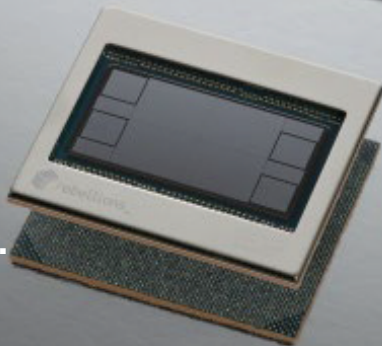


# Rebel100™

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UCleインターコネク트가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。

Rebel100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UCle-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UCle-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

<b>Architecture</b>	4-homogeneous-chiplet SoC based on UCle-Advanced	<b>Host Connection</b>	2x PCIe Gen5 x16
<b>Compute (Dense)</b>	1,024 TFLOPS (FP16) 2,048 TFLOPS (FP8)	<b>Power Consumption</b>	Up to 600W
<b>External Memory</b>	HBM3E 144GB 4.8TB/s	<b>Software</b>	Native-support of PyTorch 2.x, vLLM and Triton
<b>Chiplet (UCle-A) Interconnection</b>	16Gbps 1TB/s per channel		

# 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ズ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

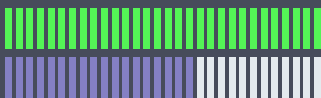
ハードウェアによってアクセラレートされた同期ファブリックは、スパースモデルやMoEモデルでもそれぞれのchiplet上での実行を同期することができます。これにより精度の高い並列演算を実現し、Expertのルーティングの際に起こる遅延を回避することができます。

## Rebel100 vs. H200

(Llama 3.3 - 70B)

Throughput  
(TPS)

~x1.6



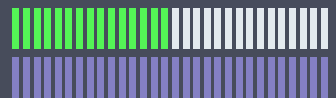
Efficiency  
(TPS/Watt)

~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.

## Rebel100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Rebel100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。