



RebelServer™

AI推論を
効率的かつ大規模に実現

RebelServerは、Rebellionsの統合型ハードウェア・ソフトウェア推論ソリューションを支えるAIコンピューティング基盤であり、四つの中核要素を中心に設計されています。



ソブリンAI

追加の電力設備や冷却設備を必要とせず、既存の空冷式データセンターへそのまま導入できます。完全なオンプレミス運用により、データ主権・セキュリティ・制御性を維持します。



使いやすさ

オープンソースのフレームワークや業界標準ツールとシームレスに統合できるよう設計されており、ベンダーロックインを回避し、既存の専門知識をそのまま活用できます。新たな技術を習得する必要がありません。



プロダクション運用の 経済性を最適化

AI推論ワークロード向けに最適化されたラックスケールアーキテクチャにより、優れたワット当たりのコストパフォーマンスを実現します。



量産検証済みの ソリューション

すでに数百台のラックが企業や政府機関に導入され、本番環境のAIワークロードを支えています。

当サーバは、急増する電力消費や運用コストなど、AI推論の展開に伴う主要な課題を解決するため、性能を最大限に発揮しながら、優れた電力効率を実現するよう設計されています。さらに、本システムは次世代のエージェント型AIや推論モデ

ルの構築を支える基盤として、MoE (Mixture of Experts) アーキテクチャ、あらゆる規模のAIモデル、さらに言語・ビジョン・音声に対応したマルチモーダル機能をシームレスに統合します。

Spec

フォームファクター	5U (RebelCard™ × 8)
CPU	AMD EPYC™ 9355 × 2 (32コア／64スレッド、3.55GHz、280W、第5世代 EPYC)
メモリ	1.5TB (64GB DDR5 × 24)
ディスク	OS: 1.92TB NVMe × 2 Data (KV Cache): 1.92TB NVMe × 2
ネットワーク	400G 1ポート QSFP112-DD × 4 100G 2ポート QSFP56 × 1 10G・25G 2ポート SFP28 (OCP 3.0) × 1
電源ユニット	2700W × 6
通常時／最大消費電力	4kW/7kW * 本仕様に基づく理論上の最大消費電力を示しています。実際の消費電力が7kWを超えることはなく、通常は実際のワークロードにおいて4～6kW程度で動作します。
サポート	3年間の標準ハードウェア・ソフトウェアサポート
重量	総重量: 100 lbs (45.3kg)、本体重量: 65.6 lbs (29.7kg)
動作環境	動作温度: 10°C～35°C (50°F～95°F) 保管温度: -40°C～60°C (-40°F～140°F) 動作時相対湿度: 8%～90% (結露なきこと) 保管時相対湿度: 5%～95% (結露なきこと)

妥協なき性能と効率

RebelServerは、FP16で petaFLOPS、FP8で 2 petaFLOPS の演算性能を実現し、AI の性能・拡張性・効率性を新たな水準へ引き上げます。1台のサーバーには RebelCard™ を8枚搭載し、各カードには Rebel100™ チップを採用しています。Rebel100™ は、UCle-Advanced インターコネクトで接続された4基のチップレットで構成される SoC に 144GB の HBM3E メモリを搭載することで、最大規模かつ複雑な推論ワークロードもシームレスに処理します。

使いやすさ

RebelServerに統合されたRebellionsのソフトウェアスタックは、導入・運用に求められる要件を満たします。

- クラウドネイティブに対応
- オープンソースフレームワークとの互換性
- 高性能な分散推論
- 幅広いモデルへの対応
- あらゆる世代のRebellionsハードウェアで一貫した利用体験を提供

Rebellionsのクラウドネイティブスタックは、プロダクションスケールのクラウドオーケストレーションに対応します。Rebellionsの推論スタックは、三つの主要コンポーネントで構成されています。

- vLLM、PyTorchとの共同設計による推論エンジン
- 高速インターコネクトと専用ソフトウェアライブラリをフルスタックで統合した高性能分散推論
- 事前に検証・最適化された多様なモデルを収録したモデルカタログ。追加作業なしで本番環境へ即座に導入できるほか、独自のAIワークロードを開発・最適化する際のリファレンスとしても活用できます。