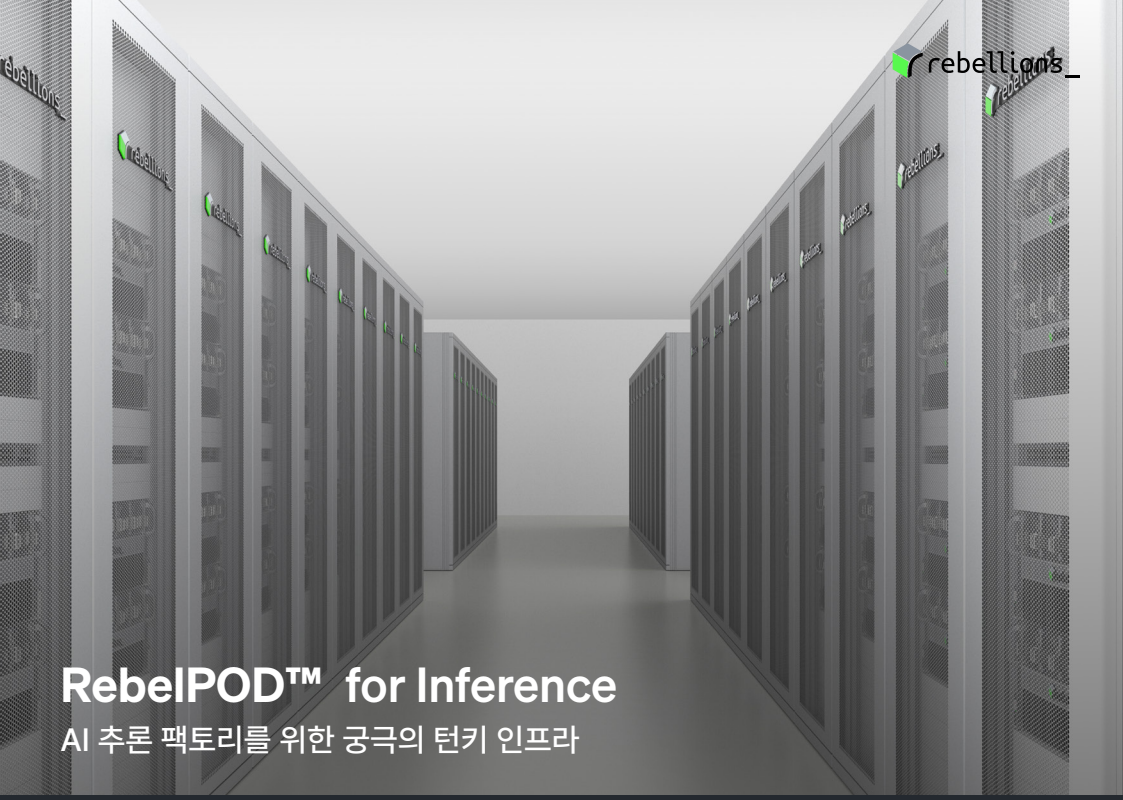

RebelPOD™ for Inference

AI 추론 팩토리를 위한 궁극의 턴키 인프라

RebelPOD는 사전 검증된 고정 구성의 배포 단위로, RebelServer를 RebelRack 단위로 집약해 구성합니다. 각 랙에는 서버와 네트워크, 전원 분배 장치가 모두 담겨 있습니다. RebelPOD는 32·64·128·256노드 구성으로 제공되며, 규모마다 예측 가능하고 반복 검증된 성능·집적도 프로파일을 보장합니다. RebelPOD는 완벽하게 설계·통합된 AI 팩토리로, 대규모 추론 배포의 복잡함을 걷어냅니다. 골치 아픈 인프라 조립은 잊고, 곧바로 모델 배포와 추론에 집중하세요. 사전 검증을 마치고 실제 양산 환경에서 입증된 구성으로 훨씬 짧은 시간 안에 가치를 실현할 수 있습니다.

RebelPOD™ 32-Node System Configuration

사양	32-노드
노드	32× RebelServer™ (5U)
가속기	256× RebelCard™
NPU 랙	8
매니지먼트 랙	1
성능	FP8: 524.3 PFLOPS / FP16: 262.1 PFLOPS
NPU 메모리	36.86 TB HBM3E, 1,228.8 TB/s 대역폭
CPU	64× AMD EPYC™ 9355 (2,048C / 4,096T)
시스템 메모리	48 TB DDR5
네트워크	Backend Fabric: 400G QSFP112-DD Frontend: 25G SFP28 Storage: 100G (선택사항)
냉각 방식	공냉식
소비 전력	224 kW max (NPU 랙 기준)

rebellions_

Key Benefits

추론을 위한 풀스택 AI 인프라

고성능 Rebel100™ 가속기와 고속 인터커넥트, Rebellions 소프트웨어 스택으로 최적화된, 전원만 켜면 되는 랙 스케일 턴키 시스템입니다. 프로덕션급 클라우드 오케스트레이션을 지원하며, 오픈소스 추론 엔진과 분산 추론은 물론 사전 검증·최적화된 다양한 모델 카탈로그까지 갖춰 프로덕션 환경에 곧바로 빠르게 배포할 수 있습니다.

업계 최고 수익성의 추론 플랫폼

대규모 에너지 효율 추론을 위해 태어난 솔루션으로, 최첨단 추론(reasoning) 모델과 에이전틱 워크로드를 탁월한 와트당 가격 대비 성능으로 소화합니다. 낮은 전력 소비를 유지하면서 처리량을 극대화해, 운영 비용은 크게 낮추고 AI 투자 수익(ROI)은 극대화합니다.

검증으로 입증된 즉각적인 가치 실현

DIY 클러스터에 으레 따라붙는 몇 주간의 설치와 몇 달간의 테스트는 이제 없습니다. 이미 수백 대의 랙이 고객의 프로덕션 환경에서 가동 중이며, 모든 구성은 실제 AI 워크로드로 혹독하게 검증됐습니다. 양산으로 입증된 이 아키텍처 덕분에, 몇 달이 아니라 며칠이면 가동을 시작합니다.

인프라 걱정 없는 소버린 AI

수냉식 설비도, 전력 개조도 필요 없습니다. 지금 쓰는 공냉식 데이터센터에 그대로 들어갑니다. POD 스케일 솔루션은 완전한 온프레미스 배포와 운영을 지원해, 민감한 정부·기업 데이터의 주권과 보안, 통제권을 온전히 지켜냅니다.

Software Edge: Rebellions Cloud-Native Stack

랙 스케일 솔루션에는 Rebellions 소프트웨어 스택이 이미 통합돼 있습니다. 새로운 기술을 배울 필요 없이, 지금의 전문성을 그대로 쓰면 됩니다:

추론 엔진

vLLM, PyTorch와 함께 설계해 호환성을 극대화했습니다.

분산 추론

고속 인터커넥트를 다루는 전용 소프트웨어 라이브러리로 멀티노드 확장이 매끄럽습니다.

벤더 종속 없음

오픈소스 프레임워크와 업계 표준 도구에 자연스럽게 녹아듭니다.

모델 카탈로그

사전 검증·최적화된 다양한 모델을 갖춰, 별도 작업 없이 곧바로 프로덕션에 투입합니다.