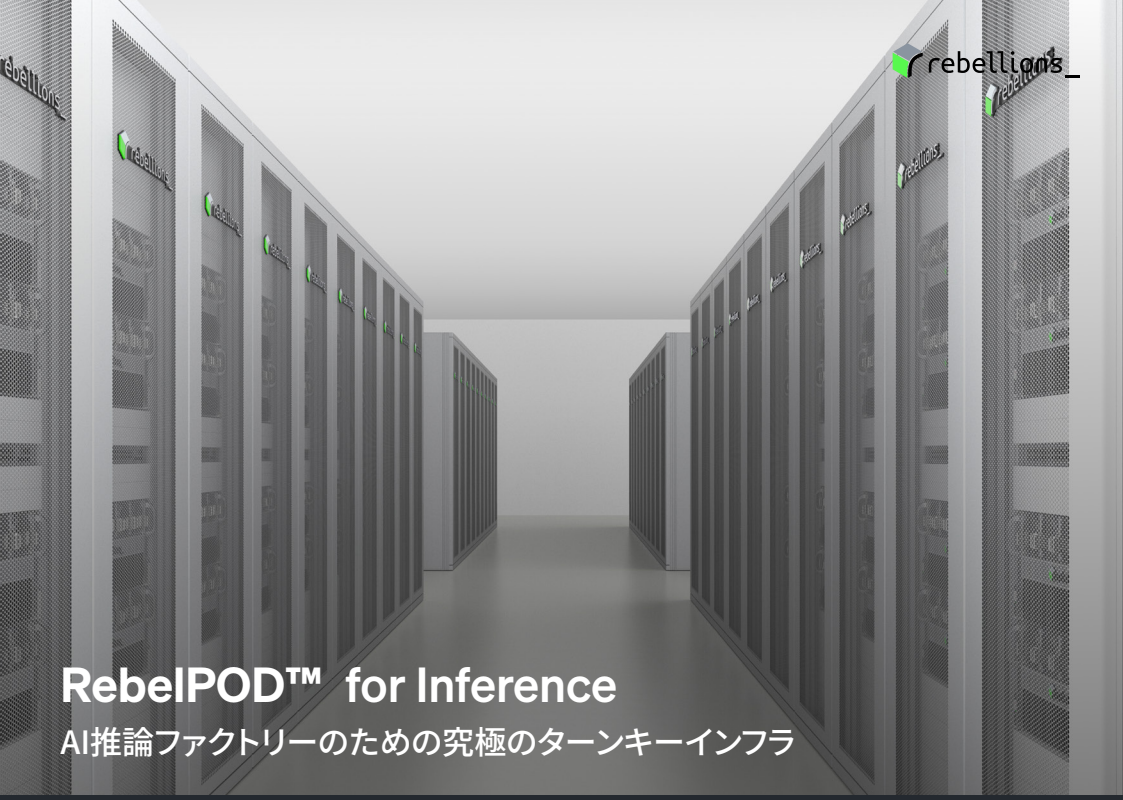

RebelPOD™ for Inference

AI推論ファクトリーのための究極のターンキーインフラ

RebelPODは、事前検証済みの固定構成を採用したデプロイメントユニットです。RebelServerをRebelRack単位で集約した構成を採用し、各ラックにはサーバー、ネットワーク、PDU（電源分配装置）があらかじめ組み込まれています。RebelPODは32・64・128・256ノード構成で提供され、構成規模ごとに予測可能で、一貫して検証された性能・高密度プロファイルを提供します。最適化されたAIファクトリーとして、大規模推論のデプロイに伴う複雑さを解消します。複雑なインフラ構築に煩わされることなく、すぐにモデルのデプロイと推論に集中できます。事前検証を完了し、本番環境で実証済みの構成であるため、数か月ではなく、短期間で価値を実現できます。

RebelPOD™ 32ノードシステム構成

仕様	32ノード
ノード	32× RebelServer™ (5U)
加速器	256× RebelCard™
NPUラック	8
マネジメントラック	1
性能	FP8: 524.3 PFLOPS / FP16: 262.1 PFLOPS
NPUメモリ	36.86 TB HBM3E、帯域幅：1,228.8 TB/s
CPU	64× AMD EPYC™ 9355 (2,048C / 4,096T)
システムメモリ	48 TB DDR5
ネットワーク	バックエンドファブリック：400G QSFP112-DD フロントエンド：25G SFP28 ストレージ：100G（オプション）
冷却方式	空冷式
消費電力	最大224 kW（NPUラックのみ）

Key Benefits

推論向けのフルスタックAIインフラ

高性能なRebel100™アクセラレータ、高速インターコネクト、Rebellionsソフトウェアスタックを統合した、電源を入れるだけで利用できるラックスケール型ターンキーシステムです。本番環境向けのクラウドオーケストレーションに対応し、オープンソースの推論エンジンや分散推論に加え、事前検証・最適化済みのモデルカタログも備えています。本番環境へ迅速にデプロイできます

業界最高水準の収益性を誇る推論プラットフォーム

大規模・高効率推論のために設計されたソリューションです。最先端のReasoningモデルやエージェント型ワークロードを、優れたワット当たり性能で高効率に処理します。低消費電力を維持しながらスループットを最大化し、運用コストを大幅に削減するとともに、AI投資のROIを最大化します。

検証済みアーキテクチャによる迅速な価値実現

DIYクラスタでは避けられなかった、数週間に及ぶ導入作業や数か月にわたるテストは、もう必要ありません。すでに数百台のラックがお客様の本番環境で稼働しており、すべての構成は実際のAIワークロードで徹底的に検証されています。本番環境で実証済みのアーキテクチャにより、数か月ではなく、わずか数日で稼働を開始できます。

インフラを選ばないソブリンAI

水冷設備や電力設備の改修は必要ありません。既存の空冷式データセンターにそのまま導入できます。PODスケールソリューションは、完全なオンプレミスでの導入・運用に対応し、政府・企業の機密データに対するデータ主権、セキュリティ、管理権を確保します。

Software Edge: Rebellions Cloud-Native Stack

ラックスケールソリューションには、Rebellionsソフトウェアスタックがあらかじめ統合されています。
新しい技術を習得する必要はなく、既存の知識やスキルをそのまま活用できます。

推論エンジン

最大限の互換性を実現するため、vLLMおよびPyTorchと共同設計されていま。

分散推論

高速インターコネクトに対応した専用ソフトウェアライブラリにより、マルチノードへのスケールアウトをシームレスに実現します。

ベンダーロックインなし

オープンソース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ます。

モデルカタログ

事前検証・最適化済みの豊富なモデルを備え、追加作業なしで、すぐに本番環境へデプロイできます。