データシート
Cisco Public

推論用 Cisco AI ポッド

2025 年 3 月

© 2025 Cisco and/or its affiliates. All rights reserved.

目次

製品の概要 3
機能と利点 4
優れた機能 4
Cisco のフル スタック ソリューションによる AI アーキテクチャ設計の変革 4
オン デマンドの拡張性 6
将来への対応 6
カスタマイゼーション 6
包括的なプラットフォーム 7
シンプルな管理 7
製品仕様 8
システム要件 10
発注情報 12
保証情報 15
製品持続性 15
シスコおよびパートナーの提供サービス 16
AI 対応インフラストラクチャ向けのCisco とパートナーのサービス 16
Cisco Capital 16
詳細情報 16
マニュアルの変更履歴 17

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI 対応 POD は、あらゆる組織が人工知能の力を利用するために設計された、専用のインフラストラクチャ ソリューションです。AI を開始する場合でも、複雑で高性能なワークロードをレガシーアーキテクチャを超えてスケーリングする場合でも、Cisco AI POD はイノベーションを推進するために必要なパフォーマンス、効率性、および制御を提供します。すぐに導入できる構成、一元化された管理、およびシームレスな拡張性を備えた Cisco AI POD は、移行のすべての段階で AI の可能性を最大限に引き出すのに役立ちます。

製品の概要

現在、すべての企業が AI ソリューションを展開する際の段階は同じではありません。これに対処するため、シスコは、最初の製品ラインである Cisco AI 対応インフラストラクチャ スタックを導入します。これらのスタックは、企業が人工知能の可能性を信頼性と容易さで迅速に活用できるようにすることを目的としています。クラス最高の管理ツールを備えた合理化されたすぐに使用できるインフラストラクチャを提供することで、AI の導入の複雑さを排除し、このテクノロジーに慣れていない企業でもアクセスできるようにしています。Cisco AI 対応インフラストラクチャ スタックは、迅速かつ検証されたセットアップと一貫した更新により、初期導入から高度なアプリケーションまでスケーラブルなパスを提供し、AI が主導する環境で組織が競争力を維持できるようにします。

Cisco AI 対応インフラストラクチャ スタックは、さまざまな AI および ML のユースケースに合わせた包括的なソリューションスイートを提供します。これらの検証済みの設計により、AI インフラストラクチャが簡素化および自動化されるため、エッジ推論から大規模なデータセンターまで、さまざまな環境に AI モデルを効率的に展開できます。Cisco® インフラストラクチャ スタックは、次のようなさまざまな要件に対応します：

- Red Hat OpenShift AI を使用した MLOps により、機械学習の運用とモデルの展開を合理化します。
- Intel®Gaudi 2 プロセッサとCisco Nexus® 9000 シリーズ スイッチを使用した Gen AI モデリングは、高性能 AI タスク用に最適化されています。
- Cisco の堅牢なネットワーキングおよびコンピューティング機能を搭載したデータ モデリング プラットフォーム。
- NVIDIA AI Enterprise および統合データシステムでサポートされる Digital Twin フレームワーク。

各ソリューションは、Cisco の検証済み設計に基づいて構築されており、AI アプリケーションのシームレスな統合と信頼性の高い実績を保証します。組織は、FlashStack や FlexPod などのコンバージド インフラストラクチャから選択できます。

小規模企業から大企業まで、さまざまなセグメントのお客様が Cisco AI インフラストラクチャに大きな価値を見出します。小規模企業は開発および基本的なワークロードにコスト効率の高い POD を利用できますが、大企業はカスタム POD を利用して集中的な AI およびデータ処理タスクを実行できます。主なビジネス上の利点には、生産性の向上、コストの削減、拡張性の向上が含まれますが、主な技術的価値は、既存のネットワーク インフラストラクチャ内でのシームレスな統合と運用を保証する POD の高い実績と信頼性にあります。

© 2025 Cisco and/or its affiliates. All rights reserved.

機能と利点

表 1. 機能と利点

機能	利点
オンデマンドの拡張性	展開サイズに合わせて構成可能なコンピューティング、メモリ、アクセラレータを使用して、進化する AI/ML ワークロードに合わせてリソースを動的に拡張します。
将来を見据えた	次世代のプロセッサ、GPU、およびストレージ ソリューションをサポートする将来性のあるモジュラ型システムで、長期的な投資保護を実現します。
カスタマイズ	特定の AI モデルとワークロードに合わせて構成を調整し、エッジ、中規模、大規模、スケールアウトの展開全体で最適な実績と効率を実現します。
包括的なプラットフォーム	Cisco Intersight® および Nexus Dashboard と統合された Cisco AI ポッド は、一元化されたプロビジョニング、リアルタイムの可視性、および自動スケーリングにより、インフラストラクチャ管理を簡素化します。
シンプルな管理	Cisco Intersight で管理を合理化し、Cisco UCS® サーバーと GPU の一元化されたプロビジョニング、リアルタイムの可視性、自動化を提供し、運用の複雑さを軽減します。
柔軟な構成	小規模なエッジ推論セットアップから大規模なマルチモデル展開まで、複数の AI ポッド サイズから選択して、特定のワークロード要件に対応します。

優れた機能

Cisco のフル スタック ソリューションによる AI アーキテクチャ設計の変革

このソリューションは、ハードウェアとソフトウェアの優れたテクノロジーを組み合わせて、さまざまなニーズに合わせた、堅牢でスケーラブルで効率的な AI 対応インフラストラクチャを作成します。

1. **NVIDIA AI Enterprise：** これらのコンポーネントは、AI ワークロードを管理するための強力な基盤を提供します。実績が最適化され、運用が合理化され、高度な管理機能が提供されるため、ユーザーは基盤となるインフラストラクチャの複雑さを気にすることなく、AI モデルの開発と展開に集中できます。
2. **OpenShift by Red Hat：** この Kubernetes ベースのコンテナプラットフォームは、コンテナ化された AI アプリケーションのオーケストレーションと展開を簡素化します。柔軟性と拡張性を提供するため、ユーザーは AI プロジェクトの成長に合わせて簡単に管理および拡張できます。
3. **Cisco Unified Computing System™ （Cisco UCS）：**Cisco はハードウェアのバックボーンとして機能し、複数のサーバー構成で高い実績と信頼性を提供します。これにより、要求の厳しい AI ワークロードに必要なコンピューティング能力をすぐに利用でき、スケーラブルになります。
4. **コンバージド インフラストラクチャ ソリューション：**
   - FlashStack（Pure Storage を搭載した Cisco UCS サーバー）と FlexPod（NetApp ストレージを搭載した Cisco UCS サーバー）はどちらも、大規模な AI および機械学習タスクの処理に不可欠な高性能ストレージとコンピューティング機能を提供します。これらのソリューションは、展開と管理を簡素化し、運用効率を向上させ、市場投入までの時間を短縮する、統合された事前検証済みの構成を提供します。

© 2025 Cisco and/or its affiliates. All rights reserved.

5. **Cisco Nexus および Nexus Dashboard :** Cisco Nexus スイッチは、高性能 AI/ML ワークロードをサポートするように設計された AI 対応インフラストラクチャを提供します。Nexus 9000 シリーズ スイッチと Nexus Dashboard の運用および自動化プラットフォームを組み合わせることで、AI/ML アプリケーションに不可欠な高帯域幅、低遅延、ロスレス ネットワークを実現できます。

Cisco のフルスタックソリューションを活用することで、お客様は次のようなメリットを得られます。

- **購入しやすい**
  - Cisco AI ポッド は、すぐに注文できる事前設計済みの検証済みバンドルとして提供されるため、お客様は特定の AI ニーズに適したセットアップを迅速に選択できます。これにより、個々のコンポーネントを選択する複雑さが解消され、意思決定にかかる時間が短縮されます。
  - 今四半期に発注可能なバンドルを提供することで、Cisco はお客様の価値実現までの時間を短縮し、不要な遅延なく AI イニシアチブを加速できるようにします。
- **展開と統合が容易：**
  - Cisco AI ポッド はプラグアンドプレイ展開向けに設計されており、最小限の労力で既存のデータ センターまたはクラウド環境に迅速に統合できます。このシームレスな展開機能は、Cisco の Intersight および Nexus Dashboard によってサポートされており、一元管理、リアルタイムの可視性、および自動化を提供して、運用の複雑さを軽減します。
  - 事前設計されたバンドルは事前にテストおよび検証されていることを表し、すべてのコンポーネントが箱から出してすぐに調和して動作することが保証されます。これにより、展開の問題のリスクが軽減され、AI ワークロードの実稼働までの時間が短縮されます。
  - Whether deployed in a converged infrastructure setup like FlashStack and FlexPod, Cisco AI PODs offer flexibility and scalability. これにより、幅広い既存のインフラストラクチャ設定に簡単に統合でき、組織のニーズに合わせて拡張できる将来に対応したソリューションが提供されます。
- **管理が容易：**
  - ソリューションの事前統合および事前テストの性質により、AI インフラストラクチャの設計の複雑さが軽減されます。最適な実績を実現するための明確で複製可能な公式を提供し、当て推量を排除します。
  - このソリューションは、小規模なエッジ推論セットアップから大規模で高性能な環境まで、さまざまな展開規模をサポートします。この拡張性により、組織のニーズに合わせてインフラストラクチャを拡張できます。
  - NVIDIA と Red Hat の統合管理ツールにより、AI インフラストラクチャの日常的な運用が簡素化され、IT チームによる AI プロジェクトのサポートが容易になります。

アイデアの段階にある個人や組織にとって、Cisco のフルスタック ソリューションは、堅牢な AI アーキテクチャを開発するための明確で構造化されたパスを提供します。最新の AI 管理、コンテナ オーケストレーション、および高性能ハードウェアを組み合わせたもので、AI 機能の構築と拡張を検討しているすべての人にとって理想的な選択肢です。

© 2025 Cisco and/or its affiliates. All rights reserved.

オン デマンドの拡張性

要件は頻繁に変更されます。また、別のリソースが必要になったときに、リソースのセットに制限されないシステムが必要です。Cisco AI 対応インフラストラクチャ スタックは、Cisco UCS X シリーズのモジュール型の性質により、本質的に成長に対応しており、生成 AI 推論のニーズに合わせてシームレスに拡張できます。Cisco Intersight を使用して、モデルの進化やワークロードの拡大に合わせて、サーバーの追加または削除、メモリ容量の調整、リソースの設定を自動化するだけで簡単に実行できます。さらに、特定のユース ケースに応じて、CPU と GPU の比率を変更したり、単一のシャーシ内で Intel Xeon® Scalable または AMD EPYC プロセッサを選択したりすることもできます。最後に、UCS X シリーズ以外のノードのサポートを通じて、POD のさらなる拡張を推進し続けます。

- **動的リソース割り当て：** コンピューティング リソースとストレージ リソースを簡単に追加または削減します。
- **柔軟な CPU/GPU 比率：** コンピューティング能力のバランスをカスタマイズします。
- **自動スケーリング：** ワークロードの増加に応じてリソースを自動的に調整します。

将来への対応

Cisco の AI 対応インフラストラクチャ スタックは、将来に備えて構築されており、新しいテクノロジーで進化し、新たなビジネス目標に適応します。UCS X シリーズ モジュラ システムは、将来の世代のプロセッサ、ストレージ、不揮発性メモリ、アクセラレータ、および相互接続をサポートするように設計されました。個別の管理モジュールやサーバを購入、設定、保守、電力供給、冷却する必要がなく、部分ごとに選択的にアップグレードできる。Intersight Software-as-a-Service モデルによって新しい機能が頻繁に提供されても、自動的にクラウドベースの管理は最新の状態に保たれます。Cisco Nexus 400G スイッチは、スケーラブルなソリューションと将来の高速化への対応力を提供し、データ センターの進化するニーズに対応する堅牢なサポートを提供します。その結果、ハードウェアのライフサイクルが延長され、時間の経過とともに継続的なコストが削減されます。

- **モジュール設計：** 大規模なオーバーホールなしでコンポーネントを簡単にアップグレードできます。
- **寿命：** 次のハードウェア世代のサポート。
- **投資の保護：** 上位互換性のあるインフラストラクチャにより、テクノロジーへの投資を保護します。

カスタマイゼーション

Cisco のカスタマイズ可能な構成により、選択したモデルをホストし、構成を編集し、さまざまなデータソースやサービスに接続できます。この柔軟性により、特定のニーズと展開規模に合わせて、AI ワークロードに対する最適な実績と制御が保証されます。

- **モデルの柔軟性：** お客様固有の要件に合わせて任意の AI モデルをホストできます。
- **構成可能なセットアップ：** 必要に応じてハードウェアのセットアップを変更します。
- **包括的な制御：** 展開環境を完全に管理および制御します。

© 2025 Cisco and/or its affiliates. All rights reserved.

包括的なプラットフォーム

AI の運用化は、すべての企業にとって困難な作業です。AI/ML の取り組みが概念実証から実稼働に移行できない最大の理由の 1 つは、実稼働対応モデルを提供するデータと機械学習の合理化と管理に関連する複雑さにあります。AI/ML の取り組みごとに技術的負債を追加するアドホックな ML の取り組みではなく、モデルを迅速かつ正確に提供および維持できるプロセス、ツール、ベストプラクティスを採用することが重要です。Cisco AI 対応インフラストラクチャ スタックは、複雑さを軽減することで AI/ML の取り組みを促進する、専用のフルスタック ソリューションを提供します。この設計では、Cisco UCS X シリーズ モジュラ プラットフォームと最新の Cisco UCS M7 サーバー、NVIDIA GPU を搭載した Cisco UCS X440p PCIe ノードを使用します。これらはすべて、Cisco Intersight を使用してクラウドから一元管理されます。Cisco Nexus スイッチとカップリングすることで革新的な輻輳管理およびフロー制御アルゴリズムに加えて、AI/ML ファブリックの設計要件を満たすために正しい遅延とテレメトリをサポートします。この高速化された高性能インフラストラクチャに基づいて構築された Red Hat OpenShift AI は、AI を一貫して効率的に加速し、運用するために不可欠なツールとテクノロジーを統合することでソリューションを強化します。NVIDIA AI Enterprise ソフトウェアはソリューションをさらに補完し、仮想 GPU、GPU オペレータ、AI 用に最適化されたツールとフレームワークの包括的なライブラリなどの主要な機能を提供します。これらの機能により、AI ソリューションの展開とスケーリングが簡素化され、企業が AI テクノロジーを簡単に運用できるようになります。

- **エンドツーエンドのソリューション：** インフラストラクチャから AI フレームワークまで。
- **生成 AI サポート：** 大規模な AI モデルの展開に最適化されています。
- **エンタープライズグレード：** 大規模な企業の要求を満たすのに十分な堅牢性を備えています。

シンプルな管理

Cisco Intersight でよりインフラストラクチャ管理を最適化します。Cisco のプラットフォームは、一元化されたプロビジョニング、リアルタイムの可視性、および自動化を提供し、UCS サーバと GPU の管理を簡素化します。これにより、運用が合理化され、手動による介入が減少し、効率が向上し、複雑さが軽減されます。

- **一元管理：** 単一のインターフェイスからすべてのリソースを管理します。
- **リアルタイムの可視性：** 実績とステータスをリアルタイムでモニターします。
- **自動化：** 自動化された管理プロセスにより、手動による介入を削減します。

© 2025 Cisco and/or its affiliates. All rights reserved.

製品仕様

Cisco AI Data Center and Edge Inference Pod は、エッジ推論アプリケーション向けに専門的に設計されており、データソースに近いネットワークエッジのユーザーのすぐ近くで計算を促進します。この戦略的な設計は、一元化されたクラウドやデータセンターに依存するのではなく、データをローカルで処理することで、遅延を最小限に抑え、効率を最大化します。Llama 2-7B、GPT-2B、その他の小規模言語モデル（SLM）などの高度なモデルをサポートするデータセンターおよびエッジ推論ポッドは、非常に用途が広く、機能が豊富です。シャーシ内に統合された X シリーズ ダイレクト ファブリック インターコネクトが含まれているため、追加のハードウェアが不要になるため、複雑さが軽減され、運用が合理化されます。

表 2. Cisco AI データ センターおよびエッジ推論ポッドの仕様

項目	仕様
コンピューティングノード	1 Cisco UCS X210c M7 コンピューティング ノード
プロセッサ	各コンピューティングノードにデュアル Intel第 5 世代 6548Y+ プロセッサ
メモリ	8X 64 GB DDR5 (4800 MT/s)、コンピューティング ノードあたり合計 512 GB のメモリ、または Cisco AI 対応データ センター Edge Inference Pod が単一ノード構成であるため合計 512 GB。
の内蔵ストレージ)	5X コンピューティング ノードあたり 1.6 TB Non-Volatile Memory Express（NVMe）2.5 インチ ドライブ
mLOM	Cisco UCS VIC 15230 X 1 100G mLOM X 2
PCIe ノード	コンピューティングノードあたり 1X Cisco UCS X440p PCIe ノード
GPU	1X コンピューティング ノードあたり NVIDIA L40S GPU（デュアル スロット）
管理	• Cisco Intersight ソフトウェア（SaaS、仮想アプライアンス、およびプライベート仮想アプライアンス） • Cisco UCS Manager：UCSM v4.3(2) 以降

Cisco AI RAG 拡張推論ポッドは、要求の厳しい AI ワークロード専用に設計されており、Llama 2-13B や OPT 13B などの大規模モデルをサポートします。強化された GPU とノード容量を搭載し、複雑なタスクを効率的に管理し、計算能力と効率を向上させます。この高度な POD は、知識ソースを活用してクエリ サービス中にコンテキストに基づく関連性を提供するRAG（Retrieval-Augmented Generation）のような最適化に対応し、幻覚を大幅に減らし、応答精度を向上させることができます。そのため、Cisco RAG 拡張推論ポッドは、堅牢な AI 機能と優れた実績を必要とする高度なアプリケーションに最適です。

© 2025 Cisco and/or its affiliates. All rights reserved.

表 3.　Cisco AI RAG 拡張推論ポッドの仕様

項目	仕様
コンピューティングノード	2 X Cisco UCS X210c M7 コンピューティング ノード
プロセッサ	コンピューティングノードごとにデュアル Intel第 5 世代 6548Y+ プロセッサ、または合計 4 つの CPU
メモリ	4800 MT/秒 の 64 GB DDR5 X 16（CPU あたり 4 DIMM）、コンピューティングノードあたり合計 512 GB のメモリ、または RAG 推論ポッドはデュアルノード構成であるため、合計で 1 TB になります。
の内蔵ストレージ)	コンピューティング ノードあたり 5X 1.6 TB Non-Volatile Memory Express（NVMe）2.5 インチ ドライブ
mLOM	コンピューティング ノードあたり 1X Cisco UCS VIC 15230 2X 100G mLOM
PCIe ノード	コンピューティング ノードあたり 1X Cisco UCS X440p PCIe ノード
GPU	コンピューティング ノードあたり 2X NVIDIA L40S GPU（デュアルスロット）、または合計 4 個の GPU
管理	• Cisco Intersight ソフトウェア（SaaS、仮想アプライアンス、およびプライベート仮想アプライアンス） • Cisco UCS Manager：UCSM v4.3（2）以降
Networking	• 2X Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス

Cisco AI Scale Up Inference Pod for High Performance は、CodeLlama 34B や Falcon 40B などの大規模モデルをサポートするように細心の注意を払って最適化されています。十分な GPU とノード容量を備えたこのポッドは、最も複雑な AI タスクに対して優れた実績を提供するように設計されています。大規模なベクトルデータベースをサポートし、高度なアルゴリズムによってベクトル検索の精度を向上させることで、取得拡張生成（RAG）機能を大幅に強化します。

表 4.　Cisco AI Scale Up Inference Pod for High Performance の仕様

項目	仕様
コンピューティング ノード	2X Cisco UCS X210c M7 コンピューティング ノード
プロセッサ	コンピューティング ノードごとにデュアル Intel第 5 世代 6548Y+ プロセッサ、または合計 4 つの CPU
メモリ	4800 MT/秒 の 64 GB DDR5 X 16（CPU あたり 4 DIMM）、コンピューティングノードあたり合計 512 GB のメモリ、または RAG 推論ポッドはデュアルノード構成であるため、合計で 1 TB になります。
内蔵ストレージ	コンピューティング ノードあたり 5X 1.6 TB Non-Volatile Memory Express（NVMe）2.5 インチ ドライブ
mLOM	コンピューティング ノードあたり 1X Cisco UCS VIC 15230 2X 100G mLOM
PCIe ノード	コンピューティング ノードあたり 1X Cisco UCS X440p PCIe ノード
GPU	コンピューティング ノードあたり 2X NVIDIA H100 GPU（デュアルスロット）、または合計 4 個の GPU

© 2025 Cisco and/or its affiliates. All rights reserved.

項目	仕様
管理	• Cisco Intersight ソフトウェア（SaaS、仮想アプライアンス、およびプライベート仮想アプライアンス） • Cisco UCS Manager：UCSM v4.3（2）以降
Networking	• 2x Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス

大規模展開向け Cisco AI スケールアウト推論ポッドは、比類のない柔軟性を提供するように設計されており、単一のシャーシ内で複数のモデルを同時に実行できます。堅牢なライフサイクル管理やモデルの高可用性を必要とする組織に最適です。このポッドは、特定のタスクごとに最適なモデルを選択できるようにすることで精度を高め、複数のユーザーをサポートするために簡単にスケールアウトします。これにより、ポッドは、多様で要求の厳しい AI ワークロードを処理するために、スケーラブルで効率的で信頼性の高い AI インフラストラクチャを求める組織にとって、優れた選択肢となります。

表 5. 大規模展開向け Cisco AI スケールアウト推論ポッドの仕様

項目	仕様
コンピューティング ノード	4X Cisco UCS X210c M7 コンピューティング ノード
プロセッサ	コンピューティング ノードごとにデュアル Intel第 5 世代 6548Y+ プロセッサ、または合計 8 個の CPU
メモリ	4800 MT/秒 の 64 GB DDR5 X 64（CPU あたり 8 DIMM）、コンピューティングノードあたり合計 1 TB のメモリ、またはスケール アウト推論ポッドは 4 ノード構成であるため、合計で 4 TB になります。
内蔵ストレージ	コンピューティング ノードあたり 5X 1.6 TB Non-Volatile Memory Express（NVMe）2.5 インチ ドライブ
mLOM	コンピューティング ノードあたり 1X Cisco UCS VIC 15230 2X 100G mLOM
PCIe ノード	コンピューティングノードあたり 1X Cisco UCS X440p PCIe ノード
GPU	コンピューティングノードあたり 2X NVIDIA L40S GPU（デュアルスロット）、または合計 8 個の GPU
管理	• Cisco Intersight ソフトウェア（SaaS、仮想アプライアンス、およびプライベート仮想アプライアンス） • Cisco UCS Manager：UCSM v4.3(2) 以降
Networking	• 2x Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス

システム要件

表 6. システム要件

特長	説明
ディスク領域	AI モデルのストレージと運用のためのコンピューティングノードあたり最小 1 TB SSD
ハードウェア	• 1X Cisco UCS X210C M7 コンピューティング ノード、1X Cisco UCS X440p PCIe ノード、データ センターおよびエッジ推論 POD 用 1X NVIDIA L40S • 2X X210C M7 コンピューティング ノード、2X X440p PCIe ノード、4X RAG 拡張推論 POD 用 Nvidia L40S

© 2025 Cisco and/or its affiliates. All rights reserved.

特長	説明
	• 2X X210C M7 コンピューティング ノード、2X X440p PCIe ノード、4X RAG スケール アップ POD 用 Nvidia H100 • 4X X210C M7 コンピューティング ノード、4X X440p PCIe ノード、8X RAG スケール アウト POD 用 Nvidia L40S
メモリ	コンピューティングノードあたり最小 128 GB の RAM（ワークロード要件に基づいて拡張可能）
ソフトウェア	一元管理用 Cisco Intersight
ネットワーク	ノードとストレージ間のデータ転送用の高速ネットワーク接続（最小 10GbE）

© 2025 Cisco and/or its affiliates. All rights reserved.

特長	説明
電源および冷却装置	高性能コンピューティングノードと GPU をサポートするための適切な電源と冷却システム
オペレーティング システム	Red Hat OpenShift AI for MLOps および AI/ML ワークロードとの互換性 注：選択した POD によって、デフォルト コントローラ ノードをオプトインできます。
GPU のサポート	最適な実績と管理を実現する Nvidia GPU ドライバとソフトウェア（NVIDIA AI Enterprise）
セキュリティ	データ保護のためのセキュアブート、暗号化、およびアクセス制御
統合	既存のデータセンター インフラストラクチャおよびクラウド サービスとのシームレスな統合
シャーシ	Cisco UCS X9508 サーバ シャーシ
ファブリック インターコネクト	Cisco UCS 6454、64108 および 6536 ファブリック
Cisco Intersight	Intersight Managed Mode（サーバーごとに Essentials ライセンス以上）

発注情報

表 7. Cisco AI データ センターおよびエッジ推論ポッドの発注ガイド

全体的な MLB PID：UCSX-AI-EDGE

項目	製品番号
ハードウェア仕様（必須）	1X Cisco UCS X210C M7 コンピューティング ノード • 2X Intel 第 5 世代 6548Y+ • 512 GB システムメモリ • 5X 1.6TB NVME ドライブ
	1X Cisco UCS X440p PCIe ノード • 1 X NVIDIA L40S • X シリーズ FI9108 100G
ソフトウェア仕様（必須）	Cisco Intersight（Essentials） NVIDIA AI Enterprise（Essentials）
デフォルトのコンポーネント（オプション）	OpenShift • OpenShift Container Platform • 単一ノードコントローラ
アドオン	アドオンの完全なブレークアウトは、注文ガイドと CCW に含まれます。

© 2025 Cisco and/or its affiliates. All rights reserved.

表 8. Cisco AI RAG 拡張推論ポッドの発注ガイド

全体的な **MLB PID**：**UCSX-AI-RAG**

項目	製品番号
ハードウェア仕様 （必須）	2X Cisco UCS X210C M7 コンピューティング ノード • Intel 第 5 世代 6548Y+ X 4 • 1 TB システム メモリ • 10X 1.6TB NVME ドライブ
	2X Cisco UCS X440p PCIe ノード • 4X NVIDIA L40S
	2X ファブリック インターコネクト • 6536 100G
ソフトウェア仕様 （必須）	Cisco Intersight（Essentials） NVIDIA AI Enterprise（Essentials）
デフォルトのコンポーネント （オプション）	OpenShift • OpenShift Container Platform • Cisco UCS X210c M7 コントローラ プレーン クラスタ Networking • 2X Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス
アドオン	アドオンの完全なブレークアウトは、注文ガイドと CCW に含まれます。

表 9. Cisco AI Scale Up Inference Pod for High Performance 発注ガイド

全体的な **MLB PID**：**UCSX-AI-LARGERAG**

項目	製品番号
ハードウェア仕様 （必須）	2X Cisco UCS X210C コンピューティング ノード • Intel 第 5 世代 6548Y+ X 4 • 1 TB システム メモリ • 10X 1.6TB NVME ドライブ
	2X Cisco UCS X440p PCIe ノード • NVIDIA H100 NVL X 4
	2X ファブリック インターコネクト • 6536 100G
ソフトウェア仕様 （必須）	Cisco Intersight（Essentials） NVIDIA AI Enterprise（Essentials）

© 2025 Cisco and/or its affiliates. All rights reserved.

項目	製品番号
デフォルトのコンポーネント（オプション）	OpenShift • OpenShift Container Platform • Cisco UCS X210c M7 コントローラ プレーン クラスタ Networking • 2X Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス
アドオン	アドオンの完全なブレークアウトは、注文ガイドと CCW に含まれます。

表 10. 大規模展開向け Cisco AI スケールアウト推論 POD の発注ガイド

全体的な MLB PID：UCSX-AI-LARGEINF

項目	製品番号
ハードウェア仕様（必須）	4X Cisco UCS X210C M7 コンピューティング ノード • Intel 第 5 世代 6548Y+ X 8 • 4 TB システム メモリ • 20X 1.6TB NVME ドライブ
	2X Cisco UCS X440p PCIe ノード • 8X NVIDIA H100 NVL
	2X ファブリック インターコネクト • 6536 100G
ソフトウェア仕様（必須）	Cisco Intersight（Essentials） NVIDIA AI Enterprise（Essentials）
デフォルトのコンポーネント（オプション）	OpenShift • OpenShift Container Platform • Cisco UCS X210c M7 コントローラ プレーン クラスタ Networking • 2x Cisco Nexus スイッチ（93600CD-GX または 9332D-GX2B） • Nexus Dashboard アプライアンス
アドオン	アドオンの完全なブレークアウトは、注文ガイドと CCW に含まれます。

© 2025 Cisco and/or its affiliates. All rights reserved.

保証情報

Cisco UCS X210c M7 コンピューティング ノードには、3 年後の翌営業日（NBD）のハードウェア保証と 90 日間のソフトウェア保証があります。

Cisco Smart Net Total Care® および Cisco Solution Support サービスは、 シスコ テクニカル サービス ポートフォリオの一環として、Cisco Unified Computing System（Cisco UCS）保証を強化するものです。Cisco Smart Net Total Care には、受賞歴のある、業界をリードするシスコの基本的なテクニカルサービスが含まれています。また、Cisco Smart Net Total Care ポータルのスマート機能を通じて、実用的かつ高度なビジネスインテリジェンスも提供されます。詳細については、https://www.cisco.com/c/ja_jp/support/services/smart-net-total-care/index.html を参照してください。

Cisco ソリューション サポートには、Cisco®製品サポートとソリューションのサポートの両方が含まれており、製品サポート単体の場合よりも平均 43% 迅速にマルチベンダー環境の複雑な問題を解決します。Cisco Solution Support は、データセンター管理の重要な要素であり、実績、信頼性、投資回収率を維持しながら、発生した問題の迅速な解決を支援します。

このサービスは、エコシステムに導入された Cisco 製品とソリューション パートナー製品の両方に対応するため、マルチベンダーの Cisco 環境全体でのサポートが一元化されます。Cisco 製品またはソリューション パートナーの製品のどちらに問題がある場合でも、Cisco にご連絡ください。シスコのエキスパートが主な連絡窓口となり、最初のお電話から問題の解決までお客様をサポートします。詳細については、https://www.cisco.com/c/ja_jp/support/services/solutions-support/index.html を参照してください。

製品持続性

Cisco の環境、社会、ガバナンス（ESG）イニシアチブおよび実績に関する情報は、Cisco の CSR および持続可能性レポートで提供されます。

表 11. Cisco の環境保全に関する情報

持続性に関するトピック		参照先
全般	製品の材料に関する法律および規制に関する情報	材料
	製品、バッテリ、パッケージを含む電子廃棄物法規制に関する情報	WEEE 適合性
	製品の回収および再利用プログラムに関する情報	Cisco 回収および再利用プログラム
	持続性に関するお問い合わせ	問い合わせ先： csr_inquiries@cisco.com
材料	製品パッケージの重量と材料	問い合わせ先： environment@cisco.com

© 2025 Cisco and/or its affiliates. All rights reserved.

シスコおよびパートナーの提供サービス

AI 対応インフラストラクチャ向けのCisco とパートナーのサービス

AI インフラストラクチャ テクノロジーを迅速に導入し、それに投資した価値を最大に活かすことでビジネス価値を促進するためには、どうすればいいでしょうか。オンボーディングと導入から次のテクノロジーの推移の最適化まで Cisco と認定パートナーは、AI インフラストラクチャをサポートする包括的なサービスを提供し、シームレスな統合と最適な実績を実現します。

推論のための AI POD 向けのシスコ ソリューション付属サービスは、サブスクリプション主導のアドバイザリおよびコンサルティングサービスを提供し、エンタープライズ環境内で推論タスクを実行するために AI インフラストラクチャおよびモデルを十分に活用します。お客様のニーズに最適なサービスレベル（シスコの専門家による戦略的ガイダンス、お客様のチームとのコラボレーション、お客様のソリューションの導入、活用、最適化を支援するすべてシスコによる支援）を選択していただきます。シスコのサービスでは、次の機能を提供することで、中断を最小限に抑えます：

- Cisco AI POD の実装とオンボーディングを支援するための計画と展開
- AI インフラストラクチャとアプリケーションの導入を支援するアドバイザリ サービス
- サブスクリプション期間中の継続的な最適化は、AI 固有の機能とアーキテクチャの拡張を構成、追加、および強化するのに役立ちます

詳細については、「Cisco Solution Attached Services」の説明をご覧ください。

Cisco Capital

目的達成に役立つ柔軟な支払いソリューション

Cisco Capital® により、目標を達成するための適切な技術を簡単に取得し、ビジネス変革を実現し、競争力を維持できます。総所有コスト（TCO）の削減、資金の節約、成長の促進に役立ちます。Cisco の柔軟な支払いソリューションは 100 か国以上で利用可能であり、ハードウェア、ソフトウェア、サービス、およびサードパーティ製の補完的な機器を、利用しやすい計画的な支払方法で購入できます。詳細はこちらをご覧ください。

詳細情報

Cisco AI ポッドで AI の未来を切り開く

Cisco の AI 対応インフラストラクチャがビジネスにどのように変革をもたらすかをご確認ください。デモをリクエストするか、無料のアセスメントにサインアップするか、仮想営業担当者に連絡して、カスタマイズされたソリューションを確認してください。詳細については、Cisco AI ソリューションの概要ページにアクセスし、ホワイトペーパー、ケーススタディなどを参照してください。生産性と成長を最適化するように設計された最先端のテクノロジーで、ビジネスをリードします。

シスコのエキスパートが推奨

- 第 5 世代 Intel Xeon スケーラブル プロセッサを使用する Cisco UCS X シリーズ M7 ブレード サーバーによる生成 AI 推論の使用例 ホワイト ペーパー
- Cisco UCS X- ファブリック テクノロジー概要
- Cisco UCS X シリーズ ダイレクト FAQ

© 2025 Cisco and/or its affiliates. All rights reserved.

マニュアルの変更履歴

新規トピックまたは改訂されたトピック	説明箇所	日付
Cisco Nexus および **Nexus** ダッシュボードと機能の詳細を追加	優れた機能	2025 年 3 月
2 つの **Cisco Nexus** スイッチ（**93600CD-GX** または **9332D-GX2B**）と **Nexus** ダッシュボード アプライアンスを追加	製品仕様	2025 年 3 月
2 つの **Cisco Nexus** スイッチ（**93600CD-GX** または **9332D-GX2B**）と **Nexus** ダッシュボード アプライアンスを追加	発注情報	2025 年 3 月

米国本社
Cisco Systems, Inc.
カリフォルニア州サンノゼ

アジア太平洋本社
Cisco Systems (USA), Pte. Ltd.
シンガポール

ヨーロッパ本社
Cisco Systems International BV
Amsterdam, The Netherlands

2023 年 11 月発行

© 2023 Cisco and/or its affiliates. All rights reserved.

Cisco および Cisco ロゴは、Cisco Systems, Inc. またはその関連会社の米国およびその他の国における商標または登録商標です。シスコの商標の一覧については、www.cisco.com/jp/go/trademarks をご覧ください。記載されているサードパーティの商標は、それぞれの所有者に帰属します。「パートナー」または「partner」という言葉が使用されていても、シスコと他社の間にパートナーシップ関係が存在することを意味するものではありません。1175152207 10/23

米国で印刷

C78-4662163-03　03/25

© 2025 Cisco and/or its affiliates. All rights reserved.