Generative AI:
Hype meets Reality

Mohannad Abuissa

CTO - Cisco Middle East & Africa

Turkey, Romania and CiS

Abundance

Scarcity

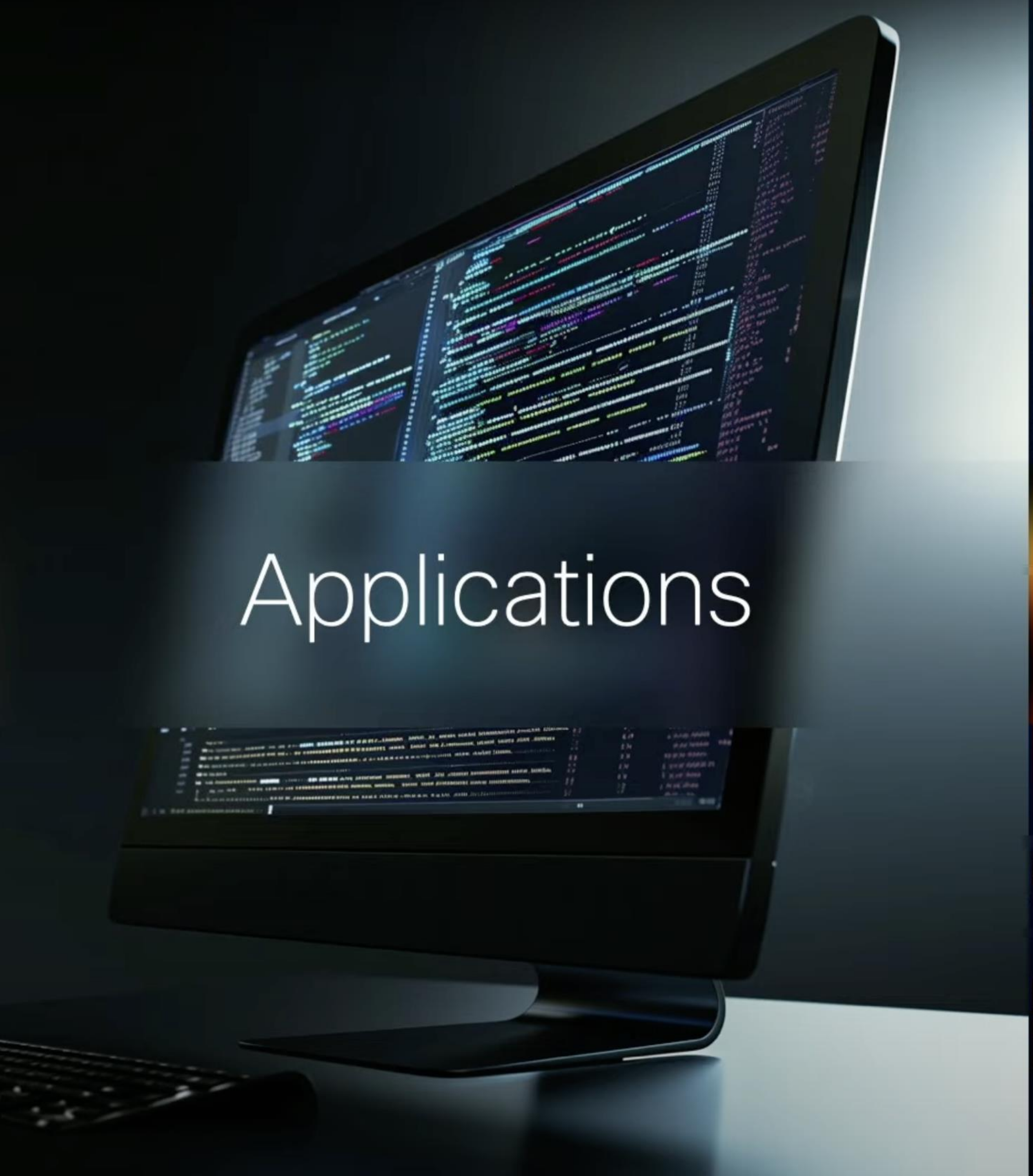
8

BILLION

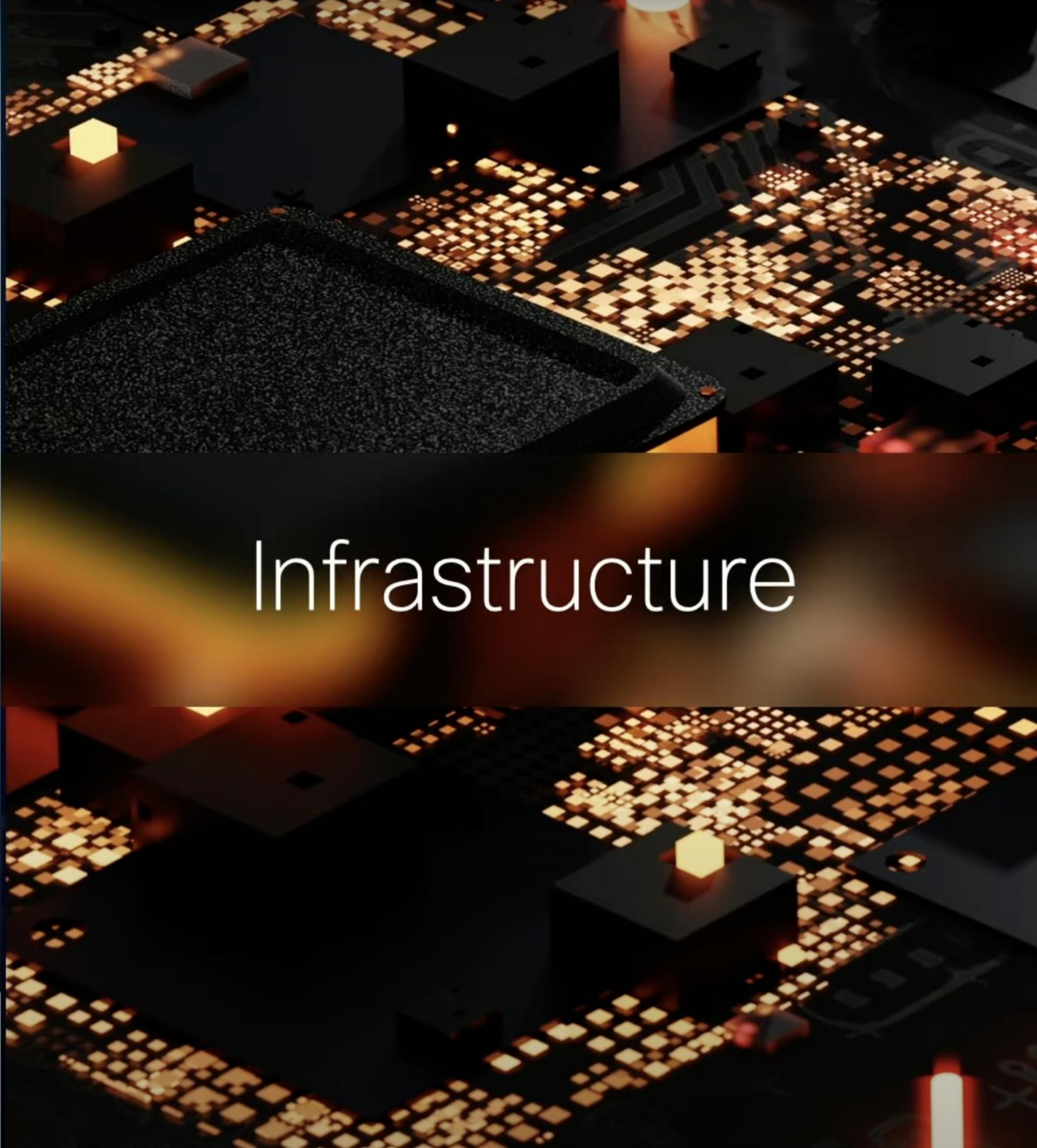
80

BILLION

Public & Private Data Centers



Applications



Infrastructure

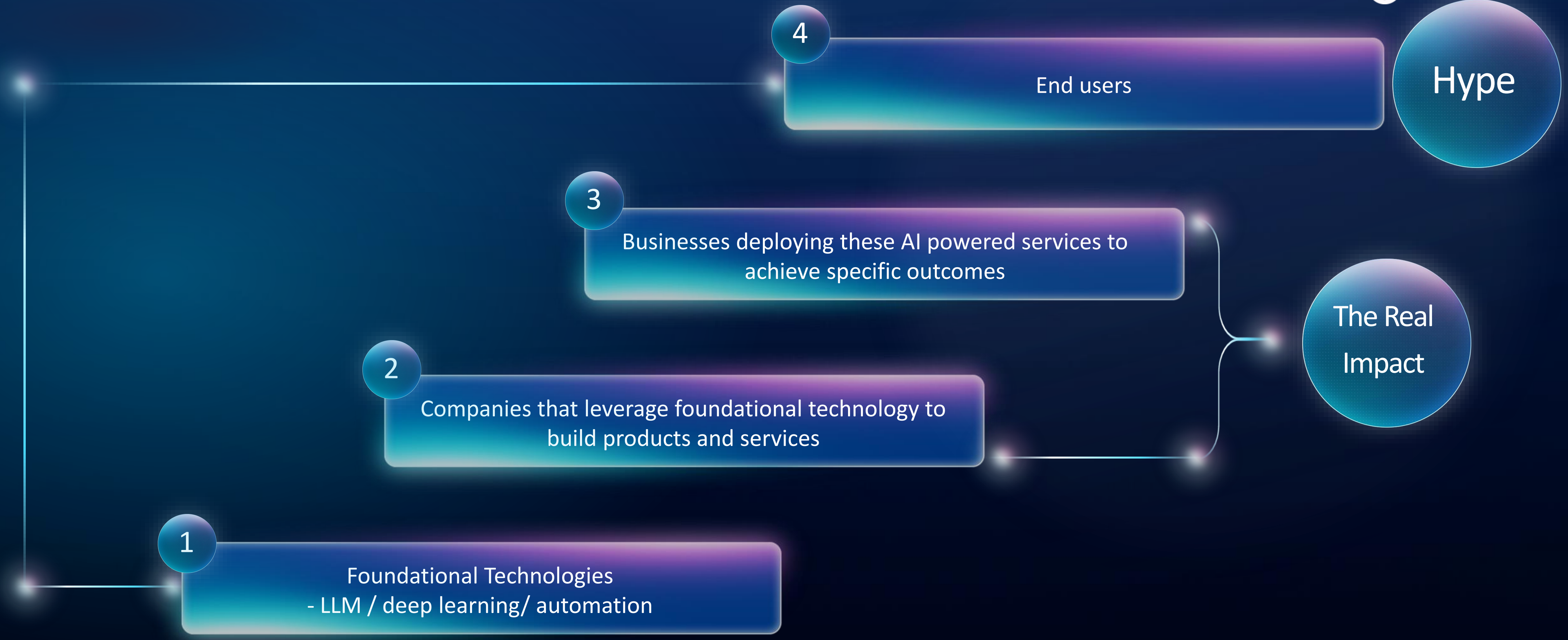
Evolving Apps & Infrastructure ▶

▶ Foundation for Generative AI Reality

Cisco AI Readiness Index 2024

AI: Hype vs Reality

ChatGPT



4 End users

Hype

3 Businesses deploying these AI powered services to achieve specific outcomes

2 Companies that leverage foundational technology to build products and services

The Real Impact

1 Foundational Technologies
- LLM / deep learning/ automation

What we set out to achieve

Measure
Readiness

Address Gaps
in readiness

What challenges, if any,
they are facing as they
address these?

Current and future
ROI on AI
Investments

AI Readiness:

The foundational building blocks

Strategy

Infrastructure

Data

Governance

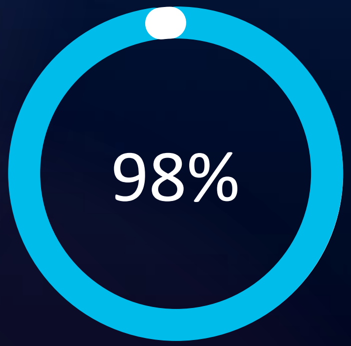
Talent

Culture

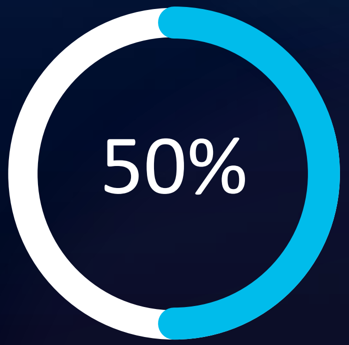
AI Readiness:

The urgency to deploy AI continues

The CEO and the Leadership team are driving the urgency, closely supported by the Board of Directors and Business Unit Leaders.



Feel that urgency to deploy AI / AI-powered technologies has increased in the past six months



Companies say CEO and the leadership team are top drivers of urgency to deploy AI

AI Readiness:

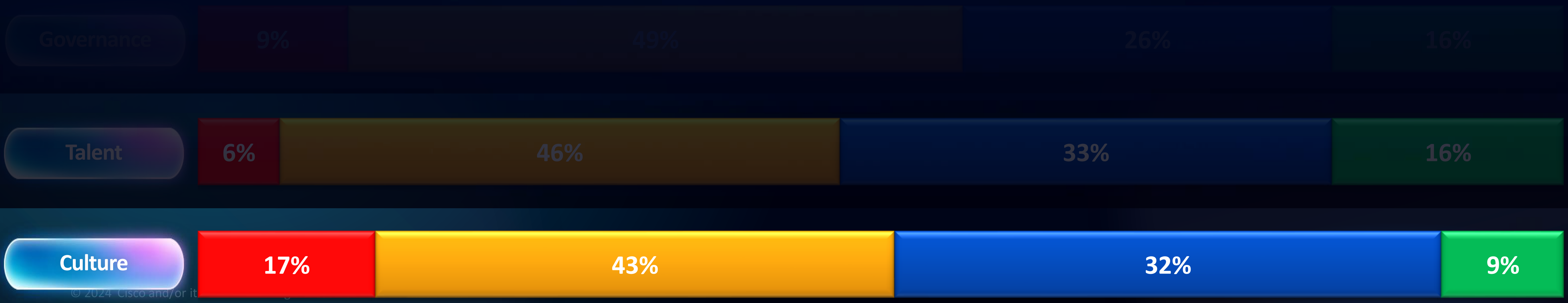
Key takeaways

Global AI Readiness is flatlining/declining

Companies are investing, but gains aren't meeting expectations

The pressure to succeed is relentless

Global AI Readiness



85%

Feel they have 18 months to show value or lose competitive advantage

59%

59% give it only 12 months

13%

Of companies are fully ready to capture AI's potential (down from 14%)

Gap

Between urgency and ability is especially startling

Networks are not equipped to meet AI workloads

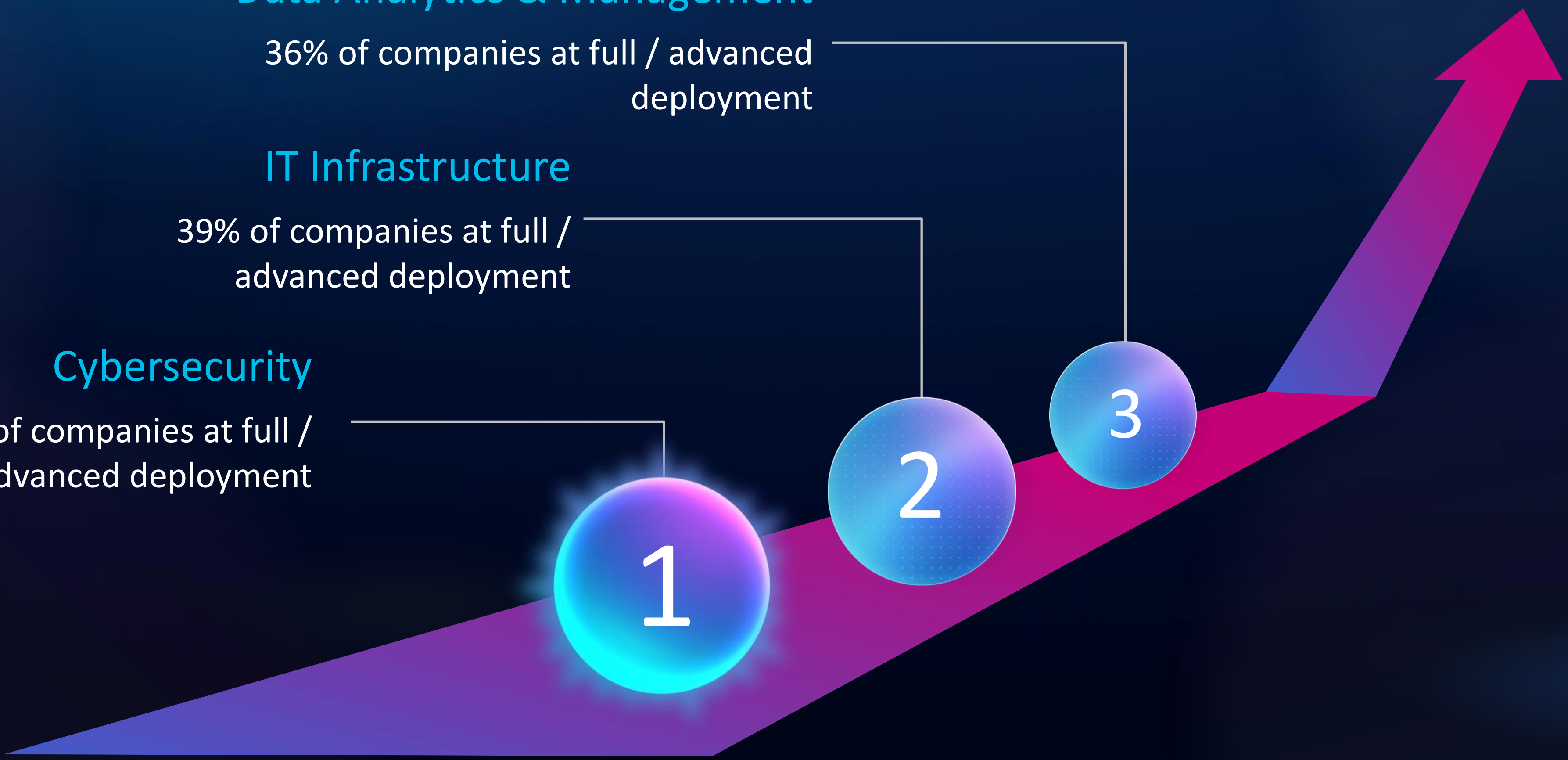
Top 3 challenges

Lack of Talent

Cybersecurity Risks

Long lead times

Top areas where businesses are deploying AI



Data Analytics & Management

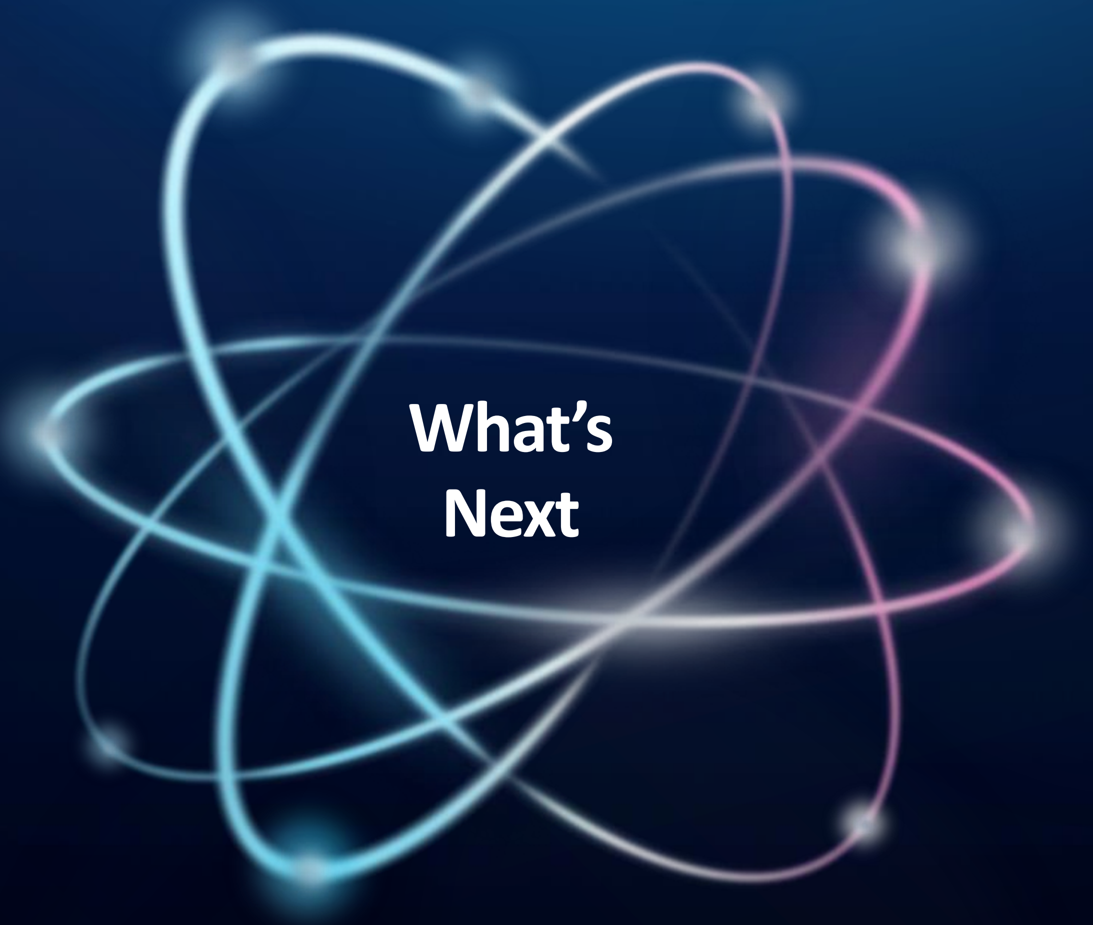
36% of companies at full / advanced deployment

IT Infrastructure

39% of companies at full / advanced deployment

Cybersecurity

42% of companies at full / advanced deployment



What’s Next

What can organizations do to boost AI Readiness?

Look long-term and think big

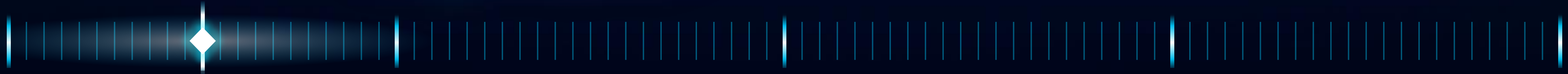
Build infrastructure for the future

Breakdown data silos

Keep people at the core

Deploy timely internal policies & protocols

Every organization's AI approach and needs are different



Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Model the world
Digital twin

Every organization's AI approach and needs are different

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Model the world
Digital twin

Every organization’s AI approach and needs are different

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Model the world
Digital twin

Every organization's AI approach and needs are different

Build the model
Training

Optimize the model
Fine-tuning and RAG

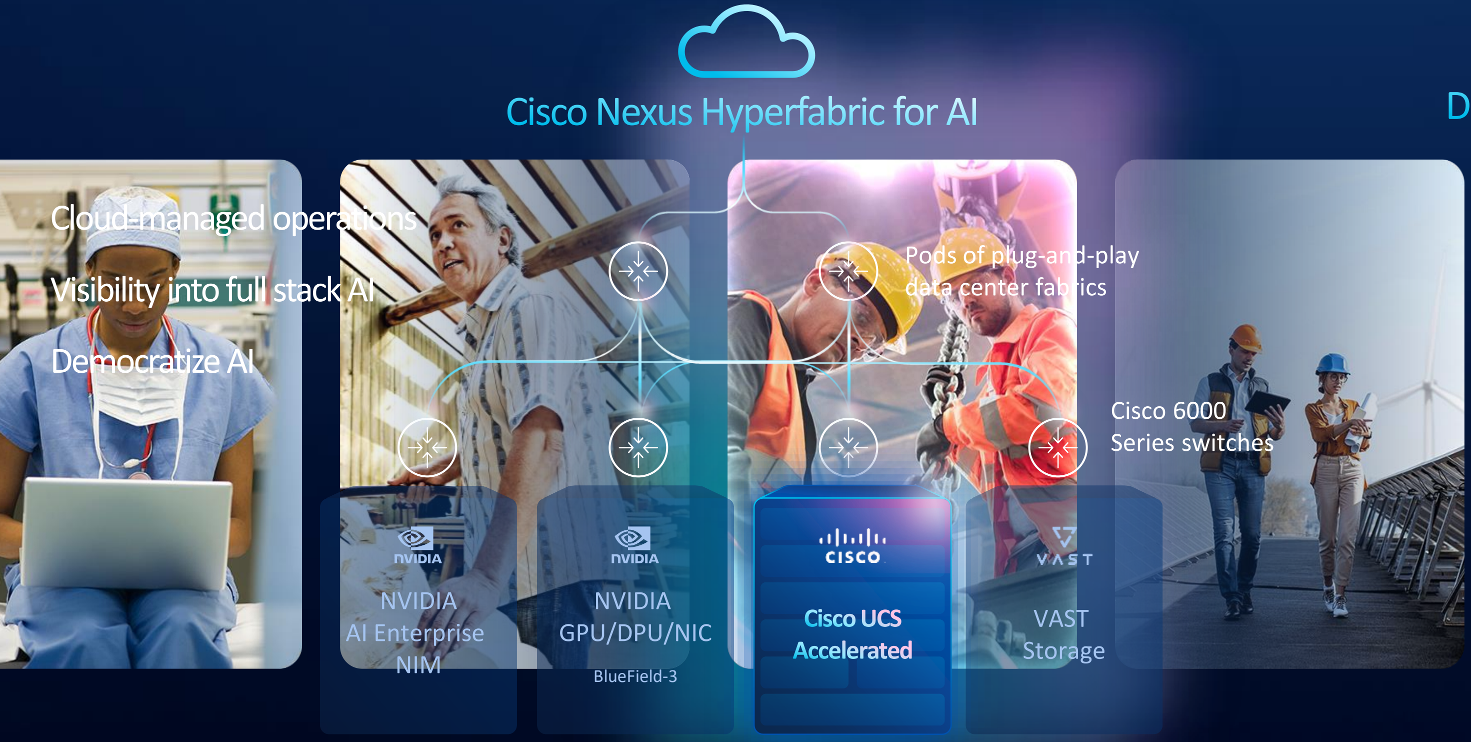
Use the model
Inferencing

Model the world
Digital twin

Every organization's AI approach and needs are different

AI Cluster in partnership with NVIDIA

Cisco Nexus Hyperfabric for AI



- Cloud-managed operations
- Visibility into full stack AI
- Democratize AI

Pods of plug-and-play data center fabrics

Cisco 6000 Series switches

Built on Cisco Silicon One and Optics innovations

Building high-density GPU servers to the Cisco UCS family & to Cisco's AI solution portfolio

Discover data-intensive use cases like model training and deep learning

UCS Accelerated

UCS C885A M8

Nvidia HGX with
8 Nvidia H100 GPUs

AMD Mi300X

2 AMD 4th Gen
EPYC™ Processors

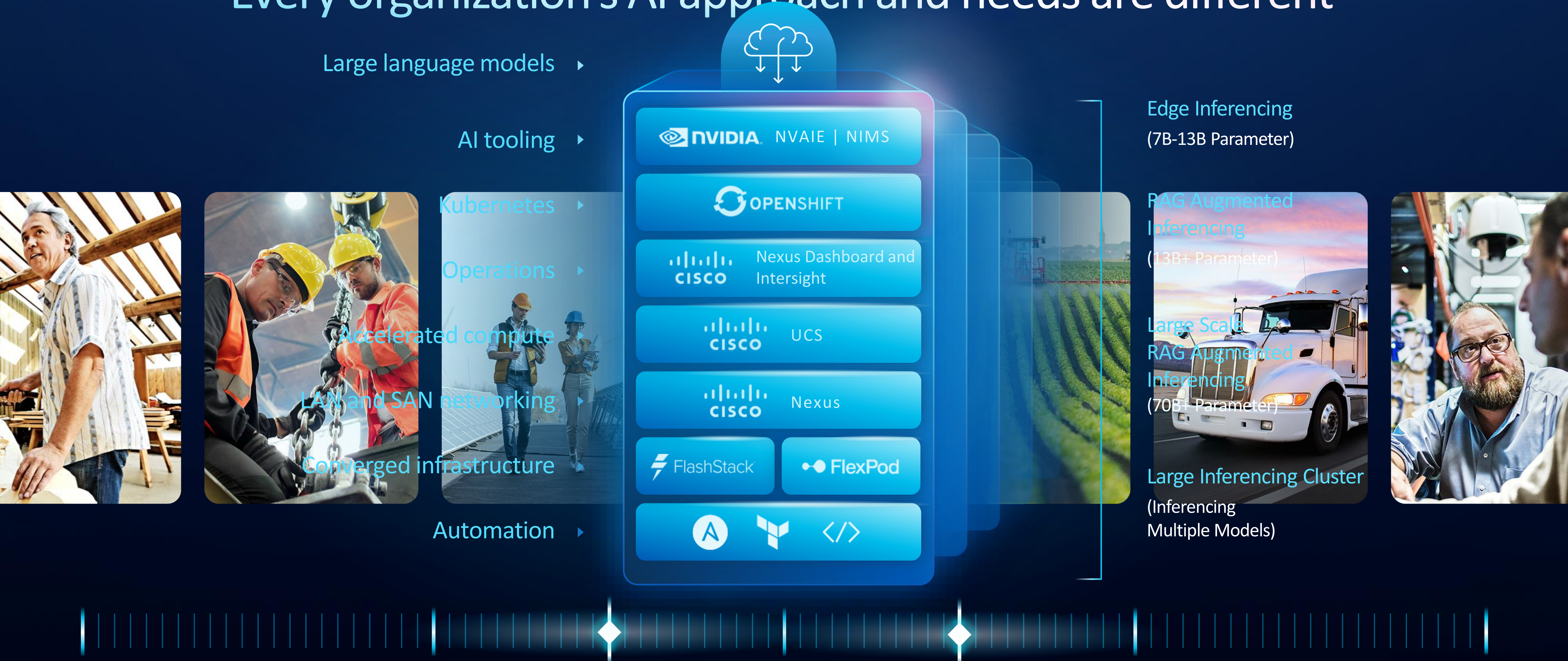
Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Model the world
Digital twin

Every organization's AI approach and needs are different



AI PODs

- Large language models
- AI tooling
- Kubernetes
- Operations
- Accelerated compute
- LAN and SAN networking
- Converged infrastructure
- Automation

NVIDIA NVAIE | NIMS

OPENSHIFT

CISCO Nexus Dashboard and Intersight

CISCO UCS

CISCO Nexus

FlashStack

FlexPod

Edge Inferencing
(7B-13B Parameter)

RAG Augmented Inferencing
(13B+ Parameter)

Large Scale RAG Augmented Inferencing
(70B+ Parameter)

Large Inferencing Cluster
(Inferencing Multiple Models)

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Model the world
Digital twin

Download the Cisco AI Readiness Report

Take the Cisco AI Readiness Assessment

cisco