

大規模モデルの高速推論実行に圧倒的な強さ

Tenstorrent製 生成AIインフラ

生成AI時代、「NVIDIA以外の有力なAIインフラ選択肢」として注目！

なぜ今、注目を集めている？

- 2016年設立の米国のAIサーバー企業
 - 生成AI需要の急拡大に伴い、大企業や投資家から高い関心を獲得
- 世界的半導体設計者 Jim Keller氏がCEO
 - 業界を代表するCPU・AIチップ開発者の一人として知られる
- オープンなAIインフラを推進
 - AIアクセラレータからサーバーシステムまでを開発
 - オープンソース/RISC-Vを活用した新しいコンセプトのAI基盤を提供



tenstorrent

大規模LLM時代に再構築された、新世代のAIサーバ。



Galaxy™ Blackhole

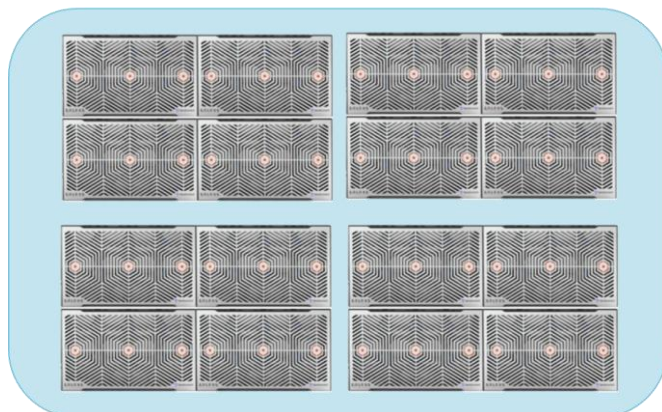
空冷型6Uサーバー/Blackhole™ ASIC x32

大きなSRAMのプール上で大規模なフロンティアモデルを稼働

1ユーザー・1トークンずつのdecodeを、Blackhole Multi-Galaxy上で極限まで速く回す

TT Galaxy Blackhole x16台(SC16)

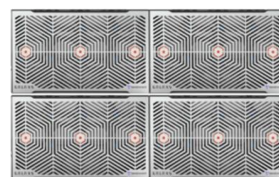
6.2GB SRAMx16台=99.2GB



400+ tokens/sec/user 級の高性能

対応モデル：GLM5.2(753B)など

TT Galaxy Blackhole x4台(SC4)



	初期実装
HW構成	SC4
動画形式	15秒間, 768P, 24fps
処理時間	60秒
状況	実装済

※DGX H100では同条件処理時間は200秒程度

対応モデル：MiniMax H3 など



Tenstorrent Galaxy™ Blackhole

アーキテクチャ優位性

- 高価なHBMではなく、一般的なGDDR6メモリを採用
- 独自ファブリックではなく、標準イーサネットを使用
- 高密度SRAMを単一ダイに統合し、データ計算ユニット近くに保持
- コンパイラからカーネルまで完全にオープンソースなソフトウェア

主要仕様（Galaxy Blackhole™ サーバー1台あたり）

項目	仕様
製品形態	スケールアウト型AIサーバー
筐体サイズ	6U ラックマウント
搭載AIプロセッサ	Blackhole® ASIC × 32基
Tensixコア総数	4,480コア （140コア × 32 ASIC）
AI性能（FP8）	最大23 PFLOPS
AI性能（BLOCKFP8）	約21 PFLOPS
DRAM総容量	1 TB GDDR6（32GB × 32 ASIC）@16 TB/s
SRAM総容量	6.2GB @2900 TB/s
ホスト構成	デュアルAMD EPYC™ 第4世代CPU
システムメモリー	1 TB DDR5
ストレージ	NVMe SSD（容量はSKUにより異なる）
ネットワーク（外部）	800GbE × 32ポート（QSFP-DD）
ネットワーク（内部ファブリック）	Ethernetベース統合ファブリック
冷却方式	空冷／液冷オプションあり
電源	冗長化電源（詳細はSKU依存）
OS / SW	Ubuntu 22.04 LTS / TT-Metalium・TT-NN・TT-Forge（オープンソース）
GA（一般提供開始）	済
想定用途	生成AI動画／LLM推論（プリフィル＆デコード）／大規模AI推論基盤

本製品に関するお問い合わせ先
株式会社マクニカ アルティマカンパニー 松本 仁兵
Email: matsumoto-j@macnica.co.jp
Mobile: 080-6615-8449