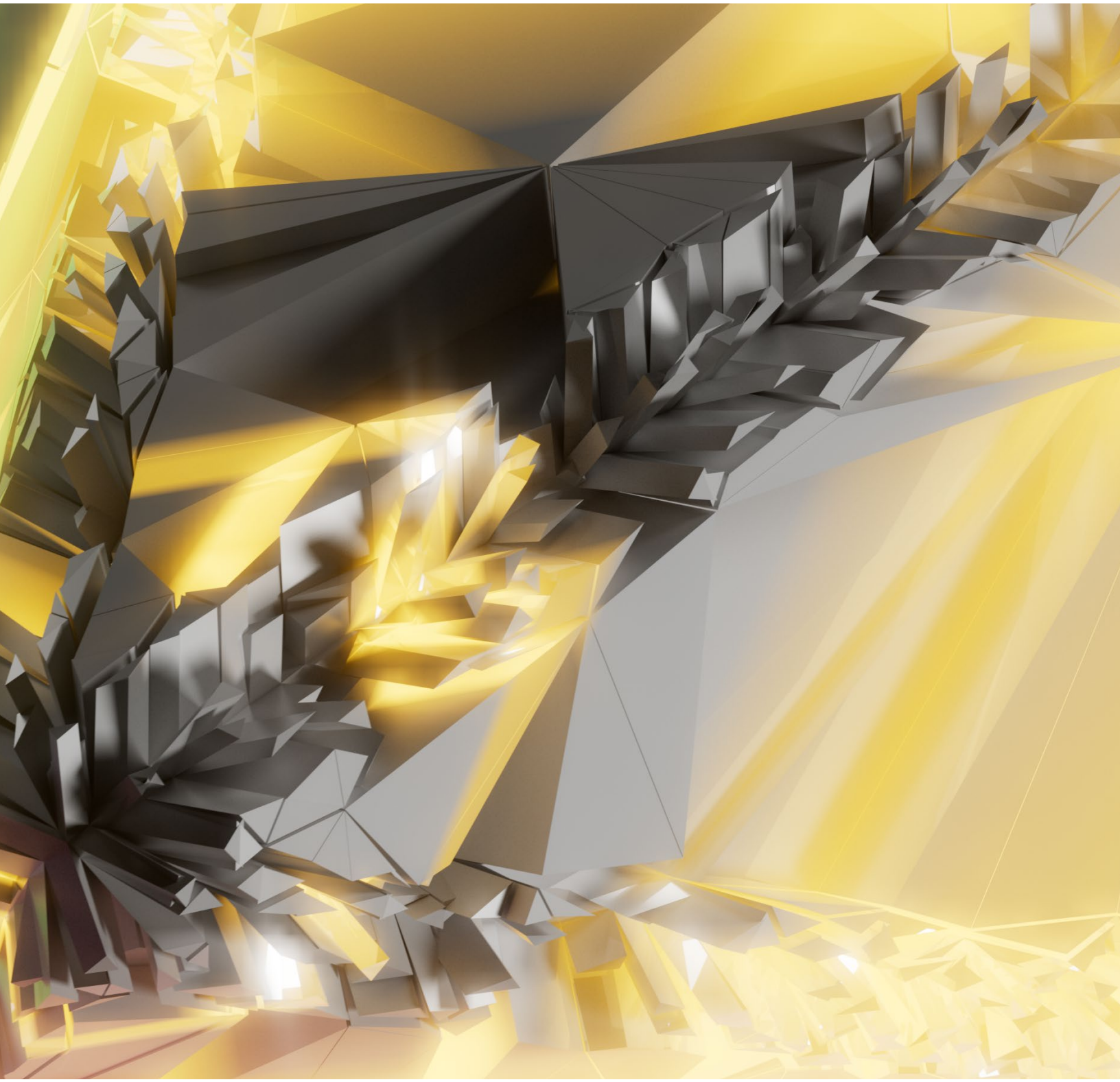


コンピュテーション

より柔軟で、よりスマートに：力任せのコンピューティングを超えて

企業は、既存のインフラのさらなる活用や、プロセスの高速化に向けた最先端のハードウェアの追加を進めている。近いうちに、バイナリーコンピューティングを超越したものに目を向ける企業も出てくるだろう。

テクノロジーが企業にとってより大きな差別化要因となるにつれて、企業はこれまで以上に複雑な計算ワークロードを構築するようになった。人工知能モデルのトレーニング、複雑なシミュレーションの実行、現実世界におけるデジタルツインの構築には、大規模なコンピューティングリソースが必要になる。このような高度なワークロードは、組織の既存のインフラに負荷をかけ始めている。一般的なクラウドサービスは、ほとんどの通常業務に十分な機能を提供しているが、競争優位性を高める最先端のユースケースでは、高度に最適化された特殊なコンピューティング環境が必要になっている[1]。

実行するハードウェアに合わせてコードベースを最適化することは、業務アプリケーションの高速化に向けた第一歩となりうる。長い間見過ごされてきた分野であるこの最適化により、大幅な性能の向上がもたらされる可能性がある。その上、AIやその他の高度なプロセスのトレーニングに特化した先進的なハードウェアは、企業の主流になりつつある。グラフィックスプロセッサ（GPU）やAIチップ、また将来的には、量子コンピューターやニューロモーフィックコンピューターが、コンピューティングの次の時代を形成し始めている。

計算性能の進歩の大部分は、回路を通じてより多くの0と1のビットをより速く処理する方法に焦点を当ててきた。それ自体にはまだ成長余地があるものの、少しずつ認知されているように、それほど長く進歩は続かないかもしれない。このため、先行する研究者やテック企業は、計算性能の制約を乗り越えるのではなく、回避するための革新的な方法を模索している。この過程において、中央処理装置（CPU）がシリコンベースのものとそうでないものを含む特殊なハードウェアと連携して動作する大規模計算の新たなパラダイムの基礎が築かれる可能性がある。

Now：将来のコンピューティング性能の向上ペースはこれまで通りにはいかない

ここ50年ほどのコンピューティングと経済の進歩は、ムーアの法則によって形作られてきた。ムーアの法則とは、コンピューターチップ上のトランジスタの数、つまり性能が2年ごとにほぼ倍増するという考え方だ[2]。

しかし、チップメーカーはますます物理的な制約に直面している。ある程度のところで、シリコンチップ1片に搭載できるトランジスタの数は限られている。ムーアの法則はすでに通用しなくなっているという見方もある[3]。これには異論があるが、少なくとも限界が見えてきた可能性はある。チップは消費電力が増え、冷却が難しくなっているため、性能が妨げられている[4]。そのため、チップメーカーがトランジスタを増やしても、性能が向上するとは限らない。

タイミングの悪いことに、このような実情に対して、企業はますます計算集約型ワークロードを志向している。多くの企業が現実世界のプロセスのデジタルツインを開発しており、産業の自動化が加速している。また、コネクテッドデバイスやモノのインターネット（IoT）の導

入も進んでおり、どちらも大量のデータを生成し、処理要件を引き上げている。機械学習、特に生成AIは、トレーニング中にテラバイト単位のデータを処理する複雑なアルゴリズムを必要とする。これらの取り組みはいずれも、企業にとって大きな競争上の差別化要因となるが、標準的なオンプレミス基盤で実行するのは現実的ではない。一方、クラウドサービスは**切望されるスケール**を実現できるが、その分コストが法外な金額となる可能性がある[5]。

CPUの性能向上ペースの鈍化は、企業の収益にインパクトを与えるだけではない。NVIDIAのCEOを務めるJensen Huangは、GTCカンファレンスの基調講演で、今日ではあらゆる企業や政府が二酸化炭素排出量を実質ゼロにしようとしているが、従来の計算の需要が高まる中では、それを達成するのは難しいことに触れ、「ムーアの法則が通用しなくなれば、計算量の急増に伴い、データセンターの電力使用量は急増する[6]」と述べた。

ある時点を過ぎると、パフォーマンスを向上させるためにデータセンターを拡大したり、クラウドへの投資を増やしたりすることは、経済的に意味をなさなくなる。従来のクラウドサービスは、顧客関係管理（CRM）、企業資源計画（ERP）、企業資産管理（EAM）、人的資本管理（HCM）などのバックオフィスプロセスを有効化および標準化するためには最良の選択肢である。しかし、AIやスマートファシリティなどの成長を促すユースケースを従来のクラウドリソースで運用することは、最終的には企業のIT予算全体を圧迫する可能性がある。そこで、新しいアプローチとして特殊なハイパフォーマンスコンピューティングシステムなどが必要だ[7]。

New：ハードウェアとソフトウェアをより柔軟で、よりスマートに動かす

従来のコンピューティング性能の進歩が鈍化しているからといって、リーダーが計画にブレーキをかける必要はない。処理を高速化する新たなアプローチは、ビジネスを前進させる上で重要な役割を果たす可能性がある。

Simple：コードの簡素化

CPUのパフォーマンスが1～2年ごとに確実かつ予測可能に向上していた時代は、コードが非効率的に書かれて少し肥大化しても、致命的ではなかった。しかし、性能の向上が鈍化している現在、エンジニアがコードを効率的に処理することがより重要になっている。企業にとっては、コードを実行するハードウェアが同じままであっても、コードを簡素化することでパフォーマンスが大幅に向上する可能性がある[8]。

この簡素化は、クラウド移行中に行うことが、通常適している。しかし、メインフレーム上のCOBOLなどの古いコードを直接移行すると、コードが肥大化して非効率になる可能性がある[9]。アプリケーションをJavaなどのより現代的なコードにリファクタリングすることで、企業はクラウドの最新の機能を活用し、この問題を解決することができる。

米ユタ州のOffice of Recovery Servicesでは最近、主要な案件管理および会計システムを完全にクラウドに移行した。自動化されたリファクタリングツールを使用してコードをCOBOLからJavaに変換したことで、その後性能が向上した。

「アプリケーションの処理速度が大幅に向上した」と、Office of Recovery Servicesの技術責任者であるBart Masonは言う。「メインフレームで動作させていた機能について、そのコードをJavaに変換することができた。今日では、メインフレームよりもはるかに高速だ[10]」

Situated：遍在リソースの統合

ベルギーの小売業者であるColruyt Groupは、コンピューティングタスクに適切なリソースを使用することで、商品を保管する倉庫の自動化、コンピュータービジョンを使用した在庫レベルの追跡と管理、顧客に商品を配送する自動運転車の開発など、イノベーションに挑戦することができた。

コンピューティングワークロードを管理する1つの方法は、使用可能なすべてのリソースを活用することだ。Colruyt Groupのdivision managerであるBrechtel Deroによると、スマートデバイスの普及のおかげで、同社には十分な計算リソースが利用可能であったという[11]。しかし、これらのリソースの多くはオペレーショナルテクノロジーに関するものであり、同社のより旧来のデジタルインフラとは結びついていなかった。当初、それらを連携させるための機能を開発することは、困難であった。しかしDeroによると、Colruytはイノベーションを推進した協力的なCEOのおかげで、恩恵を受けたという。技術面で、この会社はさまざまなソースからのデータの統合を可能にする柔軟なERP環境を運用している。これがインフォメーションテクノロジーとオペレーショナルテクノロジーの統合の屋台骨となった。

「ITとOTの間のギャップを埋めることだ。なぜなら、機械ははるかに賢くなっているからである」とDeroは言う。「IT環境、ERP環境、およびマシン間のシームレスな統合を実現し、負荷と計算が適切な場所で適切な相互作用によって行われるようにできれば、生産性を向上させるための追加のステップを踏むことができる[12]」

Specialized：特化型リソース活用

よりスマートにコーディングすることと既存のコンピューティングリソースをより良く活用することは、企業における処理の多くを高速化するのに役立つ可能性があるが、特定の種類の問題に対しては、企業はますます専用のハードウェアに頼るようになっている。GPUはAIモデルのトレーニングに欠かせないリソースとなっており、運用効率と企業のイノベーションを大きく前進させるテクノロジーとなっている。

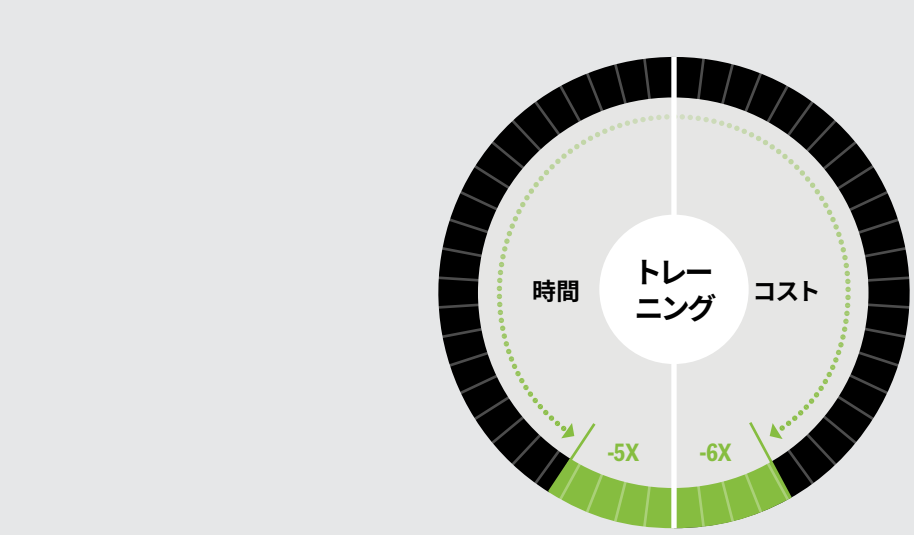
名前が示すように、GPUはもともとグラフィックスをよりスムーズに実行できるように設計された。しかし開発者たちは、GPUの並列データ処理特性を用いてAIモデルのトレーニングを効率化できることに気づいた。AIモデルのトレーニングでは、アルゴリズムを通じてテラバイト単位のデータを提供する必要があるが、これは組織が現在直面している最も計算量の多いワークロードの1つだ。GPUは問題を細かく分割し、一度に処理するが、CPUはデータを順番に処理する。数百万のデータポイントでAIアルゴリズムをトレーニングする場合、並列処理が不可欠だ[13]。生成AIが主流になって以来、モデルをすばやくトレーニングして実行する能力は、ビジネス上の必須事項になっている。

大手テック企業やソーシャルメディア企業、大手リサーチ企業、通信企業、マーケティング企業は、自社のオンプレミス環境内に独自のGPUを搭載している[14]。ただし、より典型的な企業では、クラウド上のGPUを使用することが、最も一般的なアプローチになるだろう。**調査**によると、クラウドGPUは、クラウド上の従来のCPU上のトレーニングモデルと比較して、AIモデルのトレーニングコストを6分の1に、トレーニング時間を5分の1に削減する（図1）[15]。現在、AMD、Intel、NVIDIAなど、ほとんどの主要なチップメーカーがGPU製品とサービスを提供している。

しかし、AIモデルのトレーニングに特化したハードウェアはGPUだけではない。Amazonは Inferentia と呼ばれるチップを提供しており、大規模な言語モデルを含む生成AIのトレーニングを目的としているという。これ

図1

GPUはAIモデルのトレーニング時間とトレーニングコストを削減することができる



出所：Deloitte analysis.

らのチップは、従来の処理装置よりも少ない電力で大量のデータを処理できるように作られている[16]。

GoogleもAIチップの分野に参入している。同社はTensor Processing Units（TPU）と呼ばれる製品を提供しており、Google Cloudサービスを通じて利用できる。これらのプロセッサは、ほとんどの機械学習モデルの基礎となる行列演算を処理するために最適化された、特定アプリケーション向けの集積回路のカテゴリーに分類される[17]。

企業が生成AIの価値を認識するにつれ、特化型AIチップは今後数ヶ月の間に、企業間で引き続き注目を集める可能性が高い。AIの採用が増えることは、ほとんどの組織の既存のデータセンターインフラに負荷をかける可能性があり、汎用リソースと比較してカスタムチップのパフォーマンスが高いことが、競争上の大きな差別化要因になる可能性がある。

これは、企業が一夜にしてこれらのメリットを享受できるという意味ではない。歴史的に、専用ハードウェアが広範に使えるようになるまでの期間と、ハードウェアを最大限に活用するために必要な標準やエコシステムを開発する期間の間には、常にタイムラグがあった。企業がこれらのイノベーションを本格的に採用するまでには、何年もかかる可能性がある。企業は、エコシステムパートナーシップを構築して新興テクノロジーに備え、ビジネスケースが熟し次第、これらのイノベーションを活用できるように、必要なスキルを準備することができる。

Next：バイナリーコンピューティングを超えて

CPUの優れた点は、その柔軟性にある。スプレッドシートからグラフィックデザインソフトウェアまで、あらゆるものに対応している。何十年もの間、企業はコモディティー化したハードウェア上で、何も考えずにほとんどすべてのアプリケーションを実行することができた。

しかし、研究者やテック企業は、新たなデータ処理アプローチを開発しており、その過程で新たな可能性を拓いている。最も有望な新しいパラダイムの1つは、量子コンピューティングかもしれない。この技術は何年も前から議論されており、その影響は明らかになりつつある。

量子アニーリングは、量子コンピューティングの最初のエンタープライズアプリケーションの1つとなる可能性が高く、巡回セールスマン問題などの最適化問題の新たな解決手法である[18]。この種の問題は従来は、機械学習を用いて解決されてきた。しかし、複雑な最適化問題の場合、基礎となる数学、つまり計算処理は信じられないほど複雑になるが、依然として完璧とはいえない解が得られるものとなる。

しかし、量子アニーリングでは、量子ビットの物理的な特性を利用して最適解を導くことができるため、量子コンピューターでは、宇宙ロケット打ち上げのスケジュール設定、財務モデリング、ルート最適化など、複雑なことで有名な多変数の問題を解くことができる[19]。量子アニーリングは、従来のアプローチよりも少ないデータ量と少ないエネルギー消費で、より速く解に辿りつくことができる。

量子アニーリングは、量子コンピューターで最初の広く利用可能なアプリケーションかもしれないが、それが最後になる可能性は低い。この技術は急速に成熟しており、従来のコンピューターが今日あまり適していないさまざまな問題にもすぐに適用できるようになるだろう。量子コンピューターは、従来のコンピューターとは根本的に異なる方法で情報を処理するため、異なる視点から問題を探求することができる。長期間にわたって大量のデータを扱う問題が、潜在的に適している。例えば、IBMは最近ボーイングと協力して、量子コンピューティングを適用してより強力で軽量な材料を設計し、腐食を防ぐ新しい方法の見つけ方を探求した。

IBM Quantumのdirector of theory and quantum computational scienceであるKatie Pizzolatoは、「科学的発見のためのツールとして量子コンピューターに目を向けるときが来た」と語る[20]。「従来のコンピューターの発展の歴史の中で、コンピューターが大きくなるにつれて、我々はコンピューターを使って驚くべきことを発見してきた。今日それに代わるのが量子コンピューターである。システムは従来のコンピューターに対抗できるサイズになりつつあり、今後は、システムが実用性をもたらす問題を見つける必要がある」

量子コンピューターは、現在のバイナリーコンピューティング（0と1を表す電気信号で動作する計算処理）と比較して、データの状態に基づく計算を実行する全く新しい方法だが、新しいアプローチは他にも存在する。もう一つの有望な分野はニューロモーフィックコンピューティングである。このアプローチは、人間の脳のニューロン間のシナプス結合から着想を得ている。一連のトランジスタが順番にデータを処理するのではなく、トランジスタが脳のニューロンのようにネットワーク化

され、トランジスタだけでなく接続の数に応じて計算能力が向上する。主なメリットは、電力を増やさずにパフォーマンスを向上できることである[21]。

より優れたAIアプリケーションは、ニューロモーフィックコンピューティングの最も可能性の高いユースケースである。このコンピューティングアプローチはまだ始まったばかりだが、人間の脳をモデルにしたコンピューターが認知アプリケーションにどのような影響を与えるかは容易に想像できる。自然言語理解、センシング、ロボティクス、ブレイン・コンピューター・インターフェースは、いずれもニューロモーフィックコンピューティングの有望なユースケースである。この分野はまだ比較的新しいが、IBMの開発しているTrueNorthと呼ばれるニューロモーフィックチップや、Intelが発表した研究用チップの第二世代であるLoihiなど、大手コンピューティング企業の支援を受けている[22,23]。

光コンピューティングも有望なアプローチである。ここでは、プロセッサは、電子が回路基板を這うのではなく、光波を使用してデータを移動および格納する。利点は、データが文字通り光速で移動することである。この分野は量子コンピューティングやニューロモーフィックコンピューティングに比べて発展途上だが、IBMやMicrosoftなどの大手テック企業で研究が進められている[24]。

これらすべてのパラダイムに共通する利点は、CPUやGPUよりも低い電力を使用しながら、同等もしくはより優れたパフォーマンスを実現できることである。これは、企業や国全体が炭素排出量の実質ゼロ（ネットゼロカーボンエミッション）を目指す中で、今後さらに重要になる可能性が高い。より高速でより広範なコンピューティングに対する需要は高まる一方だが、企業が目標を達成しようと真剣に考えているのであれば、従来型のクラウドインスタンスを単に立ち上げることは選択肢にならない。

だからといって、これらのテクノロジーがテクノロジー関連の気候問題の万能薬になるわけではない。量子コンピューティングには冷却や水の使用に関する懸念がまだ残っており、他のコンピューティングと同様に、膨大なコードはニューロモーフィックコンピューティングなどの技術に必要なエネルギーを増加させる可能性がある。新しいコンピューティングのオプションが展開されても、コードを簡素化する必要性は存在し続ける。

これらの技術革新がCPUに取って代わることはないだろう。従来のコンピューティングリソースは、大多数のエンタープライズワークロードにとって最も便利で信頼できるツールであり、それは今後も変わることはないだろう。しかし、企業は将来、これらの技術の一部をインフラに組み込むことで、最も革新的なプログラムを進めることができるかもしれない。そして、現在、CPUとGPUを1つの製品に統合するクラウドサービスが登場しているように、ハイパースケーラーが将来提供する製品には、量子、ニューロモーフィック、または光機能が追加される可能性があるため、エンジニアは自分のワークロードを実行しているハードウェアの種類を考える必要すらなくなるかもしれない。

今日の我々の情報世界は0と1で定義されており、間違いなくこのモデルは我々を遠くまで連れてきてくれた。しかし、未来はデジタルコンピューティングだけではない無限に近い可能性を受け入れる準備ができているように見え、これがイノベーションの新時代を牽引する可能性があり、その輪郭はまだ見え始めたばかりだ。

Endnotes

1. Shankar Chandrasekaran and Tanuj Agarwal, *The secret to rapid and insightful AI-GPU-accelerated computing*, Deloitte, 2022.
2. Brittanica, "Moore's law: Computer science," accessed October 31, 2023.
3. David Rotman, "We're not prepared for the end of Moore's Law," *MIT Technology Review*, February 24, 2020.
4. A16Z podcast, "AI hardware, explained," podcast, July 27, 2023.
5. Ranjit Bawa, Brian Campbell, Mike Kavis, Nicholas Merizzi, *Cloud goes vertical*, Deloitte Insights, December 7, 2021.
6. Jensen Huang, "NVIDIA GTC 2024 keynote," speech, NVIDIA, accessed October 31, 2023.
7. Christine Ahn, Brandon Cox, Goutham Balliappa, and Tanuj Agarwal, *The economics of high-performance computing*, Deloitte, 2023.
8. A16Z podcast, "AI hardware, explained."
9. Stephanie Glen, "COBOL programming skills gap thwarts modernization to Java," TechTarget, August 10, 2022.
10. Interview, Bart Mason, technology lead, Utah Office of Recover Services, July 28, 2023.
11. Interview with Brechtel Dero, division manager, Colruyt Group, August 18, 2023.
12. Ibid.
13. Ahn, Cox, Balliappa, and Agarwal, *The economics of high-performance computing*.
14. NVIDIA, "NVIDIA hopper GPUs expand reach as demand for AI grows," press release, March 21, 2023.
15. Ahn, Cox, Balliappa, and Agarwal, The economics of high-performance computing.
16. Amazon Web Services, "AWS inferentia," accessed October 31, 2023.
17. Google Cloud, "Introduction to cloud TPU," accessed October 31, 2023.
18. Cem Dilmegani, "Quantum annealing in 2023: Practical quantum computing," AIMultiple, December 22, 2022.
19. Deloitte, "Quantum annealing unleashed: Optimize your business operations," video webinar, August 3, 2023.
20. Interview, Katie Pizzolato, director of theory and quantum computational science, IBM Quantum, October 16, 2023.
21. Victoria Corless and Jan Rieck, "What are neuromorphic computers?" *Advanced Science News*, March 13, 2023.
22. Filipp Akopyan et al., *TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip*, IBM, October 1, 2023.
23. Intel Labs, "Neuromorphic computing and engineering, next wave of AI capabilities," accessed October 31, 2023.
24. Bert Jan Offrein, "Silicon photonics," IBM, accessed October 31, 2023; Microsoft, "AIM (Analog Iterative Machine)," accessed October 31, 2023.