IT リーダーのための AI 推論パフォーマンスガイド

この ebook は、急速に進化する今日のテクノロジー環境における AI インフラストラクチャの複雑さを乗り切るために不可欠なロードマップです。AI の導入を進める IT リーダー向けに作られたこのガイドでは、AI のユースケースがパフォーマンスの測定とインフラの最適化にどのような直接的影響を与えるのかについて分かりやすく説明しています。

AI ワークロードには、高性能、高信頼性および高効率なインフラストラクチャが必要です。しかし、自社のインフラが特定の AI ユースケースの要求を満たせるかどうかについては、どのように判断すればよいのでしょうか？

この ebook では以下の点を探ります。

- **パフォーマンス測定におけるユースケースの役割**：様々な AI アプリケーションがどのように独自のインフラ要件を生み出しているのかを理解します。
- **AI 推論の主要指標**：レイテンシー、スループット、エネルギー効率など、確実に成功するために測定すべき事項について学びます。
- **インフラ最適化のベストプラクティス**：テクノロジースタックをビジネス目標に合わせるための実行可能な戦略をご覧ください。

洞察、フレームワーク、事例が盛り込まれた、「IT リーダーのための AI 推論パフォーマンス最適化ガイド」は、AI ソリューションを効果的に評価、デプロイ、拡張するための知識を習得できる一冊です。AI 時代に、自信を持ってリードしたい意思決定者にとって必読の書となります。

目次

AI 推論の未来を解放

NVIDIA AI 推論プラットフォームを使い始めるには

推論パフォーマンスとは何か、そしてなぜ重要なのか

推論は AI を現実世界へと導き、コード作成から長文の要約、さらには画像や短い動画の生成まで、様々なアプリケーションを実現するための高度なデプロイの課題を解決します。推論技術がもたらすワークフローの自動化、新たなビジネスモデルの創出、革新的な製品の推進を可能にする力は真に変革的なものであり、特に企業内で AI イノベーションを先導する立場にある CIO や IT リーダーにとって大きな意味を持ちます。

しかし、このような可能性がある一方で、重要な課題も存在します。それは、レイテンシーとパフォーマンスのバランスを維持しながら、推論ワークロードをスケールアップする際に発生する継続的なコストをいかに管理するかということです。

トークン単価の課題を解決

オンプレミス AI インフラを維持する場合、提供される推論リクエストのすべてで、ハードウェア アクセラレーター、推論ソフトウェア、継続的な推論の最適化と管理などに生じるコストをカバーするのに十分な収益を生み出す必要があります。IaaS（Infrastructure as a Service）を利用する場合、経済的な負担は月次または年次のアクセラレーテッド コンピューティング インスタンスと推論 SaaS（Software-as-a-Service）料金へと移行します。どのデプロイ モデルを選択するのであれ、AI 推論からの投資対効果 (ROI) がこれらのコストを正当化できることを確保することは、IT リーダーにとって重要な関心事です。

NVIDIA の使命は、AI のパフォーマンスを最適化し推論コストを削減し、組織がこれらの課題に対処できるように支援することです。AI 推論システムの効率性を高めることで、NVIDIA は企業が進化し続ける AI デプロイ要件に対応するために必要な拡張性、パフォーマンス、ユーザー体験を維持しながらコストを削減できるように支援します。

推論コストは 2 つの主要な要素と密接に関連しています。その 1 つは、AI プラットフォーム自体のパフォーマンス、つまりスループットです。システムが効率的に処理できるリクエスト数が多いほど、組織は受信するリクエスト全体にコストを分散させることができ、リクエストあたりのコストを削減することができます。しかし、AI モデルがリクエストに応じてトークンを生成する際、その複雑さとタイプは大きく異なるため、**トークンあたりのコスト**はシステム パフォーマンスと全体的なコスト効率の両方を評価するための重要な指標となります。

ただし、トークンあたりのコストを下げる際には、質の高いユーザー エクスペリエンスを維持することとバランスが取れていなければなりません。ユーザー エクスペリエンスを犠牲にしてリクエスト量を最大化すると、AI を活用したアプリケーションやサービスの導入が減少する可能性があります。劣悪なユーザー エクスペリエンスは顧客離れにつながり、ひいては収益を悪化させます。したがって、トークンあたりのコストとシステム**レイテンシー**の両方を測定することが不可欠です。

レイテンシーは通常、**最初のトークンまでの時間 (TTFT)** と**出力トークンあたりの時間 (TPOT)** という 2 つの次元で測定されます。後で詳しく説明しますが、これらの次元を最適化すると、しばしばトレードオフがもたらされます。一方を向上させると他方が悪化する可能性があるため、組織が適切なバランスを取ることが極めて重要です。

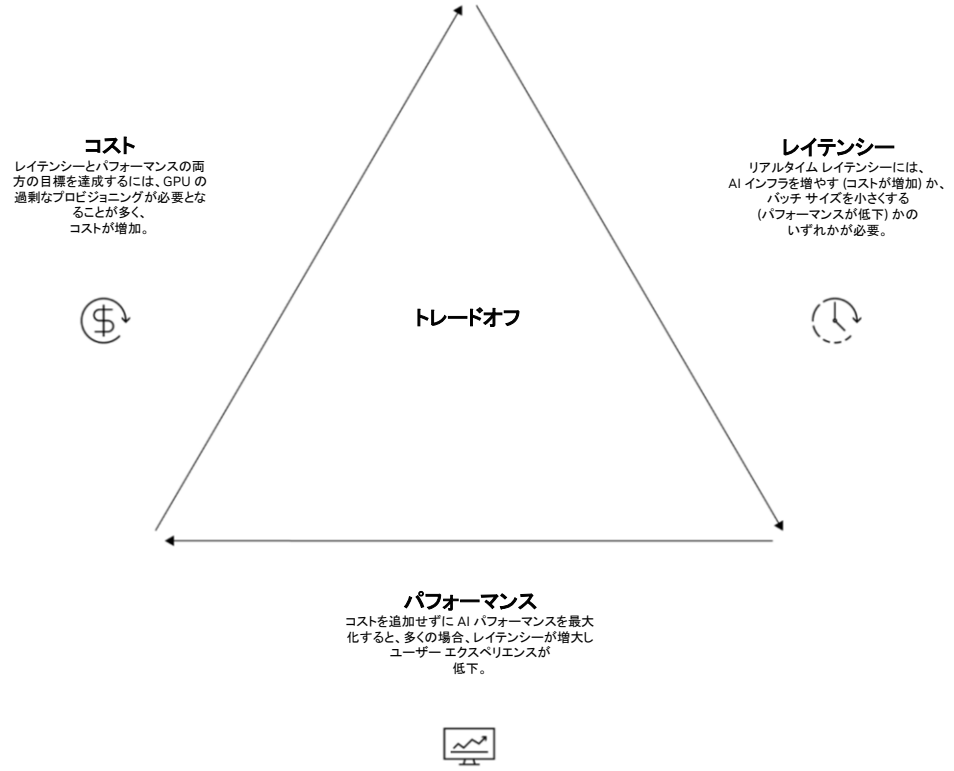

図 1.　最初のトークンまでの時間 (TTFT) とトークン間の時間 (TPOT) を最適化すると、適切にバランスが取れていない場合、3 つのトレードオフが発生することがある

これらの相互依存関係に対応するため、IT リーダーは **Goodput** (TTFT と TPOT の目標レベルを維持しながらシステムが達成するスループット) の測定を開始しています。この指標により、組織はパフォーマンスをより総合的に評価でき、スループット、レイテンシー、コストが運用効率と優れたユーザー体験の両方をサポートするように調整されていることを確認できます。

推論コストを左右する 2 つ目の重要な要因は、市場投入までの時間です。AI 業界は急速に進化しており、市場に先駆けて参入することで大きな競争優位性を得ることができます。しかし、信頼性に欠ける推論ソフトウェア スタック、十分にテストされていないもの、あるいは包括的なドキュメントや強力なコミュニティ サポートが不足しているものは、多大なコストを伴うレイテンシー、統合における課題、そして社内開発チームにとって習得に時間のかかる学習過程を引き起こす可能性があります。

このような複合的な障害が、組織が新たなビジネス チャンスを活かす能力が著しく制限しています。これらの問題が推論コストに与える正確な影響を数値化することは困難な場合もありますが、その影響は確かに存在しており、IT リーダーはこれらの要素を総合的な推論コスト管理戦略全体に取り入れる必要があります。

本書では、AI 推論のパフォーマンスを最適化し、コストを削減し、AI への投資の長期的な価値を最大化するために、これらの要因にどのように対処できるかについて詳しく説明します。

顧客事例：Wealthsimple

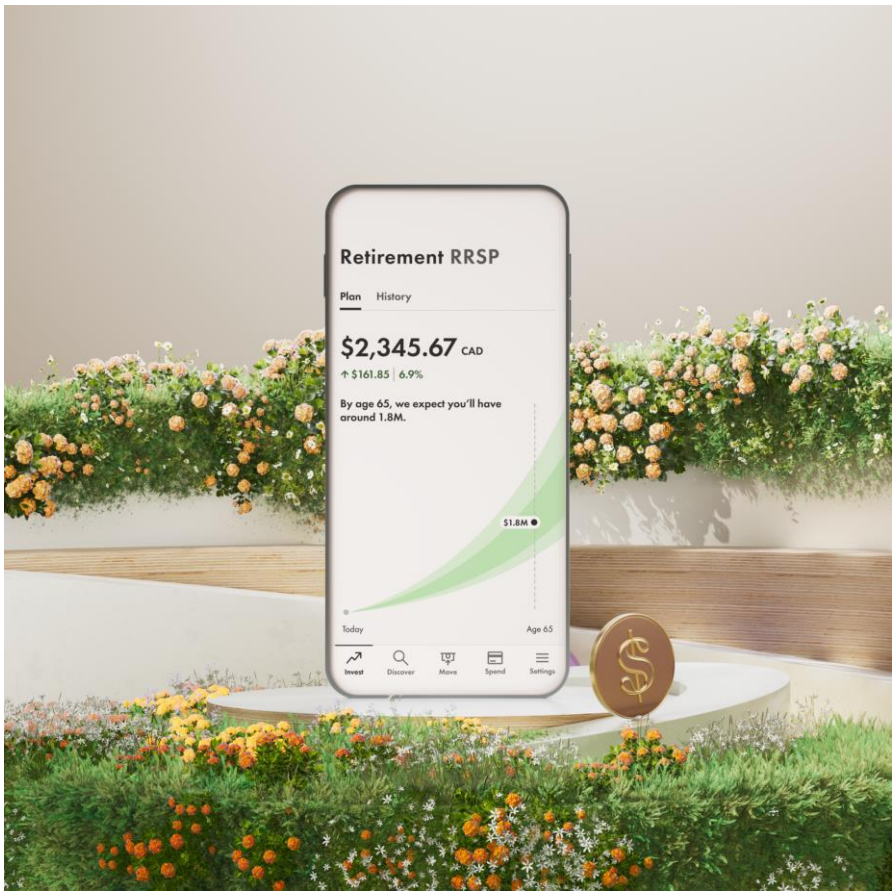

図 2.　Wealthsimple は NVIDIA AI 推論プラットフォームを使用してモデル配信時間を短縮し、全体的な顧客エクスペリエンスを向上させました。

カナダを代表する投資管理会社である Wealthsimple は、顧客エクスペリエンスのさらなる向上を目指し、機械学習 (ML) モデルの導入プロセスの強化に乗り出しました。同社は運用資産 500 億ドル以上を管理していましたが、標準化されていない AI 推論プラットフォームにより、新しいモデルのデプロイにレイテンシーが生じていました。この課題を解決するため、Wealthsimple は NVIDIA の AI 推論プラットフォームを採用し、顕著な成果を上げることに成功しました。

過去 12 か月間で、Wealthsimple は 30 を超える AI モデルをデプロイし、1 億 4,500 万件の予測を生成しました。NVIDIA のプラットフォームを導入したことで、同社のモデルの配信にかかる時間は数か月から 15 分未満へと劇的に短縮され、スピードと効率性が大幅に向上しました。このプラットフォームの信頼性は、過去 1 年半にわたり AI 推論に関連する IT サポートチケットが 1 件も発生していないことからも実証されています。

NVIDIA の Triton Inference Server により、不正検知や取引最適化などの機械学習を活用したパーソナライゼーションを提供する Wealthsimple の能力が強化され、より迅速で正確な金融プロセスが実現しました。また、このプラットフォームによりシステム稼働率が 95% から 99.999% へと向上し、顧客の待ち時間が解消されるとともに電子送金プロセスが効率化されました。

Wealthsimple は、セキュリティと安定性を確保する NVIDIA AI Enterprise のサポートを受けながら、クラウド上で NVIDIA アクセラレーテッド コンピューティングを活用して、モデルを効率的に実行しています。この変革により、Wealthsimple はシステム停止時間を最小限に抑えながら最先端のサービスを提供できるようになり、機械学習搭載の金融サービスを大きく前進させることができました。

主な成果:

- モデルのデプロイにかかる時間を数か月から 15 分に短縮
- 推論システムの稼働率を 95% から 99.999% へ向上
- 推論サービスのレイテンシーを 20% 削減
- IT サポートチケットを発生させずに 1 億 4,500 万件の予測を提供

ユース ケースが推論に与える影響

様々な AI アクセラレーターのパフォーマンスを評価する際には、単一のベンチマークに基づいてハードウェア インフラストラクチャを判断したくなることがあります。しかしながら、こういった単純にとらえ過ぎたアプローチは、時間の経過とともにシステムコストの上昇につながることがよくあります。最初に導入されたシステムが特定のユース ケースのパフォーマンス要求を満たせず、将来的に追加リソースへの投資が必要になる可能性があります。この問題が発生する理由は、異なるユース ケースからの推論ワークロードが、アクセラレーテッド コンピューティング リソースやソフトウェア スタックの最適化に対して多様な要件を持っているからであり、これらの要素は、目標とする総所有コスト (TCO) を実現し、最適なユーザー体験を確保するために極めて重要となります。

チャットボット

例えば、e コマースや顧客サービスのワークフローでよく使用される LLM を活用したチャットボットについて考えてみましょう。これらのチャットボットは、ユーザーからの短い問い合わせや懸念に対して迅速な応答を提供する役割を担っており、ユーザー離れを防ぐためには応答性が非常に重要となります。この場合、TTFT (チャットボットが応答の出力を開始する速さの指標) の高速性がポジティブなユーザー体験を提供するためには不可欠であり、人間の読書速度に匹敵するか若干上回る程度の適度な TPOT が許容範囲と考えられています。一般的なチャットボットの本番環境では、非常に多数のユーザーを対象にサービスを提供しており、複数のユーザーリクエストをまとめて 1 つのリクエストとして、推論処理のために AI システムへと送信します。

要約

その一方で、文書要約のユース ケースも様々な業界で広がりを見せています。例えば、医療分野では研究論文の要約、ニュースやメディア業界では重要なイベントや出来事の要約、Web 会議プロバイダーではビデオ会議からアクション アイテムを生成するといった用途で活用されています。これらのシナリオでは、入力テキストは数ページから数百ページにわたる長さになり、即時応答よりも完全な要約を素早く生成することに重点が置かれます。そのため、システムを高速な TPOT (トークン出力時間) が得られるように最適化する必要があります。これは人間の読書速度をはるかに超えるものであり、特に長い入力テキストの場合、ユーザーはより長い TTFT (最初のトークン生成時間) を許容する傾向にあります。高速な TPOT を実現するため、こうしたケースではバッチサイズが小さくなる傾向があるため、小さなバッチでもスループットを最大化できるように高度なソフトウェア スタックの最適化が必要となります。

質問応答

質問応答ボットはチャットボットと多くの点で類似しており、特に迅速な応答性と短いユーザークエリを処理するという点で似ています。ただし、質問応答ボットは、事前に定義済みのナレッジ ベースやデータセットから回答を生成する必要があるという点で異なります。ユーザーからの質問自体は短くても、LLM に送信される実際の推論リクエストは、知識ベースから取得した関連データが埋め込み形式で含まれるため、かなり長くなります。このワークフローは一般に検索拡張生成 (Retrieval Augmented Generation) または RAG と呼ばれています。このシナリオでは、最適な TCO とユーザー体験を達成するためには、長い入力シーケンス長と短～中程度の出力シーケンス長に対して最適に機能する AI システムとアクセラレーターを選択することが不可欠です。さらに、ナレッジベースデータは通常、GPU メモリ外にありはるかに大容量の CPU メモリに保存されるため、従来の PCIe インターコネクト速度を

超える CPU と GPU 間の高速インターコネクトを備えたシステムをデプロイすることで、ユーザー体験をさらに最適化し、TCO を削減できます。

AI エージェント

現在の AI チャットボットは、生成 AI を活用して単一のやり取りに基づく応答を提供しており、自然言語処理を用いてユーザーからの質問に回答します。AI の次なる進化形はエージェント型 AI であり、これは高度な推論と反復的な計画立案を駆使して複雑な多段階の問題を解決することで、単なる応答を超えた機能を実現します。このタイプの AI は様々な情報源から大量のデータを取り込み、自律的に課題を分析し、戦略を立て、タスクを実行することができます。フィードバック ループを通じて継続的に学習・改善することで、エージェント AI システムは意思決定と業務効率を向上させます。

エージェント AI では、推論プロセス中に追加のトークンを生成するために、関数呼び出しと複数モデルの連携が必要となります。組織内で AI エージェントを活用する IT リーダーにとって、これらの相互作用を効果的に調整するための抽象化ツールやブループリントを提供する適切な推論プラットフォームスタックの選択は、高パフォーマンスな推論を確保するために不可欠です。

これまで示してきたように、様々なユース ケースによって必要となる入力シーケンス長、バッチ サイズ、最初のトークンが出力されるまでの時間 (TTFT)、トークン出力時間 (TPOT)、CPU と GPU 間の接続性はそれぞれ異なります。適切なインスタンスを選択する際には、実際の本番環境のユース ケースに近いベンチマークに注目することが重要です。これにより、一貫したユーザーの品質保証レベルを維持し、ハードウェアのプロビジョニング不足から生じるインフラ コストが徐々に増大するのを防ぐことができます。

顧客事例：Perplexity

AI を搭載した検索エンジンである Perplexity AI は、月間 4 億 3,500 万件以上の検索クエリを処理しており、その各クエリに複数の AI 推論リクエストが伴います。増加する需要に対応しながらコストとユーザー体験のバランスを取るため、Perplexity の推論チームは NVIDIA H100 Tensor Core GPU、Triton Inference Server、および TensorRT-LLM を採用しました。このチームは様々な Llama 3.1 モデルを含む 20 種類以上の AI モデルを同時にサポートしており、ユーザー リクエストを適切なモデルとマッチングさせるにあたり、より小規模な分類器モデルを使用しています。モデムのデプロイは Kubernetes クラスタで管理されており、トラフィックが変動する状況でも品質保証レベル (SLA) が確実に満たされるようにしています。

Perplexity チームは、効果的なスケジューリングとバッチング戦略を通じてパフォーマンスとコストの最適化に専念しています。同社では Triton Inference Server を使って受信リクエストをバッチ処理し、GPU 利用率を最適化するとともに、需要に応じてデプロイの規模を調整しています。チームはトークン間のレイテンシーを最小化するためにスケジューリング アルゴリズムを絶え間なく評価、調整しており、レイテンシーに敏感なタスクのパフォーマンス向上を目指しています。さらに、A/B テストを実施して設定をファインチューニングすることで、モデルが必須となる SLA を満たしながら GPU リソースの利用率を最大化できるようにしています。

NVIDIA GPU を使用して社内にモデルをデプロイすることにより、Perplexity は大幅なコスト削減を実現しています。複数の GPU 全体でのモデルの並列処理と TensorRT-LLM の活用を含む同社のデプロイ戦略により、厳格な SLA を維持しながら、最小限のコストでのモデルの提供を実現しています。

推論パフォーマンスに影響を与えるデプロイ要因

前のセクションでは、推論ソリューションで測定および最適化できる様々なパフォーマンス指標と、これらの指標が特定のユースケースからどのような影響を受けるかについて確認しました。本セクションでは、以下のようなデプロイ要因とアーキテクチャ上の考慮事項がパフォーマンスにどのように影響するかに焦点を当てます。

1. ハードウェア アーキテクチャに関する考慮事項
2. モデル サイズ
3. モデル並行処理による利用率の最大化
4. モデルにおける複数のモダリティ
5. 高度なバッチング技術によるパフォーマンスの最適化

この ebook の目的は、特定のソリューションのデプロイやアーキテクチャ設計に関する詳細なガイダンスを提供することではなく、最適なパフォーマンスを実現するための設計に必要な知識を提供することにあります。

ハードウェア アーキテクチャに関する考慮事項

適切な GPU の選択はパフォーマンスを大きく左右する決定要因です。しかし、定格 FLOPS やメモリ仕様などのピーク チップやインスタンスの指標だけに頼っていては全体像を把握できません。実際のワークロード パフォーマンスは、ソフトウェア スタックの効率など、多くの要因に左右されるからです。同様に、GPU の絶対的な価格やクラウドでの時間当たりのコストは、単体では意味のある指標とはなりません。真に重要なのは、消費電力あたり、または投資額あたりのパフォーマンスです。これは、時間当たりのコストが高い GPU であっても、時間当たりのコストが低い GPU よりも、トークン処理あたりの総コストが低くなる場合があることを意味します。

データ センターの電力上の制約はますます厳しくなっており、限られた電力予算内で推論スループットを最大化することが極めて重要になっています。一定の消費電力に対して得られる推論スループット、すなわちエネルギー効率は、データ センターの推論スループットを最大化し、最終的に収益可能性を高めるための鍵となります。

推論向けに最適化された GPU アーキテクチャ

2012 年にデータ センターに導入されて以来、NVIDIA GPU は並列処理を実現し、多様なリソースを使うタスクの処理時間を劇的に短縮することで業界に変革をもたらしました。この転換により、従来の CPU ベースシステムと比較して、ワット当たり最大 30 倍、ドル当たり 60 倍のパフォーマンス向上が実現しています。

NVIDIA はイノベーションの限界に挑戦し続けており、世代を重ねるごとに画期的なパフォーマンスを実現しています。最新の NVIDIA Blackwell アーキテクチャは前世代と比較して、兆規模のパラメータを持つモデルの推論処理において 30 倍の高速化を実現しています。これは、第二世代トランスフォーマー エンジンの搭載や、より高速で広帯域な NVLINK インターコネクトなどの画期的な技術革新がもたらした結果であり、前世代と比較して桁違いのパフォーマンス向上を達成しています。

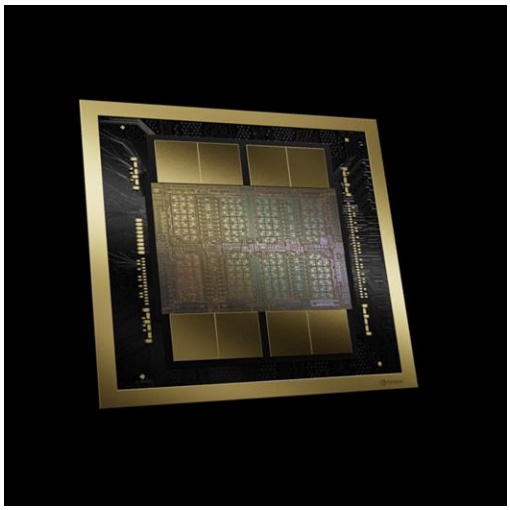

図 3.　比類なき推論パフォーマンスを実現する 2,080 億のトランジスタを搭載した NVIDIA Blackwell GPU

Blackwell は 2,080 億個のトランジスタを搭載しており、前世代の 2.5 倍以上のトランジスタ数を誇る、史上最大の GPU です。第二世代トランスフォーマー エンジンを導入し、カスタム Blackwell Tensor コアと TensorRT-LLM を組み合わせることで、大規模言語モデル (LLM) およびエキスパート混合モデル (MoE) の推論を高速化させます。Blackwell Tensor コアには、計算精度とスループットを飛躍的に向上させる新しい精度形式とマイクロスケーリング技術が搭載されています。Transformer Engine はマイクロテンソル スケーリング技術を採用することでパフォーマンスを強化し、FP4 AI を可能にすることで、FP8 と比較して計算効率、HBM 帯域幅、GPU 当たりのモデルサイズをすべて 2 倍に拡大します。

GPU と省電力 CPU を組み合わせる

NVIDIA GB200 Grace Blackwell Superchip は、NVIDIA® NVLink®-C2C インターコネクトを介して 2 基の高パフォーマンス NVIDIA Blackwell Tensor Core GPU と 1 つの NVIDIA Grace CPU を一体化し、2 つの GPU 間で 900 ギガバイト / 秒 (GB/s) の双方向帯域幅を実現します。

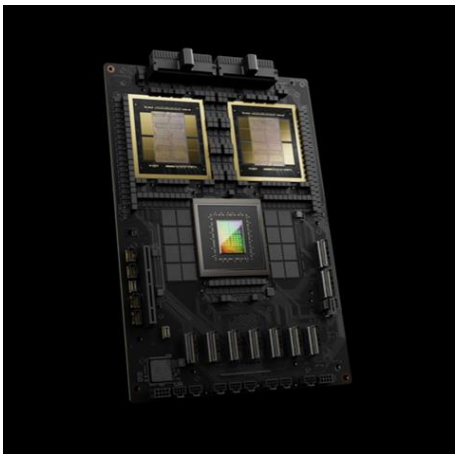

図 4.　NVIDIA GB200 Superchip は 2 基の Blackwell GPU と 1 基の Grace CPU を搭載

NVIDIA Grace CPU には 72 基の高パフォーマンスかつ省電力な Arm Neoverse V2 コアが組み合わされており、NVIDIA Scalable Coherency Fabric (SCF) によって接合されています。NVIDIA SCF は高帯域幅のオンチップ ファブリックであり、従来の CPU の 2 倍となる 3.2 TB/s の総二分バンド幅を備えています。

Grace は高速 LPDDR5X メモリを採用し、エラー訂正コード (ECC) などのメカニズムを通してサーバークラスの信頼性を提供する初のデータ センター CPU です。また、Grace は従来の DDR メモリと同程度のコストでありながら、エネルギー消費量はわずか 1/5 に抑えつつ、最大 500 GB/s のメモリ帯域幅を実現します。

これらの数々の革新により、NVIDIA Grace は画期的なワットパフォーマンス比で、卓越した処理能力、メモリ帯域幅、データ転送能力を提供します。

推論中に 1 つの GPU として複数の GPU が動作

AI のデプロイにおいて適切な GPU の選定は不可欠ですが、エクサスケールのコンピューティングや 1 兆パラメータの AI モデルには、推論時に複数の GPU が単一の巨大 GPU として協調動作するためのシームレスな GPU 間通信が求められます。

NVIDIA GB200 NVL72 は 36 基の GB200 Superchip (36 基の Grace CPU と 72 基の Blackwell GPU) をラック スケール設計で接続します。GB200 NVL72 は液冷方式を採用したラックスケールの 72-GPU NVLink ドメインであり、単一の大規模な GPU として機能します。

NVIDIA GB200 NVL72 は NVIDIA の第 5 世代 NVLink テクノロジーを搭載しており、前世代の 2 倍となるリンクあたり 100GB / 秒のパフォーマンスを達成します。これにより大規模 AI モデル向けにデータ転送の高速化と最適な拡張化を実現します。また、NVIDIA NVLink スイッチを備えており、モデルの並列処理向けに 72-GPU NVLink ドメイン (NVL72) 内で 130TB / 秒の GPU 帯域幅を実現します。NVLink と NVLink スイッチを組み合わせた場合、1.8TB / 秒の相互接続を持つマルチサーバー クラスタをサポートできるようになり、GPU 通信と演算能力が拡張されます。

図 5.　NVIDIA GB200 NVL72

GB200 NVL72 は、GPT-MoE- 1.8T のような超大規模モデルにおいて、同じ GPU 数で前世代比で推論速度を 30 倍に向上し、TCO を 25 倍削減し、省エネルギー化を 25 倍向上させます。

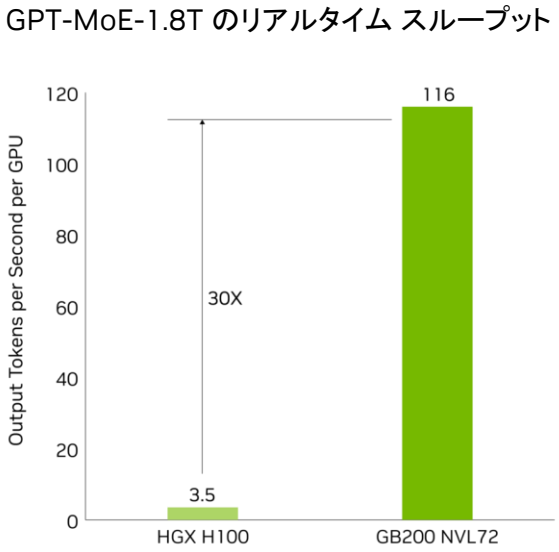

図 6.　GB200 は H100 と比較して 30 倍のリアルタイム スループットを実現

モデル サイズ

LLM をワークロードにデプロイする際、モデルのパラメータ数、すなわちモデル サイズは重要な検討事項となります。最先端 LLM のパラメータ数は時間の経過とともに増加の一途をたどり、モデル能力と応答品質に飛躍的向上をもたらしています。

様々なユース ケース、デプロイ シナリオ、予算に対応するため、モデル開発者は多様なサイズのバリエーションを用意するのが一般的です。例えば、オープン モデルの Llama 3.1 ファミリーでは、8B、70B、405B パラメータのモデルがラインナップされています。通常、同一モデル ファミリー内では、より大規模なモデルほど高精度な応答を生成します。ただし、大規模なモデルにはトークン生成において、小規模なモデルよりも多くの計算リソースが必要となることに注意する必要があります。

つまり、モデル サイズを選択する際には、想定されるユース ケースと利用可能な計算リソースの両面から判断する必要があります。複雑なタスクに対して最高水準の応答精度が求められる場合には、大規模なモデルが最適な選択となるでしょう。一方、出力トークンの生成速度が重視される場合や、計算リソースに制約がある環境では、小規模なモデルを使用した方が有利となります。

モデル並行実行による利用効率の最大化

データセンター向け GPU の進化に伴い、各世代で提供される計算能力とメモリ リソースは急速に拡大しています。これらの次世代 GPU は、1 兆パラメータ モデルの処理など、大規模な推論の高速化に極めて効果的です。しかし、この強力なパフォーマンスには注意点もあります。小規模なモデルを処理する場合、大量の計算リソースとメモリ リソースが十分に活用されないことが多々あり、総所有コスト (TCO) の上昇を招くことがあります。

このようなシナリオでは、計算能力とメモリの大部分が未使用のまま残るため、GPU リソースが効率的に使用されていません。その結果、不要なインフラ コストを生じさせ、デプロイされたリソースと実際のワークロード要件の間のバランスが取れなくなります。

この課題を解決する優れたソリューションとなるのがモデルの並行実行です。単一 GPU 上または 1 つのノード内の複数 GPU 上で、複数のモデルあるいは同一モデルの複数インスタンスを同時実行することで、インフラに追加投資することなくシステム スループットを向上させ、レイテンシーを低減することができます。このアプローチにより、組織は既存の GPU リソースを最大限に活用し、計算能力をフル活用してインフラ コストの最適化を実現できます。

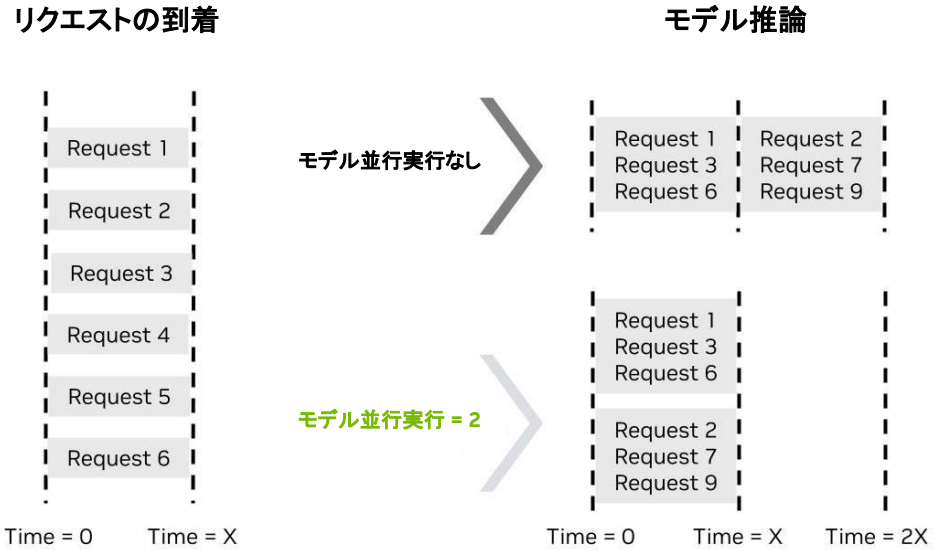

図 7.　モデル並行性の向上がスループットとレイテンシーに与える影響

NVIDIA Triton Inference Server は、並行実行を有効化しているモデルのデプロイ プロセスを効率化します。開発者はモデルの設定ファイル内にある単一のフラグを有効化するだけで、簡単にモデルの並行実行を実装できます。その後、デプロイする並行モデル インスタンスの数と使用するターゲット GPU を指定することができます。この容易さを利用して、開発者は複雑な設定や手動による拡張作業なしで GPU インフラの能力を最大限に活用できます。

複数の独立したモデルを同時に提供する必要があるが、すべてを同じインスタンスやノードに同時配置できない場合、NVIDIA Triton はマルチプレクサとして機能します。ユーザー要求に応じてモデルをダイナミックにロード・アンロードすることで、必要なモデルが適切なタイミングで利用可能な状態を確保します。これにより、組織はインフラに追加で投資することなく、高いサービス可用性とパフォーマンスを維持できます。

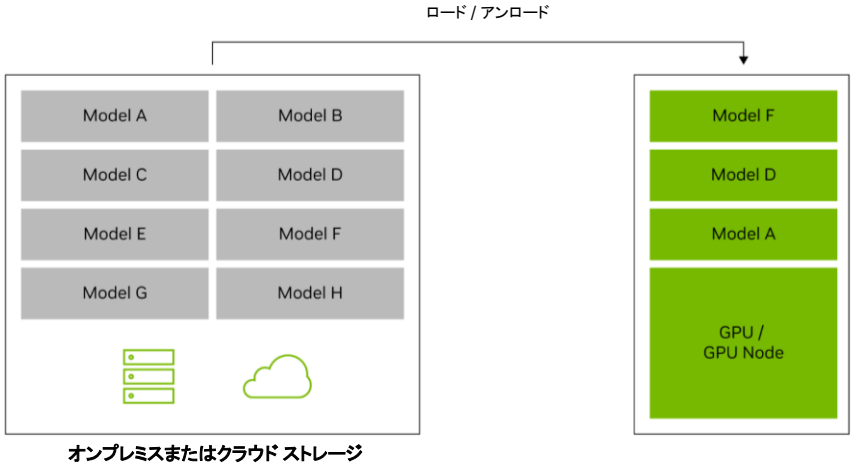

図 8.　NVIDIA Triton の GPU 上のモデルロード・アンロード機能で
サービス可用性とパフォーマンスが向上

スループットの向上とコスト削減に加え、モデルを並行実行することで、ユーザーのサービスに対する需要が事前に予測困難な状況にも対応できます。IT チームはモデルの単一インスタンスから始め、需要の増加に応じてモデルの複数並行インスタンスを段階的にデプロイすることでスケールアップできます。並行モデル数を調整できるため、未使用の計算リソースにかかるコストを最小限に抑えながら変動する需要に対応することができます。

モデル並行実行は、特にデータセンター GPU 上に小～中規模モデルをデプロイする際に強力な推論パフォーマンスツールとなります。このアプローチを活用することで、組織は以下の顕著な利点を得られます。

- **スループットの向上**：複数のモデルやモデル インスタンスを並行実行することで、システム全体のスループットが向上し、より多くの推論リクエストを同時処理できるようになります。
- **トークンあたりコストの低減**：既存のリソースを最大限に活用することで、処理トークンあたりのコストが削減され、インフラ投資効果が最大化されます。
- **レイテンシーの短縮**：モデルやモデル インスタンスの並行実行により待機時間が短縮され、サービスの応答性が向上し、ユーザー体験が改善されます。

モデルにおける複数のモダリティ

テキスト、画像、音声、動画などの複数のデータ モダリティからの情報を同時に処理、統合できる AI モデルは、マルチモーダル モデルと呼ばれています。単一の入力モダリティのみを扱う従来の単一モーダル AI モデルとは異なり、マルチモーダル モデルでは、多様なモダリティにまたがる入出力処理を実現するため、より複雑な推論サービス要件が求められます。

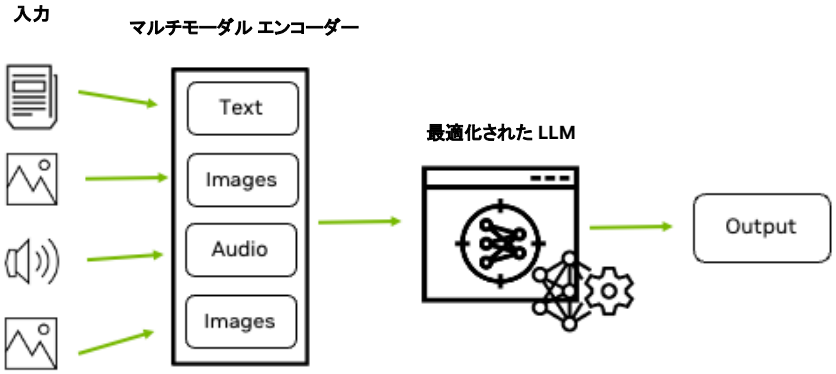

図 9.　複数モダリティの情報を融合し高スループット出力向けに
最適化するマルチモーダル LLM

NVIDIA AI 推論プラットフォームは、NVIDIA TensorRT ライブラリを用いた高パフォーマンス音声・画像入力エンコーダーと TensorRT-LLM ライブラリで最適化されたテキスト デコーダーによる最適化されたサポートをマルチモーダル モデルに提供します。これらのモデルは、NVIDIA Triton Inference Server を用いてデプロイでき、単一リクエストに複数の画像を含めることができるマルチ イメージ推論や、画像を複数のテキスト入力トークン間で共有することで KV キャッシュ再利用の長さに比例してレイテンシーを削減しパフォーマンスを向上させることができる効果的な KV キャッシュ再利用など、複数の高度な機能をサポートします。

高度なバッチ処理技術によるパフォーマンスの最適化

NVIDIA GPU などのデータセンター AI アクセラレーターは、複数のコアを活用してリクエストを並列処理します。これらのコアの潜在能力を最大限に引き出すため、リクエストは通常バッチとしてまとめられ、推論処理に送られます。しかし、バッチ サイズ、スループット、レイテンシーの間では、重要なトレードオフが発生します。これら 3 要素のうち 2 つを調整すると、残りの 1 つに悪影響を及ぼす可能性があり、最適なパフォーマンスを求める企業にとって課題となります。

例えば、小さなバッチ サイズにするとレイテンシーを低減できますが、GPU リソースの活用度が低下し、非効率性と運用コスト増加につながります。一方、大きなバッチにすると GPU リソースの使用効率を最大化できますが、レイテンシーが増加するという代償を伴い、ユーザー体験を損なう恐れがあります。

IT チームはユース ケースに理想的なバランスを見出すために A/B テストを実施することがよくありますが、これは複雑で時間を要するプロセスです。NVIDIA Triton Inference Server や NVIDIA TensorRT-LLM などの高度な推論ソフトウェア プラットフォームを活用することで、このトレードオフを緩和できます。NVIDIA AI 推論プラットフォームは精緻なバッチング技術をデプロイするため、組織がスループットとレイテンシーの間のバランスを良好に保ち、システム全体のパフォーマンスを向上させるのに役立ちます。

ダイナミック バッチング

ダイナミック バッチ処理では、最大バッチ サイズや推奨バッチ サイズ、最大待機時間などの特定基準に基づいてバッチを生成します。推奨サイズでバッチを形成できる場合はそのサイズで処理され、そうでない場合はシステムが設定された最大バッチサイズの範囲内で可能な限り大きなバッチを構成します。ダイナミック バッチングでは、リクエストが短期間キューに滞留できるようにし、追加リクエストが到着してより大きなバッチサイズを作成するまで待機します。これにより、スループットとリソース利用率が向上され、最終的にトークンあたりのコストが削減されます。

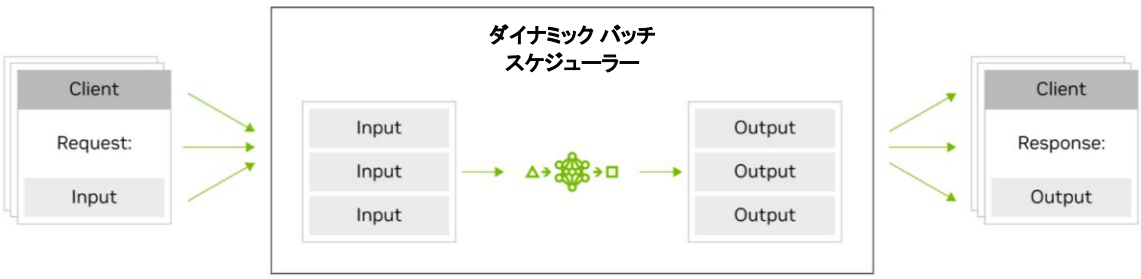

図 10. ダイナミック バッチングが複数リクエストのバッチを生成するメカニズム

シーケンス バッチング

動画ストリーミングなど、特定のリクエスト シーケンスが必要となるユース ケースでは、シーケンス バッチングで関連リクエスト (動画のフレームなど) が意味のある順序で処理されるようにすることができます。動画ストリームのように、モデルがフレーム間で状態を維持する必要があるような相関リクエストでは、シーケンス バッチングにより同一インスタンス上でリクエストが順次処理されるようにすることができます。これにより、パフォーマンスの最適化を図りながら正確な出力が保証されます。

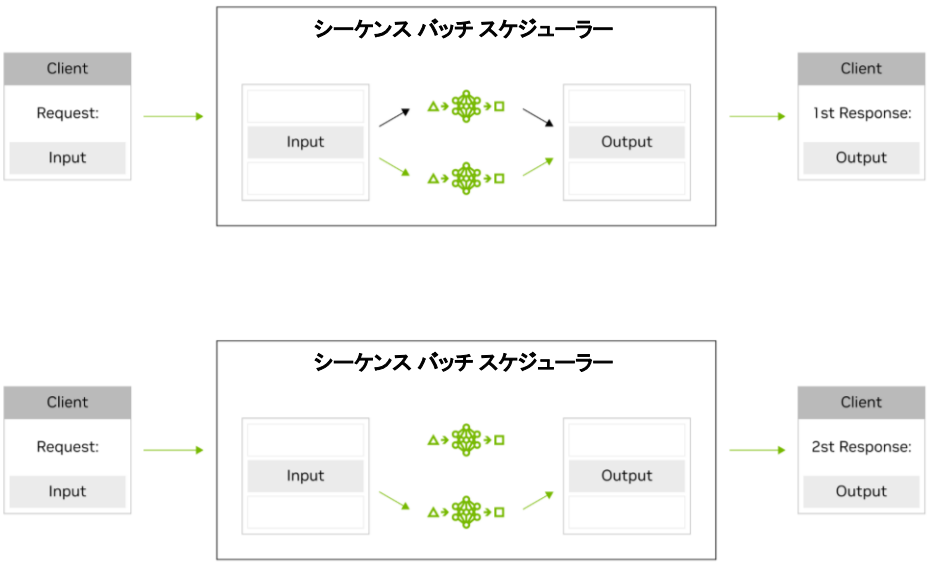

図 11. シーケンスバッチングによる、リクエストの関係性と順序を維持した異種リクエストのバッチ処理法

インフライト バッチング

インフライト バッチングは、連続的なリクエスト処理を可能にして従来のバッチ処理を強化します。インフライト バッチ処理では、TensorRT-LLM ランタイム (推論向け LLM を最適化する NVIDIA SDK) がバッチ全体の完了を待たずに、準備ができたリクエストを即時処理します。これにより、他のリクエストが処理中でも新しいリクエストが即座に開始され、処理速度とスループットが向上します。

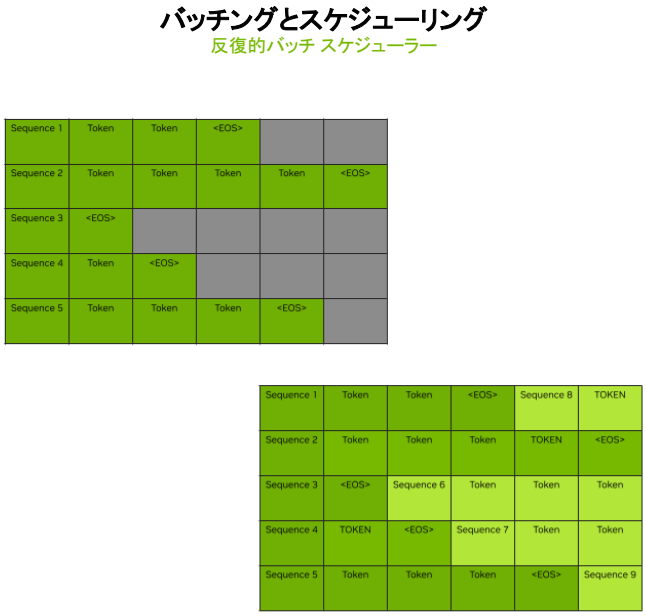

図 12.　インフライト バッチングが完了したバッチ内のリクエストを排除し、新しいリクエストへと置き換メカニズム

計算リソースを効率的に活用するには、バッチ サイズ、スループット、レイテンシーの慎重な管理が不可欠です。NVIDIA AI 推論プラットフォームが提供するダイナミック バッチング、シーケンス バッチング、インフライト バッチングなどの高度なバッチ処理技術により、組織はパフォーマンスの最適化、コスト削減、ユーザー体験の向上を柔軟に実現できます。これらの技術はスループットとレイテンシーの間のバランスを取るのに役立ち、企業が AI を活用したソリューションをより効果的に拡張できるよう支援します。

大規模モデルの推論並列化：主要な手法とトレードオフ

大規模言語モデル (LLM) のサイズと複雑さが増大するにつれ、複数 GPU 間でこれらのモデルを確実かつ効率的にデプロイすることが極めて重要になっています。モデルが単一 GPU のメモリ容量を超える場合、ワークロードを分散しパフォーマンスを最適化するために並列化技術が採用されます。経営幹部や IT リーダーにとって、主要な並列処理手法とそれらがスループットやユーザーの対話性に与える影響を理解することは、的確なインフラ決定を行う上で不可欠です。以下は、大規模モデルの推論並列化における主要な手法と、主なケース スタディからの例です。

データ並列処理 (DP)

データ並列処理では、モデルを複数の GPU または GPU クラスタに複製します。各 GPU はユーザーリクエストのグループを個別に処理し、これらのリクエストグループ間で通信が不要となる仕組みです。この手法は使用する GPU 数に比例して直線的に縮小、拡大するため、割り当てられる GPU リソースに応じて処理可能なリクエスト数も直接的に増加します。ただし、最新の大規模 LLM ではモデルの重みが単一 GPU に収まりきらないことが多いため、データ並列処理だけでは通常不十分である点に留意が必要です。そのため、DP は他の並列化技術と併用されるのが一般的です。

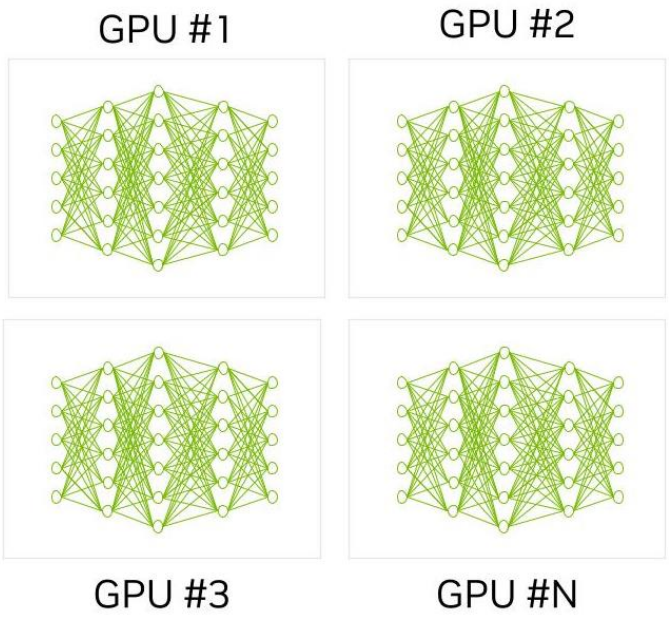

図 13. 複数の GPU 間にデータ並列処理を適用した場合

パフォーマンスへの影響:

- **GPU スループット**：モデルが複製されるため、DP 単独では影響を受けません。
- **ユーザー対話性**：リクエストは個別に処理されるため、大きな影響はありません。

テンソル並列処理 (TP)

テンソル並列処理では、モデル パラメータが複数の GPU に分割され、ユーザー リクエストは GPU または GPU クラスタ間で共有されます。異なる GPU 上で実行された計算結果は、GPU 間ネットワークを介して統合されます。この手法では各リクエストにより多くの GPU リソースが割り当てられ、処理時間が短縮されることから、特に Llama 405B や GPT4 1.8T MoE (1 兆 8,000 億パラメータのエキスパート混合モデル) などのトランスフォーマー ベース モデルにおいて、ユーザー対話性を向上させることができます。

パフォーマンスへの影響:

- **GPU スループット**：NVIDIA NVLINK などの高速相互接続なしに、TP を多数の GPU へと拡張すると、通信オーバーヘッドの増加によりスループットが低下する可能性があります。
- **ユーザー対話性**：各リクエストに割り当てられる GPU リソースが増えるため、処理が高速化し、向上します。

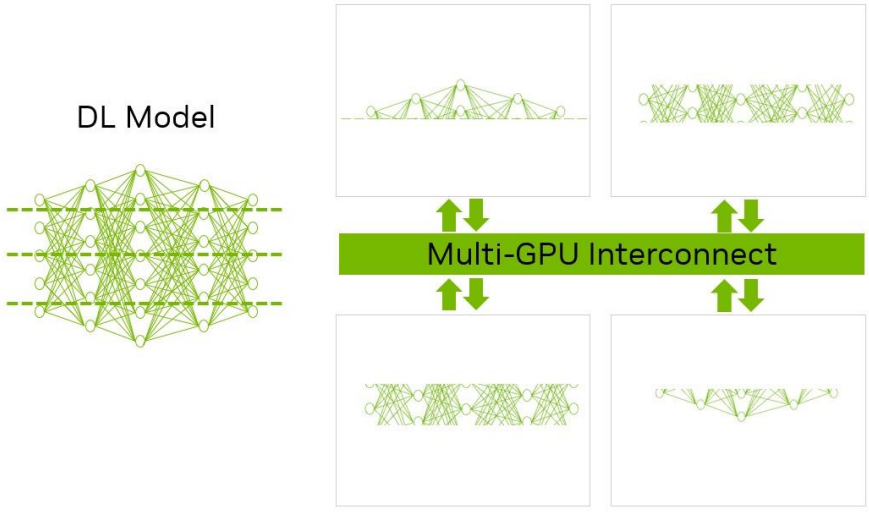

図 14. ディープ ニューラル ネットワークにテンソル並列処理を適用した場合

パイプライン並列処理 (PP)

パイプライン並列処理ではモデル層を分割し、各層のグループが異なる GPU に割り当てられます。モデルは GPU 全体でリクエストを順次処理し、各 GPU はモデルの割り当てられた部分に対して計算を実行します。この手法は単一 GPU に収まらない大規模モデルの重みを分散するのに有効ですが、効率性という面で制限があります。

パフォーマンスへの影響:

- **GPU スループット**：モデルの重みが分散されるため、効率低下を招く可能性があります。

- **ユーザー対話性**：処理が GPU 全体で順次進行する必要があるため、ユーザー対話性の大幅な最適化には至りません。

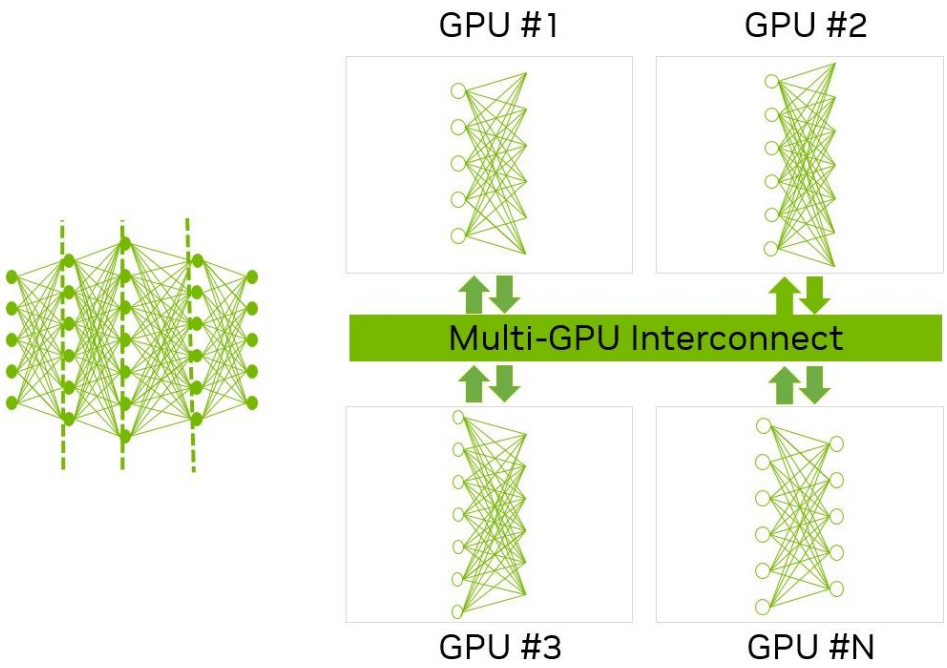

図 15. ディープ ニューラル ネットワークにパイプライン並列処理を活用した場合

エキスパート並列処理 (EP)

エキスパート並列処理では、ユーザー リクエストがモデル内に存在する特化した「エキスパート」へと振り分けられます。各リクエストを限られたモデル パラメータ群 (特定のエキスパート) に制限することで、システムの計算オーバーヘッドが削減し処理が最適化されます。リクエストは個別のエキスパートによって処理された後、最終出力段階で再結合されますが、これには高帯域幅の GPU 間通信が必要となります。

パフォーマンスへの影響：

- **GPU スループット**：エキスパート並列処理は、**特に他の技術と組み合わせたときに、スループット**に大幅な向上をもたらします。

- **ユーザー対話性**：EP 構成は他の手法で見られる大幅なトレードオフなしに、ユーザー対話性を維持または向上させることができます。

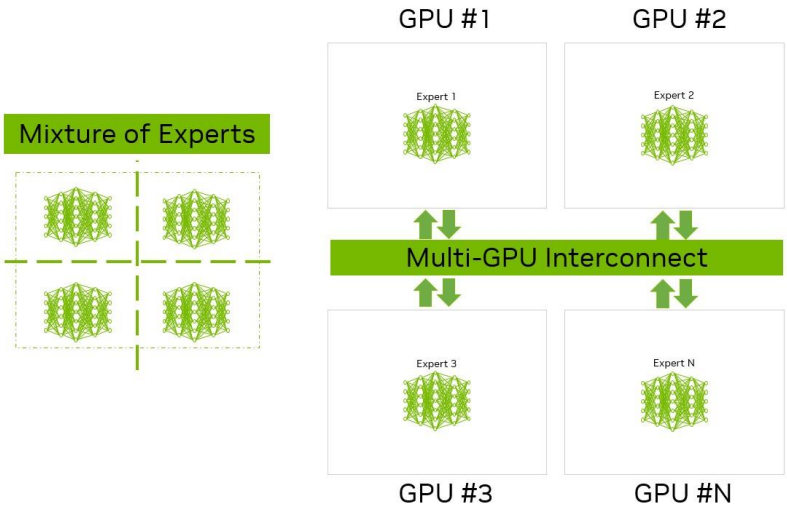

図 16. 4 つのエキスパートで構成されるディープ ニューラル ネットワークに
エキスパート並列処理を適用した場合

最適なパフォーマンスのための並列処理手法の組み合わせ

複数 GPU にわたる LLM のパフォーマンスを最適化する最も効果的な方法は、複数の並列処理技術を組み合わせることです。これにより、スループットとユーザー対話性の間に生じるトレードオフを最小化し、高パフォーマンスと優れた応答性の両方を実現したユーザー体験が得られます。

例えば、16 個のエキスパートを持つ GPT 1.8T MoE モデルを、それぞれ 192GB のメモリを持つ 64 個の GPU で使用する場合、エキスパート＋パイプライン並列処理 (EP16PP4) の組み合わせでは、エキスパートのみの並列処理を使用した場合と比較して GPU スループットをほとんど低下させることなくユーザー対話性が 2 倍向上し、テンソル＋エキスパート＋パイプライン並列処理 (TP4EP4PP4) を使用した場合には、ユーザー対話性を維持しながらテンソルのみの並列処理を使用した場合と比較して GPU スループットが 3 倍に向上しました。

LLM のデプロイを監督する CIO や IT リーダーにとって、様々な並列化戦略のトレードオフと利点を理解することは非常に重要です。データ、テンソル、パイプライン、エキスパート並列処理などの並列処理手法を組み合わせることで、組織は GPU リソースの利用を最適化し、スループットとユーザー対話性の両方を向上させることができます。実際に実行する際には、構成を十分に計画することで、複数の GPU 全体で高パフォーマンスと応答性の良いユーザー体験のバランスを取りながら、最適な結果を得ることができます。より大規模で複雑なモデルへの需要が増加するにつれ、これらの並列化技術は拡張可能で効率的な AI モデルのデプロイを確保する上で重要な役割を果たします。

クラウドにおける AI 推論パフォーマンスとコスト効率の最大化

クラウドは長らく企業にとって変革をもたらすものとして、拡張可能で効率的、かつグローバルにアクセス可能な IT インフラストラクチャを実現してきました。企業がストレージ、データベース、リソース計画などのワークロードを柔軟性とコスト削減のためにクラウドに移行したように、AI 推論も同じ道をたどっています。IDC の最新レポートによると、企業に採用されている AI 推論で最も急成長を遂げている分野はクラウド上のアクセラレーテッド コンピューティング インスタンスであり、その採用率は他のすべてのデプロイモデルを 3 倍のペースで上回ると予測されています。

IT リーダーにとって、この変化はコストを増加させることなくパフォーマンスを向上させる重要な機会となります。クラウドベースのアクセラレーテッド コンピューティングと最適化された AI ソフトウェアを組み合わせることで、組織はより高速で、拡張性があり、コスト効率の良いソリューションを実現できます。クラウドでの AI 推論パフォーマンスを最適化するためには、IT リーダーは自社のチームが多様な GPU タイプから柔軟に選択して、各ワークロードに適した GPU を割り当てることができるようにする必要があります。このような戦略的な選択は、アクセラレーテッド コンピューティング リソースの過剰供給や過少供給を回避し、クラウドにおける AI 推論の可能性を最大限に引き出す鍵となります。

また、IT チームが既にトレーニングおよび認定を受けているクラウド ツールと深く統合する AI 推論スタックを選択することも同じ様に重要です。この統合によりトレーニング コストを最低限に抑えることができ、チームは複雑なソフトウェア統合作業ではなく、推論パフォーマンスの最適化に集中できるようになります。柔軟性をさらに高めベンダー ロックインを回避するためには、IT リーダーは主要なクラウドサービスプロバイダー (CSP) 全体で利用可能な GPU と推論スタックを優先する必要があります。これによりエンド ユーザーから近い場所に GPU をデプロイすることが可能になり、AI 推論チームは特定のアプリケーション ニーズに最も適したクラウド プラットフォームを自由に選択できます。

クラウドで NVIDIA のアクセラレーテッド コンピューティング ソリューションを利用することの大きな利点の一つは、得られる柔軟性と選択肢の豊富さです。様々な世代の NVIDIA GPU が、主要なクラウド サービス プロバイダーのすべてでクラウド コンピューティング インスタンスを支えており、企業や開発者がニーズに最適なアクセラレーテッド コンピューティング ソリューションを選択できるように幅広い選択肢を提供しています。

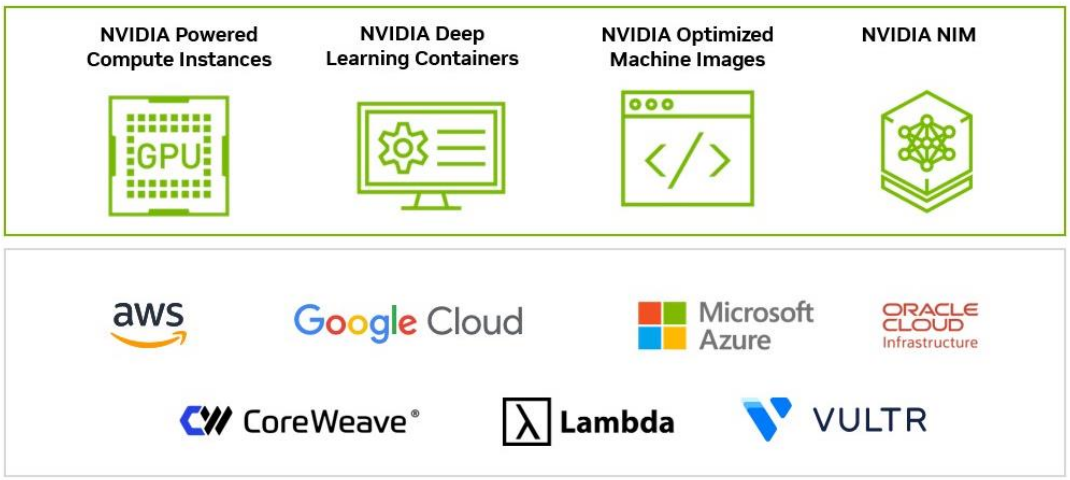

図 17. クラウドにおける NVIDIA AI 推論プラットフォームの広範なサポート

さらに、NVIDIA の AI 推論スタックは主要なすべてのクラウド MLOps ツールと完全に統合されており、AI ワークロードのデプロイと拡張を簡素化します。例えば、NVIDIA Triton Inference Server と NVIDIA NIM は主要なすべてのクラウド AI ツールに組み込まれており、IT チームは最小限の労力でモデルをデプロイできます。AWS SageMaker、Google Vertex AI、Azure ML、OCI の Data Science Cloud のいずれを使用する場合でも、企業は最低限のコーディングまたは一切コーディングすることなく、NVIDIA Triton と NVIDIA NIM を起動することができるため、時間と IT リソースを節約できます。

より合理的なアプローチを求める組織向けに、NVIDIA では必要となるドライバーと依存関係を事前にすべてインストールして構成済みのクラウド マシン イメージも提供しています。これにより手動設定なしで AI ワークロードを迅速にデプロイでき、企業はセットアップに要する時間とコストの両方を削減できます。

NVIDIA のアクセラレーテッド コンピューティングと統合済みの AI 推論ソリューションを活用することで、IT リーダーは柔軟性とコスト効率を確保しながら、最小限に抑えられたオーバーヘッドで、AI 推論のパフォーマンスをクラウド上に効果的に拡張することができます。

顧客事例：Amdocs

通信メディア プロバイダー向けソフトウェアとサービスの主要プロバイダーである Amdocs は、通信業界向けに特化した生成 AI プラットフォーム「amAIz」を開発しました。パフォーマンスの向上とデプロイの高速化を実現するため、Amdocs は NVIDIA DGX Cloud 上で NVIDIA NIM 推論マイクロサービスを使用し、リアルタイムの顧客ケアと販売問い合わせを提供するために細かく調整された LLM をシームレスにデプロイしました。

Amdocs は匿名化された通話記録、顧客のインタラクション、請求書から人間が注釈付けしたデータセットを厳選し、Meta の Llama モデル シリーズや Mistral AI の Mixtral 8x7B などのパラメータ効率に優れたファインチューニング技術である Low-Rank Adaptation (LoRA) に使用しました。チームは

NIM を使用してファインチューニング モデルのデプロイを自動化し、レイテンシーを 80% 削減してほぼリアルタイムの応答を可能にすると同時に、精度を 30% 向上させました。ファインチューニングを通してドメインもカスタマイズすることで、トークン消費量も 40% 削減させ、インフラコストを低減しています。

DGX Cloud 上の NVIDIA NIM を活用することで、Amdocs は LLM の精度と応答時間を大幅に改善しながらコストを削減し、amAIz プラットフォームが現代の通信アプリケーションの要求に応えられるようにしました。

変動する需要に対応するように AI 推論を拡張

IT リーダーが AI アプリケーションを展開する際に直面する最も困難な判断の一つは、ユーザー需要を予測し、その需要が時間の経過とともにどのように変動するかを理解することです。これらの予測はインフラ計画、特に GPU などのリソース配分に大きく影響し、結果的にコストとパフォーマンスの両面を左右します。これらの要素のバランスを取ることは、AI システムが余分なコストを抑えつつ、需要に応じて柔軟に拡張できるようにするための重要なポイントです。

AI アプリケーション、特に最適化された LLM を活用するアプリケーションには、リアルタイム推論のために相当な計算能力が必要です。これらのアプリケーションを拡張するためによくあるアプローチは、予測されるピーク需要に基づいて GPU をプロビジョニングすることです。しかし、これはピーク需要が一時的または不規則な場合には特に、非効率的になる可能性があります。

この課題に対処するため、組織は変動する需要に応じてインフラストラクチャを拡大または縮小できる柔軟性を求めています。Kubernetes は理想的なソリューションであり、IT チームは LLM のデプロイをダイナミックに縮小・拡張できます。単一 GPU から複数 GPU に拡張するのであれ、オフピーク時間帯のリソースを削減するのであれ、Kubernetes はインフラストラクチャ管理に対してコスト効率の良い、パフォーマンスが最適化されたアプローチを提供します。

Kubernetes と NVIDIA AI 推論プラットフォームによる柔軟性

企業、特に e コマースや顧客サービスなどの業界では、常に変動する推論リクエスト量に直面しています。セール期間中の大量の顧客問い合わせ処理や深夜の少ないリクエスト管理など、効率的に縮小、拡張できる能力が非常に重要です。Kubernetes で拡張することで、組織は必要な GPU 数をダイナミックに調整し、ハードウェア リソースがリアルタイムの需要と一致するようにできます。このアプローチは、ピーク ワークロードに対応するためにハードウェア リソースを過剰にプロビジョニングする場合と比較して、大幅なコスト削減をもたらします。

NVIDIA Triton や NVIDIA NIM を含む NVIDIA AI 推論プラットフォームは、縮小、拡張プロセスを容易にするために Kubernetes をサポートしています。推論リクエストが増加すると、Triton メトリクスを Prometheus が収集して Kubernetes Horizontal Pod Autoscaler (HPA) に送信され、各ポッドに 1 つ以上の GPU を持つ追加ポッドがデプロイに追加されます。Prometheus を使用して Triton から収集し、拡張に対する決定を通知するために Kubernetes HPA に送信できるカスタム メトリックの一例が、**キュー対計算比率**です。この比率は推論リクエストの応答時間を反映しています。これは推論リクエストのキュー時間を計算時間で割ったものとして定義されます。

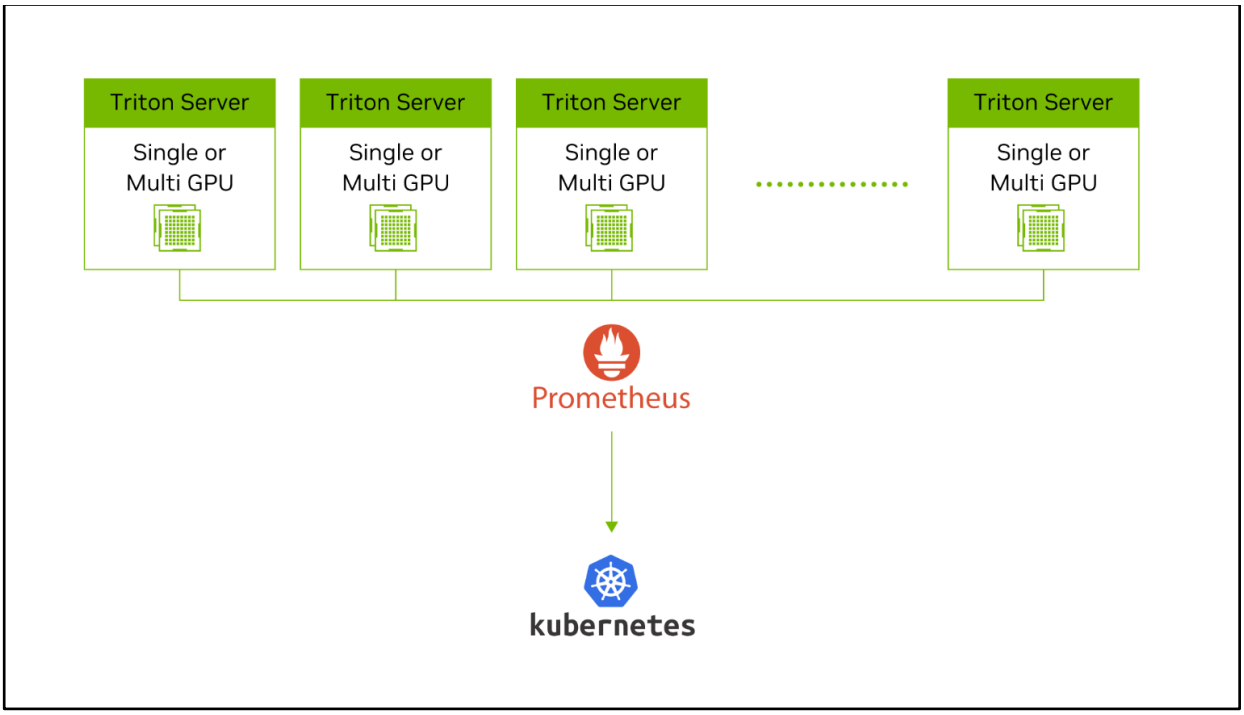

図 18. Kubernetes アーキテクチャによる NVIDIA Triton Server の拡張

拡張可能な AI デプロイの主な利点:

1. **コスト効率**:需要に基づいて拡張することで、企業はインフラ コストを最適化し、リアルタイム リクエストを処理するために必要なハードウェアのみをプロビジョニングできるようになります。このダイナミックな拡張性により、高コストになる可能性のある過剰プロビジョニングのニーズが軽減されます。

2. **パフォーマンスの最適化**:GPU リソースをダイナミックにバランシングすることで、Kubernetes と NVIDIA Triton は高負荷期間中でも高いパフォーマンスを維持します。これにより、AI アプリケーションは顧客対応環境で不可欠となる、厳格なレイテンシーと精度の要件を満たすことができます。

3. **業界全体での柔軟性**:オンライン ショッピング イベント中のトラフィック急増への対応から、コール センターの変動するリクエスト量の管理まで、拡張可能な AI をデプロイすることで、幅広い業界のニーズに対応できる柔軟性が得られます。

NVIDIA Triton と Kubernetes を使用した AI デプロイの最適化の詳細については、「Kubernetes を使用した NVIDIA Triton と NVIDIA TensorRT-LLM による LLM の拡張」をご覧ください。NVIDIA NIM と Kubernetes を使用した AI デプロイの最適化の詳細については、「Kubernetes 上の NVIDIA NIM の自動拡張」および「NVIDIA NIM Operator による Kubernetes 上の AI 推論パイプラインの管理」をご覧ください。

モデル アンサンブルによる AI パイプラインの効率化

AI および ML システムがスタンドアロン モデルとしてデプロイされることはほとんどありません。様々なモデル、前処理ステップ、および後処理タスクを統合するための、より広範で複雑なパイプラインの一部として機能することが一般的です。モデル アンサンブルとして知られるこのアプローチは、特に企業アプリケーションにおいて、現代の AI ワークフローの基盤となっています。

モデル アンサンブルとは、複数の機械学習モデル、前処理および後処理ステップを統一されたパイプラインに統合することを指します。アンサンブルでは、モデルを個別に実行する代わりに、これらのコンポーネントが連携して動作するようにして、プロセスを効率化し全体的な効率を高めます。アンサンブル内の各モデルは通常、データ処理、予測生成、出力の精緻化など、異なるタスク面を担います。

実際のシナリオ、特に企業アプリケーションでは、AI パイプラインには複数のコンピューター ビジョン モデル、画像編集ツール、およびその他の専用モデルが含まれることがあり、すべてが連携して最終結果を生成します。例えば、SDXL (生成 AI モデル) を使用してマーケティング キャンペーンを作成する場合、ワークフローでは、初めにズームイン前処理ステップ、次に SDXL を使用したシーン生成、そして最後に高解像度表示向けの画像のアップスケーリングなどの後処理ステップが必要となる可能性があります。

これらの個々のステップを一貫したパイプラインに結合することで、企業はレイテンシーを削減しリソース使用を最適化しながら、モデルを通じてデータのスムーズな流れを確保できます。

モデル アンサンブルにおける NVIDIA Triton Inference Server の役割

こうした複雑なモデル パイプラインの構築と管理は容易ではありません。しかし、Triton Inference Server などの NVIDIA AI 推論プラットフォームツールは、プロセスを自動化し、モデル アンサンブルを調整するための高度な機能を提供する強力なソリューションとなります。

Triton のモデル アンサンブル機能により、各ステップを管理するためのコードを手動で記述する必要がなくなり、複雑さが軽減されミスのリスクも最小化されます。さらに、Triton では前処理と後処理を CPU 上で実行しながら、中核となる AI モデルを GPU 上で動作させることができ、処理能力をバランス配分する際に柔軟性が得られます。加えて、Triton では条件付きロジックやループなどの高度な機能をアンサンブルに組み込むことができ、開発者はより洗練された、柔軟な AI ワークフローを構築できます。

Triton を活用したモデル アンサンブルの利点

モデル アンサンブルを活用すると、特に Triton と統合した場合に、いくつかの重要な利点が得られます。

1. **レイテンシーの低減**:前処理、推論、後処理を単一のパイプラインに統合することで、企業はモデル間のデータ移動にかかる時間を大幅に削減できます。このレイテンシーの削減は、生成 AI やコンピューター ビジョンなどのリアルタイム AI アプリケーションにとって極めて重要です。

2. **リソース利用効率の向上**:モデルアン サンブルにより、チームはリソースのデプロイを最適化できます。例えば、より軽量な前処理、後処理タスクを CPU 上で実行し、計算負荷の高いモデル タスク向けに GPU リソースを確保することで、企業はパフォーマンスを犠牲にすることなくコスト効率の良い拡張を実現できます。

3. **ネットワーク オーバーヘッドの軽減**：従来の機械学習パイプラインでは、各モデルがネットワークを介して相互と通信し、ステップ間で中間データを転送する必要がありました。しかし、モデル アンサンブルを使用することで、これらの中間データ転送を最小限に抑えることができ、ネットワーク呼び出しの回数が減少し、帯域幅の使用が最適化されます。

4. **コード不要の統合**：Triton のモデル アンサンブル機能は、異なる AI モデルを接続できる最低限のコードまたはコードを全く必要としないアプローチとなります。これにより、IT リーダーや推論チームが前処理、後処理ワークフロー (多くの場合 Python で構築) をシームレスなパイプラインに統合することが容易になります。この合理化された統合により、単一の推論リクエストで全プロセスを起動できるようになり、効率性がさらに向上します。

実用例：AI を活用した生成アプリケーション

モデル アンサンブルの仕組みをよりよく理解するために、生成 AI を用いた例を確認しましょう。入力テキストが合成画像に変換されるテキスト生成画像アプリケーションを想像してください。このようなアプリケーションのパイプラインは通常、入力テキストをエンコードするための LLM と、画像を生成するための拡散モデルという 2 つの主要コンポーネントで構成されています。

入力テキストが LLM に送信される前に、いくつかの前処理が必要です。これには、テキストのクリーニング、トークン化、または LLM と互換性のある形式へのフォーマット変換などが含まれることがあります。同様に、最終アプリケーションで使用できるようにするために、出力画像にはサイズ調整やエフェクト追加などの後処理が必要になる場合があります。

この場合、モデル アンサンブルを使用することで、テキストから画像への変換プロセスを完全に自動化し最適化することができます。Triton はテキスト前処理から画像生成、さらには後処理にいたるまでの各ステップを一つの統合パイプライン内にシームレスにつなげることができます。

AI 最適化におけるモデル アンサンブルの能力

複数のモデルと前処理、後処理ステップを一つの最適化パイプラインへと統合することで、IT リーダーは AI アプリケーションの効率を向上させ、レイテンシーを軽減し、リソース使用量を最小限に抑えることができます。

NVIDIA AI 推論プラットフォームは、これらのアンサンブルを管理する強力な基盤となり、リソースの割り当てに柔軟性をもたらすと共に、異なるコンポーネントの統合を簡素化します。少ないコードを使用するアプローチと自動化された最適化機能により、チームはより容易に AI システムを構築・拡張できるようになるため、AI モデルのパフォーマンス向上を目指す企業にとって不可欠なツールとなっています。

事例：Let's Enhance、製品写真を美しいマーケティング素材へと変換

NVIDIA AI 推論プラットフォームの能力を活用して本番環境で SDXL を提供している企業の良い例が「Let's Enhance」です。この先駆的な AI スタートアップは、3 年以上にわたり Triton Inference Server を使用して 30 種類以上の AI モデルを NVIDIA Tensor Core GPU 上にデプロイしています。

最近、Let's Enhance は最新製品「AI Photoshoot」を発表しました。この製品は SDXL モデルを活用して、シンプルな製品写真を e コマース ウェブサイトやマーケティング キャンペーン向けの美しいビジュアル素材に変換するものです。

Triton Inference Server は多様なフレームワークとバックエンドを堅牢にサポートしており、ダイナミックなバッチ処理機能セットも備えているため、Let's Enhance の創業者兼 CTO である Vlad Pranskevichus 氏は、ML エンジニア チームの関与を最小限に押さえて、SDXL モデルを既存の AI パイプラインへとシームレスに統合することに成功し、研究開発活動に専念するエンジニア チームの時間を解放することができました。

概念実証が成功したことで、この AI 画像強化スタートアップは SDXL モデルを Google Cloud G2 インスタンス上の NVIDIA L4 GPU に移行することでコストを 30% 削減できることを特定し、2024 年半ばまでに複数のパイプラインの移行を完了するためのロードマップを策定しました。

パフォーマンスを向上させる高度な推論技術

前章では、推論パフォーマンスを向上させるための基本戦略について説明しました。本章では、特に AI 推論が組織の収益創出戦略に不可欠である、あるいはすでにシステムを最適化済みで、現在 AI への投資から次のレベルのパフォーマンスを引き出そうとしている IT リーダー向けに調整された、より高度な技術について掘り下げていきます。NVIDIA AI 推論プラットフォームに搭載された最先端の最適化技術についての概要を提供し、自社チームにさらなる学習の方向性を示す方法についてリーダーらを導くことを目標としています。Triton と TensorRT-LLM から高度な戦略を探求し、チャンク プリフィル、マルチブロック アテンション、KV キャッシュの早期再利用、分散サービング、投機的デコーディングなどの、本番環境における AI 推論の境界を押し広げるために設計された技術を紹介します。

チャンク プリフィル：GPU 利用率の最大化とレイテンシーの削減

LLM 推論は通常、プリフィルとデコードという 2 つの重要なフェーズで構成されています。プリフィル フェーズでは、システムがユーザー入力のコンテキスト (KV キャッシュ) を計算しますが、これには高い処理能力を要します。続くデコード フェーズでは、KV キャッシュから導出された最初のトークンから始まり、順次トークンを生成していきます。しかし、従来の手法では、計算負荷の高いプリフィル フェーズと比較的軽い負荷のデコード フェーズのバランスを取ることに苦労することが多くありました。

チャンク プリフィルは、プリフィル フェーズを小さなチャンクに分割し、デコード フェーズとの並行処理化を向上させることで、ボトルネックを解消する最適化技術です。このアプローチは、GPU メモリと計算リソースを最大限に活用しながら、より長いコンテキストや高い同時実行レベルの処理に役立ちます。また、メモリ使用量を入力シーケンス長から切り離すことで柔軟性も提供するため、メモリ容量に負担をかけることなく大規模なリクエストを処理しやすくします。ダイナミックなチャンク サイジングにより、TensorRT-LLM は GPU 使用率に基づいてチャンク サイズをインテリジェントに調整し、デプロイメントを簡素化して手動で構成する必要性を解消します。

マルチブロック アテンション：長いコンテキストでスループットを向上

AI モデルの進化に伴い、より優れた認知理解を可能にするコンテキスト ウィンドウのサイズも飛躍的に増大しています。Llama 2 は 4K トークンから始まりましたが、最近の Llama 3.1 では、これが驚異的な 128K トークンにまで拡張されました。リアルタイムの推論シナリオでこれらの長いシーケンスを処理するタスクは、特に GPU リソースの割り当てにおいて独自の課題をもたらします。

TensorRT-LLM に搭載されているマルチブロック アテンション技術は、デコード処理を GPU 上のすべてのストリーミング マルチプロセッサ (SM) に効率よく分散させることで、これらの課題を解決します。これにより、スループットが大幅に向上し、大規模な KV キャッシュ サイズに伴うボトルネックが削減されます。この手法により、リアルタイム アプリケーションでよく見られる小さなバッチ サイズでも、GPU がフル活用され、NVIDIA HGX H200 プラットフォームで秒あたり最大 3.5 倍のトークン処理が可能になります。このパフォーマンス向上により、初回トークン生成時間 (TTFT) が犠牲になることもなく、複雑なクエリに対しても迅速な応答が保証されます。

KV キャッシュの早期再利用：初回トークン生成時間を短縮

KV キャッシュの生成は計算負荷の高いプロセスです。そのため、冗長な計算を回避してトークン生成を高速化する上で、KV キャッシュの再利用は重要な役割を果たします。しかし、従来の手法では、再

利用を発生させる前に KV キャッシュ全体が生成されるのを待つ必要があることが多く、高需要環境での非効率性につながっていました。

KV キャッシュの早期再利用は、キャッシュの完全な生成を待つのではなく、生成中のキャッシュの一部を再利用できるようにすることでこのプロセスを最適化します。この技術は、特にシステム プロンプトや事前定義された指示が必要なシナリオで、推論を大幅に高速化します。例えば、企業のチャットボットでは、ユーザー リクエストが同じシステム プロンプトを共有することが多いため、この機能によって高需要期間中の TTFT を最大 5 倍に高速化し、より迅速で応答性の高いユーザー体験を提供できます。

分散サービング：独立したリソース配分と分離されたスケーリング

従来の推論構成では、プリフィルとデコードの両フェーズが同じ GPU 上にあるケースが多く、リソース配分の非効率性や十分でないスループットという問題を引き起こしていました。NVIDIA Triton 分散サービング (DistServe) 戦略ではこれらのフェーズが分離されるため、AI 推論チームは各フェーズの特定のニーズに基づいて個別にリソースを割り当てることができます。これにより独立したリソース配分と分離されたスケーリングが可能になり、TTFT を最適化するためにプリフィル フェーズにより多くの GPU を割り当て、トークン間時間 (TBT) を改善するためにデコード フェーズ専用に追加の GPU を割り当てることができます。

さらにこの分離により、各フェーズに最適な並列処理が可能になります。計算負荷の高いプリフィル フェーズにはパイプライン並列処理を、メモリを多用するデコード フェーズにはテンソル並列処理を適用することができます。Triton DistServe は、スループット、TTFT、TBT のサービスレベル目標 (SLO) を維持しながら、推論コストを最大 50% 削減します。

投機的デコーディング：トークン生成を高速化

トークンを 1 つずつ順次生成する自己回帰的な処理方式では、時間がかかり効率性に欠ける場合があります。投機的デコーディングは、複数の潜在的なトークン シーケンスを並列に生成することでこれを最適化し、トークン生成に必要な時間を短縮します。TensorRT-LLM には、Draft Target や Eagle Decoding などの様々な投機的デコーディング手法が統合されており、システムが複数のトークンを予測し最も適切なものを選択できるようにします。

Llama 3.3 70B のようなモデルでは、投機的デコーディングによって秒あたりのトークン数が 3.5 倍と大幅に増加し、出力品質を損なうことなくスループットとユーザー体験が向上します。この技術は、各計算ステップの効率を最大化することが不可欠となる低レイテンシー、高スループット環境で特に有益です。

AI 推論の未来を解放

NVIDIA AI 推論プラットフォームは、AI 推論のあらゆる導入段階にある組織をサポートするために構築されています。まだ実験段階にあるか、市場投入の迅速化に焦点を当てている組織にとって、「推論パフォーマンスに影響を与えるデプロイ要因」の章で概説している戦略は、優れた出発点となります。その一方で、AI 推論が粗利益に影響を与える主要なコスト要因となっている組織にとっては、「クラウドにおける AI 推論パフォーマンスとコスト効率の最大化」の章で説明した高度なパフォーマンス向上とコスト削減技術が、システム効率とパフォーマンスを最大化し、コストを最小化するのに役立つでしょう。

NVIDIA AI 推論プラットフォームを使い始めるには

NVIDIA AI 推論プラットフォームでは、幅広いビジネスおよび IT 要件を満たせるように、柔軟な推論デプロイメント オプションを提供しています。

最速の価値実現を求める企業は、NVIDIA NIM を活用することができます。NVIDIA NIM は、あらゆる場所から、NVIDIA アクセラレーテッド インフラストラクチャ上で最新の AI 基盤モデルを実行できるようにする、事前パッケージ化および最適化された推論マイクロサービスを提供します。

独自の AI 推論ニーズに合わせて柔軟性、構成可能性、拡張性を最大限にしたい企業には、NVIDIA は NVIDIA Triton Inference Server と NVIDIA TensorRT を提供しており、特定の要件に合わせて推論提供プラットフォームをカスタマイズし最適化するための機能を提供します。

詳細と利用の始め方については、NVIDIA AI 推論ソリューションをご覧ください。