

2026 年 7 月 8 日
株式会社トーメンデバイス

次世代 AI アクセラレータで「電力×コスト」の改善を目指す推論基盤 PoC を開始 ～LLM 時代に求められる” 高効率 AI インフラ” の実用化に向けた実証を本格始動～

株式会社トーメンデバイス（東京都港区、代表取締役 中尾 清隆、以下「当社」）は、株式会社トゥモロー・ネット（東京都品川区、代表取締役社長 CEO 李 昌珍、以下「トゥモロー・ネット」）との協業により、次世代 AI アクセラレータを活用した高効率な AI 推論基盤の実証（PoC）を開始しました。

当社は、韓国の AI 専用開発企業である Rebellions Inc.（韓国、代表：Sunghyun Park、以下「Rebellions」）が開発した AI 推論向けアクセラレータを搭載した 4U サーバーを提供し、トゥモロー・ネットの拠点に設置し、起動および動作確認を完了しています。

本取り組みでは、従来の GPU 構成に対し、「処理効率・電力効率・コスト」の 3 つの観点における優位性の検証を通じ、新たなインフラモデルの確立を目指します。

Rebellions の AI アクセラレータは、AI 推論処理に特化したアーキテクチャにより、従来の GPU 中心のシステムと比較して高い処理効率と電力性能の最適化を実現することを特徴としており、大規模言語モデル（LLM）などの推論ワークロードにおいてもコスト効率の高い運用を可能とする新たな選択肢として注目されています。

現在は本検証環境を活用し、実運用を想定したワークロードに基づく性能評価（スループット・レイテンシ等）および運用性検証を進めています。

■背景と目的

生成 AI およびデータ活用ニーズの急速な拡大に伴い、AI インフラに求められる要件は高度化・多様化しています。一方で、従来の GPU 構成では、消費電力の増大やコスト最適化の難しさといった課題も顕在化しています。

このような状況を踏まえ、当社はトゥモロー・ネットとの協業を通じて、AI 推論専用アクセラレータを活用した高性能かつ効率的な次世代 AI インフラモデルの有効性検証を進めています。

トゥモロー・ネットは、AI インフラの設計・構築から運用支援までを一貫して提供するソリューション力を有しており、本検証においても単なる環境提供に留まらず、実運用を見据えたユースケース設計、性能チューニング、運用設計までを含めた包括的な支援を担っています。

■価値訴求

- ✓ 高い電力効率が期待される設計
- ✓ インフラコスト最適化への貢献
- ✓ AI 推論処理に特化したアーキテクチャ

「生成 AI の普及で急増する推論コスト課題への有力な選択肢を検証」

「AI は”学習”から”推論”へ。推論時代のボトルネック解消に向けた次世代インフラ基盤を検証」

■主な協業内容・役割

設計から運用までを一気通貫でカバーし、実用化を見据えた検証を推進

当社：ハード提供・AI 商材提案

トゥモロー・ネット：設計～運用

本検証では、以下の取り組みを通じて、AI 推論インフラの実用化に向けた知見を蓄積してまいります。

- ・ 顧客環境を想定した実運用レベルでの適用検証
- ・ AI 推論ワークロードにおける性能・電力効率・コスト構造の評価
- ・ 商用導入に向けた具体的な構成指針および判断材料の創出
- ・ 業種・用途別の最適な AI 推論インフラモデルの検討
- ・ トゥモロー・ネットが得意とする設計・構築から運用までを見据えた一気通貫型サービスモデルの確立

本実証を通じて、商用導入にむけた具体的な構成指針の整備および業種別の最適な AI 推論インフラモデルに関する知見の蓄積を進めるとともに、次世代 AI インフラ市場における競争力強化を目指します。

■本件に関するお問い合わせ

AI 商材担当 Email : aishohin@tomendevices.co.jp

■WEB サイト

株式会社トゥモロー・ネット : <https://www.tomorrow-net.co.jp/>

Rebellions Inc. : <https://rebellions.ai/>

以上