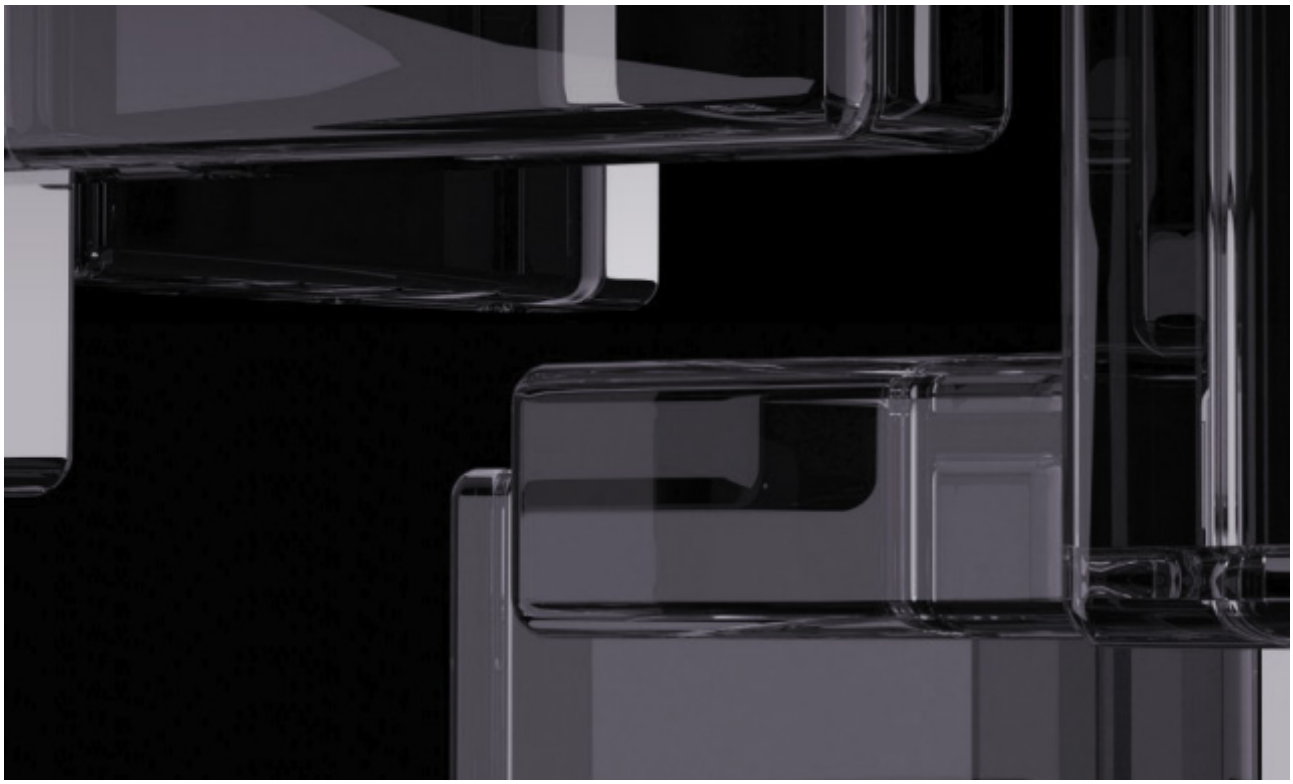
rebellions_

Rebellions Scalable Design

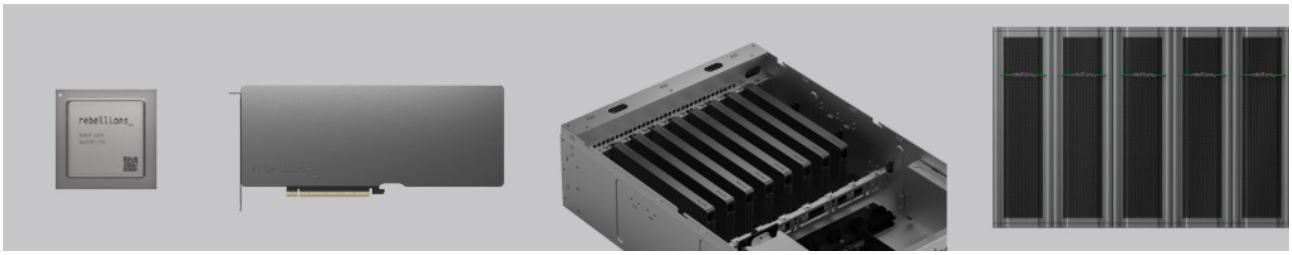
11月 15, 2024



リベリオンのエンジニアリング哲学は、拡張性とモジュール性を中心としたアーキテクチャに基づき、Rebellions Scalable Design (RSD)で実現されています。このRSDは、リベリオンの現在と未来における全製品の中核をなす基盤で、あらゆる環境で安定的な拡張性と一貫した性能を実現させる基盤を提供します。

RSDは線形的な拡張性（linear scalability）を実現するように設計されています。システムの規模が大きくなればなるほど性能もそれに比例して向上するため、効率が落ちずに拡張できます。リベリオンはこのような強みを基に、小規模なハイパースケーラ（hyperscaler）レベルの大規模な推論環境を含め、あらゆる規模のインファレンス作業に最適化した高効率なソリューションを提供しています。

大型言語モデル（LLM）をはじめ、さまざまなAIモデルに完全に対応でき、独自のソフトウェアスタックで性能と互換性を最大限に引き上げます。RSDは、小規模なCNN基盤のアプリケーションから高難度のトランスフォーマー基盤のワークロードまで、優れた性能を発揮します。



コア技術

AIモデルの規模が大きくなるほど応用の範囲もさらに広がりますが、それを実現するには、大規模な演算を複数のAIプロセッサ間で円滑に分散・同期化する高難度の技術力が欠かせません。このような課題を解決するために、RSDはテンソル処理（tensor parallelism）構造を取り入れています。これにより、大型言語モデルの演算をさまざまなプロセッサに効率的に分散でき、モデル全体が安定的に実行されます。

RBLNコンパイラはモデルを細かいレベルまで最適化する中心的な役割を担います。さらに、PCIe Gen5を統合して、高速の入力・出力とカード間のダイレクト通信を実現しました。また、システム全体のレイテンシ（latency）を最小化する一方で、データスループット（throughput）を最大化しました。このように並列処理やコンパイラの最適化、高速のインターコネクト技術を融合したRSDは、最も高度なAI演算も効率的に処理できるインフラを提供しています。

テンソル並列処理

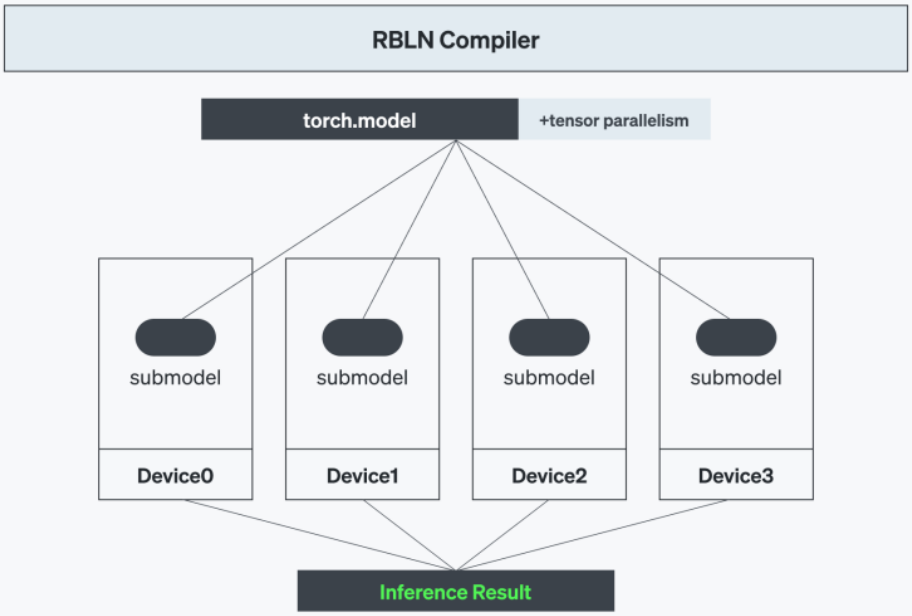
LLMをAIプロセッサで推論（inference）する場合は、さまざまな技術的な課題があります。プリフィル（prefill）段階は演算集約的であるのに対し、デコード（decoding）段階ではたくさんのメモリが求められます。この際に、テンソル並列処理は、演算負荷を複数のデバイスに分散させ、各デバイスのメモリ占有率と演算負荷を効果的に減らすdecodingの解決策を提供します。テンソル並列処理を正しく実現すれば、KVキャッシュ(KV caching)や、LLMの大規模な重みによって発生するメモリの帯域幅の制限を緩和できます。これにより、ハードウェアは高いスループットと低遅延を維持しながら、複雑なLLM推論を効率的に行うことができます。

RBLNコンパイラは、このテンソル並列処理を最大の性能と稼働率（utilization）で管理できるように最適化しています。コンパイルの段階でコンパイラは、モデルを複数のデバイスにわたってテンソル単位で細かく分割し、各チップが全体演算の一部のみを担当するように構成します。この段階で生成されるコマンドストリーム（Command Stream）は、コマンドプロセッサ（Command Processor）が実行するコマンドセットです。これには、推論をする際にチップ間通信に必要なインターデバイスのデータ移動の情報も含まれています。

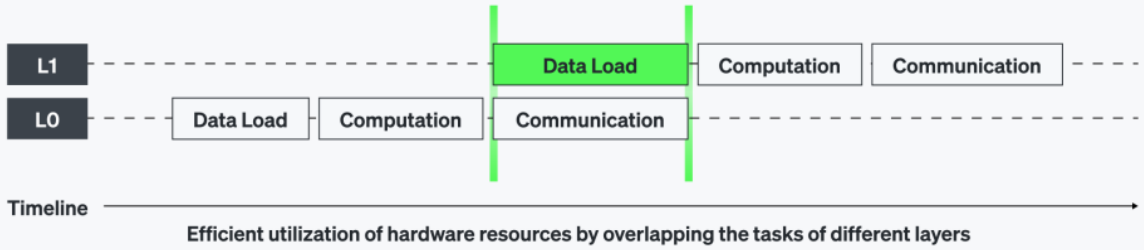
コンパイラレベルの最適化

RBLNコンパイラは、AIワークロードの複雑な拡張を効率的に管理するように設計された高性能なAI推論向けの重要なツールです。テンソル並列処理を効果的にサポートするうえ、モデルを複数のデバイスに円滑に分散し、最適なリソースの活用と実行の高速化を支えます。また、マルチデバイス間通信の最適化や自動演算分割（automatic splitting）、レイヤーパイプライン（layer pipelining）などの機能により拡張性をさらに強化しました。

1. **マルチデバイスの自動分割（Automatic Multi-Device Splitting）**
   RBLNコンパイラは、演算の分割と再結合の過程をすべて自動的に行います。ユーザーなしでもマルチデバイス間の演算が自動処理されるため、ユーザーにとって複雑なテンソル並列処理が、単純で直感的なプロセッサに切り替わります。



2. **デバイス間通信の最適化（Optimization of Inter-Device Communication）**
   RBLNコンパイラは、LLMを実行する際に発生するデバイス間通信（inter-device communication）を高度化し、ブロードキャスト（broadcast）やリデュース（reduce）、部分和（partial sum）などの集合通信パターン（collective communication pattern）を効率よく処理します。
3. **Efficient Layer Pipelining for Intra-Device Communication**
   RBLNコンパイラは、各デバイスの内部でレイヤーパイプライン（layer pipelining）を取り入れ、デバイス間通信が中断せずに（seamless）行われるようにします。これにより、すべての演算が並列で処理され、アイドル時間を最小化します。また、ハードウェアの効率を最大化し、通信のオーバーヘッドを減らすことができます。



PCIe Gen5

リベリオンのRSDは、PCIe Gen5x16のインタフェースを採用しており、ホストの連携やカード間のダイレクト通信両方で、全二重通信（full-duplex）の64GB/s帯域幅を提供します。これは、特にカード間の効率的なテンソル並列処理に欠かせず、全てのニューラルエ

ンジン（Neural Engine）間通信に対して高いスループットと低遅延をサポートし、拡張可能な速度で優れた推論性能を実現しています。リベリオンはPCIeの性能を最大限に引き上げるために、ファームウェアを専用に最適化します。また、PCIeスウィッチとCPUとの相互運用性を最大化するように設計されており、システム全体の効率をさらに向上させました。

システムソリューション

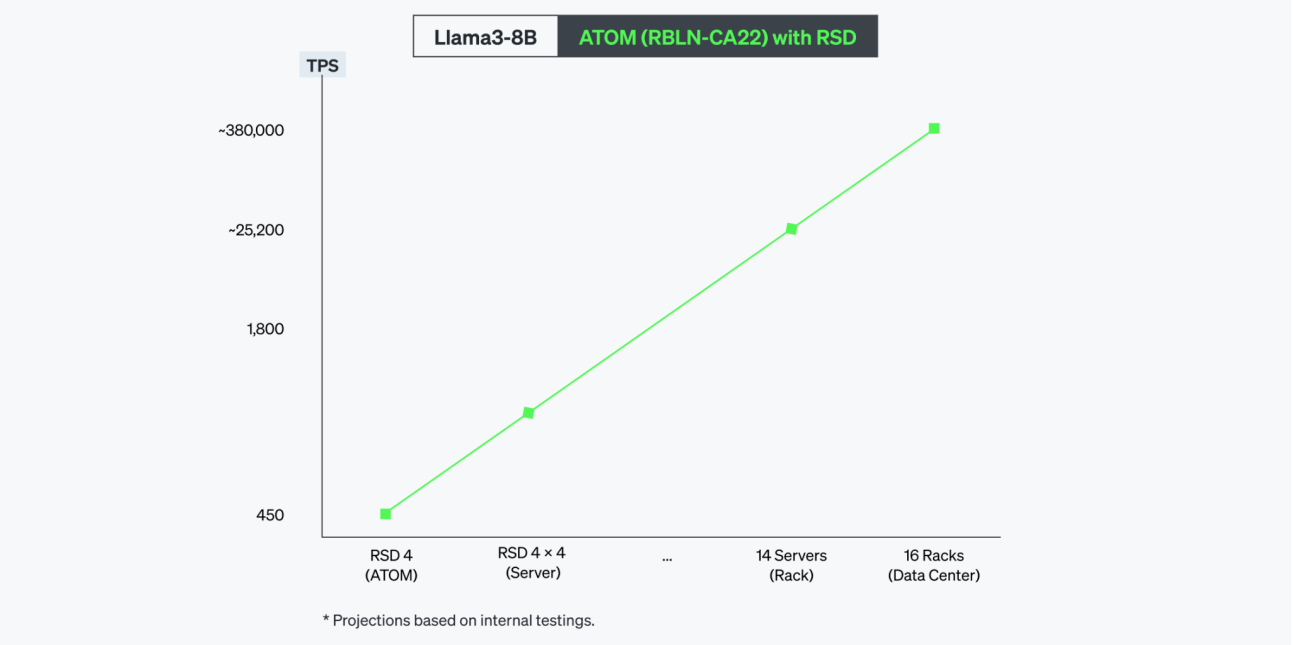
RSDは、軽量ワークロード（workstation）からSLMとLLMのような高負荷のAI作業まで、高効率のTPS/Wattを提供する優秀で経済的なシステムソリューションです。この効率性は、RSDのスケーラビリティとラックレベルの性能を最適化することで実現できます。さらに、vLLMとLiteLLMの統合により完成されました。 RSDはこれらの技術を組み合わせ、演算需要が増えても安定的に拡張できます。また、エネルギー・コスト当たりのスループットを最大限に引き上げるために、あらゆる規模のAIインフラに対して理想的なソリューションを提供しています。

スケーラビリティ

スケーラビリティは、演算資源を追加すればするほどそれに比例して性能が増加するため、ハードウェアの投入が直ちにスループットの向上につながるようサポートします。この特性は、大規模なAIモデルとデータ集約的なアプリケーションによって急増する演算需要を、高効率かつ高速度で対応するのに欠かせないものです。

RSDはカード、サーバー、ラックシステムなどさまざまなデプロイ環境において、このスケーラビリティを完全に実現しています。ノード間のデータ同期化（data synchronization）の最適化、ノード間通信のオーバーヘッドの最小化、メモリ帯域幅の管理を効率化してボトルネックを防ぎます。また、デバイスの数が増えても一貫して低遅延を維持しながら、それに比例してスループットを拡張できます。

さらに、低遅延の性能を維持するために、負荷分散（load balancing）と動的ワークロードの配分（dynamic workload distribution）を利用して、あらゆるデバイスが遅延せずに最大の効率で動作できるようにしました。RSDは、この複雑性を最先端のハードウェアアーキテクチャとAIソフトウェアフレームワークで解決するため、演算需要が増加しても拡張可能な高速処理の性能が維持できます。



ラックレベルの最適化（Rack-level Optimizations）

ラック単位のAI推論環境では、サーバー間通信とワークロードの配分を円滑にするための高効率なルーティングプロトコルが欠かせません。

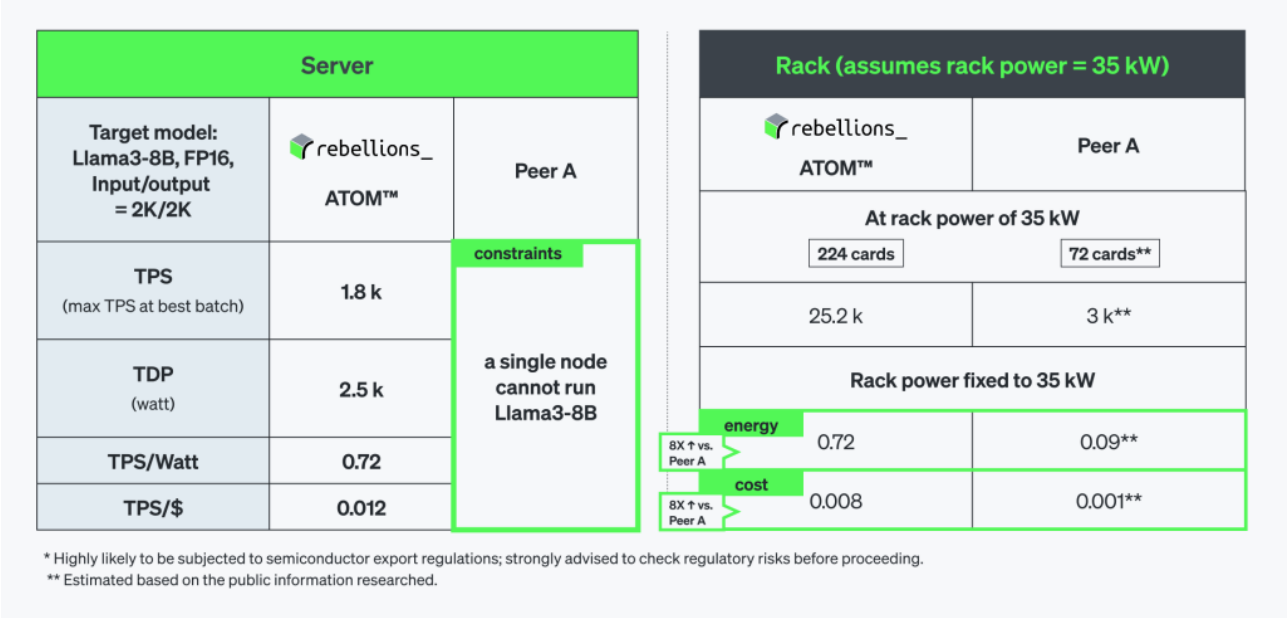
RSDはvLLMと統合したルーターサーバー（router server）を導入し、ラックレベルで最適な性能と拡張性を提供しています。このルーターサーバーは、複数のvLLMインスタンスを一つの統合システムにつなげるフレームワークの役割を果たします。サーバー間のワークロードを効率よく配分するため、ラック全体の効率を最大限に引き上げます。これにより、全てのサーバーへの負荷がバランスが取れた状態で最大のスループットを維持でき、過負荷を防ぎながらもレイテンシを持続的に短縮することができます。

そのためユーザーとしては、LLMモデルにAPIエンドポイントを通じてアクセスしやすくなり、優れた拡張性と利便性を同時に享受できます。結果的にvLLMとルーターサーバーの結合は、個別サーバーを精密に調整された高性能AIシステムに転換させる連携効果を発揮できます。

Llama3-8B

Llama3-8Bベンチマークの結果は、RSDの電力効率（energy efficiency）が競合他社の製品に比べてどれだけ優秀なのかを明確に示します。単体のATOM™サーバーが2,500W電力で1800TPSを記録しており、224枚のATOM™カードで構成されたラックシステムでは、

25,200TPSまで線形的に拡張できます。規模が大きくなればなるほどその効率性はさらに顕著になり、TPS/WattおよびTPS/$基準でエネルギー効率と費用効率が両方とも8倍以上向上しました。



結論

RSDは、複雑で資源集約型のAIワークロードのニーズを完全に満たせるように、リベリオンの最先端AIインフラビジョンを実現していきます。モジュール型・拡張型のアーキテクチャを基にしたRSDは、単体のワークステーションから大規模なラックシステムまで、一貫したスケーラビリティを提供しています。これらに、高レベルのテンソル並列処理、PCIe Gen5の統合、RBLNコンパイラの精巧な最適化技術が加わることで、RSDはあらゆる規模でも効率的かつ高性能なAI推論環境を提供することができます。