rebellions_

RebelServer™

Powering AI Inference Efficiently and at Scale

RebelServer™ is the AI compute building block behind Rebellions' integrated hardware and software inference solution, which is focused on four main pillars.

Key Features

Sovereign AI

Instantly deploy in existing air-cooled data centers without additional power or cooling requirements, with fully on-premises operation maintaining complete data sovereignty, security, and control.

Ease of Use

Built to seamlessly integrate with open-source frameworks and industry-standard tools, eliminating vendor lock-in and leveraging existing expertise. No new skills required.

Optimized for Production Tokenomics

Rack-scale architecture delivering superior price-performance-per-watt, optimized specifically for AI inference workloads.

Production-Proven Solution

Hundreds of racks deployed by enterprise and government customers running production AI workloads.

The server is designed to maximize performance while delivering energy efficiency to address the key challenges of inference deployments including sky-

rocketing 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.

Spec

Form Factor	5U (8x RebelCard™)
CPU	2x AMD EPYC 9355 (32C 64T, 3.55GHz, 280W) [EPYC 5th Gen]
Memory	1.5TB (24x 64GB DDR5)
Disk	2x 1.92TB NVMe
Network	4x 400G 1-Port QSFP112-DD 1x 100G 2-Port QSFP56 1x 10G/25G 2-Port SFP28 (OCP 3.0)
Power Supply	6x 2700W
Typical/Max Power Draw	4-6kW/7kW * Theoretical maximum power consumption based on specifications. Actual power consumption will not exceed 7 kW, typically hovering around 4-6kW at most under practical workloads.
Support	Comes with 3-year businessstandard hardware and software support
Weight	Gross Weight: 100 lbs (45.3 kg) / Net Weight: 65.6 lbs (29.7kg)
Operating Temperature Range	Operating Temperature: 10°C to 35°C (50°F to 95°F) Non-operating Temperature: -40°C to 60°C (-40°F to 140°F) Operating Relative Humidity: 8% to 90% (non-condensing) Non-operating Relative Humidity: 5% to 95% (non-condensing)

Uncompromised Performance and Efficiency

RebelServer™ breaks the limits of AI performance, scale and efficiency by delivering 1 petaFLOPs FP16 and 2 petaFLOPs FP8 throughput. A server includes 8 RebelCards™, where each card contains a Rebel100™ chip, with an SOC composed of 4 chiplets utilizing UCie-Advanced interconnect, complemented by 144GB of HBM3E memory for seamlessly serving the largest and most complex inference workloads.

Highest Flexibility and Ease of Use

The Rebellions Software Stack integrated in RebelServer™ serves end user deployment and operational requirements:

- Cloud-Native readiness
- Open-Source framework compatibility
- High-Performance distributed inference
- Extensive model versatility
- Consistent experience across all Rebellions hardware generations

The Rebellions Cloud Native Stack enables production scale cloud orchestration. The Rebellions inference stack includes three core components:

- An inference engine co-engineered with vLLM, PyTorch, and Triton
- High-performance distributed inference that integrates a full stack of high-speed interconnect and specialized software libraries
- A model catalog filled with a diverse offering of prevalidated and optimized models that can be deployed directly into production environments out of the box or serve as a reference blueprint for users to develop and optimize their own AI workloads