

DXを成功させるカギはコストと性能のバランス

「AIと言えばGPU」一択ではない CPUでAIが動く時代の サーバー・インフラストラクチャー

AI活用が広がり、「AIと言えばGPU」という認識も生まれる中で、これからのサーバー・インフラストラクチャーの在り方はどう変わのでしょうか。インテルが新世代プロセッサーとともに提供する新たな選択肢とは？

AI基盤にもクラウド/オンプレミスの選択肢

企業のサーバー・インフラストラクチャーに求められる要件は、生成AI活用、アプリケーションの多様化、コストの最適化、セキュリティ対策の強化といったニーズが高まる中で変化しつつあります。例えばAIで必要とされるGPUなどの高性能な環境はコストがかさむため、コンピューティング性能とコストのバランスをどう取るべきかが悩ましい問題です。CPUの進化も著しく、大量のデータ処理が求められる生成AIの稼働環境でさえ、クラウド上のGPUだけではなくオンプレミスのCPUサーバーで稼働させるという選択肢が登場しています。

要件と選択肢が多様化する中で、これからのサーバー・インフラストラクチャーをどう構築するべきか、2025年6月27日に開催されたオンラインセミナー「生成AIでのDX本格化を支えるサーバー・インフラストラクチャーの在り方とは」からヒントを探ります。

生成AIの“ショック期”に企業がすべきこと

基調講演に登壇したアイ・ティ・アール シニア・アナリストの入谷 光浩氏は、生成AIやAI/MLプラットフォームを中心にIT投資が活発化している現状と、企業がとるべき戦略について語りました。同氏は、生成AIを手軽に利用できるサービスが増えて幅広い業務で活用が進む一方、ビジネス価値創造へとつなげるためには、サービスの「利用」から、業務フローに組み込む「構築」、検索拡張生成(RAG)など自社に合わせた「作り込み」、独自の大規模言語モデル(LLM)の「開発」へと段階的に進める必要があると指摘します。

「現在、生成AIは初期の高揚感が失われた“ショック期”にあります。今後の本格的な回復期に備えて今のうちにAI活用の基盤を整えるべきです」(入谷氏)

さらに入谷氏は、「これまでAI環境はクラウド環境での構築が主流でしたが、オンプレミスを検討する企業も増えています」と語り、パフォーマンスやコスト、セキュリティ、拡張性などそれぞれの特長を踏まえて自社の要件にあわせて検討する必要があること、クラウドとオンプレミス両方の選択肢を持つことの重要性を強調しました。

AIインフラにも活用できる最新CPU

AIインフラストラクチャーを考える際に意識したいのがCPUの進化です。性能や電力効率が大幅に向上しており、CPUベースでAIインフラストラクチャーを構築することも現実的な選択肢となっているため、最新世代のCPUに置き換えるだけでも、かなりの総保有コスト(TCO)削減が期待できます。具体的にどこまで変わるのか、インテル株式会社インダストリー事業本部シニア・ソリューション・アーキテクト 高藤 良史のセッションを中心に紹介します。



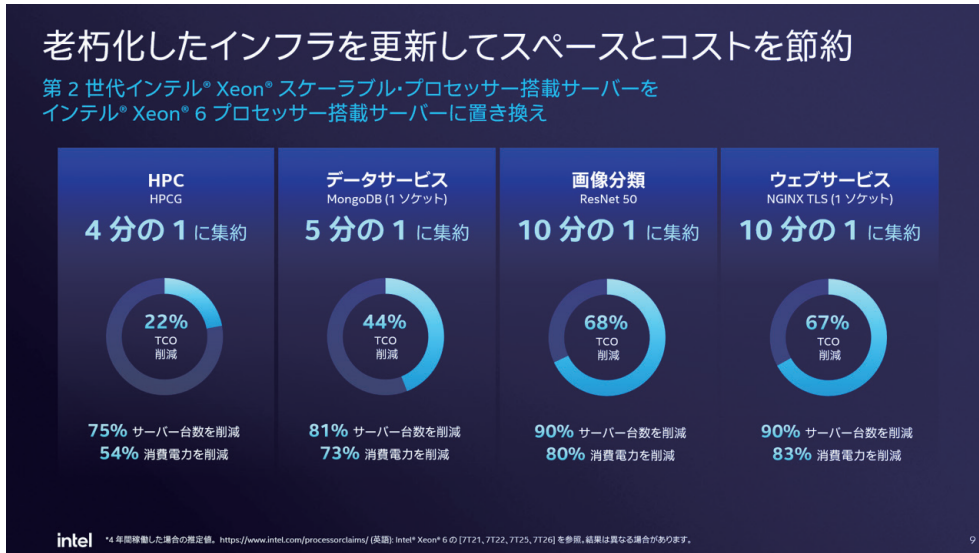


図1. インテル® Xeon® 6 プロセッサに切り替えることでスペースとコストの節約が可能

高藤は、コスト削減や電力効率の向上、スループット向上などデータセンターの課題に応えるべく、インテルのCPUは進化を進めてきたと説明。その成果として挙げたのが最新世代の「インテル® Xeon® 6 プロセッサ」です。AIやスーパーコンピューター、データベース分析、ネットワーク、コンテンツ配信など幅広いワークロードに対応し、前世代よりも格段にパフォーマンスを向上させています。データの信頼性と安全性を守る機能のほか、AIアクセラレーターを内蔵している点が特長です。インテル® Xeon® 6 プロセッサを1世代前と比較すると、同じコア数で20～50%ほどパフォーマンスを向上させています。

インテル® Xeon® 6 プロセッサは、汎用的なエンタープライズ用途に適した「P-cores (Performance-cores) 採用」と、消費電力を抑えた「E-cores (Efficient-cores) 採用」の2種類で展開。E-coresはコアの構成をシンプルにすることで集約率を高めており、マイクロサービス・アーキテクチャーやウェブ系の大規模ソフトウェア環境などに適しています。

高藤は「メディアストリーム処理の環境において、第2世代インテル® Xeon® スケーラブル・プロセッサでは200ラック必要だったものが、

インテル® Xeon® 6 プロセッサ (E-cores 採用) では、約66ラックと3分の1にまで削減できます。集約率が高まれば冷却効率も上がり、消費電力やCO₂排出量の削減にもつながります」と説明します。

第2世代インテル® Xeon® スケーラブル・プロセッサは2019年にリリース。当時導入したサーバーをインテル® Xeon® 6 プロセッサに切り替えただけで、TCOを20～70%削減できた試算もあります。サーバー更改においては従来と同じコア数のサーバーを検討する傾向が強くなりますが、コア数の多いサーバーは導入コストこそ高くなるものの、集約率が高まることでスペースや消費電力、そしてそれらにかかるコストの削減が可能です。CPU選定についても、TCOを含めた長期的な視点で検討する必要があります。

CPUレベルでもセキュリティを強化

サイバー攻撃の手口が巧妙化する今、AIが活用するデータやAIモデルそのものをいかに保護するかも重要な課題です。インテル® Xeon® 6 プロセッサはそうした状況に対応できるよう、セキュリティ対策も大幅に強化されています。

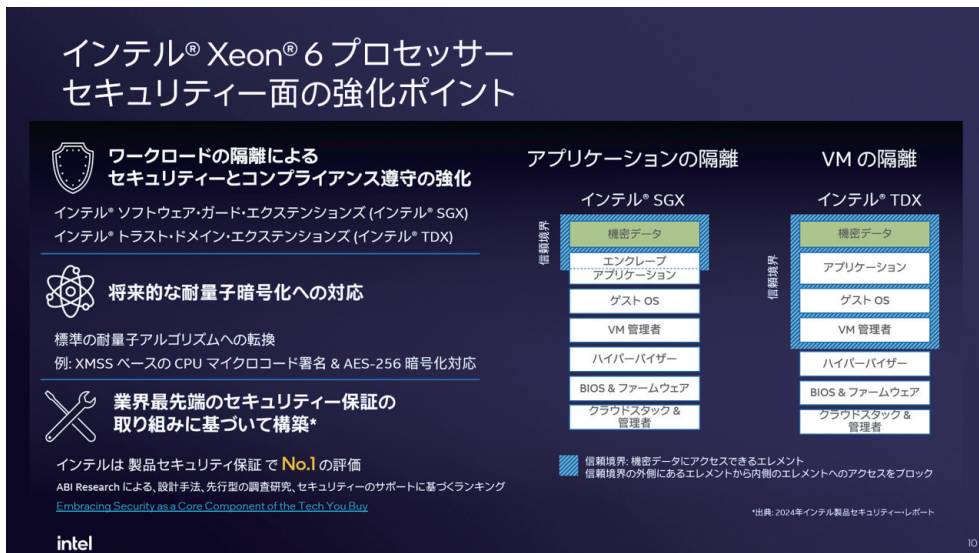


図2. インテル® Xeon® 6 プロセッサのセキュリティ面の強化ポイント

「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)」、「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)」という2つの技術を搭載。ハードウェアに内蔵された暗号化機能により、ほかのアプリケーションや仮想マシンからデータを隔離、保護し、機密データへの不正アクセスを防ぎます。「最近では量子コンピューティングが注目されていますが、インテルはその動向も見据えて耐量子暗号化への対応も視野にセキュリティを強化しています」と高藤は説明します。

AIをより手軽に使える世界へ

高藤がセッションでたびたび言及したのが、インテルが掲げる「AI Everywhere」というコンセプトです。「AI=GPUという認識を取り払い、CPUでもAIが手軽に動作する世界を実現できれば、AIの活用の幅が広がるはずですよ」と語ります。AIというと最近ではLLMや生成AIのイメージが強いものの、ビジネスの現場を見ると画像認識や画像分類の需要もいまだに根強く、こういった軽量のAI処理ではもはやGPUを使う必然性は薄れつつあると高藤は指摘しました。その中でインテルが提案するのが、CPUでAI関連のタスクを処理するアプローチです。

高藤はデータセンター向けAI製品としても、GPUではなくインテル® Xeon® 6 プロセッサーを挙げました。AIアクセラレーター「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)」を搭載しており、小規模なLLMであれば十分に実行可能。そして、700億パラメーター以上のより大規模なモデルに適しているのが、「インテル® Gaudi® 3 AI アクセラレーター」です。

では、GPUとAIアクセラレーターの違いはどこにあるのでしょうか。「GPUはハイパフォーマンス・コンピューティング (HPC) などでの利用が想定されており、倍精度浮動小数点など高精度の演算ができます。しかし、AIは比較的低精度で多くのパラメーターを保持する必要があり、高精度演算は必ずしも必要ではありません。そこで、これらの機能を排除し、AI処理に特化させたのがAIアクセラレーターですよ。」(高藤)

「RAG」の検索や、推論実行もCPUで対応可能に

CPUベースでのAI活用で、特に注目したいのがRAGです。生成AIモデルのパラメーターが増加して、複雑な処理をするにはモデルサイズの肥大化は避けられず、かといって企業ごとに大規模モデルをトレーニング

するのは現実的ではありません。そこでRAGを使い、企業が保有するデータセットを事前にベクトル化し、ユーザーの質問に近い情報を検索、参照して回答を生成することで回答精度を高めます。

RAGの仕組みにおいて、ベクトル型データベースへの問い合わせなどはインテル® Xeon® 6 プロセッサーで対応可能です。「RAGのレスポンスタイムにはプロセッサーが大きく関わります。例えば、チャットサービスのやりとりではレスポンスタイムがシビアに問われますが、このような領域でのデータ検索にインテル® Xeon® 6 プロセッサーは有効ですよ」

従来のCPUではベクトル検索の処理性能を十分に引き出すのが難しい課題となっていました。インテル® Xeon® 6 プロセッサーはAI処理高速化技術である「インテル® AMX」を組み合わせることで、RAGにおけるベクトル検索をより効率的に実行することを可能にしました。「インテル® AMXはCPUコアに内蔵されたアクセラレーターです。コア数が多ければ、それだけ多くのインテル® AMXを持つことになります。つまりコア数が多いCPUを利用することで、用途によってはGPUを使わなくても効率的な推論が可能になります」

インテルでは実際に、インテル® Xeon® 6 プロセッサーとRAGを使い、監視カメラの映像を検索する実験をしています。ユーザーが動画に「不審な人がいたか」といった問い合わせをすると、RAGが動画ログを見た上でベクトル型データベースを検索し、LLMに結果を返します。その結果をもとにLLMがユーザーへの回答を生成するという仕組みです。

AI処理でGPUが必要とされるのは主に学習のフェーズです。推論を実行するだけならばGPUは必要ない場合も多く、となれば、環境への要件も変わってきます。精度が高く商業利用可能なモデルが多く登場する今、推論処理だけを実行したいというケースもあるでしょう。その際、GPUほどコストをかけず、CPUでAIインフラストラクチャーを構築するというアプローチは魅力的です。

コスト・パフォーマンスに優れたAIアクセラレーター

インテル® Gaudi® 3 AI アクセラレーターは、1台のサーバーノードに最大8基搭載できるユニバーサル・ベースボードのほか、PCIeカードでも提供しています。128GBのHBM2Eメモリーを搭載し、1,678T(テラ)FLOPSという高いパフォーマンスを発揮します。

「インテル® Gaudi® 3 AI アクセラレーターはIBM Cloudでも採用され



図3. RAGを用いたカメラ映像検索の概要

ています。IBMのAIモデル Granite 3.1 8bでは25～40%ものコスト効率化を実現しました。コストとパフォーマンスのバランスが良い製品だと言えるでしょう。PCIe タイプもAIアクセラレーターの選択肢の1つとしてご検討いただきたいと思います」

最後に高藤は、製品のオープン性に触れて次のように締めくくりました。「特定のベンダーだけがAIビジネスに関わるのではなく、多様な企業がAIをより手軽に活用できる製品を提供することで、AIの民主化が進むとインテルは考えています。そのためにも、AIのオープン・エコシステムを活用することが重要であり、インテルの製品はその理念に基づいています」



インテル株式会社 高藤 良史

インテル® Xeon® プロセッサー :

<https://www.intel.co.jp/content/www/jp/ja/products/details/processors/xeon.html>

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>

データセンターの刷新と拡張を検討しているユーザー向けガイド「インテル® Xeon® プロセッサー・アドバイザー・スイート」:

<https://xeonprocessoradvisor.intel.com/welcome>

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。

@IT Special (2025年7月4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

366392-001JA
JPN/2507/PDF/SE/MKTG/TK