CISCO Connect

Innovate with AI. Connect with Resilience.
Cisco Connect 2026 Hong Kong

The Blueprint of Building AI Ready Data Center

July 15, 2026

CISCO Connect

The Blueprint of Building AI Ready Data Center

Louis Mah
Director, Group Information Technology, Maxim's Group

Guan Lee
Director, Cloud and AI Infrastructure, APJC, Cisco

Lawrence Ang
Consulting Solution Architect | Strategy & Solutions, APJ, Everpure

Agenda 1

01 **The Maxim’s Journey: Navigating the Agentic AI Era**

02 **Accelerating the Future: Cisco Secure AI Factory and AI POD**

03 **Fueling the Engine: The Storage Powerhouse for Cisco AI POD**

Sharing from Louis Mah

Director, Group Information Technology,
Maxim’s Group

Section title placeholder

From Maxim’s Journey : Navigating the Agentic AI Era

Bridging the Legacy Gap: System Readiness and API Integration

Powering Real-Time Operations: Dynamic Inventory and Transactions

The Secure Hybrid Frontier: Cloud Engage and Enterprise Security

cisco

Cisco Secure AI Factory & AI POD

Guan Lee
Director, Cloud & AI
Cisco APJC

Challenges with AI projects delays time to value realization

Security vulnerabilities

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

Performance bottlenecks

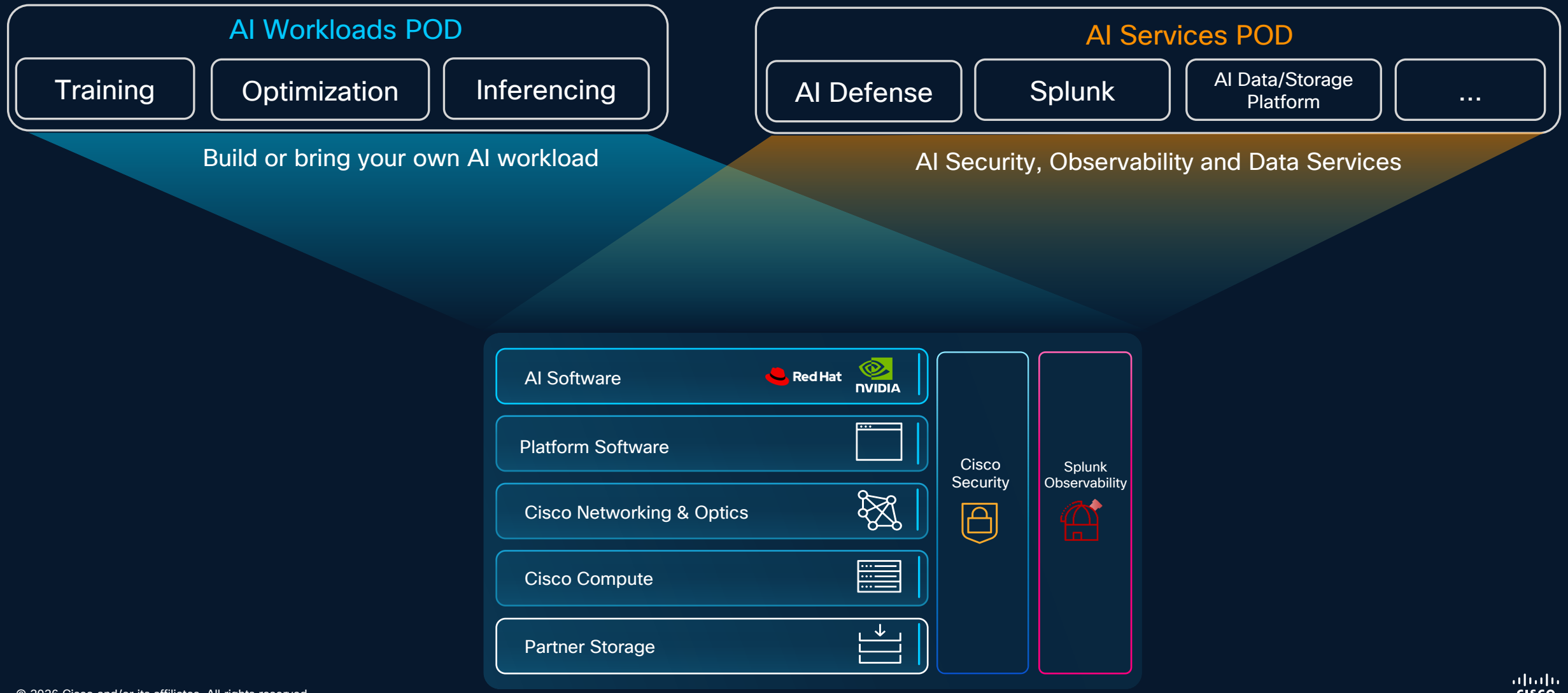
Model training and inferencing generates a lot of traffic, slow networks and delays time-to-value

Complex AI infrastructure deployment

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

cisco

Secure AI Factory Platform



Cisco AI PODs

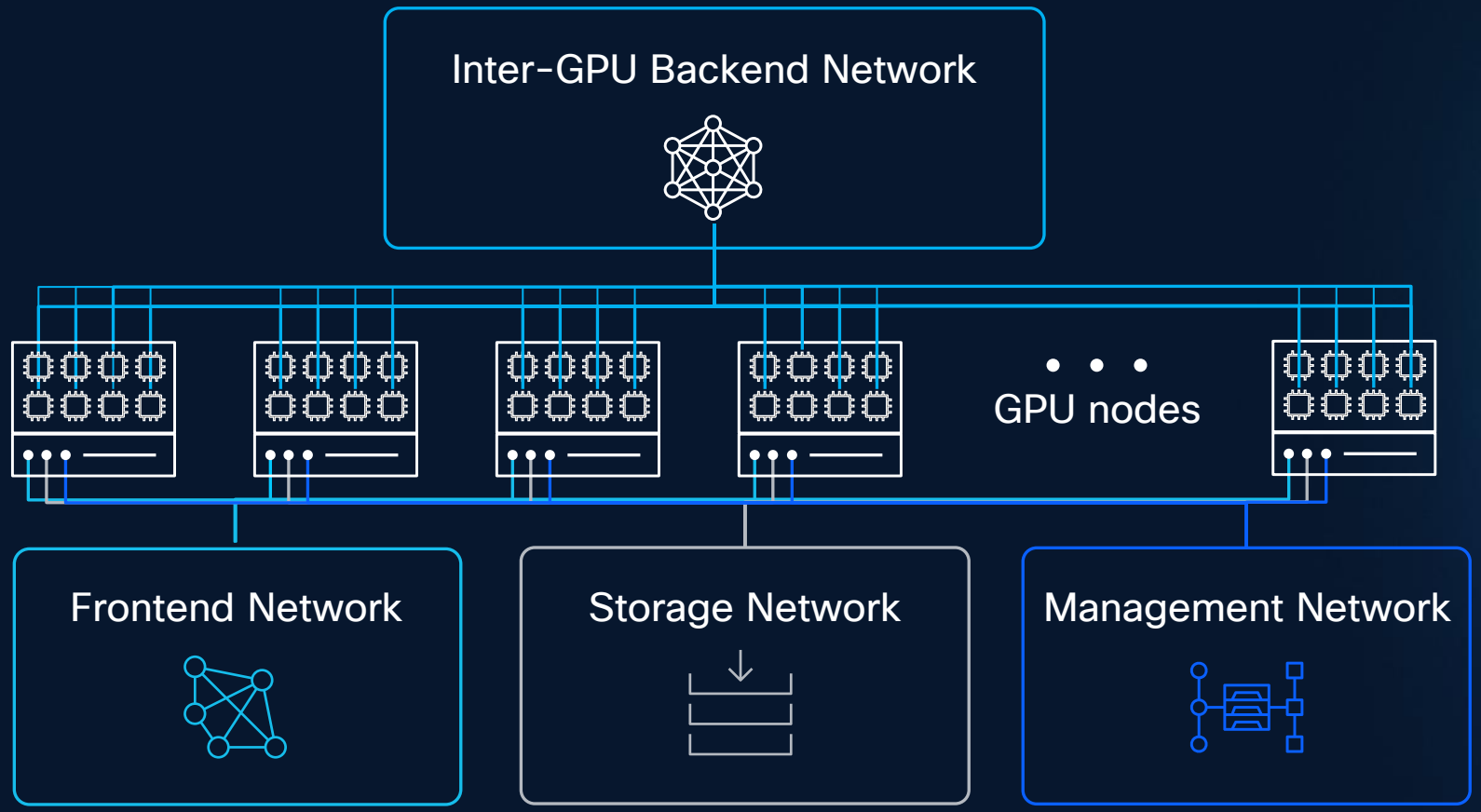
Included in Cisco Secure AI Factory with NVIDIA

Why Cisco AI PODs?

- Security-first architecture enables safe enterprise AI
- Unmatched performance AI infrastructure enables efficient model training, customization, and inferencing
- Pre-validated AI infrastructure stack for simplified deployment drastically reduces set-up time

cisco

AI Networking – The foundation for AI Performance



Cisco Nexus high density 800G & 400G fixed switches

Cisco AI PODs embedding the latest Nexus Switches

Nexus 9300 64-port 800G Switch

512-wide radix	Fully shared packet buffer	Advanced load balancing	Low Latency

N9364E-SG2-Q or N9364E-SG2-O

Compact 2RU 51.2T Switch

G200 Silicon One ASIC (5nm) | 100G SerDes | 256MB packet buffer

64 800G ports | Up to 128 line-rate 400G ports (2x400G breakout)

Choice of QSFP-DD800 or OSFP ports

Multi Core x86 CPU | 32GB RAM | 128GB SSD

Cisco NXOS spine and AI/ML spine/leaf capable

Ultra Ethernet Consortium MEMBER

Nexus 9332D-GX2B
32p 400G
8p MACsec/CloudSec

Nexus 9364D-GX2A
64p 400G
16p MACsec/CloudSec

ACI Leaf, ACI Spine, and NX-OS
25.6T, 19.2T, and 12.8T 400G switches
120MB smart buffer

Security
MACsec and CloudSec

Telemetry
FT, FTE, SSX, INT-XD

cisco

Compute AI portfolio

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

RTX PRO SERVER
RTX PRO 6000D Blackwell Server Edition GPUs

Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers

Cisco UCS blade (with GPU extensions) and rack servers

Enterprise AI edge

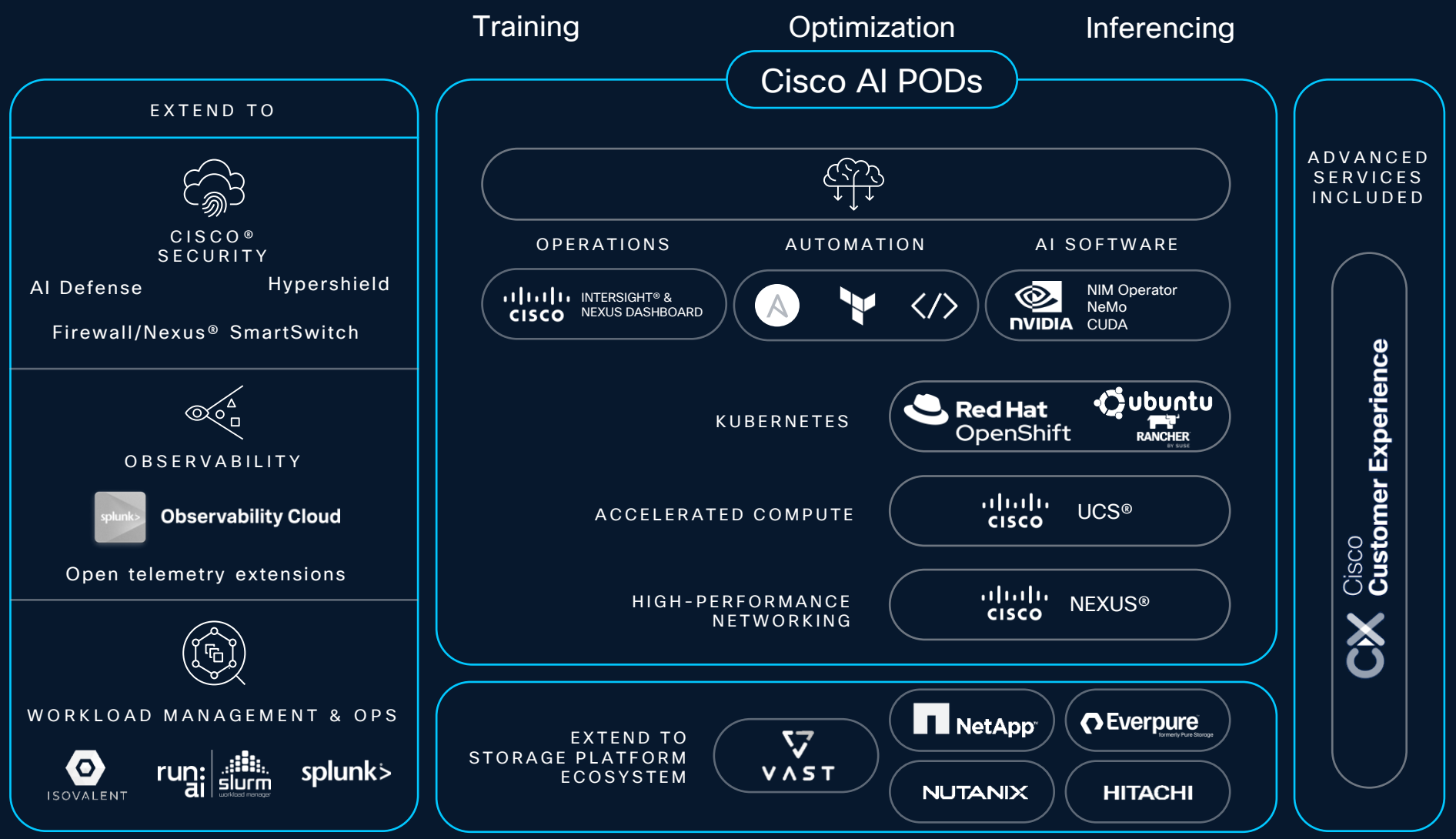
Dense compute for demanding AI

Full-stack AI with compute and networking

cisco

Cisco AI PODs

Introducing AI POD Integrated Offerings



BYO AI tools:

RAFAY
Kubeflow
jupyter
Apache Airflow
Weights & Biases
mlflow
neptune.ai
kedro
comet
ZenML
CLEARML
PREFECT
Flyte
mongoDB

FlashStack : Cisco+Everpure

Lawrence Ang
Consulting Solution Architect
Everpure

Challenges Bringing AI to Life

Lack of ROI

80% of AI projects fail to deliver intended business value—most often because success was never defined before launch.

RAND Corporation, 2025

Trust and governance

64% of leaders cite security and risk as the top barrier to scaling agentic AI—ahead of regulation or technology.

McKinsey AI Trust Maturity Survey, 2026

Data access and quality

50% of data leaders cite data quality and retrieval as their top agentic AI challenge.

Informatica CDO Insights, 2026

FlashStack for AI Factories

Cisco Secure AI Factory CVDs

NVIDIA AI Factory on FlashStack

(NVAIE | NIM | Blueprints | Cisco Validated Designs)

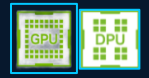
AI Software & Orchestration

NVDIA AI Enterprise | OpenShift AI | Portworx

Cisco UCS Accelerated Compute

C885A/C880A HGX System

C845A MGX System



x440

Cisco Nexus 9000 Networking | SAN | Ethernet | AI Fabric

Everpure Data Platform

FlashBlade//S200 | **FlashBlade//S500** | **Unified File + Object**

Evergreen//One for AI | SafeMode™ Snapshots | Pure1® Management

Unified Management & Automation | Cisco Intersight + Pure1® + A

AI Workload Profiles

INFERENCE

Deploy trained models for real-time predictions and responses

Use Cases

RAG | Chatbots | Agentic AI
Semantic Search | Video Analytics

NVIDIA GPUs:
RTX Pro 6000D
RTX Pro 4500

Compute

C845A
X440p nodes

Storage

FlashBlade//S200 (168–480 TB)

Performance Profile

Optimized for throughput
& cost-per-token

FINE-TUNING

Customize pre-trained models on proprietary enterprise data

Use Cases

Domain Adaptation | RLHF
LoRA / QLoRA | Instruction Tuning

NVIDIA GPUs:
RTX Pro 6000D
RTX Pro 4500

Compute

C885A/C880A
HBM3e high-memory GPUs

Storage

FlashBlade//S200–S500

Performance Profile

Balanced compute + data I/O
for iterative training loops

TRAINING

Build foundation models or full pre-training on large datasets

Use Cases

Foundation Model Training
Multi-node Distributed Training

NVIDIA GPUs:
B300
H200

Compute

C885A/C880A
Multi-node GPU clusters
NVLink interconnect

Storage

FlashBlade//S500 (960 TB+)

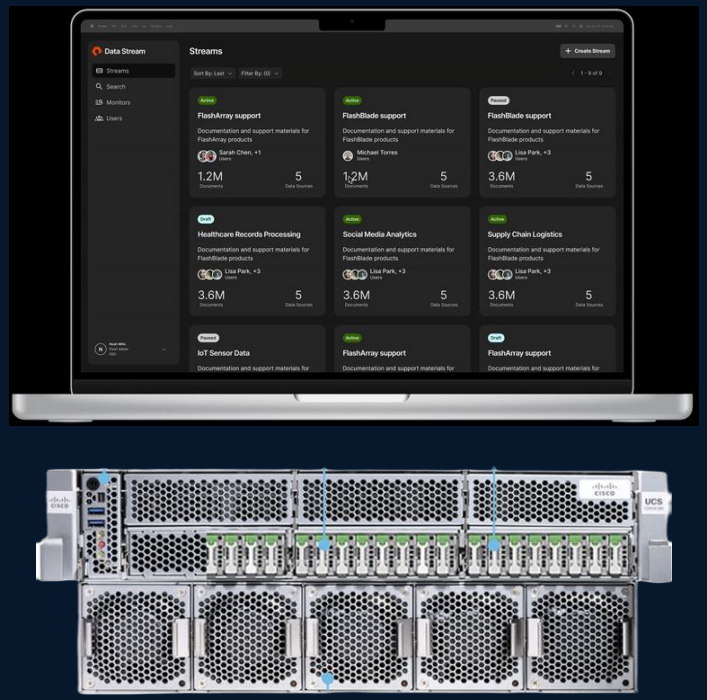
Performance Profile

Optimized for memory bandwidth
& GPU-to-GPU throughput

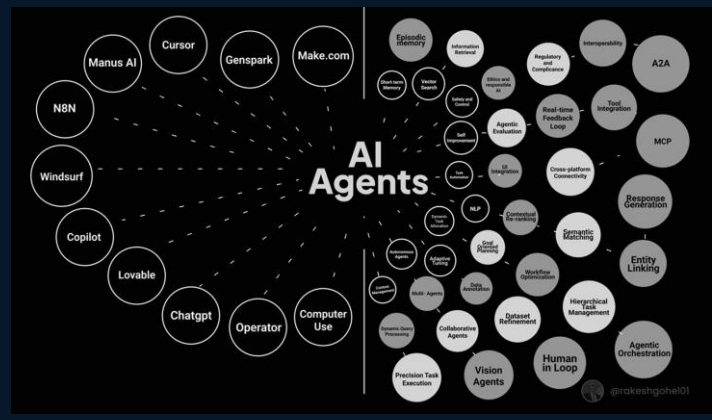
Data Stream Easy AI for the Enterprise

Raw Data

->



GPU-Accelerated Software



->

AI-Ready Data

Accelerate Time to ROI with AI-Ready Data

01 Findable

Semantic retrieval

Converted to vector embeddings that capture meaning — stored so AI can retrieve. Until this happens, your data doesn't exist from AI's perspective.

02 Trustworthy

Cleaned & classified

Duplicates, boilerplate, corrupted files, and sensitive content filtered out before they reach the model. Safety labels applied.

03 Governed

Access-controlled

Permissions match human access. An analyst reads a doc — that doesn't mean AI surfaces it to everyone. Without this, you lock everything or compromise compliance.

04 Traceable

Full provenance

You know where every answer came from, when the source was updated, and how it was processed. Non-negotiable in regulated industries.

Most RAG systems handle Findable. Everpure Data Stream with Data Intelligence handles all.

As-a-Service Consumption with Evergreen//One for AI

Think **Data**, not Capacity.
Consumption, not Acquisition.

Performance SLA
Auto-scaling or fixed throughput

Availability SLA
99.9999% uptime guarantee

Capacity SLA
25% buffer relative to usage

Energy SLA
Watts/TiB usage + buffer

Paid Power & Rack
Power consumption and rack space

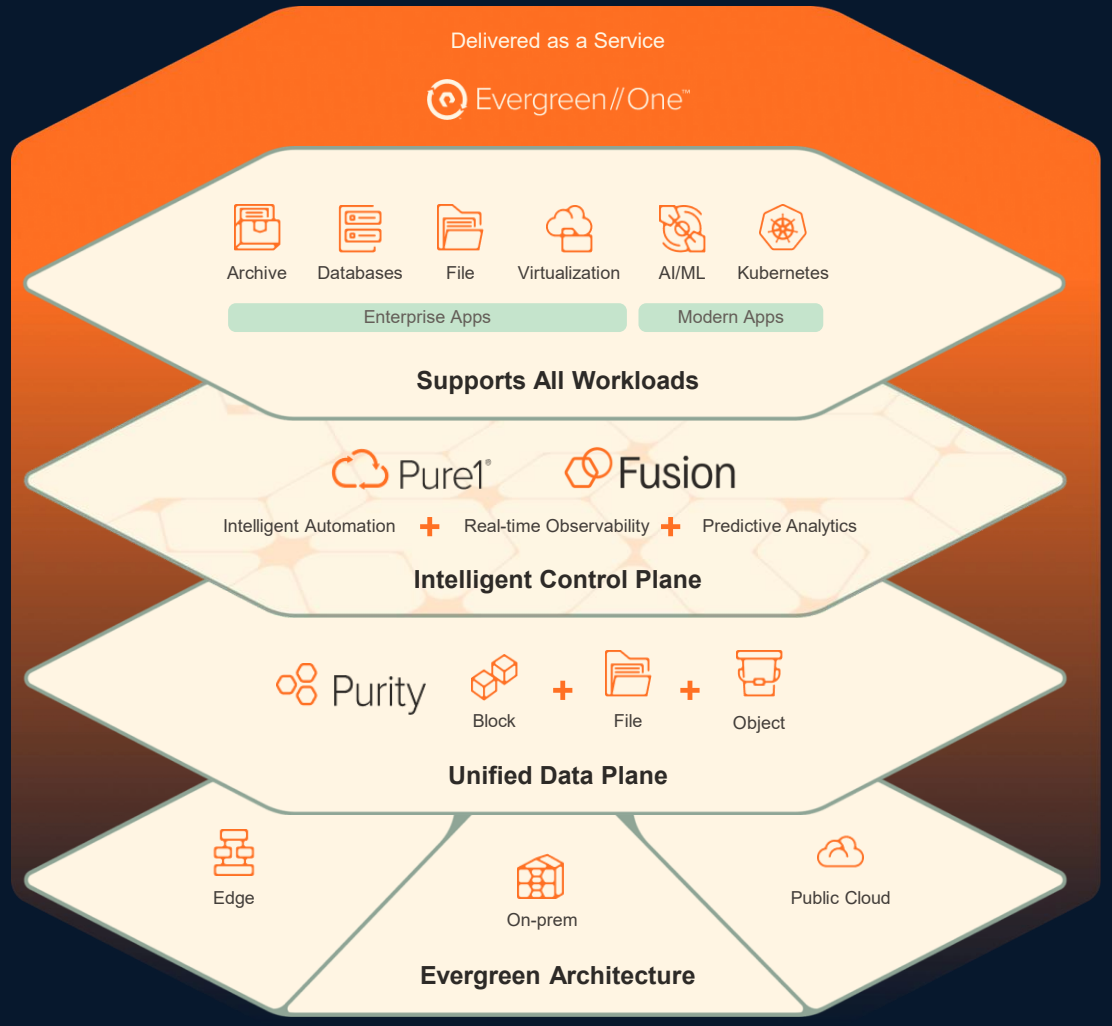
Zero planned downtime
for upgrades or maintenance

No data migration
No forklift upgrades

Zero Data Loss
Data durability

The Everpure Platform

The trusted leader in storage and data management.



- One data platform for all Enterprise AI workloads
- Accelerate ROI with autonomous operations
- Discover, Contextualize, Protect and Govern all data
- Innovation without disruption
- Simple, scalable business model

CISCO Connect

Thank you