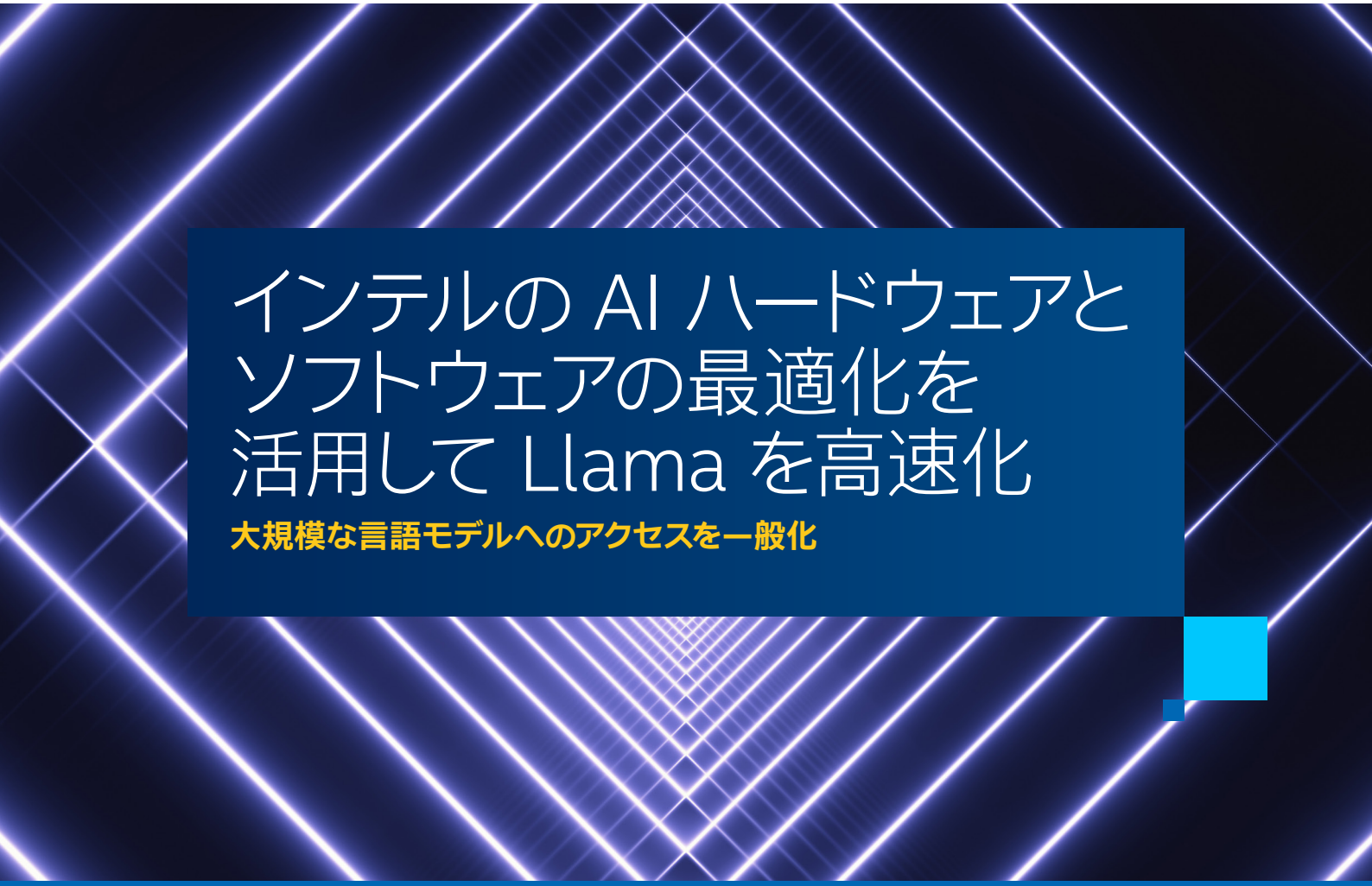

インテルの AI ハードウェアとソフトウェアの最適化を活用して Llama を高速化

大規模な言語モデルへのアクセスを一般化

Fan Zhao インテル コーポレーション AI ソフトウェア・エンジニアリング・マネージャー
Antony Vance Jeyaraj インテル コーポレーション クラウド・ソフトウェア・アーキテクト
Sree Ganesan インテル コーポレーション AI ソフトウェア・エンジニアリング・マネージャー
Susan Lansing インテル コーポレーション AI マーケティング・マネージャー
Kristina Kermanshahche インテル コーポレーション ソフトウェア・プロダクト・マネージャー
Radhika Rao インテル コーポレーション AI マーケティング・マネージャー
Patricia Mwove インテル コーポレーション クラウド・ソフトウェア・アーキテクト
Andres Rodriguez インテル コーポレーション フェロー

大規模言語モデル（LLM）へのアクセスをさらに一般化するため、Meta 社は Llama 2 をリリースしました。モデルをより広く利用できるようにすることで、AI コミュニティー全体で全世界に利益をもたらす取り組みが促進されるでしょう。LLM がテキストの生成、コンテンツの要約と翻訳、質問への応答、会話、および数学の問題を解くことや推論などのより複雑なタスクの実行において実証してきた優れた能力を考えると、LLM は、社会に利益をもたらす最も有望な AI テクノロジーの 1 つであると言えます。LLM には、新たな創造性と洞察力を引き出し、AI コミュニティーを刺激してテクノロジーを進歩させる可能性を秘めています。

Llama 2 は、開発者、研究者、組織が生成 AI を活用したツールとエクスペリエンスを構築するのを支援するように設計されています。Meta 社は、事前トレーニングおよび微調整済みの、70 億（7B）、130 億（13B）、および 700 億（70B）パラメーターの Llama 2 のモデルをリリースしました。Llama 2 で、Meta 社は微調整済みモデル全体に、中核となる 3 つの安全技術（教師あり安全微調整、ターゲットを絞った安全コンテキストの抽出、人間のフィードバックからの安全強化学習）を実装しました。これにより、Meta 社は安全実績を向上することができました。アクセスを一般化することにより、透過的かつオープンな方法で脆弱性を継続的に特定して軽減できるようになります。

インテルは、コミュニティーが Llama 2 のようなモデルを開発して実行できるように、競争力の高い魅力的なオプションを備えた AI ソリューションのポートフォリオを提供しています。インテルの豊富なハードウェア・ポートフォリオと、最適化されたオープン・ソフトウェアを組み合わせることにより、限定された計算リソースにアクセスするという課題を解決する代替手段が提供されます。この記事では、Habana® Gaudi®2 ディープラーニング・アクセラレーター、第 4 世代インテル® Xeon® スケーラブル・プロセッサー、インテル® Xeon® CPU マックス・シリーズ、およびインテル® データセンター GPU マックス・シリーズを含むインテルの AI ポートフォリオ上での、Llama 2 の 7B および 13B パラメーター・モデルの初期推論パフォーマンスを紹介します。ここで紹介する結果は、現在リリースされているソフトウェアのデフォルト設定のパフォーマンスであり、今後のリリースではさらなるパフォーマンスの向上が期待されます。現在、70B パラメーター・モデルにも取り組んでおり、近々コミュニティーに更新情報を提供する予定です。

Habana® Gaudi®2 ディープラーニング・アクセラレーター

Habana® Gaudi®2 は、ハイパフォーマンス、高効率のトレーニングと推論を提供するように設計されており、Llama や Llama 2 などの大規模言語モデルに特に適しています。LLM のメモリー要求を満たす（つまり、推論パフォーマンスを高速化する）ため、各 Habana® Gaudi®2 アクセラレーターは、96GB のオンチップ HBM2E を搭載しています。Habana® Gaudi®2 は、PyTorch* と DeepSpeed* を統合した、Habana SynapseAI* ソフトウェア・スイートにより、トレーニングと推論の両方をサポートしています。さらに、レイテンシーの影響を受けやすい推論アプリケーションに適した、HPU グラフ（英語）と DeepSpeed* 推論（英語）のサポートが最近 SynapseAI* に追加されました。2023 年第 3 四半期には、FP8 データ型のサポートを含む、さらなるソフトウェアの最適化が Habana® Gaudi®2 に提供される予定です。このアップデートにより、パフォーマンスの大幅な向上、スループットの向上、LLM 実行のレイテンシーの軽減が期待されます。

LLM のパフォーマンスを向上させるには、サーバー内とノード間の両方でネットワークのボトルネックを軽減する、柔軟かつ機敏なスケーラビリティーが必要です。すべての Habana® Gaudi®2 には、24 の 100GB イーサネット・ポートが統合されています。21 のポートをサーバー内の 8 つの Habana® Gaudi®2 の All-to-all 接続専用に、3 つのポートをスケールアウト専用にできます。このネットワーク構成は、サーバー内外の両方でスケールされたパフォーマンスを高速化するのに役立ちます。

Habana® Gaudi®2 は、最近公開された MLPerf* ベンチマーク（英語）に掲載された 384 の Habana® Gaudi®2 アクセラレーター上での 1750 億（175B）パラメーターの GPT-3* モデルのトレーニングで、大規模言語モデルでの優れたトレーニング・パフォーマンスを実証しました（詳細は、「MLCommons、AI でのインテルの強力な競争優位性を示す最新のベンチマーク結果を公開」を参照してください）。この実証されたパフォーマンスにより、Llama と Llama 2 のトレーニングと推論の両方で、Habana® Gaudi®2 は非常に効果的なソリューションとなります。

次に、単一の Habana® Gaudi®2 デバイスでの、バッチサイズ 1、出力トークン長 256、混合精度（BF16）を使用したさまざまな入力トークン長の Llama 2 7B および Llama 2 13B モデルの推論パフォーマンスを紹介します。パフォーマンス・メトリックは、（最初のトークンを除く）トークンあたりのレイテンシーです。推論の実行には、optimum-habana テキスト生成スクリプト（英語）を使用しました。Hugging Face の optimum-habana（英語）ライブラリーを使用すると、Habana® Gaudi® アクセラレーター向けにコード変更を最小限に抑えて、これらのモデルをシンプルかつ簡単にデプロイできます。**図 1** は、Habana® Gaudi®2 で入力トークン長 128 ~ 2K の推論を実行したレイテンシーが、7B モデルでトークンあたり 9.0 ~ 12.2 ミリ秒、13B モデルでトークンあたり 15.5 ~ 20.4 ミリ秒であることを示しています（ハードウェアとソフトウェアの構成の詳細は、この記事の最後に記載しています）。

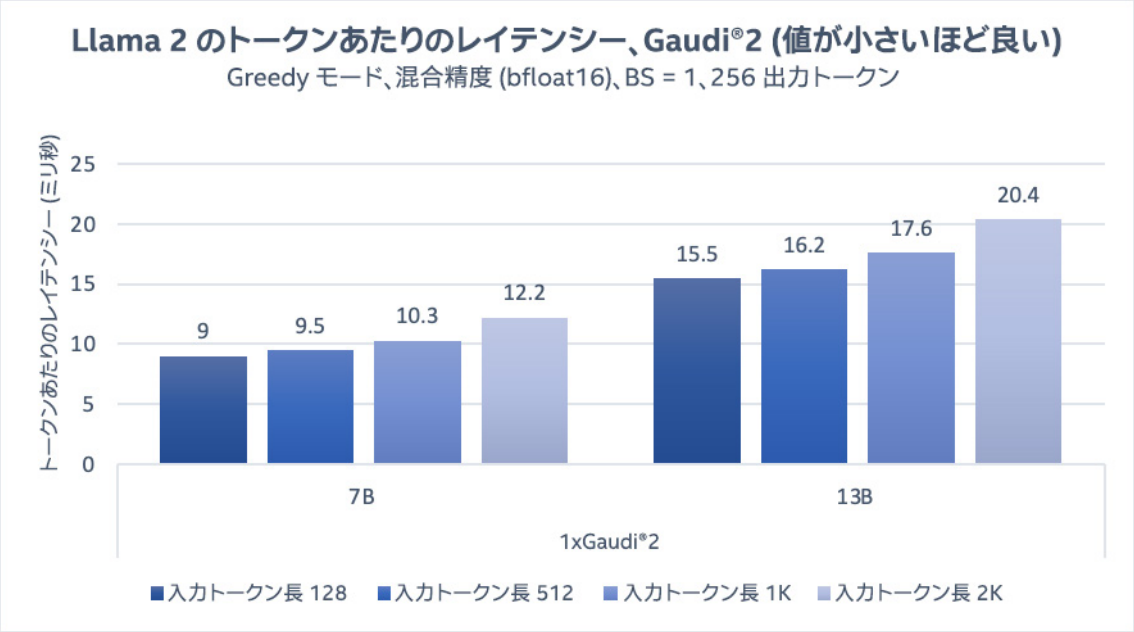

図 1. Habana® Gaudi®2 上の Llama 2 7B および 13B パラメーター・モデルの推論のパフォーマンス

Habana® Gaudi® プラットフォームで Llama 2 を使用して、生成 AI の旅を今すぐ始めることができます。Habana® Gaudi®2 にアクセスするには、インテル® デベロッパー・クラウド（英語）のインスタンスをサインアップするか、Habana® Gaudi®2 サーバー・インフラストラクチャーに関して Supermicro（英語）までお問い合わせください。

インテル® Xeon® スケーラブル・プロセッサー

第 4 世代インテル® Xeon® スケーラブル・プロセッサーには、インテル® アドバンスト・マトリクス・エクステンション（インテル® AMX）と呼ばれる AI 機能を大幅に高速化するアクセラレーターが内蔵されています。具体的には、すべてのコアに BF16 および INT8 GEMM（GEneral Matrix-matrix Multiplication、汎用行列乗算）アクセラレーターがあり、ディープラーニングのトレーニングと推論ワークロードを高速化します。さらに、インテル® Xeon® CPU マックス・シリーズでは、2 つのソケットで 128GB の高帯域幅メモリー（HBM2E）が搭載されており、ワークロードがメモリー帯域幅に制限されることが多い LLM で利点が得られます。

インテル® Xeon® プロセッサー向けのソフトウェア最適化はディープラーニング・フレームワーク（英語）にアップストリームされており、PyTorch*、TensorFlow*、DeepSpeed*、その他の AI ライブラリーのデフォルトのディストリビューションで利用できます。インテルは、PyTorch* 2.0（英語）の代表的な機能である torch.compile の CPU バックエンドの開発と最適化を主導しています。また、公式 PyTorch* ディストリビューションにアップストリームされる前でも、インテルの CPU 向けの高度な最適化を利用できるように PyTorch 向けインテル® エクステンション（英語）も提供しています。

大容量のメモリーが搭載されている第 4 世代インテル® Xeon® スケーラブル・プロセッサーは、単一ソケット内で低レイテンシーでの LLM 実行が可能であり、会話型 AI やテキスト要約アプリケーションに適用できます。この記事では、BF16 および INT8 で 1 つのソケットごとに 1 つのモデルを実行する場合のレイテンシーに注目しています。PyTorch 向けインテル® エクステンション（英語）は、INT8 精度モデルで適切な精度を確保する SmoothQuant（英語）をサポートしています。

LLM アプリケーションは高速リーダーの読み取り速度を満たす十分な速さでトークンを生成する必要があることから、調査するパフォーマンス・メトリックとしてトークンあたりのレイテンシー（各トークンの生成時間）を選択し、リファレンスとして熟練者の読み取り速度（トークンあたり 100 ミリ秒以下）を選択しました。**図 2** と **図 3** は、シングルソケットの第 4 世代インテル® Xeon® スケーラブル・プロセッサーで入力トークン長 32 ~ 2K の推論を実行したレイテンシーが、7B BF16 モデルおよび 13B INT8 モデルで 100 ミリ秒未満であることを示しています。

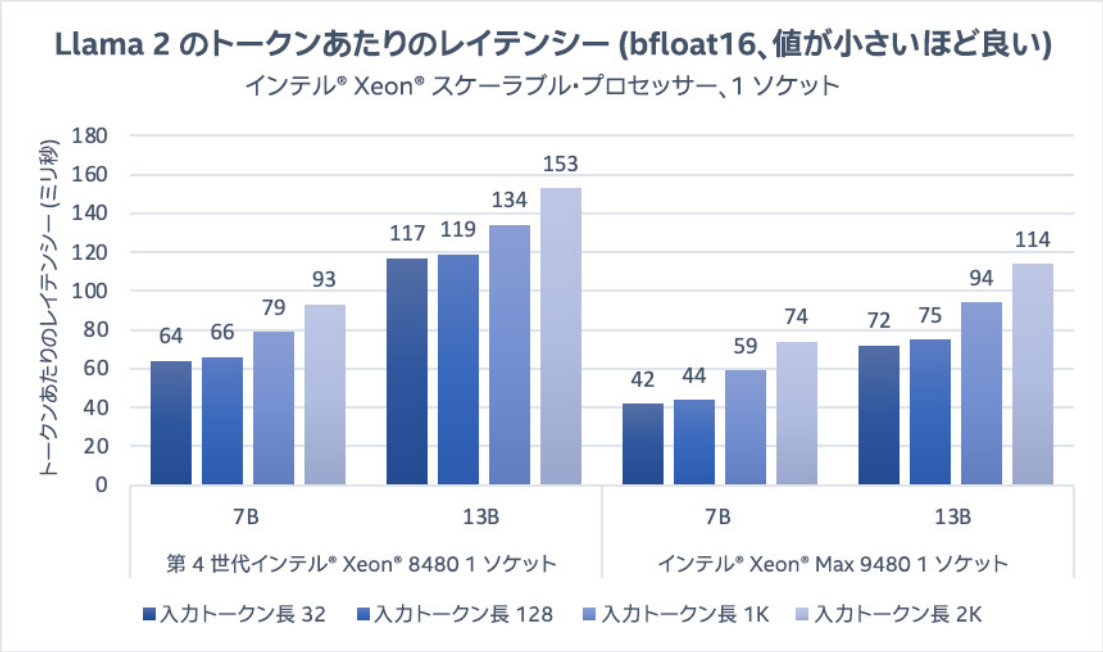

図 2. インテル® Xeon® スケーラブル・プロセッサー上の Llama 2 7B および 13B パラメーター・モデルの推論（bfloat16）のパフォーマンス

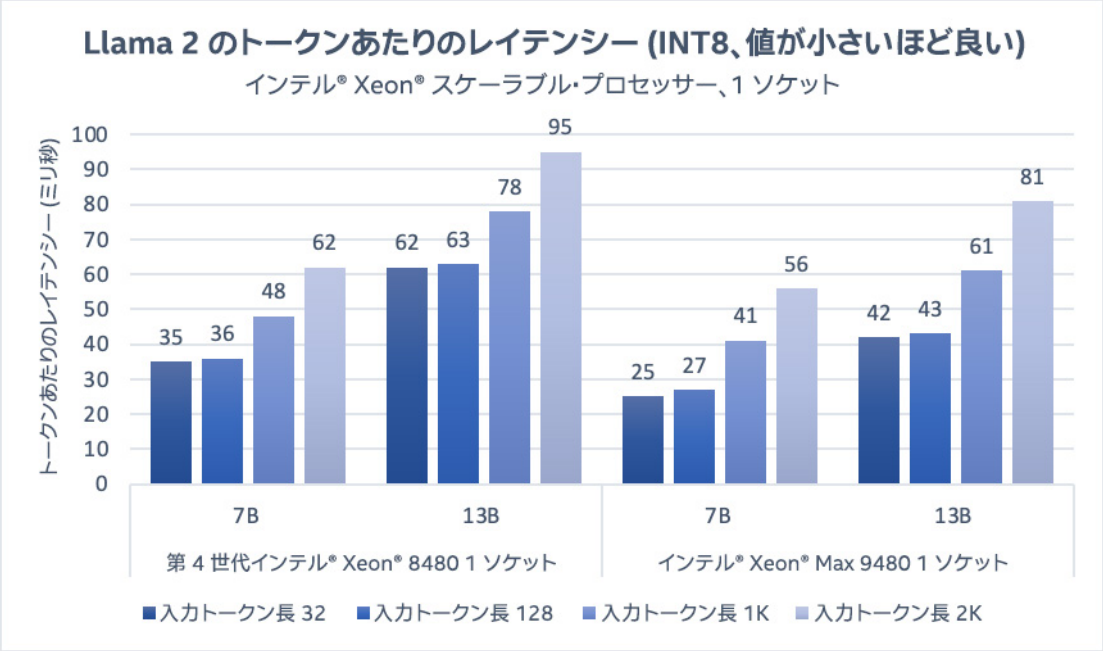

図 3. インテル® Xeon® スケーラブル・プロセッサー上の Llama 2 7B および 13B パラメーター・モデルの推論（INT8）のパフォーマンス

インテル® Xeon® CPU マックス・シリーズは、HBM2E の高帯域幅の恩恵を受け、両方のモデルのレイテンシーを軽減します。インテル® AMX アクセラレーションは、大きなバッチサイズでのスループットを向上させます。1 ソケットの第 4 世代インテル® Xeon® スケーラブル・プロセッサーで、7B および 13B パラメーター・モデルで 100 ミリ秒未満のレイテンシーを達成できます。ユーザーは、スループットを向上させ、複数クライアントの個々に対応できるように、各ソケットで 2 つの並列インスタンスを実行できます。あるいは、PyTorch 向けインテル® エクステンション（英語）と DeepSpeed*（英語）を活用して両方の第 4 世代インテル® Xeon® スケーラブル・プロセッサー上で推論を実行し、テンソル並列処理を使用してレイテンシーをさらに軽減したり、より大きなモデルをサポートすることもできます。

インテル® Xeon® プラットフォーム上での LLM と Llama 2 の実行に関する詳細は、こちら（英語）から入手できます。第 4 世代インテル® Xeon® スケーラブル・プロセッサーのクラウド・インスタンスは、AWS*、GCP* および Azure* でプレビュー版を利用できます。Ali Cloud では一般公開されています。インテルは、今後も PyTorch* および DeepSpeed* にソフトウェアの最適化を追加して、Llama 2 やその他の LLM をさらに高速化する予定です。

インテル® データセンター GPU マックス・シリーズ

インテル® データセンター GPU マックスは、並列計算、HPC、および AI ワークロードを高速化します。インテル® データセンター GPU マックス・シリーズは、インテルで最も高性能かつ高密度のディスクリート GPU であり、1,000 億個以上のトランジスターを 1 つのパッケージに搭載し、インテルの GPU 計算ビルディング・ブロックであるインテル® X^e コアを最大 128 個搭載しています。

インテル® データセンター GPU マックス・シリーズは、AI や HPC 分野で使用されるデータを多用するコンピューティング・モデルにおいて、画期的なパフォーマンスを発揮するように設計されています。

- ディスクリート SRAM テクノロジーに基づく 408MB の L2 キャッシュと 64MB の L1 キャッシュおよび最大 128GB の高帯域幅メモリー（HBM2E）。
- AI を強化するインテル® Xe マトリクス・エクステンション（インテル® XMX）は、シストリック・アレイを搭載し、単一のデバイスでベクトルと行列の機能を実現。

インテル® データセンター GPU マックス・シリーズ製品ファミリーは、生産性とパフォーマンスを最大限に引き出す、共通のオープンな標準ベースのプログラミング・モデルである oneAPI により統合されています。インテル® oneAPI ベース・ツールキットには、高度なコンパイラー、ライブラリー、プロファイラー、CUDA* コードから C++ with SYCL* への移行を支援するコード移行ツールが含まれています。

インテル® データセンター GPU マックス・シリーズ向けのソフトウェアの有効化と最適化は、PyTorch 向けインテル® エクステンション、TensorFlow 向けインテル® エクステンションおよび DeepSpeed 向けインテル® エクステンションなど、主要なフレームワーク向けのオープンソースの拡張を通じて提供されます。これらの拡張をアップストリームのフレームワークのリリースとともに使用することで、ユーザーはマシンラーニング・ワークフローのドロップイン・アクセラレーションを実現できます。

Llama 2 7B および 13B パラメーター・モデルの推論パフォーマンスはパッケージに 2 つの GPU（タイル）を搭載した 600W OAM デバイスで評価しましたが、推論の実行には 1 つのタイルのみ使用しました。**図 4** は、1 タイルのインテル® データセンター GPU マックスで入力トークン長 32 ~ 2K の推論を実行したレイテンシーが、7B モデルでトークンあたり 20 ミリ秒未満、13B モデルでトークンあたり 29.2 ~ 33.8 ミリ秒であることを示しています。ユーザーは、スループットを向上させ、複数のクライアントに別々に対応できるように、各タイルで 1 つずつ、2 つの並列インスタンスを実行できます。

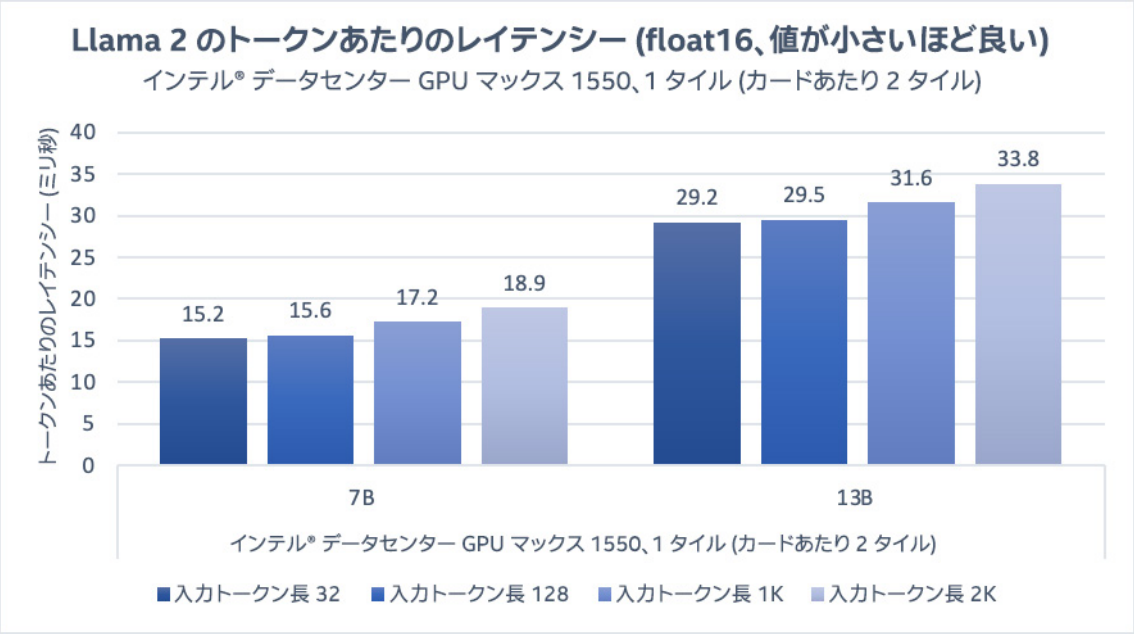

図 4. インテル® データセンター GPU マックス 1550 上の Llama 2 7B および 13B パラメーター・モデルの推論のパフォーマンス

インテル® データセンター GPU プラットフォーム上での LLM と Llama 2 の実行に関する詳細は、こちら（英語）から入手できます。インテル® デベロッパー・クラウドで、インテル® データセンター GPU マックスのクラウド・インスタンスを利用できます（記事執筆時点ではベータ版）。

インテルは、推論に加えて、Hugging Face Transformers（英語）、PEFT（英語）、Accelerate（英語）、および Optimum（英語）ライブラリーへの最適化のアップストリームによる微調整の高速化にも積極的に取り組んでいます。また、サポートされるインテルのプラットフォームにテキスト生成、コード生成、補完、要約などの典型的な LLM ベースのタスクを効率的にデプロイできる、Transformers 向けインテル® エクステンション（英語）のリファレンス・ワークフロー（英語）を提供しています。

まとめ

この記事では、Habana® Gaudi®2 ディープラーニング・アクセラレーター、第 4 世代インテル® Xeon® スケーラブル・プロセッサー、インテル® Xeon® CPU マックス・シリーズ、およびインテル® データセンター GPU マックス・シリーズを含む、インテルの AI ハードウェア・ポートフォリオ上での、Llama 2 の 7B および 13B パラメーター・モデルの初期推論パフォーマンスを紹介しました。ソフトウェア・リリースでは引き続き最適化を行っており、近々 LLM およびより大きな Llama 2 モデルについての更新情報を提供する予定です。

製品および性能に関する情報

Habana® Gaudi®2 ディープラーニング・アクセラレーター：測定は、8x Habana® Gaudi®2 HL-225H メザニンカード、2x インテル® Xeon® Platinum 8380 プロセッサー @ 2.30GHz および 1TB システムメモリーを搭載した HLS2- Gaudi®2 サーバー上で Habana SynapseAI* バージョン 1.10 および optimum-habana バージョン 1.6 を使用して行われました。パフォーマンスは 2023 年 7 月に測定されました。

第 4 世代インテル® Xeon® スケーラブル・プロセッサー：

- 第 4 世代インテル® Xeon® 8480：第 4 世代インテル® Xeon® Platinum 8480+ プロセッサー、2 ソケットシステム、112 コア（224 スレッド）、インテル® ターボ・ブースト・テクノロジー有効、インテル® ハイパースレッディング・テクノロジー有効、メモリー：16x32GB DDR5 4800MT/ 秒、ストレージ：953.9GB。OS：CentOS* Stream 8、カーネル：5.15.0-spr.bkc.pc.16.4.24.x86_64。バッチサイズ：1。1 ソケットで計測。PyTorch* nightly build0711。PyTorch 向けインテル® エクステンション tag v2.1.0.dev+cpu.llm。モデル：Llama 2 7B および 13B、データセット LAMBADA。トークン長：32/128/1024/2016（in）、32（out）。ビーム幅 4。精度：BF16 および INT8。2023 年 7 月 12 日に行われたインテル社内のテスト結果。
- インテル® Xeon® マックス 9480：インテル® Xeon® CPU マックス 9480 プロセッサー、2 ソケットシステム、112 コア（224 スレッド）、インテル® ターボ・ブースト・テクノロジー有効、インテル® ハイパースレッディング・テクノロジー有効、メモリー：16x64GB DDR5 4800MT/ 秒、8x16GB HBM2 3200 MT/ 秒、ストレージ：1.8TB。OS：CentOS* Stream 8、カーネル：5.19.0-0812.intel_next.1.x86_64+server。バッチサイズ：1。1 ソケットで測定。PyTorch* nightly build0711。PyTorch 向けインテル® エクステンション llm_feature_branch。モデル：Llama 2 7B および 13B、データセット LAMBADA。トークン長：32/128/1024/2016（in）、32（out）。ビーム幅 4。精度：BF16 および INT8。2023 年 7 月 12 日に行われたインテル社内のテスト結果。

インテル® データセンター GPU マックス・シリーズ：1 ノード、2x インテル® Xeon® Platinum 8480+ プロセッサー、56 コア、インテル® ターボ・ブースト・テクノロジー有効、インテル® ハイパースレッディング・テクノロジー有効、NUMA 2、合計メモリー 1024GB（16x64GB DDR5 4800MT/ 秒 [4800MT/ 秒]）。BIOS SE5C7411.86B.9525.D19.2303151347、マイクロコード 0x2b0001b0、1x イーサネット・コントローラー X710 10GBASE-T、1x1.8TB WDC WDS200T2B0B、1x931.5GB INTEL SSDPELKX010T8、Ubuntu* 22.04.2 LTS、5.15.0-76-generic、4x インテル® データセンター GPU マックス 1550（1 つの OAM GPU カードの 1 つのタイルのみ使用して測定）、IFWI PVC 2_1.23166、agama driver：agama-ci-devel-627.7、インテル® oneAPI ベース・ツールキット 2023.1、PyTorch* 2.0.1 + PyTorch 向けインテル® エクステンション v2.0.110+xpu（dev/LLM branch）。AMC ファームウェア・バージョン：6.5.0.0。モデル：Llama 2 7B および 13B、データセット LAMBADA。トークン長：32/128/1024/2016（in）、32（out）、Greedy 検索、精度：FP16。2023 年 7 月 7 日に行われたインテル社内のテスト結果。