シスコの最新 AI プラットフォームのご紹介
フルスタック AI インフラストラクチャの導入・運用をシンプルに

シスコシステムズ合同会社
クラウド・AIインフラストラクチャ事業
井上　景介

2024年9月27日

Agenda

- AI計算基盤を取り巻く環境
- AI基盤のシンプル化を促進するソリューションのご紹介
  - NVIDIA社と協業「Hyperfabric AI Cluster」
  - シンプルバンドルソリューション「Cisco AI Pods」
  - Nutanix社 HCIを活用した 「GPT-in-a-Box with CCHC」

AIにより期待される変革

$15.7T

2030年までに世界経済に貢献する可能性

$300B

2026年までの世界で支出されるAIへの投資金額

75%

2026年までにAIを導入したプロセスに依存するようになる企業の割合

ヘルスケアとライフサイエンス

診断
創薬
個々人に最適化された医療

金融サービス

不正検知
リスク評価
トレーディング

小売

パーソナライゼーション
在庫最適化
バーチャルエージェント

製造

予知保全
品質管理
需要予測

農業

収穫量の最適化
病害虫の予測と予防

運輸

ルート最適化
自律走行車
予知保全

エネルギー

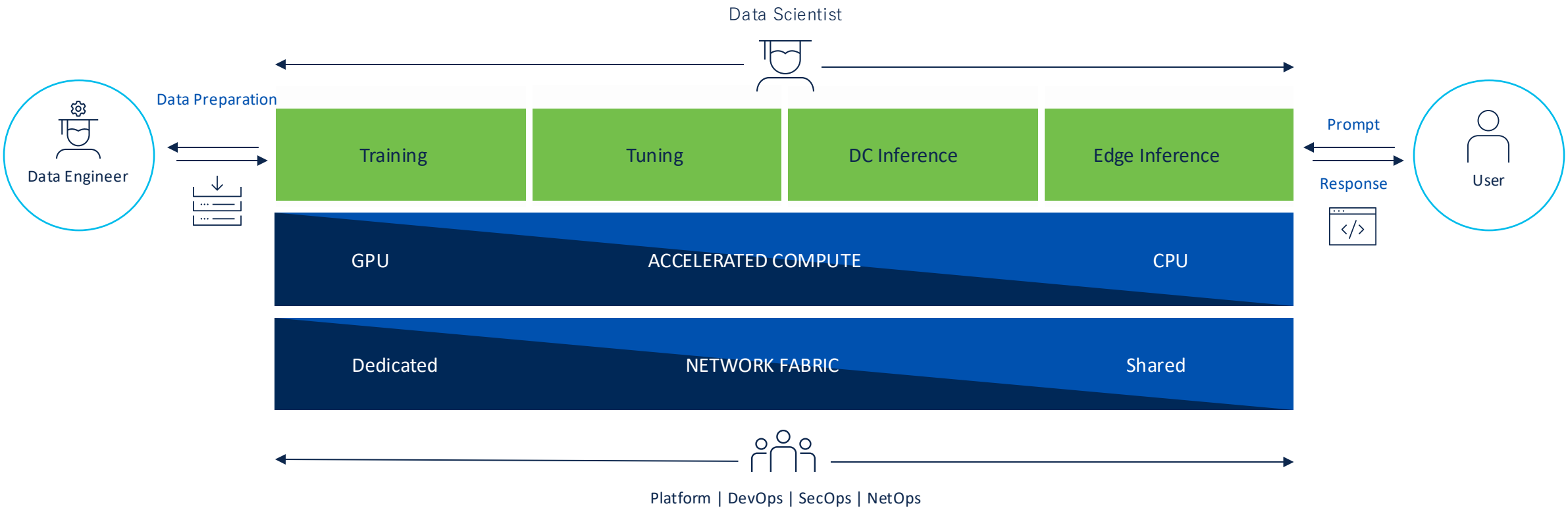
供給の最適化
故障予測
需要予測

公共

スマートシティ
セキュリティ
サービス改善

Sources: PWC, IDC

AI インフラストラクチャ



適切なフォーマットで適切な場所にあるデータを準備し、モデルの種類を選択し、準備されたデータでモデルを学習させ、モデルの精度とパフォーマンスを向上させ、本番環境へ導入する　という一連のプロセスをどこで、自ら行うのか、インフラ、スキル、時間に対してどのような投資を行うか大きく異なる。

AI共存社会 支えるインフラストラクチャ

チャレンジ

インフラ運用のサイロ化	馴染みのない断片化したアプリケーションスタック	倫理、プライバシー、コンプライアンス	複雑なインフラストラクチャのパターン
データ量、速度、多様性	技術的な専門知識の不足	参入コストが高く、ロックインの問題	多大な電力が必要

AI基盤のシンプル化の促進

Cisco Nexus HyperFabric AI Cluster	Controller	Cisco Internight/Nexus Dashboard
NVIDIA AI Enterprise/NCCL	Software Stack	Collective Communication
800G Ready Cisco 6000 series	Networking	800G Ready Nexus 9000 series
NVIDIA Tensor Core GPU	GPU(XPU)	NVIDIA / Intel / AMD
NVIDIA HGX/MGX (Cisco UCS)	Computing	Cisco UCS
NVIDIA BlueField-3 DPU, SuperNIC	NIC	NVIDIA / Intel / Cisco
VAST Data Platform	Storage	NetApp / PureStorage / Nutanix

AIフルスタックによるシンプル化

CY25 1H リリース予定

CISCO Validated Design

検証済み構成によるリスク低減

ご提供中

AI基盤のシンプル化の促進

Cisco Nexus HyperFabric AI Cluster	Controller	Cisco Internight/Nexus Dashboard
NVIDIA AI Enterprise/NCCL	Software Stack	Collective Communication
800G Ready Cisco 6000 series	Networking	800G Ready Nexus 9000 series
NVIDIA Tensor Core GPU	GPU(XPU)	NVIDIA / Intel / AMD
NVIDIA HGX/MGX (Cisco UCS)	Computing	Cisco UCS
NVIDIA BlueField-3 DPU, SuperNIC	NIC	NVIDIA / Intel / Cisco
VAST Data Platform	Storage	NetApp / PureStorage / Nutanix

AIフルスタックによるシンプル化

CY25 1H リリース予定

CISCO Validated Design

検証済み構成によるリスク低減

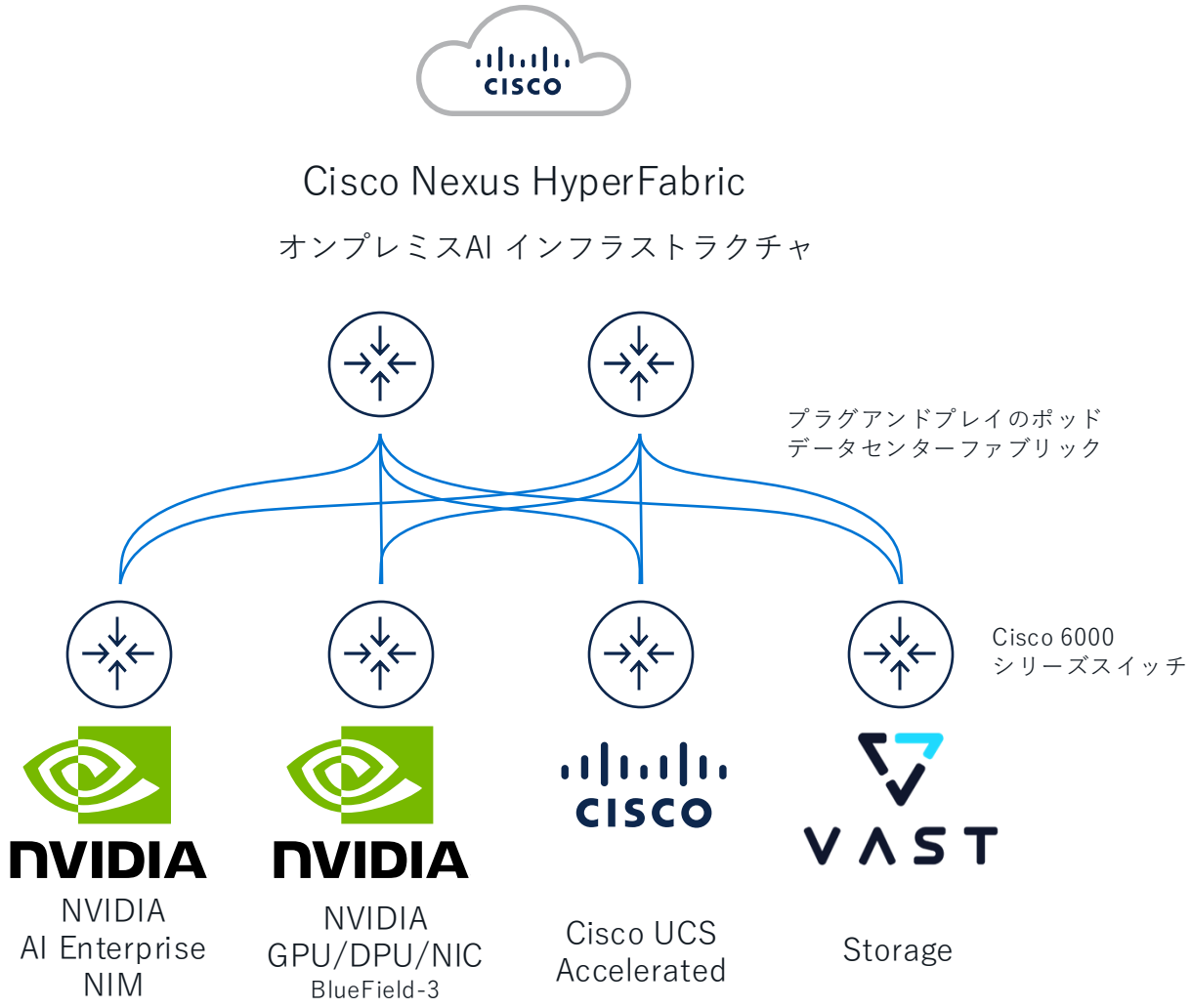
ご提供中

Cisco Nexus HyperFabric AI Cluster

in partnership with NVIDIA

AI Infrastructure の民主化	フルスタック AI の可視化
NVAIEを含む統合スタック	AIネイティブの運用モデル
ハイパフォーマンスイーサネット	クラウド管理オペレーション

IT ではなく AI イノベーションに専念できる新たなソリューション



Cisco Silicon One とOptics イノベーションを基盤として構成

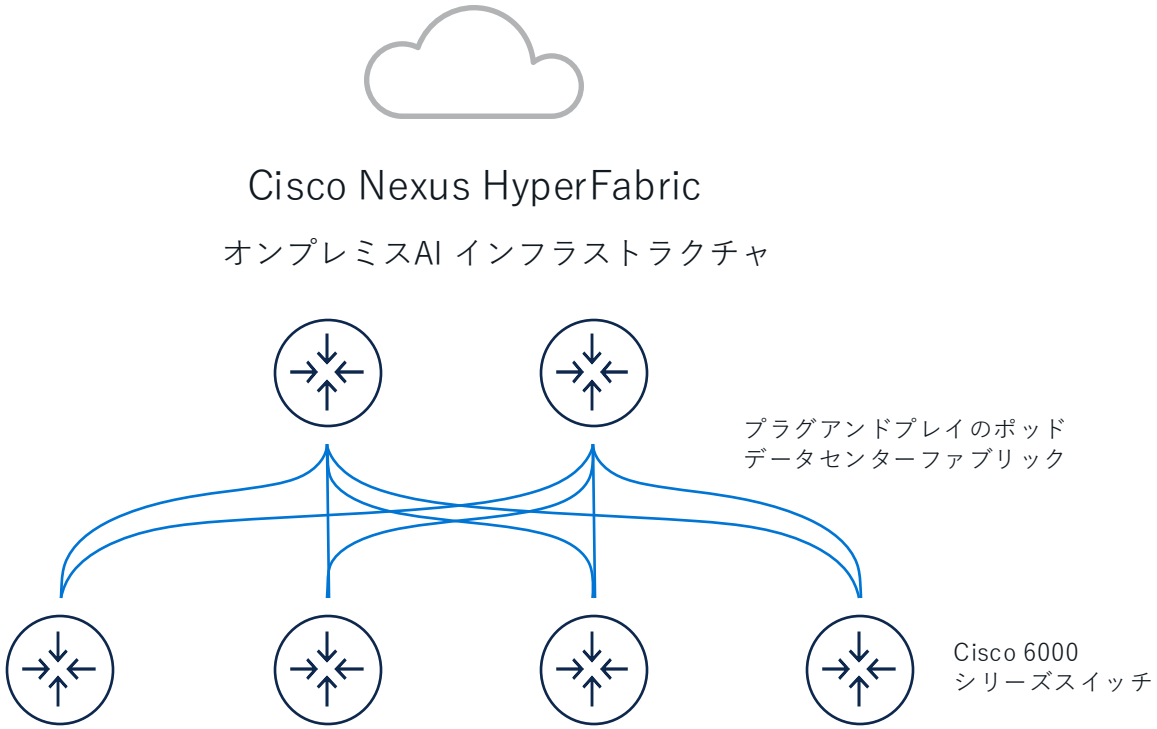
Cisco Nexus HyperFabric

Fabric-as-a-service

どこにでも設置可能なオンプレミスファブリックの設計、導入、運用

ITジェネラリスト、アプリケーションチーム、DevOpsチームにより扱いやすく

専用設計された垂直統合によってもたらされる成果



設計 › 発注 › 導入 › 検証 › 監視 › アップグレード › コラボレート

AI基盤のシンプル化の促進

Cisco Nexus HyperFabric AI Cluster	Controller	Cisco Internight/Nexus Dashboard
NVIDIA AI Enterprise/NCCL	Software Stack	Collective Communication
800G Ready Cisco 6000 series	Networking	800G Ready Nexus 9000 series
NVIDIA Tensor Core GPU	GPU(XPU)	NVIDIA / Intel / AMD
NVIDIA HGX/MGX (Cisco UCS)	Computing	Cisco UCS
NVIDIA BlueField-3 DPU, SuperNIC	NIC	NVIDIA / Intel / Cisco
VAST Data Platform	Storage	NetApp / PureStorage / Nutanix

AIフルスタックによるシンプル化

Q2FY25までにリリース予定

CISCO Validated Design

検証済み構成によるリスク低減

ご提供中

AI基盤を支えるCiscoデータセンタソリューション

Cisco ネットワーキング

Nexus & Nexus Dashboard

Cisco シリコン
&Optics

Cisco Silicon & Optics

Cisco コンピューティング

GPUアクセラレーション
UCSブレードおよびラックサーバー

cisco
Validated
Design

AI向け検証済み構成

Cisco Validated
Designs for AI

モデル構築｜トレーニング　モデルの最適化｜ファインチューニング　モデルの使用｜推論

プラットフォームのシンプル化を提供
オートメーション｜ビジビリティ｜オーケストレーション

AIユースケースに対応するUCS

現在のポートフォリオ

GPUアクセラレーション
UCSブレードおよびラックサーバー

CISCO VALIDATED DESIGN

AI向け検証済み構成

モデル構築｜トレーニング　モデルの最適化｜ファインチューニング　モデルの使用｜推論

新しい高密度
GPUサーバ

Coming Soon

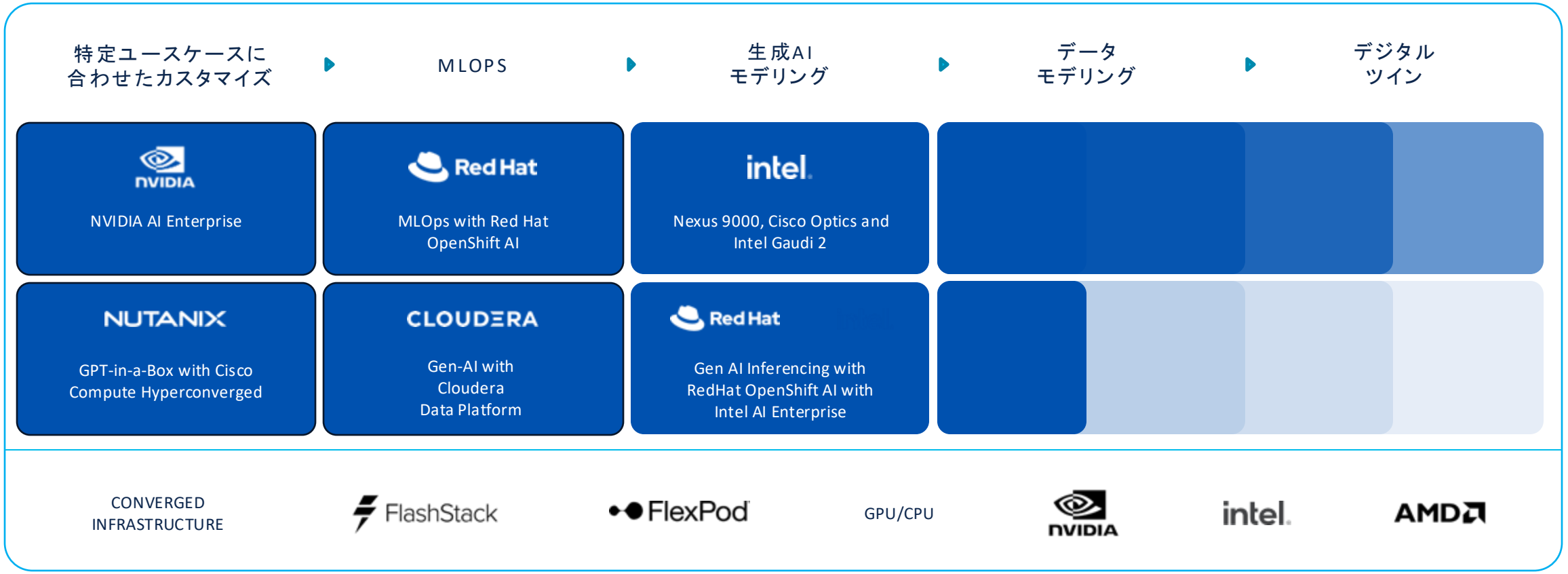
エンタープライズ向け
エッジAIプラットフォーム

エッジにおけるコンピュート、
ストレージ、ネットワーキングのための
1Boxソリューション

CY25

エコパートナー企業との検証済み構成

Validated Design



シンプルな基盤選択肢の提供 ①

Coming Soon

AI PODs

事前構成されたバンドルにより
サイジングからオーダまでの時間を短縮

安心してAIを導入

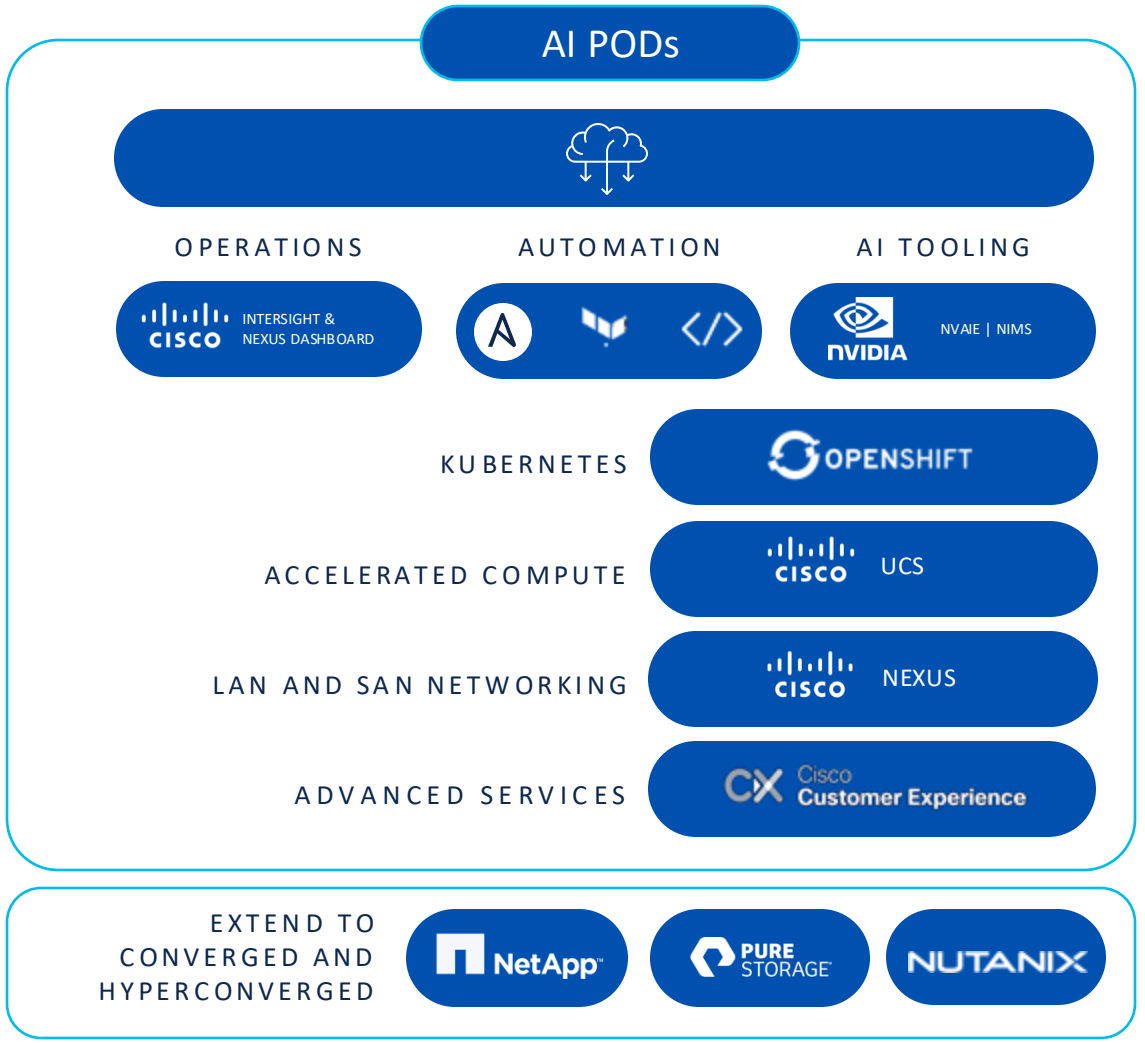
注文可能で検証済みの
AI-Readyインフラス
タック

シスコとサードパー
ティのコンポーネント
を含む完全にサポート
されたスタック

設定ガイダンスのための
AI Advisorツール

COMING SOON

Cisco AI-Ready Infrastructure Stacks



シンプルな基盤選択肢の提供 ②

リリース済み

GPT-in-a-Box with CCHC (Cisco Compute Hyperconverged)

オールインワンとは？

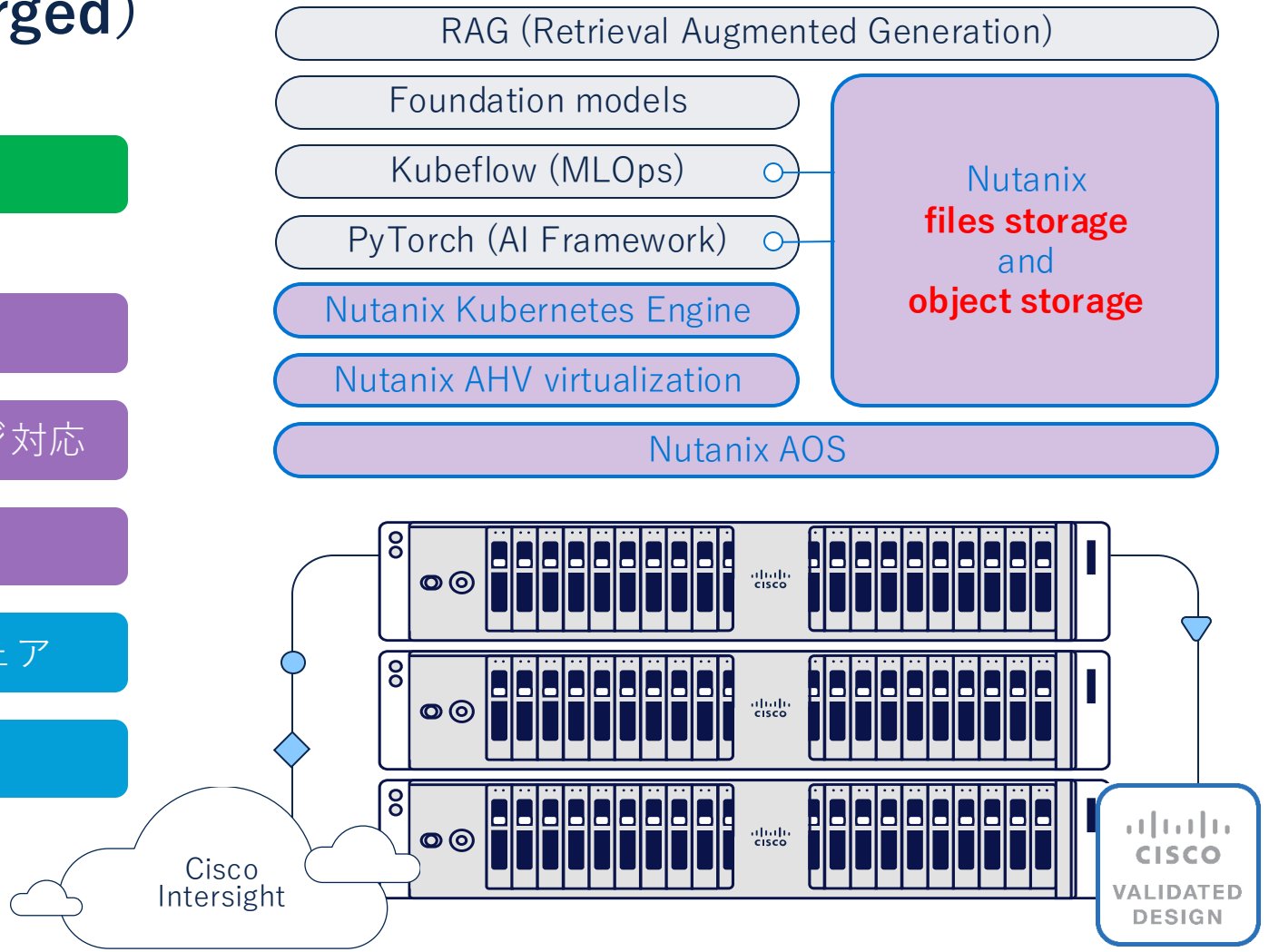
サーバ/ ストレージの一体型

ファイルサーバ/ オブジェクトストレージ対応

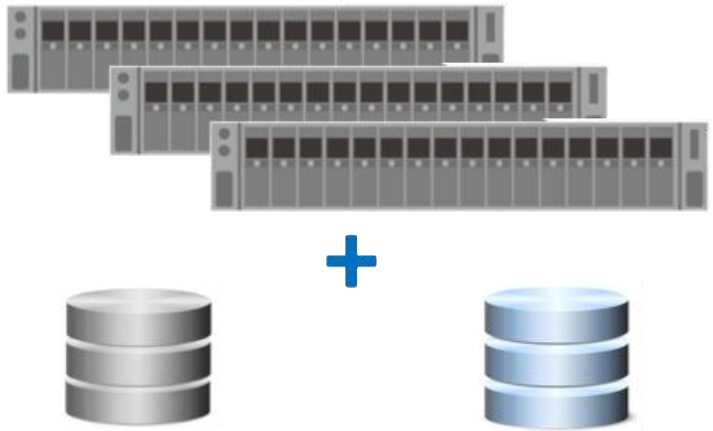
Kubernetes基盤の提供

設計・検証済オープンソースソフトウェア

設計・検証済Nutanix基盤

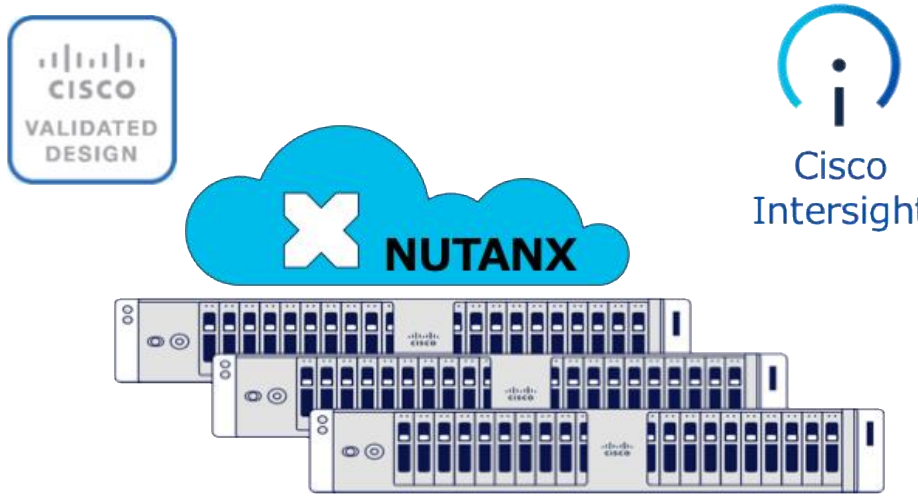


シンプルな基盤選択肢の提供 ②
GPT-in-a-Box with CCHC によるシンプル化のポイント



サーバと外部ストレージが必要で
管理もバラバラ

従来のインフラ構成



外部ストレージが不要で、
クラウドから統合管理を実現

Nutanix GPT-in-a-Box

シンプルな基盤選択肢の提供 ②
検証済み構成 - Nutanix GPT-in-a-box with CCHC

HCIAF240C M7 All-NVMe

CVD 概要

LLMの検索拡張生成（RAG）に焦点を当て、組織の内部または独自のナレッジベースを活用するオンプレミス生成AIの一般的なユースケースをもとにした構成を検証

CVD 構成

CISCO VALIDATED DESIGN

- テクノロジーの概要
- ソリューション設計
- ソリューションの導入
- ソリューションの検証

BOM 構成一覧

CCHC with Nutanix (AHV) クラスタ

- HCIAF240C M7 All-NVMe　4 台
  - Intel I6442Y 2.6GHz/225W 24C x2基
  - NVIDIA L40S x2基
  - NVMe 22.8TB
  - 1TB Memory DDR5-4800

Fabric Interconnect

- FI-6536 2 台

Nutanix サービス

- Nutanix Unified Storage（NUS）
- Nutanix Kubernetes Engine（NKE）

日本語URL
https://www.cisco.com/c/ja_jp/td/docs/unified_computing/ucs/UCS_CVDs/cisco_nutanix_gpt.html

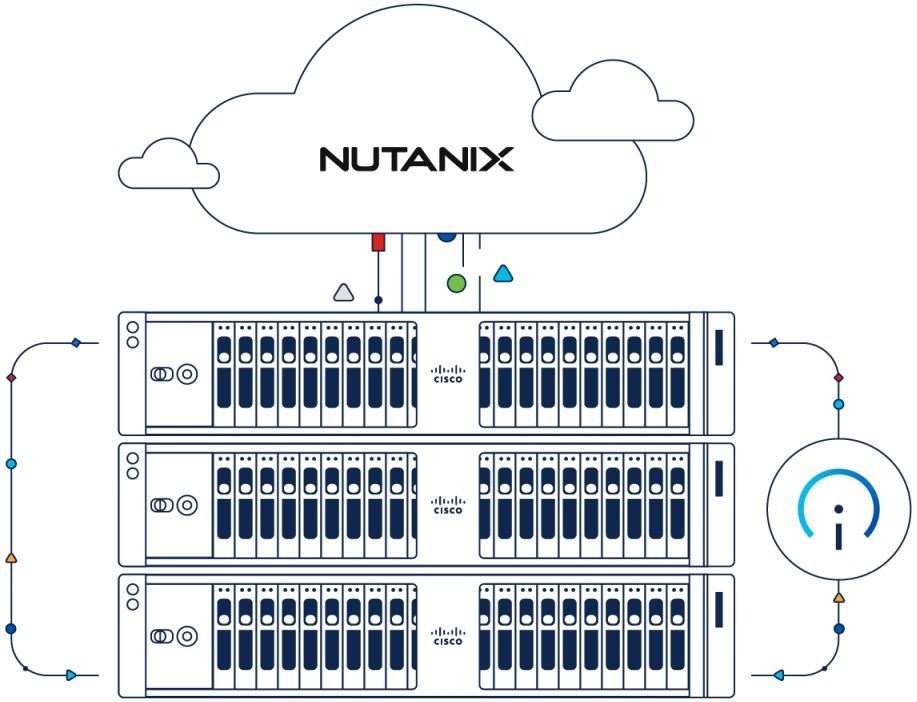
ご希望のお客様はCisco担当営業までご連絡ください

弊社CPOCの環境でローカルAIプラットフォームとなる**Nutanix GPT-in-a-Box環境**を無償で利用可能です。

弊社エンジニアのガイドのもと運用性の確認、性能の確認にご利用いただけます。ローカルAIプラットフォームをご検討のお客様は、ぜひこの機会に試してみませんか？

＜参考環境＞
- UCS-X with H100
- Intersight
- Nutanix GPT-in-a-Box

他のAIプラットフォーム環境はございますので、ユースケースとご利用期間を含めて弊社営業にご相談ください。




CISCO