AI活用の可能性を広げる新たな選択肢

組織内のナレッジを、AIで安心・快適に活用

AIの応答をもっと速く、GPUをもっと有効に
― KVキャッシュオフロードで、業務AIを快適に ―

企業や自治体では、蓄積された業務データをAIで活用する動きが広がっています。一方で、個人情報や機密情報の適切な管理に加え、利用者や参照データの増加に伴う応答遅延への対応も課題となっています。

本記事では、業務データを組織内に保持してAIを活用する環境を前提に、応答性能の向上とGPUリソースの効率的な利用を図る「KVキャッシュオフロード」に着目します。Hitachi Virtual Storage Platform One Blockを活用した構成と、想定ユースケースに基づく評価結果を通じて、その効果と活用のポイントをご紹介します。

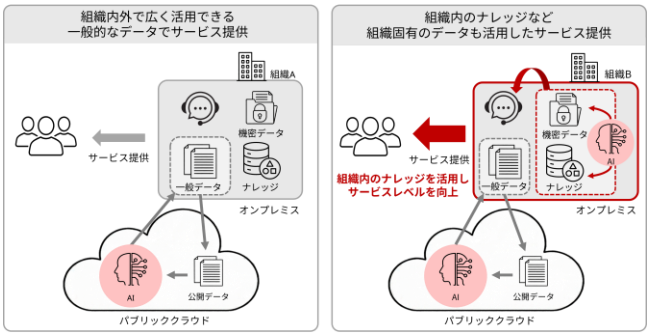
AI活用の急拡大

AIは、文書作成、問い合わせ対応、情報検索、業務マニュアルの参照など、さまざまな場面で活用が進んでいます。しかし、それはもはや、一部の先進企業だけが利用する技術ではなく、幅広い組織が業務へ取り入れる身近な存在になり、公開情報などの一般的な情報だけをAI活用するだけでは、差別化が難しくなっています。今後は、AIをより密に業務へ組み込み、かつ自組織ならではの価値創出につなげることが激しいAI活用戦争を勝ち抜くカギになります。

組織内のデータ（ナレッジ）活用における現状と課題

注目すべきは、自組織だけが保有する「組織内のデータ（ナレッジ）」活用です。「組織内のデータ」には例えば、お問い合わせの履歴、業務マニュアル、稟議書、契約書、医療記録、地域のお客さまに関する知見など、長年蓄積された貴重なノウハウが含まれています。これらの組織内のデータをAIで活用できれば、利用者は必要な情報へ短時間でたどり着けるようになり、業務効率化やお客さま満足度向上が期待されます。これはまさに、その組織だけが持っている価値であり、ほかの組織には真似ができない、その組織の価値そのものになるのです。

一方で、こうした組織内のデータには個人情報や機密情報が含まれることも多く、クラウドサービスへの保存や外部への持ち出しが難しいケースも少なくありません。また、法令や業界ガイドライン、組織内ルールへの対応が求められるため、AI活用を進めたくても踏み出せないという課題があります。



オンプレミス型AI基盤の重要性

こうした課題への有効な選択肢が、組織内の設備でAIを運用する「オンプレミス型AI基盤」です。オンプレミス環境であれば、組織内のデータを外部へ持ち出さずにAIで活用できます。セキュリティやコンプライアンスへの不安を抑えながら、組織の方針に沿ってAI利用を進められます。

AI活用が広がる中で、「安心して利用できること」は導入の前提条件になりつつあります。その観点からも、オンプレミス型AI基盤の重要性は今後さらに高まると考えられます。

組織内外でのサービス利用拡大に伴う性能への影響

一方で、AI利用が組織内に浸透すると、新たな課題も見えてきます。利用者数やAIに参照させるデータ量が増えると、回答までの待ち時間が長くなり、業務効率が低下するおそれがあります。また、長い会話や複雑な問い合わせ業務では、過去の会話履歴や参照文書を繰り返し処理するため、推論処理の負荷が高まり、さらなる応答遅延につながる可能性があります。

AIの導入時は問題にならなくても、利用部門や利用者が増えるにつれて、こうした性能面の課題が顕在化してきます。

課題解決の必要性

AIによる組織内のデータ活用は、組織の業務効率向上だけでなく、お客さま向けのサービス品質向上や競争力強化にもつながります。しかし、その価値を十分に引き出すためには、「安心して利用できること」に加え、「快適に使い続けられること」が重要です。これらの課題を解決し、安心・安全なオンプレミス環境で組織内のデータを最大限に活用できるAI基盤を整備することが、これからのAI利用における価値創出を後押しします。

また、この課題を解決する技術としてKVキャッシュオフロードがあります。ここからは、KVキャッシュの仕組みと期待される効果について解説します。

KVキャッシュとは

KVキャッシュオフロードの説明をする前に、まずは簡単にAIにおける推論時の動作についてご説明します。

AIの推論処理は、AIチャットサービスやAIエージェントから入力された情報を「読む工程（Prefill）」と「回答を作る工程（Decode）」に分けて処理しています。この「読む工程（Prefill）」において、「一度読んだ内容をメモしておく仕組み」がKVキャッシュです。この仕掛けがあることで、以前会話に出てきた内容を全部読み直すことなく、効率的に回答作成をすることが出来るのです。

もしこの「メモしておいた内容（KVキャッシュ）」がなければ、再び、過去の会話履歴を含めた入力情報を読む工程が必要となり、入力情報が増えるほど、この読み直しの作業量が増える、ということになり、回答に時間がかかります。

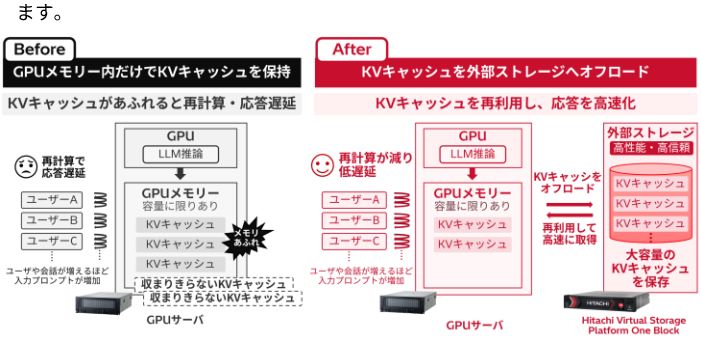
KVキャッシュオフロードの仕組み

AIの推論システムは、AIチャットサービスやAIエージェントから受け取った入力をGPUで処理します。入力を毎回最初から計算すると時間がかかるため、計算途中の結果を「KVキャッシュ」として保存し、同じ内容を再び処理するときに再利用します。これにより、回答の生成を早く始められます。

KVキャッシュは通常、GPUメモリーに保存します。しかし、入力が長くなったり同時利用者が増えたりすると、KVキャッシュがGPUメモリーを圧迫します。必要なKVキャッシュを保持できない場合は、過去に実施したPrefillを再計算するため、回答開始までの時間が長くなります。この課題を解決する技術が「KVキャッシュオフロード」です。すぐに使わないKVキャッシュをGPU外のサーバメモリーやストレージへ退避し、必要なときに読み戻します。

保存先にストレージを利用すると、GPUメモリーやサーバメモリーだけで保持する場合よりも、大容量のKVキャッシュを保存できます。一方、ストレージ障害によってキャッシュを失うと再計算が必要になり、システム全体の応答性能に影響します。そのため、容量だけでなく、性能、信頼性、可用性を備えた保存先を選ぶことが重要です。

この保存先として活用できるのがHitachi Virtual Storage Platform One Blockです。Hitachi Virtual Storage Platform One Blockは、大量のデータを高速に処理するとともに、データ保護や高可用性構成に対応しています。再利用するKVキャッシュをHitachi Virtual Storage Platform One Blockへ保存することで、GPUメモリーの容量制約を補いながら、キャッシュを安定して保持できます。

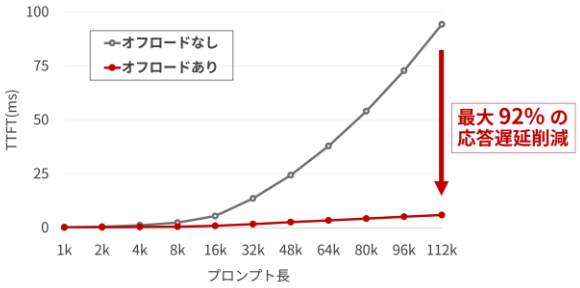


Hitachi Virtual Storage Platform One Blockを活用したKVキャッシュオフロードによる効果

Hitachi Virtual Storage Platform One BlockへKVキャッシュをオフロードし、推論性能への効果を評価しました。評価指標には、入力を受け取ってから最初のトークンの生成が始まるまでの時間を示すTTFT（Time To First Token）を使用しました。TTFTが短いほど、利用者は回答が早く始まったと感じます。

評価の結果、入力データ量（プロンプト長）にかかわらず、Hitachi Virtual Storage Platform One Blockへ保存したKVキャッシュを再利用することでTTFTが短縮しました。入力が長いほど効果は大きくなり、最大92%短縮しました。これは、過去の計算結果を読み戻すことで、長い文書や大量の入力をGPUで再計算する処理を減らせるためです。

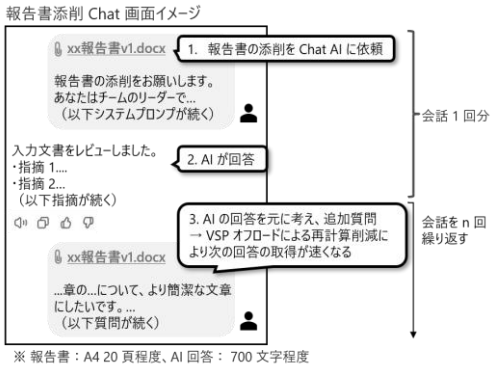
Hitachi Virtual Storage Platform One Blockを利用したKVキャッシュオフロードは、推論処理の応答性を高めるとともに、再利用するKVキャッシュを大容量かつ安定して保持する方法です。



想定ユースケースでの評価

入力が長いほどKVキャッシュ再利用の効果が大きくなることを確認しました。そこで実業務に近い条件として、組織内で保管されている研究報告書をチャットAIで添削するユースケースを評価しました。同じ長文を参照して修正や確認を繰り返すため、オフロードの効果を確認しやすい業務です。

利用者は約20ページの報告書を添付し、最初の回答後に追加の修正や確認を依頼します。初回は報告書全体と質問を処理してKVキャッシュを生成し、2回目以降はHitachi Virtual Storage Platform One Blockに保存したKVキャッシュを読み戻して追加質問を処理します。



評価では、応答開始までの時間TTFT（Time To First Token）、1時間当たりの処理人数、同等の処理能力に必要なGPUリソースを比較しました。以下に想定ユースケースでの評価結果を示します。

再参照時のTTFT （最初の回答が返るまでの待ち時間）	1時間当りの利用ユーザー （一定の応答水準で処理できる人数）	同時処理に必要なGPU
4.3 秒 ➡ 0.8 秒 約81%短縮	51人 ➡ 66人 約29%増加	9.7枚 ➡ 7.5枚 約22%削減

※上記は、評価条件(Qwen3.8ー32B、平均入力約1.4万トークン、出力1,024トークン、50ユーザー・50文書)に基づく結果です。利用ユーザー数は、低負荷時のTTFTの2倍に相当する8.6秒を応答性能の基準として算出しています。GPU数は論理値であり、実際の必要GPU枚数を示すものではありません。

KVキャッシュオフロードが効果を発揮しやすい業務

KVキャッシュオフロードの効果は、業務条件によって変わります。効果が大きくなりやすいのは、「長いコンテキスト」「多数の同時利用」「高いKVキャッシュ再利用率」が重なる場面です。長い文書や履歴はPrefillの計算量を増やし、同時利用はGPUメモリーを圧迫します。同じ情報を繰り返し参照する業務では、保存したKVキャッシュの読み戻しにより、この負荷を軽減できます。

適合しやすいのは、業務マニュアルや規程の照会、履歴を踏まえた継続対応、契約書・報告書・設計書の確認、AIエージェントです。一方、短い単発質問や参照内容が毎回変わる業務では効果が限定的です。適用時は文書の長さ、同時利用数、再利用率を確認します。

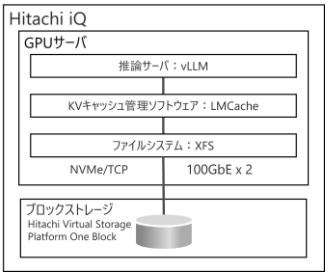
ユースケース	KVキャッシュを再利用する場面	長いコンテキスト	多数の同時利用	高いKV再利用率	総合適合度
契約書・報告書・設計書の確認／作成	同じ長文で要約、確認、修正を繰り返す	○	△	○	△
業務マニュアル・規程の検索／照会	同じ規程やマニュアルを複数の質問・利用者が参照	○	○	○	○
問い合わせ履歴を踏まえた継続対応	長い会話履歴を引き継いで追加質問する	○	○	○	○
AIエージェントによる処理	共通の指示や文書を使ってLLMを複数回呼び出す	◎	△	◎	◎
短文の単発質問	短い入力を一度だけ処理する	×	△	×	×

日立ヴァンタラからのご提案

ここまで、組織内に保管されている研究報告書をチャットAIで添削するユースケースの評価を通じて、長い文書を繰り返し参照する業務では、KVキャッシュのオフロードによってAI推論を効率化できることを確認しました。こうした業務でAIを実用化するには、推論性能の向上だけでなく、データを組織内で管理できる安心感と、AIを快適に使い続けられる性能の両立が重要です。

その実現に向けて、日立ヴァンタラは、AIの導入から開発、業務適用までを一貫して支援するAI統合プラットフォーム「Hitachi iQ」を提供いたします。Hitachi iQを活用したソリューションでは、組織内にデータを保持したままAI推論を行い、GPUメモリーだけでは保持しきれない再利用するKVキャッシュをHitachi Virtual Storage Platform One Blockへオフロードできます。これにより、データを組織外に持ち出すことなく安心・快適に使い続けられるAI環境を実現します。

この環境を支える構成は次のとおりです。質問と文書は推論サーバ(vLLM)で処理し、KVキャッシュ管理ソフトウェア(LMCache)がKVキャッシュの検索、保存、退避、読み戻しを管理します。LLMモデルとアクティブなKVキャッシュはGPUメモリー上に配置し、GPUメモリーから退避したKVキャッシュをHitachi Virtual Storage Platform One Blockへ保存します。必要時に読み戻すことで、同じ長文コンテキストに対するPrefill処理を再実行せずに済みます。



※図は想定構成です。実際の構成や設定は、モデル、利用条件によって異なります。

KVキャッシュオフロードは、GPUをストレージへ置き換える技術ではなく、GPU推論を支えながら再利用可能な計算結果の置き場を広げる技術です。

Hitachi iQ製品の組み合わせにより、組織内のデータを組織内に保ちながら、応答開始時間の短縮、同時利用時の処理能力向上、GPUリソースの効率化を実現しやすくなります。効果は利用条件で変わるため、業務適用の際には実データと想定負荷による事前評価が必要です。

日立ヴァンタラは、KVキャッシュオフロードを活用したソリューションを通じて、
お客さまのAI活用のさらなる進化を支えます。
お客さまの具体的なユースケースに基づき、価値創出の実現に向けて伴走してまいります。
本記事に関するご質問や、AI活用に関するご相談がございましたら、
右記のお問い合わせフォームまたは担当営業までお気軽にご連絡ください。

お問い合わせフォーム

Web掲載記事

日立ヴァンタラ株式会社

本資料に記載の会社名、製品名などは、それぞれの会社の商標もしくは登録商標です。
本資料の内容は2026年11月時点の情報です。
©Hitachi Vantara, Ltd. 2026. All rights reserved.