rebellions_

REBEL-Quad

Peta-Scale MoE Inference. Without the Energy Tax.

- **最新のLLMを高いユーティリゼーションで且つ低消費電力で稼 働させられるよう設計されています。**
- **統合された混合精度コア、高反応なDMA(Direct Memory Access)のスケジューリング、及びUCIeインターコネクトが実装されています。**
- **ラックの大規模実装、モジュールの柔軟性により商用デプロイを容易に可能にします。**

REBEL-Quadは、高効率、拡張性の高い推論演算、シームレスなデプロイを大規模に行えるよう設計された世界初のUCIe-AdvanceのAIアクセラレータカードです。
コアは、モジュラーな4つの同一のChipletを1つの広帯域UCIe-Advanceリンクのユニットに融合されており、1ノードから複数ラックのクラスタまでにシームレスに拡張できます。

Accelerator Card Specifications

Architecture	4-homogeneous-chiplet SoC based on UCIe-Advanced
Compute (Dense)	1,024 TFLOPS (FP16) 2,048 TFLOPS (FP8)
External Memory	HBM3E 144GB 4.8TB/s
Chiplet (UCIe-A) Interconnection	16Gbps 1TB/s per channel

Host Connection
2x PCIe Gen5 x16
Power Consumption
Up to 600W
Software
Native-support of PyTorch 2.x, vLLM and Triton vLLM PyTorch 2.x Triton

ここに提供されたすべての情報は「現状有姿」で提供され、Rebellionsはいかなる保証や確約も伴いません。Rebellionsは、この情報の利用または信用に対するいかなる責任も負いません。

SoC Architectural Pillars

統合された演算エンジン: 全ての演算を1つのコアで実現

オペランド毎のフレキシビリティを持つ混合精度のコアは、FP8やFP16を統合されたパイプラインでサポートします。これによりインストラクションのスイッチングを不要にし、演算密度を2.8倍にします。

高反応なDMAスケジューリング : メモリアクセスへの遅延を無くします

コンパイル済みのDMAエンジンが、コマンドレベル帯域のKVキャッシュアクセスを実現し、QoSコントロールを可能にします。これにより、約3.3倍高い単位コアへの帯域を提供し、32kトークンのデコードをスムースに行うことができます。

モジュール構造のUCIeインターコネクト : 複数のダイを1つに統合

4つのChipletを1チャンネル1TB/sのUCIe-Advanceリンクで接続し1つのシステムに統合しています。このインターコネクトはチップの外にも拡張でき、ラックスケールでの分散演算も可能にします。

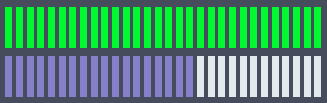
統合された同期 : Peta規模のワークロードのオーケストレーションに最適

ハードウェアによってアクセラレートされた同期ファブリックは、スパースモデルやMoEモデルでもそれぞれのchiplet上での実行を同期することができます。これにより精度の高い並列演算を実現し、Expertのルーティングの際に起こる遅延を回避することができます。

REBEL-Quad vs. H200 (Llama 3.3 - 70B)

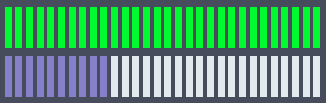
Throughput
(TPS)

~x1.6



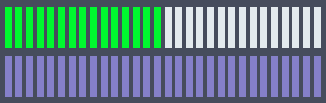
Efficiency
(TPS/Watt)

~x3.2



Power Consumption
(Watt)

~x0.5



Benchmark Condition: Performance measured on Llama 3.3 70B (TP2, FP8) with runtime input/output length 2048/2048.

REBEL-QuadはAIインフラのTCOを根幹から変革します。

弊社内の評価によると、大規模LLMの実行において演算とメモリアクセスのボトルネックを解決することでREBEL-QuadはNVIDIA社のH200クラスのGPUよりも高いスループットを実現しました。I/Oのダイを含んだchiplet開発が既に進んでおり、将来的には1兆パラメータのモデルやマルチノードのエクサスケールのデプロイを可能にします。

免責事項：全ての性能評価データはRebellions内における評価に基づいております。GPUに関するデータは、公開データ及び第3者によるベンチマークによるデータに基づいております。実際の性能・評価データは、実行する条件やソフトウェアのバージョンなどにより異なる場合があります。このデータは参照用であり、事前通知なく変更することがあります。