2024 年 9 月 26 日

報道関係各位

GMO インターネットグループ株式会社

GMO インターネットグループ、「NVIDIA H200 GPU」搭載環境の性能を実証

～生成 AI 向けクラウドサービス「GMO GPU クラウド」を 11 月下旬提供開始～

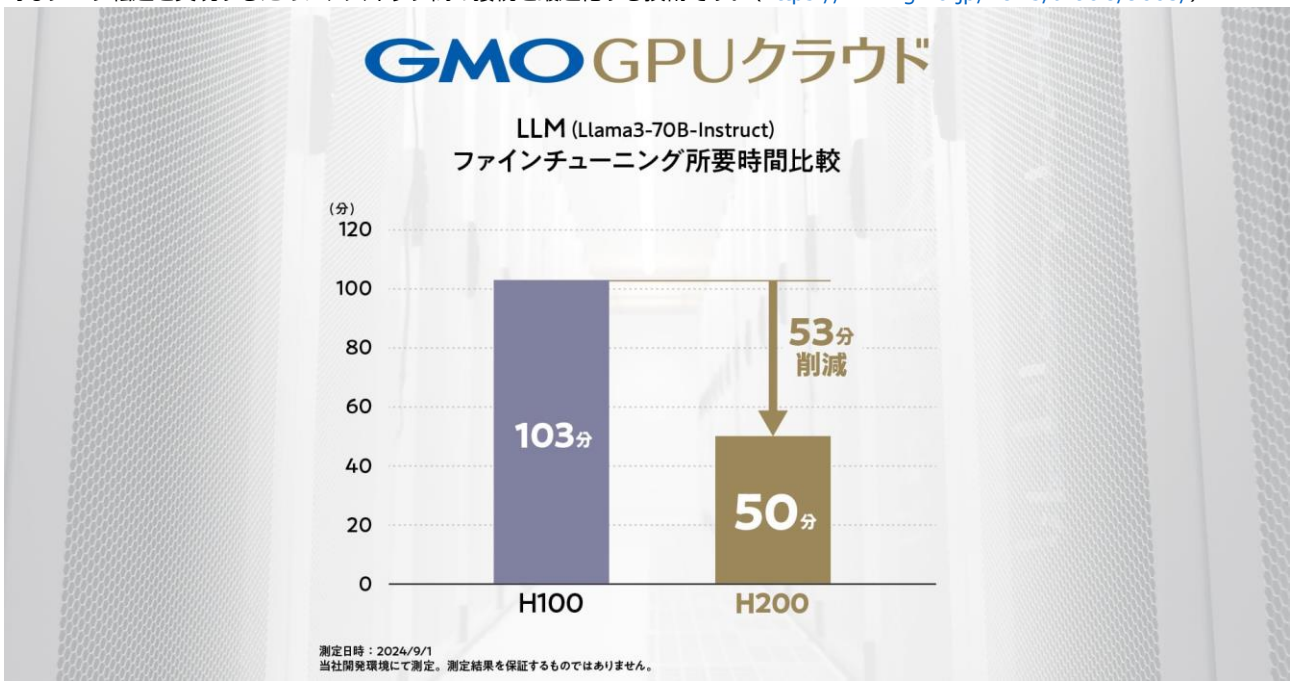
GMO インターネットグループ株式会社（代表取締役グループ代表：熊谷 正寿）は、「NVIDIA H200 Tensor コア GPU」（以下、H200 GPU）と AI ワークロード専用に設計された世界初のイーサネットファブリック「NVIDIA Spectrum™-X」（以下、Spectrum-X）(※1) 搭載環境を国内で初めて検証し、結果を開示しました。

今回の検証では、グラフニューラルネットワーク（GNN）のベンチマークと、大規模言語モデル（LLM）のファインチューニングの 2 つのシナリオを実施し、いずれのケースでも H200 GPU は「NVIDIA H100 Tensor コア GPU」（以下、H100 GPU）よりも高い性能を発揮することがわかりました。LLM（Llama3-70B-Instruct）のファインチューニングシナリオにおいては、H100 GPU 搭載機材では 103 分かかっていた学習時間が、H200 GPU 搭載機材では 50 分で完了し、処理速度が約 2 倍となりました。

なお、GMO インターネットグループは、2024 年 11 月下旬に、計算性能において AI 開発の効率化に貢献するサービスとして、H200 GPU と NVIDIA Spectrum-X を国内で初採用した生成 AI 向けのクラウドサービス「GMO GPU クラウド」の提供を開始する予定です。

GMO GPU クラウド URL：https://www.gmo.jp/gpucloud/

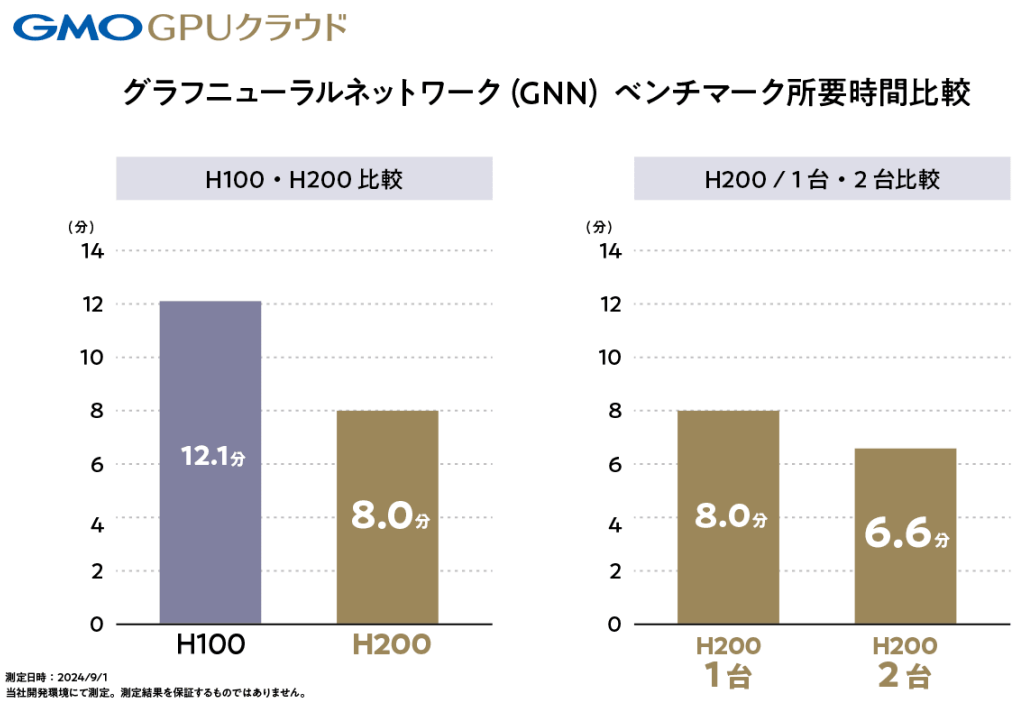
（※1）「NVIDIA Spectrum-X」は AI ワークロード専用に設計された世界初のイーサネットファブリックであり、生成 AI ネットワークのパフォーマンスを飛躍的に向上させることができます。イーサネットファブリックとは、ネットワークデバイス間の高速かつ効率的なデータ転送を実現するために、スイッチ間の接続を最適化する技術です。（https://www.gmo.jp/news/article/9005/）



【ベンチマークテストの概要と結果】

GMO インターネットグループは、「GNN のベンチマーク」と「LLM ファインチューニング」の 2 つのシナリオにおいて H200 GPU と H100 GPU の性能を比較しました。

■GNN のベンチマーク結果



超高速で複雑な計算を行う強力なコンピューターシステムである HPC（ハイパフォーマンス・コンピューティング）クラスタ全体の GNN 学習性能を測定するベンチマーク性能において、H200 GPU は H100 GPU（12.1 分）に比べて、約 1.5 倍の高速化(8.0 分)を実現しました。【表 1】

これは、H200 GPU が H100 GPU に対してメモリ搭載量が約 1.7 倍に強化されたことで、その演算性能が最適化されたためと考えられます。

また、H200 GPU の 2 台構成では、GPU 数の増加によって、ベンチマーク時間が短縮できていることも確認できました。【表 2】

【表 1：GNN ベンチマーク所要時間：H100 GPU・H200 GPU 所要時間比較】

構成	所要時間	削減時間	H100 性能比
H100（1 台）	12.1 分	—	100%
H200（1 台）	**8.0 分**	**4.1 分**	**151%**

【表 2：GNN のベンチマーク所要時間： H200 GPU/1 台・2 台所要時間比較】

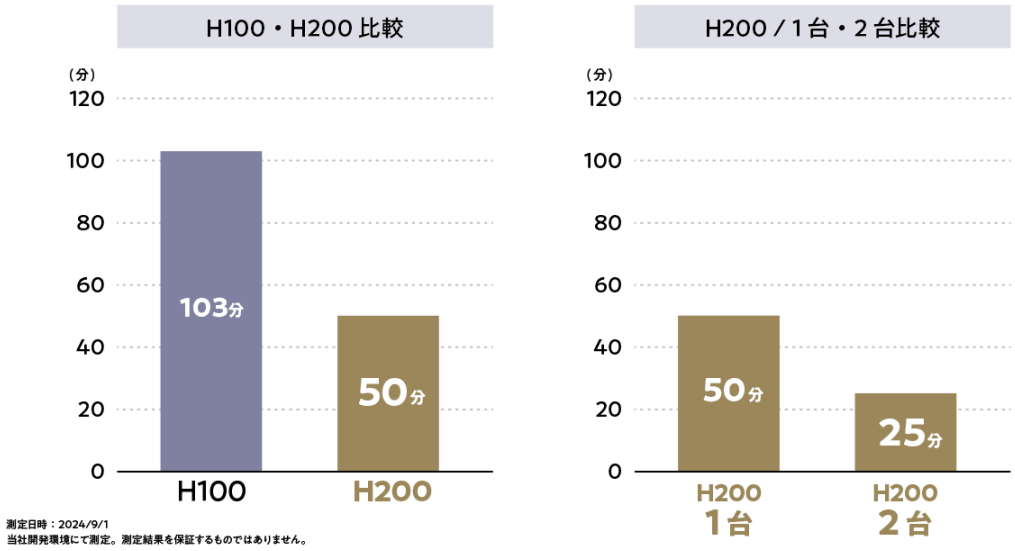
構成	所要時間	削減時間	H200（1 台）性能比
H200（1 台）	8.0 分	—	100%
H200（2 台）	**6.6 分**	**1.4 分**	**121%**

■LLM ファインチューニング結果

さらに、実際の GPU での学習時間に焦点を絞り、LLM ファインチューニングにかかる時間を測定しました。

LLM のファインチューニングにおいては、H100 GPU 搭載機材では 103 分かかっていた学習時間が、H200 GPU 搭載機材では 50 分で完了し、処理速度が約 2 倍となりました。【表 3】

LLM（Llama3-70B-Instruct）ファインチューニング所要時間比較



前述の GNN のベンチマーク結果と同様に、LLM においても、H200 GPU が H100 GPU と比べてメモリ搭載量が増加したことで、モデルがより GPU の性能を生かした学習を行えることがわかりました。

さらに、H200 2 台構成では、同 1 台構成比で処理速度が約 2 倍の 25 分で学習が完了しており、GPU が増えたことによって学習時間が短縮されることは当然ながら、Spectrum-X の採用によって複数台構成でもその性能を最大限に引き出せることがわかりました。【表 4】

【表 3： LLM（Llama3-70B-Instruct）ファインチューニング所要時間比較）

構成	所要時間	削減時間	H100 性能比
H100（1 台）	103 分	—	100%
H200（1 台）	**50 分**	**53 分**	**206%**

【表 4： LLM（Llama3-70B-Instruct）ファインチューニング所要時間比較】

構成	所要時間	削減時間	H200（1 台）性能比
H200（1 台）	50 分	—	100%
H200（2 台）	**25 分**	**25 分**	**200%**

■実施環境

	H100	H200
サーバモデル	DELL PowerEdge XE9680	同左
CPU	第 4 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー	同左
メモリ	2048GB	同左
ディスク構成	NVMe 1.92TB x 4	NVMe 7.68TB x 4
GPU（※2）	NVIDIA H100 SXM x 8	NVIDIA H200 SXM x 8

(※2)いずれも当社所有機材

- 測定日時：2024/9/1
- パフォーマンス結果は、記載された構成でのテストに基づき、当社環境で計測した参考値であり、検証結果ならびに製品の性能・動作について保証するものではありません。

■ GMO インターネットグループ インフラ・運用本部 プロジェクト統括チーム エグゼクティブリード 佐藤嘉昌 コメント

高性能な「NVIDIA H200 GPU」と先進のネットワーク「NVIDIA Spectrum-X」を採用した「GMO GPU クラウド」の提供により、日本の AI 研究開発を加速させることを目指しています。高性能かつ柔軟な GPU リソースを提供することで、研究者や開発者の皆様がより効率的に AI 開発に取り組めるよう支援してまいります。この新サービスが日本の AI 技術革新の一助となり、グローバル競争力の強化につながることを期待しています。

■ 今後の展開

「GMO GPU クラウド」は 2024 年 11 月下旬のサービス開始を予定しています。現在、H200 GPU のセットアップを進めており、順次サーバーの設置と稼働テストを行っています。今後も定期的に進捗情報を発信し、ユーザーのニーズに応じたサービス開発を進めていく予定です。GMO インターネットグループは、「GPU なら GMO インターネットグループ」と認知されることを目指し、AI 産業に欠かせないクラウドサービスとして、市場に新たな価値を提供してまいります。

■「GMO GPU クラウド」トライアルのお申し込みについて

「GMO GPU クラウド」の利用を検討されるお客様には、トライアル環境のご相談を承ります。
本サービスへのお問い合わせ・トライアルのお申し込みはこちら
URL：https://forms.office.com/r/pMi1PzTGEW

■サービス提供の背景

近年の生成 AI 開発において、LLM のモデルサイズは大規模化しており、それに伴い学習時間が増大し、開発期間の長期化やコスト増加が課題となっています。この状況下で、国内の生成 AI インフラ環境には重要な課題が浮上しています。大規模モデル開発のユースケースを正確に把握できていないため、GPU の提供が先行し、必要不可欠な広帯域ネットワークや高速ストレージの整備が遅れがちです。

従来の GPU クラウドサービスでは、これらの複合的な課題を解決することが困難でした。GMO インターネットグループは、このような課題を抱える企業に向けて、より高速で効率的な生成 AI 開発環境を提供するため、「NVIDIA H200 GPU」採用の GPU クラウドサービスの提供開始に至りました。

【「GMO GPU クラウド」について】（URL：https://www.gmo.jp/gpucloud/）

GMO インターネットグループの「GMO GPU クラウド」は、高性能な「NVIDIA H200 GPU」を搭載し、国内で最速の提供を目指します。(※3)また、AI ワークロード専用に設計された世界初のイーサネットファブリック「NVIDIA Spectrum-X」を国内クラウド事業者で初めて採用しています。この H200 GPU と Spectrum-X の統合で、生成 AI 開発や機械学習に最適化した高水準の GPU クラウド環境を実現します。

さらに、最も要求の厳しい AI/ML/DL 大規模モデルをトレーニングするよう設計された Dell PowerEdge XE9680 (NVIDIA H200 SXM 8 基搭載)を採用し、システムの構築を進めています。

GMO インターネットグループは、本サービスを通じて、生成 AI 分野やハイパフォーマンス・コンピューティング（HPC）分野に取り組む企業や研究機関に対し、インフラのチューニングが不要の高水準な計算環境を提供し、お客様の開発期間の短縮とコスト低減に貢献し、国内 AI 産業の発展を促進します。

（※3）GMO インターネットグループ、NVIDIA H200 Tensor コア GPU を採用した生成 AI 向けの GPU クラウドサービスを国内最速提供へ：https://www.gmo.jp/news/article/8933

・提供開始時期： 2024 年 11 月下旬 (予定)

■「GMO GPU クラウド」の特徴

1. 国内最速レベルの提供となる「NVIDIA H200 Tensor コア GPU」搭載

大規模言語モデルの開発・研究者向けに GPU メモリ容量とメモリバス帯域幅を大幅に拡大・最適化した H200 GPU を国内最速提供。H200 は、毎秒 4.8 テラバイト (TB/s) で 141 ギガバイト(GB) の HBM3e メモリを提供する初の GPU です。これは、H100 GPU の約 1.7 倍の容量で、メモリ帯域幅は約 1.4 倍です。

2. 国内クラウド事業者初となる「NVIDIA Spectrum-X」の採用

AI ワークロード専用に設計された世界初のイーサネットファブリックである Spectrum-X を国内で初めて採用。生成 AI ネットワークのパフォーマンスを飛躍的に向上させます。

3. DDN の超高速ストレージを採用

NVIDIA 製品と互換性のある DDN の高速ストレージを採用。強力な性能を持つ AI 開発プラットフォームをワンストップで提供します。

4. NVIDIA AI Enterprise による迅速な環境構築・管理

NVIDIA AI Enterprise は、データサイエンスパイプラインを加速し、プロダクショングレードのコパイロットやその他の生成 AI アプリケーションの開発と展開を合理化する、エンドツーエンドのクラウドネイティブなソフトウェアプラットフォームです。

5. 業界標準のジョブスケジューラーSlurm を採用

クラスタシステムのための業界標準であるジョブスケジューラーです。リソースの割り当て・ジョブの制御・モニタリング機能を提供します。

【GMO インターネットグループ株式会社について】

GMO インターネットグループ株式会社は、1995 年 12 月にインターネット事業を創業して以来、“すべての人にインターネット”をコーポレートキャッチに、インターネットの場の提供に経営資源を集中し、インターネットをより豊かに便利にするべく事業を展開してまいりました。

現在では、インターネットインフラ事業、インターネット広告・メディア事業、インターネット金融事業、暗号資産事業を展開しています。ご利用いただいているお客様の数は 2024 年 6 月末時点で 1,775 万顧客、上場企業 10 社を中心とした全 112 社、グループパートナー数約 7,500 名の総合インターネットグループに成長しています。また、「AI で未来を創るナンバー 1 企業グループへ」を掲げ、グループ全パートナーを挙げて生成 AI を活用することで、① 時間とコストの節約、② 既存サービスの質向上、③ AI 産業への新サービス提供を進めています。

以上

【報道関係お問い合わせ先】

●GMO インターネットグループ株式会社
本体事業管理本部 広報担当 川縁
TEL：03-5456-2555　E-mail：pr@gmo.jp

●GMO インターネットグループ株式会社
グループ広報部 PR チーム 山崎
TEL：03-5456-2695　URL： https://www.gmo.jp/contact/press-inquiries/

【サービスに関するお問い合わせ先】

●GMO インターネットグループ株式会社
ドメイン・ホスティング事業本部
E-mail：aicloud@gmo.jp

【GMO インターネットグループ株式会社】　（URL：https://www.gmo.jp/）

会　社　名	GMO インターネットグループ株式会社　（東証プライム市場　証券コード:9449）
所　在　地	東京都渋谷区桜丘町 26 番 1 号　セルリアンタワー
代　表　者	代表取締役グループ代表　熊谷 正寿
事 業 内 容	■インターネットインフラ事業　■インターネット広告・メディア事業 ■インターネット金融事業　■暗号資産事業
資　本　金	50 億円