2026 年 1 月 23 日

報道関係各位

GMO インターネット株式会社

GMO インターネット、「NVIDIA B300 GPU」搭載環境の性能を実証

～「GMO GPU クラウド」ベアメタルサービス搭載 GPU の実力を検証～

GMO インターネットグループの、GMO インターネット株式会社（代表取締役 社長執行役員：伊藤 正　以下、GMO インターネット）は「GMO GPU クラウド」において、2024 年 11 月より提供している「NVIDIA H200 Tensor コア GPU」（以下、H200 GPU）」、および 2025 年 12 月にベアメタル構成にて国内最速クラスで提供開始した「NVIDIA HGX B300 AI インフラストラクチャ」(以下、B300 GPU)」を導入した GPU クラウドサービスの、性能特性を検証しました。生成 AI 開発から運用までの実用性と演算性能の両面を評価するため、以下 3 つのベンチマーク（性能検証）を実施した結果を公開いたします。

【実施したベンチマークの概要】

1．大規模言語モデル（LLM）の学習ベンチマーク：「学習効率」と「演算速度」を評価する指標

LLM を実際に学習（ファインチューニング）させ、目標の品質（損失）に到達するまでの学習完了時間を測るベンチマーク

2．vLLM bench throughput による推論ベンチマーク：単位時間あたりに生成可能な「トークン量（処理スループット）」を評価する指標

LLM 推論のバッチ処理をできるだけ高速に実行し、1 秒あたりに生成できる出力トークン数（output tokens/s）など最大スループットを測る推論性能ベンチマーク

3．HPL Benchmark によるベンチマーク：高精度な数値計算の処理能力を評価する指標

密行列の連立一次方程式（Ax=b）を解く処理を通じて、浮動小数点演算性能（GFLOPS）を測定する HPC 系の基礎計算性能ベンチマーク　科学技術計算における複雑で精密な数値計算の性能を測定

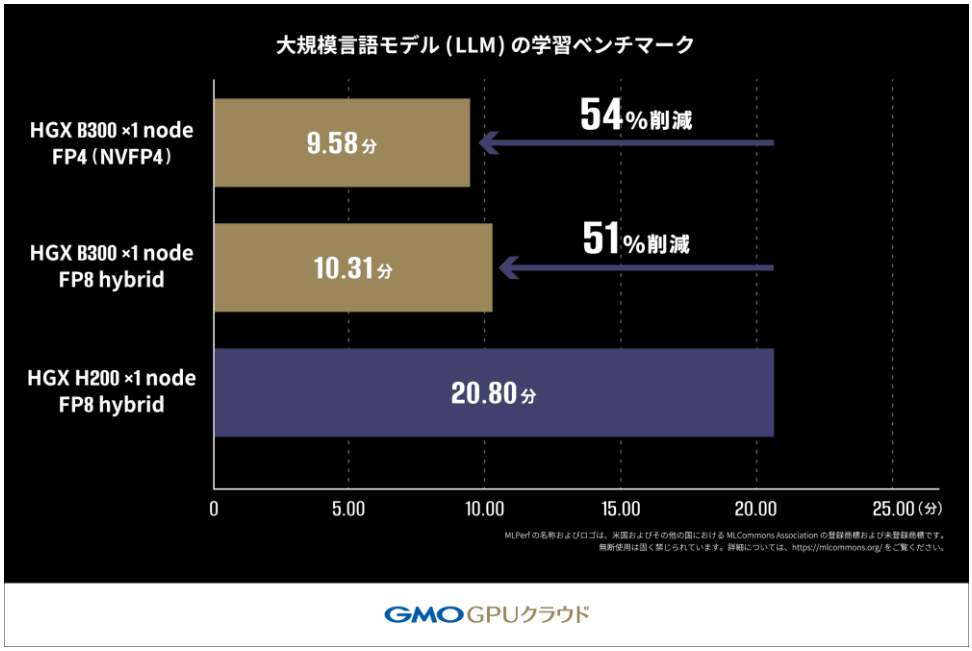
これらのベンチマークにより、生成 AI の開発から運用までの実用性能と、演算性能の両面から「B300 GPU」、「H200 GPU」の各々の特性を検証し、ワークロードに応じた最適な GPU を選択できる参考情報を提供します。

今回の検証では生成 AI ワークロードにおいて、「B300 GPU」は「H200 GPU」と比較して学習で約 2 倍、推論では最大約 2.5 倍の処理性能を発揮することが確認されました。一方、スーパーコンピュータの性能評価に用いられる HPL Benchmark では、「B300 GPU」は「H200 GPU」の 2.1%（約 47 分の 1）の性能に留まりました。

これは「B300 GPU」が生成 AI ワークロードに特化した高い性能を有している一方で、科学技術計算など計算結果の正確性を求めるユースケースにおいては依然として「H200 GPU」が適している可能性を示唆しています。

【ベンチマークテストの概要と結果】

1. 大規模言語モデル（LLM）の学習ベンチマーク



本ベンチマークでは、MLPerf ® Training v5.1(※1) が規定している Closed Division のルールに従い、Llama2 70B モデルを用いて「B300 GPU」および「H200 GPU」上での LoRA ファインチューニング(※2)にかかる学習時間を測定しました(※3)。
評価指標にははクロスエントロピー損失(※4)を用い、目標値（0.925）に達するまでの時間を測定しています。

このベンチマークにおいて、「H200 GPU」搭載機材では 20.80 分(Unverified)かかっていた学習時間が「B300 GPU」搭載機材では 10.31 分(Unverified)で完了し、約 2 倍の速度で処理が完了しました。
さらに、NVIDIA Blackwell アーキテクチャより新たに対応した FP4(※5)を用いた測定では FP8 hybrid(※6)を使用した学習よりもさらに短い時間で処理が完了しており、FP4 の高い演算性能を活かすことにより学習でもその恩恵を受けることができる可能性を示しています【表 1】。

【表 1：大規模言語モデル（LLM）ベンチマーク時間比較】

構成	GPU 数	精度	所要時間（分）	削減時間（分）	H200 比 削減割合
HGX H200 1 台	8	FP8 hybrid	20.80(Unverified)	-	-
HGX B300 1 台	8	FP8 hybrid	10.31(Unverified)	10.49	51%
HGX B300 1 台	8	FP4(NVFP4)	9.58(Unverified)	11.22	54%

（※１）MLPerf ®とは、MLCommons Association が管理する機械学習システムの性能測定における国際的なベンチマーク標準。MLPerf の名称およびロゴは、米国およびその他の国における MLCommons Association の登録商標および未登録商標です。無断使用は固く禁じられています。詳細については、www.mlcommons.org をご覧ください。

（※2）LoRA ファインチューニングとは、大規模言語モデルを効率的に学習させる手法。

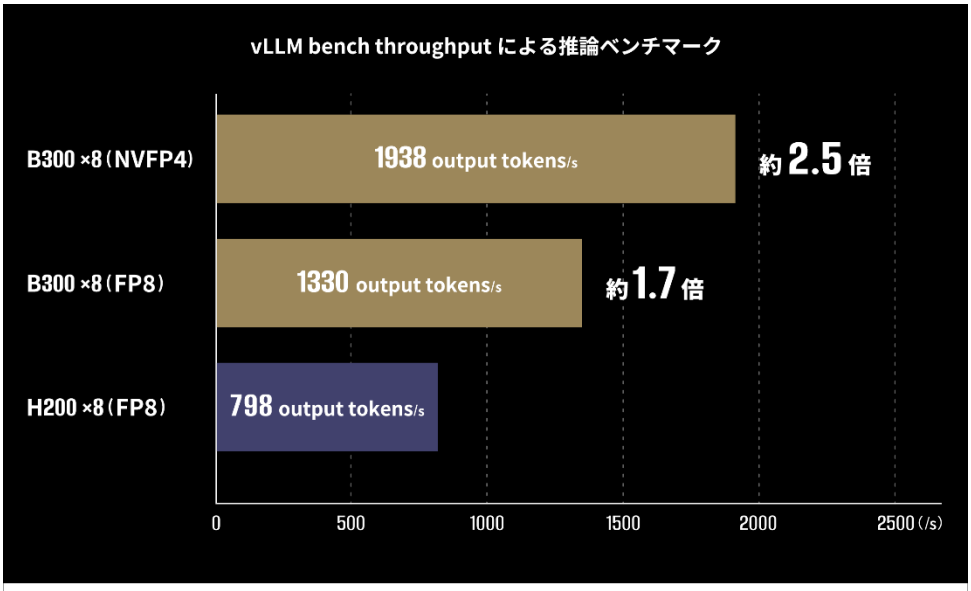
（※3）本稿で記載している結果は 非公式(Unverified)であり、MLCommons Associations に提出し、審査・承認を受けた公式結果ではありません。

（※4）クロスエントロピーとは、AI モデルの予測精度を測定する指標。モデルの予測と正解データの差異を数値化したもので、値が小さいほど学習が進み、高精度なモデルであることを示す。

（※5）FP4（4 ビット浮動小数点演算）とは、データを 4 ビット(従来の半分)で表現する演算方式。メモリ使用量を削減し処理速度を向上させることで、AI モデルの学習・推論を高速化します。NVIDIA Blackwell アーキテクチャから新たに対応した技術。

（※6）FP8 hybrid とは、8 ビット浮動小数点演算と高精度演算を組み合わせた混合精度学習手法。

2. vLLM bench throughput(※7)による推論ベンチマーク



本ベンチマークでは、vLLM の Offline Throughput Benchmark(※8)を用い、Llama-3.1-405B-Instruct モデルの推論スループットを測定しました。LLM 推論のバッチ処理における「H200 GPU」および「B300 GPU」が 1 秒あたりに生成できる出力トークン数（output tokens/s）の最大処理能力を比較しています(※9)。評価は 1 秒あたりの出力トークン数（output tokens/s）を指標としています。このベンチマークにおいて、「H200 GPU」（FP8）構成では 798 tokens/s であったスループットが、「B300 GPU」（FP8）構成では、約 170%（約 1.7 倍）の 1330 tokens/s まで向上しました。さらに、FP4（NVFP4）を適用した構成では 1938 tokens/s を達成し、「H200 GPU」構成に対し約 250%（約 2.5 倍）の性能向上を確認しました。この結果から、FP4 の活用が大規模モデルの推論パフォーマンスを向上させるための、有力な手段の一つであることがうかがえます 【表 2】。

【表 2：大規模言語モデル（LLM）推論スループット比較】

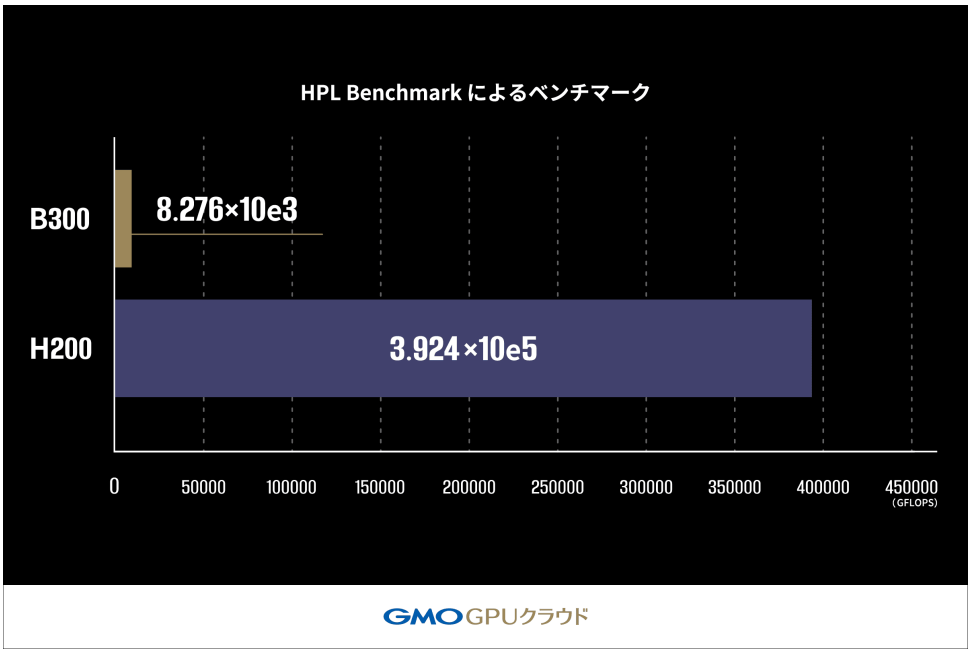
構成	GPU 数	精度	Throughput (output tokens/s)	H200 性能比
HGX H200 1 台	8	FP8	798	-
HGX B300 1 台	8	FP8	1330	約 170%（約 1.7 倍）
HGX B300 1 台	8	FP4(NVFP4)	1938	約 250%（約 2.5 倍）

（※7）vLLM bench throughput 大規模言語モデル推論エンジン「vLLM」のベンチマークツール。1 秒あたりに生成できるトークン数（スループット）を測定することで、本番環境での AI サービスの応答性能や処理能力を評価。

（※8）Offline Throughput Benchmark とは、vLLM bench throughput で実行するベンチマークモードで、バッチ処理における最大スループットを測定するモード。

（※9）計測条件：モデル：Llama-3.1-405B-Instruct, プロンプト数：2048, 入力長：2048 tokens, 出力長：256 tokens。VRAM 容量に応じてシーケンス数・バッチあたりの最大トークン数は各構成で調整。

3. HPL Benchmark によるベンチマーク



本ベンチマークでは HPL Benchmark(※10)を用いて「B300 GPU」搭載機材および「H200 GPU」搭載機材の LINPACK 性能(※11)を比較しました。HPL Benchmark では浮動小数点演算性能を測定し、1 秒間に実行できる演算回数を GFLOPS (10 億回の浮動小数点演算/秒)という単位で算出し、ベンチマークのスコアとします。この値が高いほど高性能であることを示します。

その結果、「B300 GPU」搭載機材の性能は「H200 GPU」搭載機材の 2.1%（約 47 分の 1）となりました。これは「B300 GPU」が AI ワークロードに最適化された設計であることが要因であると考えられます。科学計算など高精度な演算を必要とする場面では「H200 GPU」が依然として有用であると考えられます。【表 3】。

この結果から、「B300 GPU」は低精度演算(FP4/FP8)を用いる生成 AI ワークロードに特化した設計である一方、高精度演算(FP64)が求められる科学技術計算においては「H200 GPU」が適していることがうかがえます。

これは「B300 GPU」が生成 AI に最適な低精度演算(FP4/FP8)に特化している一方、HPL Benchmark で測定される高精度演算(FP64)は「H200 GPU」の方が優れているためです。したがって、科学技術計算など高精度な数値演算を必要とする場面では「H200 GPU」が有用であると考えられます。

【表 3：HPL Benchmark 浮動小数点演算性能比較】

構成	GPU 数	精度	GFLOPS	H200 性能比
HGX H200 1 台	8	FP64 Tensor Core	3.924x10e5	-
HGX B300 1 台	8	FP64 Tensor Core	8.276x10e3	2.1%（約 47 分の 1）

■ 実施環境

	H200	B300
サーバモデル	DELL PowerEdge XE9680	DELL PowerEdge XE9780
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	第 6 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー
メモリ	2048GB	3072GB

ディスク構成	NVMe 7.68TB x4	NVMe 3.5TB x8
GPU	NVIDIA HGX H200	NVIDIA HGX B300

（※10）HPL Benchmark は スーパーコンピュータの性能評価に用いられる国際標準ベンチマーク。

（※11）LINPACK 性能とは、複雑な数式を正確に解く計算能力。わずかな誤差も許されない科学技術計算（気象予測、創薬研究等）での性能を示す指標。スーパーコンピュータの性能評価でも使用されます。

【GMO インターネット 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】

今回のベンチマーク結果は、当社が用意した環境・条件下での検証結果となりますが、「B300 GPU」と「H200 GPU」の性能特性の違いを示す一つのデータとしてご参考いただけると考えています。「GMO GPU クラウド」は、お客様の開発目的や利用用途に寄り添い、より効率的に計算資源を活用いただけるよう、技術協力を継続的に行い、AI 開発環境における技術向上に寄り添ってまいります。このような検証情報の提供を通じて、お客様の GPU クラウドサービスの選択をサポートし、日本の AI 産業の発展に貢献してまいります。

【今後の展開】

GMO インターネットは、「GMO GPU クラウド」を通じて、生成 AI 分野に取り組む企業や研究機関に向け、ワークロード特性に応じて最適な GPU クラウドサービスを選択できる柔軟な計算環境を提供していきます。

今回の性能検証結果を踏まえ、生成 AI の学習・推論といった AI ワークロードに強みを持つ「B300 GPU」と、高精度な数値計算を必要とする用途に適した「H200 GPU」を、お客様のユースケースに応じて柔軟に組み合わせてご提案いたします。単なる GPU リソースの提供にとどまらず、お客様の開発目的や利用用途に応じた環境のカスタマイズから運用最適化まで、技術面・コスト面の両面で伴走支援を提供いたします。これにより、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。

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

「GMO GPU クラウド」は、NVIDIA H200 Tensor コア GPU を搭載し、国内初となる高速ネットワーク NVIDIA Spectrum-X と高速ストレージを実装しています。

2024 年 11 月に発表された世界のスーパーコンピュータ性能ランキング「TOP500」(※12) においては、世界第 37 位・国内第 6 位にランクインし、商用クラウドサービスとしては国内最速クラスの計算基盤を提供しています。さらに、2025 年 6 月には電力効率を競う世界ランキング「Green500」(※13) にて世界第 34 位・国内第 1 位を獲得し、高性能と省電力性の両立が国際的に評価されました。加えて、2025 年 12 月には NVIDIA の次世代 GPU「NVIDIA Blackwell Ultra GPU」を搭載した「NVIDIA HGX B300」のクラウドサービス提供を開始しました。(※14)

(※12)「GMO GPU クラウド」世界のスーパーコンピュータランキング TOP500 で 37 位にランクイン（2024 年 11 月時点 https://group.gmo/news/article/9266/https://www.gmo.jp/news/article/9266/ ）

(※13)「GMO GPU クラウド」電力効率を競う世界ランキング「Green500」で世界 34 位、国内 1 位を獲得 (https://internet.gmo/news/article/50/)

(※14)「GMO GPU クラウド」「NVIDIA HGX B300」のクラウドサービスを国内最速クラスで提供開始 (https://internet.gmo/news/article/122/)

■過去参考リリース

2024年4月19日	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
2024年6月11日	生成AI向けGPUクラウドサービスにNVIDIA Spectrum-Xを国内クラウド事業者として初採用
2024年8月29日	「GPUクラウド利用実態調査」～国内利用率わずか5.4%、約9割が海外サービスを利用～
2024年9月26日	「NVIDIA H200 GPU」搭載環境の性能を実証
2024年11月13日	「NVIDIA AI Summit」でAI・ロボティクス時代のインフラ基盤とセキュリティを解説
2024年11月19日	「GMO GPUクラウド」、世界のスーパーコンピュータランキングTOP500に初ランクイン
2024年11月22日	スパコンランキングTOP500ランクインの「GMO GPUクラウド」を提供開始
2025年2月21日	NVIDIAテクノロジを搭載した高性能GPUクラウドサービス「GMO GPUクラウド」に「マルチインスタンスGPU（MIG）機能」を追加
2025年5月7日	AIロボット協会（AIRoA）の次世代ロボット開発基盤として「GMO GPUクラウド」の正式採用が決定
2025年5月12日	「GMO GPUクラウド」がチューリングの自動運転向けマルチモーダル生成AI開発基盤に採用
2025年5月14日	「GMO GPUクラウド」の追加投資決定
2025年6月11日	「GMO GPUクラウド」電力効率を競う世界ランキング「Green500」で世界34位、国内1位を獲得
2025年7月1日	GMOインターネットとマクニカ、NVIDIAで高速化された「GMO GPUクラウド」における生成AI開発と活用支援にて協業開始
2025年8月4日	GMO GPUクラウド「NVIDIA Blackwell Ultra GPU」を採用
2025年10月2日	『GMO GPUクラウド』と低遅延回線『IOWN APN』を活用した次世代分散型AIインフラの技術実証を開始
2025年11月7日	Grafanaを活用したモニタリングダッシュボード機能を追加
2025年11月10日	プライベートコンテナレジストリ機能を提供開始
2025年11月10日	GMOインターネットとCTC、GPUクラウド事業における戦略的販売パートナー契約を締結
2025年12月12日	GPUクラウドサービス「GMO GPUクラウド」 Open OnDemand による Web ポータル機能を追加
2025年12月16日	GMO GPUクラウド「NVIDIA HGX B300」のクラウドサービスを国内最速クラスで提供開始

【GMOインターネット株式会社について】

GMOインターネット株式会社は、ドメイン、クラウド・レンタルサーバー、インターネット接続などのインターネットインフラ事業と、インターネット広告・メディア事業を展開する総合インターネット企業です。「すべての人にインターネット」というコーポレートキャッチのもと、社会基盤を支える企業として安心・安全なインターネット社会の実現と、AIで新たな未来を創る価値創造に貢献し、関わるすべての方に「笑顔」と「感動」をお届けしてまいります。

以上

【報道関係お問い合わせ先】

●GMO インターネット株式会社
広報担当　福井
TEL：03-5728-7900
お問い合わせ：
https://internet.gmo/contact/press/

●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　小犬丸
TEL：03-5456-2695
お問い合わせ：
https://www.group.gmo/contact/press-inquiries/

【サービスに関するお問い合わせ先】

●GMO インターネット株式会社
ドメイン・クラウド事業本部 GPU クラウド事業部
お問い合わせ：
https://gpucloud.gmo/form/

【GMO インターネット株式会社】（URL：https://internet.gmo/）

会　社　名	GMO インターネット株式会社　（東証プライム市場　証券コード：4784）
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 社長執行役員　伊藤　正
事 業 内 容	■インターネットインフラ事業 ドメイン登録・販売（レジストラ）事業 クラウド・レンタルサーバー（ホスティング）事業 インターネット接続（プロバイダー）事業 ■インターネット広告・メディア事業
資　本　金	5 億円

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

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	持株会社（グループ経営機能） ■グループの事業内容 インターネットインフラ事業 インターネットセキュリティ事業 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（仮想通貨）事業
資　本　金	50 億円