v.01 White Papers

rebellions_

ATOM™:
GenAI向けの最適なソリューション

7月 11, 2024

序論

生成AI（Generative AI、GenAI）があらゆる産業に変革をもたらす中、大規模な演算処理に特化したハードウェアの開発は、もはや選択ではなく必須となりました。そのため、AIワークロードに特化したAI加速器またはAIチップはコア技術となっていますが、効果的なAIチップを設計するには多くの課題が伴います。

速い処理速度とスループット（throughput）は、AIアプリケーションの性能に直接影響を与えます。大規模な演算を処理する代表的な方法の一つは、バッチ処理（batch processing）で、 多数のタスクをグループ化して連続的に実行する手法です。しかし、この方法では、スループットが増える代わりに、その分処理時間が増えます。

システム全体の性能を向上させるうえで、メモリと演算タスクのバランスは極めて重要です。しかし、柔軟性（flexibility）は重要でありながら見落とされることが多い要素です。柔軟性とは、システムレベルでメモリと演算タスクのバランスをとる能力を指します。例えば、テキスト基盤の大型言語モデル（LLM）推論は、膨大なパラメータを処理するのに、頻繁なメモリアクセスが必要なメモリ集約型タスクに分類されます。一方、テキスト－動画アプリケーションは、リアルタイムでグラフィックとデータを処理することから、演算集約型タスクと位置付けられます。結論から言うと、最適なAIチップはさまざまなアプリケーションに対応するために、レイテンシやスループット、柔軟性のバランスをとらなければなりません。

演算効率の最大化

リベリオンはATOM™を設計するにあたり、演算の高効率を最優先課題に捉え最適な均衡点を見出そうとしています。多様な機能を果たせるように、再構成が容易なCGRA（Coarse-Grained Reconfigurable Array）アーキテクチャを採用し柔軟性を最大限に引き上げました。また、アイドルリソースを最小化しながら、タスクを持続的に処理することで処理時間を減らしました。

柔軟なアーキテクチャに加え、ATOM™の同期化メカニズムは並列処理に必要なリソースだけを効率的に有効化します。これにより、処理準備にかかる手間を省き、レイテンシが減ります。さらに、 多層のメモリ階層によりデータ依存性を低減し、帯域幅を大きく向上させます。また、リアルタイム同期化により制御依存性を抑制します。これらによりレイテンシを削減しながら、高い並列性を実現しています。

ATOM™：AI推論向けのシステムオンチップ（SoC）

リベリオンの ATOM™はAI推論向けに設計されたAI加速器で、サムスンの先端5nm プロセスで製造されました。ATOM™はFP16演算で最大32 TFLOPS、INT8演算で最大 128TOPSの性能を提供し、8台のニューラルエンジンと64 MBオンチップSRAMで性能を強化します。精密に設計されたメモリアーキテクチャは、性能と効率性を最大限に引き出します。

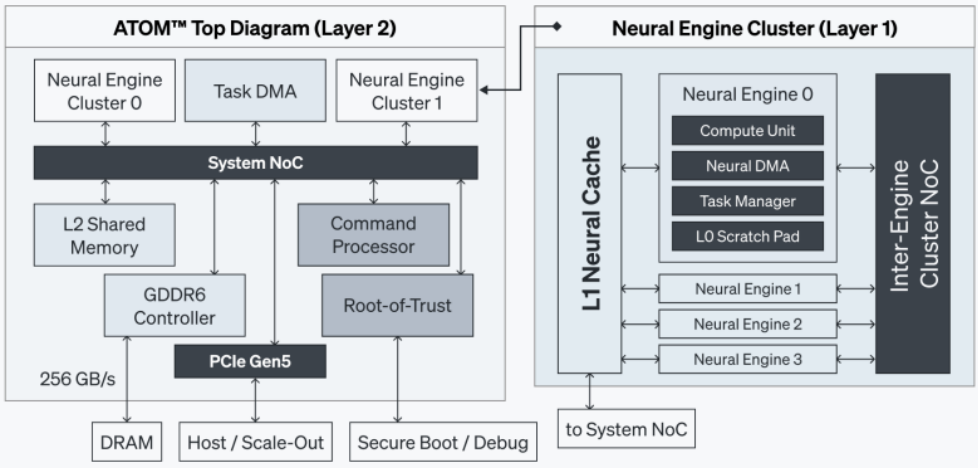
RBLN-CA12

ATOM™はRBLN-CA12というシングルスロット、FHFL（Full Height, Full Length） PCIe Gen5カードで提供されます。最大電力（Thermal Design Power、TDP）は60~130Wで、256GB/s帯域幅のGDDR6メモリとPCIe Gen5x16インタフェースを介して、ホストやカード、さらにカード間通信に対応します。また、マルチインスタンス（Multi-Instance）機能を通じて、16個の独立したハードウェア分離インスタンスで分割し、高い並列性とリソース配分を動的に管理できます。

RBLN-CA12	
AI Accelerator	ATOM™
FP16	32 TFLOPS
INT8	128 TOPS
On-chip SRAM	64 MB
External Memory	GDDR6, 256 GB/s, 16 GB
Multi-Instance	Hardware isolation up to 16 independent tasks
Thermal Solution	Passive
Mechanical Form Factor	Full Height, Full Length (FHFL) 266.5 × 111 × 19 mm
Thermal Design Power	60–130 W
Host and Card-to-Card Interface	PCIe Gen5 x16, 64 GB/s
Connectors	One CPU 8-pin power connector (2×4)
Weight	Total: 615 g

[Table 1. RBLN-CA12仕様]

ATOM™ SoC



[Figure 1. ATOM™ 多重SoCアーキテクチャ]

ATOM™は複数の中核コンポーネントを一つの集積回路に統合したマルチコアのシステムオンチップ（System-on-Chip）です。図１のように、ニューラルエンジン、コマンドプロセッサ、オンチップメモリ（SRAM）、さらにGDDR6メモリを一つのチップに集積しました。さまざまな要素を統合してチップの集積度を高め、空間の使用率と電力効率を最適化します。

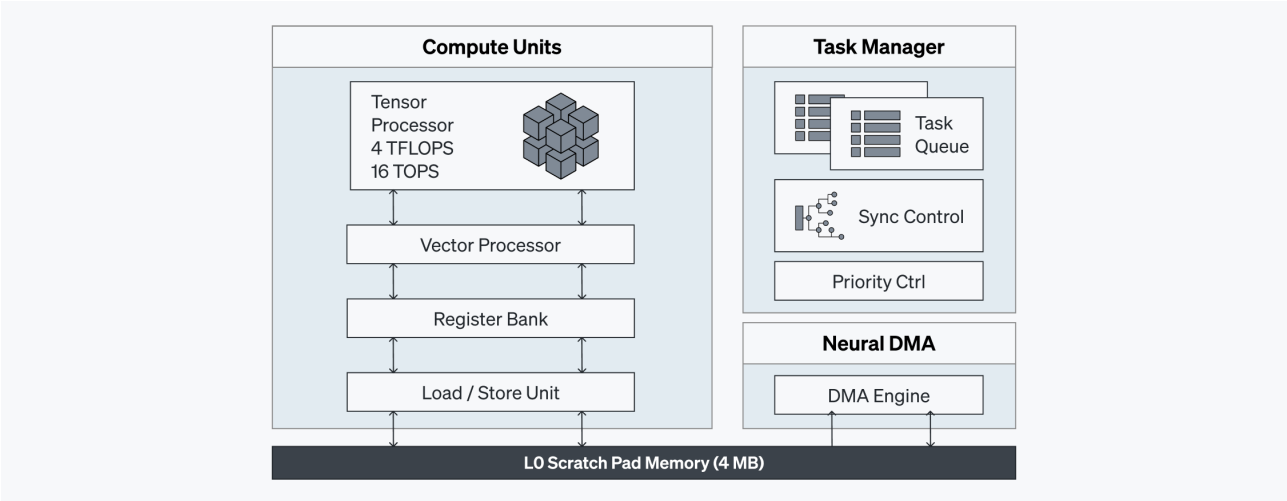
この構成だけでもコンポーネント間通信を簡素化し、レイテンシを大幅に削減できますが、さらに高帯域幅を提供するネットワークオンチップ（Network-on-Chip、NoC）を実現しました。また、複数の階層で同期できるように設計されています。

ニューラルエンジン

ATOM™のニューラルエンジンは実際の演算が行われる中核的なコンポーネントです。ニューラルエンジン内の演算ユニットは異質的なSIMD（Single Instruction、Multiple Data）およびMIMD（Multiple Instruction、Multiple Data）の要素を結合し、並列処理能力を最大限に引き出します。さらに、命令レベルでの依存関係を効率的に管理します。

4 MBスクラッチパッド（Scratch Pad）メモリが含まれた演算ユニットは、最大8TB/s速度でSRAMの中間データにアクセスでき、外部メモリへの依存性を最小化します。これにより、帯域幅の限界を補い処理速度を上げます。各ニューラルエンジンに搭載されたタスクマネージャー（Task Manager）はローカルハードウェアレベルで同期化を加速し、コマンドプロセッサ（Command Processor）と協力して稼働率を最大限に引き上げます。

このようにATOM™はニューラルエンジンの演算ユニット、スクラッチパッドメモリ、タスクマネージャーを活用して高い稼働率と低いレイテンシを実現しています。



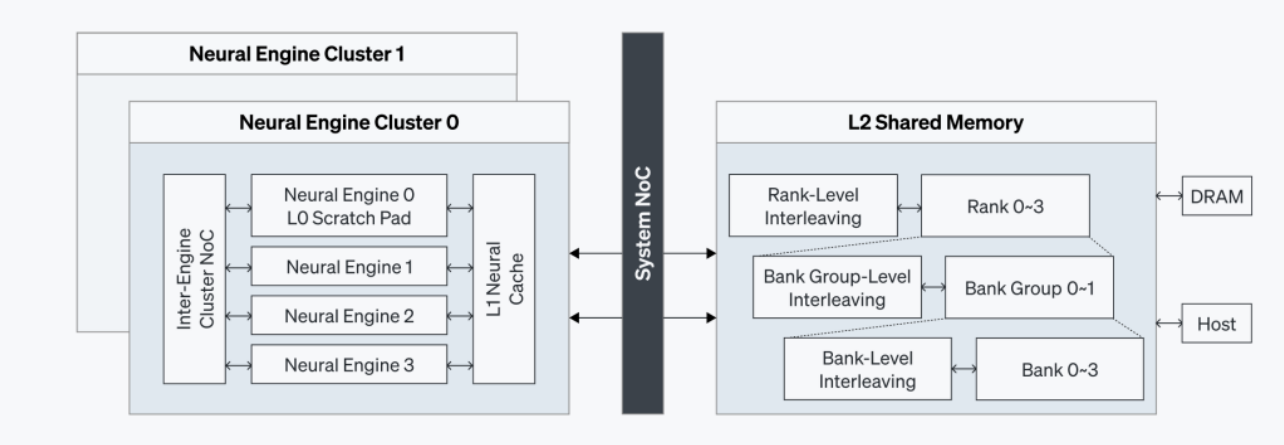
[Figure 2. ATOM™のニューラルエンジン]

ATOM™のニューラルエンジンは実際の演算が行われる中核的なコンポーネントです。ニューラルエンジン内の演算ユニットは異質的なSIMD（Single Instruction、Multiple Data）およびMIMD（Multiple Instruction、Multiple Data）の要素を結合し、並列処理能力を最大限に引き出します。さらに、命令レベルでの依存関係を効率的に管理します。

4 MBスクラッチパッド（Scratch Pad）メモリが含まれた演算ユニットは、最大8TB/s速度でSRAMの中間データにアクセスでき、外部メモリへの依存性を最小化します。これにより、帯域幅の限界を補い処理速度を上げます。各ニューラルエンジンに搭載されたタスクマネージャー（Task Manager）はローカルハードウェアレベルで同期化を加速し、コマンドプロセッサ（Command Processor）と協力して稼働率を最大限に引き上げます。

このようにATOM™はニューラルエンジンの演算ユニット、スクラッチパッドメモリ、タスクマネージャーを活用して高い稼働率と低いレイテンシを実現しています。

階層的なメモリサブシステム

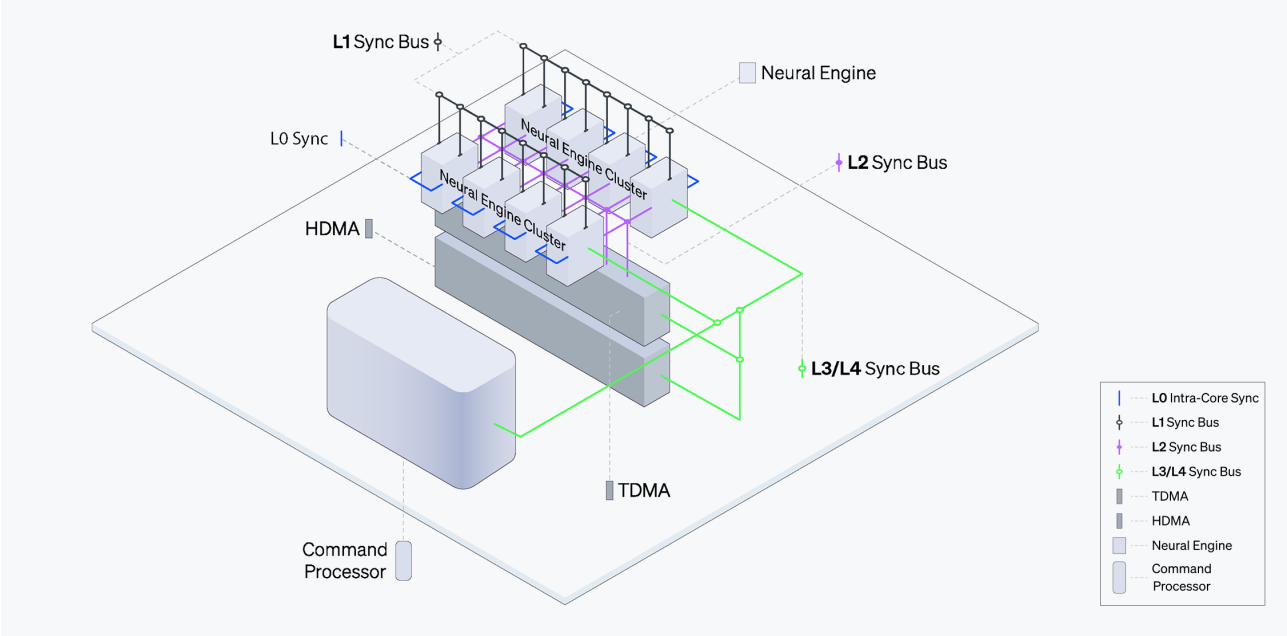


[Figure 3. ATOM™の階層的なメモリサブシステム]

ATOM™の階層的なメモリアーキテクチャは、最高水準の性能効率性を確保するように設計されています。このため、ニューラルエンジンに十分な帯域幅を提供すると同時にレイテンシを最小化します。

- **GDDR6** メモリ：16GBのGDDR6メモリを搭載し高いスループットを維持しながら、電力消費を減らします。
- **スクラッチパッド（L0）**：各ニューラルエンジンに搭載された4MBのスクラッチパッドはローカルデータへ即時にアクセスできます。
- **L1ニューラルキャッシュ：ニュ**ーラルエンジンの周りにあり、データへのアクセス速度を上げます。
- **L2共有メモリ**：64MB SRAMで、多重インターリービング技術を取り入れ、並列性に対応できます。また、帯域幅を最適化しレイテンシを減らします。

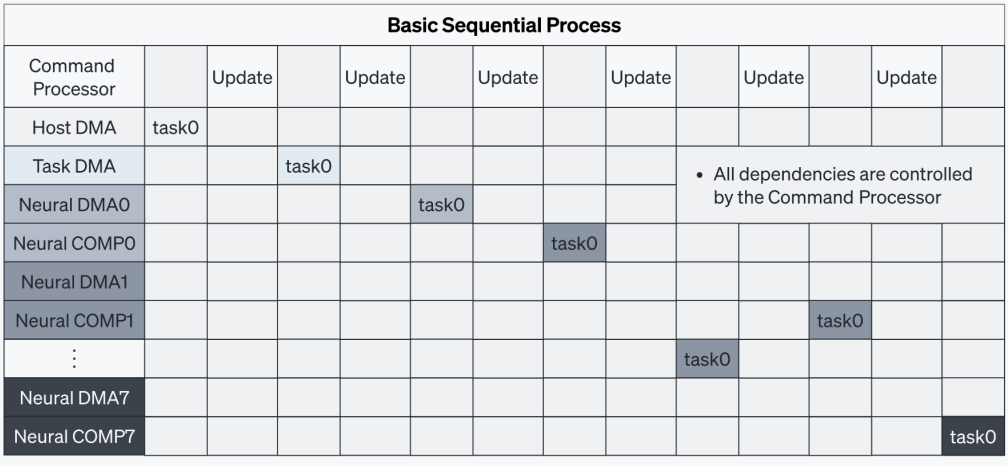
多重同期化と並列処理



[Figure 4. ATOM™の同期化構成]

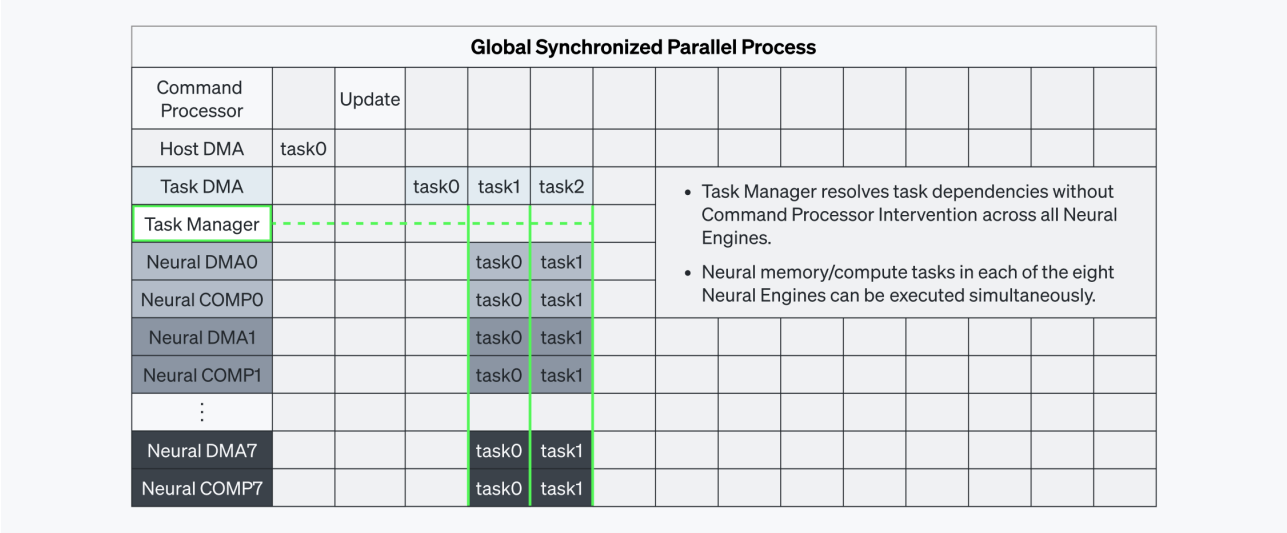
ATOM™の同期化メカニズムは、効果的な並列処理とチップの性能を拡張させます。同期化はコマンドおよびタスクレベルで行われ、コマンドプロセッサやタスクマネージャー、さらに専用のローカルバスを通じて安定的な帯域幅でタスクの流れを円滑に制御します。

ニューラルエンジンはL1同期化バス（L1 Sync Bus）を介してタスクマネージャーと通信し、このタスクマネージャーはL2同期化バス（L2 Sync Bus）を介してタスクレベルのDMAとも連携します。このような構成でシステム全体の依存性を確かめ、多様なコアを同期化することで高密度な計算タスクを実行できます。



[Figure 5-1. タスクマネージャーがない場合のコマンドの逐次実行]

コマンドプロセッサがコマンドの実行を単独で管理すると、タスクが順番に処理されるためレイテンシが増えます。タスクへの依存性が解消されるまで、コマンドプロセッサの処理を待たなければならないからです（図5-1を参照）。この方式では通信オーバーヘッドが発生しやすくなります。



[Figure 5-2. タスクマネージャーを含む並列実行]

コマンドの実行を最適化するためにリベリオンは、タスクマネージャーを採用してハードウェアレベルのローカル依存性を自動的に解消できるようにしました。図5-2でも分かるように、コマンドプロセッサが依存性を改善しなくても、各ニューラルエンジン内のDMA/COMPタスクが並列で実行できます。ニューラルエンジンのタスクマネージャーがこの依存関係を解消することで、タスクを同時に実行できます。こうした処理は、図4に示すように、専用のL1/L2データパスを通じて行われます。その結果、すべてのニューラルエンジンでタスクが効率よく調整されるため、並列実行が円滑にできレイテンシを最小限に抑えられます。

ベンチマーク

ATOM™とNVIDIA A100の性能を比較するために、代表的なAIモデルであるT5-3B（自然言語処理）とSDXL-Turbo（テキストー画像生成）を活用してテストを行いました。このような比較を通じてATOM™が最新のAIワークロードをどれだけ効果的に処理できるかを検証しました。

言語モデルのベンチマーク：T5-3B

Googleが開発したT5（Text-to-Text Transfer Transformer）はTransformerアーキテクチャを活用した革新的な大型言語モデル（LLM）です。T5モデルは、6千万から110億個のパ

ラメータ規模を備えています。

今回の比較では機械翻訳、テキスト要約、質問応答、テキスト生成といったワークロードに適した30億パラメータのモデルをバッチサイズ1で実行しました。

- **性能**：1秒当たり生成されたトークンの数で測定
- **電力消費**：ワット（W）単位で測定
- **電力効率性**：性能対比の消費電力で計算

テストの結果、ATOM™はA100と比較して最大44%高い電力効率を実現し、複雑な言語処理向けのワークロードでも優れた性能と効率性を見せました。

	Input	Output	Performance (Token/s)	Average Power (W)	Average Power Efficiency (Token/J)
ATOM™	349	512	45.0	56.1	0.80
A100	349	512	44.3	177.5	0.25

二つのテストは全てFP16の精度で行われました。ATOM™の結果は予測値に基づきます。A100の結果はHugging Face transformersのライブラリに基づきます。

テキストーイメージモデルのベンチマーク：SDXL-Turbo

Stability AIが開発したSDXL-Turboは、高解像度のイメージ生成を専門としており、既存のStable Diffusionモデルに比べ推論速度が大きく向上しました。

ATOM™はA100より遥かに少ない電力を消費しながら高性能を維持しました。つまり、少ないリソースでも優れた結果を出すため、運用コストを大幅に削減できました。さらに、サービス提供の持続可能性も高めることができます。

	Performance (img/s)	Power (W)	Power Efficiency (Performance/Power)
ATOM™	3.74	60.3	0.062
A100	7.36	192.7	0.038

*イメージのサイズ　512×512、Diffusion step：1
*ATOM™の結果は予測値です。A100の結果はHugging Face diffusersのライブラリに基づきます。

結論

業界を問わずAIへの依存度が高まる中、持続可能な拡張性を備えた最適なAIチップを求めるのは多くの企業が直面している課題です。ATOM™は柔軟性、電力効率性、高性能のバランスを取りながらも、迅速に処理できるように設計されています。革新的なニューラルエンジンと多重メモリアーキテクチャ、強力な同期化を実現しました。これにより、レイテンシと電力効率性を最適化し高い演算効率が期待できます。ATOM™はAIサービスの運用コストを飛躍的に抑え、収益性を向上できる持続可能なAIサービス向けのAIチップです。