DELL Technologies

AMD
together we advance_

2024年 | ホワイト ペーパー

AI で企業を支援：選択の時代に突入

目次

DELL Technologies

AMD
together we advance_

2024年 | ホワイト ペーパー

AIでさまざまな業界を変革する機会

今、AI によって未来のイノベーションに向けてビジネスを変革する絶好の機会が訪れています。Accenture Technology Vision 2023 の収集データによると、世界の経営幹部の 98% が、今後 3 年から 5 年の組織の戦略において、AI 基盤モデルが重要な役割を果たすことに同意しています。[1]

AI は、タスクの効率を高め、イノベーションを推進し、意思決定プロセスを改善する能力があるため、小売、医療、金融サービスなどの分野の企業にとって非常に役立つものになりました。しかし、このようなメリットがある一方で、AI の統合に関しては次のような一般的な誤解により、依然として参入障壁があると認識されています。

AI の利用を開始するには AI 開発者のチームが必要：

高度な AI ソリューションを開発し、その基盤となる原則を理解する上で、データサイエンスの専門知識は依然として有益ですが、もはや前提条件ではありません。ユーザーフレンドリーな AI ツール、Hugging Face などのプラットフォーム、AI ソリューションの開発に伴う複雑さの多くを排除するタスク固有のモデルが急増しています。

結果を得るにはハードウェアに数千万ドルの出費が必要：

この誤解は、現在利用可能な AI リソースの多様性を著しく見誤っています。これらの一般的に知られているリソースは、多くの場合強力でサポートも十分ですが、すべてのビジネスにとって常に最も適した、または費用対効果の高い選択肢であるとは限りません。

アクセラレーターを取得するにはたゆまぬ努力が必要：

アクセラレーターは負荷の高い AI ワークロードに対して優れた性能を発揮しますが、企業は AI アプリケーションにそれほど多くのコンピューティング能力を必要としない可能性があります。市場をリードするアクセラレーターを利用できるようになるまで非常に長い期間待つというのも、現実的ではありません。多くの場合、AI 向けに最適化された CPU は、AI 支援による分析と意思決定をリアルタイムで実行するために必要なパフォーマンスと効率性を実際に提供することができ、費用対効果と適応性がはるかに高いソリューションです。

1 Accenture、『Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds』（2023 年 3 月 30 日）https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

幸いなことに、AI ランドスケープは進化しつつあります。**Dell** と **AMD** は提携により、こうした迷信を打ち破っています。今日の AI ニーズに対応するよう設計されたエンドツーエンドのインフラストラクチャを利用して、より幅広いユーザーが AI テクノロジーやツールを利用できるようにしています。

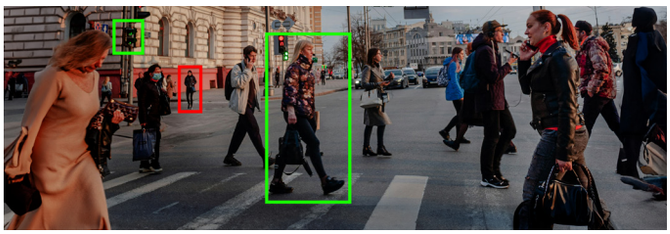
事前に最適化されたモデル、信頼性の高いソフトウェア スタック、汎用性に優れたハードウェア システムを使用して、AI の活用を始めることができます。これらはすべて、Dell と AMD のパートナーシップによって誰でも利用できます。ますます希少になりつつあるアクセラレーターや熟練した AI エンジニアの大規模なグループ、巨大なクラウド クラスターを導入するためのリソースを利用するといったことは、もはや AI を活用するための要件ではありません。

Dell と **AMD** のコラボレーションが、ハードウェアとソフトウェアの統合エコシステムを提供します。開発者は、転移学習、微調整、推論を組み込んだエンドツーエンドの AI ソリューションを簡単かつ効率的に作成できます。**Hugging Face** のサポートにより、AMD EPYC ™プロセッサーまたは AMD Instinct ™ MI300X アクセラレーターを搭載した Dell PowerEdge サーバーで実行されるモデルのポートフォリオが拡大し、開発者は微調整、転移学習の適用、推論のための展開を行うことができます。AMD ROCm ™および AMD ZenDNN ™への投資、および PyTorch、TensorFlow、ONNX Runtime フレームワークとの連携は、応用 AI 開発者が AI の民主化を経験する基本的な要素です。次のスタック図は、Dell と AMD の統合 AI エコシステムを構成するコンポーネントの詳細を示しています。

業界におけるAIの活用

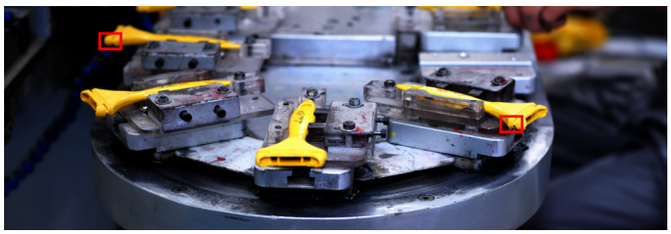
リソースの多様化とオープンソース イノベーションの重視により、AI はカスタマー サービス、金融、銀行、医療、小売など、さまざまな業界に広がっています。こうした業界全体において、AI は、データ分析、自動化、パーソナライゼーション、予測分析といった主要な機能に対応することにより、組織が自社独自のデータの潜在能力を引き出して AI ワークフローを再考することを可能にします。AMD ROCm および ZenDNN ライブラリーが、さらにこれらの AI ワークフローを高速化し、ほぼリアルタイムで結果を提供します。

AI がさまざまな業界にどのような影響を与えるかについて、以下で詳しく見ていきましょう。



自動車

AIは、自律走行車の物体検出、車線追跡、意思決定に使用されます。また、AI は車両部品が故障する可能性が高い時期を予測することもできるため、プロアクティブなメンテナンスとダウンタイムの削減が可能になります。



製造業および工業

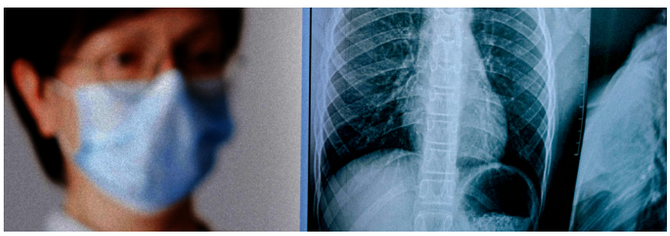
AIは、製造業や工業で予知保全、品質管理、プロセスの最適化、サプライ チェーン マネジメントに使用でき、効率性の向上とダウンタイムの削減につながります。

小売

AIは、顧客の行動を分析してパーソナライズされた製品の推奨を提供し、顧客エンゲージメントと売上を向上させることができます。また、需要を予測し、過剰在庫や在庫切れを最小限に抑えることで、在庫レベルの最適化も可能になります。

金融サービス

AIは、金融および銀行取引の不正検出、リスク アセスメント、カスタマー サービス、投資分析に使用でき、セキュリティの向上と、より多くの情報に基づいた意思決定につながります。



医療

AIは、医用画像解析、疾患診断、個々の患者に合わせた治療計画、創薬など、医療のさまざまな用途に使用でき、患者の転帰の改善とコスト削減につながります。

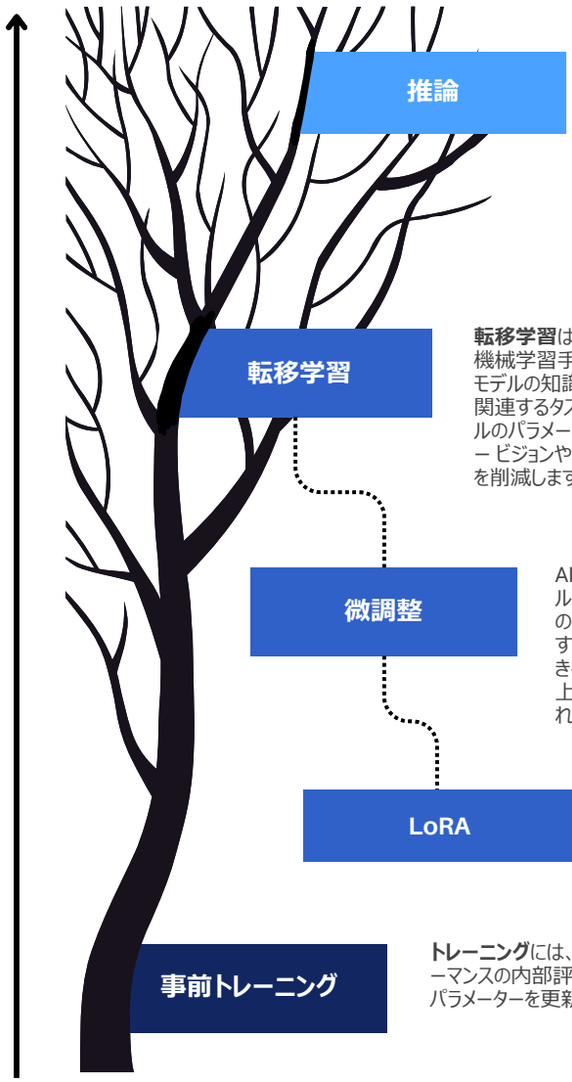
サービスの自動化

AI搭載のチャットボットは、顧客からの問い合わせに対応してサポートを提供できるため、人間の介入の必要性が軽減されます。また、AIはデータ入力やドキュメント処理などの反復作業を自動化し、効率を向上させ、エラーを減らすことができます。

IT導入決定者が考慮すべきこと

開始ステップ：AI の分解

ユース ケースを確認する前に、AI のライフサイクルについて詳しく見ていきましょう。AI（人工知能）ライフサイクルとは、AI システムの開発、導入、保守に関連する段階を指します。具体的な方法論や用語はさまざまですが、一般的な AI ライフサイクルには常にモデルのトレーニングと推論が含まれます。



推論とは、トレーニング済みのAIモデルを使用し、新しい未見データに基づいて予測、分類、意思決定を行うプロセスです。AIライフサイクルのこのステージでは、トレーニング済みのモデルを本番環境に導入して使用し、設計された特定のタスクを実行します。推論中、モデルは入力データを受け取り、トレーニング フェーズから学習したパターンと関係を使用して処理し、生成AIワークロードの予測、分類、テキストなどの出力を生成します。この出力は、ユーザーへの製品の推奨、システム内の異常の検出、画像内のオブジェクトの認識など、さまざまなアプリケーションに使用できます。自然言語処理では、取得拡張生成(RAG)によってモデルが外部のナレッジ ソースにアクセスして活用できるようにすることで、モデルを導入する際の精度を高めます。これにより、モデルが幅広いコンテキストで自然言語入力をより良く理解し、応答できるようになります。

転移学習は、再トレーニングとも呼ばれ、1つのタスクでトレーニング済みのモデルを2番目の関連タスクで再利用する機械学習手法です。一般的な考え方は、使用可能なラベル付きトレーニング データを多数含むタスクから学習されたモデルの知識を、データがあまりない新しいタスクで使用するということです。学習プロセスをゼロから始めるのではなく、関連するタスクを解決することで学習したパターンから始めます。つまり、その関連するタスク用にトレーニングされたモデルのパラメーターを、新しいタスクにより適合するよう効果的に更新するということです。転移学習は、主にコンピューター ビジョンや、感情分析などの自然言語処理タスクで使用され、これらのタスクに通常必要とされる膨大な計算能力を削減します。

AIライフサイクルの**微調整**とは、新しいデータセットまたはタスクであらかじめトレーニング済みの機械学習モデルをさらにトレーニングするプロセスを指します。この手法は転移学習で一般的に使用され、特定のタスク用の大規模なデータセットでトレーニングしたモデルを、関連するタスクまたは小規模なデータセットに適応させます。微調整を行うと、あらかじめトレーニング済みのモデルから新しいタスクまたはデータセットに知識を移転できるため、多くの場合、ゼロからモデルをトレーニングする場合と比較して融合が速くなり、パフォーマンスが向上します。コンピューター ビジョン、自然言語処理、音声認識など、さまざまなAIアプリケーションで広く使用されています。

LoRA (Low Rank Adaptation)は、ディープ ラーニング モデルを微調整するための新しい手法であり、トレーニング可能なパラメーターの数を減らして、タスクの効率的な切り替えを可能にすることで、最終的に微調整プロセスのコンピューティングの負荷を軽減します。これにより、広く利用可能なハードウェアでアクセスしやすくなります。

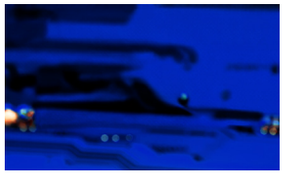
トレーニングには、適切なアルゴリズムまたはモデル アーキテクチャの選択、モデル パラメーターのチューニング、モデルのパフォーマンスの内部評価が含まれます。トレーニング対象のモデルのサイズによっては、トレーニング プロセスの各イテレーションでパラメーターを更新すると、データとコンピューティングに非常に負荷がかかるプロセスになる可能性があります。

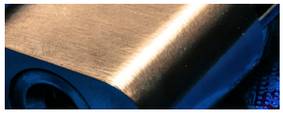
図1：AIのライフサイクル

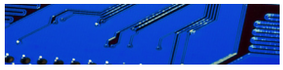
重要な選択肢

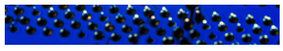
| パフォーマンス

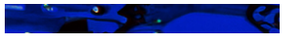
実際のアプリケーションの多くでは、リアルタイムまたはほぼリアルタイムの意思決定が成功に不可欠です。たとえば、金銭的損失を防止し、ビジネス資産を保護するためには、金融取引や保険金請求における不正行為を迅速に特定する必要があります。製造のシナリオでは、品質保証のために組み立てラインまたは工場の状態の欠陥を動的に監視する必要があります。推論ワークロードを処理するプロセッサーは実質的に、受信データ ストリームを迅速かつ効率的に処理できるよう最適化する必要があります。Dell PowerEdge サーバーと AMD EPYC プロセッサーは、汎用性に優れた組み合わせであり、エッジでの推論ワークロードの処理や、ハイパフォーマンス コンピューティング、クラウド コンピューティング、ビッグデータ分析関連のタスクに最適です。











| データ セキュリティ

データ セキュリティは、AI システム、特に生成 AI を活用したシステムの成功に不可欠であり、AI を業務に組み込むことを目指すテクノロジー リーダーにとって重要な関心事です。AI システムは通常、個人情報、財務データ、専有情報などの機密情報を含む可能性がある大量のデータに依存しています。このデータを保護することは、不正アクセスやデータ盗難を防止し、AI モデルと予測の精度、信頼性、一貫性を確保するためにも重要です。

コンフィデンシャル コンピューティングは、安全なエンクレーブでデータ処理を行い、クラウド プロバイダーやその他のユーザーを含む未承認の関係者による不正アクセスや操作からデータを保護するテクノロジーです。[2] 暗号化およびその他のセキュリティ対策を使用して、処理中にデータを分離します。AMD EPYC プロセッサーに統合された高度なセキュリティ機能セットである AMD Infinity Guard は、プロセッサーのみが認識しているキーを使用して仮想マシン (VM) を暗号化する Secure Encrypted Virtualization (SEV) を採用することで、コンフィデンシャル コンピューティングをサポートします。これらのサービスは、AMD SEV-Secure Nested Paging (SEV-SNP) を使用してハードウェア ベースの信頼できる実行環境を提供することを目的としています。これによってゲスト保護が強化され、外部の脅威に対する防御に役立ちます。

フェデレーション学習は、データ セキュリティを維持するためのもう 1 つの方法です。分散化されたデバイスまたはサーバー間で中央モデルをトレーニングします。[3] つまり、すべてのデータを中央の場所に転送するのではなく、各デバイスがローカルでモデルをトレーニングし、モデルの更新のみが共有されます。このアプローチにより、プライバシーが保護され、生のデータを共有することなく共同学習が可能になります。デル・テクノロジーズのフェデレーション AI プラットフォームを使用すると、収集されたデータセットに対して計算プロセス、AI、ML アルゴリズムを実行でき、ネットワーク経由で他のエッジ デバイス、データセンター、またはクラウドに、数学モデル、メタデータ、クエリ結果のみを共有できます。この交換は、データや知的財産を明らかにすることなく、大規模な分散データセットから実用的なインサイトをほぼリアルタイムで抽出できることから、結果を向上させます。

2 Advanced Micro Devices, Inc.、『AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings』（2023 年 8 月 30 日）、https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc.、『Data Center Solutions, Confidential Computing』（2021 年）ソリューション概要、https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

3 Analytics Vidhya、『Federated Learning: A Beginner's Guide』（2023 年 12 月）、https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
デル・テクノロジーズ、『A federated learning platform for real-time artificial intelligence』（2021 年）ソリューション概要、https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

ソリューションの拡張

| コストとイノベーションのバランス

コストとイノベーションの適切なバランスを取ることで、AIソリューションが経済的に実現可能になるだけでなく、企業とユーザーの両方に真の価値をもたらすことができます。このバランスを見つけるための重要な要素は、ユース　ケースを解決すると同時に、既存のインフラストラクチャに簡単に統合できるハードウェアを特定することです。最新のAIハードウェア市場では、生産能力の制約、物流の課題、半導体不足に加えて、さまざまな業界でのアクセラレーターの需要の高まりがその不足の一因となっています。

ただし、すでにほとんどのデータセンターではCPUが標準コンポーネントになっています。これは、まったく新しいアクセラレーター ハードウェアを追加する場合と比較して、統合がシンプルでコスト パフォーマンスに優れています。AI向けに最適化されたCPUは、既存のソフトウェアとツールを活用できるため、大規模なツールの入れ替えや再トレーニングの必要性が軽減されます。また、CPUは、AI以外の幅広いタスクに対しても優れた柔軟性と効率性を提供するため、データセンター内でリソースをより汎用的に使用できるようになります。AMD　EPYCプロセッサーを搭載したDell PowerEdgeサーバーでデータセンターを更新することで、既存のワークロードに対応しながら、AIによって推進されるさらなるイノベーションと効率化に向けた進歩に備えることができます。

| シンプルさと柔軟性

AI システムのシンプルさと柔軟性は、 長期的に効果、 適応性、 拡張性に優れた AI ソリューションを構築するために不可欠です。 ハードウェアを補完する一連のソフトウェア フレームワークと最適化機能にアクセスできるため、 クロスプラットフォームの統合に余分な時間と労力を費やすことなく、 パフォーマンスを向上させることができます。 これらの特性は、 トレーニング、 推論、 データ処理など、 さまざまなタイプの AI タスクが組み合わさった混在 AI ワークロードに取り組む場合に特に重要です。

AMD とデル・テクノロジーズは、 ハードウェアとソフトウェアのソリューションを組み合わせて、 混在 AI ワークロードに取り組んでいます。 AMD EPYC プロセッサーは、 同時マルチスレッディング (SMT) や多数のコアなどの機能を備えたハイパフォーマンス コンピューティング能力を提供し、 AI ワークロードの効率的な並列処理を可能にします。 これらのプロセッサーは AI タスク向けに最適化されており、 トレーニングと推論の両方のワークロードに対して強力なパフォーマンスを発揮します。 AMD EPYC プロセッサーを搭載した Dell PowerEdge サーバーは、 AI ワークロードを導入するための拡張性と柔軟性に優れたプラットフォームを提供します。 さらに、 Dell OpenManage Software スイートは、 混在 AI ワークロードのリソース割り当てとパフォーマンス監視を最適化する管理ツールを提供します。

また、 AMD は Unified Inference Frontend (UIF) も提供しています。 これは現在の各ソフトウェア スタックのパフォーマンス向上バージョンを利用し、 AMD EPYC プロセッサー用の AMD ZenDNN ライブラリー、 AMD Instinct アクセラレーター用のオープンソース AMD ROCm スタック、 および AMD アダプティブ SoC 用のソフトウェア スタックを活用します。 AMD ROCm は、 プロフェッショナル クラスとコンシューマー クラスの両方の製品を含む、 幅広い AMD CPU およびアクセラレーターと連携するようにも設計されています。

| 説明可能性の確保

説明可能なAIは、人工知能アプリケーションにおける透明性、信頼性、有効性を確保する上で極めて重要な役割を果たします。説明可能なAIは、AIモデルがどのように意思決定を行うかについてのインサイトを提供し、根拠と推論プロセスを明らかにします。この透明性は、特に意思決定が個人の生活に直接影響する医療、金融、刑事司法などの機密を伴う領域において、ステークホルダーの信頼を得るために不可欠です。

ヒューマン イン ザ ループAIシステムは、人間の知能を活用してAIのパフォーマンスを向上させ、アルゴリズムのバイアスを軽減します。人間による監視を統合することで、これらのシステムは複雑で曖昧な状況をより効果的に処理できるようになり、AIソリューションが倫理的および社会的規範に合致することが保証されます。さらに、人間の関与により、実際のフィードバックに基づいてAIモデルの継続的な改良と適応が可能になり、反復的な改善と長期的な信頼性が促進されます。これらのアプローチは、社会の最善の利益に役立ち、信頼性を備え、説明可能で、かつ包括的なAIシステムを構築するために不可欠です。

実際のシナリオ

Scalers AI は Dell および AMD と共同で、AMD プロセッサーを搭載した Dell PowerEdge サーバーの機能を紹介しています。これらのテクノロジーが、小売業や医療機関のシナリオでのトレーニング、転移学習、推論にどのように活用されているかをご覧ください。

小売

Scalers AI は、物体検出 AI モデルの実装を通じて小売店の棚の在庫レベルを監視および管理するよう設計されたシステム、Retail Inventory Management Reference Solution を構築しました。このリファレンス ソリューションは、SSD_MobileNet_V2 モデルを活用して店舗の棚にある製品を識別および認識し、最終的に自動在庫カウントと在庫レベルの正確な監視を実現します。このモデルは、Roboflow の 23,000 枚の画像で構成される SKU110K 画像データセットを使用して、転移学習を行いました。コンピューター ビジョンと機械学習アルゴリズムを活用することで、商品の残量が少なくなったり品切れになったりしたことを検出し、店舗の担当者にアラートを送信します。これによってタイムリーな在庫補充が可能になります。

このソリューションは、AMD EPYC 9354P 32 コア プロセッサーを搭載した Dell PowerEdge R7615 サーバーを利用しています。

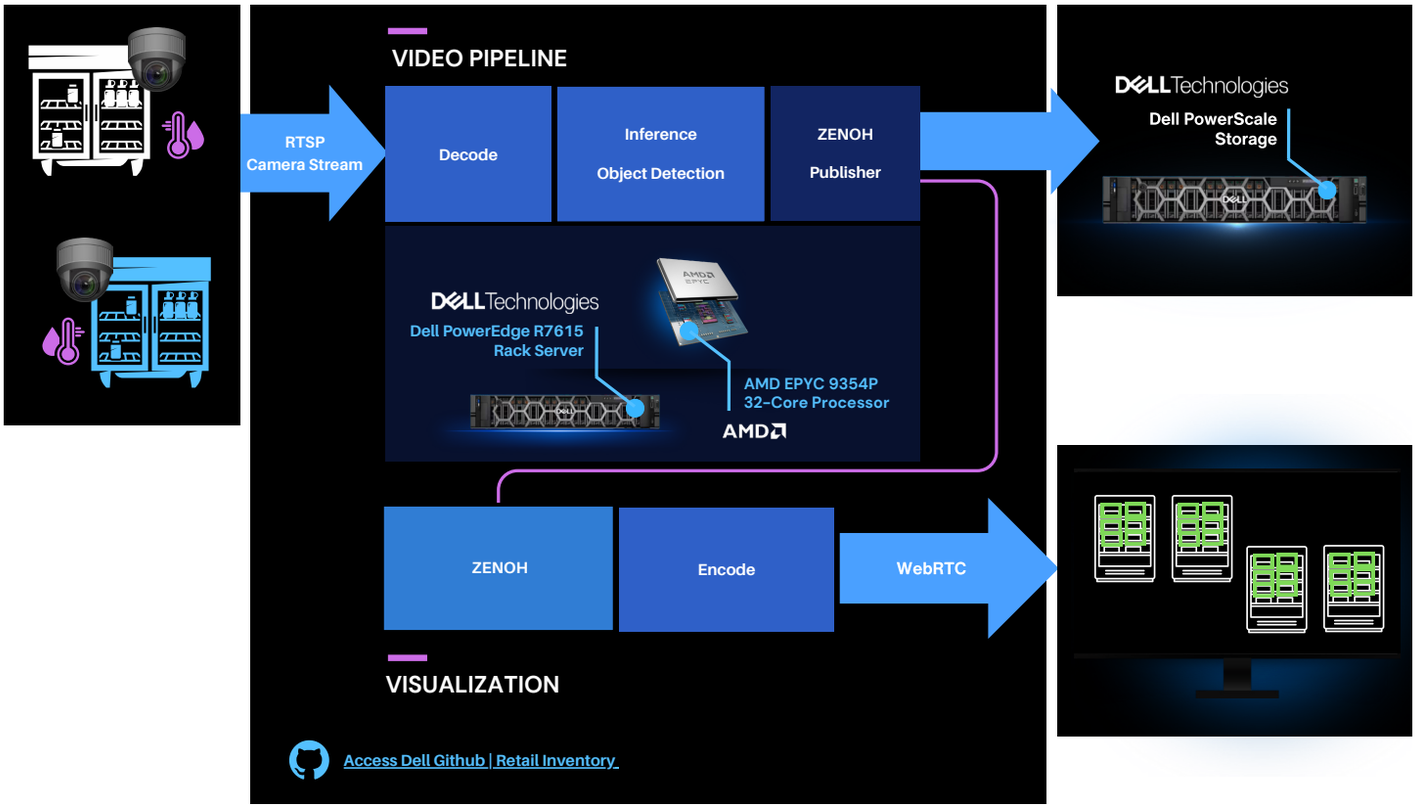


図 2：Retail Inventory Management Reference Solution のアーキテクチャ図

医療

AI を活用した医用画像は、診断の精度と効率を向上させ、肉眼では検出が困難な状態に関する正確なインサイトを医療従事者に提供し、医療を強化する能力において非常に価値があります。AI は、医用画像の分析を自動化することで、診断に必要な時間を短縮し、治療の意思決定を迅速化し、最終的に患者の転帰を改善します。

Scalers AI は、AMD EPYC 9554 64 コア プロセッサーを搭載した Dell PowerEdge R7625 サーバーの機能を活用して、肺炎検出用の AI 搭載医用画像ソリューションを構築しました。このソリューションは、高度なアルゴリズムと機械学習技術を使用して、X 線や CT スキャンなどの医用画像を分析し、患者の肺炎診断のスピードと精度を向上させるのに役立ちます。これによって最終的に、コンピューター支援レビューの追加レイヤーが導入され、医療従事者が大量の画像データをより効率的に処理するのをサポートする可能性が生まれます。

このリファレンス ソリューションでは、ResNet50 モデルを利用して、NIH Clinical Center のデータセットから取得した胸部 X 線画像を解析します。その主な目的は肺炎の有無を検出することであり、基本的に二項分類を実行します。このモデルは、NIH Clinical Center データセットの Xray DICOM データセットを使用してトレーニングされ、ResNet50 アーキテクチャによる転移学習が行われました。

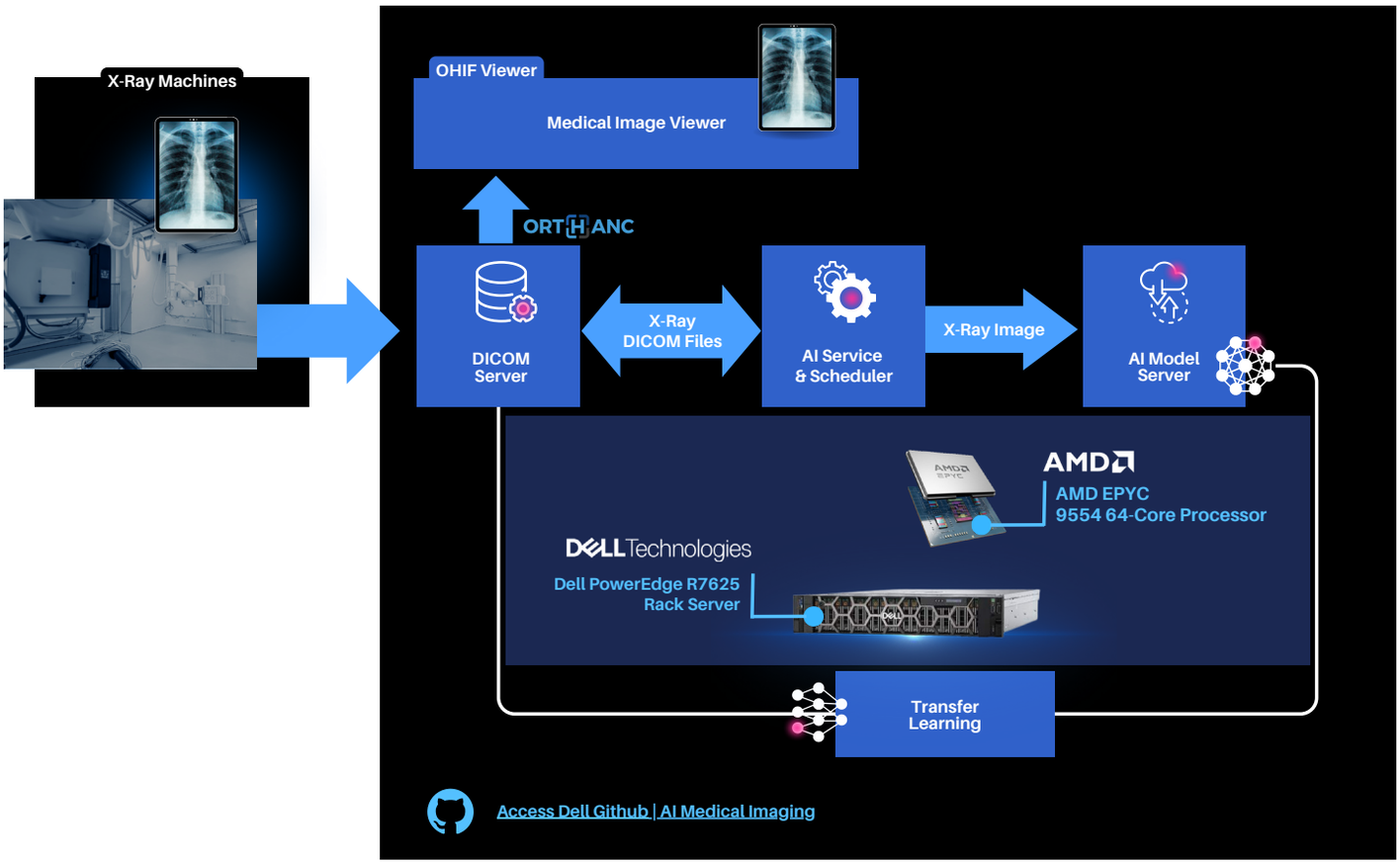
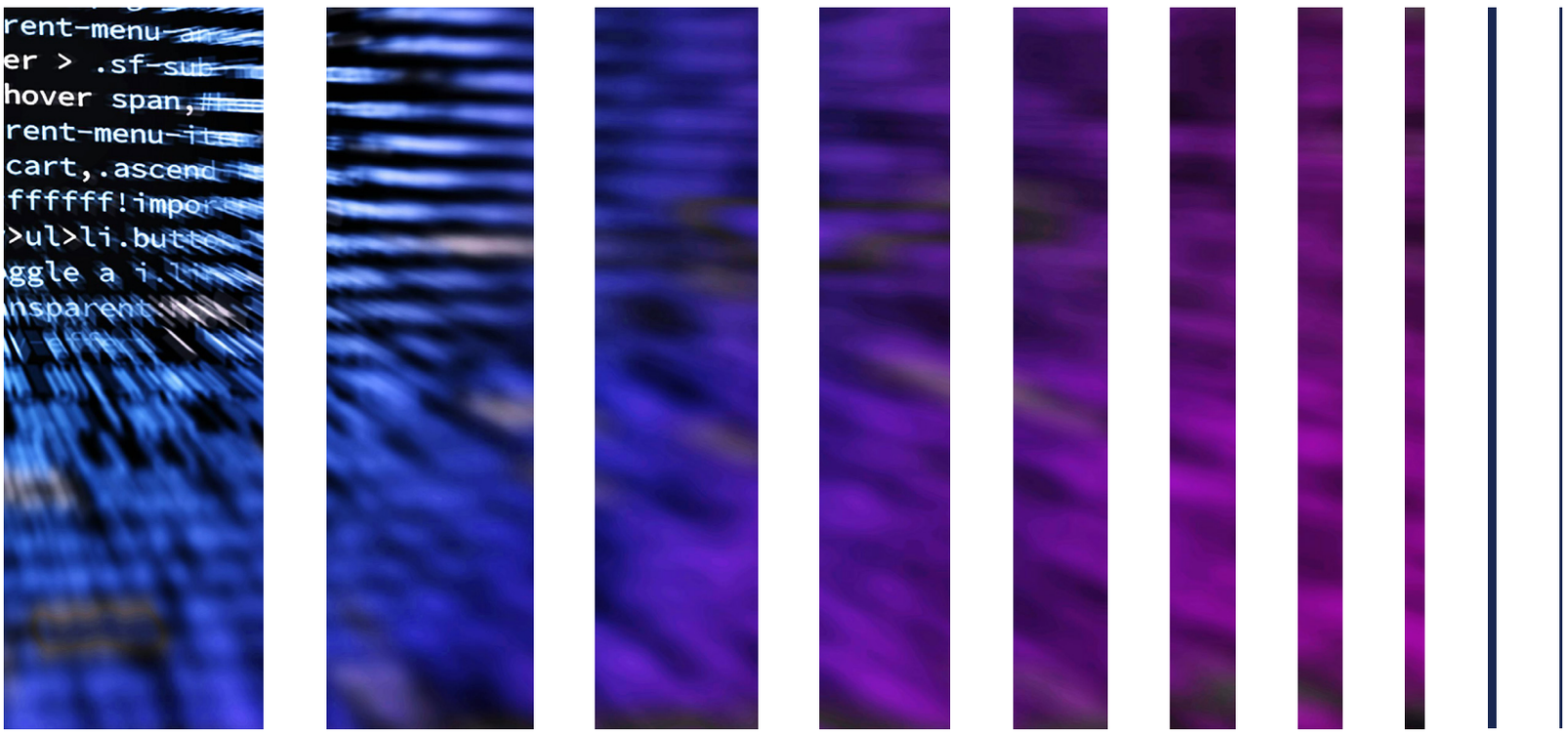


図 3：Medical AI Imaging Solution のアーキテクチャ図

ソリューション

すべての人のための AI：DELL と AMD による AI の民主化

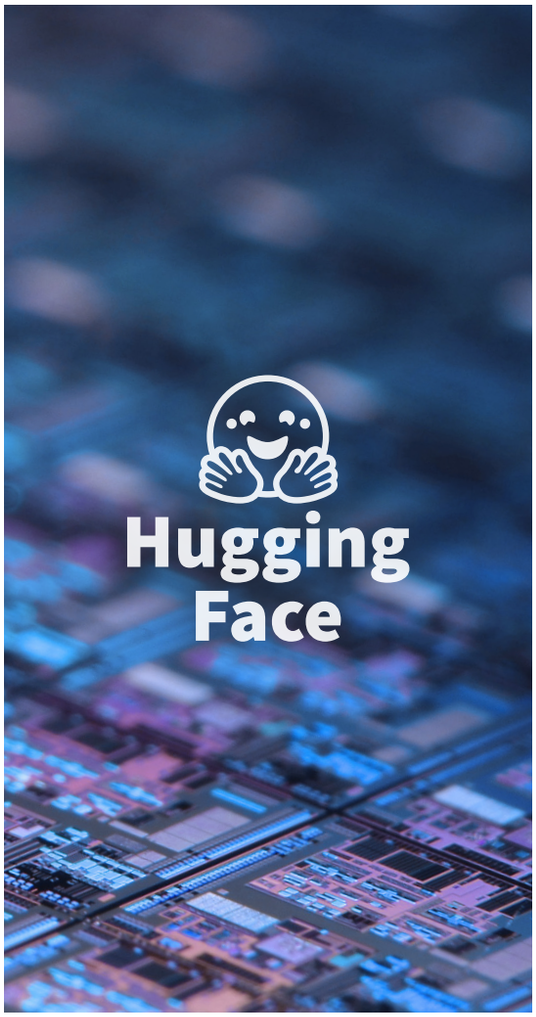
このコラボレーションは、AI エコシステムにおけるイノベーションとインクルージョンの促進に不可欠な、AI の民主化の基盤を築きます。Dell と AMD は、最先端の AMD CPU およびアクセラレーター テクノロジーを搭載したアクセスしやすい強力なサーバー スイートを利用して、個人や組織が AI を活用し、それぞれの分野で固有の課題を解決できるよう支援することで、AI の民主化基盤を構築するという成果を挙げています。AMD Instinct MI300X アクセラレーターを搭載した Dell PowerEdge サーバーは、大規模言語モデル (LLM) のトレーニングや微調整などの大規模な AI ワークロードを処理できます。一方、AMD EPYC プロセッサーを搭載した Dell PowerEdge サーバーは、エッジ推論ワークロードの処理に優れています。AMD は、基盤となるハードウェア プラットフォームに加えて、AMD CPU でのディープ ラーニング推論を最適化するための ZenDNN ソフトウェア ライブラリーと、AMD Instinct アクセラレーターでのトレーニング、微調整、推論機能を向上させる AMD ROCm ソフトウェア ライブラリーも提供しています。これらのオプションはすべて、AMD の統合推論モデル (UIF) にシームレスに結び付けられており、ユーザーはエンドツーエンドの AI ソリューションを構築でき、ソフトウェア フレームワーク、ソフトウェアの最適化、ハードウェア プラットフォームを柔軟に選択できます。

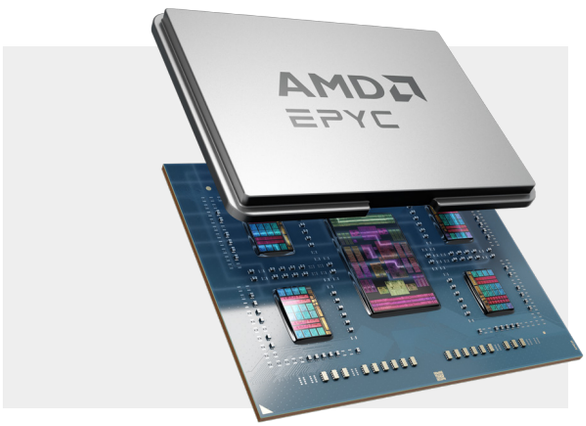


HUGGING FACE とのコラボレーション

AI の導入に意欲的な企業は、データサイエンスと機械学習に特化したオープンソース プラットフォームである Hugging Face から直接、特定のニーズに合わせてカスタマイズされた既存のモデルまたは AI ワークフローを活用することから始められます。AMD が Hugging Face とのコラボレーションを開始した目的は、すでに AMD プラットフォームとシームレスに統合されているソフトウェア ライブラリーやフレームワークに AMD 固有のソフトウェア最適化を追加して、最高レベルのトランスフォーマー パフォーマンスを提供することです。Hugging Face は、AMD のエンジニアリング チームと積極的に協力することで、主要なモデルを最適化してピーク パフォーマンスを実現します。AMD ROCm を Transformers ライブラリーに組み込み、AMD プラットフォーム専用に設計されたライブラリーである Optimum-AMD を改善するとともに、Hugging Face ユーザーが最小限のコード変更でそれらを利用できるようにしています。

また、デル・テクノロジーズは最近、Hugging Face と協力し、企業が Hugging Face コミュニティーを使用して独自のオープンソース生成 AI (Gen AI) モデルを開発、微調整、適用するプロセスをシンプルにしました。これらはすべて、業界をリードする Dell のインフラストラクチャ製品とサービスに適用されます。Hugging Face プラットフォーム上には、新しい Dell ポータルが開発されています。このポータルにはカスタムの専用コンテナとスクリプトが含まれており、ユーザーは Dell のサーバーとデータストレージ システムを使用して、Hugging Face で利用可能なオープンソース モデルを安全かつ簡単に導入できます。企業は、Hugging Face のリソースを最大限に活かして、AMD プロセッサーを搭載した Dell PowerEdge サーバーにモデルを直接導入することができます。そして、企業独自のデータを使用してエンドツーエンドの AI ソリューションを構築できるようになりました。



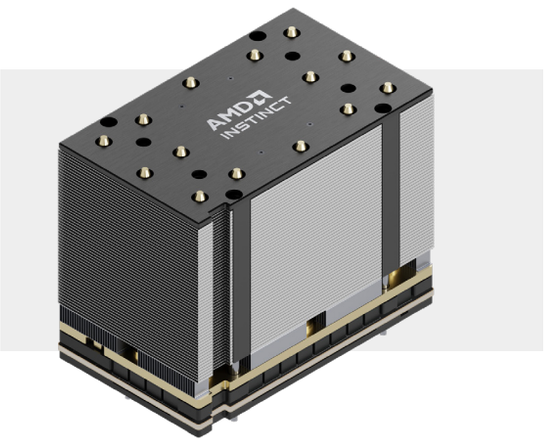


AMD EPYC プロセッサー

AMD は、AMD EPYC プロセッサーを通じて、最新のクラウドベースのデータセンターに必要な技術的進歩を提供します。これらのプロセッサーは、現在および将来のデータセンターのニーズに効率的に対応するためにゼロから設計されたシステム オンチップ (SoC) です。AMD EPYC 9000 シリーズ プロセッサーにより、データセンターは最大 128 コア、256 スレッド、ソケットあたり最大 6 TB のメモリーをサポートする 12 のメモリー チャネル、および 128 の PCIe Gen5 レーンを備えることができます。これは、業界の先駆的なハードウェア組み込み型 x86 サーバー セキュリティ ソリューションと組み合わされています。重要なコンピューティング、メモリー、I/O、セキュリティ リソースを SoC に統合した AMD EPYC プロセッサーは、最高レベルのパフォーマンスを実現し、総所有コスト (TCO) の削減を促進します。

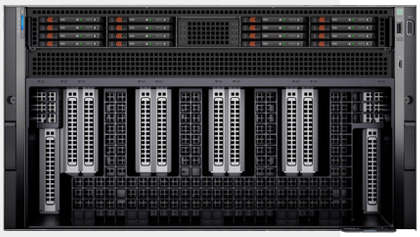
AMD INSTINCT MI300X アクセラレーター

最先端の AMD CDNA 3 アーキテクチャ上に構築された AMD Instinct MI300X アクセラレーターは、最も負荷の高い AI および HPC アプリケーションに、業界をリードする効率性とパフォーマンスを提供します。304 台のハイパフォーマンス コンピューティング ユニットを搭載し、新しいデータ型や写真および動画のデコードのサポート、単一のアクセラレーターに前例のない 192 GB の HBM3 メモリーなど、AI 固有の機能を備えています。



AMD ROCM 6オープンソース ソフトウェア プラットフォーム

AMD ROCm 6オープンソース ソフトウェア プラットフォームは、AMD Instinct MI300X アクセラレーターのハイパフォーマンス　コンピューティング(HPC)とAIワークロードのパフォーマンスを最大化するように最適化されています。また、AMD Instinct MI300Xアクセラレーターのサポートも拡張され、業界のソフトウェア フレームワークとの互換性が確保されます。AMD ROCmプラットフォームには、カーネル レベルからエンドユーザー アプリケーションまでのアクセラレーター プログラミングを容易にするさまざまなドライバー、開発ツール、APIがカプセル化されており、特定の要件に合わせてカスタマイズできます。AMD ROCmは、ハイパフォーマンス コンピューティング(HPC)、人工知能(AI)、科学計算のアプリケーションに特に適しています。さらに、AMD ROCmプラットフォームは、サーバー ノード通信用のリモート ダイレクト メモリー アクセス(RDMA)など、マルチアクセラレーター コンピューティングをサポートしています。



DELL POWEREDGE サーバー ポートフォリオ

AMD への Dell の投資は、 市場において AI を民主化するための重要な選択肢を生み出しており、 それは EPYC を搭載した 4 つのサーバー プラットフォームや、AMD Instinct MI300X アクセラレーターを搭載した主力製品の Dell PowerEdge XE9680 ラックサーバーによって証明されています。 AMD EPYC プロセッサーを搭載した最新世代の Dell PowerEdge サーバーは、ビジネスの俊敏性を高め、 市場投入までの時間を短縮します。さらに、データベースと分析、仮想化、ソフトウェアデファインド ストレージ、仮想デスクトップ インフラストラクチャ (VDI)、コンテナ化、ハイパフォーマンス コンピューティング (HPC)、AI、機械学習 (ML) などの変革ワークロードをサポートします。1 ソケット（シングル CPU） ラックサーバーは、パフォーマンスとストレージ容量をコスト パフォーマンスに優れたバランスで提供し、ビジネスに合わせてシームレスに拡張できるように設計されています。また、2 ソケット（デュアル CPU） ラックサーバーは、幅広い機能によって要求の厳しいワークロードに対応します。

Dell PowerEdge XE9680 ラックサーバーは、AI タスク向けに専用設計された堅牢なデータ処理システムです。8 基のアクセラレーターをサポートし、 機械学習 (ML)/ ディープ ラーニング (DL) のトレーニングや推論のワークロードに適しており、 特に大規模言語モデル (LLM) のトレーニングに最適です。 Dell PowerEdge XE9680 ラックサーバーは、 AMD Instinct MI300X アクセラレーターを 8 基搭載しており、 それぞれ 192 GB、 メモリー帯域幅 5.3 TB/ 秒 (HBM3) となっています。 サーバーあたりの合計 HBM3 容量 1.5 TB、 FP16 性能 21 petaFLOPS 超を実現し、 企業の生成 AI へのアクセス性をさらに拡張します。 これにより、 さらに大規模なモデルをトレーニングして、 データセンターの設置面積を最小限に抑え、 TCO を削減して競争上の優位性を得ることが可能です。

概要

AI によって推進される急速なイノベーションは、データセンターのワークロードに他のいかなる技術的変革よりも速い革命をもたらしています。これらの技術的進歩をサポートするために、Dell と AMD は、あらゆる業界の開発者がオープンソース リソースでコラボレーションして今日の生成 AI イノベーションを推進することを奨励する、より包括的かつ革新的で倫理的に開発された AI エコシステムの実現に向けて取り組んでいます。お客様の AI ソリューションのパフォーマンス要件が、AMD EPYC プロセッサーまたは AMD Instinct アクセラレーター搭載サーバーのどちらで満たされる場合でも、当社はハードウェア プラットフォーム全体で AI ワークロードの実行の柔軟性を実現し、Dell と AMD が提供するメリットの最大限の活用を可能にします。

リファレンス

AMD イメージ：AMD.com、AMD パートナー向けリソース ライブラリー、
https://www.amd.com/en/partner/resources/resource-library.html

Dell イメージ：Dell.com