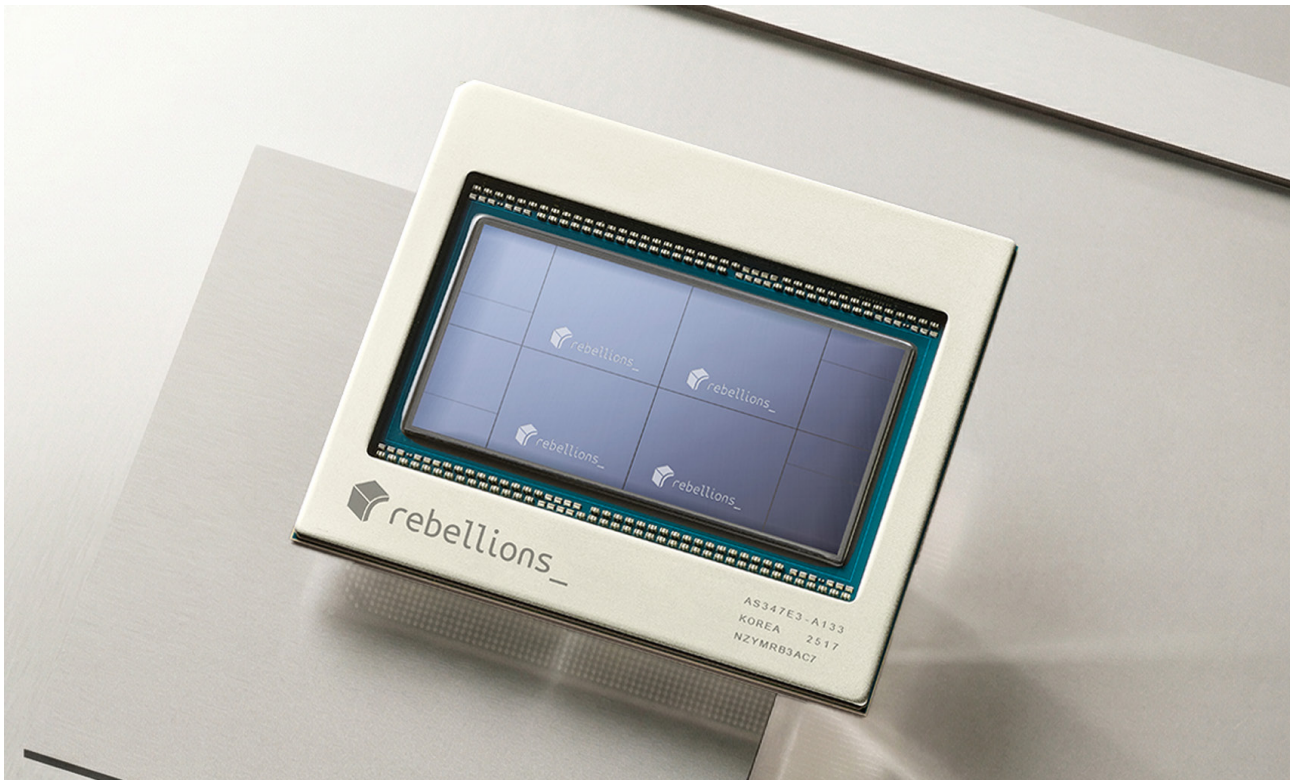
大規模なAIサービング向けのペタスケール SoC：REBEL-Quad

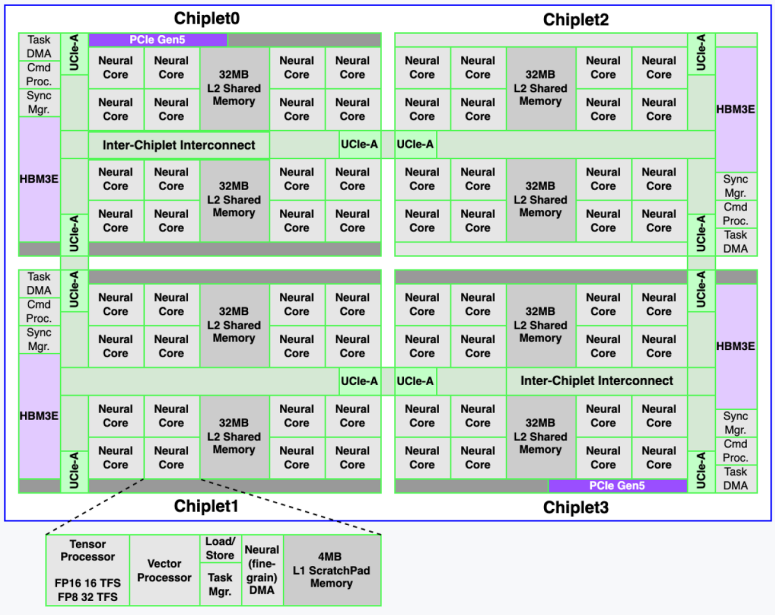
5月 02, 2025



REBEL-Quadは、超大型LLM（大規模言語モデル）のサービング向けに設計された**UCIe-Advancedチップレットアーキテクチャ基盤のAI SoC**です。ハイパースケーラー、AIデータセンター、エンタープライズ環境で求められる極めて高い演算性能とメモリ帯域幅の要求を満たすことを目的としています。

[大規模なAIサービング向けのペタスケールSoC]

REBEL-Quadはハードウェアとソフトウェアが完全統合したスタックを基盤とし、演算集約的なプレフィール（Prefill）段階とメモリ集約的なデコード（Decoding）段階の両方において、最大の稼働率と優れた電力対比性能（Performance-per-Watt）を提供しています。また、チップレット基盤の設計を通して、低遅延と一貫性（coherence）を確保しながら極めて高い拡張性を実現しています。



[Figure 1. 四つの同じチップレットで構成されたREBEL-Quadブロックのダイアグラム]

LLM推論向けのアーキテクチャ的アプローチ

REBEL-Quadは大規模AI推論で発生する**エネルギー効率および拡張性の問題を根本的に解**

決するために、構造的なソリューションを提供しています。

統合型混合精度エンジン

- FP8 / FP16 / FP32 演算を単一コアで統合的にサポート
- 従来比2.8倍の演算密度（Compute Density）を達成

予測型DMAとオンチップメッシュ

- ソフトウェアに最適化したDMAで、2.7 TB/sの有効メモリ帯域幅を確保
- 次世代メッシュファブリックを基盤とし、従来比3.3倍高速なコア間通信を実現

REBEL-Quadの優れた電力効率は、独自に設計したIPコアで実現されました。高帯域幅のオンチップメッシュインターコネクト（On-Chip Mesh）は、全てのコア間通信をリアルタイムでつなげるため、データフローのボトルネックを防ぐことができます。

広域同期化

- ハードウェア加速基盤のP2Pおよび階層型通信
- 分散ワークロード環境においても、低遅延で同期が取れる実行パスを実現

カスタムD2Dプロトコル

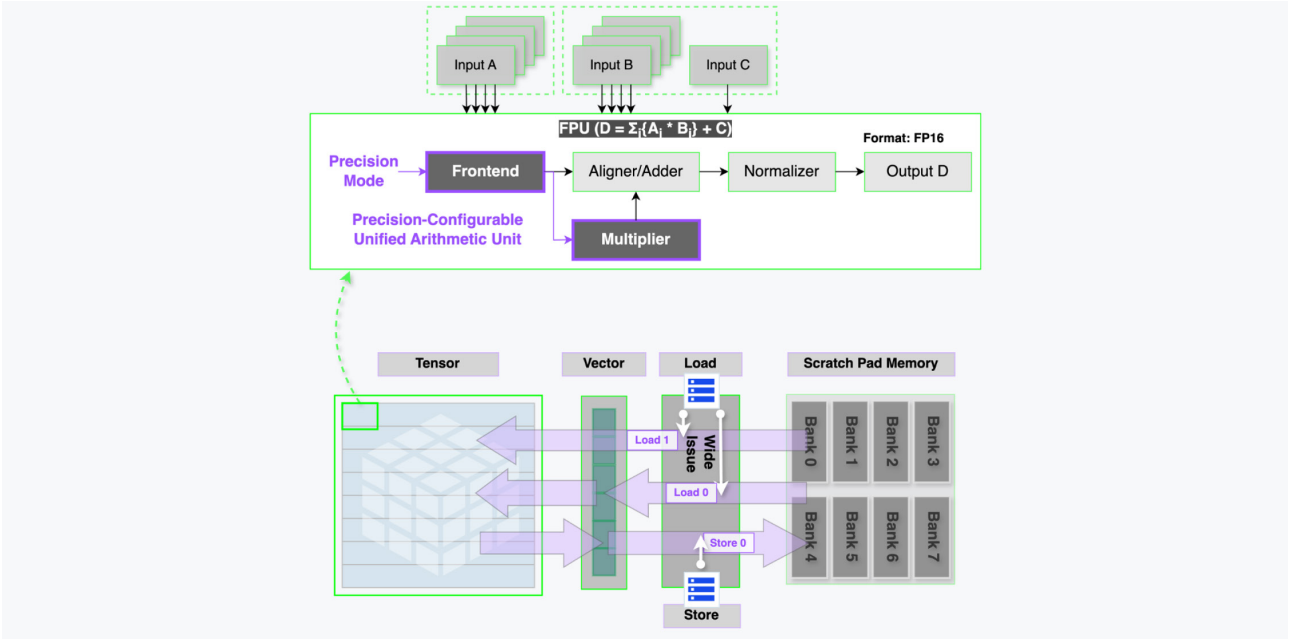
- **チャンネル当たり1 TB/sの双方向帯域幅、チップレット間11nsレベルの超低遅延**
- モジュール型の拡張性を維持しつつ、多様なチップレット構成に対応

REBEL-Quadのソフトウェアスタックは、ハードウェアと密に結合されており、**チップレット間の専用プロトコルを基に、ハイパースケーラとAIデータセンターの超大型推論負荷を安定的に処理**できます。

統合混合精度演算 & チップレット単位の拡張性

既存のNPUはFP8、FP16、BF16など精密別に独立した演算ブロックを利用していたため、**面積の非効率とデータフロー管理の複雑性**が生じました。これに対し、REBEL-Quadは**オペランド単位（operand）別に精度を設定できる統合算術エンジ**ンを搭載し、機能ブロックを追加しない**単一演算構造**を実現しました。

これにより、既存対比2.8倍高い演算密度とハードウェアレベルのWide-Issue実行が可能となり、コマンド依存性を減らしました。



[Figure 2. 統合した多層／混合精度演算コア]

Wide-Issueメカニズムはテンソルおよびベクターコア間のメモリ帯域幅をバランスよく配分し、レジスターとスクラッチパッドのメモリへのアクセスを同時に行います。これは、**FP8スループットが重要なLLM推論のプレフィル段階**で特に有利です。REBEL-Quadは**4台のチップレットパッケージ内で2 PFLOPS（FP8）**性能を達成し、単一ノードレベルでも優れた**性能対電力効**率を確保しています。

予測型 DMA & 高帯域幅メモリのアクセス

LLMのデコード段階ではKVキャッシュメモリの帯域幅がボトルネックとなり、コンテキストウィンドウが長くなるほど、処理効率が急激に低下します。これを解決するために、REBEL-Quadは予測型（Predictive）DMAエンジンを採用しています。このエンジンはソフトウェアで構成でき、以下の特徴があります。

- **2.7 TB/sの有効メモリ帯域幅**
- ローカルおよび遠隔HBMに同時にアクセス
- マルチパスのルーティングで帯域幅インターリービングを実現

このDMAエンジンは、REBEL-Quadのカスタマイズ型のメッシュインターコネクトと高度に統合され、**既存対比3.3倍高いコア別の帯域幅**を提供しています。また、**タスク別のQoSに対応**するため、ワークロードで遅延による変動を最小化できます。

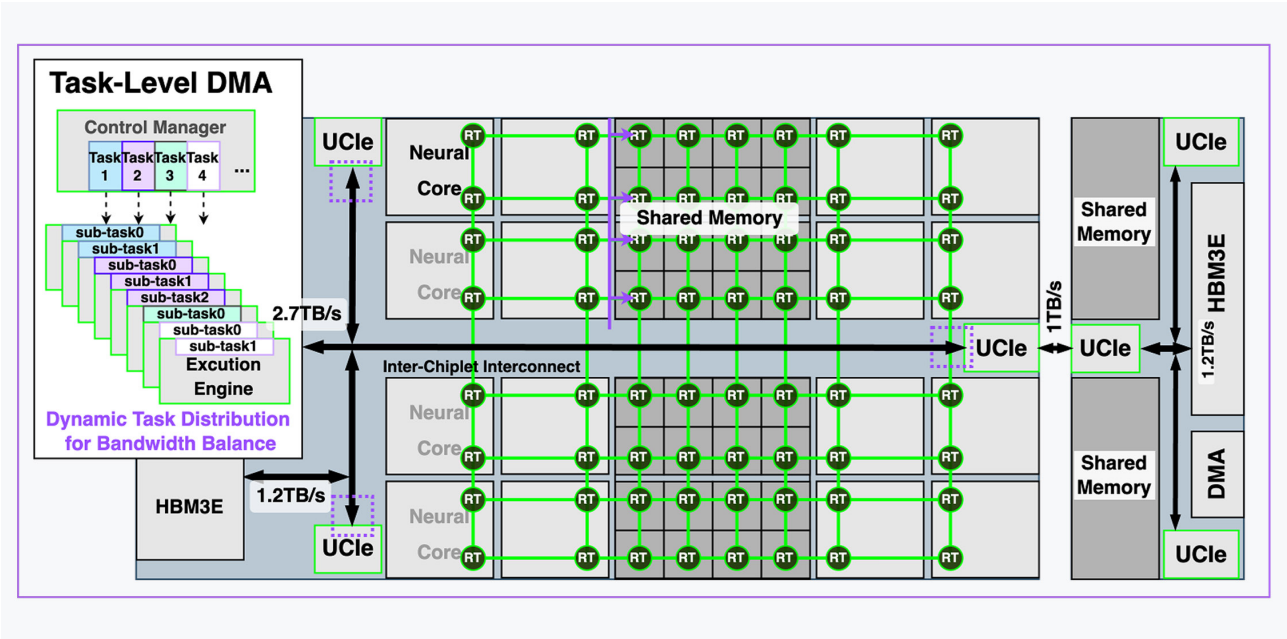
階層同期およびP2P通信

REBEL-Quadは複雑なアテンションパターンと長期的な依存性を持つモデルでも、一貫した性能を維持するために、**全体のチップ単位の同期化および通信構造**を提供しています。

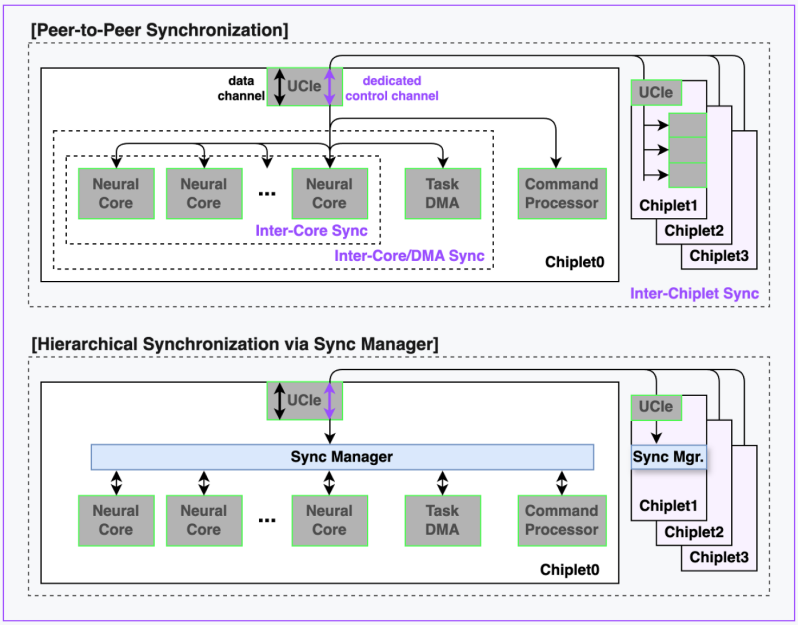
主な構成は以下の通りです。

- メッシュネットワーク全体の**制御信号専用の仮想チャンネル**
- 実行パスを制御する**中央集中型の同期マネージャー**
- コア、DMA、同期化ユニット間の**ハードウェア加速型のP2P通信**

この**階層型通信プロトコル**はチップレット内部（intra-chiplet）とチップレット間（inter-chiplet）の依存性を解決し、プレフィルとデコードが同時に行われる状況でも、稼働率を最大限に維持できます。その結果、既存のアーキテクチャの同期化で起きるボトルネックを解消し、ソフトウェアのオーバーヘッドを最小化しました。これにより、**高い演算密度、最大の稼働率、最高水準の電力効率**を実現しています。



[Figure 3. ニューラルエンジンとDMAエンジンを活用したフルチップのデータ転送構造]



[Figure 4. チップレット別の中央同期マネジャーを利用したフルチップP2Pおよび階層的な同期化方式]

モジュール型拡張向けのD2Dプロトコール

REBEL-Quadのモジュール型拡張はUCIe-Advanced基盤のカスタムD2D（Die-to-Die）プロトコールで実現されました。

- チャンネル当たり**1TB/sの双方帯域幅**、11nsチップレット間の全体パスの遅延
- チップ間**Load-Storeメモリへのアクセス**に対応
- 今後の**Scale-Up / Scale-Out構造の拡張**に備えた柔軟性を確保

このインターコネクトは、多層チップシステムを**仮想単一システム（virtually monolithic unit）**で統合し、モジュール型の拡張性を維持しながら未来のシステム設計にも対応しています。各チップレットは**３つのUCIeチャンネル**でつながり、**トポロジーに基づいたダイの回転配置（Die Rotation）**によって水平メッシュの連続性を確保できます。

また、信頼性の高い運用のために**スイッチネットワークおよびリアルタイム**のデバッグメカニズムが搭載されており、大規模なAI推論環境でも安定的かつ無誤差（Zero-Error）実行が可能です。将来的には、**I/O およびメモリ拡張チップレット（Expander Chiplet）**により、再設計することなくシステム構成を拡張できるようにする予定です。

REBEL-Quadは次世代LLMサービングに求められる**性能、効率性、拡張性**をすべて満たしています。チップレット基盤のモジュール型構造で、**柔軟なアップグレードと長期的な拡張性**を提供しています。また、**ハイパースケーラとエンタープライズAIシステムに最適な基盤**となることを目指していきます。