

ATOM™-Max Pod

RDMA 기반 고속 네트워크로 완성하는 Rack-scale AI 인프라

ATOM™-Max Pod은 노드 간 분산 처리가 요구되는 대규모 AI 추론 서비스를 위해 구성된 Rack-scale 인프라입니다. 리벨리온의 AI 가속기와 RDMA 기반 고속 네트워크, 익숙한 환경의 소프트웨어 스택까지 통합된 턴키 인프라로 제공됩니다. 서버 8대 구성의 Mini-Pod부터 기업 AI 서비스에 요구되는 스케일로 유연한 확장이 가능합니다.

Key Features



Limitless Scale Out Architecture

서버 8개로 이뤄진 Mini Pod부터 수십 개의 서버로 구성된 Pod까지, RSD를 통해 하나의 연결된 클러스터로 자유롭게 확장할 수 있습니다. 워크로드 증가에 따른 유연한 리소스 운영과 함께 선형적인 성능 향상을 제공합니다.



Ultra-Low Latency RDMA Fabric

Pod 내 서버는 400GB/s RDMA 기반 고속 네트워크를 통해 상호 연결되어 있습니다. Resource-intensive AI 모델에 필수적인 분산 처리 기능을 속도 지연 없이 운영하기 위한 최적의 인프라입니다.



All-in-One Turnkey Infrastructure

AI 가속기부터 RDMA 스위치, 노드 간 네트워크까지 통합된 인프라를 제공합니다. 검증된 하드웨어 및 소프트웨어 스택 기반으로 즉시 AI 상용 서비스에 투입할 수 있습니다. 대규모 인프라 도입을 위한 복잡성을 제거하고 운영 효율을 극대화하세요.



Ready-to-Deploy Rebellions Enterprise AI Solution

Pod에는 엔터프라이즈 환경에 최적화된 리벨리온의 SW Product, 'Enterprise AI Solution'을 도입할 수 있습니다. 기업 AI 서빙의 Full Lifecycle을 비용 효율적으로 지원하는, 이미 준비된 솔루션을 검토해보세요.

Chassis	42U
Server	8 servers
AI Accelerator	64x ATOM™-Max Cards
Management Network	1G UTP Switch
Storage Network	10G Optic Switch
RDMA Network	800G Data Switch
Power	4x redundant PDUs (2N redundancy)
Thermal	Air-Cooled

RBLN SDK

GPU의 익숙한 사용성을 제공하면서도, 차세대 AI 워크로드를 위해 설계된 Full-Stack Inference Platform을 제공합니다. PyTorch 개발부터 LLM 서빙과 배포까지, 모든 단계가 엔터프라이즈 환경에 맞춰 설계되었습니다.

Driver SDK

NPU 구동을 위한
기본 시스템 SW 및 도구 모음

- Firmware
- Kernel Driver
- User Mode Driver
- System Management Tool

NPU SDK

모델 및 서비스 개발을 위한 SW 도구 모음

- Compiler, Runtime, Profiler
- Hugging Face 지원
- 주요 추론 서버 지원
(vLLM, TorchServe, Triton Inference Server 등)

Model Zoo

리벨리온 NPU에서 곧바로 쓸 수 있는
300+ PyTorch와 TensorFlow 모델 제공

- Natural Language Processing
- Generative AI
- Speech Processing
- Computer Vision
- Physical AI

Cloud SDK

Cloud에서 NPU 자원 관리를 위한 SW 모음

- K8s Device Plugin
- Metric-Exporter
- Node Feature Discovery
- Device Installer
- VFIO Manager
- K8s Operator

엔터프라이즈 AI 서비스를 위한 Full Lifecycle 지원 솔루션

ATOM™-Max Pod에서는 엔터프라이즈 AI 서비스의 서빙 Full Lifecycle을 지원하는 Rebellions AI Serving Solution을 이용할 수 있습니다. 노드 단위 분산 서빙을 위한 개발 툴킷, 자동화된 AI 인프라 운영 도구, 여러 개발자의 독립적이고 손쉬운 개발 환경을 지원합니다.

Day 1 구축과 배포

