White Paper

ワークステーションにおける AI テクノロジーの開発およびデプロイの必要性

Sponsored by: Dell Technologies

Peter Rutten　　Dave McCarthy
July 2023

IDC の見解

AI（Artificial Intelligence：人工知能）はあらゆる業種において重要度を増し、差別化のための機能として急伸しており、AI を実行するために必要なハードウェアは急速に進化している。テクノロジー業界では、多くの先進的な AI モデルが直面する指数関数的な規模の増大に、高い関心を持っている。ここでの論点は、何百億個というパラメーター、精度の低下、拡大するメモリー、AI の学習や推論に必要な高性能コンピューティング（HPC：High-Performance Computing）およびラックに搭載したアクセラレーテッドサーバー（アクセラレーテッドサーバーは、特定の計算処理を汎用の CPU からオフロード（負荷軽減）し高速に実行するための専用ハードウェアを使用するものである。AI、可視化、自律型マシン、機械学習、データ分析などに対処できるように設計されている）についてである。しかし、こうした桁外れの大規模な AI コンピューティングは、多くの企業にとっては関係がないことであると言える。

今日、多くの企業は、スーパーコンピューターを必要としない Generative AI（生成系 AI）などの AI 戦略に注力している。多数の AI 開発、および AI のデプロイ（配置）が、特にエッジ側でますます増える兆しを見せており、高性能なワークステーション上で実行されている。AI の開発やデプロイ（配置）とって、ワークステーションには、いくつかの利点がある。1 番目は、AI サイエンティストや開発者が、サーバーを使用する時間を確保するための交渉が不要になることである。2 番目は、依然としてサーバーベースの GPU（Graphics Processing Unit）がデータセンターでは容易に利用できなくても、GPU アクセラレーションと極めて安価なサーバー群が提供されることで、クラウドのインスタンス費用の急激な上昇を気にすることなく、より少額かつ一括費用の支払いで済むことである。3 番目は、機密情報がオンプレミスに安全に保存されるという安心感である。また、ワークステーションを利用することで、サイエンティストや開発者では、AI モデルの実験だけで利用費用が増加してしまうという心配も不要になる。

AI デプロイ（配置）のシナリオは、オンプレミスやクラウドなどプラットフォーム側よりも、エッジ側で急速に拡大していると IDC ではみている。ここでも、ワークステーションは AI 推論プラットフォームとしてますます重要な役割を果たしており、GPU を必要とすることなく、ソフトウェアで最適化された CPU 上で推論実行を行うケースが増えている。エッジ側のワークステーション上で AI の推論を行うユースケースは急速に増えており、AIOps（Artificial Intelligence for IT Operations）や災害対応、放射線医学、石油ガス探査、土地管理、遠隔医療、交通管理、製造工場監視、ドローンなどがある。

本調査レポートでは、ワークステーションが AI 開発やデプロイ（配置）において果たす役割の高まりについて解説し、デル・テクノロジーズ（以下、デル）の AI 用ワークステーションのポートフォリオを簡潔に紹介する。

概況

AI の急増とインフラストラクチャへのインパクト

世界中の企業が従事している AI プロジェクトの数は急増している。すでに、あらゆる業種に渡って、多くのタスクの全体あるいは一部分が、AI モデルがデプロイ（配置）されたソフトウェアによって実行されている。IDC は多くのレベル（階層）で AI の動向を追跡している。考慮すべき有用な測定基準の一つは、ビジネスやクラウドサービスプロバイダーが、AI の開発と実行に当たってサーバーに支出する予測額である。2026 年までに、その額は 346 億ドルになり、世界中のサーバーへの合計支出額の約 22%となる。

しかし、サーバーだけでは全体像を把握できない。多くの AI の準備、開発、プロトタイピング、そしてますます増えるデプロイ（配置）が、ワークステーション上で行われているためである。小規模や大規模の企業が、「自社のアプリケーションにいくつかの AI 機能を取り入れることで、新たなビジネス機会につなげられる」ことを認識しており、AI モデルの実験事例が急増している。すぐに使用できること、およびデータとの近接性（手元でデータが利用できる）の観点から、堅牢なワークステーションは、前述の目的において理想的と言える。

AI のアルゴリズムは何十年にも渡って利用されてきたが、AI がこのように急速に広まったのはなぜであろうか。それは主に、特に普及したタイプの AI アルゴリズムであるニューラルネットワークを強化する、2 つの典型的な環境がここ数年で整備されたからである。一つは、非構造化データや半構造化データなどの広範で安価な、しかもさまざまなタイプのデータが利用可能になったことである。もう一つは、線形コンピューティングの処理能力を並列モデルで強化し、こうしたニューラルネットワークを許容時間内で処理できるようになったことである。こうした 2 つの基盤となる環境が揃ったため、データサイエンティストが驚異的に優れたタスクを実行する方法を自動的に習得するニューラルネットワークの開発が、大きな前進を遂げた。文字や数値データにとっては従来の機械学習（ML：Machine Learning）が依然として重要であるが、映像や音声、言語などに対しては深層学習（DL：Deep Learning）がより有効である。

従来の機械学習モデルは通常、数十個ほどのコアで構成されるワークステーションの CPU 上で開発できるが、ニューラルネットワークは数千個のコア上で並列処理を行うコプロセッサーを必要とする。この主な理由は、機械学習では特徴抽出や分類は手作業の処理であるのに対し、深層学習ではこの作業が自動化され、大規模なデータセットを使用した定型的な繰り返し処理を伴う計算によって学習するモデルを必要とするからである。現在、最も一般的なコプロセッサーは GPU であるが、スタートアップ企業が開発した新しい AI 専用プロセッサーも利用可能になりつつある。このタイプの高速化は、ディスクリート型（単体）のプロセッサーを並列処理に使用しており、サーバー市場やワークステーション市場に革命を起こし、IDC が大規模並列計算と呼ぶものを生み出した。

2022 年、アクセラレーテッドサーバーは世界市場で 218 億ドルになり、2026 年までには 434 億ドルに増加するとみられる。その 57%が AI を実行するアクセラレーテッドサーバーである。同時に、ワークステーションで使用されるために販売されたディスクリート GPU（単体 GPU）は 2022 年に 640 万台まで増加した。科学計算用またはソフトウェア工学用途に使用されるワークステーションの市場は、AI 開発の普及に伴い急速に拡大し、2026 年までに 20 億ドル近くに成長すると IDC は予測している。

AI 開発のステージ

前述した通り、ニューラルネットワークが実行可能になったのは、データタイプとデータ量が増加し、新しい計算方法が考案されたからである。ある説によると、この新しい計算方法の最初の部分は、データ量とデータタイプで構成されており、深層学習の AI 戦略における取り組みの最大 80%はデータ管理とデータの準備（データプレパレーション）が占める。データの収集、管理、準備を行ってからモデルの設計や学習が可能になる。IDC によると、AI の開発ステージは以下の通りである（Figure 1 を参照）。

- **データの管理**：企業が収集、生成、取得したデータセンター、エッジ、クラウドにわたる大量のデータの中から、AI モデルに関連したデータを選別して管理する（このデータはイベント駆動型、ストリーミングなど、どのようなデータタイプでも対応可能で、その大部分はガバナンスを必要とする場合がある）。
- **データの準備（データプレパレーション）**：データ（ファイル、ブロック、オブジェクト）をデータウェアハウスやデータレイクに保存し、クリーニングし、完全で高品質であることを保証する。そして次にそのデータを、たとえば、Spark や pandas などのツールで AI モデルが使用可能な形に変換する。
- **モデルの選択**：誤り率や性能という観点でプログラムされた AI タスクを、最適に実行するモデルを決定する。
- **モデルの開発**：XGBoost、LightGBM、GLM、Keras、TensorFlow、PyTorch、Caffe、RuleFit、FTRL、Snap ML、scikit-learn、H2O といったフレームワークを使用した AI モデルを設計する。
- **モデルの学習**：インフラストラクチャコンピューティング上で、十分なプロセッサーと並列用のコプロセッサーコアを使用してモデルに学習させる（また、モデルの決定を説明、検証、文書化する機能を取り入れて、公正性、説明責任、透明性を保証する事例が増加している。これにはプロトタイピング、つまり、モデル上で推論を実行させて学習済みモデルを試験することが含まれる）。
- **モデルのホスティングと監視**：運用環境にモデルを配置し、モデルを設計した目的のタスクを実行する。これは通常「AI 推論」と呼ばれ、その性能を監視する。

これら 6 つのステージのいずれにおいても、ワークステーションは、データセンターやクラウド、エッジインフラストラクチャと組み合わされ重要な役割を果たす。

FIGURE 1

AI 開発のステージ

Source: IDC, 2023

ワークステーション上で AI モデルを開発する

ワークステーション対パーソナルコンピューター

PC は一般的に、AI 開発用としては性能不足であるとみられている。データサイエンティストや AI 開発者は多くの場合、企業内で戦略的に重要なプロジェクトに関与しており、生産性を落とさないことが最も求められる。ワークステーションの方が PC よりも期待通りに性能を発揮するケー

スが多いのは、通常、ワークステーションがより高性能のコンポーネントで構成されており、そのコンポーネント上で実行されるソフトウェアに対し最適化されているからである。

こうしたコンポーネントには以下のものが含まれる。

- **高品質なプロセッサー**：一つの例は Intel Xeon Scalable プロセッサーである。
- **高性能な GPU**：一つの例は、NVIDIA RTX 6000 Ada などの NVIDIA の RTX プロフェッショナル GPU である。
- **より多くのストレージ**：一部のワークステーションでは最大 60TB の提供が可能で、入出力速度は PC に比べて格段に速い傾向にある。
- **より多くのメモリー**：ワークステーションでは最大 6TB のメモリーが利用可能である。
- **冷却**：高性能のコンポーネントでは大量の熱が発生することから、データサイエンティストには、適切な冷却機能で過熱を防止できる、最適な性能を維持するワークステーションが必要である。
- **ネットワークインターフェースカード（NIC：Network Interface Card）**：リモートのサーバー上に保存されている大量のデータセットを処理する。データサイエンティストには、データを迅速かつ効率的に転送するための高速なネットワークインターフェースカードが不可欠である。
- **ディスプレイ**：データの可視化タスクにとっては高品質のディスプレイが重要であるため、データサイエンティストには、高解像度で色精度が高く、大型の画面サイズのモニターが必要である。
- **誤り訂正コード（ECC：Error-Correcting Code）メモリー**：ECC は、最も一般的な種類の内部データの破損を検出し、訂正して、AI 学習中にハードエラー（ビット不良）またはソフトエラー（不正な値の原因となるビットの反転）に起因するブルースクリーンの発生を防ぐ。また ECC は、医療など高い安全性が求められる業務において結果の正確性を確保する。
- **特定用途のプロセッサー**：一つの例は Intel Movidius 画像処理ユニット（VPU）で、小売業、セキュリティ、産業オートメーションなどの設定に使用される、コンピューター画像やエッジ側 AI アプリケーション用の並列処理コプロセッサーである。FPGA も、たとえば金融アプリケーション用のワークステーションで使用される。
- **最適化ソフトウェア**：たとえば、oneAPI はインテルの標準ベースのプログラミングモデルで、CPU、GPU、FPGA やその他のアクセラレーターに対するデータ処理を中心とするワークロードの開発とデプロイ（配置）を簡素化するものである。また、CUDA は NVIDIA の並列コンピューティングプラットフォームおよび API（Application Programming Interface）であり、GPU 上の一般的なワークロードを実行する。

AI における CPU 対 GPU

ワークステーションは AI 開発のさまざまなステージで使用でき、一般的には多くの機能を備えている。並列処理においては GPU が重視されるが、ワークステーション上で AI を開発する際には CPU が重要な役割を果たす。GPU と同様に、CPU もデータ操作や、もちろん従来の機械学習モデルの開発にも利用できる。また CPU は、データセットを視覚的に表現して、データの特徴を解釈するプロセスであるデータ探索にも使用できる。

深層学習では、実際の学習プロセス時は GPU が引き受けるため、ホストである CPU の役割は若干軽減されるが、CPU は、OS や CUDA など重要なソフトウェア用の処理レイヤーや、GPU 群あるいは他のプロセッサー間の処理におけるオーケストレーションの役割を果たす。さらに CPU は、本番環境で AI モデルを実行するために、ワークステーションが使用される場合において、AI 推論エンジンという新しい役割を引き受ける。2024 年までに、AI 推論用のインフラストラクチャに対する支出が、AI 学習用の AI インフラストラクチャに対する支出を上回り、そうした推論のかなりの部分（39%）がホスト CPU 上で実行されると IDC は予測している。

ワークステーション対サーバー：共生の関係

ほとんどの企業において、ワークステーション、オンプレミスのサーバー、クラウドインスタンス、あるいはこれらの組み合わせを AI 開発用に配置する際は、実用の観点から経験に基づいたアプローチがとられる。AI プロジェクトのさまざまな開発ステージ用のワークステーション、サーバー、クラウドインスタンスの間には共生の関係がある。

データセンターやサーバーに対するワークステーションの利点は、データサイエンティストが希望するあらゆる場所で作業できることであり、これは現在の感染症流行期（パンデミック）はもちろん、通常時においても重要な要因である。また、データサイエンティストは AI モデル上で、自分たちが必要とする分だけ、自由に繰り返し実験ができる。現代のワークステーションのパワーは強力な GPU によって、繰り返し処理をいっそうインタラクティブに処理でき、サーバーへのアクセス要求の必要性や、他のデータセンターの制限に妨げられることなく、計算処理のフィードバックと結果を即座に提供できるからである。しかもワークステーションは、計算処理が行われる場所にデータを移動させる代わりに、計算処理をデータに近い場所で行う柔軟性を提供する。さらに、帯域幅を節減し、ネットワーク輻輳を軽減し、スループットを向上させる。また、ワークステーションは、たとえば従来の機械学習タスクやより深層学習技術が求められる作業などのさまざまなニーズに合わせて構成できる。

アクセラレーテッドサーバー市場は急速な成長を見せているものの、企業のデータセンターにおけるアクセラレーテッドサーバーの利用は、まだ限られている。本調査レポートの執筆時点で、企業のデータセンターにおけるサーバーの平均 4%がアクセラレーテッドサーバーである。つまり、多くの企業では容易に利用可能なオンプレミスの GPU 上で、AI を開発あるいは実行する手段を有していないと言える。こうした理由から見ても、アクセラレーテッドワークステーションは AI 開発の有益な代替手段である。

特に高速なアクセラレーテッドワークステーションは、今や AI モデルが極端に大規模なものでなければ、深層学習を実行するのに十分強力であり、サーバー上で学習する必要がなくなる。GPU を搭載したワークステーション上で学習されたモデルは、CPU 中の推論機能を活用し、ワークステーション上にも配置でき、また GPU を搭載していないサーバー上にも配置できる。インテルの DL ブーストや oneAPI といったソフトウェアテクノロジーは、CPU 上の AI 推論を強化し、データセンターの既存の非アクセラレーテッドサーバーが、AI アプリケーションをサポートできるようにしている。

ワークステーション対クラウド

クラウドコンピューティングは、企業のインフラストラクチャ、データ、アプリケーションの考えに革命をもたらした。ほぼ制限のないスケーラビリティを実現したことで、開発者たちはクラウドによってリソースをオンデマンドでプロビジョニングでき、リソースの制約が少ないため、イノベーションを生み出すサイクルを加速できるようになっている。広く知られている通り、クラウドは AI 開発にとって理想的なケースのように思える。

とは言え、クラウドの選択が最適でないケースもある。実際、IDC の調査では、企業はワークロードの一部をパブリッククラウドからオンプレミスのインフラストラクチャに戻す傾向にあることが分かった。この動きには以下に示す要因がある。

- **クラウドの可用性**：クラウドサービスを頼りにしている人は誰でも、クラウドプロバイダー内の問題か、ハイパースケールのデータセンターとエンドユーザー間のネットワーク接続の喪失などによって、サービスの機能停止を経験したことがあるであろう。こうした状況では、ユーザーはサービスプロバイダーが問題を解決するまで機会損失が発生し、生産性が著しく低下する。
- **セキュリティとコンプライアンス**：多くの業種では、データの送受信相手および保存場所が、企業のガバナンスポリシーに規定されており、クラウドサービスの利用が制限されて

いる。欧州の GDPR（General Data Protection Regulation：一般データ保護規則）やカリフォルニア州消費者プライバシー法などといった政府規制も、データ主権に規制を課している。

- **費用**：一般的に企業は、クラウドサービスの利用料、特に高性能の計算機能や大量のストレージを必要とするワークロードの費用が、どの程度急速に増大するかを過小評価している。クラウドエコノミクスは、データをオンサイトのインフラストラクチャに戻すことを含む、あらゆる種類のリソース消費を計測することを基本としている。
- **試行錯誤のプレッシャー**：ほとんどの AI 戦略は、初期から相当量の実験が必要となり、意図したモデルを構築できなかった場合は、開発プロセスに大きな影響を与えることになる。この開発プロセスでは、実行可能な結果が表れない間にもクラウド費用が積み増しされるため、AI サイエンティストや開発者は心理的な負担を強いられることになる。

ワークステーションはマイクロサービスベースのアーキテクチャや API 駆動型のオートメーションなどのクラウドネイティブなテクノロジーを活用しながら、こうした問題に対処できる。これによって、ワークステーションとデータセンターおよびサーバーと同様の利点が得られる。

- **どこでも作業が可能**：パブリッククラウドへの依存を排除した、ネットワーク非接続型のシナリオ（システム設計）が可能となる。高セキュリティ環境の多くは、パブリックネットワークから物理的に切り離されており、AI ワークステーションはこのニーズに対して独自に対処できる。高性能なローカルのリソースによって、高価なネットワーク接続が不要になる。
- **データの局所性**：IoT デバイスやその他の接続された装置の普及によって、エッジ側でのデータが指数関数的に増大している。多くの場合、コンピューティングリソースと専用のワークステーションを同じ場所に置くことは意味がある。また、これによって、データの移動を抑えることでコンプライアンス要件の多くを解決できる。
- **自由な実験**：AI モデルの学習と最適化には、繰り返し作業が伴い、その多くで試行錯誤が行われる。開発者は、サーバーの追加費用を気にせず、妥協することなく実験ができる自由な利用環境を求めている。また、ワークステーションはツールのカスタマイズに対し、より大きな自由度を提供する。

3 番目の「自由な実験」については、ほとんどのクラウドサービスプロバイダーでは、エンドユーザーが導入したいあらゆる構成の費用見積りを即座に提供するため、ワークステーション導入費用とクラウド利用費用は、容易に比較できる。たとえば、NVIDIA T4 を 1 つ搭載し、375GiB SSD ストレージのインスタンスを 1 つ持つ単一の通常の仮想マシン（VM：Virtual Machine）を 1 日 8 時間、週 5 日間利用した場合、ある大手のクラウドプロバイダーの場合 140 ドルとなる。VM と T4 および SSD を 2 倍にすると費用は月額で 365 ドルに増加する。VM は 2 台のまま T4 を 2 倍の 4 つにし、375GiB のストレージを 4 台としてフルタイムで学習を実施する環境にすると、費用は月額 2,700 ドルに上昇する。したがって、AI 開発用のクラウドの費用は年間で容易に数万ドルに跳ね上がり、高性能ワークステーションの年間減価償却を大きく上回ると言える。

ワークステーション上での AI のプロトタイピング

AI モデルのプロトタイピングのステージでは、オンプレミスのサーバーおよびクラウドに比べ、ワークステーションに明確な利点がある。データセンター内のサーバーはフルに利用されているかもしれないし、AI プロトタイピングやテストの用途にはミッションクリティカルすぎる（提供する信頼性が高すぎる）かもしれない。しかも、前述した通り、クラウドインスタンスをテスト環境として自由に利用すると、たちまち費用の超過につながる。AI サイエンティストや開発者は、プロトタイピングのステージでワークステーションを利用することで、サーバーアクセス権の交渉やクラウド費用の増加への危惧から解放される。ワークステーションは安価な初期費用（CAPEX）の一括払いで済むため、追加の費用なしでいつでもどこでもプロトタイピングが行える理想的で自由な利用環境を提供する。

ワークステーションにおける AI モデルのデプロイ（配置）

ワークステーション上での AI モデル開発は、ここ数年で広く普及した戦略となったが、AI モデルをワークステーション上（その多くはエッジ側）に配置するユースケースが増えていると IDC はみている。言い換えれば、AI モデルで推論を実行することで、AI モデルをワークステーション上で本番稼働させるということである。サーバー用の AI がエッジに配置されるケースが急速に増加（年間のハードウェア支出が 2020 年～2024 年の間に 3 倍以上に増大）している。また、ワークステーションのエッジでの利点をエンドユーザーが見出しており、ワークステーションもそれほど後れを取っていない。

IDC はエッジを、インフラストラクチャやアプリケーションを集中型クラウドやオンプレミスのデータセンター以外の、データの生成と消費が行われる場所にできるだけ近い場所に配置する、「分散型コンピューティングモデル」と定義している。これにはリモートや支店、工場、倉庫、病院、小売店舗などといった業界固有の場所も多く含まれる。

データ集約型およびコンピューティング集約型のワークロードは、オンプレミスまたはエッジの場所に配置されることが多くなりつつある。これは、パブリッククラウドに内在する、大規模なデータセットのアップロードにかかる時間、AI 学習を実行する際の変動費など、特に大量のデータサイエンスの実験を必要とするような状況での制約緩和を目的としている。

IDC の調査では、AI の配置シナリオでエッジが急速に増加しており、2023 年には企業は AI コンピューティングに 29 億ドルを投資し、2026 年には 69 億ドルに拡大するとみられる（『*Worldwide AI Hardware Forecast, 2022–2026*:*Strong Market Growth for AI Compute and Storage*（IDC #US49671722、2022 年 9 月発行）』を参照）。さらに、エンジニアリングや技術用途などの HPC ワークロード用の導入先としてエッジの需要が高まっており、エッジでのワークロードなどに企業は現在 10 億ドル近くを投資し、2027 年までには 24 億ドルに成長するとみている（『*Worldwide High-Performance Computing Server Forecast, 2023–2027: Enterprise Will Overtake HPC Labs*（IDC #US50525123、2023 年 4 月発行）』を参照）。AI ワークステーションの導入が有効なのはこうした分野である。

AI モデルをエッジのワークステーション上に配置する場合は、AI 開発の場合と同様に必ずしもハイエンドの GPU が必要となるわけではない。より低性能な GPU でも AI 推論を実行でき、GPU をまったく必要としない場合も少なからずある。そうした場合、特に、AI 推論などの AI ワークロードを高速に実行するように設計されたインテルのマイクロプロセッサー上のインストラクションセット（機能群）であるインテルの DL ブーストなどの最適化機能を使用すると、CPU は推論タスクを適切に実行できる。インテルによれば、DL ブーストを使用すると、DL ブーストをサポートしている第 4 世代 Intel Xeon Scalable プロセッサーでは、INT8 リアルタイム推論スループットが前の世代のプロセッサー（BERT-Large SQuAD）の 1.45 倍となることが確認できたと言う。それはまた、パワー、モビリティ、温度管理などを考慮すると、より少ない消費電力が求められる点からも、ワークステーションをエッジで設置するのが有効であると言える。インテルの Movidius Myriad（M2）は、12W という少ないエネルギー消費量のおかげで、消費電力基準にうまく適合する。

ワークステーション上に AI を配置するユースケース

ローカルのワークステーションに AI モデルを配置することが理にかなっている例がいくつかある。それらに共通した特徴は、大量のマシン生成時系列データ、動画ストリームや画像などの大量の非構造化データであるという点である。また、対象分野の専門スタッフが、人間による解釈で AI モデルを増強しなければならない場合もある。

以下はその例である。

- **AIOps**：IT システムがその規模と複雑性を増す中、事後対応型のインシデント管理から事前対応型の監視に移行する必要性が増大している。これは特に、インフラストラクチャやアプリケーションが、テクニカルスタッフがほとんどあるいはまったくいないエッジロケ

ーションに分散されている場合に当てはまる。通常性能のベースライン（基準値）をモデル化することで、異常を検知し、修正ステップの自動化が可能である。

- **災害対応**：緊急時においては、初動対応者は状況を素早く把握し、重要な設備を確認し、IT リソースを配置して、最も必要な支援をしなければならない。これは、ネットワークの接続が失われた環境下で行わなければならないことが多いため、データフィードを収集し、AI モデルで推論して、主要社員への連絡を自動化できるローカルのワークステーションが必要となる。
- **放射線検査**：画像技術の進歩によって、1 回の走査で生成されるデータの容量が増大し、また、タイムリーな分析を行うためにオンサイトで分析を行う必要がある。既存の何百万という例から学習した AI モデルは、人間の目よりも正確にパターンを検知でき、精度が向上している。
- **石油およびガス探査**：探査段階を担う石油およびガス会社はテレメトリー、地震、画像のデータを組み合わせて使用し、天然資源の埋蔵場所を特定し、採掘場所を選択して生産プロセスの設備の性能を最適化する。これは、高価な衛星通信しか利用できない場所での情報分析が必要なことが多い。
- **がん研究および創薬**：研究病院や学会の研究者たちは AI や自然言語処理を利用して、患者に対して最も効果的で、個別に最適化したがん治療をがん専門医が決定できるように支援する。また研究者たちは、機械学習とコンピューター画像処理を組み合わせて、患者の腫瘍がどの程度進行しているかを放射線科医がより正確に把握できるようにする。さらに、がんがどのように進行し、がんと戦うための最良の治療が何かを適切に判断するためのアルゴリズムを使用している。
- **保険請求評価**：人手による請求処理作業は労働集約的であり、人的エラーを起こしやすい。請求の妥当性を評価できる AI によって、保険査定者は詳細な調査が必要な業務に注力できるようになり、経費の削減につながる。これによって正確性を損なうことなく、業務全体の処理能力が向上する。
- **遠隔医療**：AI は、ウェアラブルデバイスからのリアルタイムのバイタルサインに基づいて個別に調整した治療計画を策定し、患者の回復率を向上させる。治療計画の情報は、患者の病歴や類似した事例のナレッジベースと組み合わされて生成される。これは、遠隔医療に多くを依存する地方都市において特に有効と言える。
- **小売店セキュリティ（盗難防止）**：ビデオカメラ映像にリアルタイム分析を適用することで、犯罪行為につながる人間の行動を予測する際に使用する。この予測には、複数のビデオカメラ入力を組み合わせて、店舗内の個々人の動きを追跡することが必要となる。店舗内の犯罪は短時間で行われるということから、この予測プロセスはローカルで実行するのが最適と言える。
- **交通管理**：輸送交通分野を司る政府機関は、AI を使用して信号機を制御し、車両の流れを改善するデジタルサイネージを提供して市民の安全を確保している。この取り組みには、ビデオカメラの映像、路上センサーからのテレメトリー情報などの入力を組み合わせて交通パターンを最適化することが必要となる。
- **製造工場監視**：工場管理者にとって、重要なプロセスの稼働時間を確保し、製造スケジュールを守ることが最重要である。つまり、主要な設備の予知保全、不具合の自動検知、工場のサプライチェーンの出入りの最適化が必要ということである。この分野は、安全基準を満たしながら生産性を向上できるよう人間のオペレーターを AI が支援できる分野である。
- **ドローン**：ドローンが撮影した画像を自動的に分析することで、今まで監視できなかったさまざまな状況で、これまでにはない規模の監視機能を実現する。ドローンによる監視は、ガスや電気の公益インフラストラクチャの検査、保険調査、捜索／救助活動、精密農業、漁業や野生生物保護に大きなインパクトをもたらしている。
- **日々のオフィス環境**：日常のオフィス環境は Microsoft Copilot といった AI ベースの生産性向上ツールによってますます改善されている。

- **再生可能エネルギー**：風力発電所、水力発電ダム、太陽光発電所などの再生可能エネルギーは、リアルタイムの監視、保守、データ収集を必要とし、これらはローカルで収集、分析しなければならない。

デルの AI 用ワークステーション

デルはさまざまなレベルの AI 開発やデプロイ（配置）用の広範なワークステーションを提供しており、そのすべてが Data Science Workstation（DSW）という同社のブランド下にまとめられている。このセクションでは、その仕様を簡単に述べ、データサイエンティストや Dell DSW テクノロジーの利点といった数々の AI ペルソナ／アプリケーションについて説明する。こうした AI の準備が整っているデータサイエンス用ワークステーションはデータサイエンティスト向けに特別に設計されている。最新の Precision Data Science Workstation は AI 機能を活用して、データサイエンティストが最も使用するアプリケーションの性能を最適化するように設計されたデバイスを微調整する。これによって、データサイエンティストは最も重要な作業を高速に完了できる。さらに、Dell Precision ワークステーションは独立系 ISV によってテストおよび認証がされており、デルの顧客が日々のタスクを完了するのに必要な高性能のアプリケーションをサポートすることを保証している。

デルのワークステーションが際立つ理由

NVIDIA RTX GPU を搭載した Dell Precision ワークステーションは、企業の分析や AI 戦略にとって強力なスケーラビリティと性能を提供するように設計されている。デル・テクノロジーズは、業界の最新の AI ソフトウェアを実行するように最適化された、広範なハードウェアソリューションを提供する。

- **堅牢なハードウェア構成**：Dell Precision ワークステーションは、マルチコアプロセッサー、大容量 RAM、マルチプル GPU のオプションメニューを用意しており、広範で強力なハードウェア構成を提供している。こうしたコンポーネントは AI タスクに必要な計算リソースを提供し、効率的な学習や推論を可能にする。
- **スケーラビリティとカスタマイズ性**：Dell Precision ワークステーションはスケーラブルでカスタマイズ可能であり、ユーザーがそれぞれの個別の AI 要件向けにハードウェア構成をカスタマイズできる。この柔軟性によって、ワークステーションは AI ワークロードの特定のニーズに対して最適化できる。
- **認証と最適化**：デルは NVIDIA と協業して、Precision ワークステーションの NVIDIA RTX 6000 Ada Generation カードなどの NVIDIA RTX GPU との互換性とその GPU を搭載した際の性能に関して認証している。この認証によって、AI タスク用に Dell Precision ワークステーションを NVIDIA RTX GPU と組み合わせて使用する際のシームレスな統合と最適化された性能が保証される。
- **強力な処理機能**：Dell Precision ワークステーションには、インテルのプロセッサーが搭載されており、AI タスクに必要な計算能力を提供する。マルチコアプロセッサーを搭載しクロック速度を高めることで、Dell Precision ワークステーションは AI ワークフローにおける学習や推論に必要な性能を提供する。
- **ソフトウェアとツールのサポート**：Dell Precision ワークステーションには、AI の開発やデプロイ（配置）をサポートするソフトウェアやツールがあらかじめ組み込まれている。これには、NVIDIA RTX GPU を活用するための最適化されたソフトウェアスタック、AI フレームワーク、ライブラリーなどがあり、ユーザーが AI プロジェクトを開始しやすいようになっている。

さらに、デルのワークステーションが際立つ存在である重要な別の分野のテクノロジーを次のセクションで説明する。

信頼性の高いメモリーテクノロジー

デルは ECC 上に、Reliable Memory Technology Pro（RMT Pro）というテクノロジーを提供し、稼働時間の最大化に役立てている。この RMT Pro は ECC メモリーと組み合わせることで動作し、メモリーの誤りをリアルタイムに訂正する。デルによると、RTM Pro は、DIMM がフル活用されていても不良メモリーに再びアクセスしないようにすることで、メモリーエラーを排除している。システムのリブート後、RTM Pro は不良メモリー領域を切り離して OS から見えないようにする。その結果、AI データサイエンティストや開発者は、不良メモリーへの再アクセスによって起こるクラッシュの問題に出くわすことがなくなり、生産性を大きく向上させる。

Dell Optimizer for Precision

また、デルの Dell Optimizer for Precision（DOP）は、同社のほとんどのワークステーションにインストールされている。DOP は、評判の良いさまざまな商用アプリケーションをワークステーションが可能な限り高速に実行できるようにシステムの設定を自動的に調整する。DOP によって、データサイエンティストや開発者の生産性が向上する。また、DOP は、プロセッサー、ストレージ、メモリー、グラフィックスの活用度に関する IT パフォーマンスレポートをリアルタイムに生成する。AI の開発は Linux ベースのオープンソースソフトウェアで行われる傾向があり、Linux 上で稼働することは AI のデプロイ（配置）に有用であるとみられるが、今のところ DOP は Linux には対応していない。なお、DOP は ExpressSign-in、Express Charge（モバイル機器向け）、Intelligent Audio、レポートおよび分析ツールを提供しており、ワークステーションの設定を調整するのに有益である。

課題と機会

企業における課題と機会

IDC は AI 市場の二極化が進んでいるとみている。一部の企業ではデータ戦略を展開し、AI を大規模に取り入れて競争力を維持しようとしている。たとえば、上位 100 位内のスーパーコンピューターに実際に登録されている企業向けの AI インフラストラクチャ製品を使用して、驚異的なワークロードを実行するパートナーを利用している企業がある。一方で、いくつかの企業では、十分な予算がないために、データセンターやクラウド上で、多くは性能不足のハードウェアで構成されるサーバーで、小規模な AI の取り組みが日々行われている。

多くの企業にとって、前者の大規模な AI 活用事例はほとんど関係がなく、後者の小規模事例の方がまさに現実的であろう。こうした企業における課題は、クラウドインスタンスや GPU アクセラレーテッドデータセンターサーバーなどに、高額な出費をすることなく、AI 学習を適時実行できる適切なツールを、AI データサイエンティストや開発者に用意することである。こうした企業には、自社のサイエンティストや開発者に強力な GPU で処理能力を高めたワークステーションを提供することが有効であると IDC は考えている。

デルにおける課題と機会

AI の開発やデプロイ（配置）には高額なアクセラレーテッドサーバーのハードウェアや、多くの場合クラスターが必要になるという考えには一部誤解がある。これは、数十億個のパラメーターがあるような最大規模の AI アルゴリズムについては当てはまるかもしれないが、ほとんどの企業では、そのような大規模アルゴリズムは開発されていない。多くの企業では、有益かつインパクトがあり、管理できる範囲で、AI アプリケーションを実行している。しかし、こうした規模の AI モデルがワークステーション上で開発、配置できることは、ほとんど知られていない。デルの課題は、こうした AI に対する多くの企業が持つ先入観を払拭し、同社のワークステーションのポートフォリオについて、ユーザーや企業に理解してもらうことである。

同時にデルは、同社のワークステーションが今後も引き続き、役割を果たし、テクノロジーの進化のボトルネックにならないようにしなければならない。つまり、ワークステーションを適切に

使用している（言い換えると、数十億個の大規模なパラメーターアルゴリズムは実行していない）エンドユーザーを決して失望させないように、速いペースでイノベーションを継続的に生み出す必要がある。また、急にスケーリング（規模の拡張）に着手することになった顧客や、大規模なアルゴリズムを利用することになった顧客には、デルにおいてワークステーションから AI サーバーのラインナップへのシームレスな移行という選択肢が用意されていることも示唆している。もちろん、このことは、デルにとっては市場／ビジネス機会につながる。つまり、デルに求められるのは、顧客の AI 戦略の規模を問わず、すべての顧客に適切なソリューションを提供することである。

結論

今のところ、多くのユースケースにおける AI 開発およびデプロイ（配置）の主力製品として、ワークステーションは正当な評価を得ていないと IDC はみている。ワークステーションは、AI サイエンティストや開発者にとって、サーバーよりも資本支出が少なく、クラウドインスタンスよりもはるかに運用経費を抑えることができ、AI モデル実験の自由度がはるかに高い高性能な GPU アクセラレーテッドプラットフォームを提供している。数十億個のパラメーターアルゴリズムを必要としない（ほとんどの企業が対象になる）AI 戦略に取り組む企業は、AI 開発に制約を加えずに、容易なエッジベースのデプロイ（配置）を可能にするワークステーションを導入して、AI チームを強化することを検討すべきである。

IDC 社 概要

International Data Corporation（IDC）は、IT、通信、コンシューマー向け IT 分野に関する調査／分析、アドバイザリーサービス、イベントを提供するグローバル企業です。1964 年の設立以来、IDC は、世界中の企業経営者、IT 専門家、機関投資家に、テクノロジー導入や経営戦略策定などの意思決定を行う上で不可欠な、客観的な情報やコンサルティングを提供してきました。現在、110 か国以上を対象として、1,300 人を超えるアナリストが、世界規模、地域別、国別での市場動向の調査／分析および市場予測を行っています。IDC は、IDG（インターナショナル・データ・グループ）の系列会社です。

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com