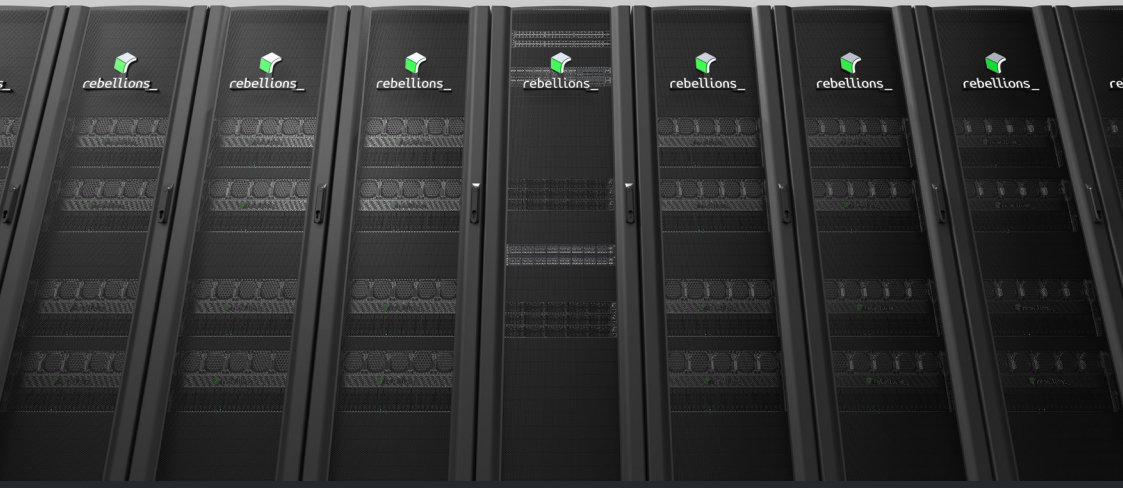
rebellions_



RebelRack™

The Definitive Turnkey Infrastructure for Generative AI and Agentic Workflows

RebelRack is a fully designed and integrated AI factory designed to eliminate the complexity of deploying large-scale inference. By combining Rebellions' high-efficiency and high performing RebelServers™ with enterprise-grade networking and Rebellions Software Stack, this solution eliminates traditional integration difficulties, 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 required by DIY clusters.

The server is designed to maximize performance while delivering energy efficiency to address the key challenges of inference deployments including skyrocketing power consumption and operational costs.
The system is a foundation that allows users to build the next generation of agentic and reasoning models seamlessly integrating Mixture of Experts (MoE) architectures, models of any size, and multimodal capabilities spanning language, vision, speech, and more.

All information is provided "as is" without any warranties or guarantees by Rebellions, Inc.("Rebellions"). Rebellions assumes no liability for use of or reliance upon this information.

Key Benefits



Full-Stack AI Infrastructure

A ready-to-run, turnkey system optimized for production-grade cloud orchestration. By pairing Rebel100™ accelerators with Rebellions Software Stack, the solution supports open-source engines and distributed inference, providing a streamlined path to production with pre-validated model support.



The Industry's Most Profitable Inference Platform

Engineered for energy-efficient inference and agentic workloads at scale, this solution delivers industry-leading price-performance-per-watt. By maximizing throughput while maintaining low power consumption, organizations can significantly reduce operational costs and maximize ROI on AI investments.



Proven Reliability, Immediate Deployment

Skip the months of configuration required for DIY clusters. With hundreds of racks already in production, our configurations are extensively tested for real-world AI workloads. Move from delivery to production 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 rack-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.

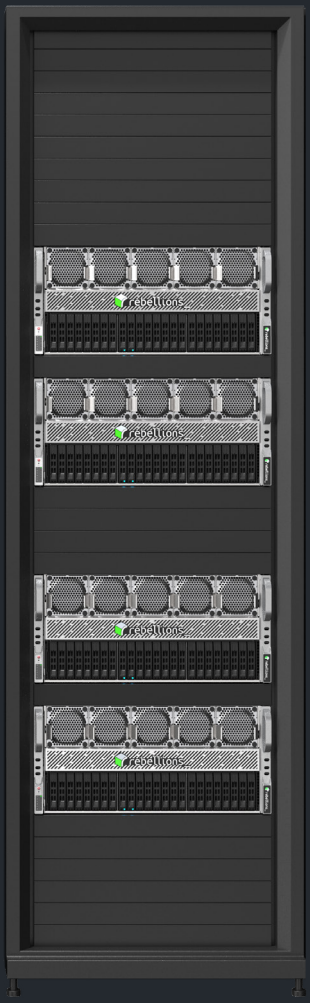
rebellions_

System Configuration

RebelRack™ Hardware

Base Configuration	1x RebelRack NPU Rack 1x RebelRack Management Rack
Scalability	After the initial base configuration, additional NPU racks can be added one at a time or as needed up to 8 total NPU racks via the network fabric. For more than 8 NPU racks, an additional Management will be required.

RebelRack™ NPU Rack



Server Model	4x RebelServer™ (5U)
Accelerator	32x RebelCard™
NPU Memory	4.61TB total, 153.6TB/s HBM3e bandwidth
CPU	8x AMD EPYC™ 9355 Processors (256C/512T)
System Memory	6TB DDR5
Network	Backend Fabric : 400G QSFP112-DD Frontend Network : 25G SFP28 Storage Network : 100G (Optional)
Performance	FP8: 64 PFLOPS FP16: 32 PFLOPS
Cooling Type	Air-cooled
Power Consumption	28kW max
Rack	Height: 42U, 78.38” (1991mm) Width: 23.62” (600mm) Depth: 47.24” (1200mm)
Weight	~773 lbs (~351 kg), unpackaged
Cooling	105,094 BTU/hour (max)

rebellions_

All information is provided “as is” without any warranties or guarantees by Rebellions, Inc.(“Rebellions”). Rebellions assumes no liability for use of or reliance upon this information.

System Configuration

RebelRack™ Management Rack



Management Servers	3x AMD CPU Servers (optional)
CPU	3x AMD EPYC 9254/9255 (24 cores)
System Memory	384GB Total (3x Management Servers, each with 128GB [4x 32GB DDR5])
Network	Backend Fabric : 800G OSFP Frontend Network : 25G SFP28 + 100G QSFP28
Storage	Flexible support for validated partner storage solutions.
Power Consumption	5.2kW max
Rack	Height: 42U, 78.38” (1991mm) Width: 23.62” (600mm) Depth: 47.24” (1200mm)
Weight	~ 558 lbs (~ 253 kg), unpackaged
Cooling	19,517 BTU/hour (max)

* Multiple rack configuration options are available to enable easier integration into existing data center environments

rebellions_

All information is provided “as is” without any warranties or guarantees by Rebellions, Inc.(“Rebellions”). Rebellions assumes no liability for use of or reliance upon this information.