

# データセンターを レベルアップし AIを活かす

intel®



# AIへの期待と インフラの現実

業界を問わずリーダーたちは、効率化、イノベーションおよび成長を促進するAIアプリケーション活用の意欲を強めています。しかし多くの場合、ビジネスリーダーがAIの活用により期待する成果と、そのワークロードに対応するデータセンターの性能との間にはギャップがあります。

より多くのパラメーターで複雑さを増したモデルは、特に古いデータセンター上ではコストや演算能力の限界が障壁になり得ます。AIの持つ可能性および課題は、企業に実存的なジレンマを投げかけます。将来のAIアプリケーションを見据えたデータセンターを準備するか、最新のAIツールを活用してイノベーションを進める企業に後れを取るか、です。

**多くの場合、ビジネスリーダーがAIの活用により期待する成果と、そのワークロードに対応するデータセンターの性能との間にはギャップがあります。**

# データセンターの AIワークロードへの対応を阻むもの

AIに対応したデータセンターの構築には、まずいくつかの重要な事項を検討する必要があります。

- どのようなワークロードの実行が必要で、それらはどのようなパフォーマンスのパラメーターが必要か
- ラックやテクノロジーをどのように最適化し、ピーク時のパフォーマンスを提供し、総保有コスト(TCO)を削減するのか
- ビジネスやインフラの、どの部分をオンプレミスにし、どの部分をクラウドにすべきか
- AIワークロードが用いる機密データを、どのようにハードウェアレベルで確実に保護するのか

これらの重要な課題と以下の考察の組み合わせは、今日および将来のAI導入に対応する、データセンターの最新化の取り組みを戦略的に進行し、ITリーダーを助けます。

## コストおよびリソースの負担

データセンターの刷新に最新のAIアプリケーションを採用しながら、TCOを最小限に抑える戦略。

## 複雑性と集約

オープン・スタンダードで柔軟なソリューションを既存のアーキテクチャーと連携させ、AI対応の工程を簡素化する運用の検討。

## データ・セキュリティおよびコンプライアンス

AIアプリケーションで処理される機密データの保存、管理および利用を、規定やプライバシー要件に準拠させるプロセス。

## 1

## コストおよびリソースの負担解消

AIアプリケーションの高度化が進むと、それらの実行に必要な消費電力と演算量が増加します。シリコンのアップグレードの間に、先進的なAIモデルの動作に必要な消費電力、そして学習や推論に用いるデータの安全確保および整理に要する労力など、組織によるAIの効果的な活用は、かなりの、そして時には継続的な投資が必要です。AIモデルの学習や推論は、かなりのエネルギー資源の消費を伴う場合もあります。

しかし、AIワークロード準備の初期段階においてコストを最小限に抑える方法や、投資収益率を最大化できるようインフラを調整する機会も存在します。

### インテル® Xeon® 6 プロセッサーを活用した AI ユースケース

インテル® Xeon® 6 プロセッサーが提供するパフォーマンス、汎用性、AIアクセラレーション機能により、データセンターの設計者は組織全体におよぶ AI ユースケースを実現できます。

- LLM チャットボット
- 推奨提案システム
- 言語処理の Q&A、トークナイザー、言語翻訳
- オブジェクト検出
- 画像分類

# 3倍

世界的な需要

主に AI のけん引により、データセンター能力の世界需要は 2030 年までに 3 倍を超える可能性がある、と、マッキンゼーは述べています。  
(英語)<sup>1</sup>

最大

# 44%

の総保有コストの低減。BERT-large LLM ワークロードをインテル® Xeon® 6 プロセッサーで実行した場合、AMD EPYC プロセッサーで実行した場合との比較。<sup>2</sup>

## CPU と GPU: AI ワークロードの最適な配分とは

AI ワークロードに対応したデータセンターの整備を急ぐ中、設計者の中には市場で最も強力かつ最も高価なソリューションを既定に求める人がいます。しかし、これは常に必要でもなく、どんな想定においても正しいとは限りません。不要なコストを避けつつデータセンターを AI に対応させるには、優先事項の評価および演算リソースの適切な組み合わせの取り入れが重要です。

CPU

従来の機械学習と  
ディープラーニング

CPU/GPU

小規模モデルの生成  
AI 推論

GPU

中規模から大規模モデル  
生成 AI の学習および推論



## AIワークロードにインテル® Xeon® 6 プロセッサのパワーと効率性を

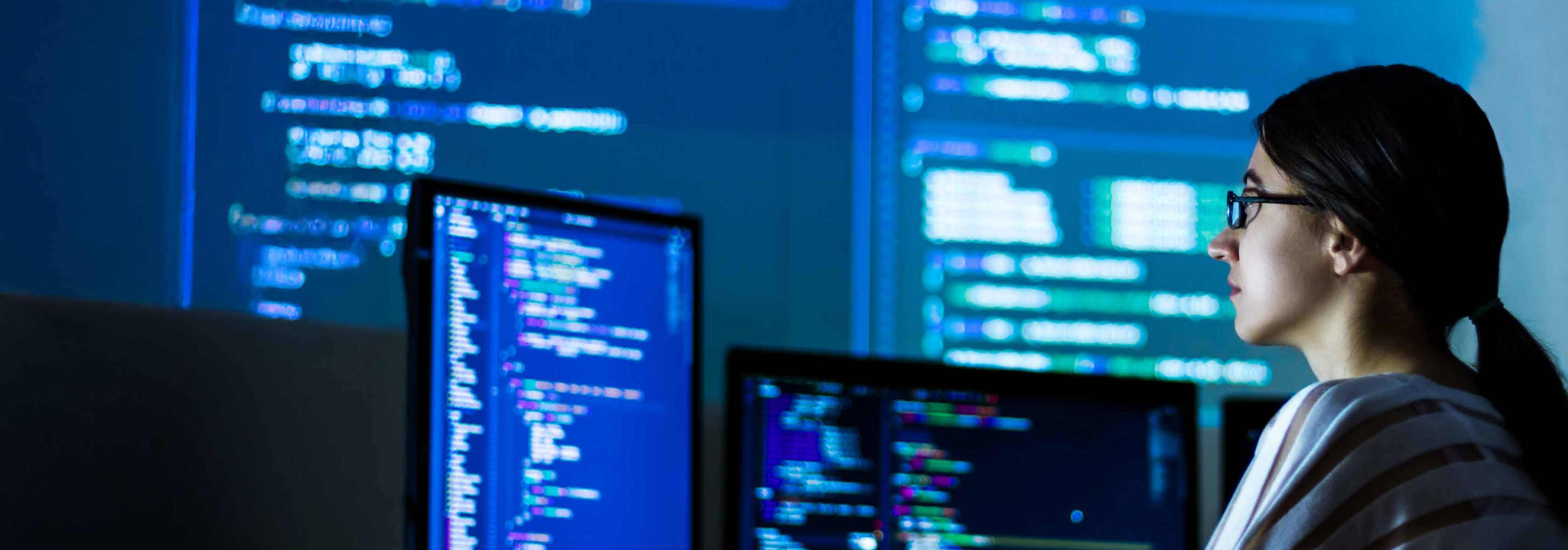
スケーラブルなAIワークロードに、設計者がどう備えるにしても、インテルのソリューションは投資収益率の最大化を助けます。パフォーマンスを最優先し、帯域幅の拡大または演算負荷の高いAIの実行には、[インテル® Xeon® 6 プロセッサ \(P-cores 採用\)](#)の導入が適切です。開発者は、[インテル® アドバンスド・マトリクス・エクステンション \(インテル® AMX\)](#)を活用し、AIの機能を高速化できます。CPUがAIタスクを別のアクセラレーターにオフロードすることなく、より効率的に処理し、CPUと外部デバイス間における電力消費の多いデータ転送を減らします。

### 導入事例:

## Devito Codes による HPC コード生成の自動化 — インテル® Xeon® 6 プロセッサ

Devito Codes は、地震学や流体力学のような複雑な物理科学をモデル化します。インテル® Xeon® 6 プロセッサにアップグレードして以来、161%のパフォーマンス向上を認めました。この向上により、科学者たちは迅速な試作およびシミュレーションを直接 Python から導入可能になり、地震モデルの開発やその他の科学的演算を効率的に行えます。ジャストインタイム・コンパイルの取り込みと混合精度コンピューティングへの対応により、Devito Codes はさまざまなハードウェアのアーキテクチャーで効率、移植性、そして性能に優れたソリューションを実現しました。

[詳細情報](#) ↗



## 2

## 複雑さと集約の合理化

データセンターを適応性がありスケーラブルなAIワークロードに対応させる開発者は、ツール、モデルおよびフレームワークなどからなる広大なエコシステムを目の当たりにするかも知れません。また、新しいコンポーネントがレガシー・ハードウェアと統合できるか否か、オンプレミスとクラウドベースのストレージの選択などを検討する必要もあります。

幸いなことに、AI導入への道のりを簡単にするリソースや戦略があります。

### 柔軟性がカギ

既存のインフラに集約できるモジュール式ハードウェアの使用は、初期導入コストを抑えながらAIアプリケーションのパフォーマンスを強化し、短時間で投資額の回収を見込めます。一例として、[インテル® Xeon® 6 プロセッサ](#)は、既存のフレームワークに接続でき、さまざまなワークロードの特性に最適化された製品ポートフォリオを提供します。また、[インテル® Gaudi® 3 AI アクセラレーター](#)は、AIワークロード全体の演算を加速させます。

インテル® Xeon® 6 プロセッサ・ファミリーは、モジュラー型のx86アーキテクチャーを活用し、データセンターの設計者は、プライベート、パブリックおよびハイブリッド・クラウド上の各組織のニーズやワークロードに特化した、インフラストラクチャーの設定、導入が可能です。

**インテル® Xeon® 6 プロセッサ  
(P-cores 採用) でラック数を  
5台から1台に集約<sup>3</sup>**

5年前のサーバー集約時の平均。

## 導入事例:

# IBM Cloud が提供する AI スケーリング — インテル® Gaudi® 3 AI アクセラレーター

IBMとインテルは長期の協力関係にあり、それぞれが、エンタープライズAIを効果的に拡張する、総保有コストの優位性とオープンなエコシステムを備えたAIシステム構築のビジョンを持ちます。IBM Cloudは、インテル® Gaudi® 3 AI アクセラレーターを顧客に提供する最初のクラウド・サービス・プロバイダーであり、オープン性、セキュリティー、そしてレジリエンスを重視した、コスト効率に優れたAIのスケールリングおよびイノベーションの推進の実現を目指し企業を助けます。

[詳細情報 \(英語\) ↗](#)

## オープンなソリューションで集約プロセスを簡素化

インテルは、エンタープライズAI向けのオープン・プラットフォーム ([OPEA \(英語\)](#)) のパートナーとして積極的に活動しています。OPEAは、設計者向けの非常に有益なリソースとなるオープンソースのコミュニティーです。OPEAを通じてベンダーに中立的なリソースを見つけたり共有も可能です。それらリソースが、既存システムへの集約およびAIモデル間の切り替えの容易な実現を助けます。

## 学習と推論をインテルで

約30年におよぶ機械学習の実績 (ディープラーニング分野も含む) と、プロセッサ製品上の命令セットの長期提供を通じて、インテルは AI アプリケーションの学習、推論、そして導入を可能にする柔軟なソリューションに対応しています。

最大

# 2.17倍

の AI 推論パフォーマンス。  
AMD EPYC プロセッサと比較して。<sup>4</sup>

## オープンソースに対する インテルの取り組み

開始時期

# 1989年

メンバーの数

# 800以上

財団および標準化団体

# 750以上

インテルが管理するオープンソースの  
プロジェクト<sup>5</sup>

### 3

## AIにおけるセキュリティ、 データ管理およびリスク対応の重視

AI導入に向けてデータセンターを整備する際、どのようなアプローチであれ、セキュリティは最優先事項です。世界中の政府がAIや個人および企業データの取り扱いを規定する新たな法律を制定し、データ主権に関する懸念の高まりにより、組織ではデータを管轄区内に留める必要が生じ得ます。

AIは、セキュリティの問題にもソリューションの一部にもなります。一方は、自動化された攻撃から学習中の機密データの不適切な取り扱いに至るまで、AI技術ならではの新たな脅威対象領域を生み出すことを示します。逆に、より迅速な脅威検出を可能にしたり対処を助けたりすることで、AIはサイバーセキュリティの守りを強化することもできます。

### セキュリティにもインテル、はいってる

インテルは、セキュリティ、プライバシー、リスク管理に対して積極的で透明性が高く、協調的なアプローチを維持し、コンピュータ・スタックのあらゆる層にセキュリティを組み込みます。継続的な報告と改善を徹底し、インテルは脅威状況の進化への対応、グローバルなコンプライアンス要件の充足、障害に強くプライバシーを重視するAIを活用したワークロードのインフラ構築において顧客を助けます。

インテル® Xeon® 6 プロセッサはインテル® セキュリティ・エンジンを内蔵し、セキュリティ機能を強化しています。最も機密性の高いデータの利用をAIの分析、学習、または処理において可能にし、プライバシーや機密性が保たれます。

### 導入事例:

## Iternal は AI 精度を向上 セキュリティを強化 — インテル® Gaudi® アクセラレーター

Iternal 社の Blockify は、データ管理のモジュラーアプローチによりAIへの信頼性を高め、AI出力の誤りによる潜在的なリスクの軽減をし組織を助けます。

[詳細情報](#) ↗

## 488万米ドル

セキュリティ侵害発生時の  
平均コスト、  
2024年 (IBMによる調査)<sup>6</sup>

## 222万米ドル

の低減。侵害の防止において  
広くセキュリティ AIと  
自動化を活用した組織が  
削減できた金額  
(活用しなかった組織との比較)<sup>6</sup>



### AIワークロード向けに 強化されたセキュリティ

インテル® ソフトウェア・  
ガード・エクステンションズ  
(インテル®SGX): セキュリ  
ティを強化し、データサイロを  
解消して、イノベーション、  
コラボレーション、新たなAI  
ユースケースを促進します。

インテル®トラスト・ドメ  
イン・エクステンションズ  
(インテル®TDX): 仮想マシン  
の堅実な分離により、  
機密性を強化し、AIデータおよ  
びモデルを保護します。

インテル®TDXコネク: 気密  
性の高い利用モデルをCPUの  
範囲を超えたPCI Expressで  
接続するデバイスにまで拡張が  
可能です。

# AIワークロードの活用 そこにもインテル、はいってる

効率を高めるサーバーの集約や、既存のAIフレームワークとのシームレスな統合、あるいは学習や推論を行う安全でコンプライアンス準拠の環境構築など、インテルはAI活用によるイノベーションがもたらす優位性を通じて企業を助ける取り組みを続けます。オープンソースのプラットフォームを通じた課題への積極的な取り組みは、費用管理やデータのセキュリティ維持の指針となります。

- **AIワークロードによる増大するデータセンター需要への対応**には、インテル® Xeon® 6 プロセッサ (P-cores 採用) の導入を検討してください。さまざまなAIユースケースを実現する最新データセンター向けの理想的なCPUです。
- **複雑な統合の効率化**には、既存のサーバーラックを拡張あるいは集約できる柔軟なプロセッサを採用します。[OPEA](#) にアクセスして、さまざまなニーズを満たし、既存のエコシステムに取り入れられるソリューションを見つけましょう。
- **堅実なプライバシーおよびセキュリティの維持**には、インテル® Xeon® 6 プロセッサと[インテル® ソフトウェア・ガード・エクステンションズ \(インテル® SGX\)](#) を。

最大

**30%**

より優れたAIパフォーマンス  
(DDR5-6400 DIMMとの比較)<sup>7</sup>

最大

**128**

コア (ソケットあたりのコア数)

最大

**38%**

の総所有コストの低減 (AMD EPYC 9654ベースのサーバーでMongoDB (1S) ワークロードを実行した場合との比較)<sup>8</sup>



## AIに対応するアップグレードを

インテル® Xeon® 6 プロセッサおよびインテル® Gaudi® 3 AI アクセラレーターで、データセンターを将来のAIワークロードに備えるチャンスとは。

[すぐに始める](#) ↗

## インテルのAI向けソリューション

インテルの生成AI、LLM、RAGのソリューションに関する取り組みの詳細は、「Building Blocks of RAG with Intel」eBookを参照してください。

[すぐに読む](#) ↗

intel®

# 脚注

1. [「AI power: Expanding data center capacity to meet growing demand」](#) (英語)、マッキンゼー、2024年10月29日。
2. 詳細は [Xeon 6 Performance Index](#) (英語) の [9T8] を参照してください。実際の結果は異なる場合があります。
3. [「インテル、AI およびネットワーキングをリードするソリューションを発表」](#)、インテル・ニュースルーム、2025年。
4. 詳細は [Xeon 6 Performance Index](#) (英語) の [9T221] を参照してください。実際の結果は異なる場合があります。
5. [「About open.intel」](#) (英語)、インテル。
6. [「Cost of a Data Breach Report 2024.」](#) (英語)、IBM。
7. DDR5-6400 RDIMMと比較した MRDIMM
8. [Xeon 6 Performance Index](#) (英語) の [9T8] を参照してください。実際の結果は異なる場合があります。

## 通知と免責事項

性能は、使用状況、構成、その他の要因により異なります。詳細については、[www.intel.com/PerformanceIndex](http://www.intel.com/PerformanceIndex) を参照してください。

性能の測定結果は構成情報に記載された日付時点のテストに基づくものです。また、公開中のすべてのアップデートが適用されているとは限りません。構成の詳細については、公開されている追加情報を参照してください。

絶対的なセキュリティを提供できる製品またはコンポーネントはありません。

実際のコストと結果は異なる場合があります。

インテルのテクノロジーを使用するには、対応したハードウェア、ソフトウェア、またはサービスの有効化が必要となる場合があります。

インテルは、サードパーティーのデータについて管理や監査を行いません。ほかの情報を参考にして、正確さを評価してください。

AI の機能には、ソフトウェアおよびプラットフォームのプロバイダーによるソフトウェアの購入、サブスクリプション、あるいは有効化が必要な場合や、特定の構成や互換の要件がある場合があります。詳細については、[intel.co.jp/aipc](http://intel.co.jp/aipc) を参照してください。

© Intel Corporation. Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。