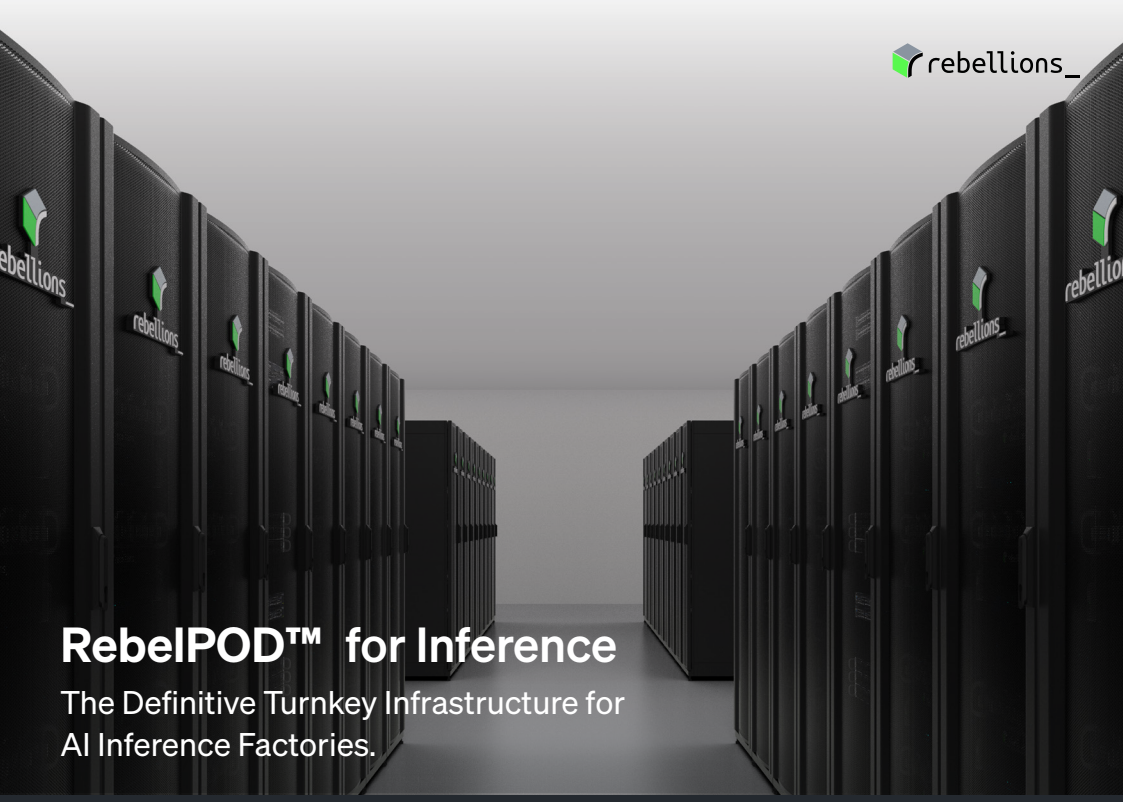

RebelPOD™ for Inference

The Definitive Turnkey Infrastructure for AI Inference Factories.

RebelPOD is a pre-validated, fixed-configuration deployment unit built from RebelServers assembled into RebelRack units, each including servers, network, and power distribution within the rack. RebelPOD is available in 32, 64, 128, or 256-node configurations, delivering a known, repeatable performance and density profile at each scale. RebelPOD is a fully designed and integrated AI factory designed to eliminate the complexity of deploying large-scale inference, shifting the focus from complex infrastructure assembly to immediate model deployment and inference. By utilizing pre-validated and production-proven configurations, customers can recognize value in a fraction of the time.

rebellions_

RebelPOD™ 32-Node System Configuration

Spec	32-Node
Nodes	32× RebelServer™ (5U)
Accelerators	256× RebelCard™
NPU Racks	8
Management Racks	1
Performance	FP8: 512 PFLOPS / FP16: 256 PFLOPS
NPU Memory	36.86 TB HBM3E, 1,228.8 TB/s bandwidth
CPU	64× AMD EPYC™ 9355 (2,048C / 4,096T)
System Memory	48 TB DDR5
Network	Backend Fabric: 400G QSFP112-DD Frontend: 25G SFP28 Storage: 100G (optional)
Cooling	Air-cooled
Power Consumption	224 kW max (NPU racks)

Key Benefits

Full-Stack AI Infrastructure for Inference

This rack-scale solution is a ready-to-run, turnkey system optimized with high-performance Rebel100™ accelerators, high-speed interconnects, and Rebellions Software Stack that enables production grade cloud orchestration with support for an open source inference engine and distributed inference along with a diverse model catalog offering pre-validated and optimized models that can be rapidly deployed directly into production environments.

The Industry's Most Profitable Inference Platform

Engineered specifically for energy-efficient inference at scale, this solution tackles state-of-the-art reasoning models and agentic workloads with a focus on superior price-performance-per-watt. By maximizing throughput while maintaining low power consumption, organizations can significantly reduce operational costs and maximize ROI on AI investments.

Validated and Proven for Immediate Time-to-Value

Eliminate the weeks of install and months of testing typically required for DIY clusters. With hundreds of racks already deployed in customer production environments, our configurations are extensively tested for real-world AI workloads. This production-proven architecture ensures you are up and running in days, not months.

Sovereign AI with Zero Infrastructure Friction

Deploy in existing air-cooled data centers without the need for specialized liquid cooling or power retrofitting. The POD-scale solution enables fully on-premises deployment and operation, ensuring complete data sovereignty, security, and control for sensitive government and enterprise data.

Software Edge: Rebellions Cloud-Native Stack

The rack solution comes pre-integrated with the Rebellions software stack, designed to leverage existing expertise without requiring new skills:

Inference Engine

Co-engineered with vLLM and PyTorch for maximum compatibility.

Distributed Inference

Specialized software libraries that manage high-speed interconnects for seamless multi-node scaling.

No Vendor Lock-in

Seamlessly integrates with open-source frameworks and industry-standard tools.

Model Catalog

A diverse library of pre-validated and optimized models ready for out-of-the-box production.