

Cisco UCS C885A M8

概要

人工知能 (AI) の急速な発展は世界を根本的に変えており、生産性やビジネス効率の向上とサステナビリティ目標の達成など、組織にとって新たな成長の原動力となっています。このような新しい AI イニシアチブのメリットを実現するために、AI ワークロード向けのインフラストラクチャを拡張することがこれまで以上に重要になっています。IT 部門は、こうした新たな要求の厳しいワークロードに対応するために積極的に関与し、データセンター インフラストラクチャを刷新するよう求められています。

AI プロジェクトには、モデルのトレーニング、ファインチューニング、エンドユーザーへのモデルの展開というさまざまなフェーズがあります。インフラストラクチャの要件はフェーズごとに異なります。トレーニングは、最もコンピューティング集約型のフェーズであり、大規模言語モデル (LLM)、ディープラーニング、自然言語処理 (NLP)、デジタルツインには、大幅に高速化されたコンピューティングが必要です。

NVIDIA HGX プラットフォーム上に構築された Cisco Unified Computing System (Cisco UCS) C885A M8 ラックサーバーは、最も要求の厳しい AI ワークロードに対応するために必要な高速コンピューティングを提供します。強力なパフォーマンスとシンプルな導入によって、AI イニシアチブによる成果を達成するまでの期間が短縮されます。



メリット

AI 対応

NVIDIA HGX アーキテクチャ上に構築され、8 基の高性能 GPU を搭載した Cisco® UCS C885A M8 は、最も要求の厳しい AI ワークロードに必要な高速コンピューティングパワーを提供します。

拡張性

Cisco UCS® C885A M8 サーバーのクラスタ全体に AI ワークロードを拡張し、ディープラーニング、大規模言語モデル (LLM) のトレーニング、モデルのファインチューニング、大規模モデルの推論、Retrieval-Augmented Generation (RAG; 検索拡張生成) に対応します。

一貫性のある管理

通常のワークロードを管理する場合と同じツールを使用して AI サーバーを管理することで、AI インフラストラクチャのサイロ化を防止します。

機能

Cisco UCS C885A M8 は、要求の厳しい AI ワークロード向けに設計された高密度 GPU サーバーであり、モデルのトレーニング、ディープラーニング、推論のために強力なパフォーマンスを提供します。NVIDIA HGX プラットフォーム上に構築されているため、スケールアウトしてコンピューティングパワーのクラスタを提供し、きわめて野心的な AI プロジェクトを実現できます。

構成

UCS C885A M8 では、8 基の NVIDIA HGX H100/H200 Tensor コア GPU または 8 基の AMD MI300X OAM GPU を選択できるため、1 台のサーバーで非常に高速なコンピューティング パフォーマンスを実現できます。さらに、GPU ごとに NVIDIA ConnectX-7 NIC または NVIDIA BlueField-3 SuperNIC を 1 つ搭載し、高密度 GPU サーバーのクラスタ全体に AI モデルのトレーニングを拡張できます。各サーバーは、GPU によるデータアクセスを高速化する NVIDIA BlueField-3 DPU も備えています。また、第 4 世代または第 5 世代の AMD EPYC™ プロセッサ 2 基を搭載しています。

管理

Cisco UCS C885A M8 は、クラウドベースの IT 運用プラットフォームである Cisco Intersight® によって管理されます。このため、IT 運用チームは、Cisco UCS® インフラストラクチャの場所にかかわらず、ライフサイクル全体を通じて一元的に確認、制御、自動化できます。

Intersight を使用することで、一貫性と制御性を備えた運用、セキュリティ態勢の強化、エネルギー効率の向上を実現し、イノベーションと成長を促進することができます。

詳細情報

Cisco UCS C885A M8 の詳細については、[データシート](#)を参照してください。

AI 向けのデータセンター ソリューションについては、<https://www.cisco.com/site/jp/ja/solutions/artificial-intelligence/infrastructure/index.html> をご覧ください。