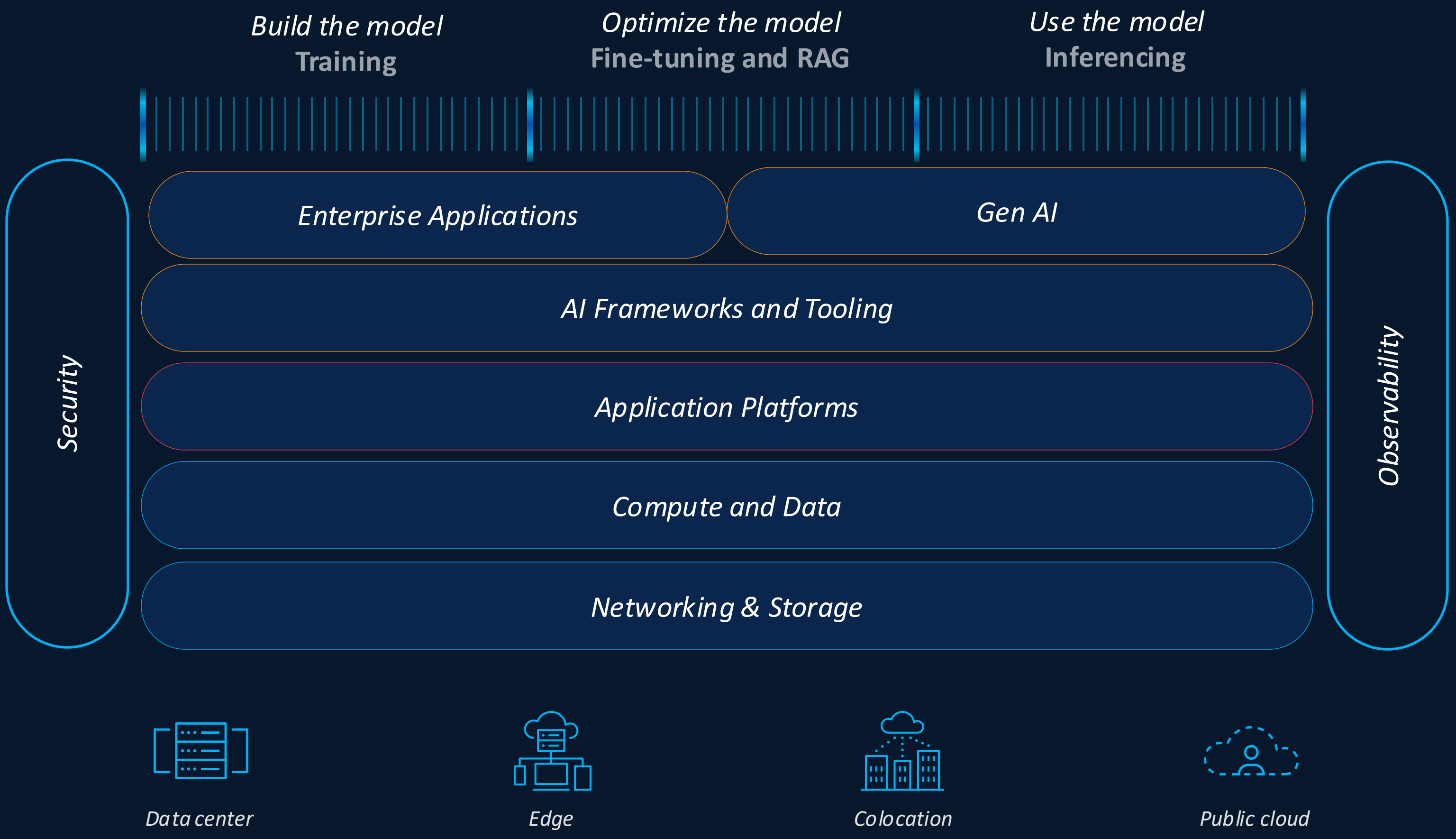
- Nexus HyperFabric AI
- AI PODs
- Nutanix GPT
- NVIDIA & AMD GPU
- Virtualization (partners)
- AI SW Tools (partners)

AI infrastructure operations

- Intersight
- Nexus Dashboard
- Nexus HyperFabric
- NVIDIA AI Enterprise & NIMs
- Splunk
- AppDynamics
- ThousandEyes
- Accedian(PCA)

Cloud protection

- Cloud Protection Suite: Hypershield, Secure Workload, Multicloud Defense
- AI Defense & Perimeter Firewall



Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Security

Enterprise Applications

Gen AI

AI Frameworks and Tooling

Application Platforms

Compute and Data

Networking & Storage

Observability

Data center

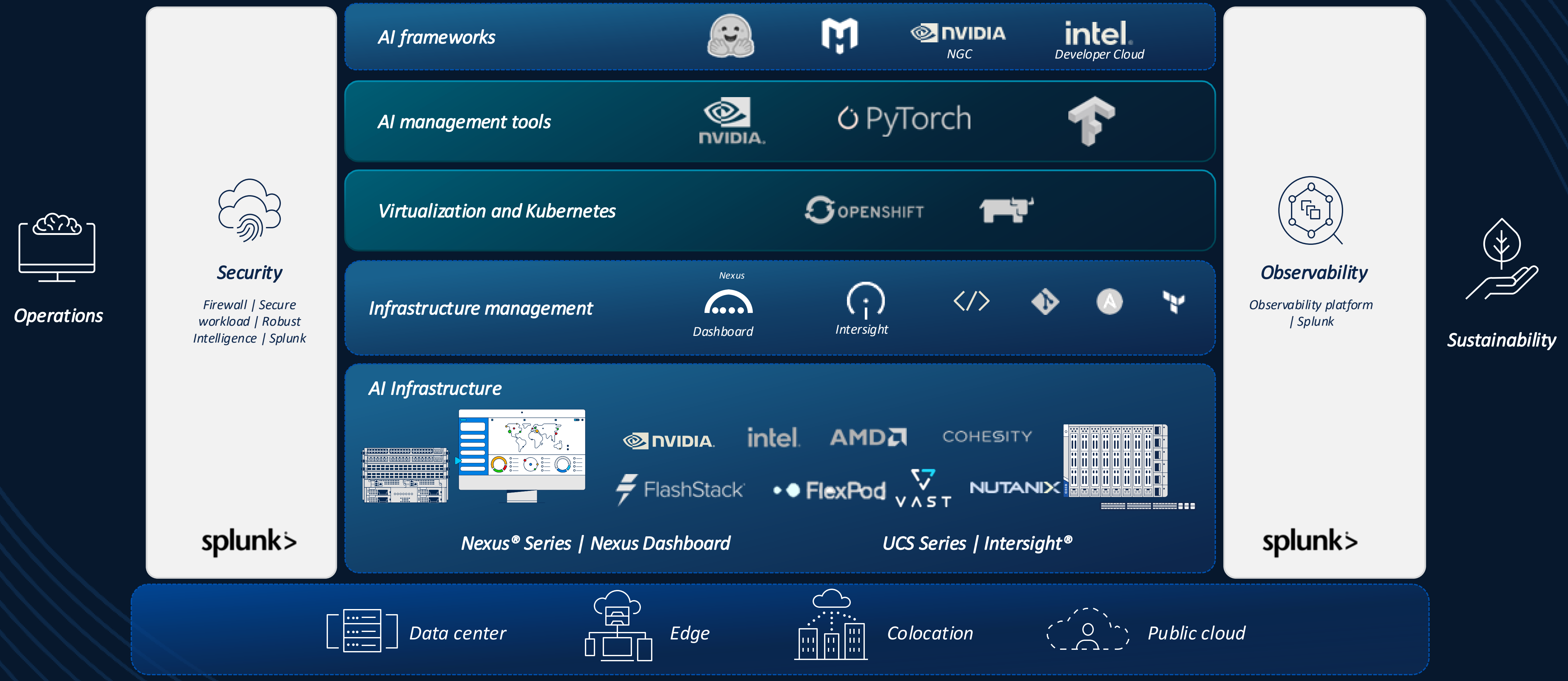
Edge

Colocation

Public cloud

Cisco AI-Ready Data Center

Ecosystem Stack



Operations

Security

Firewall | Secure workload | Robust Intelligence | Splunk

splunk>

AI frameworks

NVIDIA NGC

intel. Developer Cloud

AI management tools

NVIDIA.

PyTorch

Virtualization and Kubernetes

OPENSHIFT

Infrastructure management

Nexus Dashboard

Intersight

AI Infrastructure

NVIDIA. intel. AMD COHESITY

FlashStack FlexPod VAST NUTANIX

Nexus® Series | Nexus Dashboard

UCS Series | Intersight®

Observability

Observability platform | Splunk

splunk>

Sustainability

Data center

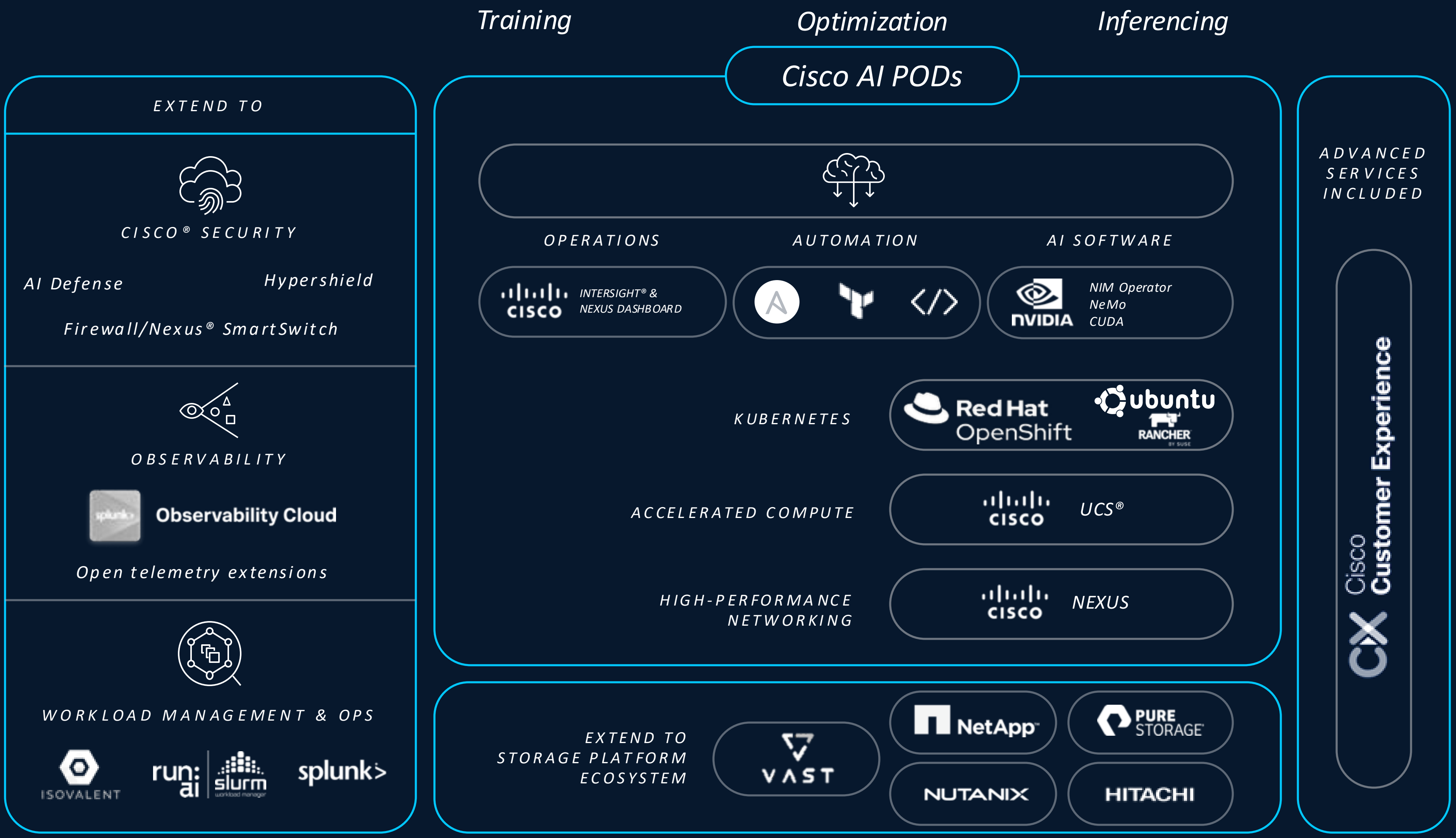
Edge

Colocation

Public cloud

Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently



Training

Optimization

Inferencing

EXTEND TO

CISCO® SECURITY

AI Defense

Hypershield

Firewall/Nexus® SmartSwitch

OBSERVABILITY

Observability Cloud

Open telemetry extensions

WORKLOAD MANAGEMENT & OPS

Cisco AI PODs

OPERATIONS

INTERSIGHT® & NEXUS DASHBOARD

AUTOMATION

AI SOFTWARE

NIM Operator
NeMo
CUDA

KUBERNETES

ACCELERATED COMPUTE

UCS®

HIGH-PERFORMANCE NETWORKING

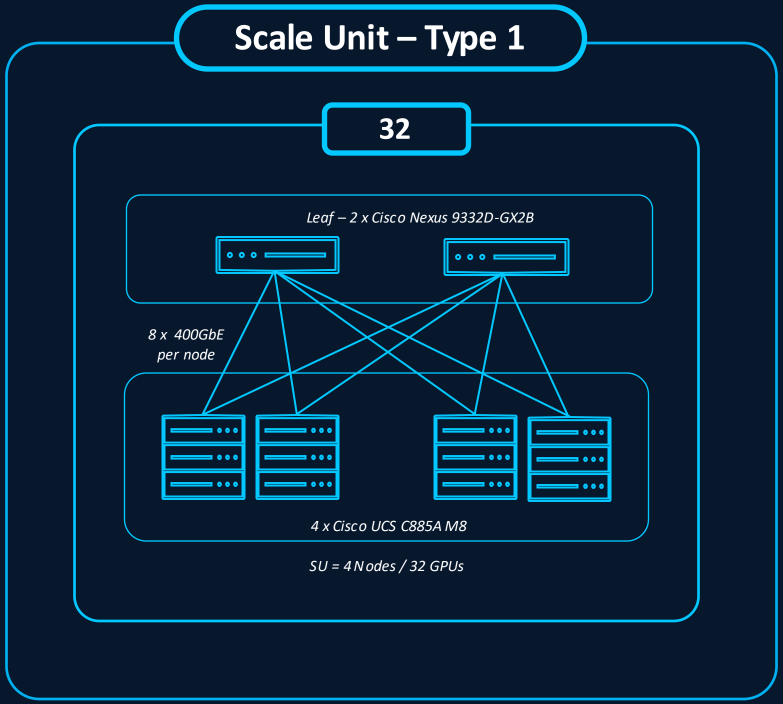
NEXUS

EXTEND TO STORAGE PLATFORM ECOSYSTEM

ADVANCED SERVICES INCLUDED

Cisco AI POD

Scaling Units



Scale Unit – Type 1

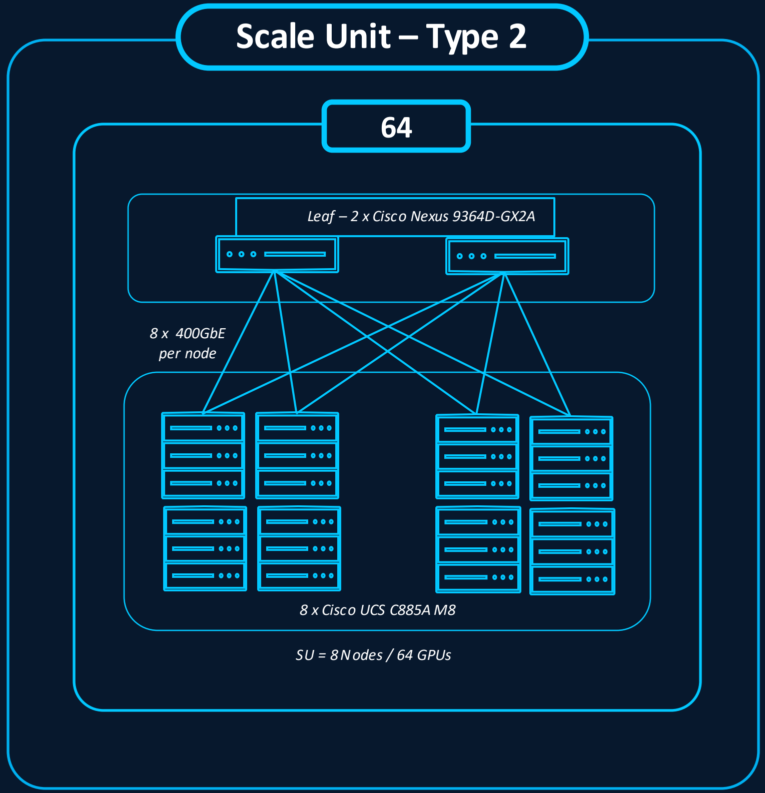
32

Leaf – 2 x Cisco Nexus 9332D-GX2B

8 x 400GbE per node

4 x Cisco UCS C885A M8

SU = 4 Nodes / 32 GPUs



Scale Unit – Type 2

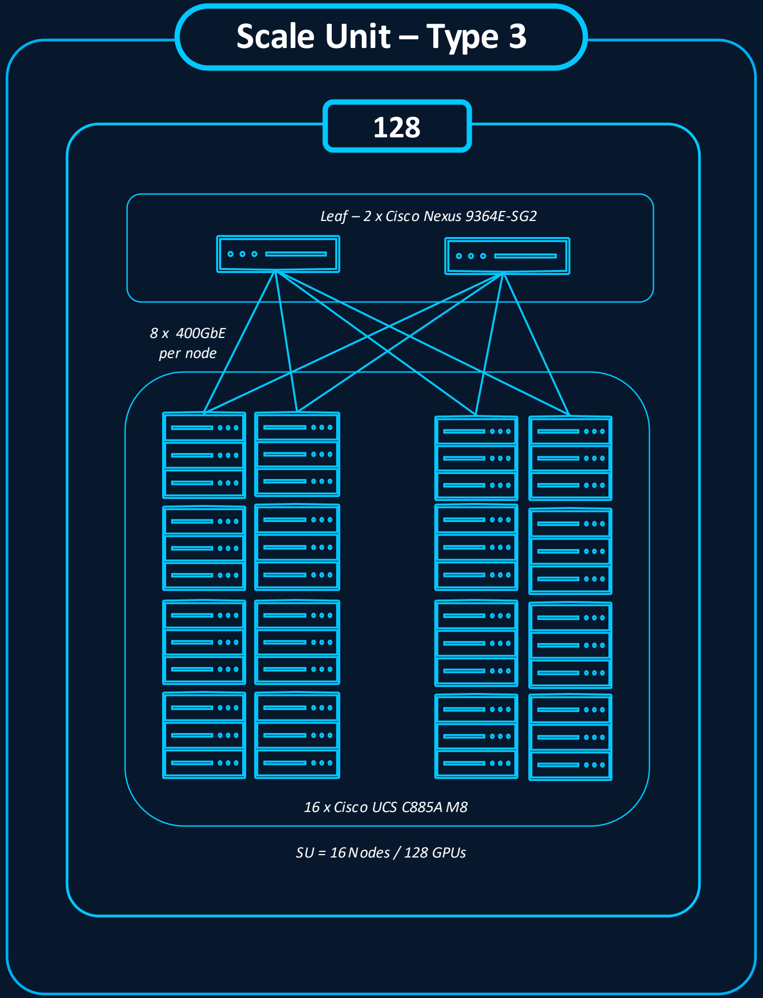
64

Leaf – 2 x Cisco Nexus 9364D-GX2A

8 x 400GbE per node

8 x Cisco UCS C885A M8

SU = 8 Nodes / 64 GPUs



Scale Unit – Type 3

128

Leaf – 2 x Cisco Nexus 9364E-SG2

8 x 400GbE per node

16 x Cisco UCS C885A M8

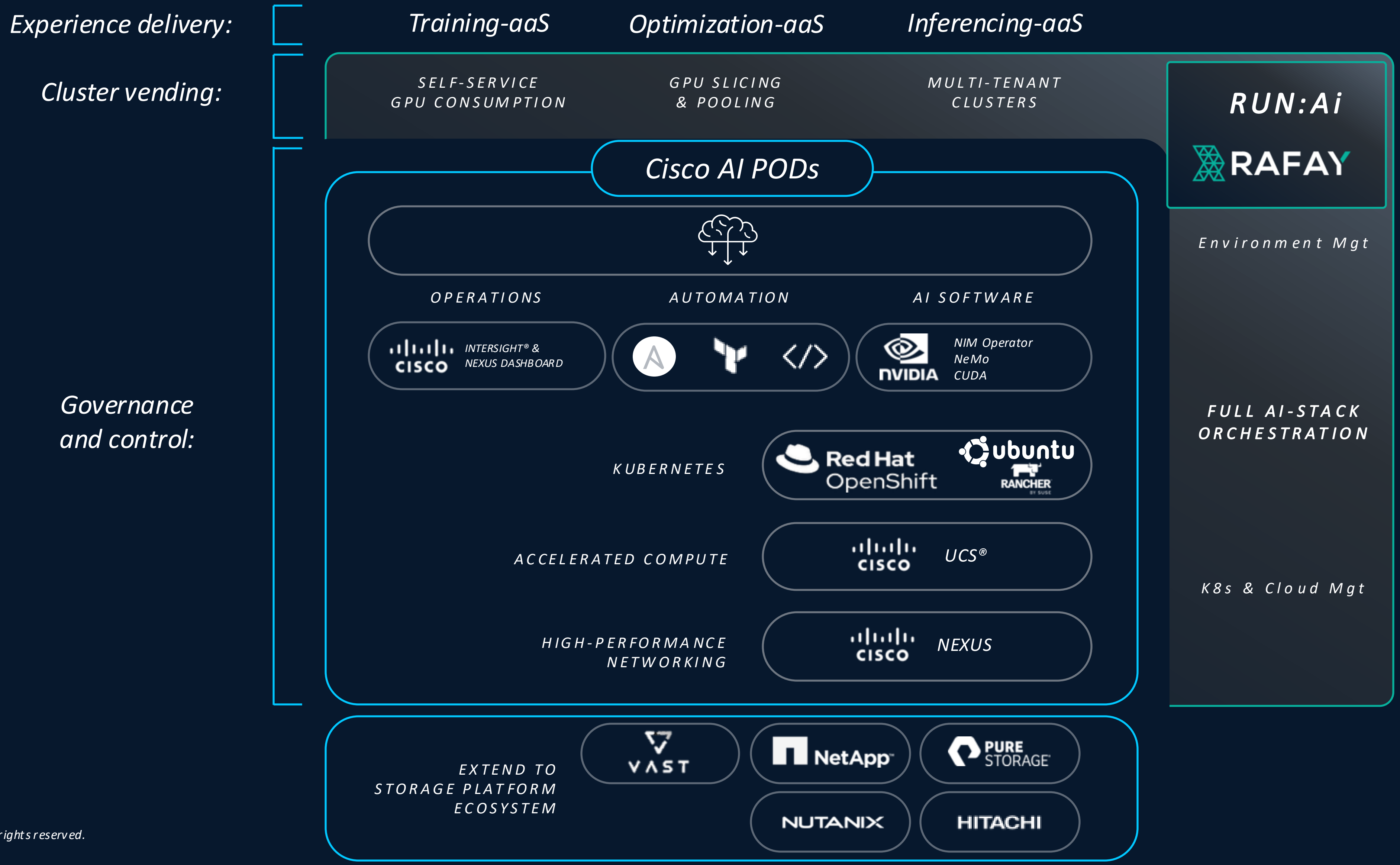
SU = 16 Nodes / 128 GPUs

Note:

- Scale Unit Type: A pair of leaf switches + UCS nodes(leaf uplinks excluded)
- Non-blocking design for max GPU performance

Transform Cisco AI PODs into a GPU cloud

Deliver sovereign and enterprise AI clouds



Experience delivery:

Training-aaS

Optimization-aaS

Inferencing-aaS

Cluster vending:

SELF-SERVICE GPU CONSUMPTION

GPU SLICING & POOLING

MULTI-TENANT CLUSTERS

RUN:Ai

RAFAY

Environment Mgt

FULL AI-STACK ORCHESTRATION

K8s & Cloud Mgt

Governance and control:

Cisco AI PODs

OPERATIONS

AUTOMATION

AI SOFTWARE

cisco INTERSIGHT® & NEXUS DASHBOARD

NVIDIA NIM Operator NeMo CUDA

KUBERNETES

Red Hat OpenShift

ubuntu

RANCHER

ACCELERATED COMPUTE

cisco UCS®

HIGH-PERFORMANCE NETWORKING

cisco NEXUS

EXTEND TO STORAGE PLATFORM ECOSYSTEM

VAST

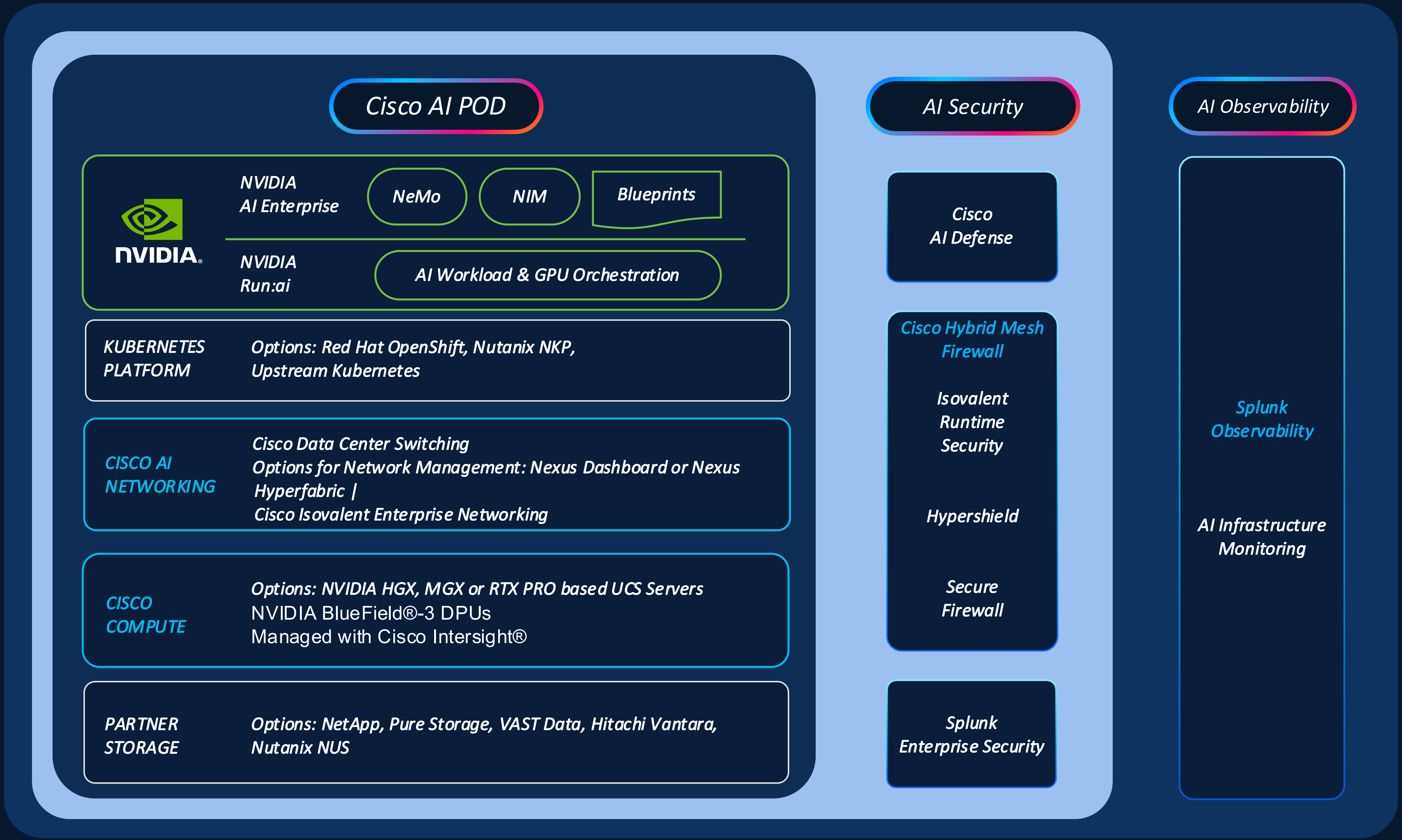
NetApp

PURE STORAGE

NUTANIX

HITACHI

Product view: Cisco Secure AI Factory with NVIDIA



Cisco AI POD

NVIDIA

NVIDIA AI Enterprise: NeMo, NIM, Blueprints

NVIDIA Run:ai: AI Workload & GPU Orchestration

KUBERNETES PLATFORM: Options: Red Hat OpenShift, Nutanix NKP, Upstream Kubernetes

CISCO AI NETWORKING: Cisco Data Center Switching; Options for Network Management: Nexus Dashboard or Nexus Hyperfabric | Cisco Isovalent Enterprise Networking

CISCO COMPUTE: Options: NVIDIA HGX, MGX or RTX PRO based UCS Servers; NVIDIA BlueField®-3 DPUs; Managed with Cisco Intersight®

PARTNER STORAGE: Options: NetApp, Pure Storage, VAST Data, Hitachi Vantara, Nutanix NUS

AI Security

Cisco AI Defense

Cisco Hybrid Mesh Firewall

Isovalent Runtime Security

Hypershield

Secure Firewall

Splunk Enterprise Security

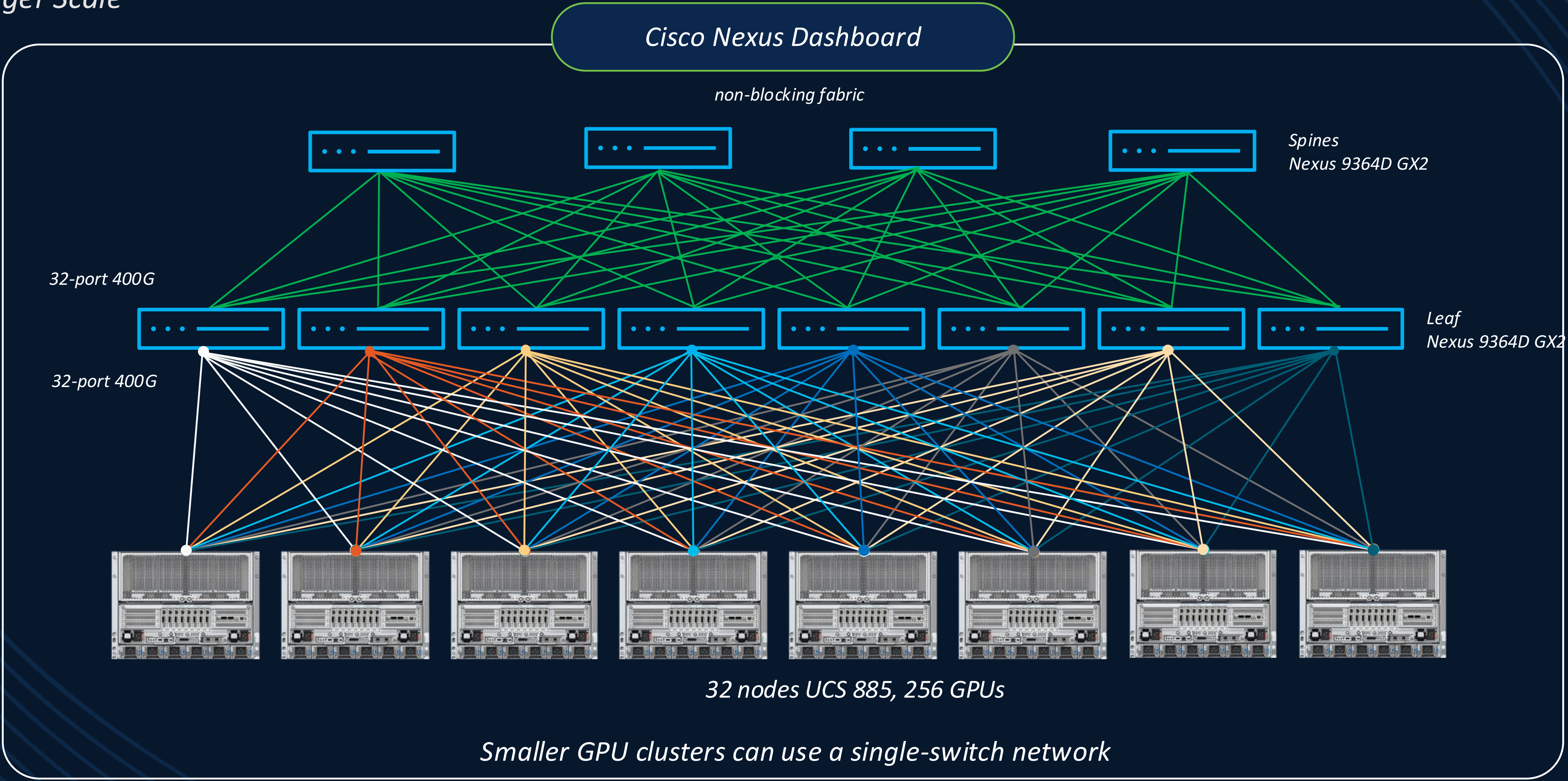
AI Observability

Splunk Observability

AI Infrastructure Monitoring

AIML Network

Larger Scale

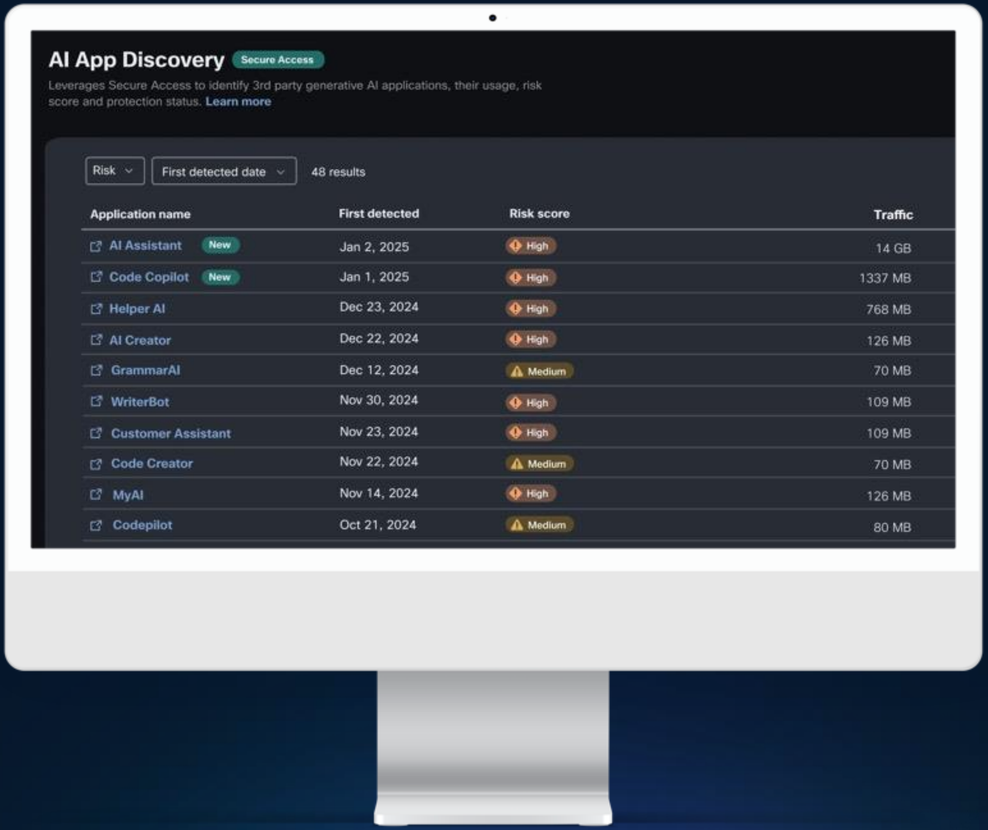


32 nodes UCS 885, 256 GPUs

Smaller GPU clusters can use a single-switch network

Security at every layer of the stack

Cisco AI Defense



AI App Discovery

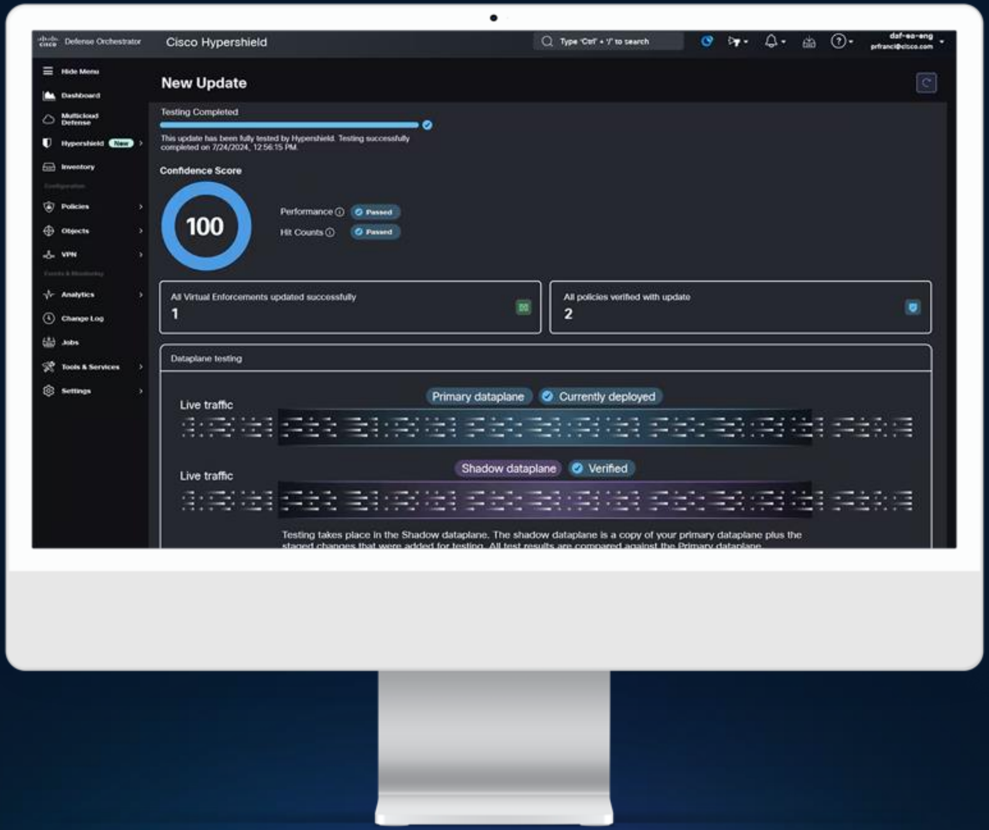
Leverages Secure Access to identify 3rd party generative AI applications, their usage, risk score and protection status. Learn more

48 results

Application name	First detected	Risk score	Traffic
AI Assistant New	Jan 2, 2025	High	14 GB
Code Copilot New	Jan 1, 2025	High	1337 MB
Helper AI	Dec 23, 2024	High	768 MB
AI Creator	Dec 22, 2024	High	126 MB
GrammarAI	Dec 12, 2024	Medium	70 MB
WriterBot	Nov 30, 2024	High	109 MB
Customer Assistant	Nov 23, 2024	High	109 MB
Code Creator	Nov 22, 2024	Medium	70 MB
MyAI	Nov 14, 2024	High	126 MB
Codepilot	Oct 21, 2024	Medium	80 MB

Secure what AI is doing

Cisco Hypershield



New Update

Testing Completed

This update has been fully tested by Hypershield. Testing successfully completed on 7/24/2024, 12:56:15 PM.

Confidence Score

100

Performance: Passed

Hit Counts: Passed

All Virtual Enforcements updated successfully
1

All policies verified with update
2

Dataplane testing

Testing takes place in the Shadow dataplane. The shadow dataplane is a copy of your primary dataplane plus the staged changes that were added for testing. All test results are compared against the Primary dataplane.

Secure where AI is running

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

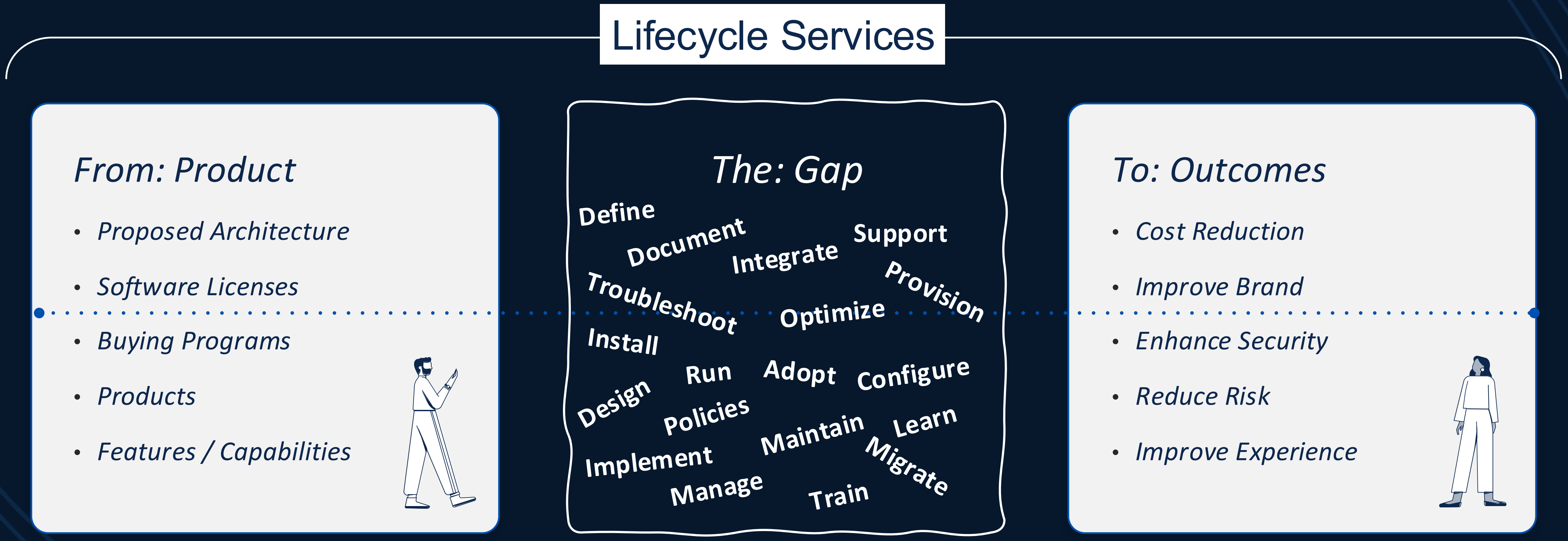
Cisco Services

Customer Experience CX

The empowering force to help you achieve AI Ready DC

Cisco Services

CX Bridge the Gap



Lifecycle Services

From: Product

- *Proposed Architecture*
- *Software Licenses*
- *Buying Programs*
- *Products*
- *Features / Capabilities*

The: Gap

Define
Document
Integrate
Support
Troubleshoot
Optimize
Provision
Install
Run
Adopt
Configure
Design
Policies
Maintain
Learn
Implement
Migrate
Manage
Train

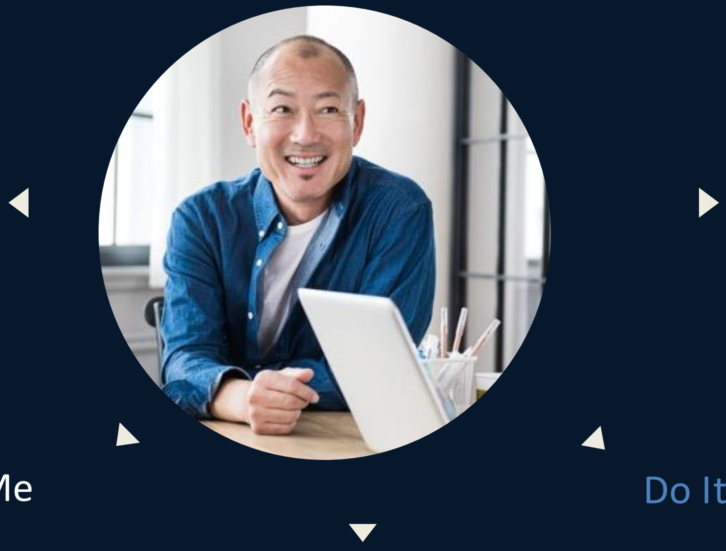
To: Outcomes

- *Cost Reduction*
- *Improve Brand*
- *Enhance Security*
- *Reduce Risk*
- *Improve Experience*

CX and One Cisco

CX Offers to AI-Ready Data Centers

Together with our partners, we get you to outcomes faster



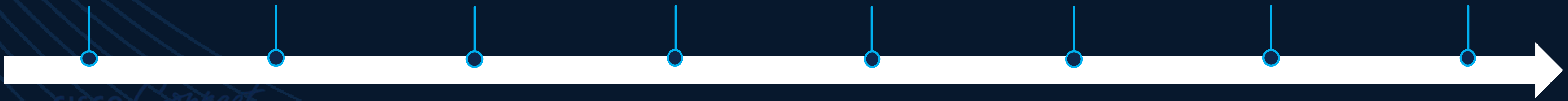
Empower Me

Do It For Me

Advise Me

Do It With Me

Guide Me



Discovery Phase

Deployment Phase

Development Phase

Qualification

Ideation

Business Case

Proof of Concept

Pilot/ Prototype

Production

Postproduction

Expansion and Scaling

Cisco CX Unifies : AI Cluster Build & Scale

The empowering force to help you achieve AI Ready DC

AI Platform – Plan & Design

Implementation, Test & Validate

AI Infrastructure Optimization & Support

Infrastructure Fine Tuning

AI Infra Consulting Support

CISCO SERVICES

<<<<<< *Accelerated by Cisco CX* >>>>>>

AI Ready Data Center - CX Services

AI infrastructure Readiness

- AI Infrastructure Readiness

AI POD Inference

- AI POD Adoption Service

Private AI Foundation

- Customize Private AI for Training and Inference
- AI Ready Full Stack Infrastructure
- AI ML Network Services

AI Digital Resilience

- AI Observability
- AI Security

AI Ready Data Center - CX Services

Comprehensive Approach

Design | *Implement* | *Fine Tune* | *Operations*

AI Compute & Networking

AI Infrastructure Readiness

Review and Analyze existing Infra
Identify gaps for AI Infra

AI POD Deployment Services

Leverage standardized architecture for AI POD deployment and adoption for Inference

Private AI – Compute & Network Services

- Infra Readiness, POCs
- Customized approach for a combined AI Compute & Networking env based on use cases
- Work alongside with teams to Fine tune & Optimize the AI Ready infrastructure

AI Security

Securing AI

Comprehensive approach to protecting the development, deployment and use of AI applications.

Secure AI Ready DC Services

- Safeguard AI systems against evolving threats by implementing comprehensive strategies that address data integrity, model protection
- Advanced measures such as zone firewalling and segmentation to ensure robust protection against data breaches and cyberattacks

AI Observability

AI Observability

Cisco's AI Observability services provide a comprehensive framework for ensuring resilience and performance with real-time insights, secure operations, and actionable data

AI Operations

Optimize AI infrastructure while preventing degradations and downtime

AI Ready Data Center

Defining Infrastructure Requirements

Requirements

- *What is the use case?*
- *Am I training? Fine tuning? Inferencing?*
- *Am I using private data?*
- *Do I need a dedicated network?*
- *Who is responsible for management?*
- *Where will the workload live?*

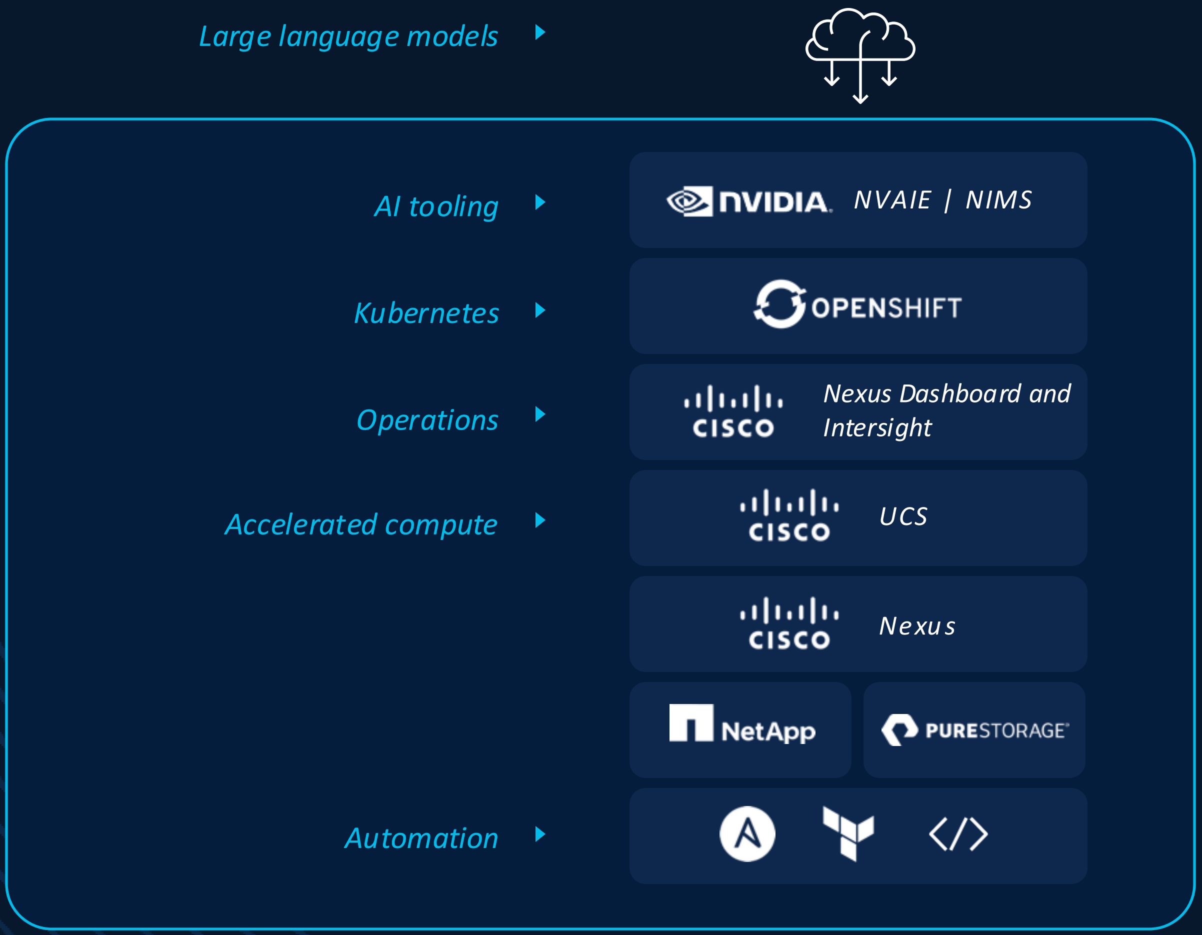
Considerations

- *Cost*
- *Accuracy*
- *Model Size*
- *User Experience (response time)*
- *Concurrent users*
- *Data Fidelity*
- *Power Requirements*
- *Cooling and Power Management*
- *Software Ecosystem*
- *Data Management Strategy*

AI Compute Stack

CX Services

AI POD for Inferencing



Large language models

AI tooling — NVIDIA *NVAIE | NIMS*

Kubernetes — OPENSHIFT

Operations — cisco *Nexus Dashboard and Intersight*

Accelerated compute — cisco *UCS*

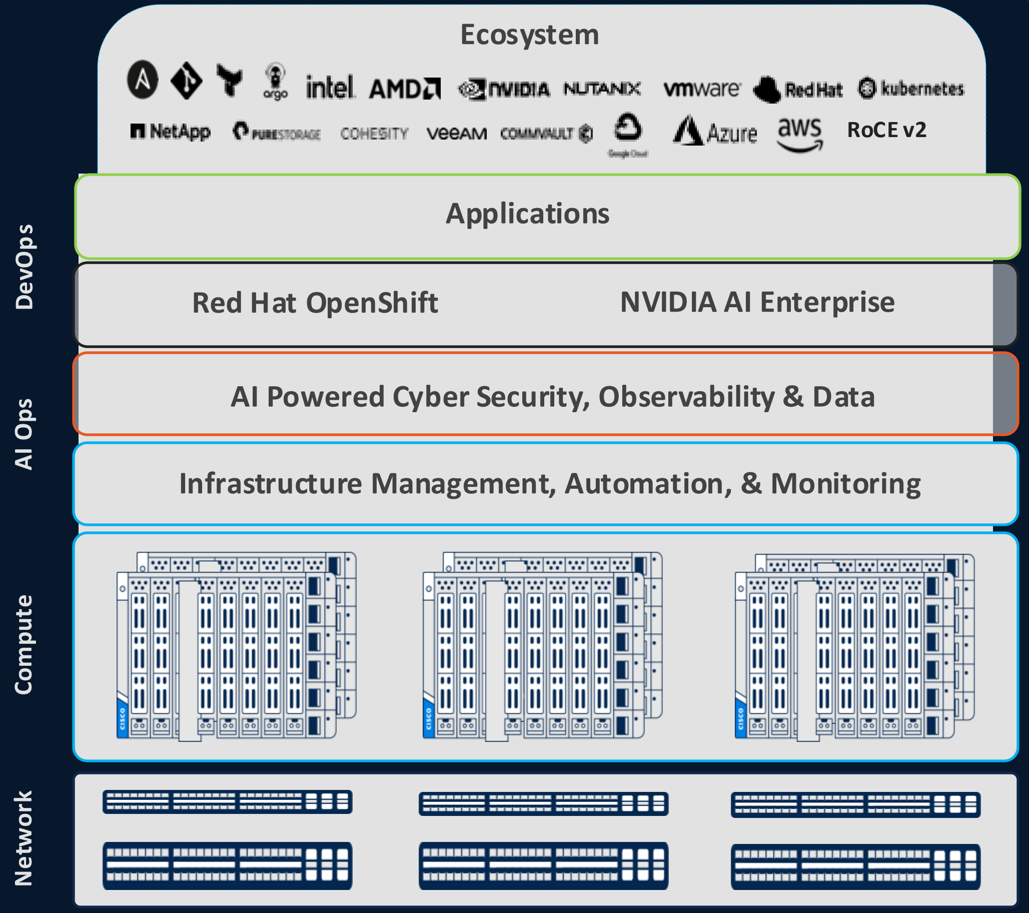
cisco *Nexus*

NetApp

PURESTORAGE

Automation

Private AI for Training/Inferencing



Ecosystem

RoCE v2

DevOps

Applications

Red Hat OpenShift

NVIDIA AI Enterprise

AI Ops

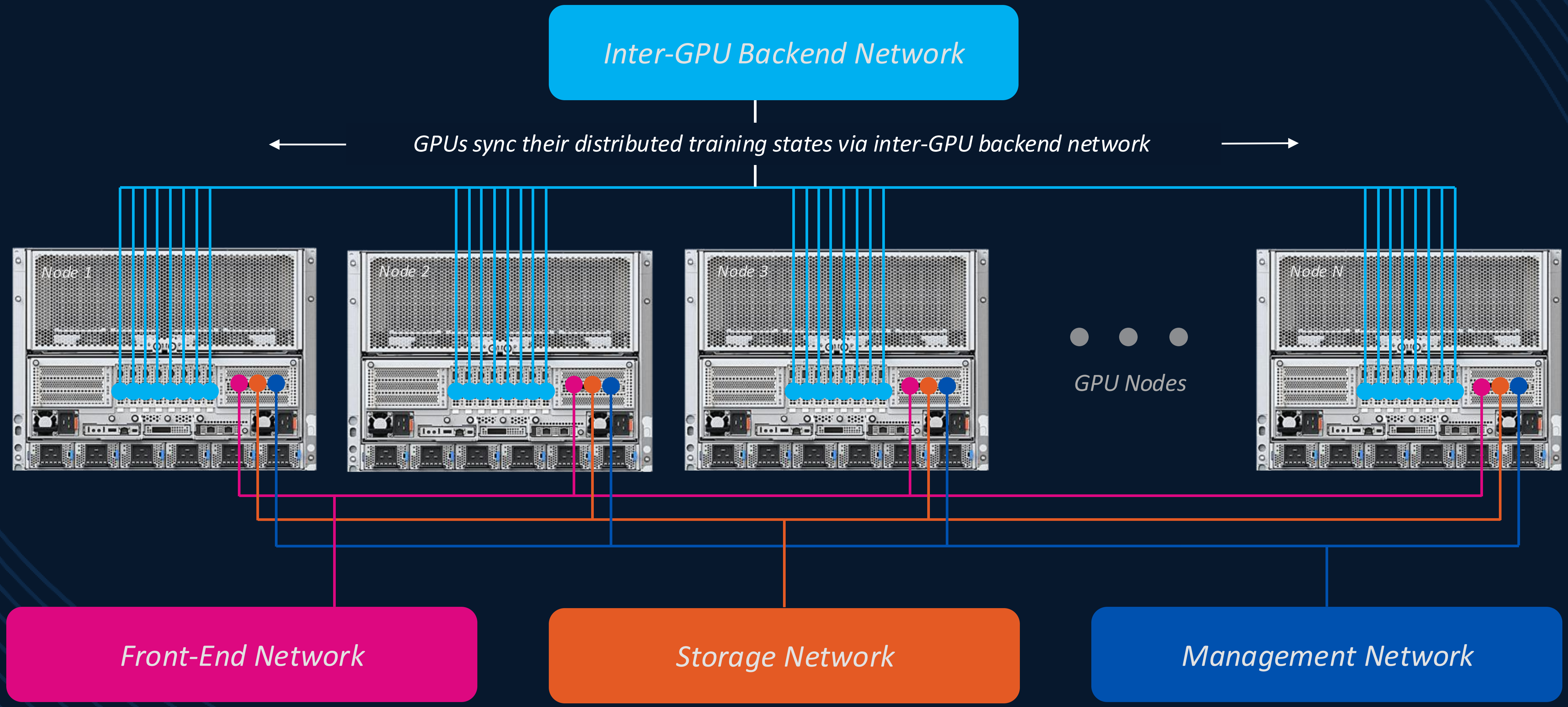
AI Powered Cyber Security, Observability & Data

Infrastructure Management, Automation, & Monitoring

Compute

Network

Cisco AI Cluster



Inter-GPU Backend Network

GPUs sync their distributed training states via inter-GPU backend network

GPU Nodes

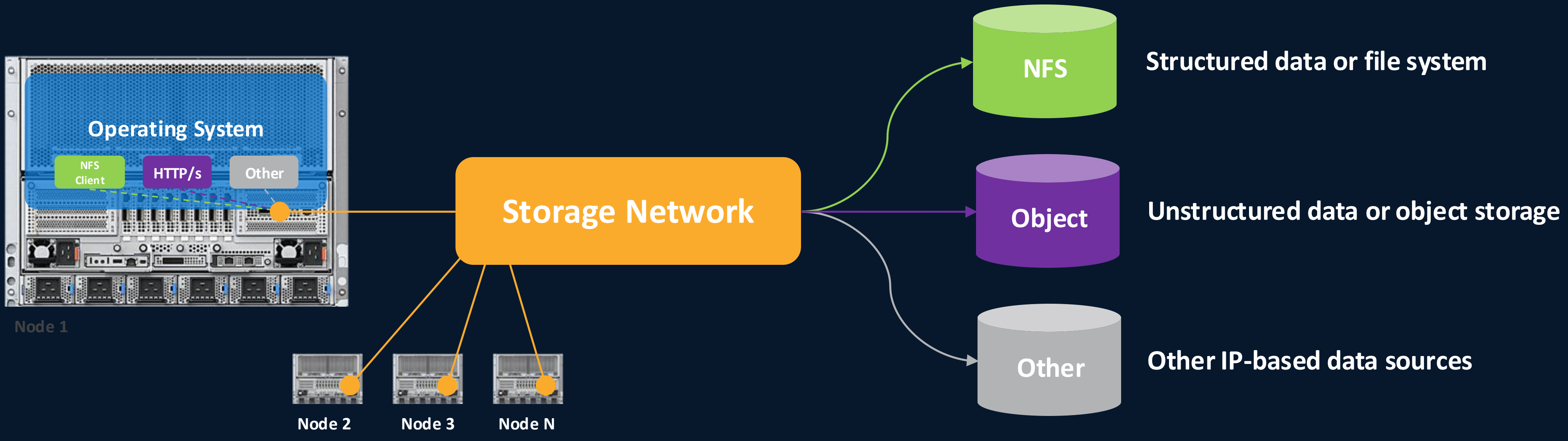
Front-End Network

Storage Network

Management Network

Storage Connectivity

- Support for standard IP-based storage protocols leveraging OS/software clients
- Connects to Shared storage for Loading (reading), Checkpointing (writing) model parameters



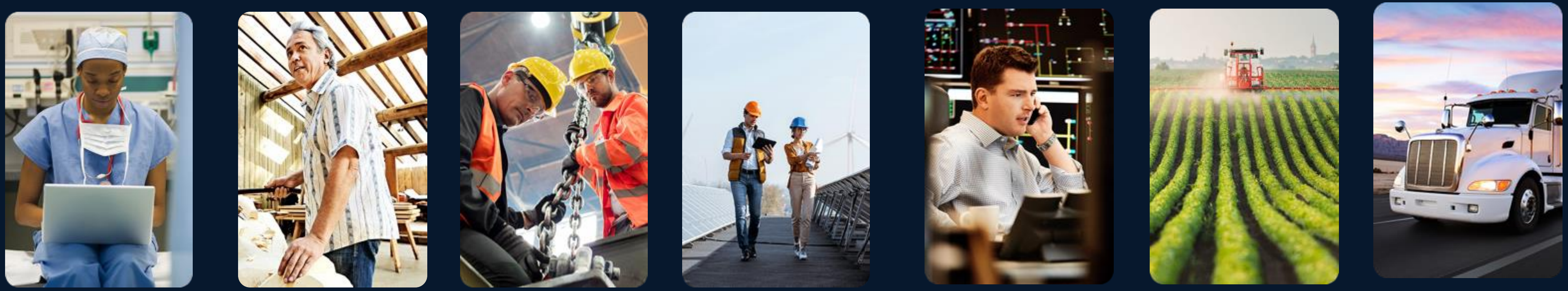
Operating System
NFS Client
HTTP/s
Other
Node 1
Storage Network
Node 2
Node 3
Node N
NFS
Structured data or file system
Object
Unstructured data or object storage
Other
Other IP-based data sources

Note: Most customers leverage a dedicated storage network, but storage traffic can be aggregated with other north-south networks as well

PARTNERSHIPS WITH

PURESTORAGE® NetApp

Every AI/ML Network is different



Build the model
Training

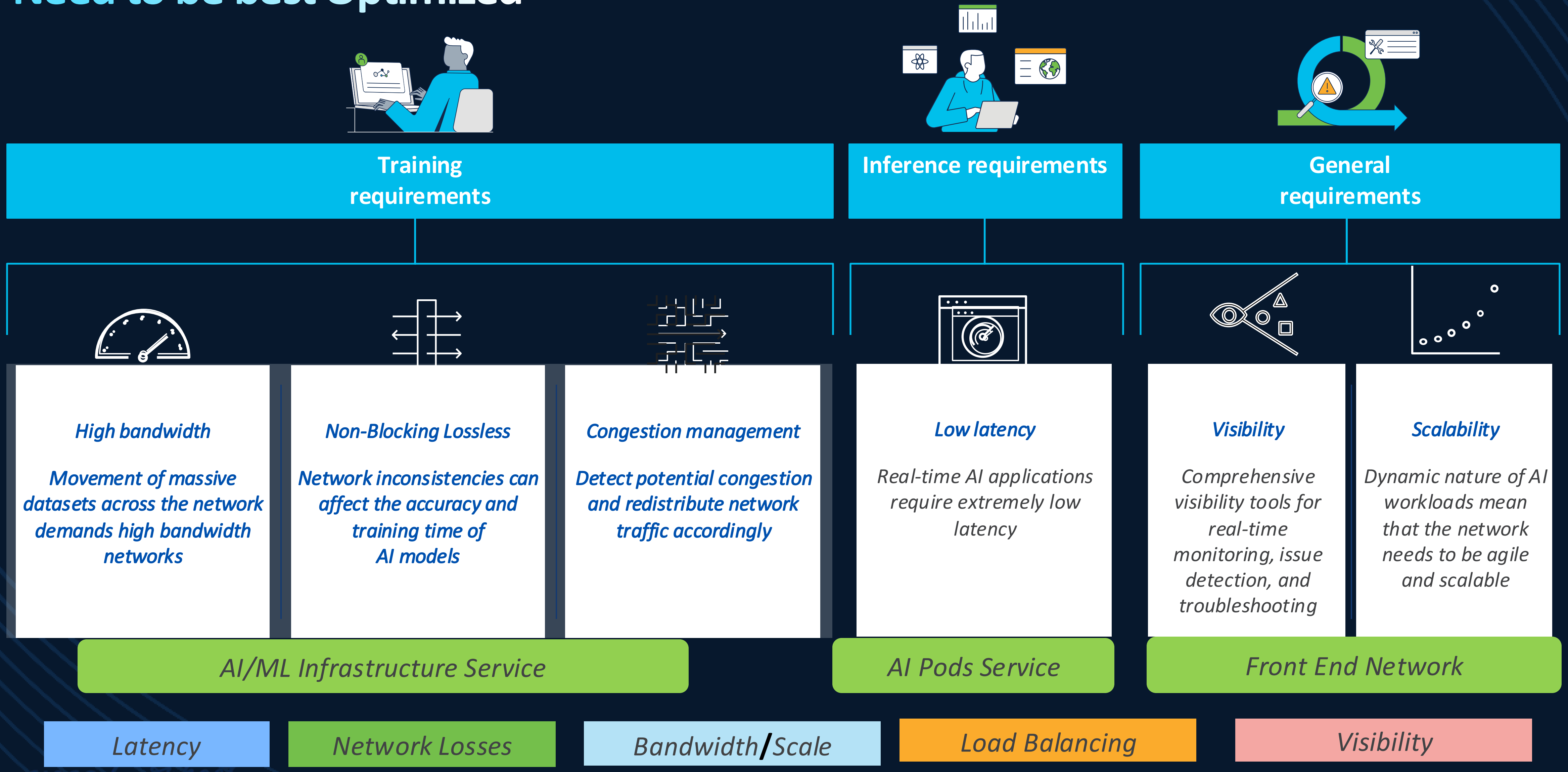
Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

From Training to Inferencing: A Songwriting Analogy

AI/ML Network - CX Services

Need to be best Optimized



Training requirements

Inference requirements

General requirements

High bandwidth

Movement of massive datasets across the network demands high bandwidth networks

Non-Blocking Lossless

Network inconsistencies can affect the accuracy and training time of AI models

Congestion management

Detect potential congestion and redistribute network traffic accordingly

Low latency

Real-time AI applications require extremely low latency

Visibility

Comprehensive visibility tools for real-time monitoring, issue detection, and troubleshooting

Scalability

Dynamic nature of AI workloads mean that the network needs to be agile and scalable

AI/ML Infrastructure Service

AI Pods Service

Front End Network

Latency

Network Losses

Bandwidth/Scale

Load Balancing

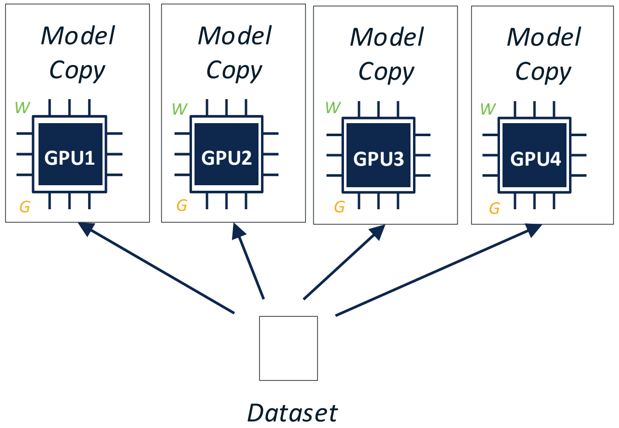
Visibility

Optimizing AI/ML Cluster

Parallelism Techniques - Distributed Training and Inference

About running AI smarter and more efficiently

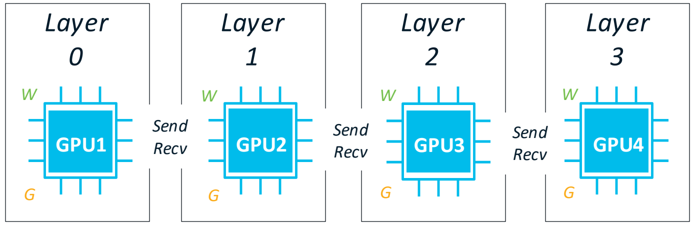
Data Parallelism



Uses All-Reduce to sum all gradients together for the full batch, update weights, learn, repeat

Model Parallelism

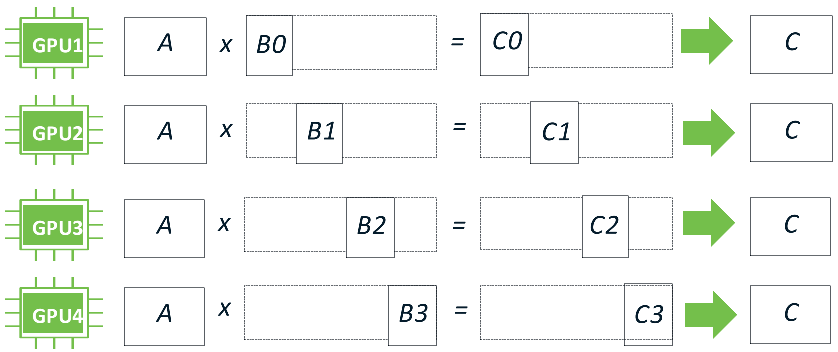
Pipeline



It splits a model into stages or groups of layers across multiple GPUs.

The dataset goes from one GPU to the next using send/receive operations.

Tensor



Groups of parameters (weights, gradients) within a layer are split across GPUs.

Uses All-Gather for all GPUs to get aggregated results

Digital resilience for AI-ready data centers

Digital resilience

Security

Assurance

Observability

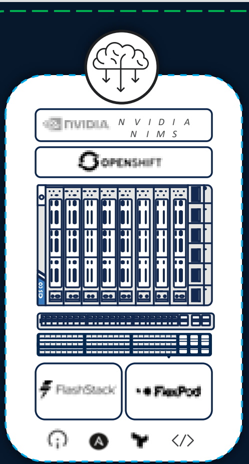
AI-ready data centers

Build the Model I Training

Optimize the Model I Fine-tuning and RAG

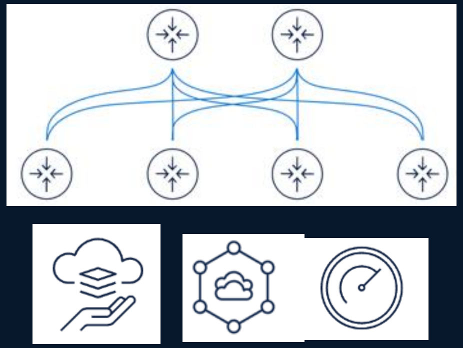
Use the Model I Inferencing

Model the World I Digital Twin

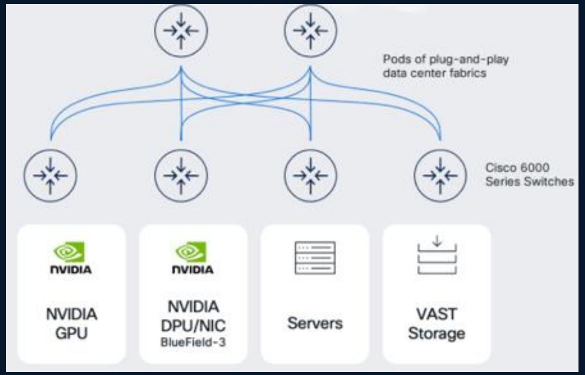


AI PODs Data Center (inferencing suite)

Cisco UCS C-Series Accelerated (GPUs)



Cisco Nexus HyperFabric


Pods of plug-and-play data center fabrics
Cisco 6000 Series Switches
NVIDIA GPU
NVIDIA DPU/NIC BlueField-3
Servers
VAST Storage

Cisco Nexus HyperFabric + AI Cluster

Private cloud managed AI Infrastructure

Cloud managed AI Infrastructure

PROTECT ASSETS | OPTIMISZE PERFORMANCE | MAINTAIN TRUST

Security

Observability

Data Protection

Protecting data from breaches and misuse

Competitive Advantage

Regulatory Compliance

Model Integrity

Ensure models remain untampered and resistant to attacks

Business Continuity

Customer Retention

AI Operations

Optimize operations to drive digital resilience

Rapid Detection and Resolution

Controlled AI Costs

Digital Experience

Realtime insights assure the user AI experience

Accelerated adoption

Improved Brand Loyalty

Cisco Secure AI

Changing the fundamentals of data center security

Visibility | *Zone firewalling* | *Segmentation* | *Exploit protection*

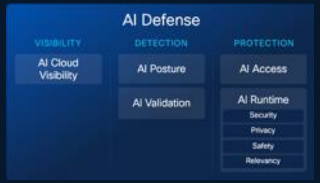
Cisco Security Cloud Control

Secure Firewall

Secure Workload

Isovalent

Hypershield



AI Defense

VISIBILITY

AI Cloud Visibility

DETECTION

AI Posture

AI Validation

PROTECTION

AI Access

AI Runtime

Security

Privacy

Safety

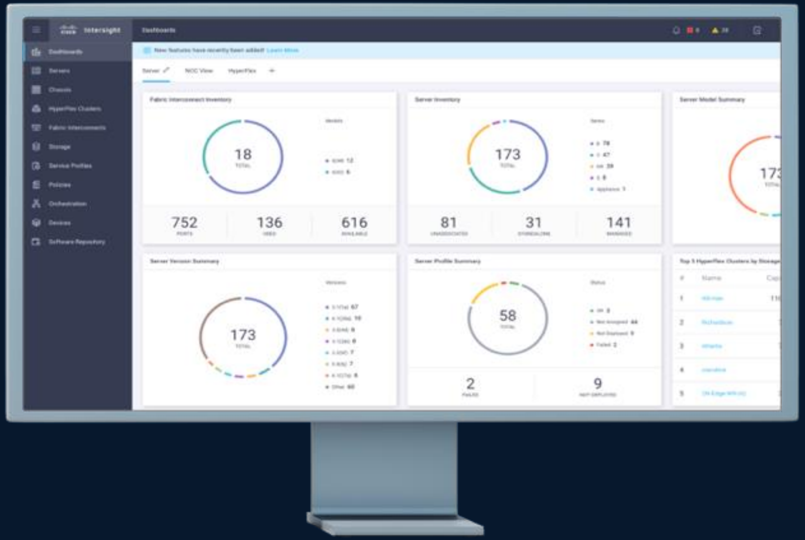
Relevancy

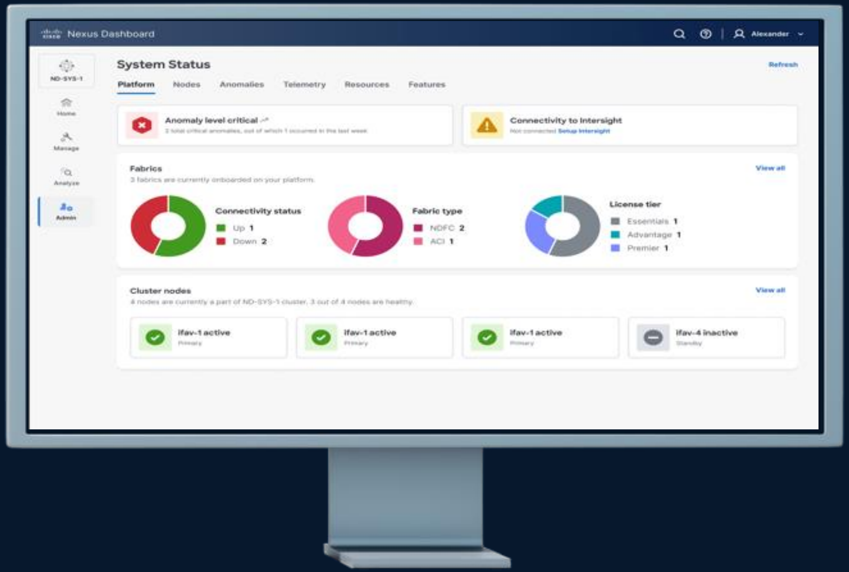
AI Defense

Security infused into the network through a fabric of enforcement points

Observability for AI

Unified Experience for Infrastructure Operations





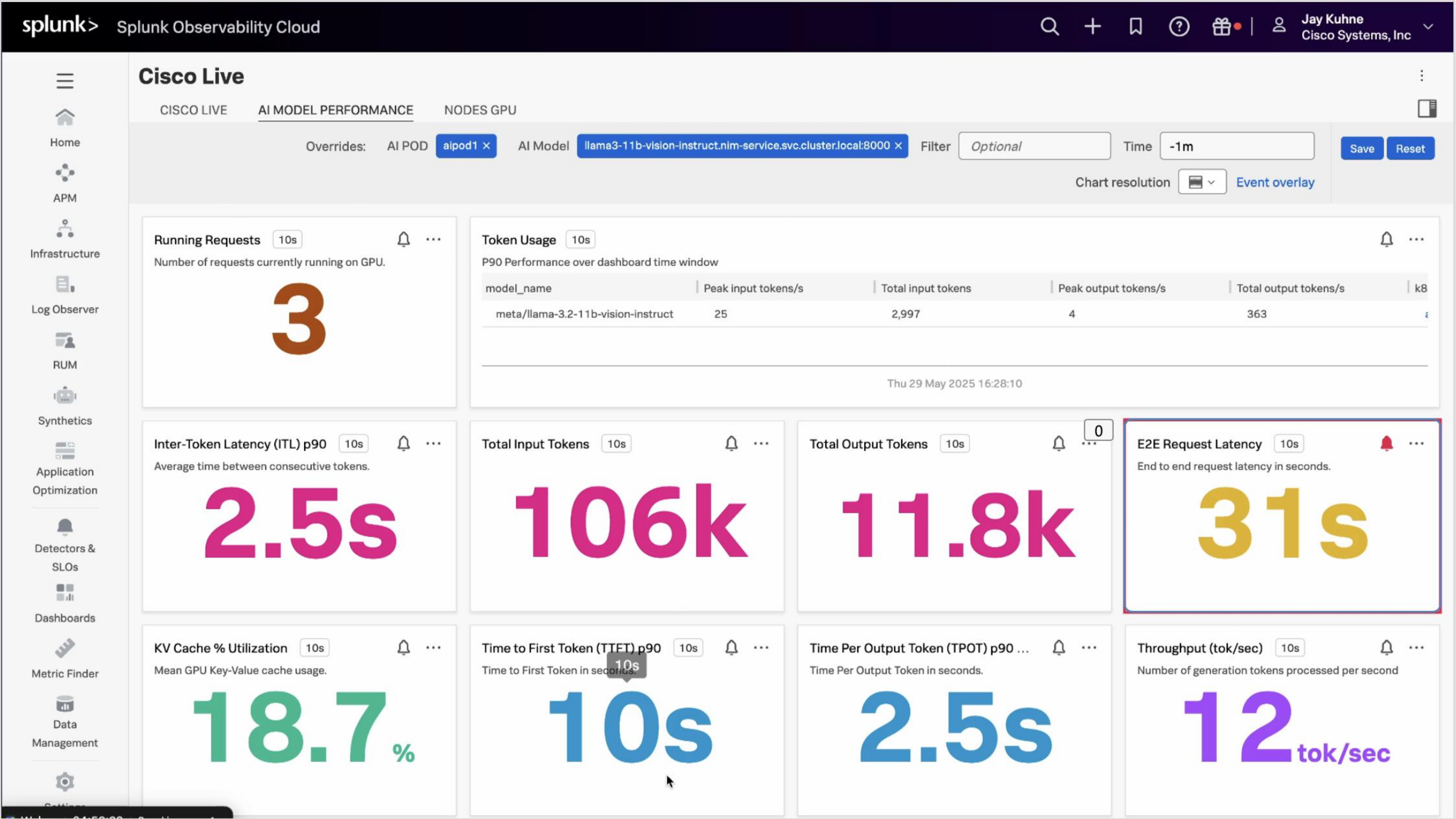
Cisco Intersight splunk> APPDYNAMICS **Cisco Nexus Dashboard**

See Connect Secure Automate

Simplified Accelerated Insights

Powerful Integrations with industry leading platforms

Splunk Observability for AI



Cisco Live

Overrides: AI POD aipod1 × AI Model llama3-11b-vision-instruct.nim-service.svc.cluster.local:8000 × Filter Optional Time -1m

Running Requests 10s
Number of requests currently running on GPU.
3

Token Usage 10s
P90 Performance over dashboard time window

model_name	Peak input tokens/s	Total input tokens	Peak output tokens/s	Total output tokens/s	k8
meta/llama-3.2-11b-vision-instruct	25	2,997	4	363	a

Thu 29 May 2025 16:28:10

Inter-Token Latency (ITL) p90 10s
Average time between consecutive tokens.
2.5s

Total Input Tokens 10s
106k

Total Output Tokens 10s
0
11.8k

E2E Request Latency 10s
End to end request latency in seconds.
31s

KV Cache % Utilization 10s
Mean GPU Key-Value cache usage.
18.7%

Time to First Token (TTFT) p90 10s
Time to First Token in seconds.
10s
10s

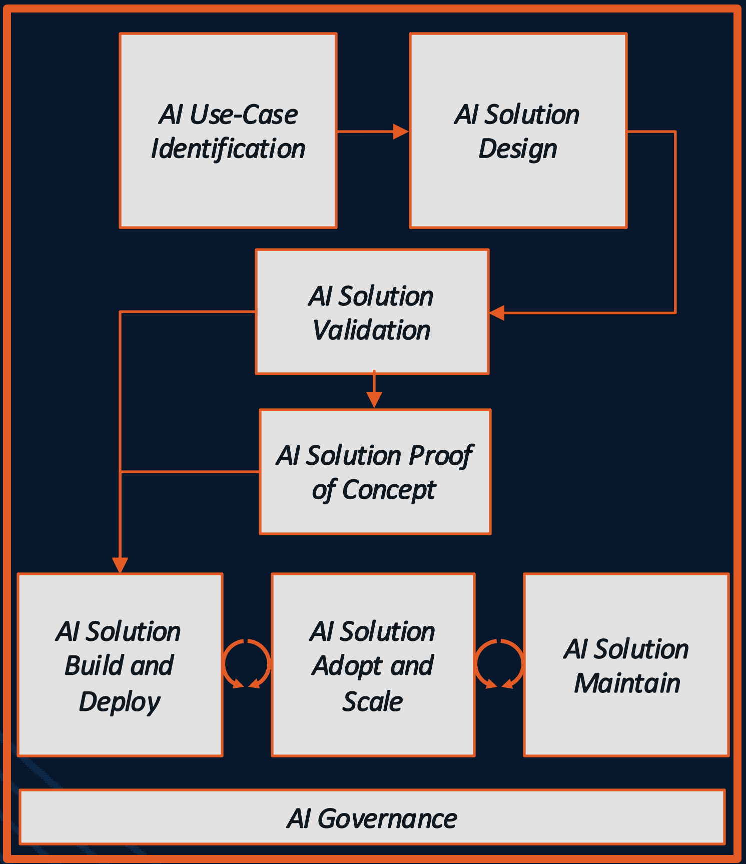
Time Per Output Token (TPOT) p90 ... 10s
Time Per Output Token in seconds.
2.5s

Throughput (tok/sec) 10s
Number of generation tokens processed per second
12 tok/sec

AI Ready Data Center Services Portfolio

AI Adoption Journey/Framework

STRATEGY



AI Use-Case Identification

AI Solution Design

AI Solution Validation

AI Solution Proof of Concept

AI Solution Build and Deploy

AI Solution Adopt and Scale

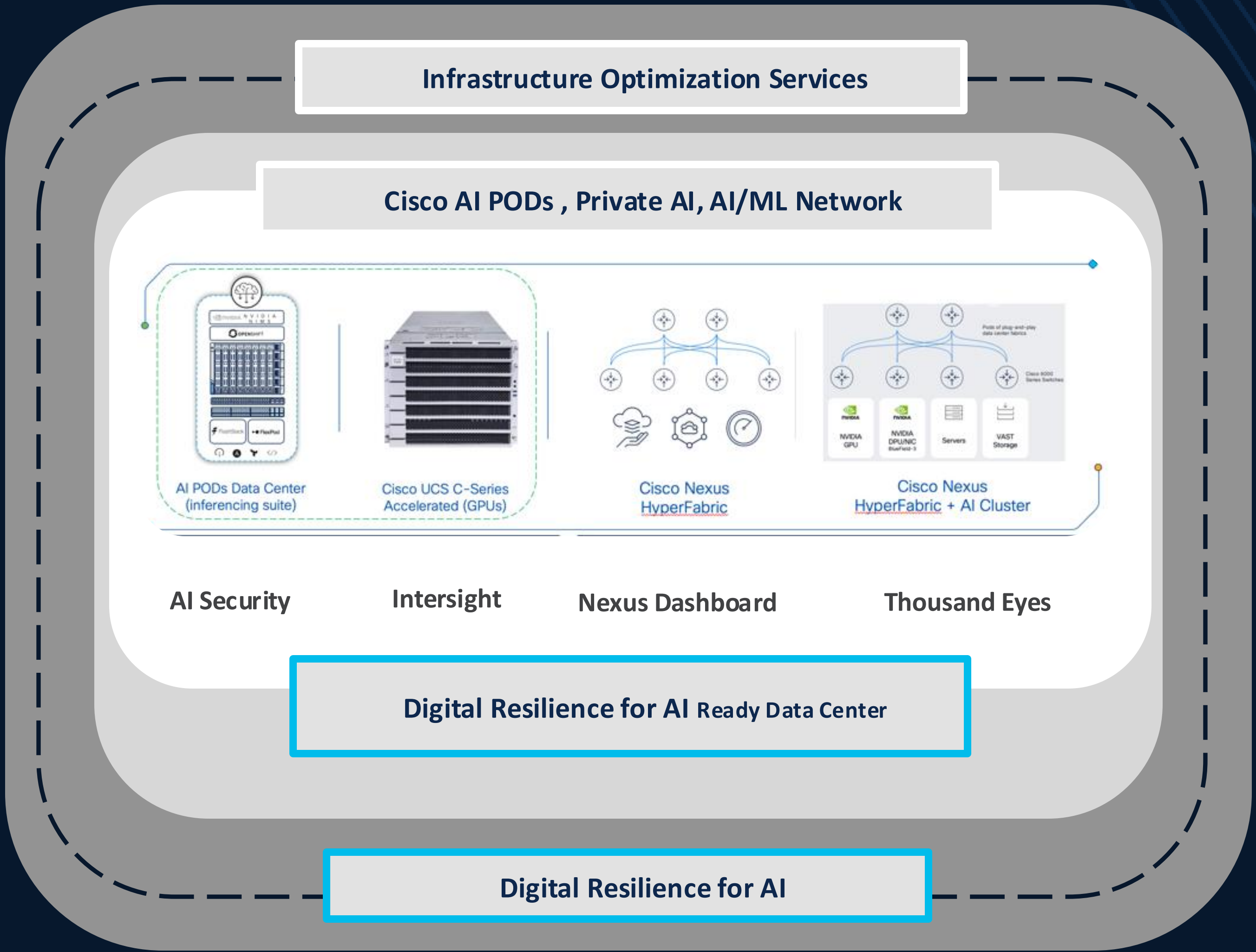
AI Solution Maintain

AI Governance

+

Design , Deploy & Observ

CX Services Portfolio for AI Ready DCs



Infrastructure Optimization Services

Cisco AI PODs , Private AI, AI/ML Network

AI PODs Data Center (inferencing suite)
Cisco UCS C-Series Accelerated (GPUs)
Cisco Nexus HyperFabric
Cisco Nexus HyperFabric + AI Cluster
NVIDIA GPU
NVIDIA DPU/NIC
Servers
VAST Storage

AI Security

Intersight

Nexus Dashboard

Thousand Eyes

Digital Resilience for AI Ready Data Center

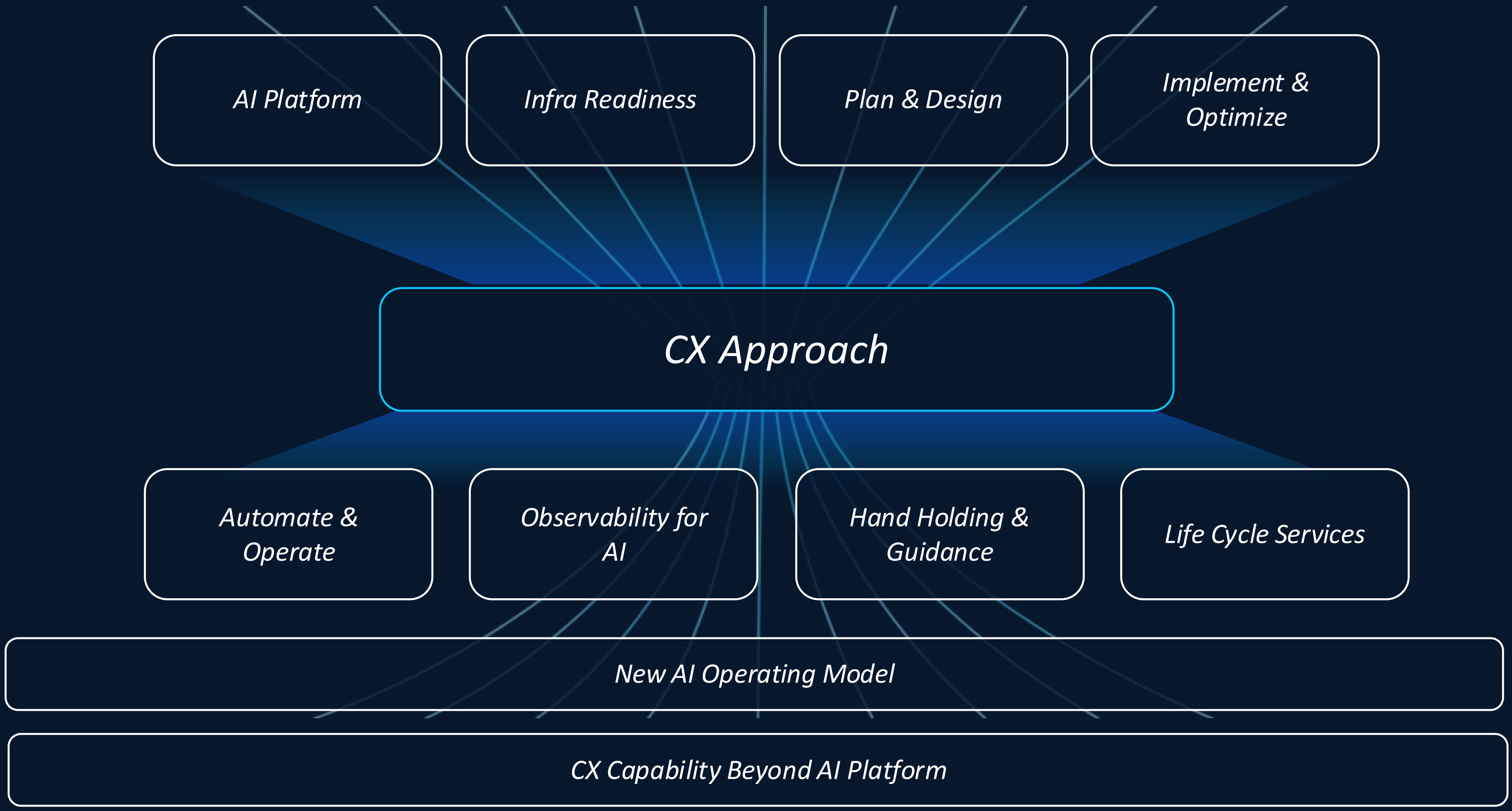
Digital Resilience for AI

□ *Product*

▨ *Implementation Services*

■ *Premium Services*

AI Ready Enterprises



AI Platform

Infra Readiness

Plan & Design

Implement & Optimize

CX Approach

Automate & Operate

Observability for AI

Hand Holding & Guidance

Life Cycle Services

New AI Operating Model

CX Capability Beyond AI Platform

* Ripple Effect - Beyond DC – Ex: DCI, SDWAN , Campus, Security , Ops, D2 Services ...

AI Ready Enterprises

Enterprise AI Use cases

Knowledgebase copilots

AI assistants

Content & code generation

Text | Images | Video | Code

Security

Physical | Virtual

Language translation

Multilingual real-time communication

Virtual agent & chatbots

Specialized domain specific chatbots

Detection, prediction & reporting

Analytics | Forecasts | Anomalies | Insights

AI Ready Data Center – CX Services

Customer Asks

Early Alignment	Rapid Prototype	Security for AI	E2E Services	Storage & Data
AI Governance	As a Service	Adoptable Network	Expansion	Constraint Handling
Resource Management	Customer Ref	Cloud vs OnPrem	Convergence	Operational Challenges

AI Is Rising — The Real Wave Is Still Ahead

Agentic workflows will power the next wave of demand

LAYER	DRIVER	IMPACT
Adoption	Commoditization 💸	↑ Lower cost & demand growth
X	Better LLMs 🧠	↑ Addressable tasks
Capability	Multi-modal 📊	↑ Payload size
X	Reasoning 🧮	↑ Tokens & Compute
Interaction	AI Agents 🤖	↑ Transactions per request
	Multi-agent 🌐	↑ Requests per session

1448%
Increase in tokens processed by AI models in the last 12 months[1]

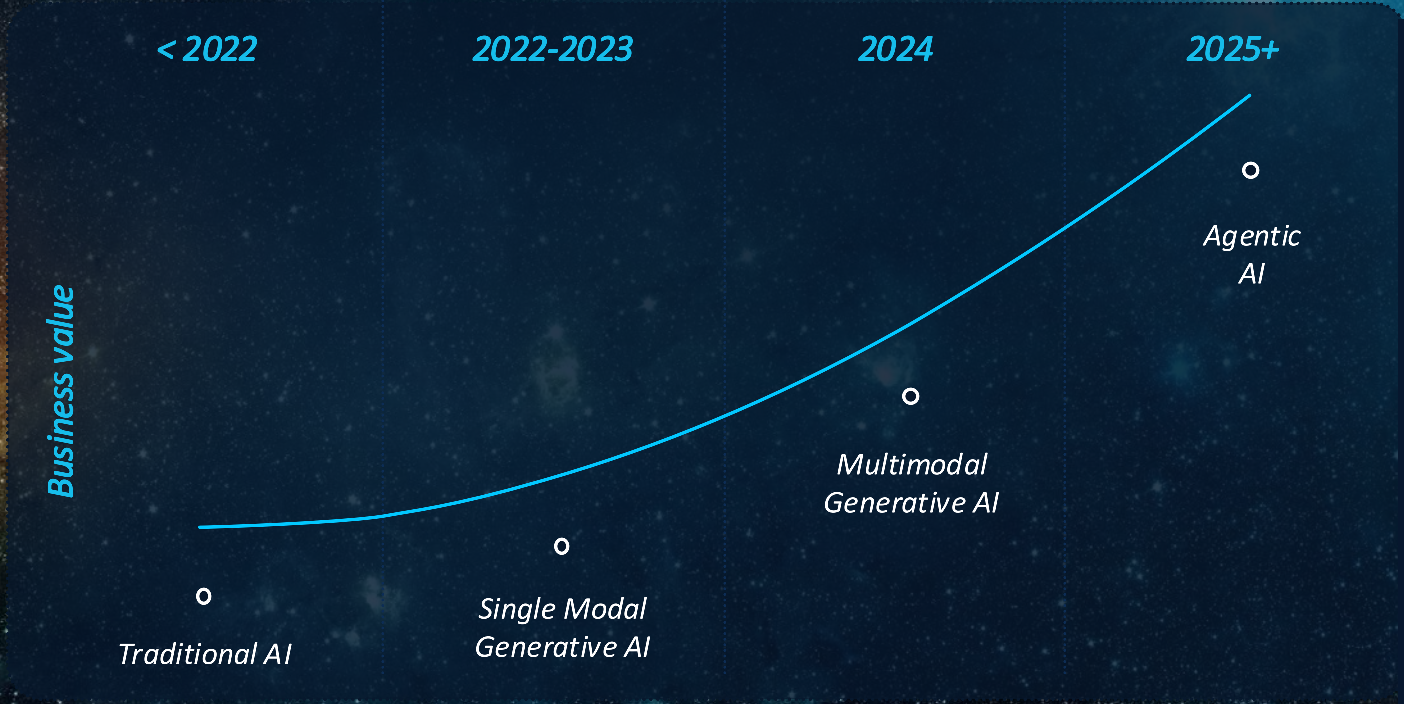
75%+
Of inference-driven data creation and processing at the edge by 2030[2]

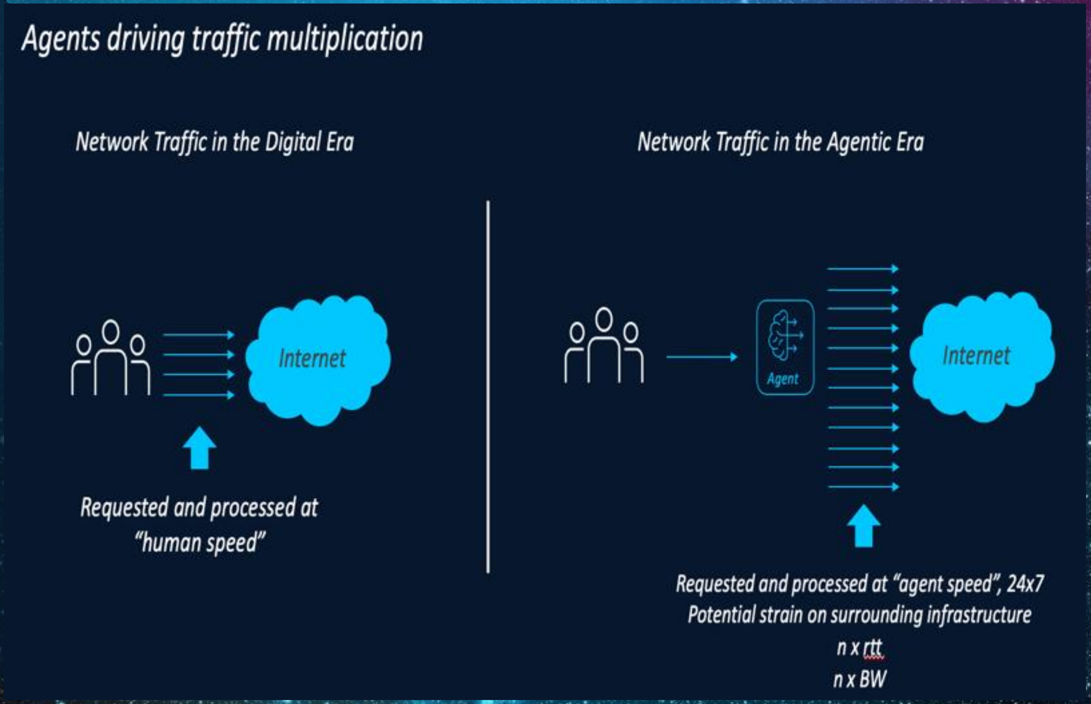
10x
Bandwidth required upstream compared to downstream

36x
Increase in AI traffic as early as 2023-2024[1]

1: Openrouter.ai/rankings 2: KPMG

The Rise of Agentic AI





Agents driving traffic multiplication

Network Traffic in the Digital Era

Requested and processed at "human speed"

Network Traffic in the Agentic Era

Requested and processed at "agent speed", 24x7
Potential strain on surrounding infrastructure
n x rtt
n x BW

"33% of enterprise software applications will include agentic AI by 2028, up from less than 1% in 2024"

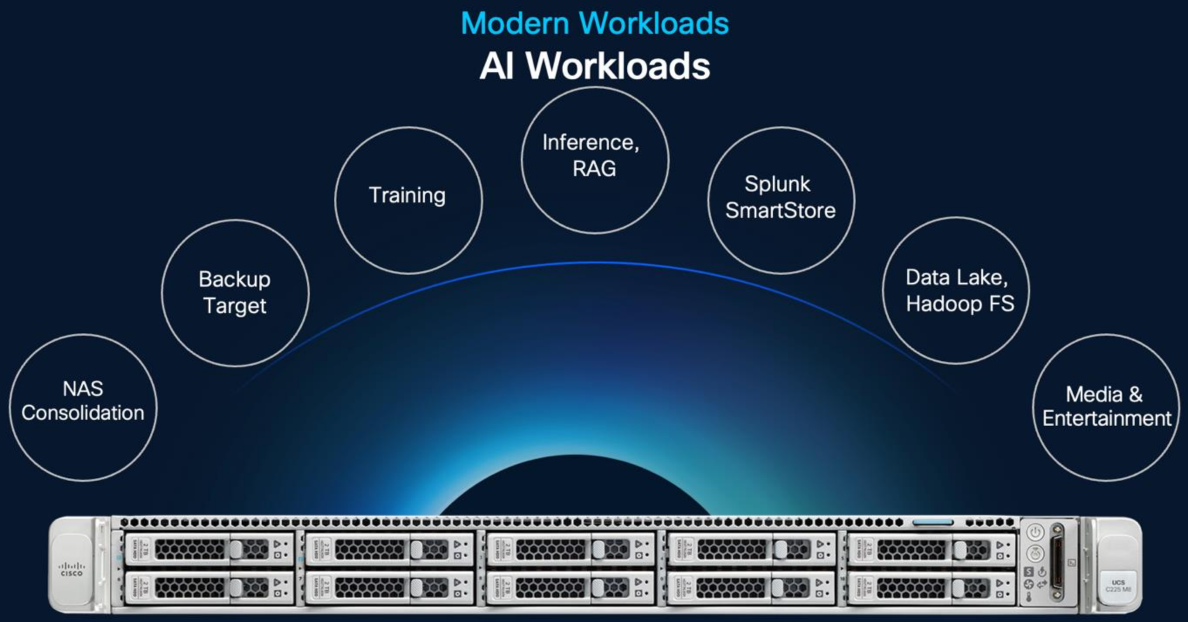
— Gartner

AI is re-shaping DCs

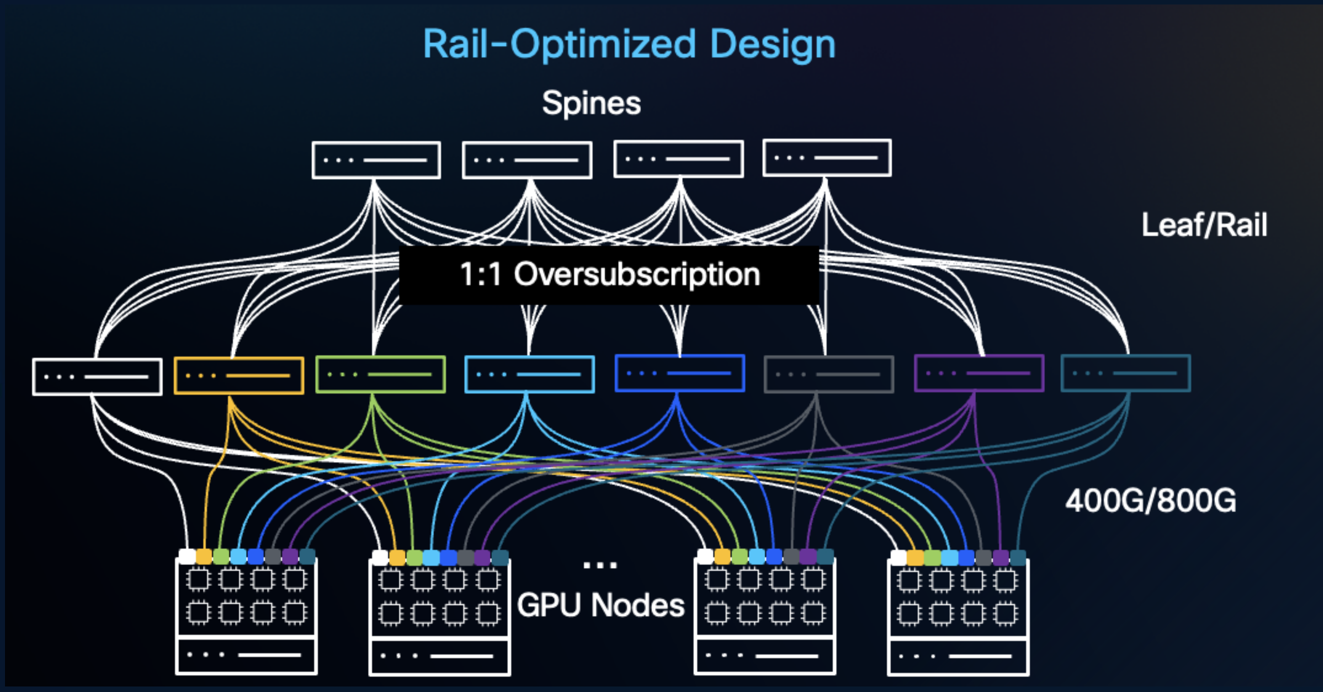
AI infrastructure for enterprise

Enterprise data center AI + traditional

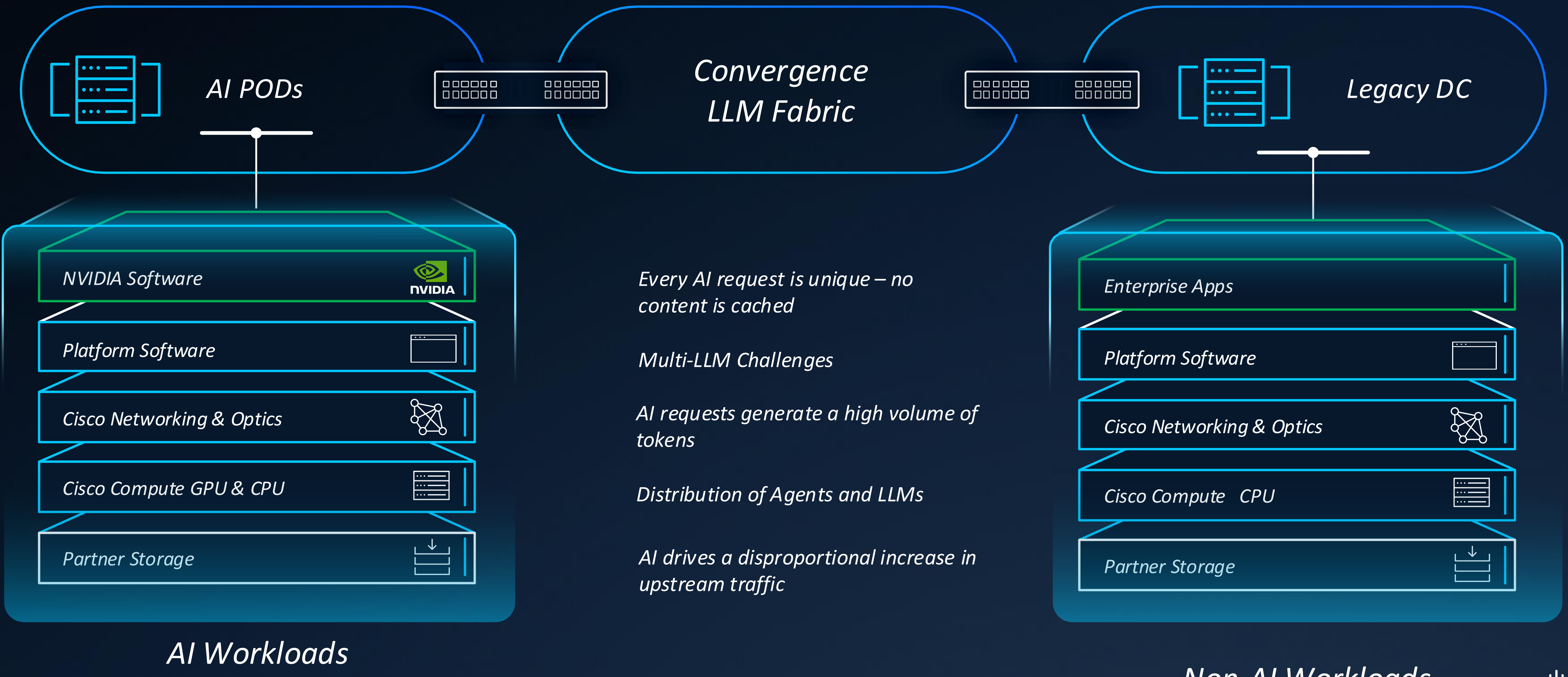
AI infrastructure for hyperscalers



Key	Training	Fine Tuning	Inference	RAG/Embedding	Data Analytics	Simulation	Visual Computing
	Model Size: Medium-→ Small	Model Size: Medium-→ Small	AI Agents, SDG, VSS, CV, ReccSys	Multimodal Retrieval Pipelines	Spark RAPIDS, Data Science	CAE, CFD, Scientific Simulation	Digital Twins, CAD, Rendering, VDI



Enterprise AI Convergence



Every AI request is unique – no content is cached

Multi-LLM Challenges

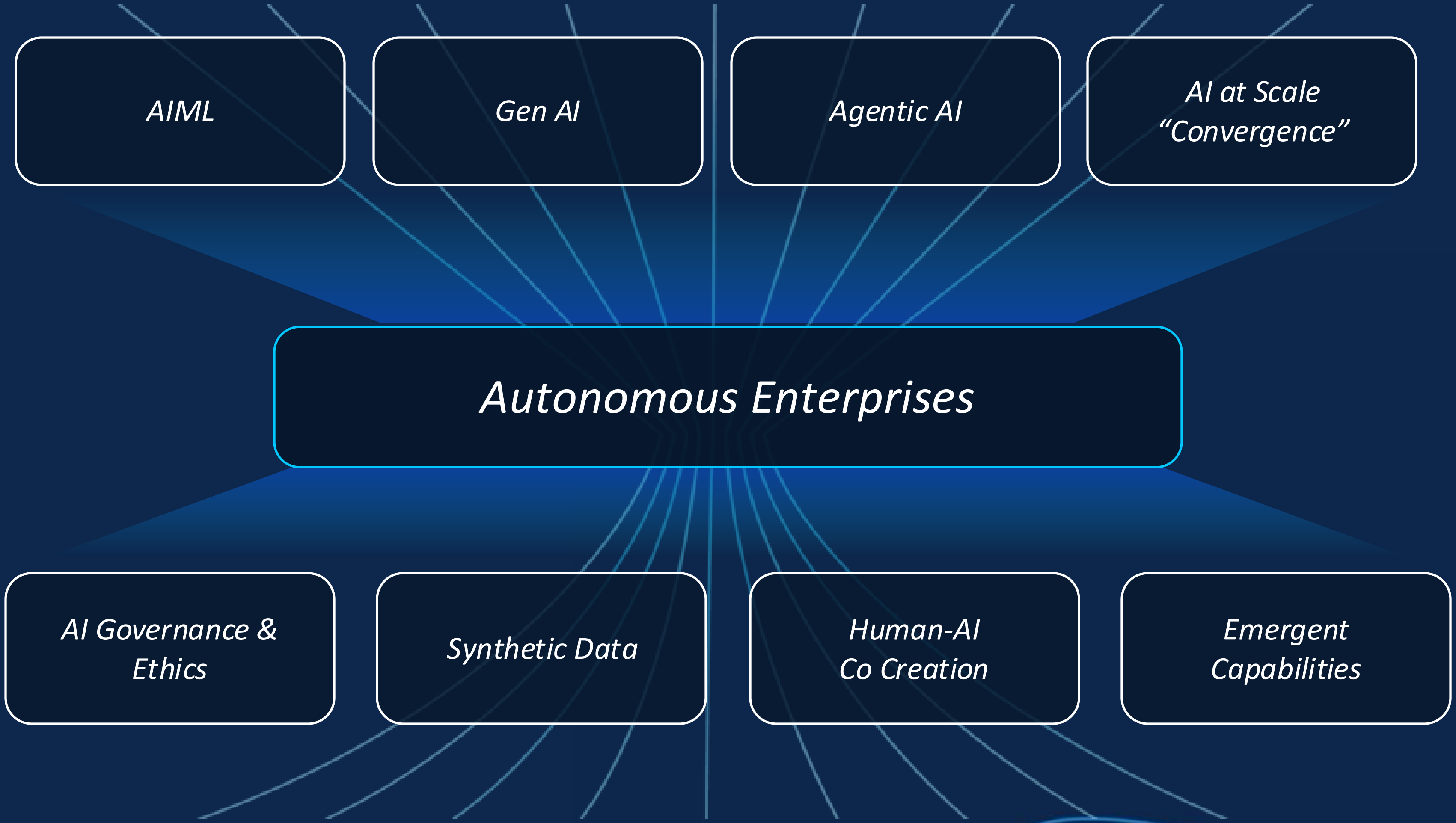
AI requests generate a high volume of tokens

Distribution of Agents and LLMs

AI drives a disproportional increase in upstream traffic

Future AI Ready Enterprises

Token Demand Inflation



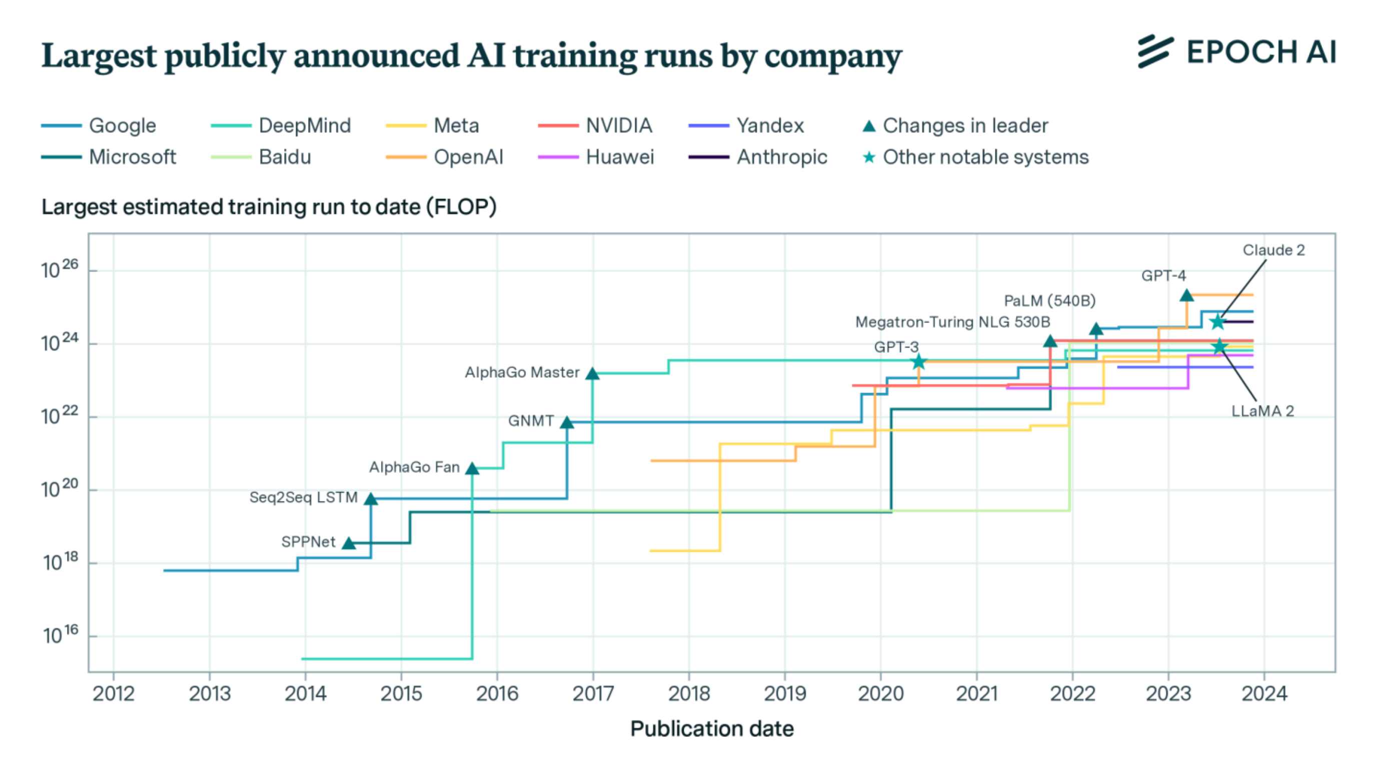
* Ripple Effect - Beyond DC – Ex: DCI, SDWAN , Campus, Security , Ops, D2 Services …

Future of AI !

Industry view

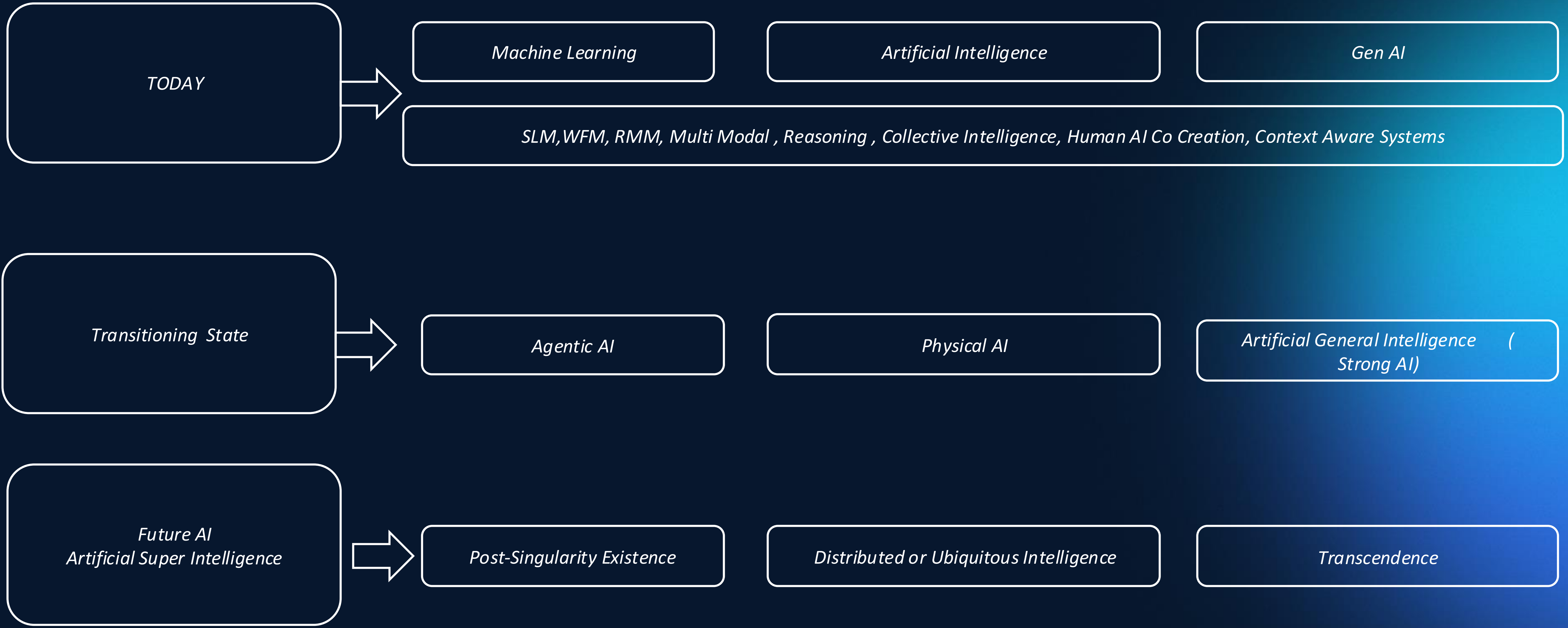
No analyst firms have an overall AI report

- *No industry standard of the AI market*
- *AI made up of several market segments*
- *No magic quadrant for AI*
- *AI is Reported in all segments*
- *AI effects all market segments*
- *Market research suggest Enterprise AI will be consumed as a service*



Largest publicly announced AI training runs by company

The Future !



TODAY

Machine Learning

Artificial Intelligence

Gen AI

SLM,WFM, RMM, Multi Modal , Reasoning , Collective Intelligence, Human AI Co Creation, Context Aware Systems

Transitioning State

Agentic AI

Physical AI

Artificial General Intelligence (Strong AI)

Future AI
Artificial Super Intelligence

Post-Singularity Existence

Distributed or Ubiquitous Intelligence

Transcendence

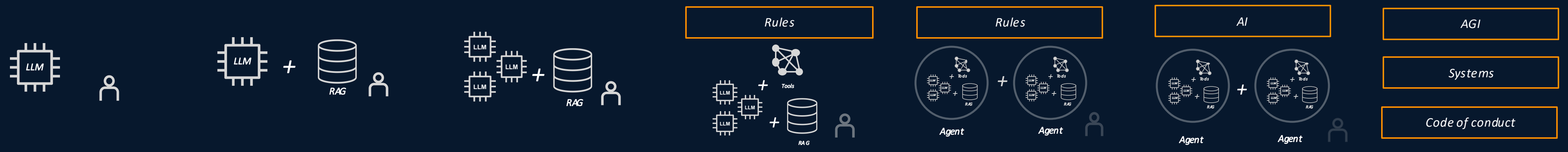
The Roadmap

AI Chat

Attended AI Agents

Unattended AI Agents

	2022	2023	2024	2024/2025	2026/2027		2030+?
	Standalone AI Chat	RAG Powered AI Chat	Reasoning AI Chat	Standalone AI Agent	Multi AI Agent System	Built for purpose	General purpose
Value	Low (Micro efficiencies)	Low/Medium	Medium	Medium/High	High	Very High	Revolutionary
Need for low latency	Low	Medium	High	High	Very High	Very High	Unknown
Volume / tokens	Medium	Low/Medium	High	High	Very High	Extremely High	Unknown
Risk exposure	Medium	High	High	Very High	Very High	Extremely High	Unknown



Definitions: LLM = Large Language Model, LMM = Large Multimodal Model, UI = User Interface, RAG = Retrieval Augmented Generation, API = Application Programming Interface,

AI will make our world of *8B* people feel like one with the capacity of *80B*

“80B Agents + 8B”

Cisco CX along with our Partners unifies networking, compute, security, and observability to deliver AI-ready Accelerated data centers.

Thank you

CISCO