v.05 White Papers

rebellions_

ATOM™-Max: 大規模なAI推論向けの性能を革新する。

5月 02, 2025

AIワークロードが徐々に複雑になり規模が大きくなるにつれ、既存のGPU基盤のシステムは**効率性と持続可能性**の両方でさらに大きな限界に直面しています。過度な電力消費とインフラへのニーズにより、大規模なデーターセンターの長期的な運営において、**性能および費用面でボトルネック**が発生しているからです。

ATOM™-Maxはこのような限界を解消するために、大規模な推論に特化したアーキテクチャで設計されており、従来の加速器に比べ**より高い効率性と拡張性を提供**しています。これにより、最も高度なエンタープライズおよびデーターセンター規模のAIワークロードでも、高い稼働率とスループットを保証できます。また、**電力効率の最適化と炭素削減のニーズ**にも応えられる持続可能なインフラを構築していきます。

拡張可能な高効率の推論

ATOM™-Maxは大規模なAI推論に特化した構造で、優れた演算性能と高帯域幅メモリへのアクセス性能を兼ね備えています。

- **128 TFLOPS (FP16)、最大1024 TOPS (INT4)**の性能で、LLMとエンタープライズ規模のAIワークロードを安定的に処理します。
- **PCIe Gen5 x16基盤のカード間直接通信（card-to-card communication）**により低遅延と高速のデータ交換ができ、ノード間の水平拡張をしやすくします。

大規模な推論環境向けに設計されたATOM™-Maxは、**高い稼働率、予測可能なレイテンシ、簡単なデプロイ**を保証し、今日のデーターセンターのインフラとも完全互換できます。

総保有コスト（TCO）削減の中核

大規模な推論環境では、**スループットや電力効率（Performance-per-Watt）**、デプロイの柔軟性とのバランスが欠かせません。既存のGPUシステムは高いCapEx·OpExを求めるため持続的な拡張が難しいのが現状です。これに対し、**ATOM™-Max (RBLN-CA25)**は運営環境でも**Tokens-per-Second per Watt (TPS/W)**基準でL40Sを上回る性能効率を提供することができます。

データ・メモリの管理を密に統合して**ハードウェアの稼働率を最大限**に引き上げ、アイドルオーバーヘッドと資源ロスを最小限に抑えます。 このような**アーキテクチャの効率性はTCOの削減効果につながり**、デプロイの規模が大きくなるにつれ、このメリットはさらに大きくなります。

ATOM™-Max：スケーリング可能な電力効率

既存のインフラは電力効率やメモリのボトルネック、デプロイのコストなどさまざまな課題に直面しています。ATOM™-Maxはこうした構造的な問題を解決します。

より大規模なスケール、さらに大幅な削減　—　高集積の演算効率

350Wの電力の予算内で

- **128 TFLOPS (FP16)**
- **1024 TOPS (INT4)**
  性能を提供し、ラック当たりのスループットを最大化します。

同じワークロードを行うために必要なサーバー数を減らすことで、**インフラコストを削減しシステムの効率を向上**させます。

ワークロードが大きくなればなるほど、**規模の経済（economies of scale）**の効果も大きくなります。したがって、ATOM™-Maxは大規模なAIインフラに最適な選択肢といえます。

ハードウェアーソフトウェアの共同最適化

ATOM™-Maxは**ハードウェアとソフトウェアの協調最適化（Co-optimization）**により、**高レベルの同期化および共有メモリ（SHM）**の構造を採用しており、メモリ効率と稼働率を最大限に引き上げます。コンパイラはAIモデルを**ATOM™-Maxのアーキテクチャに最適化した実行コマンドに自動変換**し、レイテンシと演算オーバーヘッドを最小限に抑えた状態で、最高の性能を実現できます。

[ATOM™-Max加速器を搭載した高集積のAIサーバー]

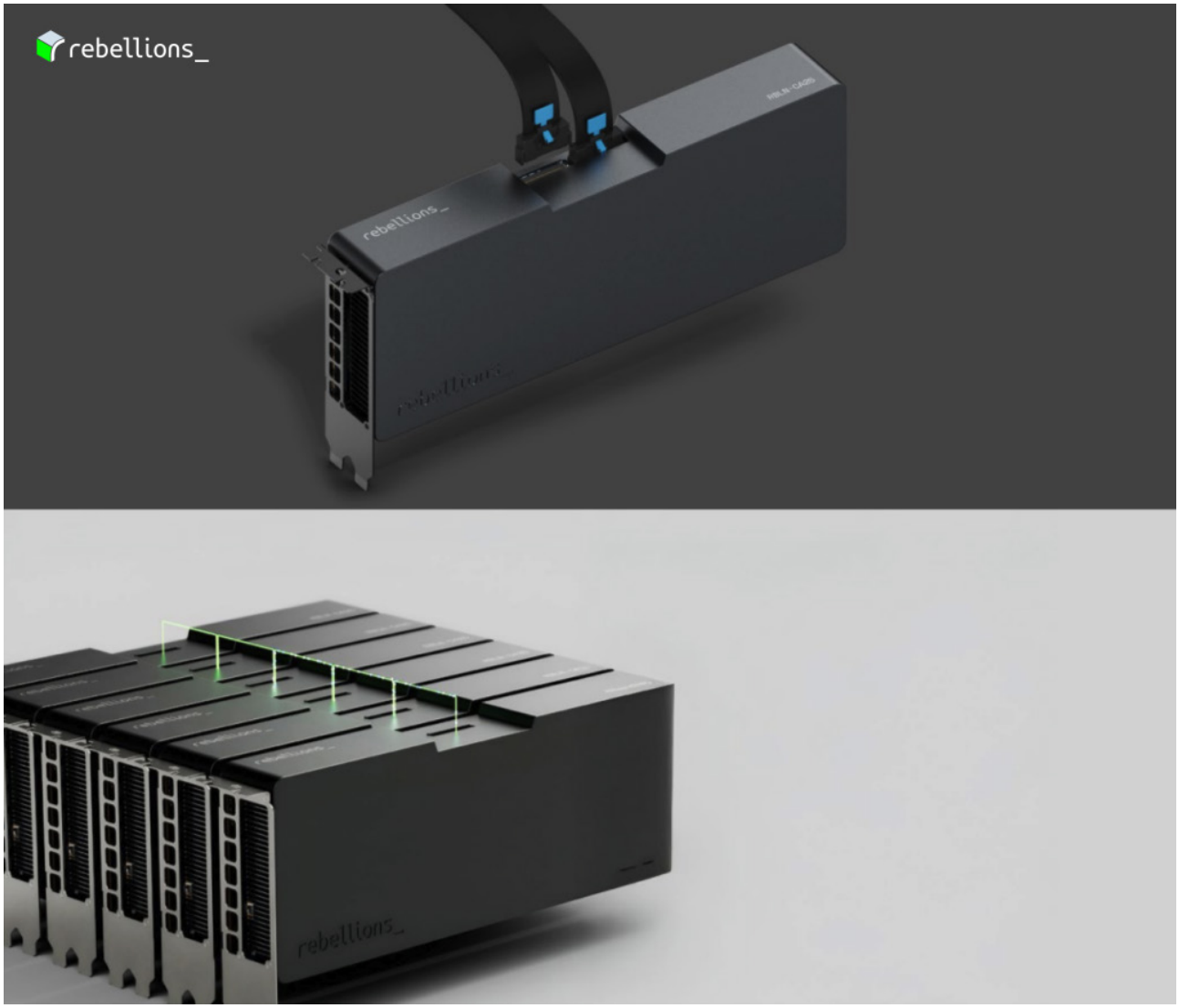
次世代AIサーバーのアーキテクチャ

高帯域幅メモリ

64GBの高帯域幅メモリ（HBM）と**1024 GB/sのメモリ帯域幅**を提供し、LLMおよび大規模なAI推論に必要な**データ処理速度と容量**を同時に確保します。最適化したデータフロー設計でボトルネックを解消し、スループットを最大限に引き上げます。

円滑な拡張性

ATOM™-Maxは単一カードの性能を超え、最大8枚まで多重カードの拡張（Multi-card Scaling）に対応できます。専用の高速インターカードコネクタ（High-speed intercard connector）により、カード間のデーター転送遅延を最小化するため、大規模なAI演算環境でも優れた拡張性能を発揮できます。



[マルチカードの拡張向けの高速インターカードコネクタ]

結論

AI推論への需要が学習（Training）より多い時代に、既存のGPU中心のインフラは徐々に非効率になっています。高い電力費用とインフラ負担、運用制限が成長の妨げとなっています。**ATOM™-Max**はこうした構造的な限界を解決する**専用のAI推論加速器**であり、以下のメリットを提供しています。

- より高い演算密度で効率的かつ大規模な推論
- 低い電力消費で**TCOおよび環境負担を軽減**
- 次世代AIモデル向けの**円滑な拡張性**

電力効率がよく高集積設計を採用した**ATOM™-Max**は、持続可能性と優れた費用対効果を両立し次世代AIインフラの拡張において欠かせない存在です。