

導入事例

インテル® Xeon® プロセッサ



GPUなしでも生成AIが使える!?

IIJのビジネスを支える IT基盤の運用負荷をAIで軽減

インテル® Xeon® プロセッサで推論を行う生成AIとは

ネットワークからクラウド、セキュリティ、モバイル、システム・インテグレーションまで幅広く事業を展開し、多種多様なITサービス・ソリューションで社会インフラ基盤を支えてきた株式会社インターネットイニシアティブ（以下、IIJ）。同社が膨大なサービス群を安定稼働させるためのインフラ基盤として構築/運用している「Next Host Network (NHN)」は、2008年10月より提供を開始しました。市場の変化やサービスの拡大に合わせて強化を重ね、現在は9,000台以上のサーバーで構成される大規模サービス基盤へと進化を遂げています。

株式会社インターネットイニシアティブ

所在地：東京都千代田区富士見2-10-2
飯田橋グラン・ブルーム

設立：1992年12月

資本金：23,023百万円(単体)

事業内容：インターネット接続サービス、WANサービスおよびネットワーク関連サービスの提供、ネットワーク・システムの構築・運用保守、通信機器の開発および販売

<https://www.ij.ad.jp/>

NHNのサーバー群には、開発当初からインテル® Xeon® プロセッサが使われており、2023年の新サーバー選定においても、第4世代インテル® Xeon® スケーラブル・プロセッサを採用。同社では、CPUでの推論パフォーマンスを向上させる新しい拡張命令セット「インテル® AMX」が同プロセッサに搭載されていることに着目し、GPUを用いずにCPUで推論を行う生成AI活用の取り組みを開始しています。

本稿では、NHNの運用および今回の生成AI活用に携わっているIIJ基盤エンジニアリング本部基盤技術部 システム基盤技術課 課長代行 早田 祥弘 氏に話を伺い、同社のサービス基盤と生成AIの活用において、インテル® Xeon® プロセッサが担う役割を紐解いていきます。



株式会社インターネットイニシアティブ
基盤エンジニアリング本部
基盤技術部 システム基盤技術課
課長代行
早田 祥弘 氏



IIJサービスの全体像



株式会社インターネットイニシアティブ 営業資料より

第4世代インテル® Xeon® スケーラブル・プロセッサーに実装された「インテル® AMX」が“GPUいらず”の生成AI開発のカギに?

IIJが提供する膨大なITサービス・ソリューションは、サービスの運用効率向上とコスト圧縮を目的に開発されたサービス基盤インフラ「NHN」上で稼働しています。まさに同社のビジネスを支える根幹であるNHNですが、その分、運用チームにかかる負荷は大きく、基盤運用業務の効率化は喫緊の課題となっていました。NHNの構築/運用を担当しているシステム基盤技術課の早田氏は次のように語っています。

「IIJが展開するサービスは多岐にわたり、その基盤であるNHNに対するニーズも多種多様です。サービス自体が日々進化し続けていることもあり、毎日何かしらの調整を行っているような状況でした。ITサービスはスピード感が重要になりますので、“いつまでに作業を終わらせてほしい”という要求も多く、優先順位付けとスケジュール管理の部分に課題を感じていました」(早田氏)

適切かつスピーディーなサービス基盤運用を実現するためには、サービスを提供している事業部とのコミュニケーションを密にする必要があったと早田氏は話します。しかし、限られたメンバーで24時間365日対応していくことは困難だったと課題を口にし、近年のビジネストレンドといえる「生成AI」の活用に着目したと話を続けました。

「当時は、生成AIのビジネス活用に取り組む企業も多く、AIブームが起きていた時期でしたが、そのほとんどがGPU(グラフィック・ボード)の利用を前提としたもので、IT基盤の運用を行っている立場からはリスク

も大きいと感じていました。NHN自体が、さまざまなニーズに対応するための“汎用性”を重視しており、それがCPUにインテル® Xeon® プロセッサーを採用してきた理由の1つにもなっています。そのNHNの運用を効率化する手法として、調達性や互換性、さらにコスト面でも不安が残るGPUを前提としたアプローチは採用しづらく、CPUで推論処理を行えないかという観点で検討を開始しました」(早田氏)

互換性(汎用性)とパフォーマンス(性能)の観点からインテル® Xeon® プロセッサーを採用してきたNHNでは、2023年の新サーバー選定にあたって、第4世代 インテル® Xeon® スケーラブル・プロセッサーを採用しています。早田氏は、このインテル® Xeon® プロセッサーに搭載されている拡張命令セット「インテル® AMX」に注目し、GPUを用いずにCPUで推論処理を行うチャットボットの開発をスタートさせました。

「利用者からの“より強い計算リソースを”という要望に対応したサーバーメニューを開発するため、インテル® AMXが実装された第4世代インテル® Xeon® スケーラブル・プロセッサーを採用したという背景もあり、NHNの運用効率化にあたってインテル® AMXの活用を第一に考えました。具体的には、社内システムに書き込まれたNHNへのリクエストをAIが解析し、必要な情報を抽出してチャットに流すという仕組みを構築することで人的ミスを減らし、読み落としや対応遅れの抑制を図っています」(早田氏)

量子化技術で計算量を削減し、CPU推論のパフォーマンスを担保

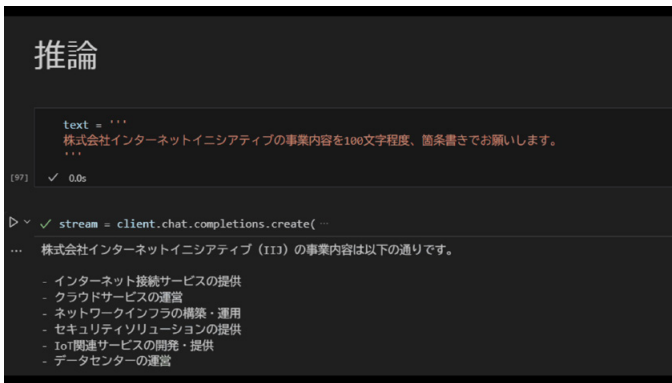
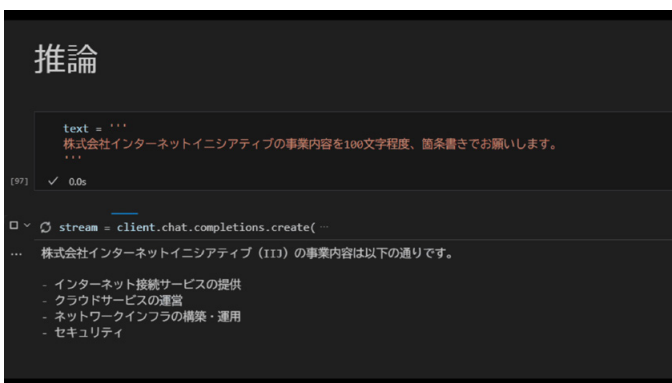
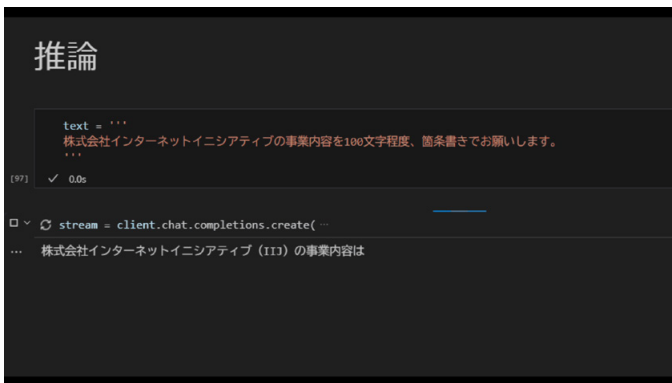
こうしてインテル® AMXの活用により、CPUで推論を行うチャットボット・サービスの開発に着手した早田氏は、「インテル® PyTorch 拡張パッケージ」というライブラリーを用いて推論性能を調査。小規模言語モデル(SLM)の「Phi-4」や、大規模言語モデルの「Llama」などを用いて検証を重ね、高いリアルタイム性を必要としないチャットボット用途では十分な推論速度が得られることを確認しました。

「AI関連の技術は日々進化を続けており、計算量を減らす量子化技術も実用化されてきています。今回の取り組みでも、インテル® AMXを使った量子化技術が出てきたことで、文章を生成する速度が大幅に向上しました。当初は1つの文章を生成するのに5~6分かかっていたのですが、現在は1分以内で作れるようになっています」と早田氏。

GPUの推論処理速度と比較すると多少の差異はあるものの、リアルタイムのAIチャットなどでなければ問題ないパフォーマンスを得られていると手応えを口にします。

またインテル® AMXの活用にあたっては、インターネットのコミュニティなどから情報を収集しながら、ほぼ独学で進められたと早田氏は話します。「試行錯誤を続けるなかで、コミュニティ内のプロジェクトにコミットしているインテルのエンジニアから提供された最新情報や、ナレッジには助けられました」とインテルからの間接的なサポートにも言及しました。

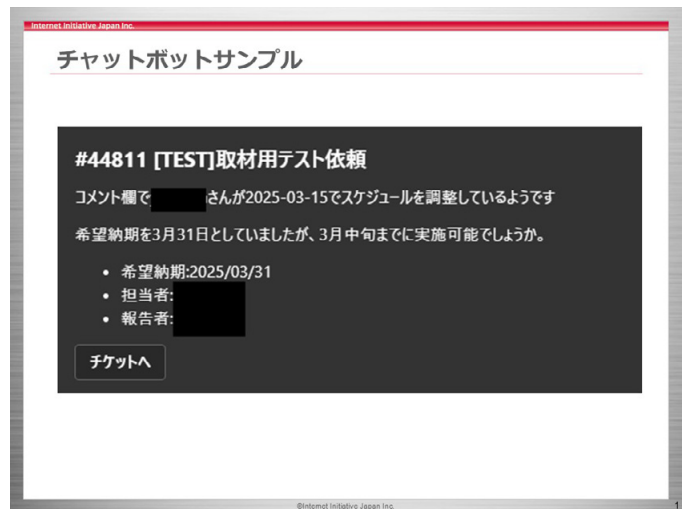




実際にインテル® Xeon® プロセッサーを用いて推論を行っている様子

こうして、GPUを使わないチャットボット・サービスの稼働が開始されました。現在はPoC段階ですが、すでに実際の業務で使われており、想定どおりの効果が得られているといいます。

「これまで人が文章を読んで判断していた業務が、AIの自動分析で問題点を抽出し提示する仕組みを導入することで大幅に効率化できました。ミッション・クリティカルなサービスが稼働しているNHNの運用では、週末や深夜に対応が必要になるケースも当然あり、そういった人手が不足する時間帯でもAIがしっかりと問題点に気づき、必要な



チャットボットのサンプル画面

情報を教えてくれるのは非常に大きな効果といえます。管理職の立場から見ても、簡潔な文章で情報を把握できるようになり、大変助かっています」(早田氏)

ただ、推論速度が向上した結果、運用チームに膨大な量の通知が届くようになり、運用負荷の軽減という意味では完全に課題を払拭できたわけではないとし、早田氏は「今後は運用する“人”と“機械(AI)”のバランスを考えながら、改良を進めていきたい」と力を込めます。

「今回の取り組みもそうですが、AI活用では定量的な効果を測定しづらいケースが少なくありません。ただし、今後もAI活用を促進していくことを考えれば、統計的に調べるための手段を確保する必要があり、サービスの開発と併行して取り組みを進めているところです」(早田氏)

生成AI活用における選択肢の幅を広げるため、IIJの挑戦は続く

今回の取り組みを踏まえ、IIJでは今後も幅広い領域で生成AIの効果的な活用に取り組んでいくといいます。NHNの運用をはじめとする社内業務への適用はもちろん、事業部門においてもAI活用の気運が高まっており、NHNに対する要望も増えてきている状況です。そのなかで、インテル® Xeon® プロセッサーとインテル® AMXによるCPU推論のポテンシャルを明確化したチャットボット・サービスは、事業部門のAI活用における選択肢を増やすという効果も生み出しています。

「提供中のサービスをAIで強化したいので、高性能なGPUを入れてほしいといったリクエストも増えてきていますが、今回の開発で、インテル® Xeon® プロセッサーでもこれだけの推論処理が可能だと提案できるようになりました。実際、当社の社員が自由に記事をアップできるサイトで今回の成果を公開したところ、かなり反響が大きく、AI活用は必ずしもGPUありきではないということが認知されてきたように感じています」(早田氏)

同社では、GPUを用いたAI活用も含め、さまざまな選択肢を提供できるようNHNの整備を進めていく予定です。そのなかで、AIの推論処理をはじめ、多様なワークロードに対応するインテル® Xeon® プロセッサーの有用性は高まっていくと早田氏は見えています。インテル® Xeon® 6 プロセッサーの導入も視野に入れ、生成AIの業務活用に向けた活動を続けていきたいと今後の展望を語りました。

「AIは、チャットだけでなくいろんな形で我々を支えるひとつの手段として利用されていくものと捉えています。もちろん、IIJが提供するサービス・ソリューションにおいてもAIの活用は不可欠になっていくはずで、その意味では、さまざまなプラットフォームで運用できる知見を蓄積することが大切になります。冒頭で話したように、我々が運用しているサービス基盤は、汎用性と性能を重視して構築しており、AI活用においても1つの技術に依存することなく、ニーズに合わせて最適な機材の提供を目指して検証を進めていきたいと考えています」(早田氏)

AI活用にはGPUが不可欠という固定観念を打ち破り、インテル® Xeon® プロセッサーの拡張命令セットを用いたCPU推論の可能性を切り開いたIIJの取り組みは、生成AIによる業務効率化やイノベーションの創出を目指す企業と組織にとって重要な“気づき”を与えてくれるはずです。

※取材内容は2025年3月時点での情報です。



株式会社インターネットイニシアティブ

<https://www.ij.ad.jp/>

インテル® Xeon® プロセッサー

<https://www.intel.co.jp/content/www/jp/ja/data-center/what-is-xeon-processor.html>

intel
XEON

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3月28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4月