Sponsored by: IBM

Authors:
Peter Rutten
David Schubmehl

September 2017

人工知能の導入時に遭遇するサーバーインフラの壁

IDCの見解

ディープラーニング（深層学習）を使用して新たな分析と洞察を得るための人工知能（AI）アプリケーションに対し、どのように取り組むべきかを判断するため、企業は膨大な数の検討課題を前に奮闘している。そこには並はずれて大きなビジネスチャンスが約束されている。かつては利用できなかったデータ資産を競合企業が手に入れ、顧客基盤を成長させていく状況では、何もしないことがビジネス上の大惨事につながる恐れがある。大半の企業はこうした問題に気付いており、ビジネス部門（LOB）、ITスタッフ、データサイエンティスト、開発者はAI戦略を確立しようと躍起になっている。

企業が重要な決定を下す必要があるにもかかわらず、新たに出現したこの環境で勝ち抜くためのAI戦略は、現在まったく確立されていないとIDCはみている。企業は社内で開発を行うべきか、もしくはVAR、システムインテグレーター、あるいはコンサルタントを使うべきか。開発は、オンプレミス、クラウド、ハイブリッド環境の、どれで進めるべきか。既存のインフラを使用できるのか、あるいはAIアプリケーションやディープラーニング向けの卓越した能力を持つ新規サーバーが必要になるのか。

IDCでは、こうした質問に対する答えの多くは、十分に検討、調整された小規模な社内プロジェクトからスタートし、そこから発生するさまざまな影響や変化を注意深く観察し、徐々に規模を拡大する過程を通して得られると考えている。

こうした道筋の中で企業は、AIアプリケーションで先行する企業がすでに経験したことを改めて体験することになる。サーバーのパフォーマンスという壁に突き当たる。指数関数的に増加していくデータを解析するAIアプリケーション、とりわけディープラーニングが要求するパフォーマンスは極めて高く、強力な並列処理能力が必要であり、標準的なCPU (Central Processing Unit) では、十分に実行できないことが明白になってきている。初期段階と発展段階にいるAIユーザーは、必要な処理能力を確保するため、インフラストラクチャの見直しを実施しなければならない。

このためAIの能力を開発しようとしている、あるいは既存のAIの能力を拡張しようとしている企業は、十分に制御できる状況の下で、あえて「壁にぶつかる」ことが必要であると、IDCは提言する。よく理解

し、詳細を完全に把握して、次のインフラへの移行を進める必要がある。また、ビジネス全体を通して、初期段階から高度な開発段階に至るまで、AI能力の完全な活用を実現できるサーバーベンダーと密接に協力しながら実行に移すことを推奨する。

本調査レポートの以下のセクションは、AIアプリケーションのためにコンピューティングインフラでアクセラレーターを採用する北米企業100社に対する大規模なIDCによるアンケートと、アクセラレーテッドコンピューティング（AIに対応したアクセラレーターを装備するシステム）上でAIを実行している企業8社に対する詳細なインタビュー結果をまとめたものである。

概況

人工知能を導入する上での検討課題

世界中の企業が、AIのワークロードで提供される新しいチャンスに積極的に対応している。AIのワークロードには、マシンラーニングやディープラーニングに基づいたアプリケーションが含まれており、これらを駆動するため非構造化データや情報を活用している。まさに、AIのワークロードを展開中の企業もあれば、試行錯誤段階の企業もある。さらに、AIアプリケーションが自社にとってどのような意味があるかをいまだに評価中の企業もある。これら3つのいずれの段階にあったとしても、適切に対処すれば課題の克服は可能であり、その結果得られるビジネス拡大に向けたソリューションは無数に存在する。

こうした検討課題を把握するために、数多くの企業のIT部門とビジネス部門のトップは、AI活用による新たなチャンスに対して、特別委員会などを含め、組織での取り組みを積極的に検討している。ここでの根本的な疑問の1つは、「検討しているAIのイニシアティブにおけるビジネス上の目的は何か」である。これは重要な問いではあるが（AIのためにAIに投資する者などいない）、最初から検討をやり直す必要もない。業種を超えて適用できる、明確に定義されたユースケースも数多く存在する。たとえば、IDCでは以下のようなAIのユースケースを特定している。

- 不正分析／調査（銀行、その他の業種）
- プログラムアドバイザーやリコメンデーションシステム（多数の業種）
- 法規制インテリジェンス（多数の業種）
- 脅威インテリジェンス／防止システムの自動化（多数の業種）
- ITの自動化（ほとんどの業種）
- 販売プロセスのリコメンデーションと自動化（小売、その他の業種）

- 診断／治療（医療）
- 品質管理調査やリコメンデーション（製造業）
- 供給およびロジスティクス（製造業）
- 資産／車両管理（ほとんどの業種）
- 貨物管理（運輸）
- 専門性の高いショッピングアドバイザーや製品リコメンデーション（小売業）

こうした多くのユースケースは社内で開発することも、市販のソフトウェアを利用して実現することもできる。クラウド上のSaaSとして利用できるものもある。その結果、第2の主要な検討課題に結び着く。企業が目的にあったAIユースケースを特定できたとすると、通常次に起こる疑問は、社内で開発すべきか、市販のソリューションを手に入れるべきか、サードパーティベンダーに委託すべきか、あるいはクラウドソリューションを探すべきかである。社内開発の場合は困難を伴う可能性があるが、克服できないわけではなく、実現できる可能性も大きい。IDCの調査では、23%の企業が、自社が検討しているAIソリューションに適したソフトウェアまたはアルゴリズムが何であるかを理解できていない。この比率は大きいようにも見えるが、大半の企業が適切なアルゴリズムを特定できているということも示唆している。

社内でAIアプリケーションを開発するメリットは、AIソリューションをビジネスのニーズに合わせて細かく調整できることである。また、こうした取り組みでは、サードパーティサービスを使う場合のコストが不要となる。企業の32%は、外部のサービスが高すぎると考えている。多くの企業にとっての疑問は、AIソリューションの開発にオープンソースフレームワークや商用ソフトウェアを使用すべきかどうか、ということである。31%の企業が業界大手のコグニティブソフトウェアのコストが高すぎると考えていることから、やはりコストは重要な意味を持っている。その代わりとして、バラエティに富んだオープンソースフレームワークから選定している。それらはダウンロードして使えるか、あるいはAI向けに調達したサーバーにあらかじめパッケージ化されている。

多くの場合、次に検討すべきトピックは、企業がAIソリューションの有効化に適したデータを保有しているかどうかである。この点についてIDCの調査では、ほとんどの企業が必要なデータをよく理解しているかに見えるが、その一方で4分の1の企業はAIアプリケーションに投入される膨大な量のデータを管理するだけではなく、データのクレンジング、ラベリング、変換などの極めてリソース集約的で労働集約的なデータの準備に悪戦苦闘している。AIソリューションのトレーニングに使用する機密データをセキュアに保つことも問題となる。データ準備のサポートは、さまざまなプロバイダーから受けることができる一方、適切なサーバーハードウェアを選択することは、データをセキュアに管理する能力を確保する上で、決定的な役割を果たす。

そして、もちろん、AIソリューションの有効化に関する議論では、それに必要なスキルセットは何か、そのスキルセットは自社に備わっているのか、という疑問に発展することになる。多くのケースで、開発者はAIフレームワークに対応できるように自己をトレーニングする。この方法は成功が証明されているアプローチである。AIを開発する完全なスキルセットを持つ開発者の雇用は高くつくことがある。約3分の1の企業では、AIソフトウェアの人材採用にかかるコストは高すぎるという。データサイエンティストについても同じことが当てはまる。データサイエンティストはより複雑なソリューションのために必要とされているが、そのような人材を見付けるのは困難であり、雇用するとなれば通常は高額になる。このため、企業が、社内のエバンジェリストを介してAIの取り組みを始めることは、よくあることである。エバンジェリストには、開発者に必要なスキルの習得が奨励される。エバンジェリストは検証を行うだけのこともあれば、インフラチームと連携し、新しいAIアプリケーションを開発、テスト、展開するために、環境の一部を適合させる労力も求められる。

こうした検討課題の最上位に、企業はそれぞれのAIイニシアティブに対してどのようなインフラストラクチャが必要かという、問題が存在している。その問題とは、AIイニシアティブを既存のインフラストラクチャの上で展開すべきか（端的には、これで始めるのが妥当であるというのが我々の見解）、あるいは新たなインフラストラクチャを構築すべきか（最終的にはこれが必要となる）、そしてもし新設するとすれば、それはどのようなものであるべきかという問題である。当然、企業はAIソリューションをクラウドで実行できるかどうかについても自問している。

AIアプリケーション用途の多様性

多くの企業がAIの能力を組織で活用しようと努めている。IDCの調査では、2021年までに、コグニティブソフトウェアの売上額は100億ドルに、コグニティブサーバーインフラは90億ドルに成長すると予測している。企業は、このように急速に成長しているテクノロジーに投資を行い、新しい競争力を獲得しようとしている。本番稼働中またはテスト中のAIアプリケーションを持つ企業へのIDCの詳細なインタビューで得られたいくつかの事例では、ソリューションと導入計画の多様性が示されている。

- 中堅不動産企業では、不動産センサーデータの解析のためにCognitiveScaleを使用している。同社では、変動する不動産賃貸用のリコメンデーションエンジンである Dato (前身はTuri) に加えPTC ThingWorx、GE Predix、そして、異常検知や制御管理、メンテナンス予測用のマシンラーニングモデルが実装された業界向けPaaSも使用している。

- ある大手銀行では、顔検出、音声認識、および感情分析にAuthenticIDを、サイバーセキュリティ対策と顧客の行動とパターンを判定してパーソナライズされたサービスを作成するためにIBM Watsonを使用している。同行は、ロボアドバイザーやファイナンシャルプランニングのプラットフォームであるMarstoneと、小型デバイスでインテリジェントローカルアナリティクスの実行を可能にするソリューションであるSaffronを使用して概念実証（POC）も行っている。

- ある中堅医療企業は、社内でさまざまなアプリケーションを開発している。Alerterは、機械で症状を特定するマシンラーニングアプリであり、Responderでは、こうした症状に対して、人が介在なしで治療できるよう自動化を進めてている。医療企業は、サービスとして提供予定の、脅威インテリジェンスについてディープラーニングを行うアプリケーションも開発中であり、HPE Havenのみならず、マイクロソフトのコグニティブサービスも使用している。

- タイの病院では、IBM Watson for Oncologyを使用して、患者ごとのデータを何千もの症例と突き合わせて分析し、医師と医療スタッフに癌の症例に関する包括的な情報を提供している。この情報には、ニューヨークのMemorial Sloan Kettering Cancer Centerで腫瘍医たちが行った5,000時間のトレーニング、300種類の医学雑誌、200冊のテキスト、1,200万ページの文書から得られた情報が含まれている。

AIインフラの壁

しかしながら、IDCの調査では、AIやディープラーニングのアプリケーションのPOCや本番稼働中の企業の大半が、ある時点でこの「インフラの壁」に遭遇することがある。別のインフラへ移行した後、一度ならず2度までもこの壁に遭遇することさえある。「壁に突き当たる（Hitting the wall）」というフレーズは、持久力を必要とするスポーツの用語である。この種のスポーツのアスリートは、グリコーゲンの欠乏によって、突然疲労感に襲われ、急速にエネルギーを失う。企業がAIワークロードのインフラを通じて経験するものの比喩として、極めて的確な表現と言える。

IDCは、既存のオンプレミスインフラ上でAIアプリケーションの稼働を始めたときの経験を企業にたずねた。その回答はシビアなものであった。回答企業の77.1%が、オンプレミスでのAIインフラの限界を1度もしくは複数回経験したと答えている。コグニティブのクラウドユーザー企業では90.3%もが同様の限界に直面したのである。Table 1は、オンプレミスとクラウドのインフラでAIアプリを使用して発生した限界をまとめたものである。

TABLE 1 AIアプリ使用時のインフラの限界（発生頻度の高い順にランク付け）

AIアプリ使用時のオンプレミスインフラの限界	AIアプリ使用時のクラウドインフラの限界
管理が難しい	拡張が難しい
拡張が難しい	パフォーマンスの限界
パフォーマンスの限界	ストレージが不十分
SLAの範囲内でタスクを完了できない	管理が難しい
ストレージが不十分	問題の判別が難しい
問題の判別が難しい	負荷分散が難しい
サーバーの仮想化が困難	SLAの範囲内でタスクを完了できない
データセンターでの相互運用性の欠如	消費電力が著しく大きい
消費電力が著しく大きい	データセンターでの相互運用性の欠如
メモリーの限界	サーバーの仮想化が困難

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

こうした困難によって、企業は短い期間でインフラの世代交代を経験している。AIアプリケーションとディープラーニングが市場に投入されてまだほんの数年間であるが、IDCの調査では、すでに22.8%の企業がAIアプリケーションの第3世代インフラに世代交代しており、37.6%が第2世代、39.6%が第1世代のサーバーインフラを稼働させている。これらの割合は、適切なインフラを模索する必要性を示している。Table 2は、最も高い頻度で発生しているAIのサーバーインフラの世代交代の要因をまとめたものである。

TABLE 2 AIサーバーインフラで最もよくある世代交代の要因 (発生頻度の高い順にランク付け)

高性能なプロセッサーへの移行
スケールアウトからスケールアップへの移行
VMから専用サーバーへの移行
スケールアップからスケールアウトへの移行
I/O帯域幅の増強
専用サーバーからVMへの移行
アクセラレーターの追加

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

より優れたプロセッサー性能（最も多く取られるアクション）、より高性能なI/O帯域、さらにアクセラレーターを備えるシステムへの移行が合理的な決断と言える。しかし、このデータは、理想的な構成が確立していないことも示している。スケールアウトを試み、スケールアップへの移行を行った企業もあれば、この逆を行った企業もある。VMで開始してから専用サーバーに移行した企業もある一方で、正反対のことを行った企業もある。

相反するこれらの動きは、見た目ほど奇妙なわけではない。企業は、AIソフトウェアだけではなく、それを実行するためのインフラでも検証を行っている。スケールアウト構成から始め、ソリューションが成熟するにつれて、さらにパフォーマンスが必要であると判断した企業もある。こうした企業は、既存のスケールアップシステムで、そのことに気付いている。スケールアップシステムのパーティションでPOCを開始し、次のステージへのソリューションを採用したときに、1ソケットもしくは2ソケットサーバーのクラスターに移行することを決めた企業もある。同様に、VMで開発され、少し隔離された環境でさらなる開発を行うために専用サーバーへ移行するソリューションもある。

初期の検証と開発ではこうしたすべての移行に合理性があるとIDCでは考えている。既存環境の活用は、正しい構成とはどのようなものであるべきかが明確になるまで新しいサーバーインフラへの投資を遅らせることを意味する。しかし、アプリケーションがリリース時期に近付き、本格稼働の準備ができたら、インフラの壁にぶつからないように、インフラ選定上の意思決定を適切に行う必要がある。

AIアプリケーションを稼働させている企業の回答に基づき、コグニティブアプリケーションにとって理想的なインフラ構成は、アクセラレーターを備えた1ソケットもしくは2ソケットサーバーのクラスターであるとIDCは考えている。なお、アクセラレーターは後に必要になったときに追加することもできる。また、中規模システムのクラスターでもよいが、その選択はAIワークロードが急速に拡大する場合にのみ妥当性がある。さらに、他にも実現可能な構成があるかもしれないが、ユーザー調査で判明したことは、ハイパーコンバージドシステムとVMは、コグニティブアプリケーションでは、あまり有効ではないということである。

何をすべきか?

AIイニシアティブを現在検討中の企業や検証段階からより成熟した段階へ移行しようとしている企業は、本セクションで説明するAI開発アプローチのいくつかを試してみることが有効であるとIDCは考えている。

中小企業におけるAIイニシアティブ

中小企業におけるAIイニシアティブでは、社内でソリューションを開発することが推奨される。これにはいくつかのメリットがある。共同作業による検証を通じて、開発者、ビジネス部門、データアナリスト、またはデータサイエンティスト（それが獲得可能な場合）とインフラチームは、重要な新しいスキルセットを習得しながら、企業専用のオーダーメイドのソリューションを生み出すことになる。データアナリストとデータサイエンティストは、データセットや関連するモデルを準備することができ、開発者はフレーム

ワークをテストできる。また、インフラチームは開発対象のハードウェアや、本番稼働における有用性を評価できる。ビジネス部門はソリューションが満たすべきパラメーターを設定する機会を得られる。ただし、こうした取り組みを行うのはユニークなAIプロジェクトのみにすることをお勧めする。必要なソリューションが市販のソフトウェアで利用可能な場合、社内開発のメリットよりも、市販パッケージが実現できるスピーディな導入のメリットのほうが勝る。

IDCでは、小規模なオンプレミスから始めることを推奨している。その場合、他の環境からできるだけ隔離された専用サーバーで始める傾向になりがちであるが、最終的には他のシステムとの統合が重要であることに気付かなければならない。AIトレーニング用のコンポーネントがあるのなら、環境内でトレーニング用途のデータにアクセスできるようにする必要があり、強力な並列処理が可能なハードウェアを採用する必要がある。この場合、グラフィック処理ユニット（GPU）などのアクセラレーターが十分な数だけ実装されているのが理想的である。こうした環境は、AIソリューションで好まれるクラスターや、複数ノードを備えたコンバージドシステムでも構成できる。ただし、初期段階のAIインフラでは、スケールアップサーバー上の物理パーティションを活用することもできる。逆にVMまたはハイパーコンバージドシステムはあまり好ましくない。ビジネスクリティカルなデータの場合、企業はデータを安全な環境から移動させる必要がないため、データをホスティングするエンタープライズクラスのスケールアップサーバーの物理パーティションが役に立つ場合がある。また、Linuxでのみ稼働する豊富なAI開発用のオープンソースフレームワークの恩恵を享受できることにも留意すべきである。

インフラチーム、開発チーム、データサイエンティストがこうしたソリューションに慣れ、本番環境でソリューションを実行して、ソフトウェアとハードウェアの能力と限界を体験すると、企業は次のステップでの判断をしやすくなる。次のステップには、インフラのアップグレードや拡張、クラウドコンポーネントの追加、またはVARやコンサルタントなどの参加を含むオンプレミスでの自社開発能力の継続的強化も含まれる。

こうした試行錯誤の段階では、インフラチームによる新しいインフラソリューションの徹底的な調査が必要不可欠である。上述の通り、AIシステムは、コア当たりのパフォーマンスが高く、GPUなどのアクセラレーターと組み合わせたI/Oパラメーターを持つシングル／デュアルソケットサーバーのクラスターで稼働させたほうがよい。インフラチームは、すでに取り引きのあるベンダーから入手できるサーバー製品を検討するだけではなく、他のサーバーベンダー、特に全方位の完全なAIハードウェア／ソフトウェアスタックを提供するベンダーにも目を向ける必要がある。こうしたベンダーの中には、ハードウェアの選定から、導入完了までに発生するソフトウェアスタックを介した最適化やコンサルティングサービスまで、AIシステム導入のすべてのステージでサポートを提供している企業もある。AIやディープラーニングのインフラ要件を深く理解しているベンダーを選ぶことを推奨する。

たとえ既設のハードウェアであろうと、そのベンダーが最初の検証段階で助言を行うことができ、さらに組織をオンプレミスまたはハイブリッドオンプレミスクラウドの拡張へと導くことができるベンダーであるかを確認する必要がある。数件から始まって、小さいものから大きいものまですべてのシナリオをこなせるベンダー、つまり、小さなイニシアティブではアドバイザーになり、より大きなAIイニシアティブとなる次の段階ではコンサルタントにもなれるベンダーが理想的である。

より大きなAIイニシアティブ

より規模の大きいAIイニシアティブでは外部サポートから恩恵を受ける。膨大なリソースを擁する大手企業は別として、ビジネスクリティカルなイノベーションを企業にもたらすことを目的とした包括的なAIソリューションを開発するために必要な時間、コスト、および複雑さは、社内での試行錯誤を通じて取り組むにはあまりにも規模が大きすぎる。サードパーティのAIソリューションプロバイダーは、VARやシステムインテグレーターと同様に、ソリューションを速やかに実行するのに役立つが、柔軟性に欠け、固有のビジネスニーズ向けに調整することは容易ではない。非常に大きなイニシアティブでは、コンサルティングパートナーからベネフィットを享受できる。コンサルティングパートナーは、高価であり、長期的な依存関係が生じやすく、初期の導入時間は長くなる傾向にある。その一方、結果として得られたソリューションは企業のニーズに合わせて完全に調整され、適切に実行されれば、データセンターとの統合も可能になる。

大規模なイニシアティブでは、AIの専門技術を持ち、ハードウェア／ソフトウェアスタック全体を含むAI製品／サービスを提供できるサーバーベンダーと連携することにも、明確な利点がある。サーバーベンダーは、一般的にサードパーティのコンサルティングパートナーほど費用がかからず、他のソリューションプロバイダーに比べ、自社が提供するハードウェアの最適化やスケーリングについて深い知識を持っているはずである。しかし、この2番目の点については、必ずしも自明とは言えない。そのサーバーベンダーがAIアプリケーションやディープラーニングのインフラをスケーリングする能力を開示できているか確認する必要がある。なぜならアクセラレーテッドコンピューティングノード（アクセラレーター）を使ってスケーリングさせることは、CPUだけを使ってコンピューティングノードをスケーリングさせるほど単純ではないからである。

ビジネス部門、開発チーム、およびインフラチームは、AIソリューションをできる限りカスタマイズし、（トレーニングを通じて）スキルセットが生み出されるように密接な連携を続けることを推奨する。企業はサーバーベンダーやソリューションプロバイダーのみが理解する「ブラックボックス化」したソリューションに終わらせないようにすべきである。さもないと、拡張もデータセンターとの統合も思うようにならず、さらにトランザクションやデータ量が増加し始めるとパフォーマンスの制約が生じることになる。つまり、こうした取り組みでは、インフラチームの負荷軽減につながらない。したがって、AIのサーバーベンダー、ソリューションプロバイダー、コンサルタントは、パフォーマンスの加速、I/O、管理のしやすさ、スケーラビリティといった、社内開発と同じ厳しい基準で精査されるハードウェアのリコメンデーションの作成が求められることになる。

アプローチと配備の観点で、いくつかのシナリオを組み合わせることができることに留意すべきである。たとえば、社内で構築したソリューションをクラウドのSaaSソリューションと組み合わせてハイブリッドソリューションを実現できる。あるいは、社内でソリューションを構築した後に、VARによって大規模な実装を行うこともできる。IDCの調査では、大半の企業が、AIの取り組みに必要なインフラやソフトウェアに対して明確な費用の見積りをしていない。企業は、ソフトウェア、インフラ、労働力のコストを含め、AIプロジェクト用の指標を用意する必要がある。(生産性の向上、コストの削減、もしくは収益の増加のいずれかによる) 潜在的な回収効果を試算し、プロジェクトの進行と並行してこうした指標に関するデータを確実に収集する必要もある。

オンプレミスか、クラウドか

より大きなAIイニシアティブ向けに、SaaSソリューションが提供されている場合もあるが、クラウドベースのソフトウェアソリューションと同様に、カスタマイズ性に限界があり、スケーラビリティやパフォーマンスはプロバイダーのインフラに依存する。また、データ量やトランザクション数が急に増えるとコストが逼迫することがある。ビジネスクリティカルなデータの場合、機密データや法規制の対象となるデータ、SaaSソリューションのセキュリティを評価する必要がある。

IDCの調査では、AIアプリケーションのためにアクセラレーテッドインフラを持つ企業の65%が、オンプレミスでソリューションを実行している。内訳は、22%がオンプレミスのみであり、43%がオンプレミスとクラウドの両方で実行している。大部分の企業が、これまでに納得できるクラウドエクスペリエンスを見出しており、AIワークロードをクラウドに移行する予定であるという。しかし、こうした移行が、今後24か月の間に発生し得るすべての展開において、コグニティブワークロードの比率全体に影響を与えることはないであろう。つまり、オンプレミスの比率に変化はないということである。一部の例外はあるにせよ、AIに関する特定のユースケースがオンプレミスとクラウドのいずれかに、より適しているとみなされることもない。診断や治療などのAIのユースケースは、データセキュリティへの懸念などによって、クラウドよりもオンプレミスの方が普及している傾向がある。しかしながら、オムニチャネルオペレーション向けのマーチャンダイジングでは、クラウドのほうが普及率は若干高い。それにもかかわらず、オンプレミス、クラウド、そしてもちろんハイブリッド戦略にはそれぞれ明確な役割がある。その中で、ハイブリッド戦略は、最も有利な展開手法になる可能性が高い。

アクセラレーター

本調査レポートでは、AIシステムでのインフラのパフォーマンス制約を乗り越える重要な方法として、さまざまな場面でアクセラレーターを取り上げている。このため、本セクションではアクセラレーターについて簡単に説明する。これは、トレーニングのために大量の演算能力を必要とするディープラーニングアルゴリズムを使用したAIシステムで特に効果的である。あるケースでは、アクセラレーターを使用したことによって、ディープラーニングアルゴリズムのトレーニングに必要な反復を、数日から数時間にまで短縮している。

IDCではアクセラレーテッドコンピューティングを次のように定義している：(CPU) 処理の一部をGPU (グラフィック処理ユニット) や FPGA (書き換え可能ゲートアレイ) など隣接する半導体サブシステムに移すことで、アプリケーションやそのワークロードの演算を加速させる能力およびその方式。アクセラレーテッドコンピューティングは、AIアプリケーションなど、ワークロードがCPUの限界を乗り越えるためのソリューションを求める企業に注目されている。

GPUは、特に、市販品として調達可能で、アプリケーションに容易に組み込める標準的ライブラリーを活用できるため企業にとって魅力的である。とはいえ、FPGA、メニーコアプロセッサー (many-core processor)、特定用途向け集積回路 (ASIC) など、ワット当たりの性能がさらに向上する可能性のある他のテクノロジーも関心が向けられている。

- GPUはニューラルネットワークレイヤーでのベクトルおよび行列演算を実行する。GPUはこれを並列処理するため、エネルギー効率を高く維持したままトレーニングスピードを大幅に向上させる。
- メニーコアマイクロプロセッサーは、外部のアクセラレーターを使用することなく、並列処理またはベクトル化を最適化する。同プロセッサーは、通常のマルチコアCPUよりもコア数が多く、プロセッサー、キャッシュ、メモリー間のデータ伝送速度を最大化することを目的としたアーキテクチャの一種である。また、通常のCPUとしての処理も行う。
- コプロセッサーは、並列のワークロードを加速させるために使用するPCIeカードである。これには、メニーコアプロセッサーが組み合わされ、専用キャッシュ、メモリー、オペレーティングシステムのカーネルが含まれているが、ブートストラップのためのCPUが必要である。
- FPGAは、製造後に顧客がハードウェア記述言語や高水準言語を使用して設定できるように設計された集積回路である。FPGAは、プログラマブルロジックブロック、相互接続、I/Oブロックの配列で構成されている。これらは再構成可能である。
- ASICは、製造後に再設定できない、使用目的があらかじめ設定済みの集積回路である。
- 相互接続 (インターコネクト) は、GPU、FPGA、またはASICとCPUとの間のデータ接続のことである。PCIeによる相互接続では一方向の最大帯域幅はおよそ16Gbpsであるが、NVIDIAのNVLink 2.0では150Gbpsとなっている。

小規模な企業の大半は、サーバーベンダーからサーバーの一部としてアクセラレーターを購入することを選ぶ。これは、有力なサーバーベンダーが、アクセラレーテッドサーバーを提供している場合には手頃な手段である。また、大規模企業では、VARやシステムインテグレーターに相談するか、アクセラレーターベンダーから直接購入する。これは市場全体に対し、より大きな柔軟性をもたらす。VARやシステムインテグレーターはよりカスタマイズされたソリューションを提供できる。一方、アクセラレーターベンダーからの直接購入によって、アクセラレーターのインストールは十分に柔軟性のあるものになる。

サーバーの一部としてアクセラレーターを購入した場合、価格プレミアムを想定する必要がある。アクセラレーターが、元のサーバーに対しどれほどのパフォーマンス向上を実現できたかを判断するベンチマークは、今まではほとんど存在していなかったが、IDCの調査では、こうしたシステムを調達している企業は、平均的に、総合的に得られたパフォーマンスの増加に対して価格プレミアムは許容できる範囲とみていることが明らかになっている（Table 3を参照）。

アクセラレーションは非常に有効であるが、インフラの限界に対する究極のソリューションであるとは限らない。その多くが、サーバーのコアパフォーマンス、選択されたアクセラレーターのタイプ、相互接続のタイプ、そして、ソフトウェアやデータなどのさまざまな他の要因に依存するためである。したがって、企業は、どのアクセラレーターをいくつ使用すべきかを検討するだけではなく、コア当たりのパフォーマンスやI/O帯域幅など、インストールしたサーバーのタイプも検討することが必要不可欠である。とりわけ、さまざまなAIのモデルを試している段階の企業にとっては、それぞれのモデルがシステムに与える負荷は異なっているため、バランスの取れたシステムの選択が重要である。

TABLE 3 提供されたパフォーマンス向上に対して受け入れ可能な価格プレミアム

パフォーマンスの上昇率（%）	価格の上昇率（%）
25	19
50	25
75	31
100	36

Source: IDC's Cognitive Server Infrastructure Opportunities Survey, June 2017

将来の展望

AIアプリケーションは、特定分野のソリューションとして急速に普及するだけではなく、他のすべてのワークロードにも採用され始めるとIDCでは予測している。長期的には、あらゆるワークロードにAIコンポーネントが組み込まれ、不可分な要素としてアプリケーションに統合されていくであろう。これは、ディープラーニング技術を使用して、学習とそれを維持する必要のあるアプリケーションがますます増えることを意味する。したがって、当社の予測では、データやアルゴリズムが大幅に増加し、それに見合うインフラ能力を、タイムリーな方法（リアルタイム、あるいはリアルタイムに近い方法）で効率的に実行する必要が出てくるであろう。

このような変化は、均質なプロセッサーによる従来型データセンターが終焉を迎えつつあるというIDCの見解と大いに関係がある。AIアプリケーションで求められる処理能力を満たしパフォーマンスギャップを埋めるために、従来のx86プロセッサーに加えて、さまざまなタイプのプロセッサーが利用されるようになってきている。さまざまなプロセッサーとは、x86とは異なるCPUであり、アクセラレーターであり、これら2つを組み合わせたものである。

課題と機会

課題

- **混迷の時代：**AIのユースケースが自社にもたらすビジネスメリット、AI能力を社内にもたらすために必要なスキルセット、こうしたアプリケーションを開発するために必要なソフトウェア、インフラ／開発モデルはどのようなものであるべきか、今日のサーバーインフラの性能面での制約を乗り越えるためにどのようなアクセラレーターを選択すべきかについて、企業は確信を持てずにいる。

機会

- **効果的かつ効率的なAIコンピューティング：**この混沌とした環境から抜け出すために、効果的かつ効率的なAIコンピューティングを実現するためのモデルが登場するであろう。具体的には、顧客と密接に連携して検証やスケーリングを行い、最終的にはAIの能力を企業全体にもたらすベンダーが、ハードウェア、ソフトウェア、そして導入の観点を含めて、適切なAIモデルを特定する手法を開発し、広く市場に投入することになるとIDCは確信している。こうしたベンダーは、AIコンピューティングの分野で将来のリーディングカンパニーになるであろう。企業はベンダーの動向を注視して、AIイニシアティブにおける新興のリーダー企業を見付け出す必要がある。

結論

IDCは調査を通じて、AIやディープラーニングへの取り組みのスタート段階にある企業では、試行錯誤の期間を、既存のハードウェア上で実施するのが一般的であることを確認した。ディープラーニングアルゴリズムやAIアプリケーションはまだ発展途上であるため、こうした新しいワークロードを通常のサーバーインフラでの実験の継続が奨励されるべきであり、サーバーベンダーとしては、少し利他的になって、こうした複雑な段階にある顧客を積極的にサポートすべきである。しかし、それと同時に、企業のインフラチームは、開発中のAIアプリケーションの本稼働を開始する段階では、次のステージに向けた準備も進める必要がある。AIおよびディープラーニングでは、サーバーインフラに関する要求が極めて高く、特定の構成、CPU特性、I/Oの能力、アクセラレーター、CPUとアクセラレーターの間の相互接続の影響を受ける。多くのAIイニシアティブでは、企業規模の大小を問わず、またさまざまな部門やグループがサーバーベンダーのサポートを受ける必要があるであろう。おそらく最も効率的な進め方は、初期の実験段階から、十分に統合されたスケーラブルなAIソリューションの実装に至るまで、すべてを含めるというサポート戦略でパッケージ化された完全なAIハードウェア／ソフトウェアスタックを提供するサーバーベンダーからのサポートを受けることである。

IDC Global Headquarters

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

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。

IDC ANALYZE THE FUTURE