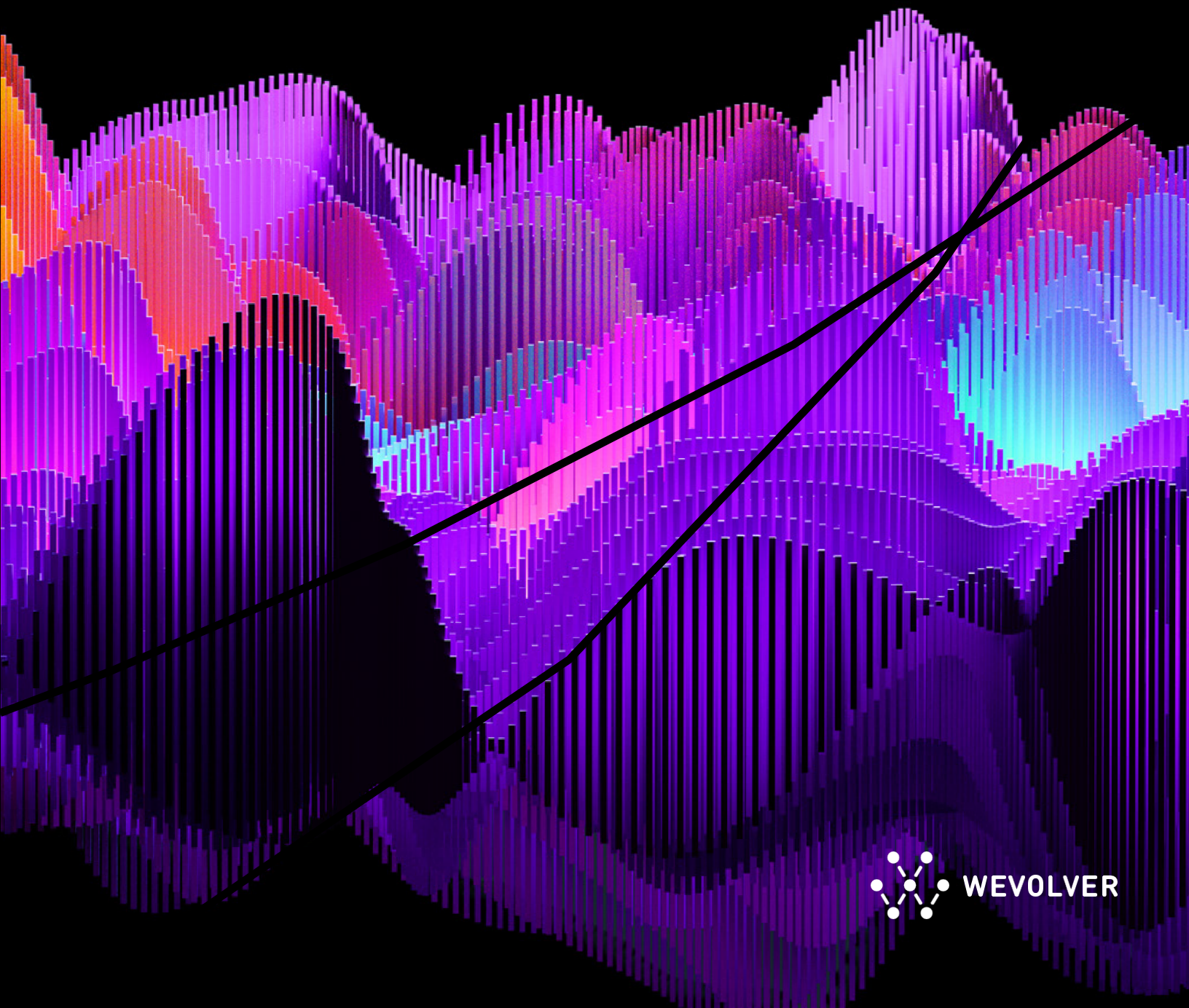
arm

Arm AI Readiness Index

変化する世界における進歩と機会

［日本語要約版］



WEVOLVER

CHAPTER 1

AIへの準備度に関する世界的状況: データドリブンな分析

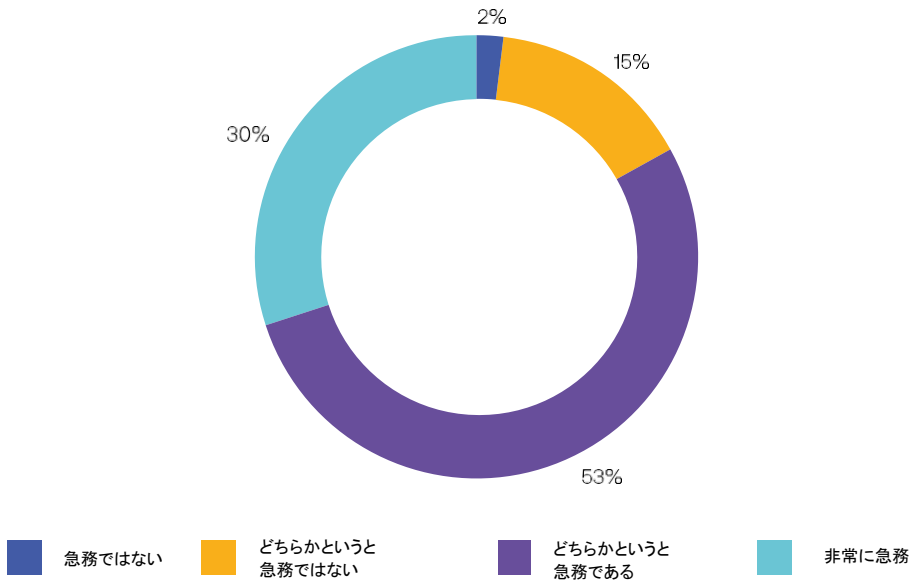
METHOD
COMMUNICATIONS

世界中の企業の役員室や戦略会議において、人工知能（AI）は遠い憧れから差し迫った優先事項へと変わっています。ビジネスリーダーを対象とする私たちの広範な調査からは、注目すべき現実が明らかになりました。AIの採用は急速に加速しており、グローバル企業の82%はすでに日常業務でAIアプリケーションを導入しているのです。このような人工知能の採用は、さまざまな業界と地域で広く浸透しており、AIがもはや巨大テクノロジー企業の専売特許ではなく、あらゆる種類の企業にとって必須技術となっていることが分かります。

このような採用の勢いの中心的な存在がカスタマーサービスであり、企業の63%が顧客インタラクションの強化に向けてAIを導入しています。文書処理（54%）、ITオペレーション（51%）、セキュリティ・アプリケーション（51%）が僅差でこれに続いており、AIがいかに急速にビジネスの中核機能に浸透しているかが分かります。これは単なる実験ではなく、大規模な変革といえます。

AIに対する熱意は、組織的な階層の中に深く浸透しています。リーダーの実に90%は、自社がAI主導の変化を受け入れていると報告、83%はAIの採用を「急務」と認識しています。

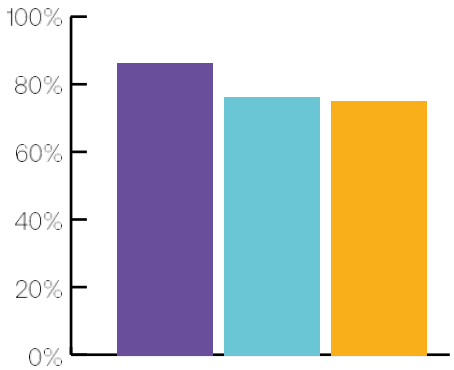
83%のグローバルリーダーは、組織においてAIを活用することを急務と認識しており、
30%は非常に急務であると回答している



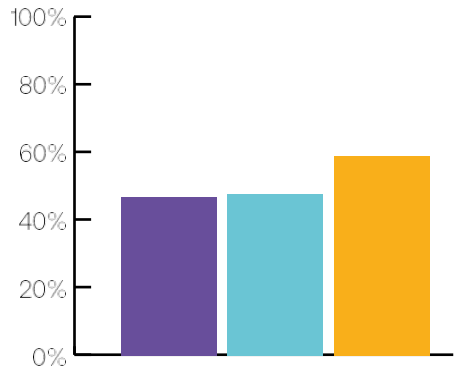
最も重要な点として、企業の82%は最高レベルのエグゼクティブ・スポンサーシップを保証しており、CEOと上級管理職はAIイニシアチブを個人的に支持しています。このようなトップダウンのコミットメントからは、AIが技術的な好奇心の枠組みに留まらず、戦略的な必須事項になっていることが分かります。

財務的なコミットメントは、こうした戦略的な移行を強固にするものです。8割の企業がAI専用の予算を設定していますが、地理的な差異によって投資アプローチは明確に異なります。予算配分ではアジア太平洋（APAC）地域が最も高く、欧州の76%、米国の75%に対し、アジア太平洋地域の企業は86%がAIイニシアチブにリソースを割いています。一方、投資については米国企業が積極的で、57%はIT予算の10%以上をAIに投じることを表明しており、欧州（46%）とアジア太平洋（45%）の割合を上回っています。

AI専用の予算を持っている企業の割合



IT予算の10%以上をAIに割いている企業の割合



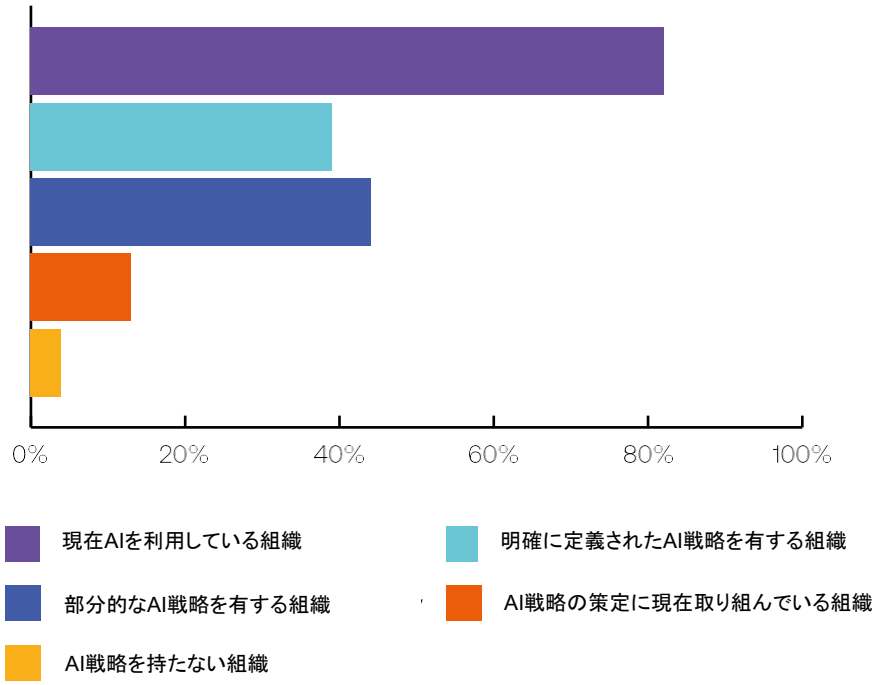
ビジネスリーダーの87%は、今後3年間でAI予算の増加を予想しており、この財務的なコミットメントには揺らぐ気配がありません。

「AIによって反復作業がなくなることで、社員はより戦略的・創造的なタスクに専念できます。」
– ビジネスリーダー調査の回答者

このようなAI推進の原動力とは何でしょう。端的に言うなら効率化です。グローバルリーダーの実に63%が、AIに期待する第一のメリットとして業務の効率化を挙げており、80%は2025年のAI戦略の中心的なフォーカスと考えています。あるビジネスリーダーは、「AIによって反復作業がなくなることで、従業員はより戦略的・創造的なタスクに従事できます」と回答しました。別のリーダーは、「業務全体で意思決定とイノベーションを効率化することで、AIの進化は自社に影響を及ぼす一方、新たなテクノロジーや規制が変化する中、継続的な適応が求められる」状況を説明しました。

しかし、こうした勢いの背後には、憂慮すべきパラドックスが存在します。

採用と戦略のパラドックス



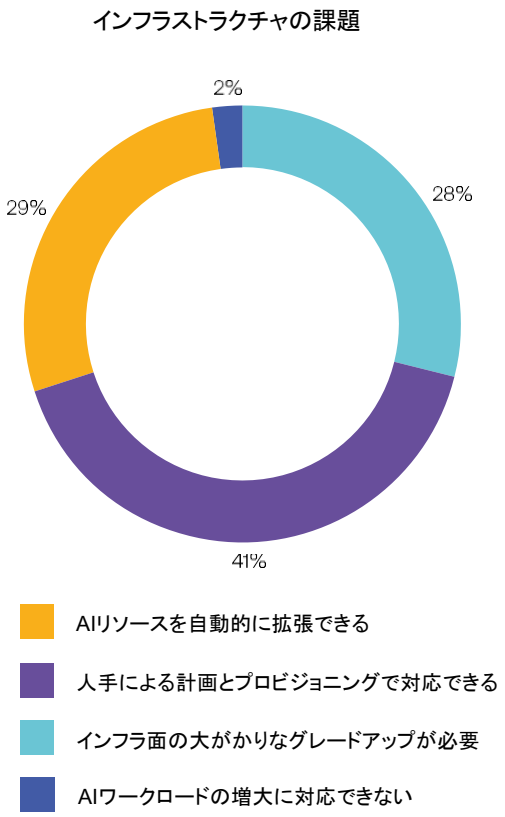
広く採用され、経営陣の支持を得ているに関わらず、明確に定義された包括的なAI戦略の策定率は39%に留まっています。さらに、全社的なAI導入の指針となる、堅牢な変更管理計画を導入している企業はさらに少なく、37%に過ぎません。こうした溝からは、AIが導入される一方、多くの企業がAIの効果的な統合方法を十分に理解しておらず、今後の変化に向けた従業員の準備も不十分であることが分かります。

金融業界のIT部門に従事するバイスプレジデントは、以下の点を認めました。「私たちに必要なのは、こうしたツールセットを正式に導入するための業務です。ツールセットの利用はビジネスユニット内で有機的に拡大しており、重複が存在します。目標には近づきつつあると思いますが、こうした状況を把握することが私の役割であることは明白です。」

このようにまとまりを欠いた戦略により、現行のAIイニシアチブの持続可能性や業績に及ぼす長期的な影響に関して、重大な疑問が提起されます。

慎重に進める

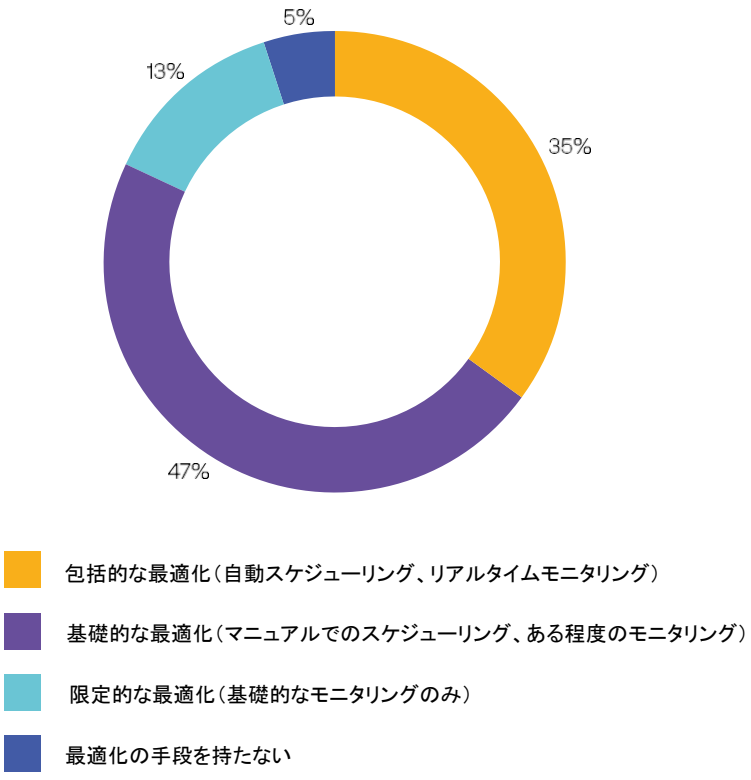
AIへの熱意は、過酷な現実によって冷却する必要があります。というのも、多くの企業は未だにAIイニシアチブの効果的な拡張体制が十分ではありません。私たちの調査によると、**インフラストラクチャの準備、人材の確保、データの品質**という3つの重要な分野で重大な欠陥が明らかになっており、それぞれがAIの可能性をフルに引き出す上での潜在的な障壁となっています。



インフラストラクチャの制限は特に懸念されます。AI需要の高まりに対応するためのシステムやストレージのリソースを保有している企業は29%に過ぎません。AIワークロードのエネルギー要件の増大に対処すべく、専用の電源インフラストラクチャを

保有する企業はさらに少なく、23%に留まっています。つまり、電力管理戦略を設定している場合も、**企業の77%は設備のアップグレードを開始したばかり**なのです。さらに、リーダーの65%は、AIシステムに対する包括的なエネルギー効率の最適化対策が欠けていることを認めており、こうしたワークロードの増大に伴い、運用コストと環境への影響について疑問が生じています。

AI効率性の最適化レベル



人材の溝も同じく重要な課題です。ビジネスリーダーの3分の1（34%）は、AIの専門知識に関して自社が「著しく人材不足」または「人材不足」だと報告しており、半数近く（49%）は、AI導入を成功へと導く上での第一の障壁として、高スキル人材の不足を挙げています。製造業に従事するある経営幹部は、現状を以下のように嘆きました。「これは人的資源に関わる問題です。AIを熟知し、AIを語れる人材を見出すのは非常に困難です。」

こうした人材の溝を認識しているにも関わらず、問題解決へのアプローチは一貫性を欠いています。リーダーの3分の2（66%）は、AI統合の拡大に適応するための既存従業員のスキルアップの意向を示しているものの、39%は、従業員のAIスキルを育むための専用プログラムを設けてはいません。このように、ニーズを認識しつつも決定的なアクションにはつながっていないことで、人材の溝は解消されるどころか拡大する恐れもあります。

おそらくは最も根本的な問題として、企業は効果的なAIアプリケーションにとって不可欠な基盤としてのデータの準備に悪戦苦闘しています。企業の半数近くが顧客データ（49%）と業務データ（48%）をAIイニシアチブで活用しています。一方、データ管理プラクティスの多くは未だに初歩的です。AI/MLモデル用の基本的なデータ自動化プロセスの導入の割合は約半数（53%）に過ぎず、18%は手動または場当たり的なデータクリーニング手順に依存しています。驚くなかれ、リーダーの46%が、AI導入の成功を妨げる大きな障壁として、データの品質とアクセシビリティの問題を挙げています。

「私たちに必要なのは、こうしたツールセットを正式に導入するための業務です。ツールセットの利用はビジネスユニット内で有機的に拡大しており、重複が存在しています。」

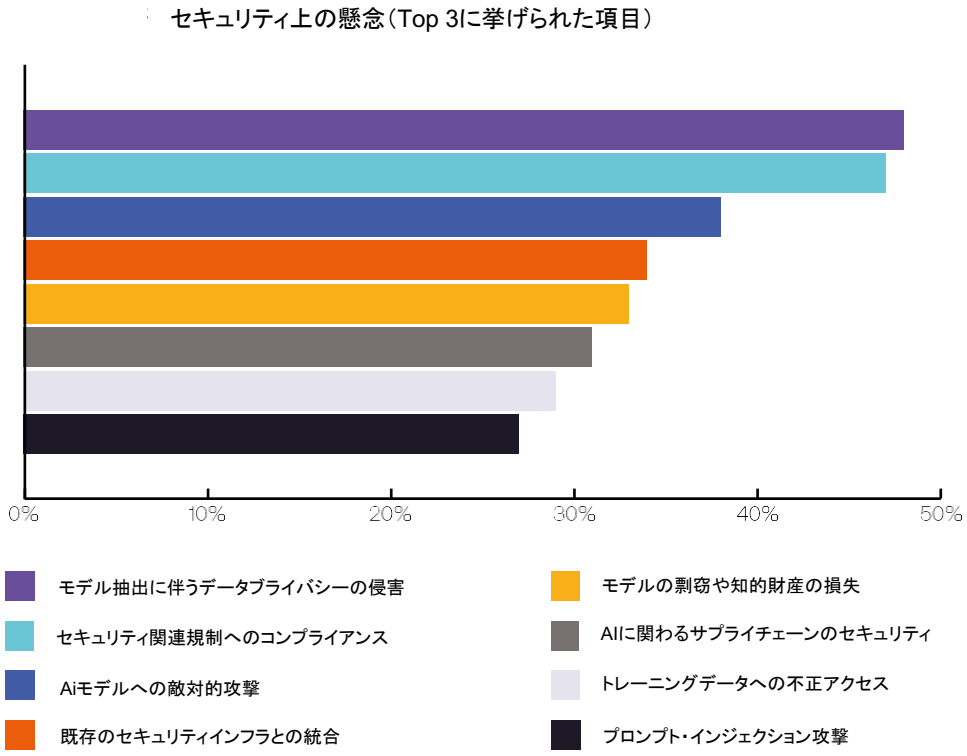
— 金融業・情報技術部門のバイスプレジデント

業界リーダー各社はこうした課題を認識しています。金融業界のIT部門のあるバイスプレジデントは、以下のように説明しました。「AIに関しては、データの品質とカタログ作成が非常に重要です。（中略）これは私たちが直ちに取り組むべき課題です。」製造業に従事する別のIT経営幹部は、統合型データプラットフォームの必要性を以下のように強調しました。「AI利用時に他の機能の活用も試せるよう、データを単一のプラットフォームに統合する必要があります。」

こうしたインフラストラクチャ、人材、データの課題は看過できません。このような根本的な能力を解決できない場合、初期の概念実証デプロイ後の拡張に向けた試みの際には、AI投資によるリターンの減少リスクが発生します。

セキュリティへの感度

自社の業務にAIを組み込む企業が増加する中、データのセキュリティとプライバシーという、緊急の注意を要する新次元のリスクが台頭しています。個人を特定できる情報（PII）が急速にAIイニシアチブの中心的な存在となる中、これに比例してセキュリティへの懸念も拡大しています。



企業の半数近く（49%）はすでにAIイニシアチブで顧客データを利用しており、このトレンドは現在も加速しています。今後については、リーダーの56%が、将来のAIアプリケーションで個人を特定できる情報を活用する計画を示しています。このような機密データへの依存の高まりは、セキュリティ上の重大な懸念を伴います。

ライフサイエンス/バイオテクノロジー業界のある取締役は、以下のように説明しました。「当社もそうですが、データセットを保有する企業は、これまで以上に安全策を講じるべきだと考えます。

そのすべてを公の知識として共有できる訳ではありませんが、AIによって、こうしたデータを閲覧し、解釈し、分析することは、1,000倍容易になっています。」

AIシステム内で機密データの利用が拡大することで、セキュリティ上の懸念のみならず、偏りや公平性に関する倫理的な懸念も発生します。AIシステムの客観性が、トレーニング対象のデータのそれを超えることはなく、AIアプリケーションが偏りのない公平な決断を下すことに関し、重要な問題が提起されています。

「業務全体で意思決定とイノベーションを効率化することで、AIの進化は自社に影響を及ぼす一方、新たなテクノロジーや規制が変化する中、継続的な適応が求められます。」
– ビジネスリーダー調査の回答者

私たちの調査では、この分野での重大な溝が明らかになっています。ビジネスリーダーの半数近く（47%）は、自社のAIシステムのバイアス検知・修正プロセスが不十分なことを認めています。さらに憂慮すべき点として、17%が正式なバイアス修正プロセスを策定しておらず、場当たり的なバイアスチェックに頼っています。ビジネスの重大な意思決定にAIがより深く組み込まれる中、このような一貫性を欠いたバイアス軽減アプローチは重大なリスクを伴います。

未来志向のリーダーは、こうした課題を認識しています。金融業界のあるバイスプレジデントは、以下の見解を述べました。「私たちはデータのバイアスを排除しなければならず、そのためには、データサイエンティストやデータスチュワードのサポートによって、社内で導入すべきポリシーや基準のコントロールを考案すべきです。」こうした認識により、リーダーの44%は、AIの倫理やデータエンジニアリングを、今後5年間の企業が最も必要とする最重要スキルと考えています。

好材料として、企業は自社のAIシステムにセキュリティ対策を導入しつつあります。具体的なAIセキュリティ対策が存在しないと回答した企業は、わずか5%でした。大多数の企業は何らかの予防策を講じており、56%はAIモデルとインフラストラクチャの定期的なセキュリティ監査を実施し、56%はAIを活用したアプリケーションのセキュリティテストを実行しており、51%はAIシステムの脆弱性評価を実施しています。

上記の対策にも関わらず、重大な懸念が残っています。ビジネスリーダーの半数近く（48%）は、モデルの抽出によるデータプライバシーの侵害をAIセキュリティの最大の懸念事項に挙げています。小売業界のあるマーケティング/セールス担当バイスプレジデントは、以下の見解を述べました。「ITセキュリティについては、大企業が心配するような、未解決の重大な問題があると考えます。」

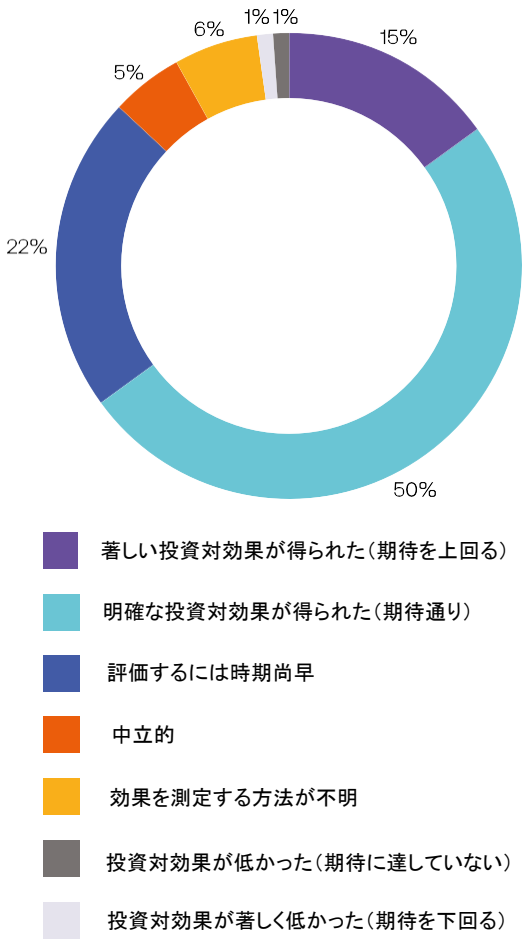
機密性の高いビジネスデータや顧客データとAIシステムがより深く統合される中、企業は基本的なセキュリティ・プラクティスから脱却し、包括的なセキュリティ戦略を策定することで、AI技術に関連する独自の脆弱性に対応する必要があります。

AIを活用した未来に備えて

今後の展望を見据えた際、以下の予測が考えられます。AIの採用は今後もあらゆる業界で加速し続けるでしょう。現時点で効果的な準備をしている企業は、こうした技術革命に乗じたポジションを獲得できるはずです。一方、根本的な準備の溝に対処できない場合、企業は遅れを取るリスクがあります。

AIの使用企業の間で、AIのビジネスケースはますます強力になっています。回答者の65%は、自社のAIアプリケーションの投資対効果について、期待通り、あるいは期待以上の実績を残していると答えています。このようなプラスの体験は採用拡大の原動力となっており、ビジネスリーダーの92%は、何らかの形でAIの使用を拡大する計画を示唆しています。

AIへの期待と現実



しかし、AIイニシアチブを成功裏に拡大するには、準備に関わる複数の重要な溝に対処する必要があります。ビジネスリーダーの3分の1（34%）は、AIのセキュリティに関する最大の懸念事項として、既存のセキュリティ・インフラストラクチャとの統合を挙げています。一方、43%の企業では、従業員がAIについてせいぜい基本的な理解しか示しておらず、トレーニングの追加が急務となっていました。こうした課題により、ビジネスリーダーの3分の2（67%）は、企業が今後5年間で育むべき最も重要なAI関連のスキルとして、AIのセキュリティを挙げています。

AIを活用した未来への準備に際し、企業は熱意と戦略的な計画のバランスを取る必要があります。インフラストラクチャの制限に体系的に取り組み、人材不足を

解消し、データ品質を向上し、セキュリティ対策を強化することで、企業は初期の実験段階を超えて、AIの変革的な可能性を大規模に実現できます。

課題はあるものの、今後の道筋は明確です。企業に求められるのは、AIへの熱意を包括的な戦略へと転換し、基本的な準備の溝に取り組むことです。移行を成功へと導くことで、AIの約束する効率化と競争上の優位性を享受できるようになります。これに失敗した場合、AIへの野心とAIの準備の溝を埋めることは、ますます困難になることが実証されています。

82%

現在すでにAIアプリケーションを使用している世界のビジネスリーダー

39%

明確に定義された包括的なAI戦略を策定している企業

29%

AI需要に合わせて、演算リソースやストレージを自動的に拡張可能

23%

AIワークロードの電力需要の増加を管理するため、専用の電力インフラストラクチャを導入済み

95%

何らかの形でAIセキュリティ対策を導入済み

arm

CHAPTER 2

技術的基盤: AIを成功へと導くテクノロジー要件

DR JOHN SOLDATOS,
グラスゴー大学名誉研フェロー

2.1 AIのテクノロジー要件の概要

2.1.1 堅牢な技術インフラストラクチャの必要性

人工知能（AI）システムは急速に進化しており、より柔軟でスケーラブル、強力になっています。しかし、その可能性をフルに引き出すには、確固たる技術基盤が必要です。複雑化、大規模化、高度化の進むAIアプリケーションを管理する際には、堅牢なインフラストラクチャは重要です。インフラストラクチャは、モデルの効率的な実行を保証しつつ、膨大なデータの処理、保存、分析をサポートする必要があります。AIワークロードが増大する中、インフラストラクチャは、多様なタスクを処理し、レイテンシを削減し、大規模なデータ量を処理しつつ、モダン・アプリケーションの要求するスピードと信頼性を維持する必要があります。

2.1.2 技術的要件：主要な技術分野の主な課題

堅牢なAI技術インフラストラクチャの開発とデプロイでは、ハイパフォーマンス・コンピューティング・アーキテクチャ、電力効率に優れたシステム、スケーラブルなインフラストラクチャ、エッジAIとクラウドAIのデプロイの適切なバランスなど、

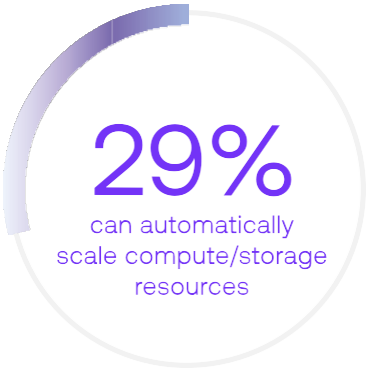
最先端の進化が幅広く活用されています。上記の分野の最近の進化は、パフォーマンス、スケーリング、エネルギー効率の重要な課題の克服に寄与しています。具体的な課題は以下の通りです:

- **コンピューティング・アーキテクチャの進化**:この30年間で、単一層のシンプルなシステムは、アプリケーションのさまざまな側面(例:ユーザー・インターフェイス、データアクセス、ビジネスロジック)を異なる機能層に分割した複雑な多層アーキテクチャへと移行してきました。このような移行を後押ししたのは、柔軟性、適応能力、拡張性を高めたシステムの必要性です。こうした進化を受けて、初期のAIシステムは、大規模なデータセットを効率的に処理できない、あるいは、複雑な演算を実行できないという限界がありました。このような限界に対処するため、多層システムへの移行が進んでいます。こうした移行によって、複数の学習層とサービス指向の設計が導入され、システムの機能とパフォーマンスが強化されています。

- **インフラストラクチャのスケーリング**:AIモデルの大規模化・高度化に伴い、高いコスト効果で拡張可能なインフラストラクチャが必要となっています。その顕著な例として、大規模言語モデル(LLM)とトレーニングデータセットは、かつてないスピードで拡大しています。これにより、インフラストラクチャには、トレーニングデータとデータ処理ワークロードの両方に対応するスケーリング能力が求められます。最近の分散型コンピューティングの進歩は、こうした課題への解決策をもたらしており、特筆すべき点として、スピードや効率性を損なうことなく、大規模なワークロードの処理が可能です。とはいえ、コスト効率を維持しつつリソースの利用を最適化する、効果的なリソース管理戦略が必要であることは変わりません。

- **エッジとクラウドのコンピューティングの課題の比較**:近年の企業は、エッジからクラウドまでのコンピューティングの連続体でワークロードを管理しています。この方向では、局所的な処理のメリットと、集中的なリソースのメリットが対立しています。原則として、エッジ・コンピューティングは、ソースに近い場所でデータを処理し、レイテンシを削減するため、リアルタイム・アプリケーションにとって理想的です。これとは対照的に、クラウド・コンピューティングは、事実上あらゆる数の演算リソース

IT infrastructure readiness

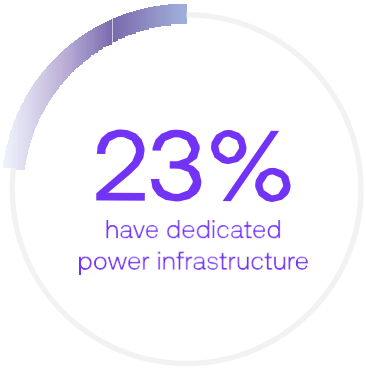


のスケーラブルなアクセスを提供する一方、特定のアプリケーションではレイテンシの問題が発生します。

– **消費電力とパフォーマンスのトレードオフ**：エネルギー消費量は現在、非常に大規模なAIシステムのデプロイを取り巻く大きな懸念事項の1つとなっています。AIモデルは徐々に複雑化しており、より高い演算能力が要求されるため、エネルギー消費量も増加しています。そのため、高性能とエネルギー効率のバランスを取る必要があり、これは一般的に重大な課題です。これと同時に、企業は特定のワークロードの要件と制約に基づき、異なる処理アーキテクチャ間のトレードオフを慎重に検討する必要があります。Armテクノロジーをベースとするものなど、エネルギー効率に優れたCPUアーキテクチャは、データセンターやエッジ・コンピューティングの環境で大規模なAIワークロードを実現する上で不可欠な存在です。こうしたプロセッサーは、汎用コンピューティング機能とAI演算向けの最適化機能を組み合わせることで、汎用性と電力効率が最も重要な推論ワークロード向けの効率的な基盤を実現します。

– **用途特化型AIハードウェア**：特定の演算課題に対処するため、用途特化型のAIハードウェアも登場しています。グラフィックス・プロセッシング・ユニット（GPU）は、AIのトレーニングで一般的な並列処理タスクに秀でています。テンサー・プロセッシング・ユニット（TPU）とニューラル・プロセッシング・ユニット（NPU）は、AIワークロードに特化した設計であり、特定の種類のニューラルネットワーク演算向けに高いワットあたりパフォーマンスを発揮します。これらの技術はそれぞれ、演算能力、エネルギー効率、柔軟性、コストの間で異なるトレードオフが存在します。そのため、モダンな産業エンタープライズには、多数のAIハードウェア・オプション（例：AIに最適化されたCPU、GPU、TPU、NPU）が提供されており、これらを組み合わせることでAIのデプロイ要件に対応できます。こうした状況下で、企業はそれぞれのオプションの機能とともに、特定のニーズに対応するためこれらを統合する方法を理解しなければなりません。

Power consumption management



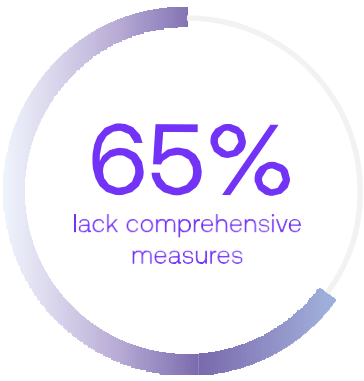
2.1.3 ユーザードリブンな技術：インフラストラクチャとエンドユーザーの要件を対応

拡張性、スピード、効率性を保証するには、技術インフラストラクチャをエンドユーザーの要件に対応させることも重要です。これには、レイテンシの感度やデータ量など、AIワークロードの特定のニーズを理解することが含まれます。原則として、AIインフラストラクチャは、パフォーマンス、拡張性、レイテンシ、エネルギー効率などの要件を確実に満たすために、さまざまなAIアプリケーション・プロファイルに合わせたソリューションを提供する必要があります。新たなAIインフラストラクチャは、AI技術の進化に後押しされるのと同様、エンドユーザーの要件にも後押しされるものでなければなりません。

2.2 モダンなAIのアーキテクチャ基盤

2.2.1 エッジからクラウドまでの統合：AIインフラストラクチャの新たなパラダイム

Energy efficiency optimization



AIシステム・アーキテクチャの進化は、エッジからクラウドまでのAIシステムや生成AIシステムなど、最先端のAIシステムの開発、デプロイ、運用を実現する上で不可欠な役割を果たしてきました。一例として、大規模な生成AIシステム（例：ChatGPT）と分散型の学習システム（例：連合学習）は、旧式のモノリシックな単一層アーキテクチャでは、ほとんどデプロイできません。むしろこれらは、分散型、多層型、ハイブリッド型アーキテクチャの出現によって可能になっており、AIシステムの拡張性、適応能力、モジュール性、パフォーマンスの面で新たな可能性が開かれています。

特に、AIアーキテクチャは、過去数十年間で重要な変革を遂げてきました。こうしたAIアーキテクチャの進化の主なマイルストーンとしては、以下が挙げられます：

— **単一層システムと多層システム**：すでに説明した通り、AIシステムが一枚岩であることは、ほとんどありませんでした。AIシステムは、ユーザー・インターフェイス、ロジック、データストレージのコンポーネントを単一の実行可能モジュールに統合したものです。これにより、開発とデプロイは簡単になりましたが、拡張性と

適応能力には限界がありました。そのため、単一層アーキテクチャでは、非自明なAIシステムやタスクには不向きでした。こうした制限に対処するため、さまざまな機能を機能特化型の層に分割した、多層（n層）アーキテクチャが登場しました。多層システムは、モジュール型の開発と分散型の処理を実現し、効率性と拡張性を向上させます。

クラウドベースのアーキテクチャ：過去20年間、クラウド・コンピューティングの台頭により、演算リソースへのオンデマンド・アクセスが可能になったと同時に、サービスの弾力性、容量、品質も向上しました。後者の特長は、非自明なAIアプリケーションなど、エンタープライズ・ワークロードの拡張性と柔軟性で新たな可能性を切り開いています。

「モダンなAIインフラストラクチャが一般的に採用するヘテロジニアスな演算アプローチは、異なる種類のプロセッサーを異なるタスクに活用します。」

— **エッジ・コンピューティング**：IoTデバイス、リアルタイム・アプリケーション、スマートフォンなどのモバイル機器にホストされたモバイル・アプリケーションの出現により、クラウドコンピューティング・インフラストラクチャではほぼ対応不能な新たな要件が登場しています。これにより、クラウドではなくデータソースに近い場所でデータを処理するという、エッジ・コンピューティングの概念が誕生しました。エッジ・コンピューティングは、レイテンシを削減し、エネルギー効率を向上させて、機密データセットの機密性を保証する上で効果的なソリューションとなっています。

現在、AIアーキテクチャは純粋なクラウドベースではなく、ハイブリッドモデルをサポートしています。ハイブリッド型アーキテクチャは、オンプレミス・インフラストラクチャとクラウドリソースを組み合わせることで、柔軟性とコントロールのバランスを取ります。一例として、ハイブリッド環境の場合、オンプレミス・システムを使用することで、機密データや低レイテンシの求められるタスクを処理できる一方、クラウドリソースを使用すれば、大規模モデルのトレーニングや膨大なデータセットの保存で拡張性を実現できます。高度なハイブリッド型ソリューションは、クラウドとエッジのコンピューティングを統合し、パフォーマンスを最適化します。一例として、エッジ機器は、ローカルでデータを前処理してからクラウドに送り、より深い分析を行います。これにより、エッジとクラウドの両方のメリットを活用できます。特に、エッジからクラウドのデプロイは、クラウドのメリットである拡張性、コストの柔軟性（従量課金）、インテリジェントかつ集中管理型のリソース管理機能と、エッジのメリットであるレイテンシの削減やプライバシーの強化の特性を兼ね備えています。

2.2.2 AIモデルとハードウェア

AIシステム・アーキテクチャの進化は、さまざまなAIモデルの開発・デプロイの実現で不可欠な役割を果たしています。具体的には、モダンなAIシステムは以下をはじめ、異なるアーキテクチャを採用したさまざまなモデルを統合しています。

- **ニューラルネットワーク**：ニューラルネットワークは、多くのAIシステムの基盤となるアーキテクチャです。ニューラルネットワークは、人間の脳を模倣した、相互接続したノード（＝ニューロン）で構成されます。ニューラルネットワークは主に以下に大別されます。
- (i) **フィードフォワード・ニューラル・ネットワーク（FNN）**：データは入力から出力へと一方向に流れます。FNNは、基本的な分類タスクと回帰タスクで使用されます。
- (ii) **畳み込みニューラルネットワーク（CNN）**：画像処理とパターン認識に特化しています。その顕著な例として、AlexNet CNNは、深い畳み込み層を用いて画像分類に革命を起こしました。
- (iii) **リカレント・ニューラル・ネットワーク（RNN）**：記憶を保持するためのフィードバック・ループを持つシーケンシャル・データ用に設計されています。その顕著な例として、長・短期記憶（LSTM）ネットワークは、時系列予測や自然言語処理（NLP）で広く使用されています。
- (iv) **残差ネットワーク**：ディープネットワーク内の消失勾配の問題に対処するため、スキップ接続を導入しました。一例として、ResNet50は、物体検知などのコンピュータ・ビジョン・タスクに幅広く使用されています。全体的には、ニューラルネットワークは、画像認識、音声・テキスト変換システム、医療診断などの幅広いアプリケーションとタスクで使用されています。
- **トランスフォーマー・モデル**：非常に効率的な方法でシーケンシャル・データを処理するため、トランスフォーマーは並列処理と自己注意のメカニズムを実現することで、AIに革命を起こしています。アーキテクチャの観点からは、トランスフォーマー・モデルは、エンコーダーとデコーダーの構造で構成されます。エンコーダーは、入力シーケンスを表現に変換します。同時に、デコーダーは、こうした表現に基づき出力を生成します。

トランスフォーマーのコア・コンポーネントには、マルチヘッドアテンション層やフィードフォワード・ネットワークも含まれます。トランスフォーマー・モデルの中で最も顕著な例の1つは、テキスト分類や感情分析などのタスク向けのエンコーダーのみのモデルである、**BERT（トランスフォーマーによる双方向エンコーダー表現）**です。もう1つの例は、翻訳や要約などのNLPタスク向けのエンコーダー・デコーダーモデルである、**T5（テキスト・テキスト変換トランスフォーマー）**モデルです。最も重要な点として、ChatGPTの背後にあるモデルなど、非常に人気の高いGPT（生成型事前トレーニング済みトランスフォーマー）モデルファミリーは、テキスト生成AIや会話型AI用のデコーダーのみのモデルでもあります。一般的に、トランスフォーマーは、機械翻訳、テキスト要約、質問応答などのNLPタスクに秀でています。また、タンパク質の折り畳み（例：AlphaFold）や、物体検知のようなコンピュータ・ビジョンのタスクにも適用されています。

- **生成AIモデル**：生成モデルは、トレーニングされた入力データに類似する新たなデータを生成します。その顕著な例として、敵対的生成ネットワーク（GAN）は、合成データを作成するジェネレーターと、その真正性を評価するディスクリミネーターで構成されます。GANの注目すべき事例は、画像生成用のDCGAN（深層畳み込み敵対的生成ネットワーク）と、写真を絵画風に変えるような領域移動用のCycleGANがあります。生成AIモデルのもう1つのクラスとして、潜在的な表現を学習して新たなデータポイントを生成する、変分オートエンコーダー（VAE）が挙げられます。このVAEは、リアルな画像の生成や、データポイント間の補間で使用されます。全体的には、生成モデルは芸術作品の生成、合成データの生成、創薬のような創造的なタスクに使用されます。
- **マルチモーダル・モデル**：こうしたモデルは、複数のタイプのデータ（例：テキスト、画像）を処理・統合します。一例として、**CLIP（対照的言語画像事前トレーニング）**は、画像のキャプション作成やゼロショット分類のようなタスク向けに、視覚入力とテキスト記述を連携させます。もう1つの例として、**DALL-E**は、NLPとコンピュータ・ビジョン機能を組み合わせることで、テキストプロンプトに基づいて画像を生成します。

マルチモーダル・モデルは一般的に、コンテンツ作成、拡張現実、クロスモーダル検索の各種システムに適用されます。

– **その他の用途特化型アーキテクチャ**：特定のドメインやタスクに合わせたモデルも存在します。一例として、**ビジョン・トランスフォーマー（ViTs）**は、画像をパッチに分割することで、画像処理にトランスフォーマーを適用させます。これらは物体検知やセグメント化などのタスクに使用されます。別の例として、**深層強化学習モデル**は、ニューラルネットワークと強化学習の手法を組み合わせたものです。AlphaZeroモデルは、チェスや囲碁などのボードゲームを習得したことで有名な深層強化学習モデルです。

全体としてAIシステム・アーキテクチャは、基本的なニューラルネットワークから、トランスフォーマーや生成モデルなどの高度な設計へと進化しています。それぞれのカテゴリーには、独自の目的があります。ニューラルネットワークは、基本的なAI機能を提供する一方、トランスフォーマーは、NLPとマルチモーダル・アプリケーションを支配します。生成AIモデルは創造的な出力を実現し、マルチモーダル・モデルは異なるデータタイプを橋渡しします。

さらに、モダンなAIアーキテクチャと上記のAIモデルは、用途特化型ハードウェアの恩恵を受けています。その詳細は以下の通りです。

– **CPU**：AIインフラストラクチャ、とりわけ、汎用性と電力効率が極めて重要な推論ワークロードにとっては、基本的な存在です。AIワークロード向けに最適化されたモダンなCPUアーキテクチャは、高度なベクトル処理機能とエネルギー効率を兼ね備えており、データセンターとエッジ機器でAIを大規模にデプロイするのに理想的です。前処理から推論まで、多様なワークロードを効率的に処理する能力により、CPUは、柔軟性とコスト効果が極めて重要な実環境のAIデプロイに必須です。
– **GPU**：並列処理に秀でています。行列演算を効率的に処理できることで、深層学習モデルのトレーニングには不可欠です。現在のAI企業各社は、自社の製品やサービスを大規模に提供するため、GPUを大量にデプロイしています。

– **NPU**：ニューラルネットワークの演算を高速化するために特別設計された専用プロセッサーです。畳み込みや行列乗算のような一般的なAI演算のための専用回路を実装することで、NPUは、トレーニングと推論の両方のタスクで高いワットあたりパフォーマンスを発揮します。NPUのアーキテクチャにより、消費電力の制約が致命的な意味を持つ用途特化型のエッジAIアプリケーションでは、GPUは特に有効であり、モバイル機器やIoTセンサーなどのリソースの制約のある環境で高度なAI機能を実現します。
– **TPU**：テンサーベースの演算向けに最適化されており、大規模な深層学習タスクに高スループットを提供します。GPUがさまざまなアプリケーションに汎用的である一方、TPUは、ニューラルネットワーク演算の高速化に特化する傾向があります。

AIに最適化されたハードウェア・ソリューションによって、複雑なAIモデルのトレーニング・推論時間と演算コストは大幅に削減されています。

2.2.3 AIシステムとアーキテクチャのトレンド

AIシステムの最も顕著なトレンドの一部と、AIハードウェアへの影響は以下の通りです。

– **機械学習フレームワークとの統合**：現在、分散型コンピューティング・プラットフォーム（例：Apache SparkやHadoop）は、TensorFlowやPyTorchなどのフレームワークとの統合が進んでいます。この組み合わせは、分散型環境でのモデルのデプロイを簡素化しつつ、拡張性を強化します。
– **モジュール型データセンター**：モジュール型の設計により、データセンターは、必要に応じてコンポーネントを追加・交換して、段階的な拡張が可能です。このアプローチは、エネルギー消費量と建設コストの両方を削減するため、拡張性と持続可能性の向上に寄与します。
– **AIモデルの分散型トレーニング**：分散型トレーニングの手法は、複数の演算ノードを活用することで、モデルトレーニングを高速化します。これは、GPT（生成型事前トレーニング済みトランスフォーマー）やBERT（トランスフォーマーによる双方向エンコーダー表現）などの大規模なトランスフォーマー・モデルに

特に有益です。

- **連合学習**：連号学習により、分散化した機器は、生データを共有することなく、グローバルかつより正確なモデルを共同でトレーニングできます。このアプローチは、帯域幅の使用を削減しつつプライバシーを強化するものです。
- **微調整**：事前トレーニング済みのモデルを適応させることで、より高い精度で用途特化型タスクを実行します。微調整では、モデルをゼロからトレーニングするのではなく、GPTやBERTなどの事前トレーニング済みのモデルに組み込まれた知識を活用し、タスク固有のデータを用いてこれを改良します。このアプローチは、演算コストとトレーニング時間を削減しつつ、ドメイン固有のタスクのパフォーマンスを向上します。微調整は、すべてのモデル層の完全な最適化から、特定の層やパラメーターのみを更新する部分的な微調整まで多岐にわたります。チャットボットでの会話のトーンのカスタマイズ、法律や医療のアプリケーションでのドメイン固有の知識の強化、事前トレーニングでカバーできなかったエッジケースへの対処など、微調整はすでに各種タスクで幅広く使用されています。
- **検索拡張世代（RAG）**：RAGは、リアルタイムの情報検索を統合することで、生成モデルの機能を強化します。静的なトレーニングデータのみに依存する従来型の大規模言語モデル（LLM）とは異なり、RAGは、外部データソース（例：データベース、API、ドキュメント・リポジトリ）を応答に組み込みます。この動的な検索プロセスにより、RAGシステムは、より正確で文脈を認識した最新の出力を提供できます。このプロセスでは、関連データをベクトル・データベースにインデックス化することで、ユーザークエリに基づいて適切な情報を検索し、このデータで入力を強化し、検索された知識と事前トレーニング済みの知識の両方を組み合わせた応答を生成します。RAGはすでに、高度な質問応答システム、コンテンツの要約、会話型エージェント、パーソナライズされた教育ツールなど、多数のアプリケーションで使用されています。RAGの新たなトレンドには、構造化データと非構造化データの検索を組み合わせたハイブリッド型の検索手法、テキストと画像や音声を統合するマルチモーダルRAG、プライバシーを強化し、レイテンシを削減するオンデバイスRAGなどがあります。

「国際エネルギー機関の最近の調査によると、AIワークロードは現在、データセンターの総電力使用量の10～15%を占めており、この割合は2030年には25～30%に到達すると予測されています。」

RAGは、以下の主要なハードウェア・コンポーネントに基づき、効率的なデータ処理と推論を要求します。(i) GPU（例：NVIDIA HopperやBlackwell GPU）、AIに最適化されたCPU（例：Google Axionプロセッサーなど、クラウド環境のArm NeoverseベースCPU）またはCPUとGPUを組み合わせた、エンベディング生成とLLM推論用のヘテロジニアスな演算ソリューション（例：Arm NeoverseベースのNVIDIA Grace Hopper）。上記は、リアルタイムの検索と応答生成で、高スループットと低レイテンシに寄与します。(ii) エンベディング生成とLLM推論用のGPU。リアルタイムの検索と応答生成で、高スループットと低レイテンシに寄与します。(iii) 埋め込みデータとベクトル・データベースを適切に処理するためのランダムアクセスメモリ（RAM）（例：64GB以上）(iv) 検索ステップでの迅速なアクセスに役立つ方法で、大規模データセットとベクトルインデックスを保存するための高速NVMe SSD。(V) 拡張性を維持しつつ、広範なオンプレミス・ハードウェアの必要性を削減するため、ストレージと検索のワークロードをオフロードできるクラウドベースのソリューションなど、拡張性ソリューション。

– **エージェント型AI**：エージェント型AIとは、人間のエージェントに似た意思決定能力を発揮しつつ、自律的に動作するよう設計されたAIシステムのことです。このシステムは、受動的なデータ処理に留まらず、自社の環境と積極的に相互作用し、フィードバックに基づき適応し、独自に目標を追求します。エージェント型AIは、強化学習やマルチエージェント・コラボレーションなどの要素を採用することで、動的な環境のパフォーマンスを最適化します。一例として、エージェント型AIシステムは、サプライチェーン・ロジスティクスを自律的にナビゲートすることや、リアルタイム・データに基づいて金融ポートフォリオを管理することが可能です。エージェント型AIは、強化、微調整、適応学習などの分野の進歩と密接に結びついています。こうした手法により、AIエージェントは、事前に定義された目標との整合性を維持しつつ、時間の経過とともに意思決定プロセスを洗練させることが可能です。エージェント型AIシステムは、自律的な意思決定、学習、行動のモジュールを含むため複雑です。マルチモーダル・データを処理し、リアルタイムでアクションを実行し、動的に適応する必要性によって、こうしたハードウェアの需要は高まっています。

(i) AIに最適化され、電力効率に優れたCPU（Arm NeoverseベースのCPUなど）と、NVIDIA HopperやBlackwellなどのGPUの組み合わせにより、知覚モジュール（コンピュータ・ビジョンなど）や、強化学習に依存する意思決定エンジンをサポートします。

(ii) モジュール型アーキテクチャで複数のエージェントにまたがる同時タスクを管理するため、**大容量メモリ（128GB RAM以上）**が必要です。

(iii) **NVMe SSD**は、コグニティブ・モジュールで使用される知識グラフ、トレーニングデータセット、履歴データへの高速アクセスを保証します。

(iv) 分散したエージェント型AIシステムとしてのネットワーキングと電力には、エージェント間のコミュニケーションを促進するための堅牢なネットワーキング・インフラストラクチャが必要です。そのため、電力効率に優れた設計は、自律環境での長期的な運用をサポートする上で不可欠です。

2.3 消費電力と性能のソリューション

2.3.1 演算需要

大規模言語モデル（LLM）や生成AIシステムなどの大規模AIモデルは、指数関数的な大規模化・複雑化が進んでいます。こうしたモデルは、数千億、あるいは数兆のパラメーターで構成され、効率的なトレーニングとデプロイのために膨大な演算リソースが必要です。

モダンなAIインフラストラクチャが一般的に採用するヘテロジニアスな演算アプローチは、異なる種類のプロセッサーを異なるタスクに活用します。この基盤は一般的に、データの前処理、推論、複雑なAIワークロードのオーケストレーションに必要な汎用的な演算機能を提供する、高性能CPUに端を発します。エネルギー効率に優れたCPUアーキテクチャは、電力効率と一貫したパフォーマンスが重要な、大規模な推論で特に有益であることが実証されています。

トレーニング負荷の高いワークロードの場合、企業はCPUインフラストラクチャを用途特化型のハードウェア・アクセラレーターで補強するのが一般的です。GPUは、モデルトレーニングに必要な並列処理に秀でており、一方のNPUは、特定のAI演算に最適化されたパフォーマンスを提供します。この汎用プロセッサーと用途特化型アクセラレーターの組み合わせにより、企業はAIワークフロー全体でパフォーマンス、効率性、コスト効果のバランスを取ることが可能です。

このように演算要件の規模は非常に大きく、既存インフラストラクチャの限界を押し上げ続けています。GPT-4やMeta Llama 3などの最先端LLMのトレーニングには、通常は複数のタイプのプロセッサーを連携させる、ハイパフォーマンス・コンピューティング（HPC）の設定を使用して膨大なデータセットを処理します。こうした背景を受け、上記の各種処理エレメントの効果的な調整に対応する分散型の演算フレームワークと高度なハードウェア・アーキテクチャは、こうした需要増への対応で必要です。

2.3.2 トレーニングと推論の演算需要

AIモデルは、トレーニング段階と推論段階の両方で相当の演算量が必要です。モデルの複雑性、データセットの規模、リアルタイム・パフォーマンスの必要性を考えると、こうした需要は重要です。先に紹介したハードウェア・オプションは、こうした需要への対応で重要な役割を果たしており、AIワークフローのさまざまな段階に最適化された、用途特化型コンポーネントを提供します。

AIトレーニング向けのハードウェア・サポート

AIモデルのトレーニングには大規模なリソースが必要です。その際、大規模な浮動小数点演算（FLOP）を伴います。一例として、GPT-4などのLLMのトレーニングには、数万個のものGPUが必要で、数十ギガワット時のエネルギーを消費します。演算需要に影響を及ぼす主な要因としては、以下が挙げられます。

- **モデルの複雑性**：数十億ないし数兆ものパラメーターを持つより大規模なモデルの場合、さらなる演算能力が必要です。
- **データセットの大きさ**：データセットが大きくなると、トレーニングの時間とリソースのニーズが増大します。

– **並列処理**：複数のプロセッサーを効率的に使用することは、大規模演算を処理する上で重要です。

この場合、AIトレーニング用のモダンなハードウェアには、以下が含まれます。

– NVIDIA A100やNVIDIA GeForce RTX™ 4090などのGPUは、高い並列処理機能と1秒間に数兆回の演算（TOPS）能力により、幅広く使用されています。
– GoogleのTPUは、データセンターで高いパフォーマンスを発揮しつつ、深層学習のタスクに最適化されています。
– ASIC（特定用途向け集積回路）は、特定のAIタスク向けに設計されたカスタムチップのことで、エネルギー効率とスピードを向上させます。
– NVIDIAのArmベースのGrace BlackwellやGrace HopperなどのCPUコアとGPUを組み合わせた演算ソリューションは、AIトレーニングのパフォーマンスを大幅に向上させます。

AI推論のハードウェア・サポート

推論では、予測や決定を行うため、トレーニングされたモデルを新しいデータに適用します。トレーニングよりも演算量は少ないものの、この作業には低レイテンシと高スループットが要求されます。後者は、自動運転車や医療診断などのリアルタイム・アプリケーションのケースで非常に重要です。この場合、推論ハードウェアは、以下をはじめ、スピードと効率性を重視しています。

– NVIDIA Jetson Nano™などのコンパクトで電力効率に優れたデバイスは、エッジでのリアルタイム推論を実現します。
– NPUは一般的に推論タスクに特化しており、パフォーマンスとエネルギー効率のバランスを取ることができます。
– GPUとCPU（例：NVIDIA Grace Hopperスーパーチップ）を組み合わせたソリューションは、比較的大規模なモデルの高スループット・リアルタイムの推論に最適な選択肢です。
– CPU（例：ArmベースのAWS Gravitonプロセッサー）は、多くの推論シナリオの

代替手段であり、比較的小型のノードを中心に、強力な価格性能比とエネルギー効率を発揮します。

このほか、推論パフォーマンスを向上させるための最適化手法も、以下の通り幅広く存在します。

- **量子化**：スピードを向上させ、メモリ使用量を削減するため、モデルの精度を（例：32ビットから8ビットに）削減することです。
- **プルーニング**：冗長なモデル・パラメーターを削除することで、精度を犠牲にすることなく演算の複雑性を軽減します。
- **バッチ処理**：CPUの利用率を最大限に引き上げるため、複数の入力を同時に処理します。

AIのトレーニングと推論の未来を形成する要素として、以下に挙げるハードウェア開発の新たなトレンドが台頭しています。

- **分散型コンピューティング**：需要の拡大に対処するため、分散型システムでの推論が増加しています。
- **エネルギー効率**：メモリアーキテクチャと用途特化型チップの進化は、トレーニング・推論時のエネルギー消費量の削減を目指しています。
- エッジ・コンピューティング・ベースの**リアルタイム・アプリケーション**：即時応答が必要なアプリケーション（例：自動運転）にとって重要になりつつあります。

2.3.3 エネルギー効率

AIワークロードは、膨大なエネルギー消費で悪名高く、演算と冷却に多大な電力を消費するデータセンターを活用しています。AIワークロードは、データセンターの総電力使用量の10～20%を占めていると推定されます。この需要は、関連する二酸化炭素排出量による持続可能性の懸念を引き起こします。全体的には、AI関連のエネルギー効率の主な課題としては以下が挙げられます。

- **データセンターの消費電力**：大規模なAIモデルのトレーニングに必要な電力は、途方もない量になる可能性があります。一例として、単一LLMのトレーニングには、数百世帯の1年間の電力に相当するエネルギーが消費されます。さらに悪いことに、冷却システムも大量のエネルギーを消費します。

- **二酸化炭素排出量**：AIが環境に及ぼす影響は甚大で、広範な採用を阻む大きな要因の1つとなっています。特に、国際エネルギー機関によると、モダンなデータセンターは、世界の温室効果ガス排出量の約1%を占めています。

「単一LLMのトレーニングは、そのライフサイクルで自動車5台分の二酸化炭素を排出することになります。」

こうした消費電力の課題に対処するため、AIベンダーとインテグレーターは、以下のソリューションを活用します。

- **エネルギー効率に優れたチップ**：チップ設計の最近のイノベーションは、エネルギー効率の課題を軽減しています。一例として、演算能力を犠牲にすることなくエネルギー消費量を最小限に抑えることで、低消費電力のArmアーキテクチャは、ワットあたりパフォーマンスを最適化します。

- **持続可能なデータセンター**：再生可能エネルギー源を使用したモジュール型データセンターが、AIインフラストラクチャの二酸化炭素排出量を削減するソリューションとして台頭しつつあります。さらに、高度な冷却技術と仮想化の手法の使用によっても、運用効率は向上します。

2.3.4 レイテンシへの感度

自動運転車、医療用監視、産業オートメーション・アプリケーションなど、リアルタイムの意思決定を実装したクラスのAIアプリケーションを中心に、AIアプリケーションではレイテンシも重要な要素です。一例として、自動運転車は運転時の判断

を下すためミリ秒単位でセンサー・データを処理する必要がある一方、バイタルサインを監視し、ほぼリアルタイムでユーザーにアラートを発するため、医療用ウェアラブル端末には低レイテンシが要求されます。

レイテンシの課題に対処すべく、AIベンダーとインテグレーターは、AIデプロイのエッジでのローカルなデータ処理を採用しています。エッジ・コンピューティングは、ソースのより近くでデータを処理するため、ラウンドトリップ時間は短縮され、応答性は向上します。リアルタイムの制約があるAIアプリケーションの普及により、ゲートウェイやエッジサーバーなどのエッジ機器に強力なAIアルゴリズム（例：ディープニューラルネットワーク）をデプロイするエッジAIのパラダイムが台頭しています。最も顕著なエッジAIアプローチの1つが、TinyMLパラダイムです。これは、マイクロコントローラなどの超低消費電力機器で比較的小規模なAIモデルをデプロイすることで、レイテンシを最小限に抑えた軽量機械学習タスクの実現を目指します。エッジAIシステムは一般的にハードウェアレベルの最適化機能を統合しており、メモリ帯域幅の進化と専用AIアクセラレーターを活用することで、AI演算時のレイテンシを最小限に抑えます。リアルタイムの応答性を向上させるだけでなく、エッジAIシステムでは、AIインフラストラクチャのネットワーキング部分で転送されるデータ量を大幅に抑えることで、プライバシーを強化し、帯域幅を効率化します。

2.4 インフラストラクチャのスケーリング要件

2.4.1 拡張性へのニーズ

AIデータ量の増加とAIモデルの複雑化に対応するため、AIインフラストラクチャには、コスト効果の高い方法でのスケーリングが必要です。大規模言語モデル（LLM）やリアルタイム推論システムなどの大規模AIワークロードには、膨大な演算能力と効率的なリソースが必要です。データ需要と演算需要の大幅な増加は、ボトルネックを招き、ストレージ、処理速度、ネットワーク帯域幅を増大させます。さらに、インフラストラクチャをスケーリングしつつシステム性能を維持することは、高コスト化や運用の複雑化を招く可能性があります。その解決策として、分散型コンピューティングは、複数のノードやマシンにワークロードを分割することで、

スケーラブルなアプローチを実現します。並列処理も、大規模なAIタスクをより効率的に処理する上で重要な要素です。この方向性で、Kubernetesなどのオープンソース・プラットフォームは、トレーニング時間と推論プロセスの最適化に向けて分散型リソースのオーケストレーションを実現します。

2.4.2 コスト管理

AIインフラストラクチャのスケーリングに関しては、パフォーマンスとコスト効果のバランスも考慮する必要があります。モダンなエンタープライズは、クラウドベースの弾力性とオンプレミスのコントロールのトレードオフに対応し、リソース利用率を最適化する必要があります。クラウド・コンピューティングは、動的なスケーラビリティを実現するため、企業は多額の先行投資をすることなく、ワークロード需要に基づきリソースを調整できます。しかし、こうした柔軟性は、インフラストラクチャに対するコントロールの低下や、使用量が多い場合の長期的な出費の可能性という対価を伴います。一方、オンプレミス・ソリューションは、コントロールとカスタマイズを向上させる一方、多額の設備投資と継続的な保守コストが発生します。

これと同時に、リソースのプーリングは、演算の利用率を最適化する上で効果的な戦略です。複数のユーザーや地域のリソースのアグリゲーションに基づき、企業は効率性を最大限に高めつつ、待機時間を最小限に抑えられます。一例として、ピーク需要が重複しないワークロードをプールすることで、共有のリソースをより有効活用できます。このアプローチにより、ピーク時の演算能力を確保しつつ、過剰なプロビジョニングを確保する必要を抑えられます。

2.4.3 モジュール型のアプローチ

モダンなモジュール型インフラストラクチャ設計は、全面的なオーバーホールが不要で、AIシステムを段階的にスケーリングするための柔軟なソリューションを提供します。具体的には、モジュール型データセンターによって、企業は必要に応じてコンポーネントを追加またはアップグレードし、需要の増大に応じて段階的なスケーリングが実現します。モジュール型データセンターは、設置面積と消費電力に相当の制約がある、エッジ・コンピューティング環境で特に有益です。

2.5 AIインフラストラクチャの新たなトレンド

2.5.1 持続可能性への取り組み：より環境に配慮した技術への移行

環境的なパフォーマンスと持続可能性が、政治とエンタープライズの最重要課題となっているこの現代において、AIインフラストラクチャ関連の二酸化炭素排出量を削減するため、より環境に配慮したテクノロジーへの移行が進んでいます。こうした移行の背景には、以下のトレンドが存在します。

- **再生可能エネルギー源と環境に配慮したデータセンター**：AIデータセンターは、太陽光発電や風力発電などの再生可能エネルギー源へと移行しつつあります。この移行は、温室効果ガス排出量を削減するだけでなく、世界的な持続可能性の目標とも合致しています。GoogleやMicrosoftなどの企業は、自社のデータセンターを100%再生可能エネルギーで運用することを公約に掲げており、業界の他のインフラストラクチャ・プロバイダーにとっては、極めて高い基準となっています。さらに、Googleは最近、Kairos Powerと提携し、AIドリブンなデータセンター用に小型原子炉を導入しました。同時に、2028年の電力供給を目的とし、Microsoftは、スリーマイル島で操業停止中の原子炉の再稼働に資金を拠出しています。
- **エネルギー効率に優れたチップ**：すでに説明した通り、チップ設計のイノベーションは、エネルギー消費量の削減に貢献しています。一例として、Armはこれまで、ワットあたりパフォーマンスを最適化する低消費電力アーキテクチャを開発してきました。これらのチップは、エネルギー消費量を削減しつつ複雑なAIタスクを処理できるよう設計されており、データセンターとエッジ機器の両方にとって理想的となっています。

2.5.2 インフラストラクチャ管理の自動化

AIインフラストラクチャを管理する上で、自動化の重要性はますます高まっています。生成AI（GenAI）ツールは最近、こうした自動化を実現する上で突出した役割を果たしてきました。これらのツールは、リソースのプロビジョニング、ワークロードの最適化、スケーリングの意思決定を強化し、運用効率の自動化を向上します。より具体的には、ここ数年の間で生成AI技術は自動スケーリング機能を通じてリソース割り当てに革命を起こしてきました。K8sGPTなどのツールによって、リアルタイムの

ワークロード要件に基づきリソース・プロビジョニングを動的に調整する、生成AIを活用したリソース管理が可能となっています。これにより、ピーク使用時には演算リソースを効率的に割り当てつつ、需要の少ない期間にはスケールダウンを行えます。こうしたレベルの自動化により、手作業による介入の必要性が減ることで、ITチームは日常的な保守タスクではなく戦略的なイニシアチブに集中できます。全体的には、自動化によって、パフォーマンスを損なうことなく運用コストが削減されます。

結論

AIアプリケーションの大規模化と複雑化が進む中、AI技術インフラストラクチャを進化させる必要があることは明白です。私たちは20年以上にわたり、ハードウェア、ソフトウェア、アーキテクチャに関連する事項など、AIインフラストラクチャの多様な側面で著しい進化を目の当たりにしてきました。一例として、モノリシックな単一層システムから分散型、多層、ハイブリッド型アーキテクチャへの移行は、AIシステムの拡張性、適応能力、パフォーマンスの課題の解決で重要な役割を果たしてきました。現在、モダンなアーキテクチャは、CPU、GPU、TPU、マイクロサービスベースの設計など、最先端のテクノロジーを統合することで、演算の効率性と柔軟性を強化しています。

「エッジ・コンピューティングは、ソースのより近くでデータを処理するため、ラウンドトリップ時間は短縮され、応答性は向上します。」

消費電力とパフォーマンスのトレードオフのバランス、大規模データ処理の管理、レイテンシの最小化など、インフラストラクチャに関連する重要課題は、ハードウェアとインフラストラクチャの設計のイノベーションを通じ、これまで効果的に対処されてきました。近年では、拡張性とプライバシーを維持しつつ、低レイテンシ処理を実現するための要として、エッジからクラウドまでの演算パラダイムが台頭しつつあります。さらに、AI運用が環境に及ぼす影響を軽減するイニシアチブに基づき、持続可能性は話題の中心となっています。こうしたイニシアチブの主な例としては、再生可能エネルギーを活用したデータセンターや、エネルギー効率に優れたチップ設計が挙げられます。

さらに、自動化の最新の進化によって、インフラストラクチャ管理を取り巻く環境は

変化しています。今後数年間で、リソースの割り当てを動的に最適化する生成AIツールが、ピーク時のパフォーマンスを維持しつつ運用コストの削減をもたらす見通しです。また、インフラストラクチャのスケーリングに対するモジュール型のアプローチは、既存システムのオーバーホールなしに段階的な成長に柔軟に対応できます。

上記の進化により、企業は今後、技術的な障壁を克服し、AIドリブンなソリューションの可能性をフルに引き出せるようになります。Armなどの企業は、革新的なチップ設計とエコシステムのパートナーシップを通じてエネルギー効率と演算の需要に対処しており、こうした変革で重要な役割を果たしています。AI技術インフラストラクチャの現状と展望からは、AIの未来を担うのが技術的なブレークスルーだけでなく、より効率的で環境的な責任を負ったデジタルエコシステムも重要であることが分かります。

ARM SIDEBAR

AIをあらゆる場所で: コンピューティングの未来に向けたビジョン

著: ケヴォーク・ケシシアン(Kevork Kechichian), ソリューションエンジニアリング担当EVP, Arm

人工知能（AI）によってコンピューティングのあり方は変化しており、クラウド規模の処理からパーソナル機器でのリアルタイム推論まで、あらゆる要素に影響を及ぼしています。AIの採用が加速する中、高性能かつスケーラブルでエネルギー効率に優れた演算プラットフォームへの需要は高まり続けています。ここでの課題は、クラウドのデータセンターから電力に制約のあるエッジ機器まで、多様な環境で効率性やセキュリティを損なうことなくAIを実現することです。

Armの演算プラットフォームは、こうした変革で重要な役割を果たしており、さまざまな業界のAIワークロードのアーキテクチャの基盤となっています。CPU、GPU、NPUでAIを効率的に実行することで、Armは、AIの大規模なデプロイを可能にするエコシステムをサポートしています。このアプローチは、クラウドからエッジまでのAIの加速、エネルギー効率に優れたAI推論の実現、AIワークロードでのセキュリティと信頼の保証という、3つの基本的な柱に基づいています。

Armにより、AIをクラウドからエッジまでのあらゆる場所で実現

AIワークロードは、クラウドベースのモデルトレーニングからオンデバイス推論まで、極めて広範なコンピューティング環境を対象としています。こうした連続体でAIを効率的に実行することは、AIアプリケーションを拡張し、リアルタイムのパフォーマンスを保証する上で重要です。Armの演算アーキテクチャは、あらゆるレベルのAIワークロードをサポートできるよう設計されており、クラウドからエッジまでのAIを加速できます。

クラウド環境において、Armを活用したソリューション（NASDAQ：ARM）がスケーラブルなAIの演算機能を提供することで、データセンターはパフォーマンスを最適化しつつエネルギー効率を維持できます。Arm CPUは、従来型アーキテクチャに比べて低消費電力で高い演算スループットを発揮するため、AI推論用のクラウドインスタンスで採用が進んでいます。AIモデルの複雑化が進む中、クラウドプロバイダーは、大規模なワークロードを処理しつつ運用コストを管理可能な、効率的な演算プラットフォームを必要としています。

低レイテンシの処理とプライバシーの強化の必要性に伴い、AI推論はクラウドの枠組みを超えてエッジへと迅速に移行しています。スマートフォン、産業用センサー、組み込みシステムのいずれ環境でも、デバイス上でAIを実行することで、クラウドリソースへの依存が低減され、リアルタイムの意思決定が実現します。柔軟な基盤の提供を通じ、多様なAIワークロードをサポートすることで、Armプロセッサー・アーキテクチャはこうした移行を促進しており、さまざまなハードウェア構成で最適なパフォーマンスを保証します。

AI演算を効率化するScalable Vector Extension（SVE）や、ベクトル処理機能を強化するCompute Matrix Engine（CME）など、Armv9アーキテクチャにはAIアクセラレーションの重要な強化機能が採用されています。さらに、Armのヘテロジニアスな演算アプローチによって、CPUがGPUおよびNPUと連携することで、特定のワーク

ロード需要に応じてバランスの取れたAIの実行フレームワークが提供されます。

AIが拡大を続ける中、クラウドからエッジまでの環境でワークロードをシームレスに実行する能力は、今後ますます重要になるでしょう。Armの演算プラットフォームにより、幅広いデバイスでAIアプリケーションをデプロイできると同時に、効率性、柔軟性、拡張性が維持されます。

エネルギー効率に優れ、広く普及した演算プラットフォームが生成AIに対応

AIワークロードの顕著な増加により、消費電力と持続可能性について懸念が高まっています。大規模なAIモデルのトレーニングには、膨大な演算リソースが必要ですが、大規模なAI推論には、過剰なエネルギー使用を回避するための効率的な処理が求められます。Armの演算アーキテクチャは、こうした課題に対処できるよう設計されており、AIワークロードは、パフォーマンスを犠牲にすることなく高い効率性で実行できます。

Armの電力効率に優れた設計理念は、数十年間に及ぶイノベーションに根ざしています。Armv9アーキテクチャは、エネルギー消費量を最小限に抑えつつ、AI演算を最適化する高度な機能を統合しています。IoT機器で機械学習のアクセラレーションを実現するHeliumや、AI推論パフォーマンスを向上させるSVEなどのテクノロジーによって、AIアプリケーションは、電力オーバーヘッドを削減しつつ実行できます。これらのイノベーションにより、データセンター、エッジ機器、バッテリー動作の家電製品のいずれの環境でも、AIワークロードの効率的な拡張が可能です。

Armベースのソリューションの主な優位性の1つとして、高性能のAI推論を提供しつつ、低消費電力の要件を維持できます。一例として、モバイル業界では、サードパーティ・アプリケーションのAIワークロードの70%以上がすでにArm CPUで実行されています。こうしたトレンドはPC、産業オートメーション、自律システム

へと広がっており、AIドリブンなアプリケーションでは、厳格な電力制約の中で継続的な処理が必要となっています。

持続可能性は、AIのデプロイ、とりわけ大規模なクラウド運用で中心的な存在となっています。AIワークロードは、データセンターの電力需要の拡大に寄与しています。Armベースの演算プラットフォームを活用することで、クラウドプロバイダーはエネルギー効率を最適化し、消費電力を削減しつつ、高い演算パフォーマンスを維持できます。

AIワークロードをあらゆる場所で実行するための、開発者が信頼するプラットフォーム

AIの採用が拡大する中、AIワークロードの信用と信頼性と保証する上で、セキュリティは重要な要因になっています。AIモデルは、膨大なデータを処理しており、機密情報や独自情報を取り扱うことが多く、セキュリティの脅威、情報漏洩、不正アクセスからの保護は必須です。

「Armの演算プラットフォームは、こうした変革で重要な役割を果たしており、さまざまな業界のAIワークロードのアーキテクチャの基盤となっています。」

Armはセキュリティファーストのアプローチを採用しており、堅牢なセキュリティ機能を演算アーキテクチャに統合することで、さまざまな業界でAIデプロイの信頼できる基盤を提供します。Armv9アーキテクチャは、Armのコンフィデンシャル・コンピュート・アーキテクチャ（CCA）とRealm Management Extension（RME）を採用しており、演算の行われる場所に関わらず、シリコンによってAIのデータとコードを保護します。AI推論がクラウドからエッジへと移行する中、こうしたセキュリティ強化機能は重要であり、デバイスは外部の脅威にさらされることなく、機密データをローカルで処理する必要があります。

ハードウェアベースのセキュリティ機能を統合することで、Armアーキテクチャは、AIのプライバシーを保護し、静止時および転送中のデータを保護します。医療、金融、自律システムなど、AIモデルが機密情報を取り扱う業界では、これは特に重要です。さらに、クラウド環境では、Armのセキュリティフレームワークが

ハードウェア強制型の保護機能を提供することで、マルチテナント・インフラストラクチャに関連したリスクを軽減し、不正なモデルのアクセスを阻止します。

AIが拡張を続ける中、セキュリティは、演算設計の中心的存在であり続ける必要があります。Armがセキュリティをアーキテクチャレベルに組み込むことで、AIアプリケーションは信頼され、回復力に優れ、プライバシーを重視した環境で動作し、企業はAIワークロードを確実にデプロイできます。