

Cisco Nexus 9000 スイッチを 使用した AI インフラストラク チャ

Cisco リファレンス アーキテクチャ

目次

はじめに	3
ハードウェア	4
ネットワーク トポロジ	6
ストレージ アーキテクチャ	12
ソフトウェア	13
セキュリティ	14
テスト/認定	14
要約	14
付録 A : コンピューティング サーバの仕様	15
付録 B : 制御ノード サーバの仕様	15
参考資料	16

NVIDIA HGX™ H200 および NVIDIA Spectrum™-X を搭載した Cisco UCS C885A コンピューティング サーバを特徴としています

はじめに

この Cisco リファレンス アーキテクチャは、オンプレミスの **Nexus Dashboard** プラットフォームによって管理される ネットワーキング AI クラスタ用の **Cisco Nexus 9000** スイッチに基づいています。これは、**NVIDIA Spectrum™ -X** ネットワーキングを備えた **NVIDIA HGX™H200** 向け **NVIDIA** エンタープライズ リファレンス アーキテクチャに準拠しています。

Cisco Nexus 9000 スイッチは、**Cisco Silicon One** および **Cloud Scale** アーキテクチャを利用して、**AI** および **High-Performance Computing (HPC)** ワークロードに高速で確定的、低遅延、および電力効率の高い接続を提供します。**NX-OS** オペレーティング システムの複数のフォーム ファクタ、光ファイバ、および豊富なソフトウェア機能を使用できるため、**Nexus 9000** スイッチは、バックエンド、フロントエンド、管理、およびストレージ ネットワークに統合されたエクスペリエンスを提供します (図 1 を参照)。

Cisco Nexus Dashboard は、**Nexus 9000** スイッチベースのファブリックを管理するための運用および自動化プラットフォームです。組み込みのテンプレートを使用して構成を簡素化することで、**Nexus 9000** スイッチのデータプレーン機能を補完します。輻輳、ビットエラー、トラフィック バーストなどのネットワークの正常性の問題をリアルタイムで検出し、自動的に異常としてフラグを立てます。これらの問題は、一般的に使用されるツール (**ServiceNow** や **Ansible** など) との統合を使用してより迅速に解決でき、**AI** クラスタのネットワークを組織の既存のワークフローと整合させることができます。

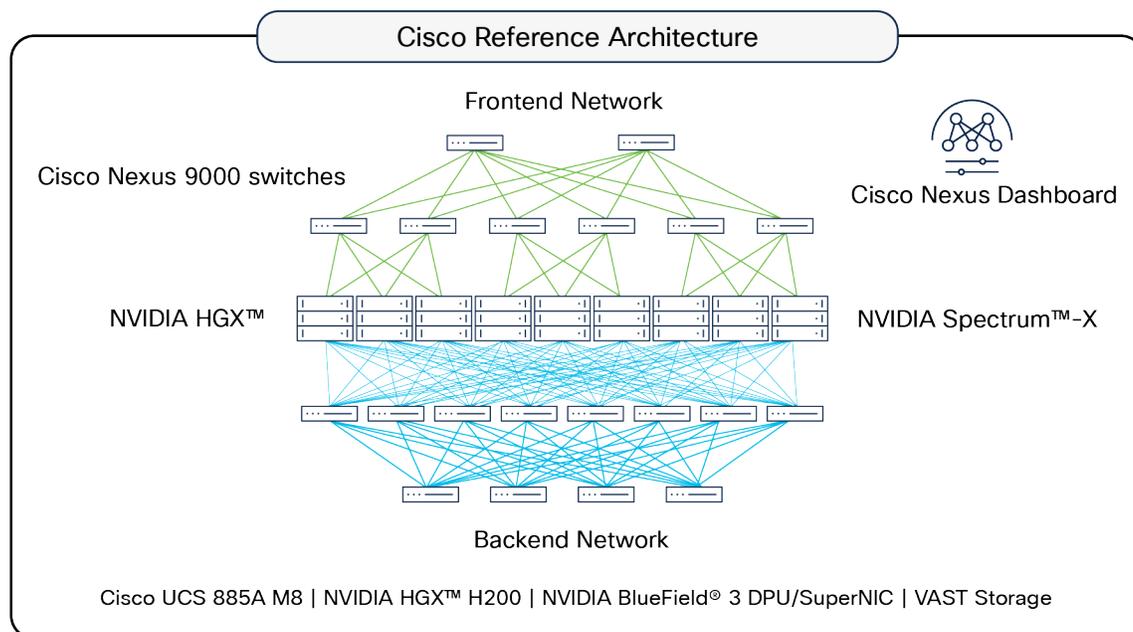


図 1. Nexus Dashboard プラットフォームによって管理される AI クラスタのネットワーク用の Cisco Nexus 9000 スイッチ

ハードウェア

Cisco UCS C885A M8

Cisco UCS C885A M8 ラック サーバは、大規模言語モデル (LLM) トレーニング、ファインチューニング、大規模モデル推論、取得拡張生成 (RAG) などの AI ワークロード向けに、大規模でスケーラブルなパフォーマンスを提供する 8RU 高密度 GPU サーバです。C-G-N-B 命名規則は、2-8-10-40 (C-G-N-B) 内の NVIDIA HGX™ リファレンス アーキテクチャに基づいてが次のように定義されています：

- C : ノード内の CPU の数。
- G : ノード内の GPU の数。
- N : 以下のように分類されているネットワーク アダプタ (NIC) の数：
 - North/South : ノードと外部システムとの間の通信。
 - East/West : クラスタ内の通信。
- B : GPU あたりの平均ネットワーク帯域幅 (ギガビット/秒 (Gb/s) 単位) 。

サーバ内の 8x NVIDIA H200 SXM GPU は、高速 NVLink インターコネクトを使用してインターコネクトされています。East-West トラフィックは、8x NVIDIA BlueField®-3 B3140H SuperNICs、そして、North-South トラフィックは、NVIDIA BlueField®-3 B3240 DPU NICs (1x400G モード内) を使用して、他の物理サーバに GPU 接続します。コンピューティング用に、各サーバには 2 つの AMD EPYC CPU、最大 3 TB の DDR DRAM、30 TB の NVMe ローカルストレージ、およびホットスワップ可能なファントレイと電源が含まれます。サーバの詳細な仕様については、付録 A を参照してください。

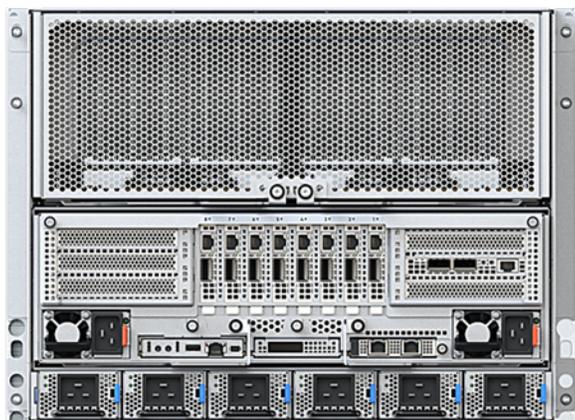


図 2.
NVIDIA HGX™ を搭載した Cisco C885A M8 サーバ

この参照アーキテクチャは、2-8-10-400 (CGNB) 構成の NVIDIA HGX™ に基づく他のサーバに適用されますが、次のセクションでは図のために Cisco UCS C885A M8 を使用します。

Cisco Nexus 93108TC-FX3 スイッチ

Cisco Nexus 93108TC-FX3 スイッチ（図 3 を参照）は、48 個の 100 Mbps または 1/10 Gbps 10GBASE-T ポートと、1RU フォームファクタで 6 個の 1/10/25/40/100 Gbps QSFP28 ポートを提供します。このスイッチは、管理ネットワークで使用できます。

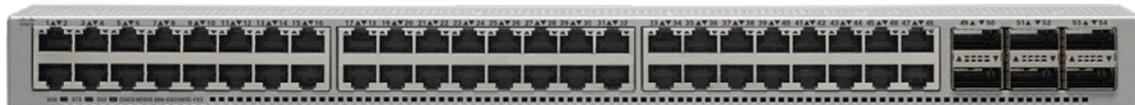


図 3.
Cisco Nexus 93108TC-FX3 スイッチ

Cisco Nexus 9332D-GX2B

Cisco Nexus 9332D-GX2B スイッチ（図 4 を参照）は、1RU フォームファクタで 32 個の 400G QSFP-DD ポートに 10/25/50/100/200 Gbps ブレックアウト サポートを提供します。このスイッチは、リーフまたはスパインロールで使用できます。



図 4.
Cisco Nexus 9332D-GX2B スイッチ

Cisco Nexus 9364E-SG2

Cisco Nexus 9364E-SG2 スイッチ（図 5 を参照）は、QSFP-DD と OSFP の両方のポートタイプで使用可能な 2RU フォームファクタで、64 個の 800G ポートまたは 128 個の 400G ポートを提供します。このスイッチは、リーフまたはスパインロールで使用できます。

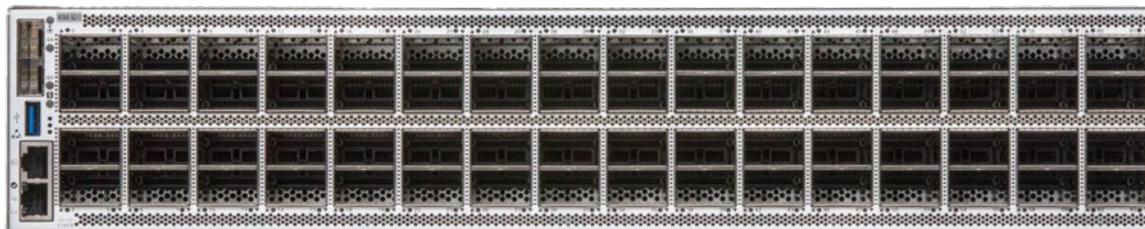


図 5.
Cisco Nexus 9364E-SG2 スイッチ

図 7.
NVIDIA HGX™ (96 GPU) を搭載した 12 台の Cisco C885A サーバ用エンタープライズ RA

表 1 に、NVIDIA HGX™ サーバを使用した 12 ノード Cisco C885A クラスターの部品表 (BOM) を示します。

表 1. NVIDIA HGX™ (96 GPU) を搭載した 12 ノード Cisco C885A クラスターの BOM

PID	説明	数量
UCSC-885A-M8-HC1	NVIDIA HGX™ を搭載した Cisco UCS C885A M8 サーバ	12
N9364E-SG2-O	Cisco Nexus スイッチ、64x800Gbps OSFP	2
N9K-93108TC-FX3	Cisco Nexus スイッチ、48 1/10G BASE-T 6 QSFP28	2
N9K-C9332D-GX2B	Cisco Nexus スイッチ、32x400Gbps QSFP-DD	2
OSFP-800G-DR8	OSFP、800GBASE-DR8、SMF デュアル MPO-12 APC、500m	114
QDD-400G-DR4-S	400G QSFP-DD トランシーバ、400GBASE-DR4、MPO-12、500m パラレル	10
QSFP-400G-DR4	400G QSFP112 トランシーバ、400GBASE-DR4、MPO-12、500 m パラレル	118
QSFP-100G-DR-S	100GBASE DR QSFP トランシーバ、500 m (SMF 使用)	8
CB-M12-4LC-SMF	ケーブル、MPO12-4X デュプレックス LC、ブレイクアウト ケーブル、SMF、各種長さ	2
CB-M12-M12-SMF	MPO-12 ケーブル	204
CAT6A	10G 用銅ケーブル	24
CAT5E	1G 用銅ケーブル	36

800G から 400G への接続では、デュアル 2x400G MPO-12 コネクタで光ファイバを使用できます (図 8 を参照)。



図 8.
Cisco OSFP-800G-DR8

各接続は、ブレイクアウト ケーブルを必要とせずに独立して 400G をサポートします (図 9 を参照)。

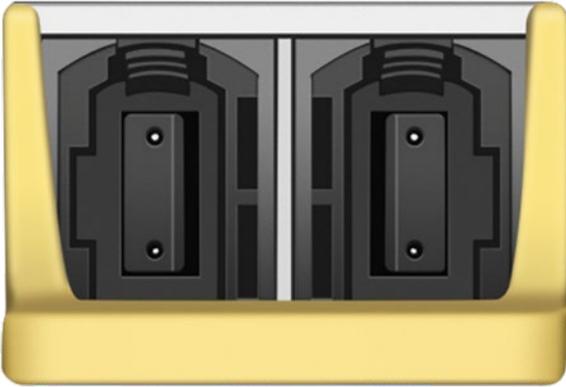


図 9.
Cisco OSFP-800G-DR8 プラグホール ビュー

図 7 に、NVIDIA HGX™を使用した最大 16 の Cisco C885A コンピューティング ノードのクラスタ トポロジを示します。EW ネットワークは、レール 1 ~ 4 が左側の EW スパインに、レール 5 ~ 8 が右側の EW スパインに配置されています。

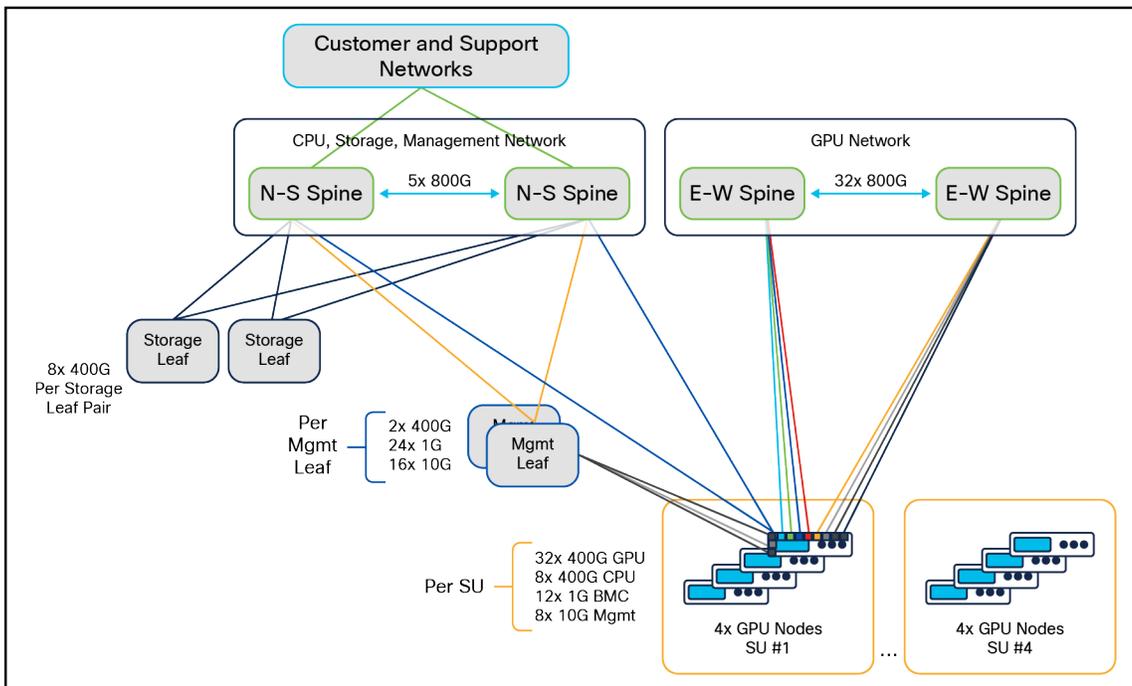


図 10.
NVIDIA HGX™ を搭載した 16 台の Cisco C885A サーバ用エンタープライズ RA (128 GPU)

表 2 に、NVIDIA HGX™ サーバを使用した 16 ノード Cisco C885A クラスタの BOM を示します。

表 2. NVIDIA HGX™ (128 GPU) を搭載した 16 ノード Cisco C885A クラスタの BOM

PID	説明	数量
UCSC-885A-M8-HC1	NVIDIA HGX™ を搭載した Cisco UCS C885A M8 サーバ	16
N9364E-SG2-O	Cisco Nexus スイッチ、64x800Gbps OSFP	4
N9K-93108TC-FX3	Cisco Nexus スイッチ、48 1/10G BASE-T 6 QSFP28	2
N9K-C9332D-GX2B	Cisco Nexus スイッチ、32x400Gbps QSFP-DD	2
OSFP-800G-DR8	OSFP、800GBASE-DR8、SMF デュアル MPO-12 APC、500m	144
QDD-400G-DR4	400G QSFP-DD トランシーバ、400GBASE-DR4、MPO-12、500m パラレル	12
QSFP-400G-DR4	400G QSFP112 トランシーバ、400GBASE-DR4、MPO-12、500 m パラレル	158
QSFP-100G-DR-S	100GBASE DR QSFP トランシーバ、500 m (SMF 使用)	8
CB-M12-4LC-SMF	ケーブル、MPO12-4X デュプレックス LC、ブレイクアウト ケーブル、SMF、各種長さ	2
CB-M12-M12-SMF	MPO-12 ケーブル	198
CAT6A	10G 用銅ケーブル	32
CAT5E	1G 用銅ケーブル	48

クラスタ サイズが 16 を超える場合、East-West コンピューティング ネットワークはスパイン リーフ ファブリックに拡張されます。クラスタ サイズが最大の場合、128 UCS C885A M8 ノードクラスタの図 11 に示すように、north-south ネットワークもスパインリーフになります。E-W ネットワークは、各 EW リーフ 1~8 にある各レール 1~8 とレールに沿っています。

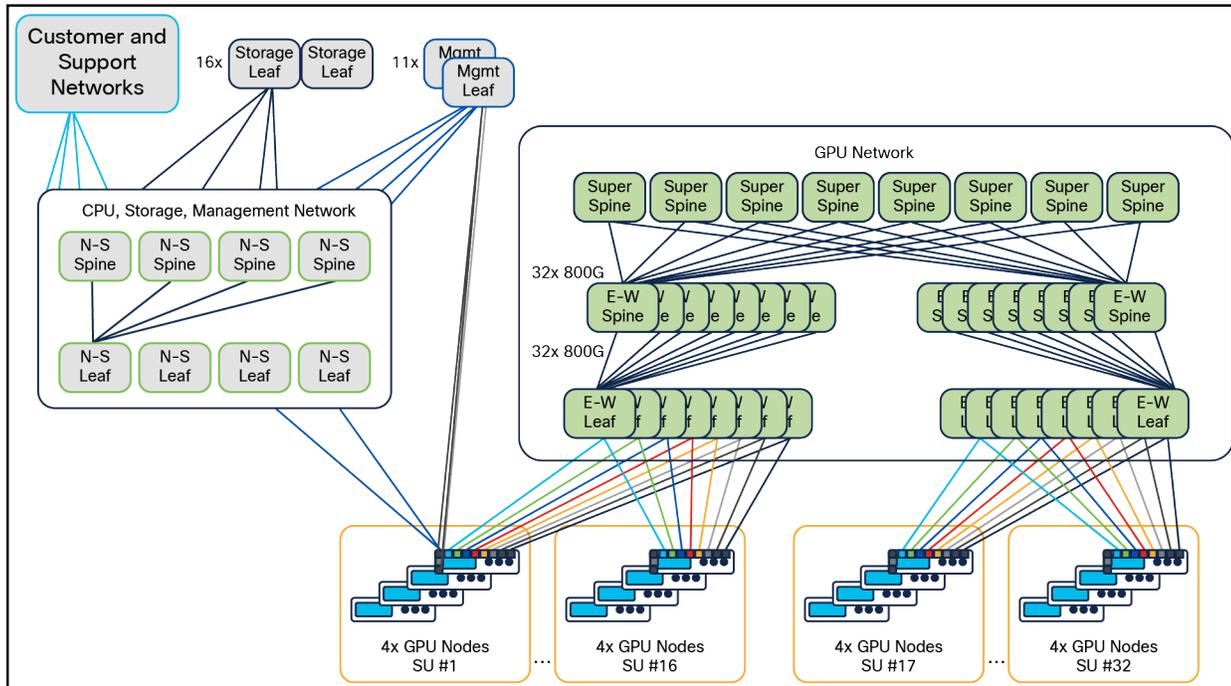


図 11. NVIDIA HGX™ (1024 GPU) を搭載した 128 台の Cisco C885A サーバ用エンタープライズ RA

クラスタ ファブリックのサイジング テーブル C885A でのサイジング

次の 2 つの表は、NVIDIA HGX™ を搭載した Cisco C885A コンピューティング システム、8 EW B3140H NVIDIA BlueField®の3 基の SuperNIC と 2 NS B3240 NVIDIA BlueField®-3 DPU NIC を使用したさまざまなクラスタ サイズに必要なさまざまなユニットの数量を示しています。

表 3. NVIDIA HGX™ サーバ を搭載した Cisco C885A の東西コンピューティング ファブリック テーブル: スイッチ、トランシーバ、ケーブル数

コンピューティング数		スイッチ数			トランシーバ数			ケーブル数	
ノード	GPU	リーフ	スパイン	SuperSpine	ノードからリーフ		スイッチ間 (800G)	ノードからリーフ	スイッチ間
					ノード (400G)	リーフ (800G)			
12	96	2	N/A	なし	96	48	48	96	48
16	128	2	N/A	なし	128	64	64	128	64
32	256	4	2	該当なし	256	128	256	256	256
64	512	8	8	該当なし	512	256	1024	512	1024
128	1024	16	16	8	1024	512	2048	1024	2048

表 4. NVIDIA HGX™ サーバを搭載した Cisco C885A の North-South ファブリック テーブル : スイッチ、トランシーバ、ケーブル数

コンピューティング数		スイッチ数				トランシーバ数								ケーブル数				
ノード	GPU	リーフ	スパイン	管理リーフ	ストレージリーフ	ノードからコンピューティングリーフ		ISL ポート	ノードから管理リーフ (1/10G)		リーフからスパインへの管理		ストレージリーフからスパイン		スパインからお客様およびサポートへのスパイン		SMF MPO-12	CAT6A + CAT5E
						ノード (400G)	リーフ (800G)	800G	ノード	リーフ	リーフ (100G)	スパイン (800G)	リーフ (400G)	スパイン (800G)	お客様 (800G)	サポート (800G)		
12	96	East-West でコンバージド		2	2	24	12	N/A	なし	なし	8	2	8	4	8	4	60	60
16	128	2	該当なし	2	2	32	16	10	N/A	なし	8	2	8	4	8	4	78	80
32	256	2	該当なし	4	4	64	32	16	N/A	なし	16	4	16	8	16	4	144	160
64	512	2	該当なし	7	8	128	64	30	N/A	なし	36	7	32	16	32	4	274	320
128	1024	4	4	14	16	256	128	256	N/A	なし	56	14	64	32	64	4	756	640

光ファイバおよびケーブル

第 1 フェーズでは、次の光ファイバとケーブルがシステム内のさまざまなデバイスで使用されます。

表 5. さまざまなデバイスでサポートされる光ファイバおよびケーブルのリスト

デバイス	光ファイバおよびケーブル
B3140H、B3240	QSFP-400G-DR4 と SMF MPO-12 ケーブル
B3220	QSFP-200G-SR4 と MMF MPO-12 ケーブル
N9364E-SG2-O	OSFP-800G-DR8 (デュアル SMF MPO-12 ケーブル付属)
N9K-C9332D-GX2B	SMF MPO-12 ケーブルを使用した QDD-400G-DR4 QSFP-200G-SR4 と MMF MPO-12 ケーブル
N9K-93108TC-FX3	SMF デュプレックス LC ケーブルを使用した QSFP-100G-DR-S CAT5E ケーブル CAT6A ケーブル

ストレージ アーキテクチャ

シスコは **VAST Data** と協力して、EBox アーキテクチャの **Cisco UCS C225 M8** ラック サーバ上にストレージ ソフトウェアをオンボードし、**Cisco Hyperfabric AI** クラスタのストレージ サブシステムを提供しています。**VAST** データは、サーバを段階的に追加することでストレージ容量と読み取り/書き込みパフォーマンスを水平方向にスケールできる「分散および共有 (DASE) アーキテクチャ」をサポートします。**AI** データ パイプラインのすべてのステージをサポートするために、**NFS、S3、SMB** などのすべてのプロトコル サーバが有効になっています。

次の図は、2つのストレージ リーフを備えた単一の **EBox** のストレージ サーバと **BOM** の全体的なネットワーク接続を示しています。データ パスの場合、各サーバは2つの **NVIDIA BlueField®-3 B3220L 2x200G NIC**を使用します。**NIC0** はサーバ内の内部ネットワークに使用され、他のサーバからストレージ ドライブにアクセスできるようにします。**NIC1** は、**NFS、S3、SMB** などのクライアント トラフィックをサポートする外部ネットワークに使用されます。すべてのサーバがすべてのリーフに接続するため、内部ネットワーク トラフィックはリーフでローカルに切り替えられることに注意してください (スパインには到達しません)。クライアントが外部トラフィックに面する場合、**EBox** ごと、スパイン上の最小要件は **11 x 200G** または **6 x 400G** です。**1G BMC** および **10G x86** 管理ポートは管理リーフ スイッチに接続されます。

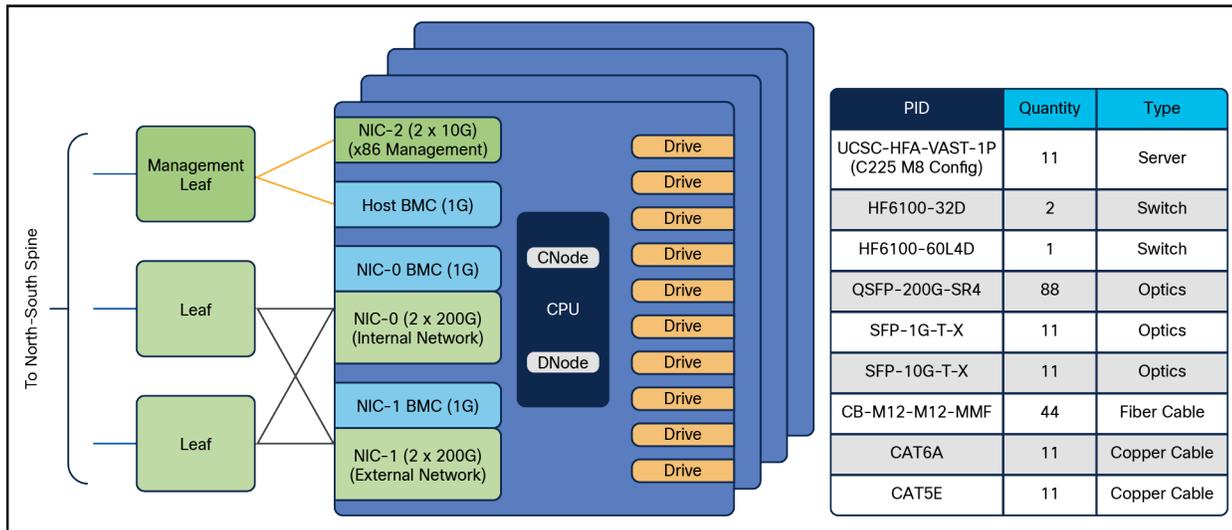


図 12. ストレージ サブシステムのブロック図と BOM

クラスタ サイズが大きくなると、ストレージ リーフ スイッチと **Ebox** の数は、クラスタのサイジング テーブルにあるとおりに直線的に増加します。

VAST データに加えて、他の **NVIDIA** 認定ストレージ パートナーもこのリファレンス アーキテクチャで使用できます。

ソフトウェア

NX-OS と Nexus Dashboard

前述のように、NX-OS は Nexus 9000 スイッチで実行されるオペレーティング システムですが、Nexus Dashboard はファブリックを管理するための運用および自動化プラットフォームです。

NX-OS および Nexus Dashboard では、次のことは行いません。

- 何らかの方法でコンピューティングまたはストレージを構成する
- サーバの BMC またはホスト CPU ソフトウェア ライフサイクルを管理する
- NVIDIA BlueField®-3 NIC でカーネルおよびディストリビューションを管理する

サーバの構成および管理機能は、他の方法で管理する必要があります (Intersight は Cisco のオプションです)。これらのツールの展開および使用は、お客様側で全責任を負います。ネットワーク コントローラの適切な範囲であることに加えて、この懸念事項の分離は、ネットワーク運用をコンピューティングとストレージからセグメント化するという主要な運用パラダイムに合致しています。

NVIDIA AI Enterprise

このリファレンス アーキテクチャには、NVIDIA 認定の Cisco UCS C885A M8 サーバまたは別の NVIDIA HGX プラットフォームで展開およびサポートされている NVIDIA AI Enterprise が含まれています。NVIDIA AI Enterprise は、実稼働対応の AI エージェント、生成 AI、コンピュータ ビジョン、音声 AI などの開発と展開を合理化するクラウドネイティブなソフトウェア プラットフォームです。エンタープライズ レベルのセキュリティ、サポート、および API の安定性により、プロトタイプから実稼働へのスムーズな移行が保証されます。

NVIDIA NIM™ マイクロサービスは、オープンソース コミュニティ モデル、カスタム モデル、および NVIDIA AI Foundation モデルの実稼働展開のための完全な推論スタックを提供します。スケーラブルで最適化された推論エンジンと使いやすさにより、モデルが加速し、TCO が改善され、実稼働展開が迅速化されます。

コンピューティング サーバ スタック

クラスタ ソリューション全体が、Ubuntu Linux 22.04 LTS および NVIDIA Cloud Native Stack (CNS) バージョン 12.3 を実行しているコンピューティング ノードで検証されています。これには、Kubernetes (K8) 環境内の互換性のあるドライバ、GPU、およびネットワーク オペレータが含まれます。Slurm バージョン 24.11.1 は、ワークロード オーケストレーション エンジンとして検証されています。NVIDIA® NGC™ カタログの下にあるコンテナは、Kubernetes と Slurm の両方で起動できます。

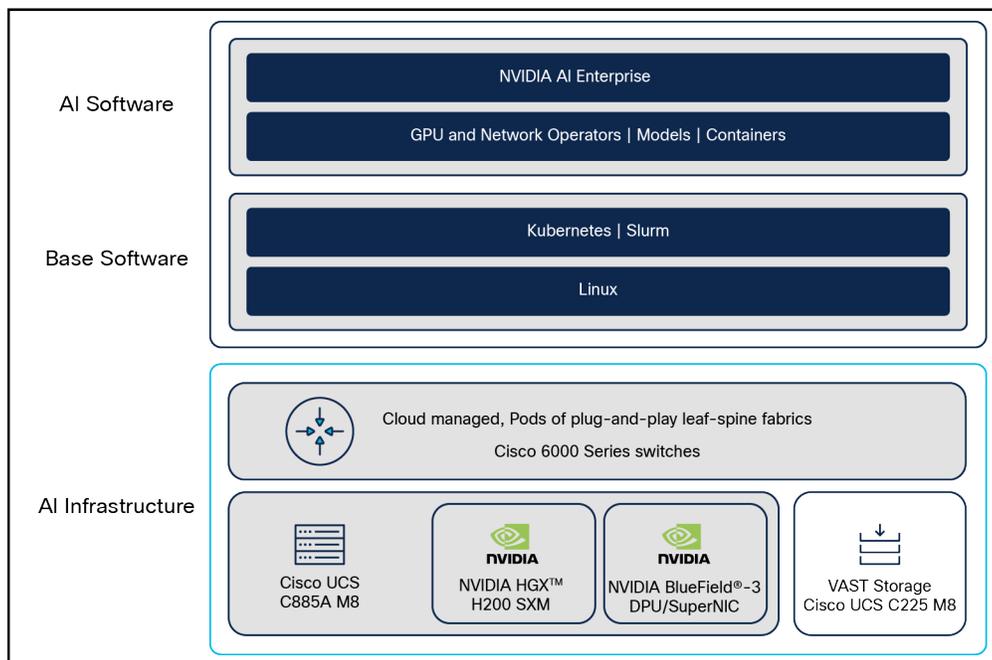


図 13. コンピューティング サーバ ソフトウェア スタック

お客様は、NVIDIA が公開している NVIDIA AI Enterprise、ドライバ、および CNS 互換性マトリックスに従って、OS ディストリビューションと SW バージョンを選択して実行できます。

セキュリティ

Cisco のネットワーク セキュリティおよびオペラビリティ サービスは、必要に応じて、クラスタのさまざまな HW (スイッチ、ホスト、NIC) および SW コンポーネントに統合できます。

テスト/認定

管理プレーン、コントロールプレーン、およびコンピューティング、ストレージ、ネットワークを組み合わせたデータプレーンのあらゆる側面を考慮し、全体的なソリューションを徹底的にテストしています。また、HPC Benchmark、IB PerfTest、NCCL Test、MLCommons Training、Inference ベンチマークなど、多数のベンチマークテストスイートも実行され、パフォーマンスを評価し、調整を支援しています。NVIDIA AI エンタープライズエコシステムのさまざまな要素とエンティティが投入およびテストされ、微調整、推論、RAG に関する多数の企業中心のお客様のユースケースを評価します。ネットワークを使用したシングルノードおよびマルチノードの両方で NVIDIA 認定システム™ のテストスイートバージョン 3.5 を実行した結果、C885A は合格しました。

要約

Cisco Nexus 9000 スイッチと Nexus Dashboard プラットフォームは、NVIDIA アクセラレーションコンピューティングを利用した AI インフラストラクチャに、スケーラブルで管理が容易な高パフォーマンスのネットワークを提供します。

付録 A : コンピューティング サーバの仕様

Cisco UCS C885A M8

表 6. Cisco UCS C885A M8 8RU ラック サーバ

エリア	詳細
フォームファクタ	8RU ラック サーバ (空冷)
コンピューティング + メモリ	第 5 世代 AMD EPYC 9575F X 2 (400W、64 コア、最大 5GHz) 24x 96GB DDR5 RDIMM、最大 6,000 MT/S (推奨メモリ構成) 24x 128GB DDR5 RDIMM、最大 6,000 MT/S (サポートされる最大メモリ構成)
ストレージ	RAID サポート付きデュアル 1 TB M.2 NVMe (ブート デバイス) 最大 16 台の PCIe5 x4 2.5 インチ U.2 1.92 TB NVMe SSD (データキャッシュ)
GPU	8 X NVIDIA HGX™ H200 (各 700W)
ネットワーク カード	8 PCIe x16 HHHH NVIDIA BlueField®-3 B3140H East-West NIC 2 PCIe x16 FHHL NVIDIA BlueField®-3 B3240 North-South NIC 1 つのホスト管理用の OCP 3.0 X710-T2L
冷却	システム冷却用の 16 ホットスワップ可能 (N+1) ファン
前面 IO	2xUSB 2.0、1xID ボタン、1x電源ボタン
背面 IO	1x USB 3.0 A、1x USB 3.0 C、mDP、1x ID ボタン、1x 電源ボタン、1x USB 2.0 C、1x RJ45
電源モジュール	6 X 54V 3kW MCRPS (4+2 冗長性) および 2 X 12V 2.7kW CRPS (1+1 冗長性)

付録 B : 制御ノード サーバの仕様

汎用性の高い Cisco UCS C225 M8 1RU ラック サーバは、サポートサーバ、Slurm および Kubernetes (K8) などの制御ノードサーバとして使用できます。以下は、サーバの最小仕様です。

表 7. Cisco UCS C225 M8 1RU ラック サーバ

エリア	詳細
フォームファクタ	1RU ラック サーバ (空冷)
コンピューティング + メモリ	1x第 4 世代 AMD EPYC 9454P (48 コア) 32GB DDR5 RDIMM X 12 (4800MT/s)
ストレージ	RAID 搭載デュアル 1 TB M.2 SATA SSD (ブート デバイス) 最大 10 台の 2.5 インチ PCIe Gen4 NVMe PCIe SSD (それぞれ容量 1.9 ~ 15.3 TB) - オプション

エリア	詳細
ネットワーク カード	1 PCIe x16 FHHL NVIDIA BlueField®-3 B3220L (DPU モードで構成) または 1 PCIe x16 FHHL NVIDIA BlueField®-3 B3140H (DPU モードで構成) x86 ホスト管理用の 1 OCP 3.0 X710-T2L (2 x 10G RJ45)
冷却	システム冷却用の 8 ホットスワップ可能 (N+1) ファン
電源モジュール	2x 1.2KW MCRP PSU (N+1 冗長構成)
BMC	ホスト管理用の 1G RJ45

2 ソケット CPU を使用する展開では、B3220 DPU NIC とともに UCS C245 M8 2RU バリエーションを使用できます。

参考資料

- [NVIDIA AI Enterprise ソフトウェア リファレンス アーキテクチャ](#)。
- [NVIDIA HGX](#)。
- [NVIDIA Spectrum-X ネットワーキング](#)。
- [Cisco Nexus 9000 スイッチ](#)。
- [Cisco Nexus Dashboard](#)。

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/go/ciscologotrademarks 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

