White Paper

AI スケーリングの準備はできているか？ 絶対的なコア不足の影響を受けてはいけない

Sponsored by: IBM

Peter Rutten
May 2019

IDC の見解

AI によって獲得できるビジネスチャンスは極めて豊富に存在する。競合相手が AI を活用してこれまで入手不可能であったさまざまなデータや機能を手に入れ、成長に必要な力を備え、彼らの顧客を満足させられるのであれば、逆に AI を活用しないということは、将来、ビジネスにおいて最悪の事態を招く可能性があると、企業などの組織は認識している。現在、「AI は役に立たない」とか「AI はほとんど虚構である」と認識している企業はほぼ皆無であろう。むしろ、どのような業種や規模の組織であろうと、世界全体で AI に真剣に取り組んでいる。

多くの組織の事業部門（LOB）、IT スタッフ、データサイエンティストおよび開発者は AI について学び、ユースケースを理解し、自からのビジネスに対する AI 戦略を作成し、初期の AI への取り組みを開始し、それらに基づいて、機械学習（ML）アルゴリズム、特に深層学習（DL：ディープラーニング）を用いて、新たなインサイトや機能をもたらす AI アプリケーションを開発し、テストしてきた。

組織は現在、これらの取り組みや、新たに発生する課題を解決するために AI の活用領域を拡大する準備ができている。実際のところ組織は、これまでの体験から、一般的で、複数の目的に対応するために用意されたインフラストラクチャでは、AI ワークロードに十分対処できないことを学んでいる。また、組織は、AI の学習（AI モデルを訓練すること）と AI の推論（イベントを理解する、または予測を行うために訓練済みのモデルを使用すること）には、異なるコンピュート処理が必要であることを明確に認識するに至っている。しかし、その異なるコンピュート処理とは何であろうか。これとは別に、組織は、オンプレミス、クラウド、あるいはハイブリッドクラウドモデルのどちらを採用すべきであろうか。

大量のデータに対して指数関数的に解析を進める AI アプリケーション、特に DL システムは、処理要求のレベルが極めて高く、多数のコアを備えた強力な並列処理機能が求められる。このため、標準的な CPU ではこれらの AI タスクを実行するには不十分である。IDC の調査は、CPU に実装可能なコア容量を考慮すると、提供可能な CPU コア容量と必要とされるコア容量との大きなギャップは、次の数年間でさらに拡大することを示している。既存のインフラストラクチャを利用して AI 活用のトライアルを行い、また AI の活用領域を拡大する準備ができている AI ユーザーは、上述したギャップを埋め、必要な並列処理性能を確保するために、インフラストラクチャを見直す必要がある。これは、GPU、高速インターコネクト、大容量メモリーおよび高度な I/O 機能が組み合わされたマルチスレッド CPU によって達成されることになる。

概況

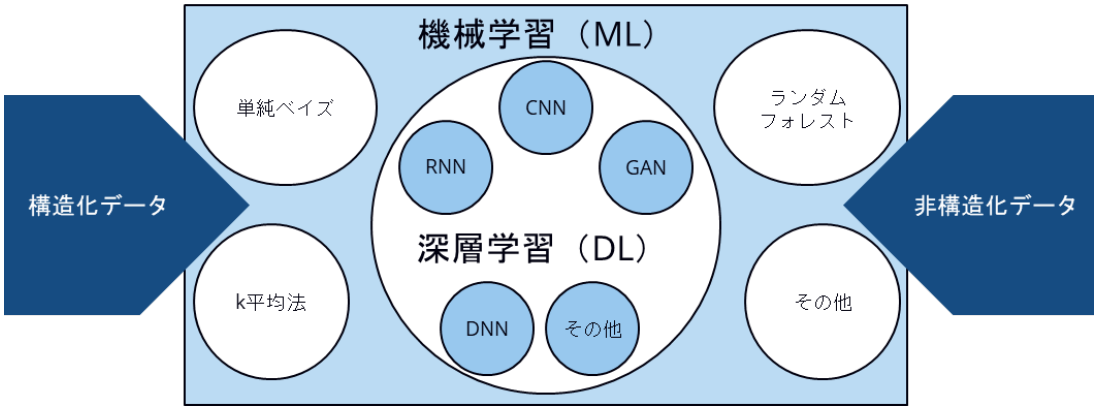
世界中の企業は AI のワークロードによってもたらされる新たなチャンスに積極的に対応している。IDC は AI を自然言語処理（NLP）、画像／動画アナリティクス、ML、ナレッジグラフに加えて、問題への回答、インサイトの発見および提言を行う技術を利用する一連の技術と定義している。これらのシステムは、入手可能なエビデンスを基に仮説を立てて答えを導き出す。また、膨大な

コンテンツを取り込んでシステムの学習ができる。システムは再学習や人による訓練を通してシステム自体の誤りや不具合に適応し、学習する。

ML は AI 技術の一つであり、人がプログラムする必要なく与えられたタスクについてコンピュータシステムが自ら対処方法を学習し改善できるようにするものである。ML モデルは、大量の構造化データや非構造化データを用いて、タスク（人の顔認識など）を「学習した」と思われるまで何回もテストを繰り返し長い時間をかけて改善できるアルゴリズムである。Figure 1 は DL が ML の一部であることを示している。一般的な DL アーキテクチャにはディープニューラルネットワーク（DNN）、畳み込みニューラルネットワーク（CNN）、回帰型ニューラルネットワーク（RNN）、敵対的生成ネットワーク（GAN）があり、その他にも多数が存在する。

FIGURE 1

機械学習と深層学習（DL：ディープラーニング）



Source: IDC, 2019

AI ソフトウェアプラットフォームには以下のものがある。

- 対話型 AI ソフトウェア（デジタルアシスタントなど）
- データの間に隠れている関係性を発見し、予測する予測的アナリティクス
- テキストを認識、理解し、価値を抽出するテキストアナリティクスと自然言語処理
- 音声を認識、特定し、情報を抽出する音声アナリティクス
- 画像および動画を認識、特定し、情報を抽出する画像および動画アナリティクス。これはパターン認識、物体、色、その他人間、顔、感情、車および風景などの属性を含む。

多くの組織において、AI への取り組みは着実に進んでおり、本格運用レベルで、AI を採用する準備が整う段階に到達した。AI を採用する準備が整っていないその他の企業は、AI の実験段階にあったり、組織にとって AI アプリケーションがどのような意味を持ち得るのかを評価する段階にあったりする。

最初のグループ（本格運用準備完了）について、IDC は企業、政府およびその他の組織が実装を開始したさまざまな AI のユースケースを調査している。現在の最も一般的な 5 種類のユースケースは以下の通りである（ハードウェア、ソフトウェアおよびサービスに関し、ユースケースに企業が支払った費用を基準に降順で示す）。

- **自動カスタマーサービスエージェント**：たとえば銀行業において、AI アプリケーションは顧客のニーズや問題を理解し、銀行が顧客の問題解決に要する時間やリソースを低減し

サポートする学習プログラムを用いて、カスタマーサービスを提供する。これらのエージェントは業界を超えて広く使われるようになってきている。

- **販売プロセスレコメンデーションおよび自動化：**さまざまな業界で使用されているもので、これらはリアルタイムで顧客のコンテクストを理解し、顧客関係管理（CRM）システムと連携して関連する対応を販売エージェントに提言する AI アプリケーションである。
- **自動化された脅威インテリジェンスおよび防止システム：**政府や産業全体に渡る脅威の防止において重要になりつつあるもので、AI アプリケーションはインテリジェンスレポートを処理し、そこから情報を抽出し、さまざまな情報同士の関係性を分析し、データベース、システム、Web サイトなどへの脅威を特定する。
- **不正分析と調査：**保険業界において、AI アプリケーションはルール学習を用いて不正なトランザクションを特定し、さまざまな保険関連の不正スキームを特定するため自動的に学習する。保険業界に限らず他の業界でも広く使われている。
- **自動予防的メンテナンス：**製造業では、AI アプリケーションが、プラントおよび機械の将来の故障に関する正確な予測モデルを構築する機械学習アルゴリズムを基にしており、ダウンタイムの短縮やメンテナンスコスト低減を行う。

AI のユースケースとして、新たにエンタープライズにおいて導入が進みつつある項目には以下が含まれる（ハードウェア、ソフトウェアおよびサービスの費用を基準に降順で示す）。

- プログラムアドバイザー／レコメンデーションシステム
- 診断処置システム
- インテリジェントプロセッシング自動化
- 品質管理調査とレコメンデーションシステム
- IT 自動化
- エンタープライズナレッジワーカーのためのデジタルアシスタント
- 専門的なショッピングアドバイザーと製品の提言
- 供給と物流
- レギュラトリーインテリジェンス
- 資産／フリート管理
- 自動クレーム処理
- デジタルツイン／最新のデジタルシミュレーション
- 公衆安全と緊急対応
- 適応学習
- スマートネットワーキング
- 貨物管理
- 薬物研究と創薬

クラウドとオンプレミス

これらのユースケースに対応するアプリケーションは、組織自身がカスタム開発するか、市販の AI ソフトウェアを基にするか、AI の SaaS として提供されるかのいずれかである。カスタム開発や市販のソフトウェアを採用する際に考慮することは、オンプレミス、IaaS でのクラウドまたはハイブリッドクラウドのいずれで使用するのかということである。ここでのオンプレミス環境はパブリッククラウド環境と連携する。

さまざまな採用シナリオにおいて、ソリューションは次の点が考慮されていなければならない。

- 極めて高パフォーマンスで AI モデルを学習するのに必要となる大量のデータの安全な処理。DL 学習のパフォーマンス要件は、高帯域幅でのデータ取り込みと GPU を使用した大規模な並列処理を実行する能力も含む。

- 極めて高パフォーマンスで AI モデルの推論に使用される大量データの安全な処理。推論に関するパフォーマンスは、学習済み AI モデルでデータを処理し、ほぼリアルタイムの AI によるインサイトまたは決定事項を提供する能力を意味する。

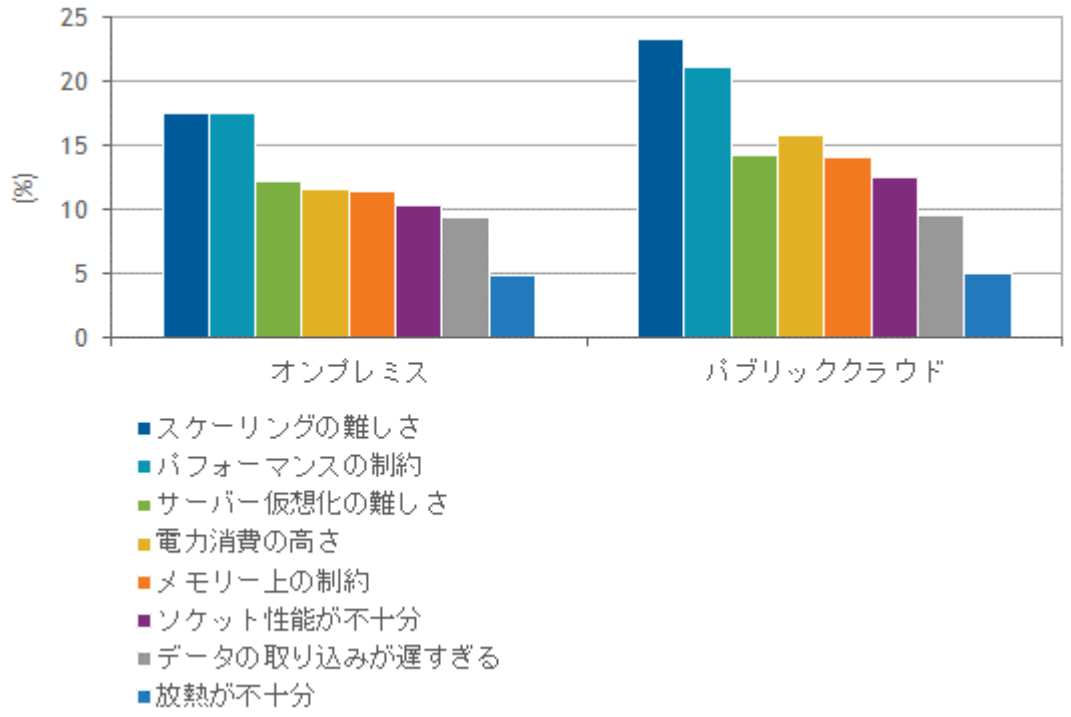
データサイエンティストや開発者にとって、AI への取り組みをクラウドで始める方が負担は少ない。DL 用にアクセラレーテッドコンピューティングインフラをオンプレミスで自ら準備しなくてよいからである。アクセラレーテッド AI クラウドインスタンスはほとんどのパブリッククラウドで提供されており、オープンソースの AI スタックによるものが一般的である。当然のことながら、AI 学習用アクセラレーテッド AI インスタンスのコンポーネント、つまりプロセッサー、コプロセッサー、インターコネクト、メモリーサイズ、I/O 帯域幅などは、クラウド SP（サービスプロバイダー）が決めることになる。すべてのクラウド SP がこれらのコンポーネントの最適な組み合せを提供するわけではないことから、最終的に、データサイエンティストが学習モデルを作成する上でのスピードと品質を決定しているのは、クラウド SP ということになる。その結果、多くの組織がオンプレミスの採用を選択している。

過去数年間の AI の実験を通して、多くの組織は、標準的なインフラストラクチャや基本的なクラウドインスタンスでは、彼ら自身が「行き詰まること」に気づいた。モデルの学習には時間がかかり、推論はゆっくりすぎる。IDC の調査では、回答者の 77.1%がオンプレミスの AI インフラストラクチャによって 1 つもしくはそれ以上の制約を受け、90.3%がクラウドにおいてコンピューティング上の制約を受けたことが示された。

Figure 2 はオンプレミスやクラウドで最も多く見かけるハードウェアに関わる制約を示す。多くの組織ではこれらのハードルが組み合わさって生じる。オンプレミスのハードルとして多く生じる順に回答を並べた。

FIGURE 2

実施中の AI ユースケースにおいてサーバーインフラストラクチャで遭遇するハードウェアの制約の上位



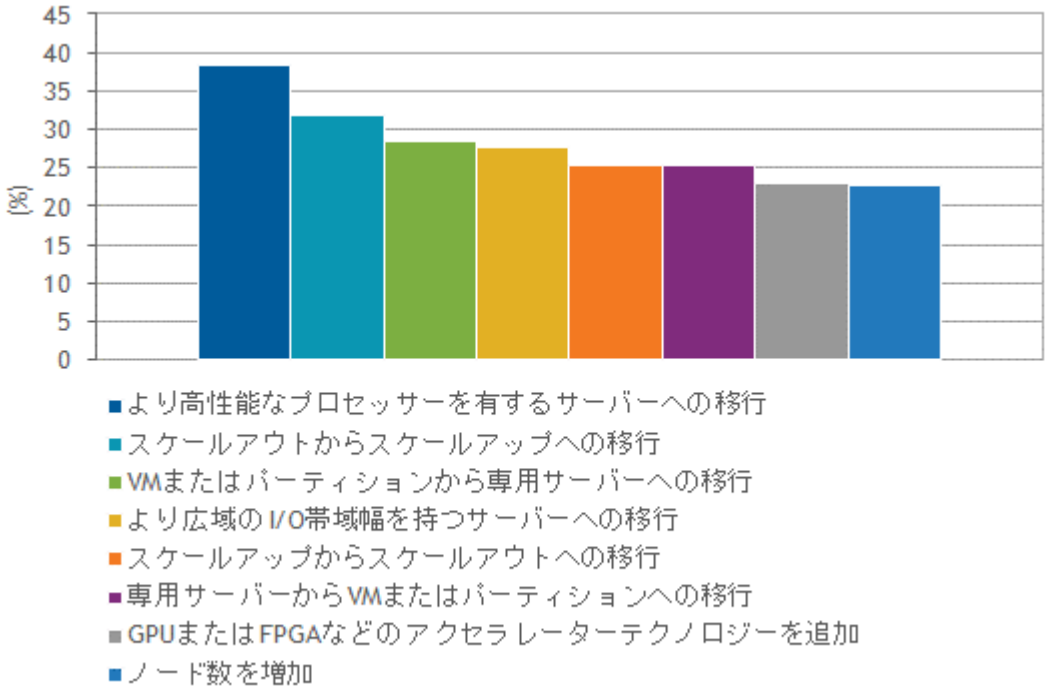
Source: IDC 2019

ハードウェア関連の制約を解消するために、IDC は企業が自社の AI インフラストラクチャを全面的に刷新するのを目の当たりにしてきた。しかも数年の間に 2 回である。Figure 3 は AI インフラストラクチャの刷新が持つ特徴を示している。さらに、ある企業が実施した AI インフラストラクチャ刷新の方向性は、他の企業が実施した刷新の方向性と相反する逆の方向に進んでおり、時として矛盾することを示している。

Figure 3 はまた、AI のパフォーマンスを高めるために行われたインフラストラクチャの刷新のうち最も多いものは、グラフの一番左の棒グラフが示すように、より高い性能のプロセッサーへの移行であることを示している。4 番目に多いのは、I/O パフォーマンスを改善し、AI のデータ取り込みを高速化することである。アクセラレーターの追加は、より一般的になってきており、ノード数の拡大も同様に一般的になりつつある。グラフの凡例に示した移行内容は、互いに排他的ではないことに留意いただきたい。たとえば、複数の回答者は自社のインフラストラクチャをスケールアウトすると同時に、アクセラレーターも採用していた。また、移行内容が正反対となっているインフラストラクチャの見直しもいくつかあるが（「スケールアップからスケールアウト」や「スケールアウトからスケールアップ」）、これらは業界に AI が導入された最初の数年において実施されてきた実験のいくつかを示している。

FIGURE 3

AI インフラストラクチャの世代的な移行の内容



Source: IDC, 2019

コアの絶対的な不足

企業が遭遇する制約の根本的な理由に、IDC が「コアの絶対的な不足」と呼んでいるものがある。AI は高度な数学と統計の計算を基にしている。たとえば、画像や動画のアナリティクスを考えてみよう。画像は各ピクセルを数字で表し行列に変換される。相関を見るため数百万の行列とその

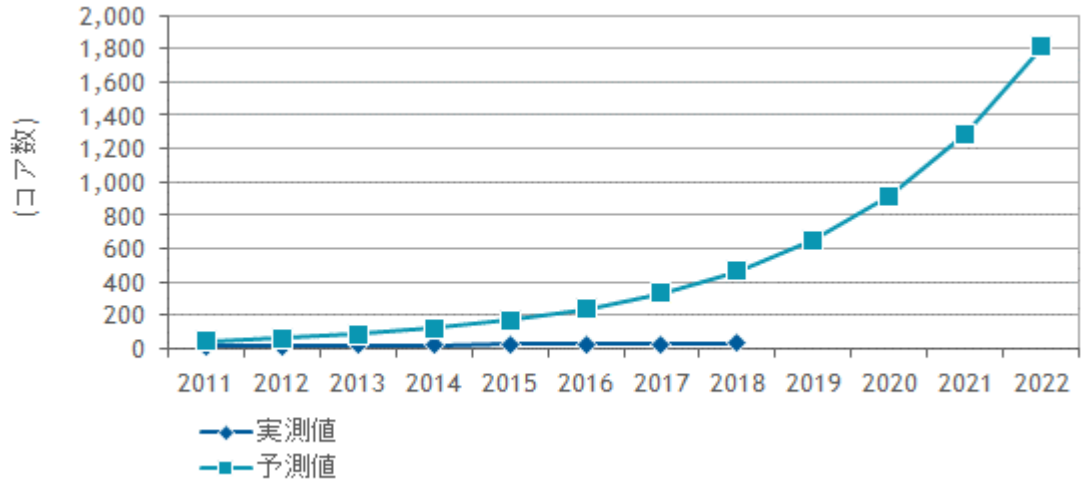
分類をニューラルネットワークへと送り込む。行列はその後お互いに掛け合わされ、正しい結果を導き出す（たとえば「犬」や「ソーダの缶」）。

このプロセスを高速化するためには、CPU に搭載できる数よりもはるかに多いコアを使って並列処理しなければならない。CPU は逐次処理用に設計されており、材料の物理的特性の制約によって潜在的な最大値にまもなく達してしまう。現在、すべてのプロセッサーメーカーは、ムーアの法則は終焉し、CPU の改良以外の補完的な方法によって数十年間続いてきた性能向上の軌跡を維持する必要があると認めている。コアのサイズとコストによって CPU のコア数が制限（数千単位ではなく数十単位）されることが主な理由である。

Figure 4 は 2011 年から 2018 年の CPU の性能の伸びの履歴、および 2011 年から 2022 年において CPU の性能が物理的制約に抑制されないと仮定して対数関数モデル化した CPU の性能曲線を示している。モデル化された CPU 性能は CPU 性能に物理的な制約がないと仮定（現状の逆）している。つまり CPU 性能に対するニーズがどのように発展していくかについての見解を示している。また、Figure 4 は CPU 性能のニーズと実現可能な CPU 性能の間のギャップを示している。

FIGURE 4

すべてのワークロードにおける世界的に必要なコア数の実際値と予測値



Source: IDC, 2019

今後、GPU（数千のコアを伴う）やカスタム設計のプロセッサー（ASIC、FPGA）が増加すれば、実際の CPU 性能のニーズと予測性能のニーズ間のギャップが埋められる。これらのアクセラレーターは、必要な並列コンピューティングの性能を手ごろな価格で提供できるよう、半導体のダイ上に数百または数千にも及ぶコアを搭載している大規模な並列アーキテクチャを有している。これらのコプロセッサーの影響は、コプロセッサーの形式による他のベネフィットと組み合わさり、明確に性能を向上させてきた。同時に、これらのコプロセッサーに対し膨大な量のデータを供給する技術（コプロセッサー同士および CPU とコプロセッサーのインターコネクト、メモリーサイズの拡大、および高速ストレージなど）は、重要性が高まった。

IDC はアクセラレーテッドサーバーの世界市場が 2022 年には 256 億ドルまで成長し、CAGR は 31.6%になるとみている。これらは、オンプレミスおよびクラウドにおいて設置され、GPU、FPGA または ASIC のいずれかで高速化されているサーバーである。実際、この市場は急速に成長しており、IDC はアクセラレーテッドコンピューティングが市場において非アクセラレーテッド

コンピューティングを侵食し始め、2021 年までにアクセラレーテッドコンピューティングが世界的なサーバーに占める割合は 12%に至ると予測している。

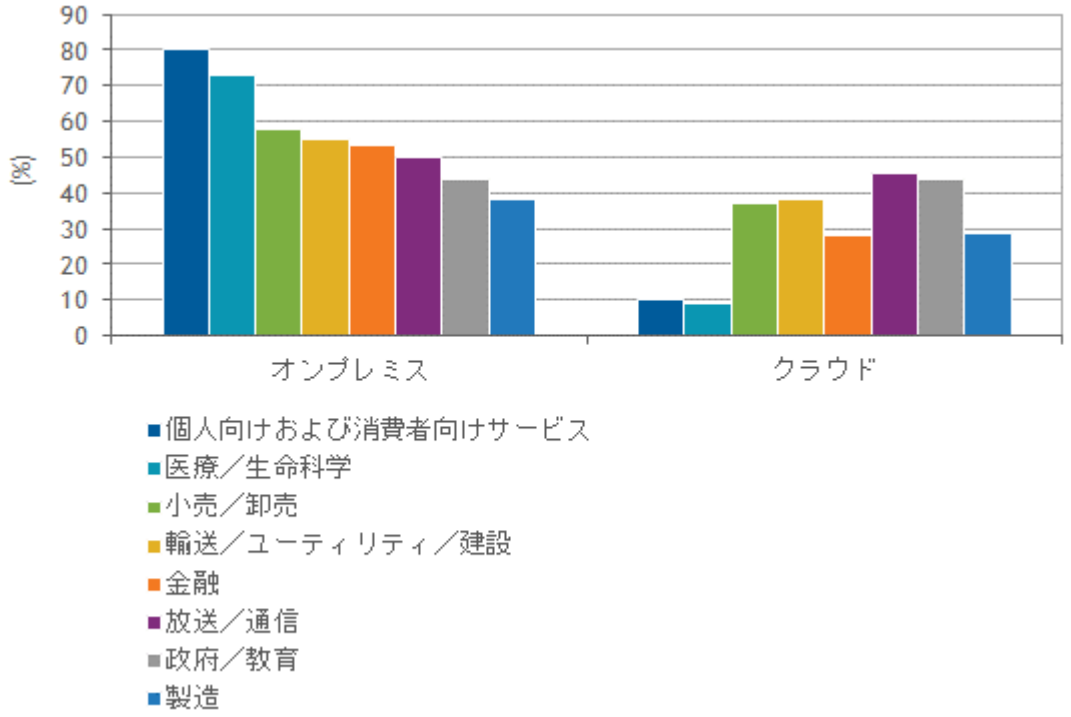
しかし、高速化は比較的新しいテクノロジーである。IDC の調査では、自社のインフラストラクチャの性能を向上させようと組織がアクセラレーターを使用する話は、ここ約 2、3 年のことである。IDC は、量子コンピューティングなどの本当に新しいコンピューティングの形式が主流となるまでの間は、複数ワークロードのアクセラレーターによる高速化が標準的な使い方といて定着し、アクセラレーターを搭載したインフラストラクチャの性能向上がコンピューティングにおいて長く続くと予測している。さらに、量子コンピューティングなどの新しいコンピューティングが主流になったとしても、アクセラレーター搭載の従来型コンピューティングは、多くのワークロードにとって標準となり続けるであろう。

高速化が最高レベルとなっているネットワーキング、暗号化、セキュリティおよび圧縮などのワークロードは別として、高速化のため、コプロセッサーの浸透が高まっているワークロードはリアルタイムアナリティクス、AI の DL、Hadoop およびデータベース、動画／画像／音声認識、シミュレーションおよびモデリング、AI 推論である。

今日まで、オンプレミスはアクセラレーテッドサーバーに適した配備環境とされてきている。Figure 5 は、GPU アクセラレーテッドサーバーの業種別にみたオンプレミスおよびクラウドでの導入率を示している。また、Figure 5 はオンプレミス環境で最も多く GPU を使用しているのは個人向けおよび消費者向けサービス業界であり（回答者の 80%）、クラウド環境では放送／通信業界である（45.5%）ことを示している。

FIGURE 5

産業ごとの GPU アクセラレーテッドサーバーのオンプレミスおよびクラウドにおける採用の比較



Source: IDC, 2019

興味深いことに、オンプレミス環境での採用は今後 12 か月で増加するとみられ、クラウド SP はこれが示す問題点を認識しなければならない。たとえば、高速化のために GPU を使用する企業の 50.8%が、現在オンプレミス環境での採用と回答しているが、この割合は今後 12 か月のうちに 54.2%まで増加する。同様にオンプレミス環境での ASIC や FPGA の使用も約 5%増加するとみられる。

クラウド SP がさらに留意すべきことは、かなり多くの企業がパブリッククラウドで運用していたアクセラレーテッドワークロードをオンプレミス環境へ戻している（「回帰」とも呼ばれる）という事実である。組織がワークロードをオンプレミスへと戻している理由は、クラウドの使用時に経験したコスト、セキュリティ、スケーリングおよびパフォーマンスの問題である。組織の 66%が過去にアクセラレーターを必要とするワークロードをクラウド上で開始したが、その後自社のオンプレミスのデータセンターにそのワークロードを移行させたと話している。

アクセラレーテッドサーバーによって得られる費用対パフォーマンス向上は業種ごとに異なっており、ある業種では、平均 2 倍となっている。一方、オンプレミスの CAPEX（アクセラレーテッドサーバーの取得費用）やクラウドの OPEX（アクセラレーテッドインスタンスの運用費用）の増加率も業種によって異なるが 26%から 33%である。実際には CAPEX や OPEX の増加率は採用するインフラストラクチャに大きく左右される。ある種のアクセラレーテッドサーバーを採用することによって、パフォーマンスの向上率はより高くなるからである。いずれにせよ、組織はこれが高い ROI であると考えなければならず、それと同時に AI が現状に対して 2 倍を超えるパフォーマンスを求め始めている。このような状況に対処すべく、膨大な数の新たなプロセッサーおよびコプロセッサーの研究がハードウェアのスタートアップ企業や IBM、インテル、AMD、Xlinx、NVIDIA などで進行中である。また、IDC の調査によると、高速化技術について、それをベンダーが取り入れることをほとんどの組織が期待しており、少数の組織がシステムインテグレーターや VAR 主導での活用を検討し、さらに少数が自社の IT チームによる導入を望んでいることが明らかになっている。

GPU 用に CUDA、OpenACC または OpenMP などを使用してアクセラレーターのプログラミングを行うには、スキルを持つスタッフが必要であり、業種によるばらつきはあるが（アクセラレーターのタイプは無関係）、組織はこのプログラミング用に平均で 2.4 FTE（Full-Time Equivalent：常勤換算での仕事量）から 3.7 FTE のスタッフが必要であったと回答した。GPU は最もプログラムがしやすく、FPGA が幾分難しく、ASIC は開発期間が長引く傾向にある。多くの企業はアクセラレーターのプログラミングスピードに非常に満足していると回答している。言い換えると、これは障害と考えるべきではないということである。

企業は、最終的にパフォーマンスを向上させることが、アクセラレーテッドサーバーを選択するための最も重要な目標であると述べている。パフォーマンスは、データサイエンティストが AI モデルの学習終了を待たなければならない時間、AI モデルの詳細さおよび正確さ、学習された AI モデルに対して行われる推論のスピードに直接影響を与える。言い換えれば、パフォーマンスの高さは結果が出るまでの時間や結果の正確さを意味する。コストは二の次で、新たなスキルセットの必要性はそれほど重要ではない。誰に確認しても、組織は AI の力を欲しており、投資に際して、コストや新たなスキル獲得の必要性については懸念していない。

推論

AI モデルの学習と、そのモデル上での推論に求められる要件の違いについては意見が分かれている。IDC では、推論に使用されるサーバーインフラストラクチャへの需要は、学習用のサーバーインフラストラクチャよりも次第に大きくなるとみている。これには、「旧式の」機械学習（近傍法または単純ベイズ分類器など）や DL（たとえば、音声または画像や動画アナリティクス向け）などの、すべての形式の ML 用のインフラストラクチャが含まれる。世界サーバー市場の支出額で見ると、推論用は 2020 年までに学習用を超えるであろう。推論に使用されるサーバーインフラストラクチャについても、ハイパフォーマンスコンピューティングや、多くの場合にアクセラレーテッドコンピューティングを必要とするようになる。

AI 推論を AI 学習とは異なるワークロードとして捉えるべき要因にはさまざまなものがある。推論は学習された AI モデルにデータを入力する。多くのアプリケーションにおいて、推論は、ほぼリアルタイムで結果を出せるように最小限のレイテンシーで行う必要がある。データは小規模のもの（1 回の画像認識タスクなど）もあれば大きいもの（公道上で常時行われるリアルタイムの顔認識タスクなど）もある。推論で用いられる全体的なサーバーインフラストラクチャの要件は、学習に対するものとは大きく異なる場合があると一般に理解されている。具体的には、学習では特定のタイプのアクセラレーション（GPU などを使用）が大多数を占めているのに対し、推論では非常に繰り返しの多い推論処理（FPGA や ASIC などを使用）が極めて大規模に実行される。

推論の種類によっては、高性能なホストプロセッサー上で十分実行可能な程度に負荷が軽いため、コプロセッサーによる高速化は必要ではないことがある。これは、AI イネーブルド（AI 機能を備えた）アプリケーションが該当する（完全 AI アプリケーションとの対比で用いる用語）。AI イネーブルドアプリケーションでは、アプリケーションの一部を AI 機能が補完する。たとえばビジネスソフトウェアの調達機能の一部で対話型インターフェースを使用する。高性能 CPU であれば、そのようなアプリケーションに対して、アクセラレーターである GPU などのコプロセッサー上に AI 処理を移行させる必要はないかもしれない。AI 推論は、クラスタリングによって性能が向上するという面もあるが、マルチソケットのスケールアッププラットフォーム上でも実行できる。

AI インフラストラクチャでの検討項目

ここ 2、3 年において、AI アプリケーションに適したインフラストラクチャを見極めるため、多くのトライアルが実施された。企業は、ハイパーコンバージドから、スケールアップ、スケールアウトまですべてを試した。その後、AI の実装が成熟し、スケーリング（実運用での用途拡大）が進むにつれて、AI アプリケーションに適したインフラストラクチャはクラスタリングからまず性能上のベネフィットを得られるとの合意が形成されたと IDC では見ている。AI は並列処理される性質を持っているため、1 つのアクセラレーターで数百のコア、1 サーバーノードで複数のアクセラレーター、および 1 サーバークラスターで複数のサーバーを利用できることは、パフォーマンス上の利点となる。

AI インフラストラクチャのスケーリングを始める企業は、程度は異なるものの、一般にはオンプレミスでの導入やクラウドにて実現可能な検討項目のリストに目を通す。したがって、企業は、オンプレミス、クラウドまたはハイブリッドクラウドにおけるアクセラレーテッド AI インフラストラクチャをスケーリングしたいかどうかを判断する前に、最も重要なものは何かを判断することに時間をかける。

考え得る最高の AI インスタンスを顧客に提供したいと考えるクラウド SP にとって、これらの検討事項は重要であり、それは特に、アクセラレーションが必要なアプリケーションのオンプレミス展開の増加が見込まれる場合に重要である。したがって、クラウド SP からの「acceleration-as-a-service」製品は、クラウド SP が最適な AI クラウドプロバイダーと評価されたい場合には、考え得る最適なオンプレミス製品と同等である必要がある。

Table 1 は、アクセラレーテッドシステムにおける複数のハードウェア、ソフトウェアおよびデータセンターの検討項目の概要を示す。Table 1 を用いることで、個々の項目の重要度（アクセラレーテッドシステムユーザを対象とした IDC による調査を基にしている）を踏まえて、オンプレミスとクラウド SP のいずれで達成可能かを検討できる。

パフォーマンス、メモリー、セキュリティ、高可用性、仮想化およびインターコネクトの帯域幅は重要で、現状、クラウドよりもオンプレミスの方が実現しやすいと考えられるサーバー特性である。現実的なアプローチとしてハイブリッドクラウド環境も挙げられる。

TABLE 1

アクセラレーテッドサーバーの検討事項：重要度とオンプレミスおよびクラウドにおける達成可能性

	重要度	オンプレミスでの達成可能性	クラウドでの達成可能性
ハードウェア			
ホスト CPU の性能	●	●	◕
アクセラレーターが使用できるメモリーの数	●	●	◕
アクセラレーテッドサーバーのセキュリティ	●	●	◕
アクセラレーテッドサーバーの高可用性	●	●	◕
アクセラレーテッドシステムのサーバーの仮想化	●	◕	●
サーバーノード内でのアクセラレーターのスケールアップ	◕	●	◕
アクセラレーターとホスト CPU 間のインターコネクトの帯域幅	◕	●	◒
アクセラレーターによる性能向上	◕	●	●
アクセラレーテッドサーバーの能力要件	◕	◒	●
アクセラレーテッドサーバーノードのスケールアウト	◕	●	●
アクセラレーテッドサーバーからの放熱	◕	◒	●
ソフトウェア			
アクセラレーターのプログラミングの容易さ	●	●	●
アクセラレーテッドサーバー上での診断	●	●	◒
API、ライブラリー、ソフトウェア開発キット、ツールキット、フレームワーク、プログラミング言語などが利用できること	◕	●	◕
アクセラレーターのプログラムに必要な時間	◕	◕	◕
アクセラレーターのプログラムにかかるコスト	◕	◕	◕
OpenCL、OpenMP および OpenACC などのオープンソースへの対応	◕	●	◕
データセンター			
他のインフラストラクチャとのアクセラレーテッドサーバーの相互運用性	●	◕	◕
運用環境に必要なスキルのレベル	●	◕	◕
アクセラレーテッドサーバーの管理容易性	◕	●	◕

Note: ●（高い）、◕（やや高い）、◒（平均）

Source: IDC, 2019

IBM POWER SYSTEM AC922

IBM Power System AC922 は、世界最速のスーパーコンピューターのビルディングブロックで、オークリッジ国立研究所の Summit で採用されている。IBM がこの功績について 2018 年中ごろに発表したが、非常に控えめな発表であったと IDC は考えている。Summit は、200 ペタフロップスの能力を持ち、exaops に到達した最初のスーパーコンピューターである。これは 1 秒間に 10^{18}（exa）回の計算を行うことを意味する。Summit はまた、世界で 3 番目に地球にやさしいスーパーコンピューターである。興味深いことに、Summit はシミュレーションやモデリング用に構築されている他の多くのスーパーコンピューターとは異なり、AI 用に構築された。IBM は、膨大な AI ワークロードをスケーリングする企業向けに高速コンピューティングプラットフォームを提供している。一般にはこれを「mini-Summit」と呼んでいる。

このような規模で運用していない組織においても、1 台の Power System AC922 または小規模から中規模の Power System AC922 クラスターの優れたパフォーマンスを活用できる。高密度設計によって PCIe Gen4 と InfiniBand を組み合わせることで、組織は 1 台の Power System AC922 ノードから始めることができ、その後は、ほぼ線形のスケーリング効率を持ったラックまたは数千ノードへのスケーリングも可能であると IBM は主張している。

IBM は数年前に Linux ベースのスケールアウト（型）Power Systems の構築を開始しており、その際、極端にデータ集約型コンピューティングのためのシステム構築に焦点を当てていた。その戦略によって市場に先駆けたが、そのとき AI やアナリティクスはまだ出始めたばかりのワークロードであった。同時に、IBM は、ミッションクリティカルデータのセキュリティで信頼性が高いプラットフォーム Power Systems の評判を、これらの新たな Linux ベースのスケールアウトシステムにまで広げた。

これらの取り組みは、1 ソケットおよび 2 ソケットの Linux サーバーに幅広いポートフォリオをもたらした。これらは業界で最も高いコア性能を示している。Power System AC922 は、このラインナップのモンスターマシンであり、4 個または 6 個の NVIDIA Tesla V100 GPU を搭載し、POWER9 プロセッサーと、NVIDIA のインターコネクト NVLink2（NVLink 第 2 世代）によって統合した 2 ソケット 2U システムである。NVLink2 はシームレスな CPU-GPU 接続とコヒーレントなメモリーシェアリングを実現する。コヒーレンスによって、システムはシステムメモリーを GPU メモリーと同様に処理でき、プログラミングの簡素化とはるかに大規模な AI モデルが可能となる。現在、CPU-GPU 間の超高速接続用プロセッサーに直接 NVLink を組み込んでいるサーバープラットフォームは他に存在せず、これは、このリンクを通した DRAM への GPU の高帯域幅アクセスを可能にしている。

Power System AC922 についての特筆すべきポイントは他にもある。コア当たりのレベル 2 キャッシュは 512KB で、レベル 3 キャッシュは 10MB 以上、システムメモリーは 256 から 2,048GB まで使用可能である。スレッド数は 8 で、POWER プロセッサーのコアは x86 ソリューションのスレッド数の 4 倍となっている。これは、AI のような並列ワークロードにおいて大切なポイントである。Power System AC922 は高スループットのオンチップファブリックを有しており、毎秒 7TB（テラバイト）でデータを移行できるオンチップスイッチを含む。各コアとのデータの出し入れは毎秒 256GB となっている。NVLink2 だけではなく、コヒーレントかつ極めて高帯域幅の ASIC や FPGA または外部フラッシュストレージへのアタッチメント用として CAPI 2.0、および業界初の PCIe デバイスに接続するための PCIe Gen 4 も含まれている。

冷却は、4GPU 構成の Power System AC922 では空冷または水冷、6GPU 構成では水冷のみが用意されている。これもまた、IBM が他のベンダー数社と共に市場に先駆けて提供している。水冷は現在のアクセラレーテッドコンピュータ時代において、ノード当たりの GPU が増え、ラック内のノード数が増えるという高密度化を考慮すると、新たな、ある意味では再注目された技術である。

AI に関する Power System AC922 のベネフィットは、このシステムによって AI の学習時間が短縮できることにある。IDC は、学習時間は、組織が AI ソリューションを使って本番へ進める際の重大な懸案事項となっていると捉えている。データサイエンティストは、学習モデルが完成するまで数日から数週間も待たなければならないことが多く、微調整や、やり直しによって多大な時間を消費しなければならない。
Power System AC922 はデータサイエンティストがモデルの微調整や、やり直しをはるかに速くできるようにしている。その上、GPU からシステムメモリーにほぼ直接的にアクセスできるため、データサイエンティストは IBM の Large Model Support（LMS）を用いて学習できる。これはデータサイエンティストがより大規模で複雑なモデルを使用し、また、GPU メモリーのみに頼る場合よりも正確に作業できることを意味する。

スケーリングを可能にするため、IBM は Distributed Deep Learning（DDL）というライブラリーを開発した。これによって、データサイエンティストは完全にシームレスな方法で数百台のサーバー全体をほぼ線形にスケールアウトすることが可能となる。データサイエンティストはシンプルに呼び出し、必要な数の GPU のリクエストをしさえすればよい。DDL は、TensorFlow、Caffe、Torch および Chainer などの ML フレームワークとつながっており、これらのフレームワークを複数の GPU へスケーリングすることが可能になる。その結果、データサイエンティストは数十の GPU 全体で学習プロジェクトをスケーリングしやすくなる。これは DDL がなければ実現が困難である。また、データサイエンティストはシステムに動的な管理を委ねることができる。

Table 2 において、エンドユーザが重要と考えるハードウェアの機能に対して、Power System AC922 の対応状況を検証する。

TABLE 2

アクセラレーテッドサーバーの重要なハードウェア機能への IBM Power System AC922 対応状況

ハードウェア	
ホスト CPU の性能	✓
アクセラレーターが使用できるメモリー数	✓
アクセラレーテッドサーバーのセキュリティ	✓
アクセラレーテッドサーバーの高可用性	✓
アクセラレーテッドシステムの仮想化	✓
サーバーノード内でのアクセラレーターのスケールアップ	✓
アクセラレーターとホスト CPU 間の相互接続の帯域幅	✓
アクセラレーターによる性能向上	✓
アクセラレーテッドサーバーの能力要件	✓
アクセラレーテッドサーバーノードのスケールアウト	✓
アクセラレーテッドサーバーからの放熱	✓

Source: IDC, 2019

Power System AC922 は、オンプレミスであるか IBM クラウドであるかを問わず、IBM の新たなオープンソースベース AI スタックのハードウェア基盤を表している（ただし、同様に IBM は x86 ベースのハードウェアで使用できるこのスタックの多くを作成してきたことに留意）。Power System AC922 において、スタックは高度に最適化されている。DL はアクセラレーターに大きく依存するため、IBM はソフトウェアを最適化し、Power System AC922 で 4GPU または 6GPU、NVLink および Power System AC922 サーバーのクラスターを活用してきた。このスタックは IBM クラウドを含む、プライベートクラウド、パブリッククラウドで利用でき、次の機能も提供している。

- **データプレパレーションおよびモデル開発のための Watson Studio、Jupyter および RStudio：**Watson Studio には現在、Watson Machine Learning Community Edition（Watson ML CE）が含まれ、本質的には PowerAI を Watson Studio に統合。これは無償で提供されている。
- **プライベートクラウドおよびパブリッククラウドの両方において AI モデルの学習、導入および管理を行う実行環境である Watson Machine Learning：**これは、IBM Cloud Private または Kubernetes ベースのアプローチのいずれかであるプライベートクラウドのオプションによって、他のモデル管理のパブリッククラウドソリューションから IBM を差別化している。Watson Machine Learning は、たとえば、Spark、TensorFlow、PyTorch、Chainer、Keras、および IBM の新たな「従来型」機械学習性能ブースターの SnapML を含む。SnapML はロジスティック回帰、決定木、およびランダムフォレストのようなデータサイエンス手法として非常に人気があることを示してきた。また、ここには、Watson ML Community Edition や、現在 Watson ML Accelerator と呼ばれているソフトウェアも含まれている。Watson ML Accelerator は、IBM の DL モデルの学習用のソフトウェアである PowerAI Enterprise というブランドであった。Watson ML Accelerator は、複数のデータサイエンティストが同じインフラを共有することを想定したリソース管理や、インフラストラクチャ全体で単一ジョブの弾力的なスケーリングに焦点を絞っている。
- **AI モデルの指標と AI モデルのバイアスおよび公正なモニタリングを提供する Watson OpenScale：**OpenScale は、AI モデルパフォーマンスモニタリングを除けば、アプリケーションパフォーマンス管理と同類のものである。これは、モデルの正確性を追跡でき、またはバイアスおよび公正に関する一定の指標を実装してその指標についてモデルをチェックできる。

このスタック上で、組織はさまざまな AI アプリケーションを実行できる。興味深い例は、IBM PowerAI Vision である。これによって、組織は非常に容易かつ迅速にさまざまな種類の画像の分類や検出、特にコンピュータビジョンの DL のニューラルネットワークモデルを開発できる。PowerAI Vision は、インストールおよびコンフィギュレーションの完全なライフサイクル管理、データラベリング、モデルの学習、推論およびモデルの本番環境への移行など、包括的なワークフローのサポートを行う。PowerAI Vision は、ドローンによる調査、職場や工場における安全規定の執行、製造品質検査、都市交通管理、その他多くのユースケースに活用できる。

将来の展望

IDC はいわゆる IDC FutureScape やその他の調査レポートにおいて、AI の将来、予測の公開、要因、IT への影響、および提言に対する幅広い調査を行ってきた。本調査レポートの紙面は限られており、これらの予測のほんの一部でさえ把握するのに十分であるとは言いにくい。あえて言うなら 2024 年までに、AI は私たちがどのように生活し、業務を遂行し、あるいはデータセンターを運用するか、これらすべての様相を劇的に変化させるであろう。

短期的には、最も重要なステップを踏むことで、データサイエンティストがソフトウェアやハードウェアツールを用いて、適切で高品質な AI ソリューションを素早く構築できるようになる。データサイエンティストが AI に対する期待を実現するのに必要な計算は実に驚くべき数である。アクセラレーテッドコンピューティングはデータ駆動型ワークロードについての例外というよりも

基準となることが予測される。GPU（および FPGA や ASIC）がこれまで想像できなかった処理性能をもたらしていると同時に、新たな AI プロセッサーが、スタートアップ企業や IBM などの既存のテクノロジーカンパニーの両方において発明されつつある。これらの新たなプロセッサーは 10 倍（100 倍ともされる）の処理性能をもたらすと主張している。同時に、新たなソフトウェアモデルはすべてのものへの AI の導入を促進するために発明の途上にある。

AI モデルの学習は弱まることなく続き、モデルはより大規模に、複雑になり、バイアスをなくすなどさらに正確さを要求されていく。これらのモデルでの推論は間もなく、エッジにおいて最大のワークロードとなり、すでにエッジにおいては再学習が行われている。このすべてを要約すると、ユースケース設計やモデル開発に関する AI の勢いは止めようがない（つまり、組織がその勢いをサポートする適切なインフラストラクチャの構築および運用に失敗しない限り）AI インフラストラクチャによって実験的な「行き詰まり」となる日々は終了する。ビジネス上重要な AI アプリケーションのスケーリングを初めている多くの組織では、インフラストラクチャへの投資を最優先事項にしなくてはならない。

課題／機会

IT バイヤーにとって

本ホワイトペーパーでは、AI アプリケーションの本稼働を拡大する準備が整った際に、組織が直面するさまざまなインフラストラクチャの課題について議論してきた。モデル開発に対するデータの準備から、AI モデルの学習、導入および実行環境の運用管理まで、基礎的なインフラストラクチャの要件は、一般的な用途のハードウェアに求められる古いモデルとは異なるものである。データ集約型ワークロード用に設計されたインフラストラクチャのみが、優れたコア性能、複数の GPU、高速インターコネクト、大量のコヒーレントメモリー、非常に優れた I/O 帯域幅を有し、DL の学習ワークロードいついては処理時間に求められる制約を満たした上で実行できる。組織では既存の一般用途のハードウェアを AI 専用ハードウェアに置き換えるか、既存ハードウェアをその AI 専用ハードウェアで補うかの判断が求められている。明らかに、AI によってもたらされる市場機会は、組織が最新の AI アプリケーションを開発し、実行できるようにする処理能力を備えることによって実現されるといえよう。

IBM にとって

IBM Power Systems の課題は常に市場での認知度にまつわるものである。IBM は、綿密に考え抜いた AI ソフトウェアスタックでパッケージ化された優れた AI インフラストラクチャソリューションを提供しているが、潜在顧客の多くは、誤った認識、つまり「自分たちの求めるソリューションではない」、あるいは割高である、と受け取っている。その結果、膨大なデータ集約型ワークロードに対して、大規模なコモディティハードウェアベンダーの中から 1 社を選択するというワンパターンなリアクションをとることで、組織が本当にベネフィットを得られる AI インフラストラクチャソリューションをその組織から奪っている。たとえば、Power System AC922 はどのデータセンターにおいてもスーパーコンピューターのビルディングブロックとなり得る。IBM は Watson ブランドの下で自社の AI インフラストラクチャと AI ソフトウェアのブランディングを整理統合している。長期的には、これは理にかなっており、IBM と顧客の双方にとってベネフィットになる。ただし、短期的には、新たなサブブランドを明確にし、顧客や業界のアナリストの混乱をなくすためにさらに多くのことに取り組まなければならない。IBM にとっての市場機会は、IBM の優れた技術的能力にある。現在は、新たな AI ワークロードが現状のオンプレミスおよびクラウドのインフラストラクチャが抱える性能面での課題を顕在化させている。これは IBM にとって潜在顧客の認識を変え、市場機会を捉える好機といえよう。

結論

ここ数年において、IDC は多くの組織がさまざまな AI 機能の開発に着手したのを目の当たりにしてきた。これらの取り組みは比較的経験の浅いスタッフによる試みとして、利用可能なさまざま

なインフラストラクチャで実施されるところから始まり、現在はクリティカルマスを超えようとしている。多くの組織では、広範囲に渡る AI の専門知識を身につけ、組織の AI 機能が自社ビジネスの重要な側面を担えるように、スピード感を持って直接経験を積み上げている。

同時に、IT も AI が実行されるインフラストラクチャに関する学習曲線に沿って進展してきた。現在、DL 学習や推論について、また製品について、これらの環境をスケーリングすることがインフラストラクチャにとって重要な要件であるということが、かなり明確になってきている。この DL 学習が他のワークロードとは異なるインフラストラクチャを必要とすることは、ほぼ既知の事実である。DL 学習は、高い能力のプロセッサー、高性能なコプロセッサー、高速インターコネクト、高 I/O 帯域幅および大量のメモリーを有するクラスター化ノードを必要とする。

現在、IT が決断しなければならない最大のポイントは、前述のコンポーネントや、これらをどのように相互接続し最適化するかを考慮した上で、AI ワークロードを稼働させる各ノードのパフォーマンスをいかに向上させるかということである。なぜならそのパフォーマンスは、単に大量の GPU を使うことではなく、GPU の持つ処理能力を最大限に引き出すプラットフォームを使って初めて達成可能になるためである。IBM の Power System AC922 は、AI 学習実行時のノードごとのパフォーマンスを最大にする優れた選択肢であると IDC は考えている。つまるところ、Power System AC922 は世界最速のスーパーコンピューターのビルディングブロックである。

IDC 社 概要

International Data Corporation（IDC）は、IT および通信分野に関する調査・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。50 年にわたり、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、1,100 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 は世界をリードするテクノロジーメディア（出版）、調査会社、イベントを擁する 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

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com