From Idea to Impact

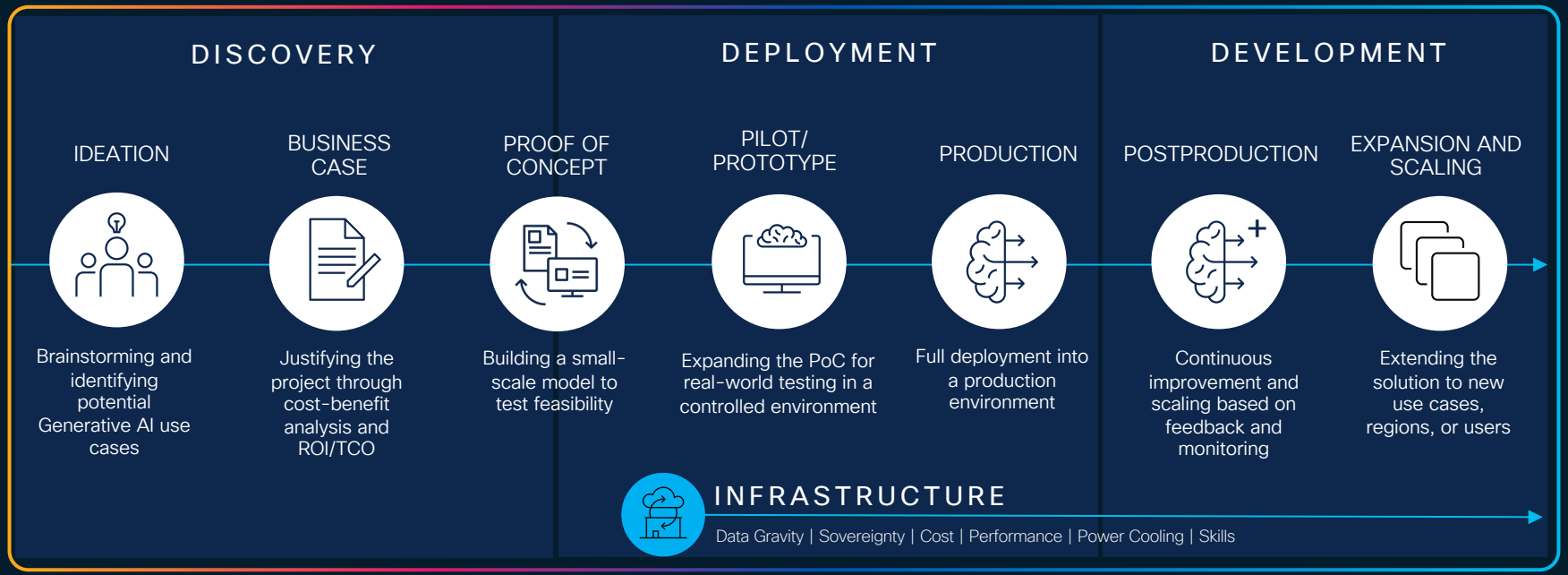
How to securely build your AI infrastructure from the ground up

Roger Dickinson
@DCgubbins

Legal Disclaimer

Any information provided in this document regarding future functionalities is for informational purposes inly and is subject to change including ceasing any further development of such functionality. Many of these future functionalities remain in varying stages of development and will be offered on a when-and-if available basis, and Cisco makes not commitment as to the final delivery of any such future functionalities. Cisco will have no liability for Cisco's failure to deliver any of all future functionalities and any such failure would not in any way imply the right to return any previously purchased Cisco products.

Typical AI Project Lifecycle



AI Design Considerations

Inference-led repatriation

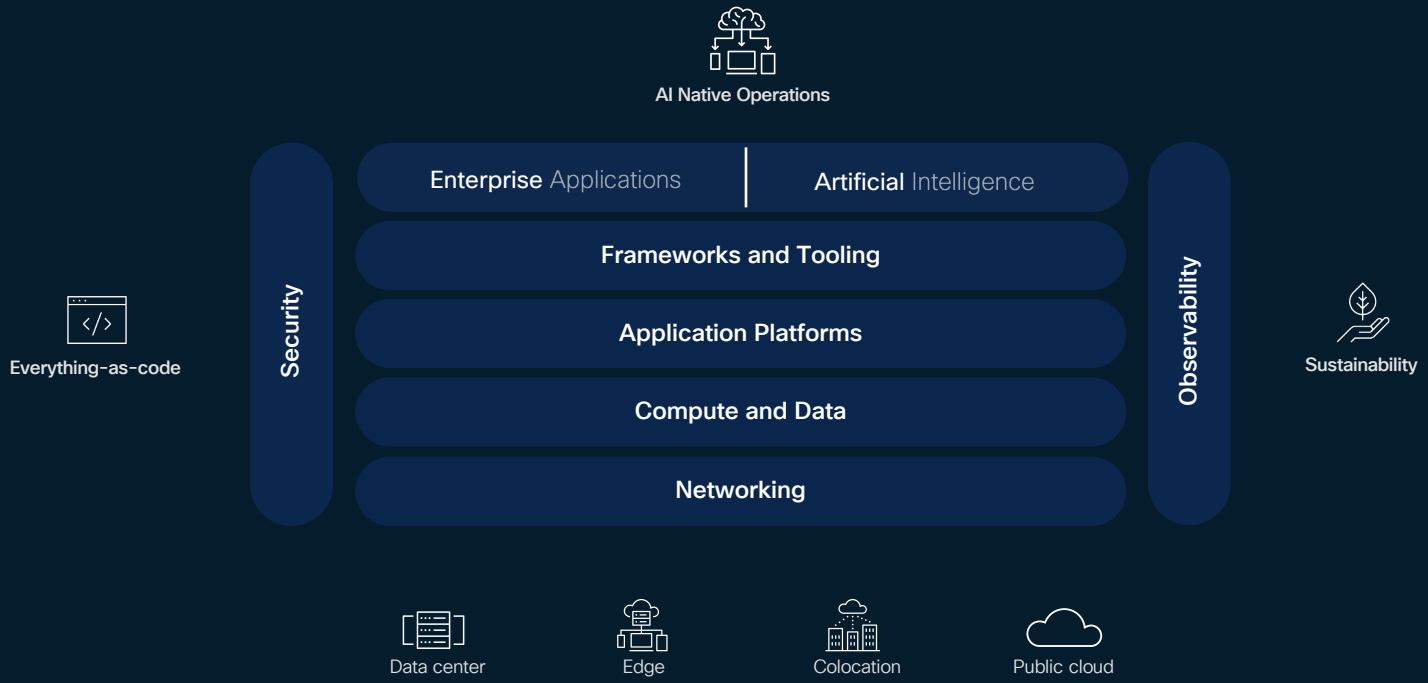
AI is network bound

Common substrate for AI security

Hyper-distributed security

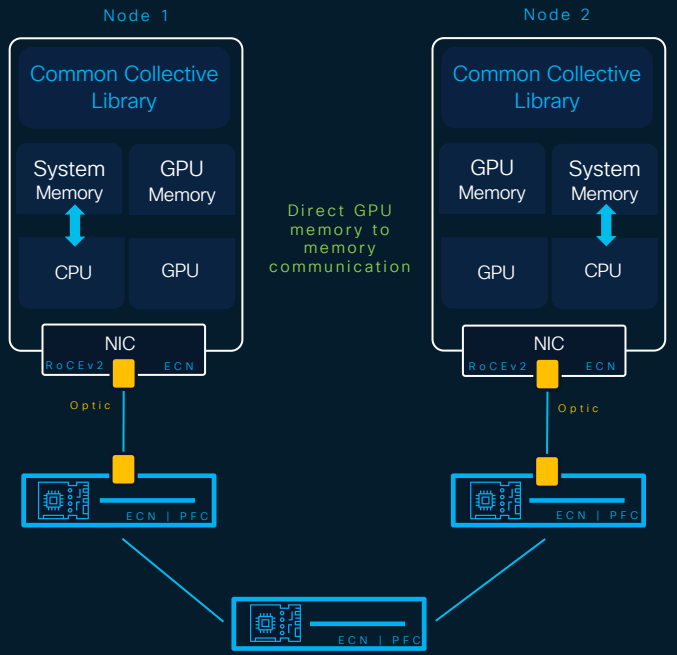
Customer choice in silicon

Inference-led repatriation



AI is network bound

Cisco AI Networking



Lossless Ethernet Fabric

- RDMA over converged Ethernet (RoCEv2)
- Common Collective Library (CCL)
- Data Center Quantized Congestion Notification (DCQCN)

Custom Silicon

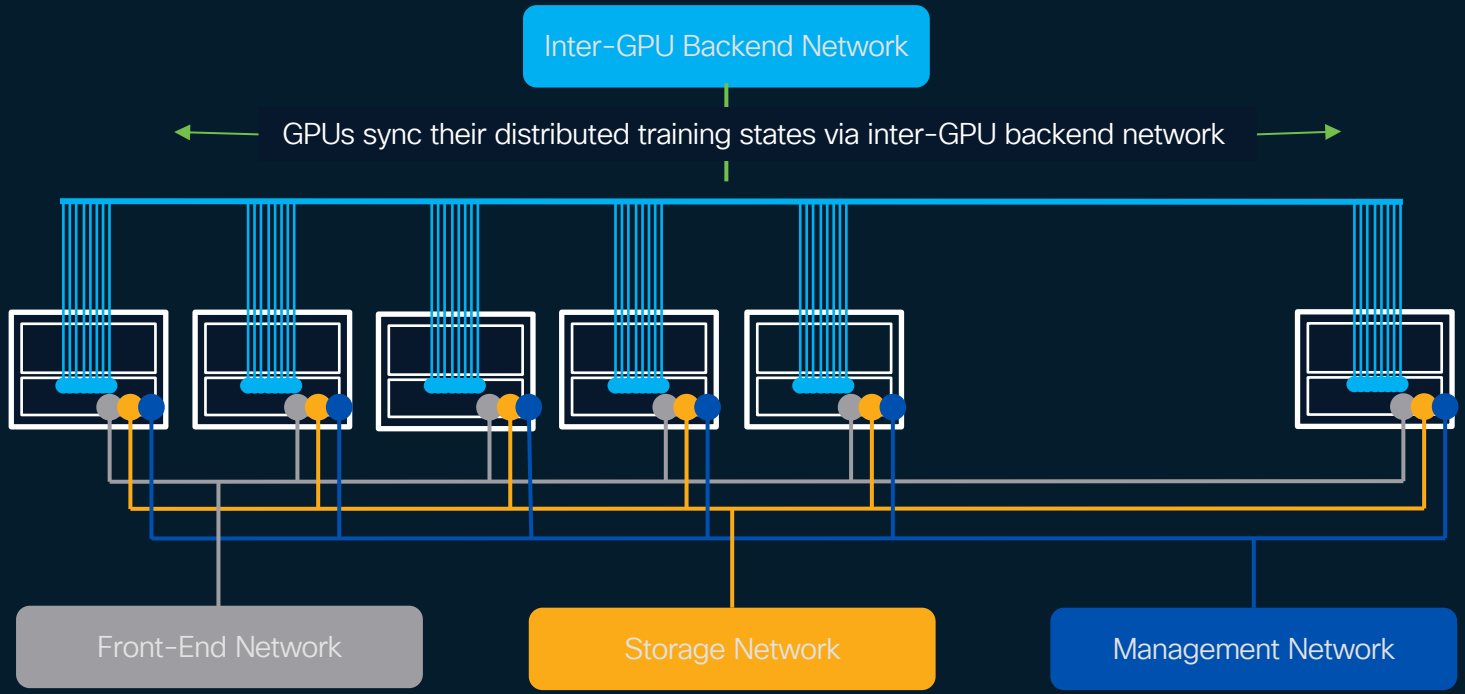
- Enhanced network utilization with telemetry
- Programmable SerDes for optimal AI performance
- Visibility and insights into energy consumption

Custom optics

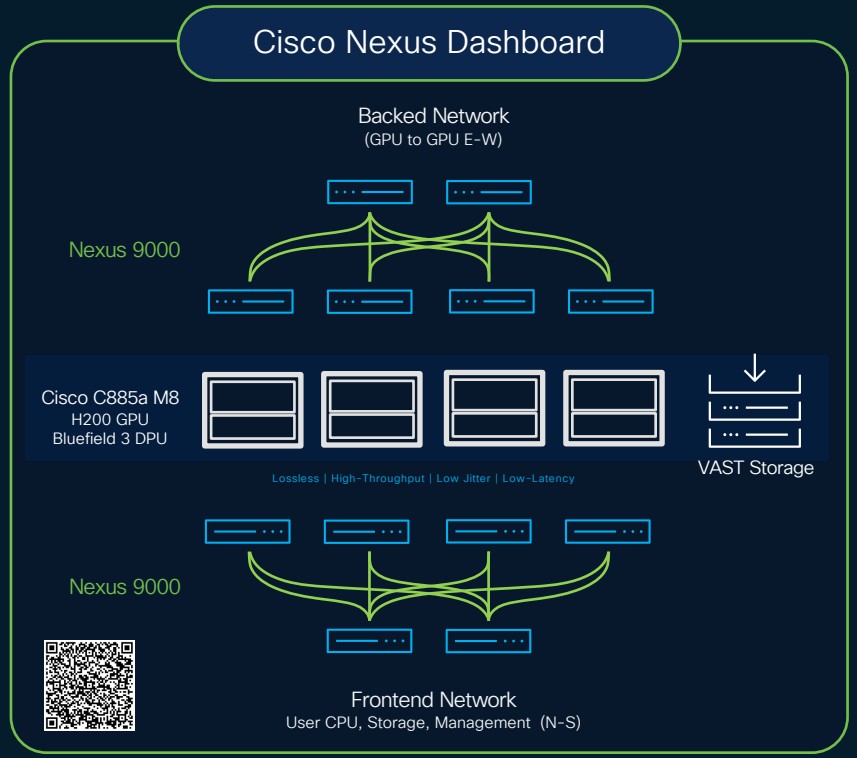
- Multivendor qualification and compliance
- Low failure rates with full TAC support
- Data Center Intconnect with coherent optics

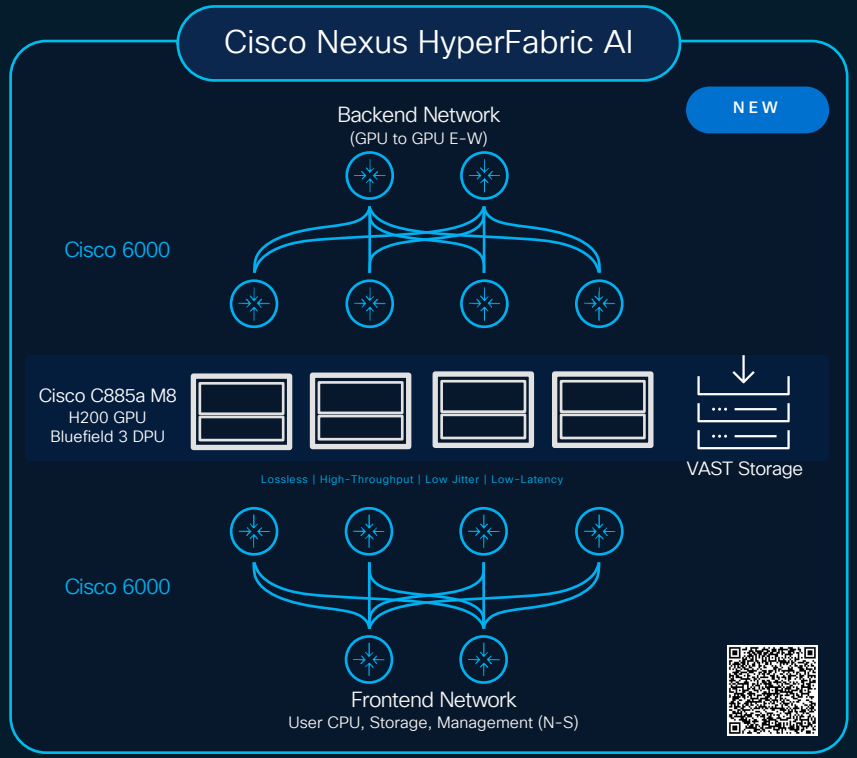
AI is network bound

Dedicated Backend and Frontend Networks



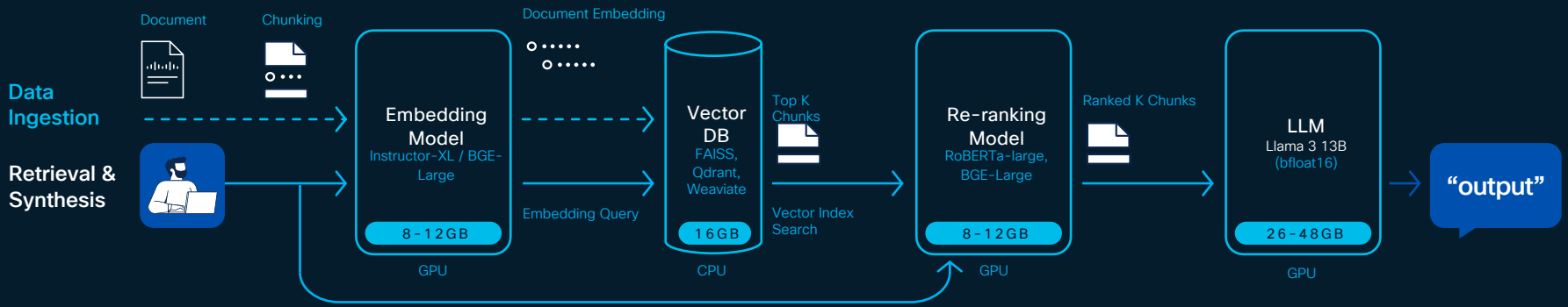
Cisco Data Center Networking for AI





A Simple Sizing Example

Retrieval Augmented Generation (RAG)



Total KV-cache size (bytes) = num_layers × num_heads × seq_len × head_dim × 2 × dtype_size

Compute AI Portfolio

Customer choice in silicon

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

UCS Dense GPU Servers

UCS Blade (w/GPU Expansion) and Rack Servers

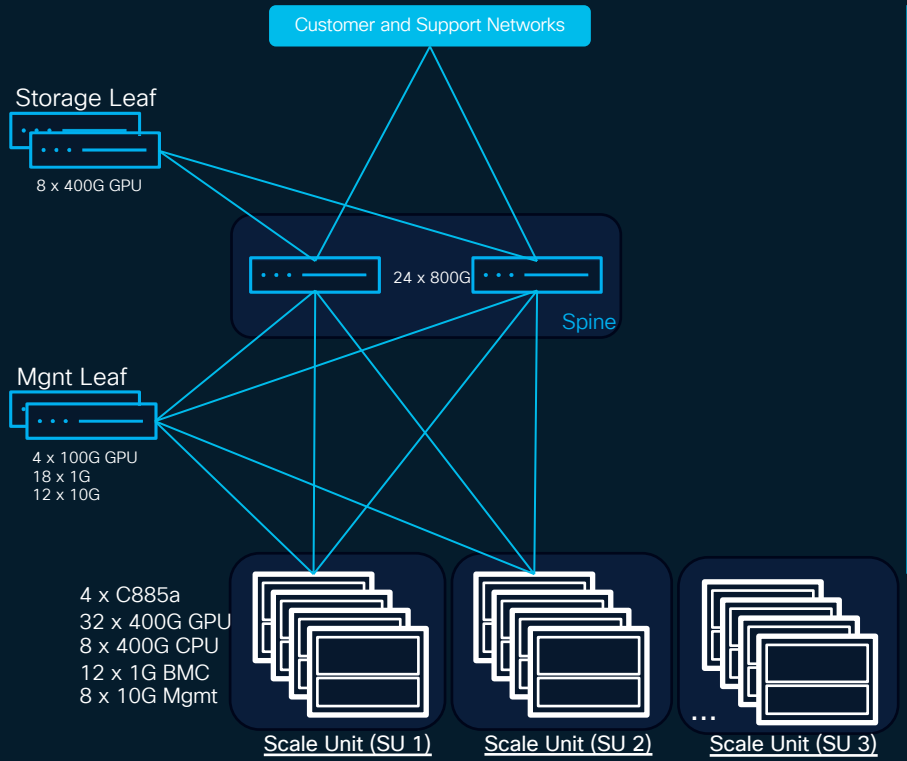
Enterprise AI Edge

Dense compute for demanding AI

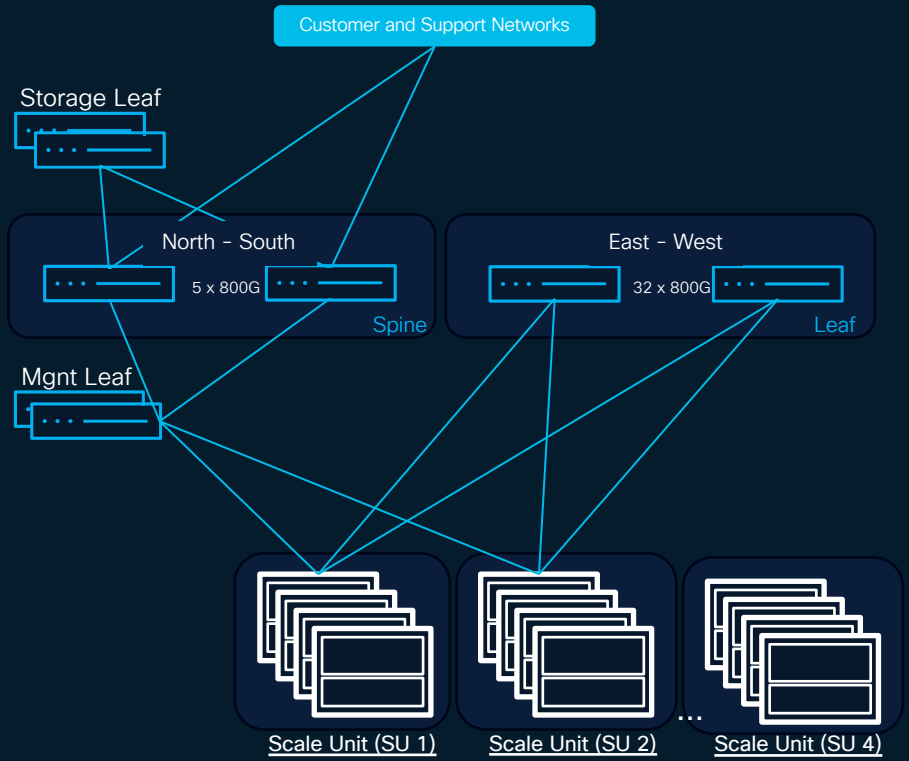
Full stack AI PODS with compute and networking

NVIDIA Enterprise Reference Architecture

Cluster Topology up to 12 Nodes



Cluster Topology up to 16 Nodes



AI Security requires a cohesive and integrated toolset

Data Source risks include data poisoning, exfiltration and accidental leaks.

Training risks include open-source model contamination, insecure APIs and unknowingly consuming 3rd party IP

Inference risks include denial of service attacks, prompt injection and data mining

AI-Native. Ever aware. Everywhere

Cisco Hypershield

Closing the Exploit Gap

In minutes versus months, AI-native rule engine prioritises vulnerability and then deploys surgical compensating controls.

Segmentation

An extended network that segments itself continuously adapting to current realities. Informed by process behaviours, file changes, learned policy preferences

Self Qualifying Updates

Powerful solution that allows you to validate upgrades and policy changes against live production traffic with our innovative dual dataplane approach.

AI-Native. Ever aware. Everywhere

Cisco Hypershield

Closing the Exploit Gap

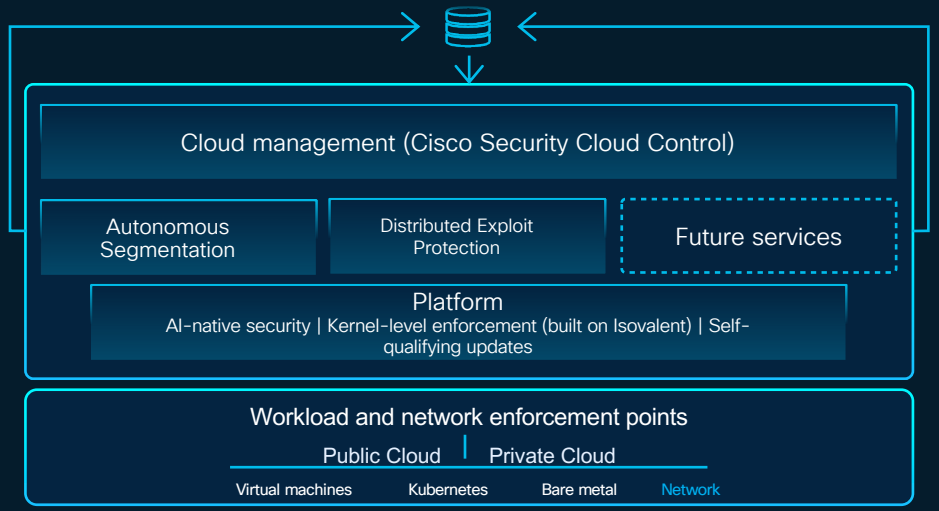
In minutes versus months, AI-native rule engine prioritises vulnerability and then deploys surgical compensating controls.

Segmentation

An extended network that segments itself continuously adapting to current realities. Informed by process behaviours, file changes, learned policy preferences

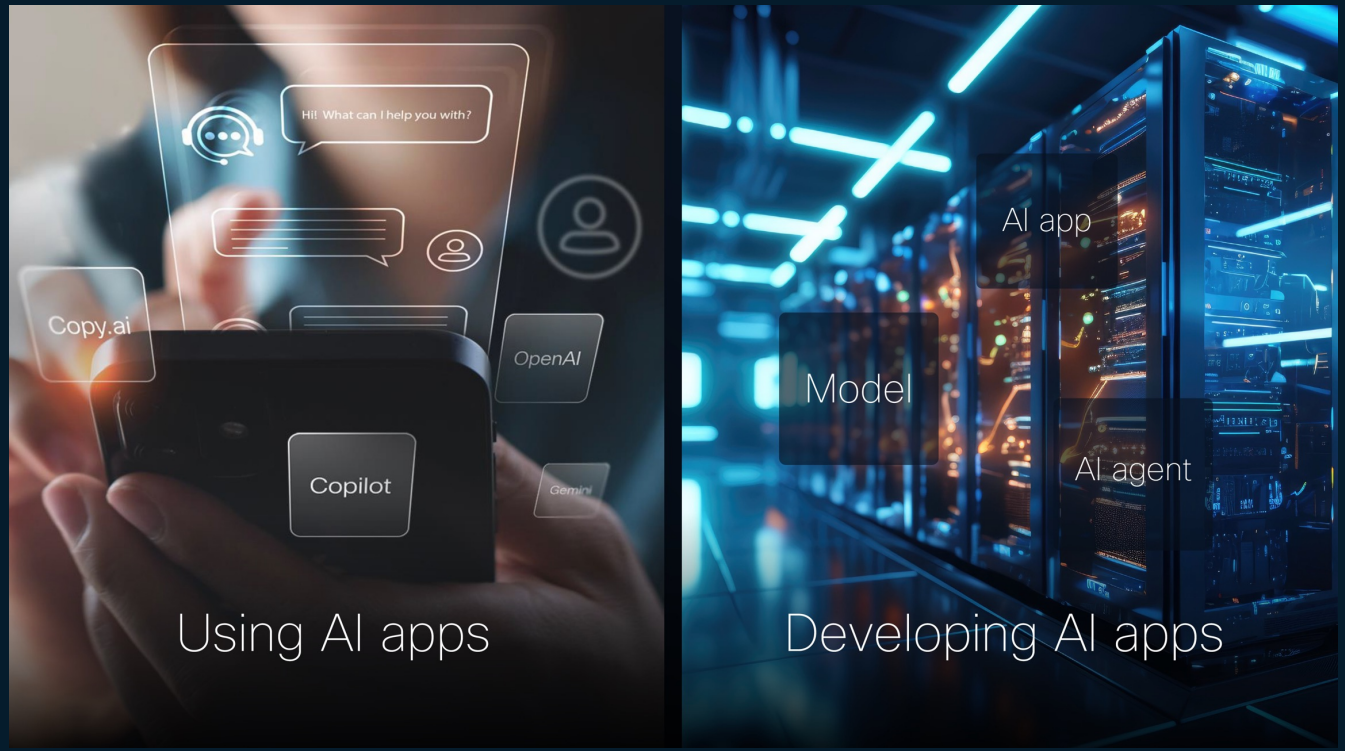
Self Qualifying Updates

Powerful solution that allows you to validate upgrades and policy changes against live production traffic with our innovative dual dataplane approach.

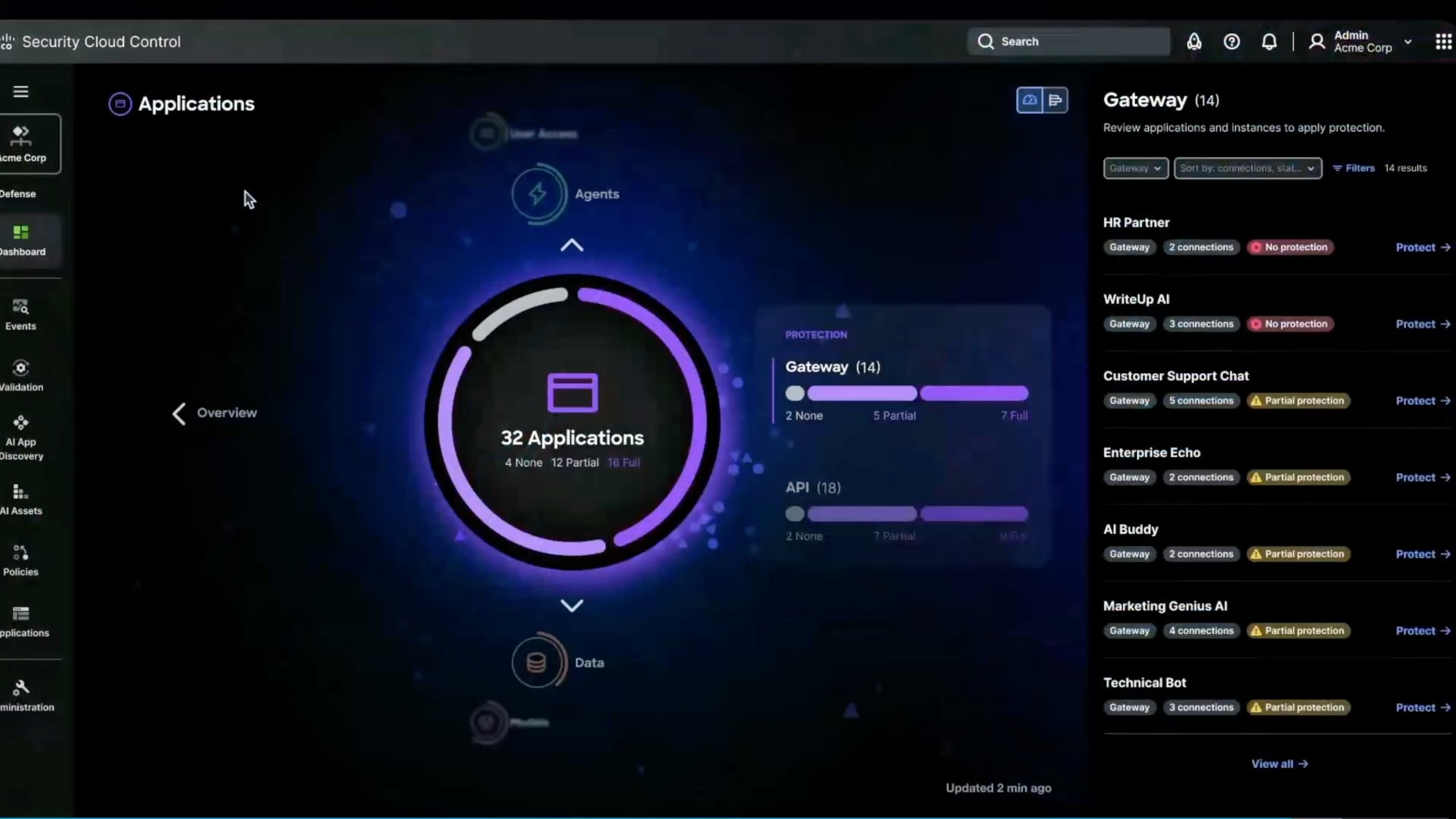


Hyper-distributed security

Cisco AI Defense

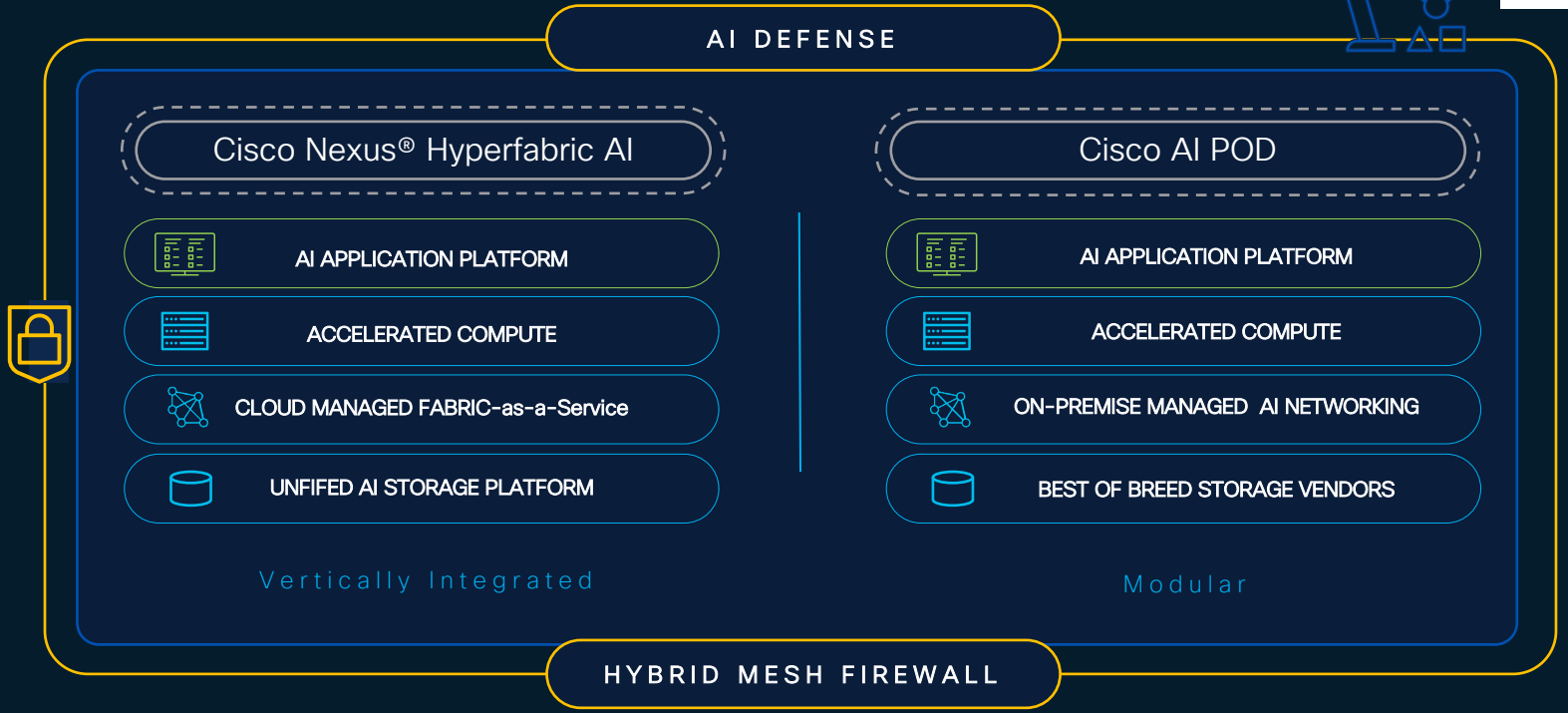


Common substrate for AI security


Security Cloud Control
Search
Admin
Acme Corp
Applications
Acme Corp
Defense
Dashboard
Events
Validation
AI App Discovery
AI Assets
Policies
Applications
Administration
Agents
Overview
32 Applications
4 None 12 Partial 16 Full
Data
PROTECTION
Gateway (14)
2 None 5 Partial 7 Full
API (18)
2 None 7 Partial
Updated 2 min ago
Gateway (14)
Review applications and instances to apply protection.
Gateway
Sort by: connections, stat...
Filters
14 results
HR Partner
Gateway 2 connections No protection Protect
WriteUp AI
Gateway 3 connections No protection Protect
Customer Support Chat
Gateway 5 connections Partial protection Protect
Enterprise Echo
Gateway 2 connections Partial protection Protect
AI Buddy
Gateway 2 connections Partial protection Protect
Marketing Genius AI
Gateway 4 connections Partial protection Protect
Technical Bot
Gateway 3 connections Partial protection Protect
View all


CISCO
NVIDIA
Bringing Secure AI to the Enterprise.

The solution: Cisco Secure AI Factory with NVIDIA



Cisco AI-ready data center

Cloud
Protected

Nexus 9000	Hyperfabric	Unified Computing System	AI Pod	Hyperfabric AI	Secure AI Factory
					
Data Center Networking	Cloud managed Network fabric	Cloud managed compute fabric	Validated hardware and software AI system	Cloud managed integrated AI System	Secure and integrated AI full stack system

AI READY INFRASTRUCTURE

AI FULL STACK SYSTEMS

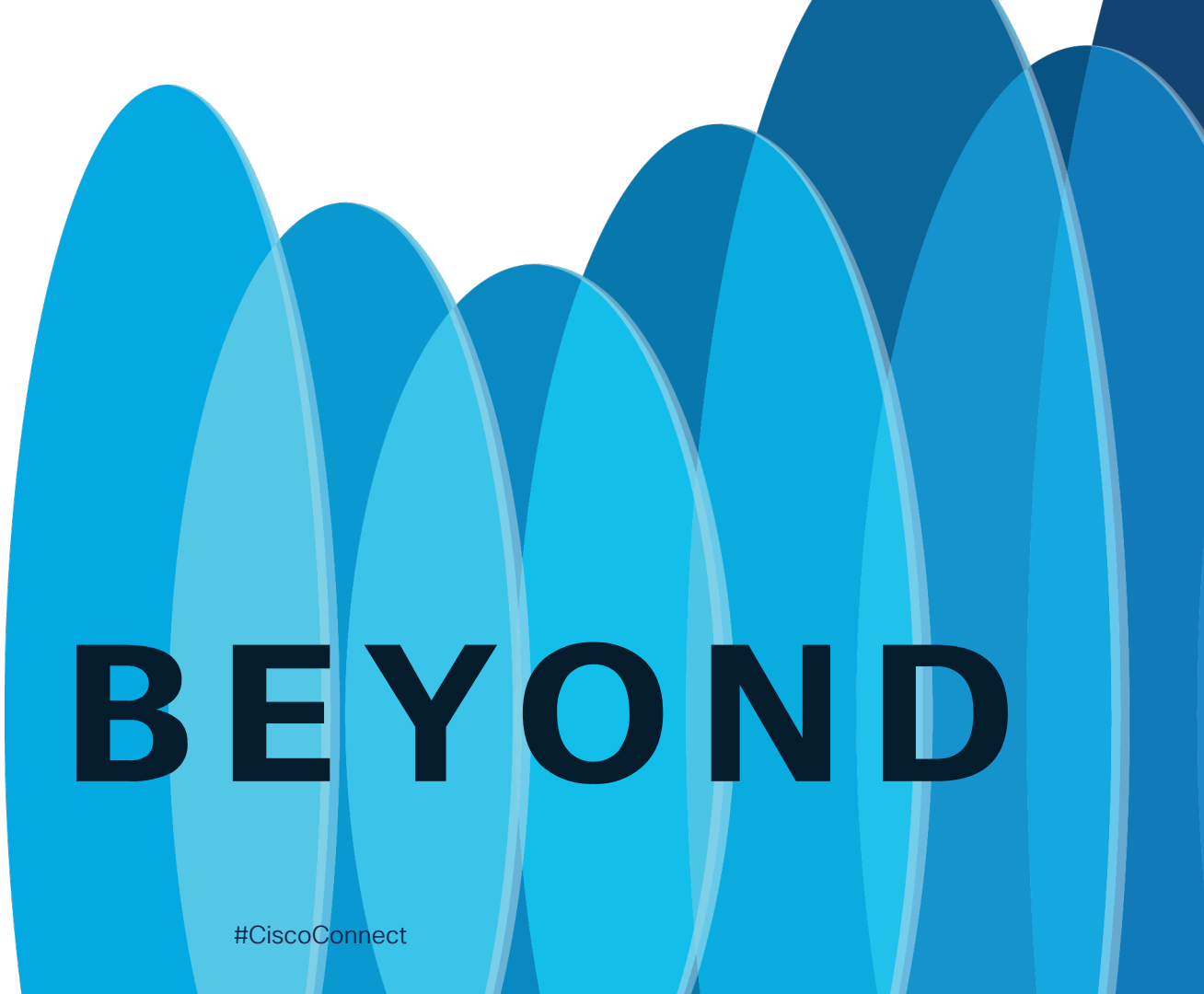
Digital
Resilience

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

Stay Connected

Thank you

cisco Connect



GO BEYOND

#CiscoConnect