

# RebelServer™

Powering AI Inference  
Efficiently and at Scale



RebelServer™는 Rebussions의 통합 하드웨어 및 소프트웨어 기반 AI 추론 솔루션을 구성하는 핵심 인프라로, 네 가지 핵심 가치에 기반해 설계되었습니다.



## Sovereign AI

기존 공랭식 데이터센터에 추가 전력이나 냉각 설비 없이 즉시 배포 가능합니다. 온프레미스 환경에서 데이터 주권, 보안, 운영 전반에 대한 완전한 통제를 제공합니다.



## Ease of Use

오픈소스 프레임워크 및 산업 표준 도구와 자연스럽게 통합됩니다. 벤더 종속성 없이 기존 역량을 그대로 활용할 수 있으며, 별도의 학습 없이 즉시 적용 가능합니다.



## Optimized for Production Tokenomics

랙 단위 아키텍처를 기반으로 와트당 성능을 극대화합니다. AI 추론 워크로드에 최적화된 효율을 제공합니다.



## Production-Proven Solution

기업 및 공공기관의 실제 운영 환경에 대규모로 배포되어 검증되었습니다. 안정적으로 AI 워크로드를 지원합니다.

성능과 에너지 효율을 동시에 확보하여, 급증하는 전력 소비와 운영 비용 등 AI 추론 환경의 주요 과제를 해결합니다. 또한, 차세대 에이전트형 및 추론 모델 구축을 위한 기반을 제공

하며, MoE(Mixture of Experts) 구조, 다양한 규모의 모델, 그리고 언어·비전·음성을 포함한 멀티모달 기능을 유연하게 통합할 수 있습니다.

## 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 business standard 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)

## 성능과 효율의 극대화

RebelServer는 FP16 기준 1 petaFLOPs, FP8 기준 2 petaFLOPs의 처리 성능을 제공하여 AI 성능, 확장성, 효율의 한계를 확장합니다.

하나의 서버는 8개의 RebelCards로 구성되며, 각 카드에는 Rebel100™ 칩이 탑재되어 있습니다. 해당 칩은 UCIe-Advanced 인터커넥트를 활용한 4개의 칩렛 기반 SoC 구조로 설계되었으며, 144GB의 HBM3E 메모리를 통해 대규모·고복잡도 AI 추론 워크로드를 안정적으로 처리합니다.

## 사용 편의성

RebelServer에 통합된 Rebellions Software Stack은 사용자 배포 및 운영 요구사항을 지원합니다.

- 클라우드 네이티브 환경 지원
- 오픈소스 프레임워크 및 산업 표준 도구와의 통합
- 고성능 분산 추론
- 다양한 모델 지원
- Rebellions 하드웨어 전 세대에 걸친 일관된 사용자 경험 제공

Rebellions Cloud Native Stack은 프로덕션 규모의 클라우드 오케스트레이션을 지원합니다. Rebellions Inference Stack은 세 가지 핵심 구성 요소로 이루어집니다.

- vLLM, PyTorch, Triton과 공동 설계된 추론 엔진
- 고속 인터커넥트와 특화된 소프트웨어 라이브러리를 통합한 고성능 분산 추론
- 사전 검증 및 최적화된 다양한 모델로 구성된 모델 카탈로그로, 즉시 프로덕션 환경에 배포하거나 AI 워크로드 개발 및 최적화를 위한 기준(reference)으로 활용 가능

