Artificial Intelligence:
Was kommt und wie bereitet man sich vor?

Uwe Mueller

Director Cloud & AI Infrastructure, CAI

September 2024

Any Similiarities between Browser and AI ?

Quelle: ChatGPT,
https://www.mozilla.org/de/firefox/browsers/browser-
history/#:~:text=Der%20britische%20Informatiker%20Tim%20Berners,Zugang%20zum%20Internet%20„WorldWideWeb".

Cisco GSX

Cisco AI Readiness Index

Overall | Strategy | Infrastructure | Data | Governance | Talent | Culture

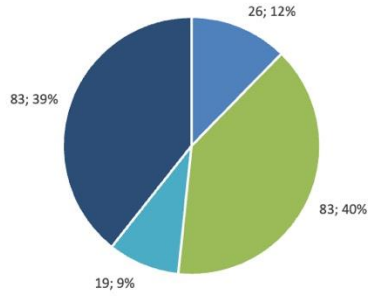| Europe ⌄ | Select an industry [GLOBAL] ⌄ | Germany ⌄ |

Global readiness

Pacesetters 14% Chasers 34% Followers 48% Laggards 4%

Europe

Pacesetters 8% Chasers 26% Followers 61% Laggards 5%

Germany

Pacesetters 7% Chasers 29% Followers 60% Laggards 4%

Cisco GSX

Habt Ihr konkrete Aktivitäten rund um Künstliche Intelligenz (KI)?

- 26; 12%
- 83; 39%
- 83; 40%
- 19; 9%

- ■ Implementierung von privater Infrastruktur dafür ist abgeschlossen/im Gange
- ■ Ist in die Planungen mit einbezogen
- ■ Machen wir in einer Public Cloud Infrastruktur
- ■ Nicht bekannt

**Planen Sie eine durchgehende (Security) Policy für Campus, Data Center und Cloud?**

- 41; 21%
- 62; 32%
- 30; 15%
- 9; 5%
- 54; 27%
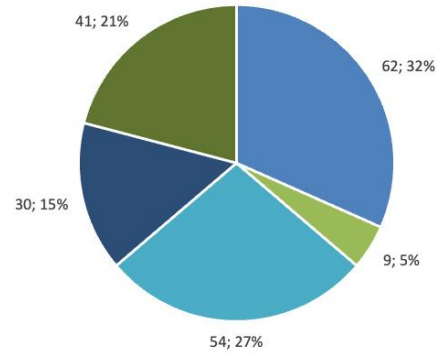
- ■ Ja, alles zusammen ■ Ja, für Campus & Cloud ■ Ja, für Campus & DC ■ Ja, für DC & Cloud ■ Nein

Welche Aspekte von KI nutzt Ihr bzw. wollt Ihr nutzen? (Mehrfach Auswahl möglich)

- 33; 14%
- 60; 25%
- 17; 7%
- 43; 18%
- 88; 36%
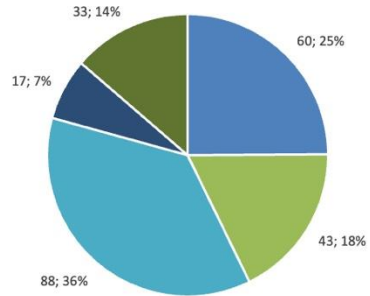
- ■ Fertige Modelle von der Stange anwenden (Inferencing)
- ■ Grundmodelle leicht anpassen und anwenden (Fine-Tuning + Inferencing)
- ■ Keine KI Anwendungen in Sicht
- ■ Wir trainieren eigene große Modelle über GPU Grenzen hinweg (LLM Training)
- ■ Wir trainieren eigene kleine Modelle (SML Training)

Wie in vergangenen Jahren möchten wir erneut wissen, wie Ihr mit dem Thema Automatisierung weitergekommen seid?

- 46; 23%
- 88; 45%
- 62; 32%
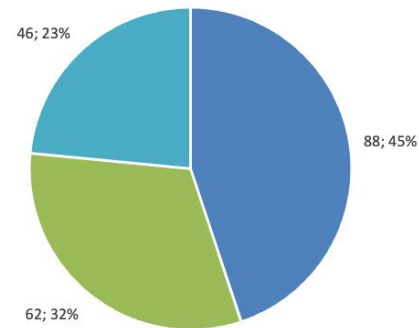
- ■ Implementierung läuft gerade ■ Ist aktiv im Einsatz ■ Nein

1. 79% are planning or implementing AI Infrastructure

2. 43% are Inferencing or Tuning

3. Security Solution: 32% for DC+ Campus+Cloud, 27% DC+Campus

4. 77% need more Automation

Source: Datacenter University 2024, Germany

Agenda

- AI: Hype or Real?

- Welche neue Fragen stellen sich?

- Cisco AI Ready Datacenter

- Wie kann Cisco unterstützen?

Technology Trends Driving Investment

GENERATIVE AI

80% adoption in 3-years[4]

Biggest disruption since cloud

APPLICATIONS

750M new workloads deployed[1]

Dense, and Diverse

AI-Ready
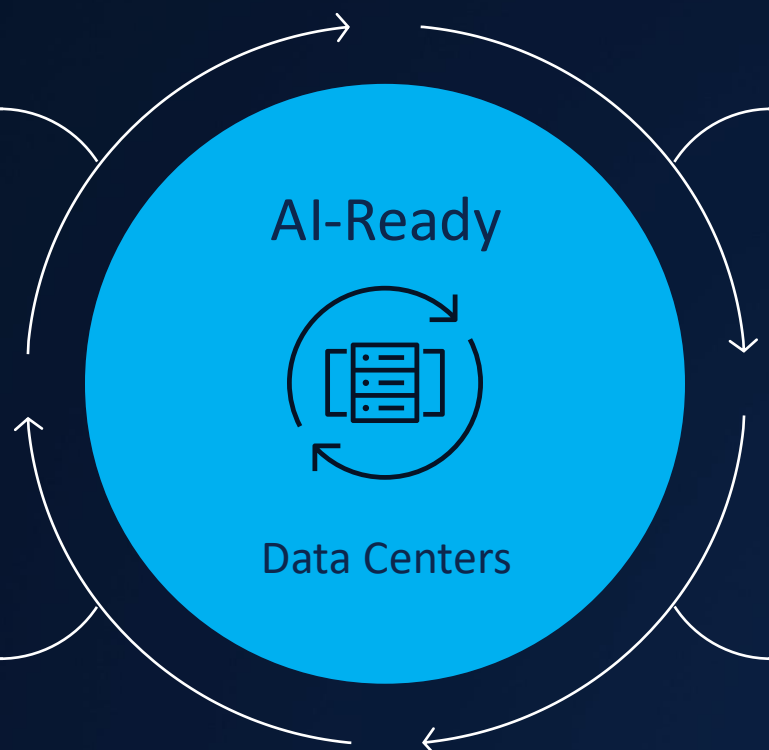
Data Centers

DIGITAL RESILIENCE

$336K cost of an hour of downtime[3]

Prevent and respond to disruption

HYBRID MULTICLOUD

49% of workloads on-prem[2]

Distributed and Optimized

Cisco GSX

1 Modernize infrastructure
Build modern, resilient infrastructure

2 Cybersecurity
Protect against the threats of today and tomorrow

3 AI and data
Harness the power of AI and data

Accelerate growth across AI, hyperscalers, rest of cloud

Hyperscalers[1]

	TAM CY27	TAM CAGR CY25-27
	$20.5B	**28%**
– AI[2]	$8.8B	18%

Technology leadership

- 20 use cases in production, 25% in AI
- 17 design wins, 65% in AI
- 12 use cases in pipeline

Selling models

- Fully integrated systems
- Grey box
- Optics
- Optical
- Silicon

Enterprise and rest of cloud[3]

	TAM CY27	TAM CAGR CY25-27
	$64.9B	**9%**
– AI[2]	$9.8B	29%

- HyperFabric for AI
- Digital Experience Assurance
- Secure Connectivity: private to public cloud

Selling models

- Orchestration & automation
- Fully integrated systems
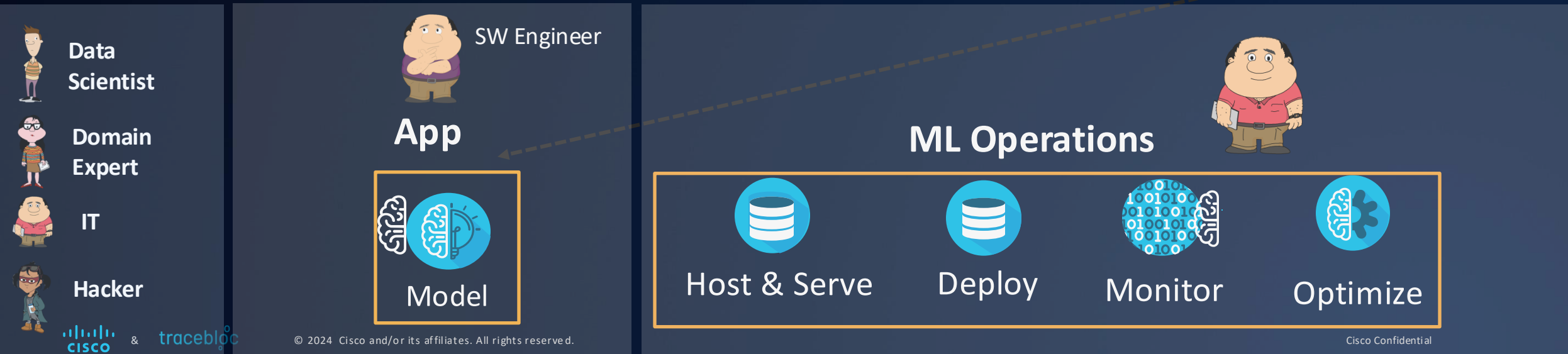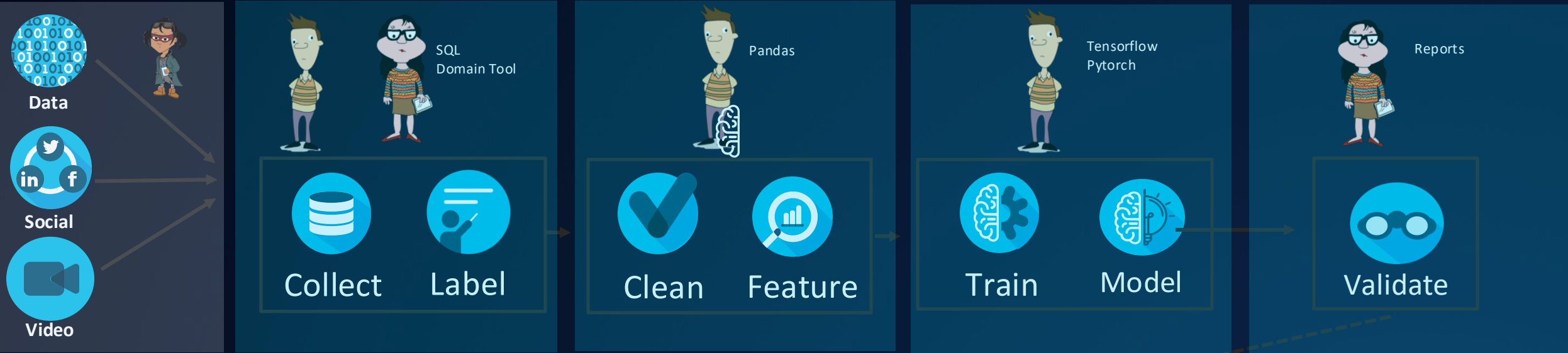- Compute (GPU and x86)
- Optics
- Optical

1. Includes DC Switching (Managed, Fixed-Unmanaged, Infiniband) Silicon & Optics. Does not include Server, Storage, Software and Services
2. Includes AI DC Switching (Mgd + Infiniband) and Direct Purchase Ethernet Optics
3. Includes Enterprise Data Center and second-tier cloud providers. Numbers include Mgd DC Switching , x86 Blade Servers, Silicon & Optics, Cloud related M&O S/W, Compute aaS

Welche neue Fragen stellen sich?

Es geht darum, Daten nutzbar zu machen, Zusammenhänge zu erkennen, neue Informationen zu generieren, um damit intelligente Automatismen für Menschen in positiver Weise zur Verfügung zu stellen.

AI brings New Roles and Responsibilities

Data
Social
Video

SQL
Domain Tool

Collect **Label**

Pandas

Clean **Feature**

Tensorflow
Pytorch

Train **Model**

Reports

Validate

Data Scientist

Domain Expert

IT

Hacker

SW Engineer

App

Model

ML Operations

Host & Serve   Deploy   Monitor   Optimize

Which Model is the right one ?
tracebloc enables AI partners to fine-tune their models on companies' data

AI Consulting

Startups

Researcher

Supplier

custom AI model development and fine-tuning

CISCO

turnkey, on-prem AI server

AI Partners
train models on proprietary company data without accessing it

tracebloc

orchestrates AI trainings and keeps data save

Companies
remain in full ownership of their data

Cisco GSX

Which Infrastructure for which model?
Inferencing, Tuning, or Training

MLOps App Integration

API App Integration

AI Projektteam

Eigener Kontext

Eigene Daten

API App Integration

(S-M) AI Infrastruktur

(M-XXL) AI Infrastruktur

Konsumption
(**Inferencing**)

Tuning
(Data Center Inferencing)

Eigenentwicklung
(Training)

Model exist's

Model tuned

Model trained

Cisco GSX
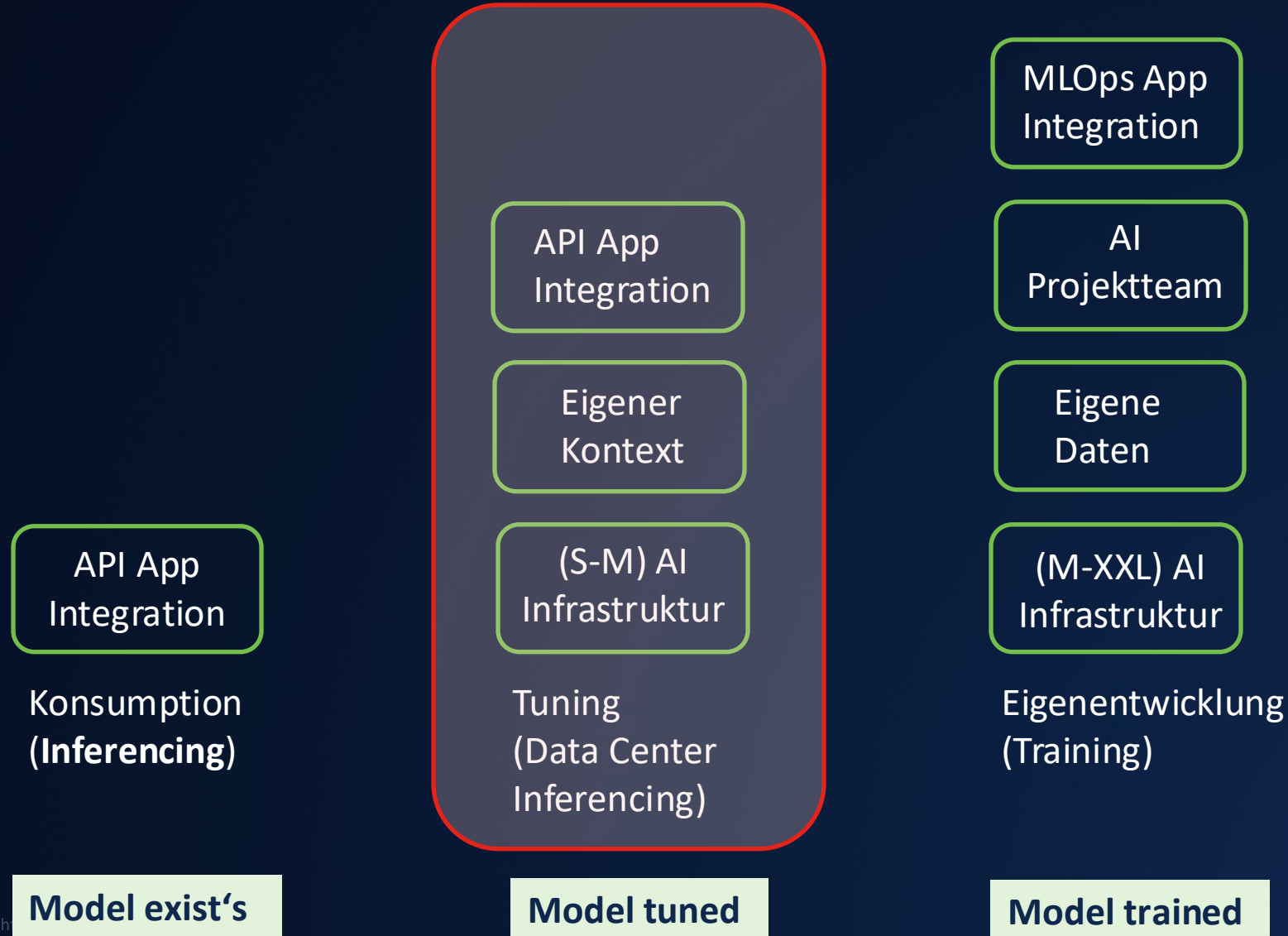
Which GPU is the right one for AI?

Power Consumption

AI use-case	Data Ingest and Preparation	Edge Inferencing	Data Center Inferencing	Fine Tuning	Large Foundational Training

General Purpose Compute (CPU) ← → AI Acceleration (GPU)

Compute matrix	CPUs	0-2 GPU	1-4 GPU	4-64 GPU	64-10K+

Network Considerations	Shared Fabric				Dedicated Fabrics

Cisco GSX
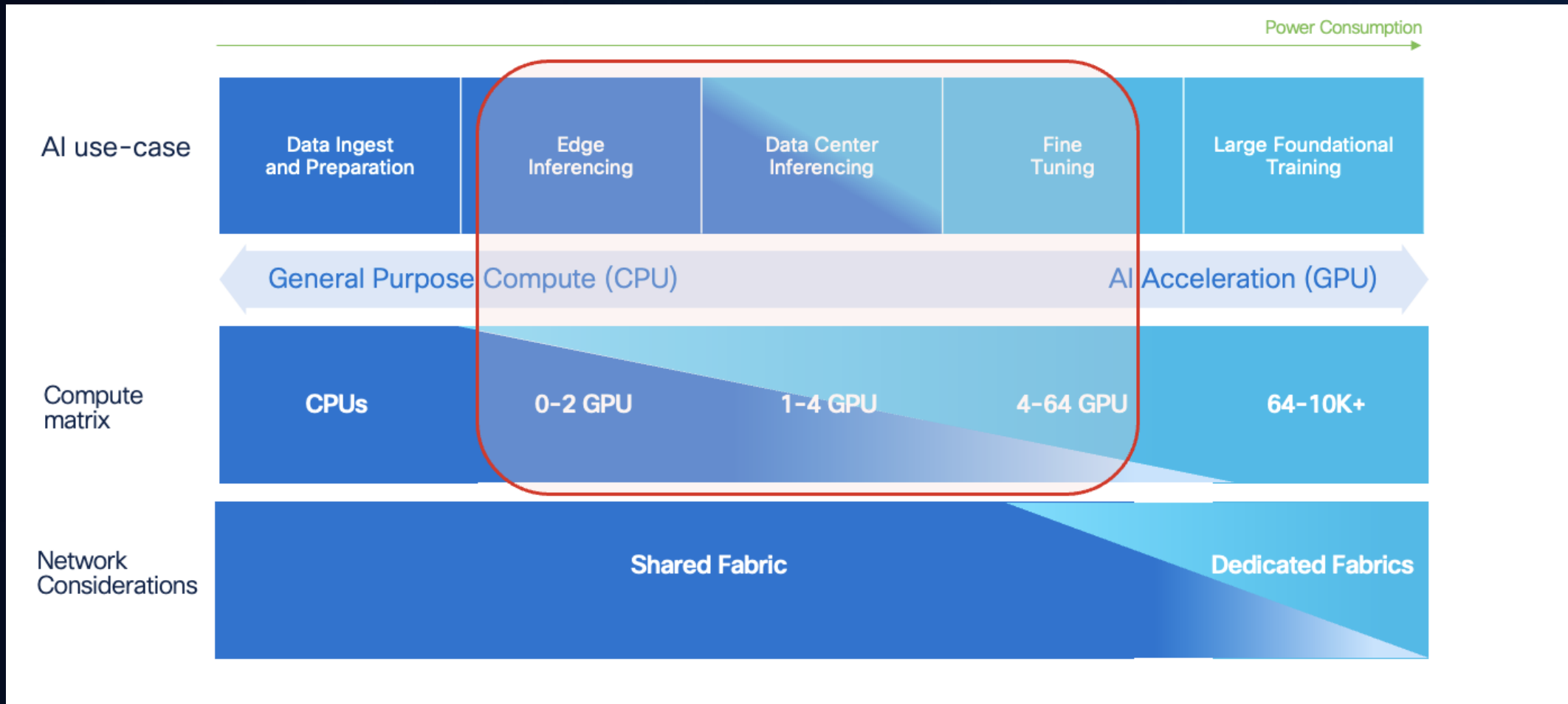
Cisco AI Ready Datacenter

AI-Ready Data Centers

Transform data centers to power AI workloads anywhere

Distributed AI Infrastructure

Full Stack Systems

AI Infrastructure Operations

Cloud Protection

Digital Resilience: security, reliability, and performance

Cisco GSX

Data Center Infrastructure

For Enterprise Applications

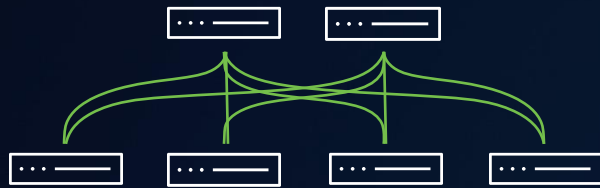Infrastructure for multicloud supporting a diverse and distributed application landscape

For Generative AI

Accelerated Data Center Networking, Compute and storage for AI workloads

Cisco GSX
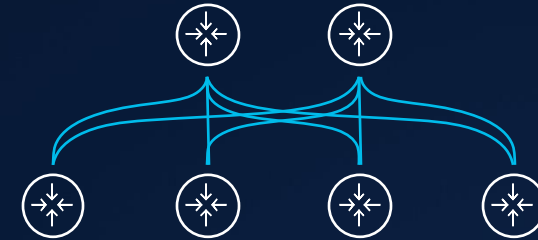
Cisco Data Center Networking

Cisco Nexus Dashboard

Private cloud managed, flexible solution

General purpose data center solution

Greenfield and brownfield deployments

Any size spine-leaf data center (NX/ACI)

Nexus Dashboard for simplified operations

CloudScale and Silicon One based switches
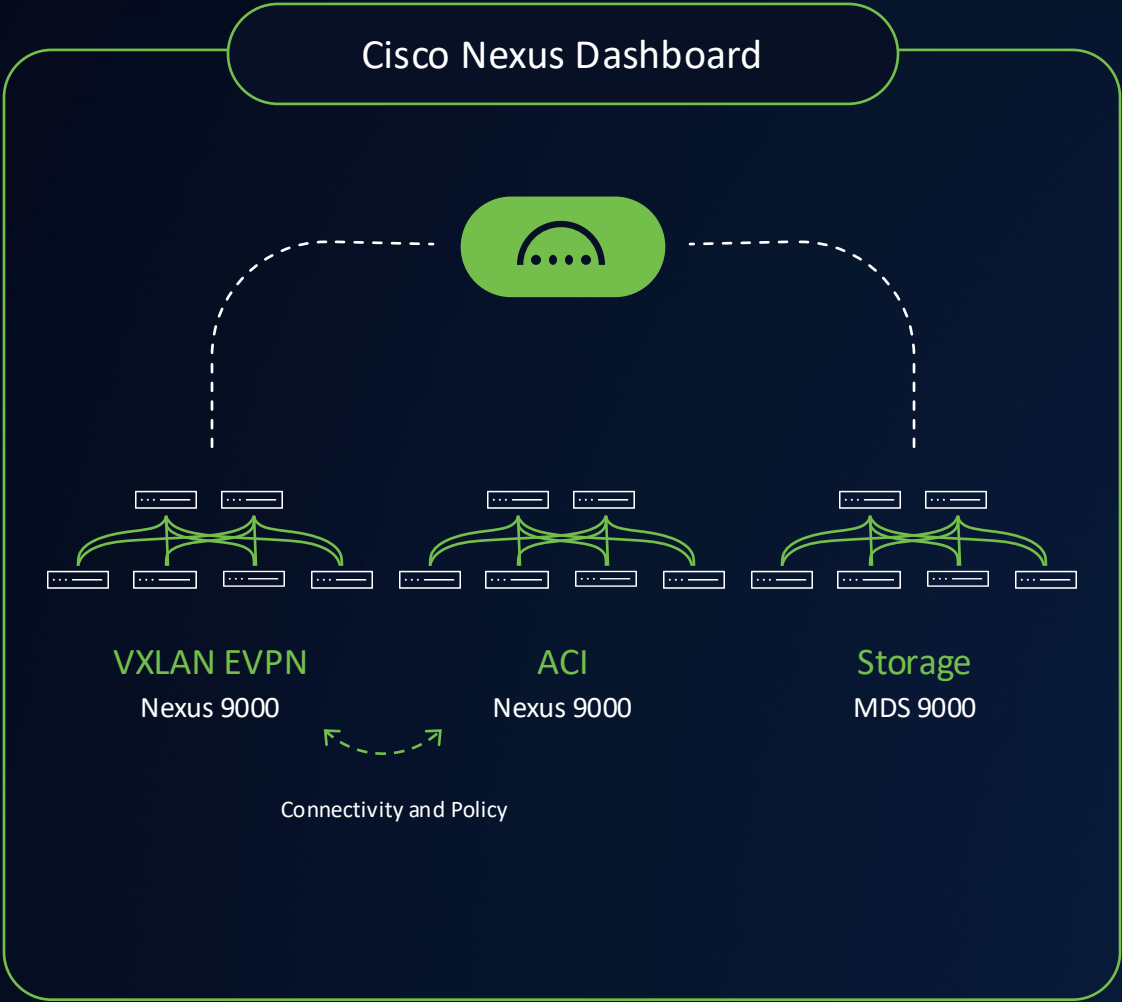
* Shipping

Cisco Nexus Hyperfabric

Cisco cloud-managed, fully integrated

Easily design, order, deploy, monitor and upgrade fabrics

Purpose built vertical stack

Greenfield deployments only

Cisco 6000 series switches (Silicon One)

Cisco GSX

Nexus Network Innovation

Cisco Nexus Dashboard

VXLAN EVPN
Nexus 9000

ACI
Nexus 9000

Storage
MDS 9000

Connectivity and Policy

Nexus ONE

Evolve fabric architecture for consistent user experience using open networking

Standards Based

Configure and connect fabrics using open networking protocols

Cisco GSX

Simplified Orderability

AI PODs

Faster time to value with pre-configured bundles

ORDERABLE THIS QUARTER

Deploy AI with confidence

Fully supported stack including Cisco and 3rd party components

Orderable, validated AI-Ready infrastructure stacks

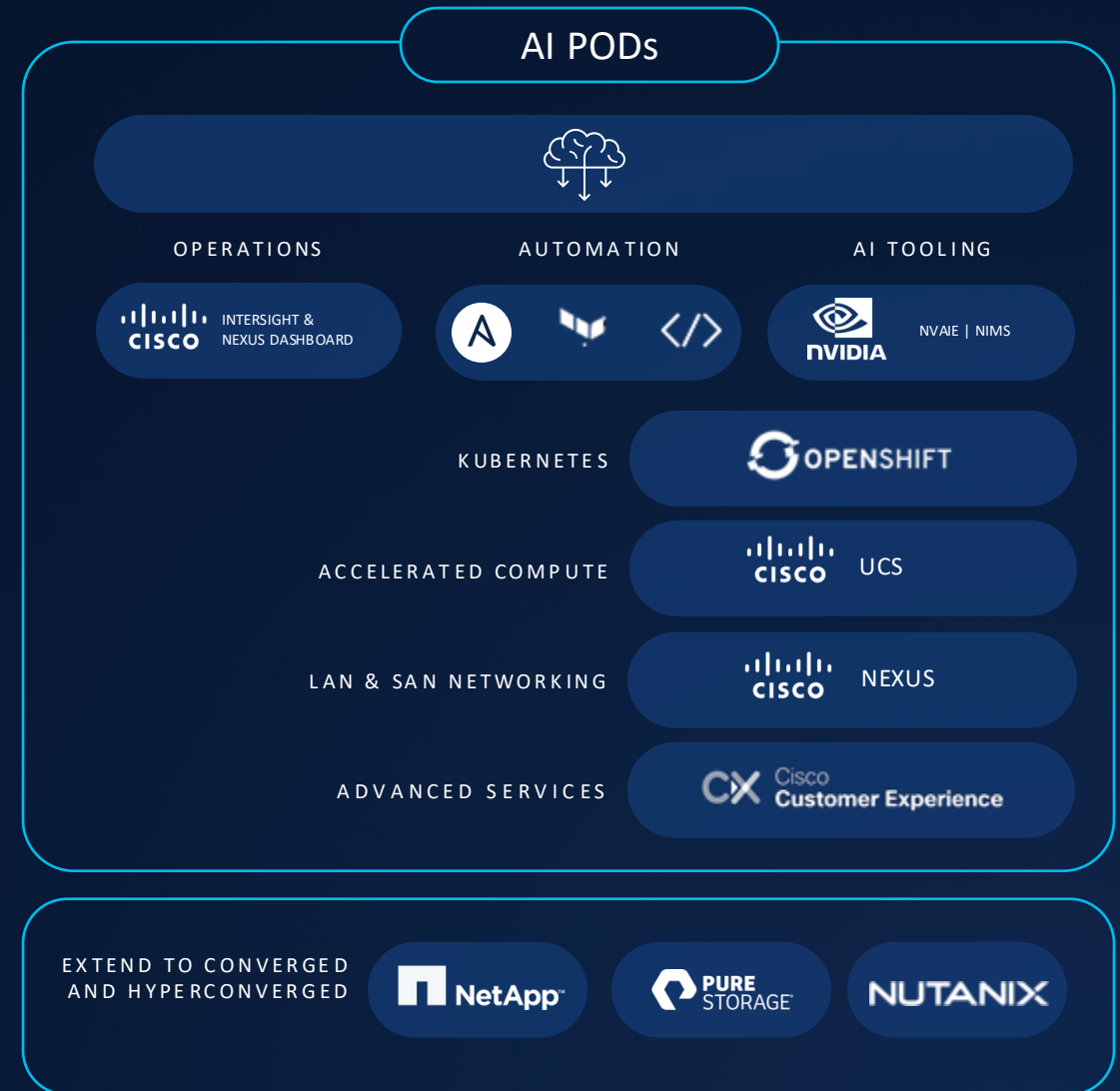**AI Advisor tool for configuration guidance**

COMING SOON

AI PODs

	OPERATIONS	AUTOMATION	AI TOOLING
	CISCO INTERSIGHT & NEXUS DASHBOARD	A, </>	nVIDIA NVAIE | NIMS

KUBERNETES	OPENSHIFT
ACCELERATED COMPUTE	CISCO UCS
LAN & SAN NETWORKING	CISCO NEXUS
ADVANCED SERVICES	CX Cisco Customer Experience

| EXTEND TO CONVERGED AND HYPERCONVERGED | NetApp | PURE STORAGE | NUTANIX |

Cisco GSX

How Cisco can help to build your AI ready Datacenter?

Cisco AI Infrastructure
(for the Generative AI era)

| Build the Model \| Training | Optimize the Model \| Fine-tuning & RAG | Use the Model \| Inferencing | Model the World \| Digital Twin |

Data Center Networking Blueprint for AI
CVD Available NOW

AI PODs
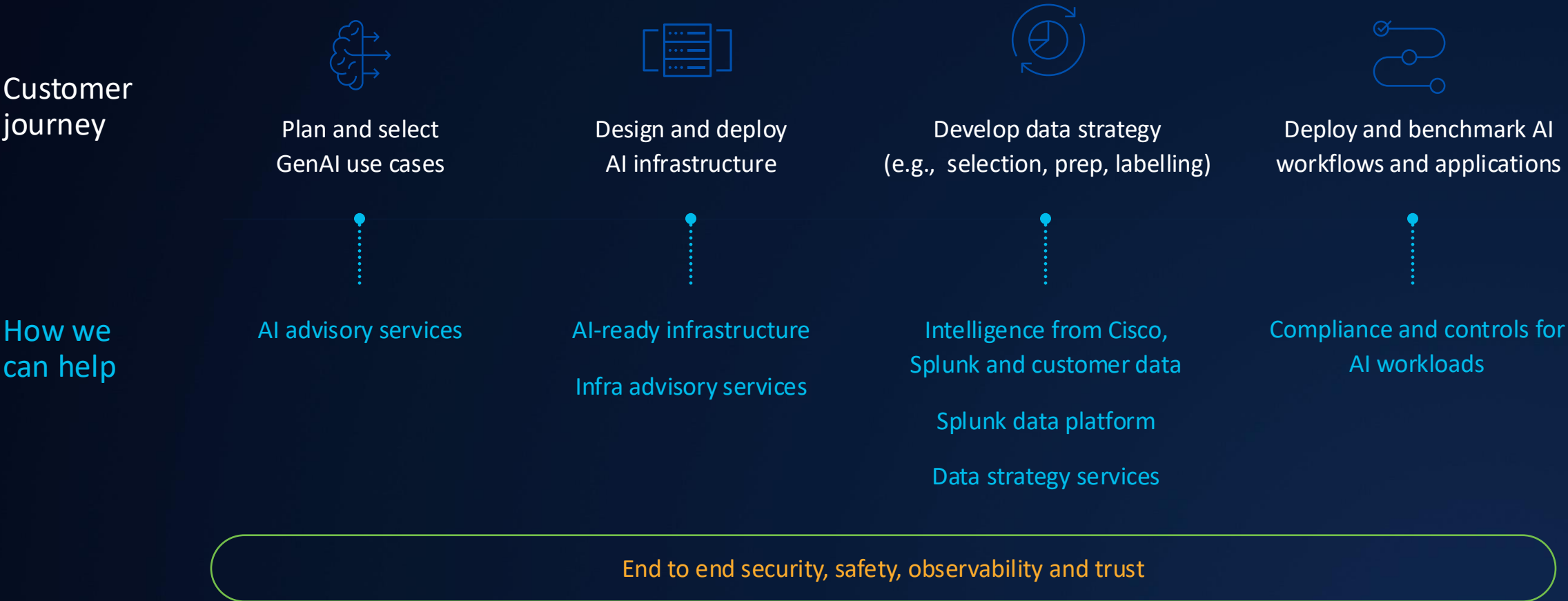**CVD Available NOW**

Cisco Nexus HyperFabric AI Cluster
Available by Q2FY25

PRIVATE CLOUD MANAGED AI INFRASTRUCTURE

CLOUD MANAGED AI INFRASTRCUTURE

Cisco supports customers along any starting point of their AI journey

Customer journey

Plan and select
GenAI use cases

Design and deploy
AI infrastructure

Develop data strategy
(e.g., selection, prep, labelling)

Deploy and benchmark AI
workflows and applications

How we can help

AI advisory services

AI-ready infrastructure

Infra advisory services

Intelligence from Cisco,
Splunk and customer data

Splunk data platform

Data strategy services

Compliance and controls for
AI workloads

End to end security, safety, observability and trust

Cisco can support you in AI:

- **Use Cases Development**: Cisco + Partner Ecosystem

- **Education**:

  - Business Level : Events, EBC

  - Technical Ebene : AI Seminars, Certs, Test Drives

- **Technology**: AI Fabrics, AI PODs, Cross Architecture

- **Demonstration**: Next Gen AI Lab, POC

- **Statement of Work**: Partner, Cisco Services (CX)

Cisco GSX

CISCO

The bridge to possible