

導入事例

インテル® Gaudi® 3 AI アクセラレーター



IBM Cloud上でインテル® Gaudi® 3 AI アクセラレーターの利用が可能に

選択肢を増やして多様なニーズに対応するIBMのハイブリッド・クラウド/AI戦略とは



日本アイ・ビー・エム株式会社

所在地：東京都港区虎ノ門二丁目6番1号
虎ノ門ヒルズステーションタワー

設立：1937年6月

資本金：1,053億円

事業内容：情報システムに関わる製品、サービスの提供

<https://www.ibm.com/jp-ja/>

空前の生成AIブームを経て、ビジネスにおけるAI活用は新たなステージに足を踏み入れています。AIありきの業務変革に取り組む企業や組織は増加傾向にあり、パフォーマンスはもちろん、コスト/可用性/セキュリティといった要素を加味してAI開発基盤の最適化を図る動きは加速しています。

「オープンなハイブリッド・クラウド」「ビジネスのためのAI」という戦略を掲げ、企業のAI活用を支援するソリューションを提供し続けるIBMでは、2025年3月31日～4月1日に開催したイベント「Intel Vision 2025」において、同社が運営するクラウドサービス「IBM Cloud」上で「インテル® Gaudi® 3 AI アクセラレーター」の利用が可能になったことを発表。GPUを前提としたAI開発基盤に新たな選択肢を提供し、インテルとの協業により多様化するエンタープライズAIのニーズに応えるソリューションの拡充を図っています。

IBM Cloud

インテル®Gaudi® 3 AI アクセラレーターを広く採用する最初のグローバル・クラウド・プロバイダー

IBM Cloudの各種サービスでインテル® Gaudi® 3 AI アクセラレーターをご利用いただけます。



watsonx

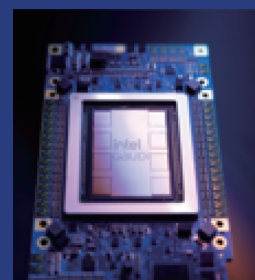


Red hat
OpenShift



Virtual
Servers

intel、インテル、IntelID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



本稿では、日本アイ・ビー・エム株式会社 テクノロジー事業本部 クラウド・プラットフォーム事業部の森 大輔 氏と青野 めい 氏、インテル株式会社 マーケティング本部の近藤 太郎に話を伺い、IBM Cloud上でインテル® Gaudi® 3 AI アクセラレーターの提供を開始した経緯と目的、導入のメリットについて紐解いていきます。

IBM のビジョン : お客様の経営課題をテクノロジーで解決する 変革パートナーになること

IBM の戦略

1. オープンなハイブリッド・クラウド
2. ビジネスのための AI

戦略遂行上の重要タスク

- ・ テクノロジーを活用したお客様との共創
- ・ パートナー企業とのエコシステム拡大
- ・ 技術革新をリード

コンサルティング	パートナー様との協業 ・販売・再販 ・オフリング連携、組込み ・システム開発・運用
オートメーション (自動化) プラットフォーム	アプリケーション開発とシステム間連携、システム基盤とネットワーク基盤の運用、システム資産の統治と管理
データ・プラットフォーム watsonx	AI アシスタントと AI エージェント、AI マネジメント、AI と機械学習のツール群、データ・ファブリック、データ・セキュリティ
ハイブリッド・クラウド・プラットフォーム Red Hat	サーバーとクラウドのオペレーション・システム、アプリケーションとコンテナ管理、仮想化、自動化、AI/ML プラットフォーム
トランザクション処理	他社クラウド AWS, Azure, その他
インフラストラクチャー	企業向け各種基盤
メインフレーム基盤 (IBM Z)、分散基盤 (TBM Power, IBM Cloud)、ストレージ基盤 (IBM Storage DS8000, FlashSystems)、ソフトウェア・ストレージ (Fusion など)	エッジ

ビジネスにおける AI 活用は新たなステージへ進み、 AI 開発環境への要望も変化

パフォーマンスと信頼性、セキュリティを担保し、ミッション・クリティカルな基幹業務システムの運用にも対応するクラウドサービス「IBM Cloud」を中心に、ビジネスのための AI プラットフォーム「IBM watsonx」や「Red Hat OpenShift AI」など AI 開発に必要なソリューションを展開する IBM。日本国内においても業界資格を有するコンサルタントが 5,000 人を超え、高度な AI スキルを有するスペシャリストは約 700 人を擁するなど、AI を用いた企業の業務変革を全方位で支援しています。主にアーキテクトとして、既存環境を踏まえつつ最新のクラウド/AI 技術を取り入れたシステムのデザインや提案を行ってきたテクノロジー事業本部 クラウド・プラットフォーム事業部の森氏は、同社の強みについて次のように語りました。

「IBM は“世界をより良く変えていく「カタリスト (触媒) 」になる”というミッションを掲げ、オンプレミス環境との親和性も考慮したオープンなハイブリッド・クラウドや、単に AI を業務に適用するのではなく、AI を前提として業務を再構築する“AI ファースト”のアプローチで企業を支援しています。上流のコンサルティングからアプリケーション、プラットフォーム、インフラストラクチャーまでをワンストップで提供できることが、ほかのベンダーにはない強みと捉えています」(森氏)

オープンなハイブリッド・クラウドを戦略に掲げる同社では、特定のベンダーにロックインしない環境の提供を目指しています。「お客様が運用しているオンプレミスの業務システムを、シームレスに最適な形でクラウド上に配置できるソリューションを展開しています」と森氏は語り、ベンダーロックインされない形でモダナイゼーションやイノベーションの実現を支援していると説明しました。

こうした IBM のアプローチは、近年の AI 開発基盤に対するニーズにマッチしています。IaaS や AI の分野で同社の最新技術動向や IBM Cloud の新機能といった情報発信に取り組んでいるクラウド・プラットフォーム事業部テクニカルセールス担当の青野氏は、AI のビジネス活用におけるトレンドと最新動向について話を展開しました。

「生成 AI がビジネストレンドとなってしばらく経ちますが、とにかく取り

入れて何ができるのか模索する、いわゆる PoC のフェーズは終わりを告げました。現在は、例えばチャットボットに画像を読み込ませて回答精度を上げるといった、具体的な活用を推進するフェーズに入っていると考えています。このため、AI モデル開発のライフサイクルでみると、新しいモデルを作る学習だけでなく、推論を行う部分の需要も高まっています。また当社でも取り組んでいますが、特定の業界や業務に特化した専用の AI モデルを作るという流れが加速しており、特化型モデルを作るためのファインチューニングに関してニーズが増えていると感じています」(青野氏)



日本アイ・ビー・エム株式会社
テクノロジー事業本部
クラウド・プラットフォーム事業部
青野 めい 氏

こうした状況を踏まえ、インフラ面でのニーズにも変化が見られると青野氏。「これまでは GPU 中心に、学習から推論までを行うのが一般的でしたが、推論の需要が高まるにつれて、コスト・パフォーマンスを考慮して GPU 以外の選択肢を求める企業も増えてきました」と現状を分析しています。森氏も「トランザクションに応じて課金されることを考えると、AI アプリケーションの利用拡大に伴い、コストの最適化が求められるのは必然的です」と AI 活用における課題に言及しました。

コスト・パフォーマンスに優れた インテル® Gaudi® 3 AI アクセラレーターは、 需要が高まる推論とファインチューニングの領域に最適

こうした課題が顕在化するなか、IBM とインテルは、IBM Cloud 上でインテル® Gaudi® 3 AI アクセラレーターの提供を開始したことを発表しました。インテル® Gaudi® 3 AI アクセラレーターは、Tensor プロセッサ・コア (TPC) と行列乗算エンジン (MME) を実装し、ディープ・ニューラル・ネットワーク (DNN) の演算処理を加速させます。この AI アクセラレーターは、大規模生成 AI 向けに最適化されており、コストを

参考) AI のライフサイクルによって異なる 必要なコンピュータ・リソース

準備	開発	デプロイ	
データの準備	分散トレーニング / モデルの検証	モデルの適用	推論
ステップのワークフロー (重複排除、ヘイトや冒流の除去など) AI のライフサイクルによって異なるコンピュータ需要	大きなインフラで長時間実行するジョブ	後続のタスクのためのカスタム・データセットによるモデル・チューニング	レイテンシー / スループットに敏感で、かつコストに敏感
処理時間: 時間 - 日 10-2000+ 低 - 中レベルの CPU コア	処理時間: 週 - 月 10-500+ 高レベルの GPU (ジョブ毎)	処理時間: 分-時 1+ 中 - 高レベルの GPU (ジョブ毎)	処理時間: 秒未満の API リクエスト -1 (キーもしくは一部) 低~中レベルの GPU

抑えながら高性能 GPU と同等の推論処理を実行可能です。インテルの近藤はインテル® Gaudi® 3 AI アクセラレーターを「エンタープライズ AI 向けの優れた選択肢」と説明します。

「青野さんが話されたように、現在の AI インフラは GPU ありきのイメージがありますが、GPU は価格の高騰をはじめ、入手性の問題、消費電力 / 発熱の問題と課題は山積です。インテル® Gaudi® 3 AI アクセラレーターは、こうした GPU の課題を解決する特徴を備えています。コスト・パフォーマンスに優れているのはもちろん、消費電力の面でも GPU と比べてアドバンテージがあります。先ほどから話が出ているように、昨今需要が高まっている推論処理やチューニングといった領域では十分なパフォーマンスを発揮でき、学習は GPU、推論処理はインテル® Gaudi® 3 AI アクセラレーターといった使い分けも有効と考えています」(近藤)



インテル株式会社
マーケティング本部
近藤 太郎

IBM も、こうしたインテル® Gaudi® 3 AI アクセラレーターの特徴を高く評価しており、企業それぞれのニーズに応えるエンタープライズ AI の選択肢として採用。IBM Cloud VPC (仮想プライベート・クラウド) 環境においてインテル® Gaudi® 3 AI アクセラレーターを活用したスタンドアロン・サーバーの提供を開始しました。

「インテル® Gaudi® 3 AI アクセラレーターが非常にコスト・パフォーマンスに優れた計算資源であることは理解しており、お客様により多くの選択肢を提供したいという目的で採用を決定しました。私たちはグローバルのクラウドベンダーでインテル® Gaudi® 3 AI アクセラレーターを採用した最初のベンダーとなります。VPC の仮想サーバーでインテル® Gaudi® 3 AI アクセラレーターが選択可能になったことを皮切りに、マネージド型コンテナ基盤を活用したいお客様向けには Red Hat OpenShift AI クラスタで提供、ビジネスのための AI プラットフォームの watsonx はソフトウェア版ですすでに対応し、IBM Cloud

上の SaaS 版も近日対応する予定です」(青野氏)

現状はワシントン D.C、フランクフルト、ダラスの 3 リージョンでの提供となっており、グローバルではすでに導入し、確かな効果を得られたという企業の事例も出てきています。森氏は、IBM Cloud 上でインテル® Gaudi® 3 AI アクセラレーターを導入するメリットについて、実際のベンチマーク結果を踏まえて説明しました。

「前述したように、AI 開発のライフサイクルにおいて、今後は推論やチューニングの需要が増えていくと考えており、大規模な言語モデルを作るハイスpek的な環境と、コストを抑えながら数多くの推論処理を行える環境を適材適所で使い分けことが重要と捉えています。インテル® Gaudi® 3 AI アクセラレーターは後者の選択肢として非常に魅力的です。IBM Cloud 上で提供され、as a Service の形で利用できることを評価する企業も多く、すでにグローバルにビジネスを展開されている保険会社様が株式や債券などクロスアセット投資のトレンド分析を AI で行うため、IBM Cloud × インテル® Gaudi® 3 AI アクセラレーターを導入。8 並列の GPU と同等のパフォーマンスを半分のコストで実現できたというベンチマーク結果も出ています。もちろん、大規模言語モデルの学習などでは GPU が優れたケースも多いですが、推論処理では GPU と比べて約 2 倍のコスト・パフォーマンスが期待できます。また、当社が提供する基盤モデル「Granite-8b」でベンチマークを取ったところ、シングルカードで 1 秒間に 5,000 トークンの処理を実行できるという結果が得られており、100 名のユーザーが同時に使っても十分な AI 推論処理能力があることが確認できています」(森氏)



日本アイ・ビー・エム株式会社
テクノロジー事業本部
クラウド・プラットフォーム事業部
森 大輔 氏

また、IBM Cloud で採用しているインテル® Xeon® プロセッサとの親和性も見逃せないメリットと森氏。IBM Cloud × インテル® Gaudi® 3 AI アクセラレーターの有用性は高いと手応えを口にしています。



AI技術の業務利用に必要な要素を網羅した「IBM Cloud×インテル® Gaudi® 3 AI アクセラレーター」がAI市場を牽引

IBMとインテルでは、今後も協業しながらIBM Cloud×インテル® Gaudi® 3 AI アクセラレーターの組み合わせで企業のニーズに応じていくといいます。Red Hat OpenShift AIクラスターおよびwatsonx SaaS版での提供は2025年第2四半期より開始の予定です。

「現在、国内リージョンでの提供も前向きに検討しています。今後もIBM Cloud並びにインテル® Gaudi® 3 AI アクセラレーターの最新情報をさまざまな形で発信して行きたいと思います」と青野氏は今後の展望を語りました。

森氏も「AIに関しては、本当にさまざまなお客様からお声がけいただいております。2024年の実績では400以上のお客様との実証プロジェクトがあり、そのプロジェクト数は増加傾向にあります」と、国内のAI活用における機運の高まりを語り、インテルとのパートナーシップを強化し、IBM Cloudとインテル® Gaudi® 3 AI アクセラレーターの組み合わせで多様なニーズに応じて行きたいと力を込めています。

IBM Cloud

<https://www.ibm.com/jp-ja/cloud/>

インテル® Gaudi® 3 AI アクセラレーター

<https://www.intel.co.jp/content/www/jp/ja/products/details/processors/ai-accelerators/gaudi.html>



性能は、使用状況、構成、その他の要因によって異なります。詳細については、<https://www.Intel.com/PerformanceIndex/>(英語)を参照してください。

性能の測定結果は、構成に示されている日付時点のテストに基づいています。また、現在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、補足資料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やコンポーネントはありません。

実際のコストや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っていません。ほかの情報も参考にしてデータの正確さを評価してください。

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。

テックプラス (2025年6月27日) に掲載されたコンテンツから抜粋し、再構成したものです。

インテル株式会社

〒100-0005 東京都千代田区丸の内1-4-1 丸の内永楽ビル25階

<http://www.intel.co.jp/>

©2025 Intel Corporation. 無断での引用、転載を禁じます。

2025年7月

「AI活用は持たなしの状況で、活用シーンは加速度的に増えていくと思います。AI活用が本格化すると、データの機密性やシステムの信頼性も重要になり、今後はコスト/可用性/セキュリティを意識して取り組んでいく必要があると考えています。その意味でも、IBM Cloudとインテル® Gaudi® 3 AI アクセラレーターの組み合わせは非常に魅力的な選択肢になると思います」(森氏)

コンサルティングからインフラ、プラットフォームまで、AI開発環境の構築と運用をワンストップで支援するIBMと、コスト・パフォーマンスに優れながら高性能なAIアクセラレーターであるインテル® Gaudi® 3 AI アクセラレーターを提供するインテル。両社の協業により生まれる先進的なソリューションには、今後も注視していく必要があります。

