



2026年10月02日

各位

株式会社フィックスターズ
代表取締役社長 CEO 三木 聡
(コード番号: 3687 東証プライム)
問合せ先 取締役管理本部長 堀 美奈子
(電話番号: 03-6420-0751)

フィックスターズとマクニカ、Tenstorrent製品の実機評価から AIモデルの最適化、導入・運用までを一貫支援

共同PoCで、OpenAI「gpt-oss-120b」の1トークンあたりの生成時間を31.3%短縮

パフォーマンスエンジニアリング技術のリーディングカンパニーである株式会社フィックスターズ（東証プライム：3687、代表取締役社長 CEO：三木 聡）と株式会社マクニカ（本社：神奈川県横浜市、代表取締役社長：原 一将）は、Tenstorrent製AI製品の導入を検討する国内企業向けに、AIモデルの実機評価から性能最適化、AIインフラの構築、導入および運用を共同で支援します。



また、両社の連携による技術支援の実効性を検証する取り組みとして、TenstorrentのAIワークステーション「TT-QuietBox」上で、OpenAIのオープンウェイト大規模言語モデル「gpt-oss-120b」の推論性能を解析・最適化する共同PoC（概念実証）を実施しました。その結果、デコード処理における主要ボトルネックとなっていたMoE（Mixture of Experts）

のexpert行列積を3.68倍高速化し、TPOT（応答開始後の1トークンあたりの生成時間）を31.3%短縮しました。

本連携により、企業は新しいAIハードウェアについて、製品導入を前提とせず、導入予定のAIモデルを実機上で検証できます。さらに、検証と性能最適化の結果を踏まえ、自社の用途や性能要件に適したAI基盤であるかを定量的に判断できる環境を提供します。

連携の背景

生成AIの企業利用が広がる中、機密性やガバナンス、運用の柔軟性などの観点から、AIモデルを自社が管理する環境で利用したいというニーズが高まっています。一方、企業が新しいAIハードウェアを導入するには、製品選定だけでなく、実際に利用するAIモデルの動作確認、実機上での性能検証、ソフトウェアの調整、AIインフラの構築、導入後の技術支援が必要です。

両社が提供する支援

そこで両社は、マクニカが持つAIインフラ構築・製品評価支援と、フィックスターズが持つパフォーマンスエンジニアリングを組み合わせ、実機評価から性能最適化、導入・運用までを共同で支援します。

＜両社による主な支援内容＞

- 顧客が利用するAIモデルやアプリケーションの実機評価
- 評価環境およびモデル実行環境の構築
- AIモデルの移植、動作検証および性能分析
- 処理上のボトルネックの特定
- 計算処理やデータ受け渡し方法を含むソフトウェア実装の改善
- AIモデルの推論性能の最適化
- AIインフラの構築、導入および運用支援
- Tenstorrentとの技術連携

これにより、顧客は用途や性能要件に基づいてAI基盤を比較・検証し、性能最適化の結果を確認した上で、自社に適した環境の導入を検討できます。

共同PoCの概要と成果

両社は、TenstorrentのAIワークステーション「TT-QuietBox」（Wormhole チップ8基、GDDR6メモリ合計96GB）上で、OpenAIのオープンウェイトモデル「gpt-oss-120b」の推論性能を改善する共同PoCを実施しました。

マクニカはTT-QuietBox、リモート評価環境、AIモデルの実行環境および技術支援を提供し、フィックスターズは、AIモデルの処理状況を詳細に分析しました。その上で、処理時間が長い箇所を特定し、計算処理やデータの受け渡し方法を改善しました。

その結果、デコード処理における主要なボトルネックとなっていたMoEのexpert行列積を3.68倍高速化しました。これにより、モデル全体のTPOTは144.85 msから99.49 msへ31.3%短縮されました。

あわせて、1ユーザーあたりのデコード速度は6.9 tok/sから10.05 tok/sへ45.7%向上しました。また、リクエスト全体の平均応答時間（E2EL）は25.0秒から19.3秒へ23.0%短縮されました。

今回のPoCを通じ、Tenstorrent製品上で大規模なAIモデルを分析・最適化するための技術的な知見を両社で蓄積しました。詳細な評価条件、解析手法、最適化内容および測定結果は、フィックスターズが公開する技術レポートで紹介します。

フィックスターズ パフォーマンスエンジニアリングシリーズ06:

[Tenstorrent TT-QuietBoxにおけるgpt-oss-120b推論高速化の実践](#)

顧客に提供する価値と今後の展開

本連携では、マクニカのAIインフラ構築・製品評価支援と、フィックスターズのパフォーマンスエンジニアリングを組み合わせ、AIハードウェアを提供するだけでなく、顧客が実際に利用したいAIモデルを実機上で動作させ、その処理内容を詳細に分析します。さらに、分析結果に基づいてソフトウェア実装を改善し、AIハードウェアの性能をより有効に活用できるよう支援します。

今後両社は、今回のPoCで得た知見を活用し、gpt-oss-120bに限らず、顧客が利用を希望するさまざまなAIモデルやワークロードを対象に支援を展開します。また、AI基盤の選定から性能最適化、導入・運用までを支援することで、さらなる企業のAI活用推進に貢献していきます。

両社からのコメント

本連携および今回のPoCについて、両社の代表者は次のようにコメントしています。

株式会社フィックスターズ

代表取締役社長 CEO 三木 聡

フィックスターズが得意とするパフォーマンスエンジニアリングは、ハードウェアの特性に合わせてAIモデルの性能を最大限に引き出す技術です。同規模・同性能の計算資源を用いて、

より効率よく生産性を向上させることができれば、設備投資や運用コストは抑えられ、企業のAI投資効果はさらに高まります。

当社はマクニカ、Tenstorrentとの連携を通じ、お客様のAI基盤の選定・活用を支援してまいります。

株式会社マクニカ

アルティマカンパニー バイスプレジデント 尾崎 英臣

AIモデルとAIハードウェアの選択肢が広がる中、企業には、自社の用途や性能要件に適した組み合わせを見極めることが求められています。マクニカは、フィックスターズのソフトウェア最適化技術とTenstorrentのオープンなAIプラットフォームをつなぎ、製品選定、実機評価、性能改善、導入・運用を支援します。3社の連携を通じて、お客様が新しいAI技術を安心して活用できる環境づくりに貢献してまいります。

株式会社マクニカについて

マクニカは、半導体、サイバーセキュリティをコアとして、最新のテクノロジーをトータルに取り扱う、サービス・ソリューションカンパニーです。世界26か国/地域92拠点で事業を展開、50年以上の歴史の中で培った技術力とグローバルネットワークを活かし、AIやIoT、自動運転など最先端技術の発掘・提案・実装を手掛けています。

<https://www.macnica.co.jp/>

株式会社フィックスターズについて

フィックスターズは、“Speed up your AI”をコーポレートメッセージとして掲げるテクノロジーカンパニーです。計算資源を最大限に活用するソフトウェア最適化技術を駆使し、AIモデルの推論処理と学習プロセスの両面で圧倒的な高速化を実現します。医療、製造、金融、モビリティをはじめ様々な分野で、次世代AI技術の進化を推進しています。

<https://www.fixstars.com>

報道関係のお問い合わせ

株式会社フィックスターズ 広報担当

Email: press@fixstars.com

Tel: 03-6420-0751

※本資料に記載の会社名、製品名、サービス名は、各社の商標または登録商標です。